2018 | OriginalPaper | Buchkapitel

An Involuntary Data Extraction and Information Summarization Expending Ontology

verfasst von: R. Deepa, R. Manicka Chezian

Erschienen in: Artificial Intelligence and Evolutionary Computations in Engineering Systems

Verlag: Springer India

An Involuntary Data Extraction And Information Summarization Expending Ontology

R. Deepa[1], Dr. R Manicka Chezian[2]
[1]Research Scholar, NGM College Pollachi,
Coimbatore
India
deepaharini2015@gmail.com

[2]Associate Professor, NGM College Pollachi,
Coimbatore
India
Chezian_r@yahoo.co.in

Abstract: The World Wide Web is the repository for huge data that are the web pages. The web pages are acquired using a Query given by the user. The web Pages may sometimes be unstructured and unequal. The Main objective of the study is information extraction and summarization using Ontology. The System proposes a new method named as SSDO (Structural Semantic Domain Ontology) for effective information retrieval. The proposed system automatically extracts the unstructured information from the repository and stores it in the Search Buffer. The information Extraction will be performed using domain ontology.

The main disadvantage of the existing system is that the information which is extracted from the various sources will not be aligned properly. The system may fail to know where the exact information is located on the website. The current proposal overcomes the above problem by adopting the technologies which are named as pair alignment, top down alignment and loop structure algorithms. The proposed system will acquire things such as if the user need to know any data, then the user will type the detail known as a label. Then the web page will extract the information with a proper description and additional details.

Keywords: Domain Ontology, Search Buffer, Query, Extraction, Structure Algorithms.

I Introduction

This study focus on to overcome the existing approaches and it works a new method called Structural Semantic Domain Ontology (SSDO)[10] for extracting the regulated data from the web pages that are generated relevant to the users[1] query with the use of ontology. This performs unit level data extraction that successfully extracts data units and aligns them in an ordered format, finally summarizes the data by means of clustering. Auxiliary information which was eliminated in the existing method is maintained and stored by means of Ontology Domain Storage.

The main goal of data extraction is to extract SR (search result) from the query[13] result pages and line them in tabular format based on the criteria that the tag and value are similar to tables assigned to the data units. It aims to provide higher precision and extraction results compared with the existing state of the art methods.

This study aims to automatically extract wrappers, tags and values from a raw HTML file by introducing a semantic method called Structural Semantic Domain Ontology (SSDO), which automatically assigns meaningful labels to the data units in SRs. It has a main objective of finding relative labels for SR result set alignment.

- This Study provides a solution against three major problems in dealing of SR
- The query result records extracted from result pages may have lots of unwanted data.
- The web page extraction may vary based on the construction of web pages.
- Another problem is the alignment and summarization of the needed results into an ordered form.

The implementation of SSDO aims to provide optimal extraction, alignment and summarization of SR data attributes. The extraction time and memory usage can be minimized to enable the fully automatic extraction and summarization. The result pages have to be automatically obtained and the SRs need to be automatically extracted.

II Methodology

2.1 Structural Semantic Domain Ontology (SSDO)

The proposed system extracts and summarizes the SRs in web databases if there are at least two records in the page. The new technique is proposed to handle the case when there are not contiguous in the dataset, it extracts the available data and calculates priority and makes it outline for the fast summarization. The result shows that noncontiguous data region problem can be addressed easily due to the presence of auxiliary information such as a comment, recommendation, and advertisement. Data can be referenced further if needed. Existing DELTA eliminates the auxiliary information for data labeling. Compared with existing data extraction methods, SSDO improves data extraction accuracy in many ways as follows.

- A data region identification technique is proposed to identify the noncontiguous SRs that have the same parents according to their tag similarities followed by data region, merge that merges the data region that contains similar data records into one. Tag identification and classification method are used to identify similar tags [12].
- A new method is proposed to align the data values in the identified SRs first pair wise, then top down and finally prioritizing they can be put into a table with the

data values belonging to the same attribute arranged into the same table column. Both tag structure similarity and data value similarity are used in the pair wise alignment.

- Dissimilar existing nested structure processing algorithms that rely on only the tag information [12], but SSDO uses both tag and data value similarity information to improve the accuracy of Middle out ontology structure processing.
- This approach considers other important features shared between data units, such as their priority based information. Finally the system integrates the interface schema over multiple web databases with the common cluster to enhance data summarization.

2.2 Search Result Extraction

Search result records are obtained from the result pages by identifying the data regions in the web pages followed by merging the data regions that exist in several pages. For a query result page, the tag tree construction module constructs a tag tree for the page rooted in the HTML tag[2]. Every node within the tag tree represents a tag within the hypertext markup language page, its derivatives are tags authorized within it. Each internal node n of the tag tree includes a tag string (tsn) and a tag path (tpn) [4], which integrates the tags from the basis node to n. Next, information region identification module identifies all potential knowledge regions in websites that contain dynamically generated data, prime down ranging from the basis node to n. The information region merges section helps to merge totally different data regions that contain similar records into one. The record segmentation section then segments the known knowledge regions into knowledge records in step with the tag patterns within the knowledge reigns. Query result sets identification part helps to pick one among united region that contains the SR and eventually SR are extracted from this region [2]. Earlier works on webpage segmentation were on free texts and intended by raising performances of the knowledge retrieval. In info retrieval documents are extracted with the values of the linguistics of a web page to the queries [12][13][7][8].

2.3Middle out Ontology Structure Processing

Top down data value alignment constraints a data value in a SR to be aligned to at most one data value from another SR. If SR contains middle out ontology structure such that an attribute has several values, then certain of the values may not be associated with any other values [12].

The middle out ontology structure identification algorithm steps

Step1:

Identifies the nested column set C and so creates a replacement row for every combination of a repetitive part**.**

Step 2:

For all found SRs the tag tree for the question result page and also the SRs prime Down Aligned columns as input.

Step 3:

For each SR with record root node t in T, the procedure nest column determines is i

nvoked to spot any nested columns within the SR.

Step 4:

After all the nested columns are known, a new row is generated by bridge the remaining components additionally because the repetitive information worth.

Step 5:

Given n records with a most of m data values and a most tag string length of 1, the time complexity of the middle out ontology structure processing algorithm is O (nl2m2) [2].

Middle out Ontology Structure Processing Algorithm[2]

```
Procedure nets processing (SRs, T, Top Down Align)
C==Φ
for each SR with record root t
nest column identify (t,T, Top down align, C)
for each column pattern cp in C do
create a new row for each repeated subpart
Procedure nest processing (SRs, T, Top down align)
if(t contains more than one data value) then
for each child ti of  t do
nest column identify(ti, T, Top down align, C)
for each repetition p of any consecutive maximum repetitive tag pattern
    found in t's children
Cp= data columns for p in the Top down align
if cp Є C and nested (cp, Snest ) then
add nested column (cp, C)
 Function Boolean nested (cp, Snest)
Simintra== intra- column similarity within cp
Siminter==inter- column similarity with in cp
if(Siminter/Simintra>Snest) then
return true
else return false
Procedure add nested column (cp, C)
for each element ci in C do
if(cp ∩ ci ≠ ϕ) then
C==C –ci + cp U ci
break
if no element in C shares a common column with cp then  C == C+ cp
```

2.11 Data Summarization

In this data summarization technique clustering algorithms are used. The first identifies all information units within the SRs so organize them into totally different teams with every cluster equivalent to a distinct conception.

Grouping data units of the same semantic can help identify the common patterns and features among these data units. These common features are the basis of the labeling technique for summarization. A tag node corresponds to an HTML tag surrounded by "<" and ">" in HTML source, while text node is the text outside the "<" and ">".<b>stero</b>

2.12 Process included in Data Summarization

The proposed enhanced data alignment rule is predicated on the idea that attributes seem within the same order across all SRs on the identical result page though the SRs might contain completely different sets of attributes. The following are the steps involved in the enhanced alignment for data summarization [3].

Step 1:

Merge text nodes. This step detects and removes decorative tags from each search result record to allow the text nodes corresponding to the same attribute (separate by decorative tags) to be merged into a single text node.

Step 2:

Align the text nodes. This step aligns the text nodes into groups so that eventually each group contains the text nodes with the same concept (for atomic nodes) or the same set of concept. (For merged nodes)

Step 3:

Split composite text nodes. This step aims to split the values in composite text nodes into character data units. This step is passed out based on the text nodes in the same group Top Down. A group whose values need to be split is called composite group

Step 4:

Align data units. This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept [10].

Data Summarization Algorithm [10]

```
Grouping Data Units
Input: a set of query terms  T,  A data section  block B
Output: a set of data  Record R
1. set R, a set of leaf nodes Nl, a set of starting leaf nodes Ns, a set of data unit
groups G, a set of leaf nodes groups Gl and a set of horizontally expanded data
unit groups G all to {}
2. Add every text nodes in B to Nl
3. for every leaf node nl Є Nl do
4. if nl contains a query term t Є T then
5. Add nl to NS
6. remove nl from Nl
7. for every starting leaf node ns Є Ns do
8. set a data unit group g to {ns}
9. for every leaf node nl Є Nl do
10. if nl is horizontally aligned with ns then
```

```
Add g to G
repeat
remove a leaf node nl from Nl
set a leaf node group gl ={nl}
for each leaf node nl Є Nl do
if nl is horizontally aligned with nl then
add nl to gl
remove nl from Nl
add gl to Gl
until Nl= {}
repeat
remove a data unit group g from G
for each data unit group g Є G do
if g is horizontally aligned with g then
set g to g U g
remove g from G
add g to G
Until G={}
Repeat
Remove a horizontally expanded data unit group g from G
for each horizontally expanded data not group g Є G do
if g is vertically adjacent of g then
set g to g U g
Remove g from G
for each leaf node group gl Є Gl do
if gl vertically adjacent to g then
set g to g U gl
remove gl from Gl
Add g to R
Until G={}
Return to R
A set of query terms- as T, Data section block- as B, Set of data records- as R
```

III Results and Discussion

3.1Data Summarization

SR results will be summarized by highlighting the data attributes. Based on the details such as Author Edition, Publication Date, the links extract the necessary data. In case of mobile dataset, model, version, prices is summarized. In case of laptop, price, configuration, name, brand are described [2].

3.2 Performance evaluation

Performance evaluation of the proposed approach is conducted based on the working of SRs context scenario. Precision, Recall, and F1 Score plays an important role in the ontology based semantic web query matching performance.

$$\text{Precision}=\frac{tp}{tp+fp}\ . \tag{1}$$

$$\text{Recall }=\frac{tp}{tp+fp}\ . \tag{2}$$

$$\text{Accuracy}=\frac{tp+tn}{tp+tn+fp+fn}\ . \qquad (3)$$

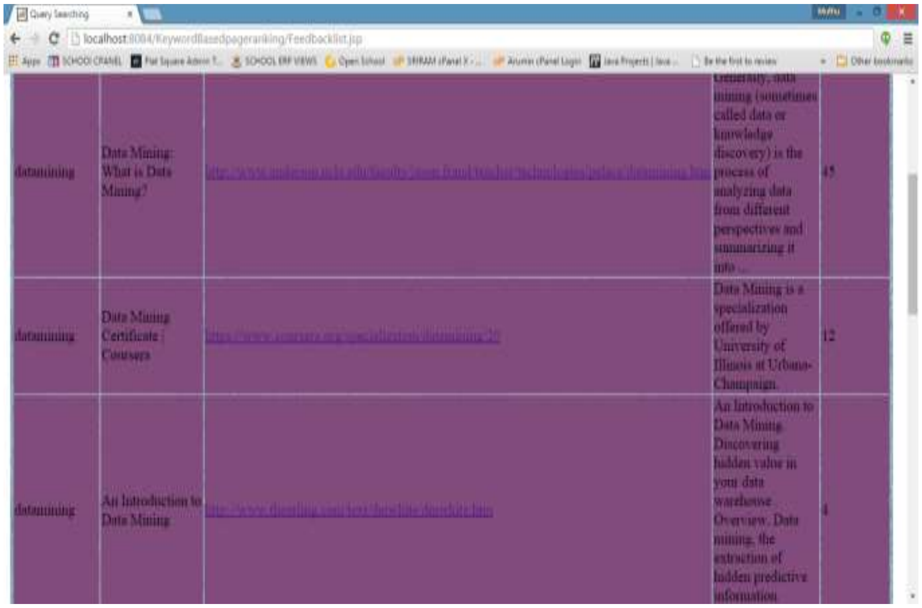
$$\text{F-Measure}=2.\frac{precision.recall}{prrecision+recall}\ . \qquad (4)$$

Where
tp – True Positive (Correct result)
tn – True Negative (Correct absence of result)
fp – False Positive (Unexpected Result)
fn – False Negative (Missing result)

Noncontiguous SR analysis compares the performance for query result pages during which the SRs area unit for contiguous and noncontiguous [10]. The fig 1 compares the contiguous and non contiguous search results for the existing DELTA and the proposed SSDO Methods. The Comparison notifies that SSDO performs better as it provides the SRs needed result.

Table 1. Search Result Extracted

	Website1		**Website 2**	
	DELTA	SSDO	DELTA	SSDO
Contiguous Extracted SRs	500	330	350	335
Non-Contiguous Extracted SRs	480	325	330	335



datamining	Data Mining: What is Data Mining?	http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm	Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into ...	45
datamining	Data Mining Certificate | Coursera	https://www.coursera.org/specializations/datamining?0	Data Mining is a specialization offered by University of Illinois at Urbana-Champaign.	12
datamining	An Introduction to Data Mining	http://www.thearling.com/text/dmwhite/dmwhite.htm	An Introduction to Data Mining Discovering hidden value in your data warehouse Overview. Data mining, the extraction of hidden predictive information	4

Fig 3. User Query based top web pages

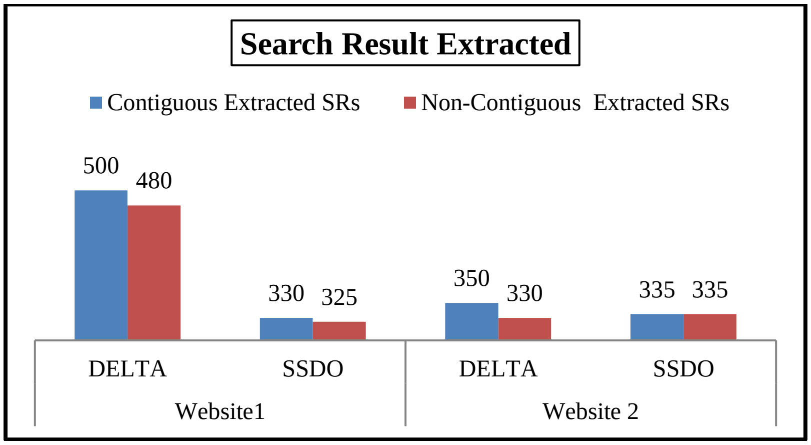


Fig 1. The extracted results and currently extracted SRs for DELTA and SSDO are compared

The simulation results for the evaluation of the proposed approach against the existing approach for various performance measures like Precision, Recall and F-Measure are shown in the Table 2. The results of the performance measure are plotted in Fig.2

Table 2 Performance Analysis

	Website1		**Website 2**	
	DELTA	SSDO	DELTA	SSDO

Record level Precision (%)	72.6	81	76.6	91.9
Record Level Recall (%)	83	89.5	86.3	96.9
Accuracy (%)	91.8	94.5	93.8	94.3
F-Measure	77.4	85	81	94.3

From Table 2 of the performance measures, it is seen that the accuracy of the proposed system is 94% which is high in comparison with the available approaches in the literature. Then SSDO shows the best recommendation to users. The recommendation result shown in fig 2.

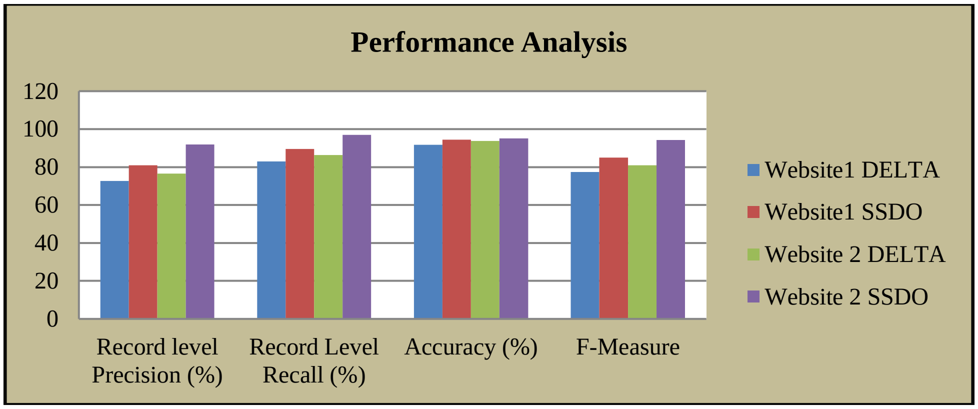


Fig 2. Performance Analysis Measure for the existing DELTA and the proposed SSDO for the factors Precision, Recall, Accuracy, F-Measure

IV Conclusion

A new design of query based recommendation system based on improved SSDO (Strutural Semantic Domain Ontology) with nested structure and data alignment using improved clustering Algorithms has been proposed for the implementation of an efficient web search. The salient feature of this proposed approach is that it is based on semantic domain ontology, which decides the relevance between keyword and user query words. SSDO first discovers the data regions from multiple pages and merges the data region that contains similar data records. Finally, it aligns the data values in SR by the following methods: Pairwise, Top down, Priority basis and Middle out Ontology structure processing. SRs with a similar tag and value will be stored in the tables. The Ontology Domain Storage technique is used to store the auxiliary information.

SSDO extracts desired data from various SR pages. The experiments on many collections of SR, drawn from several well-known knowledge wealthy sites, this means that SSDO is extraordinarily smart in extracting and orienting the info from the online page sources. The opposite fascinating feature of the planned system is that's doesn't complexes fail to extract any knowledge, even once a number of the idea created by non mandatory tag aren't met by the input assortment. SSDO inclined to offer higher accuracy compared with the prevailing ways.

References

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121.

2. Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997) 415–438.

3. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995).

4. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996).

1. AravindArasu, Hector Garcia-Molina, Extracting Structured Data from Web Pages, SIGMOD 2003, June 9-12, 2003, pp. 337-348, San Diego, CA.

2. Deepika.J, Non-Duplicate Data Extraction in Web Databases by Combining Tag and Value Similarity, International Journal of Advanced Information Science and Technology (IJAIST), ISSN: 2319:2682 Vol.9, No.9, pp. 16-22, January 2013.

3. Eduardo J. Ruiz, VagelisHristidis, and Panagiotis G. Ipeirotis, Facilitating Document Annotation Using Content and Querying Value, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 2, pp. 336-349, FEBRUARY 2014.

4. M. Jude Victor, D. John Aravindhar, V. Dheepa, Web Data Extraction and Alignment , International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, pp0 129-132, Volume 2 Issue 3, March 2013.

5. KeRui Chen , Fan Zhang , Feng Lin He, Extracting Data Records Based on Global Schema, pp. 553-558, Applied Mechanics and Materials (/AMM), Volumes 20-23.

6. Lidong Bing Wai Lam Yuan Gu, Towards a Unified Solution: Data Record Region Detection and Segmentation, CIKM’11, October 24–28, Glasgow, Scotland, UK, 2011.

7. K. Manonmani, M.Kalidass, AUTOMATED DATA EXTRACTION AND ARRANGEMENT USING SEGENTATION BASED TAG AND VALUE RESEMBLANCE ANALYSIS, International Journal of Computer Science and Management Research , Vol 2 Issue 4 April 2013, pp. 2211-2216, ISSN 2278-733X.

8. Miguel Gomes da Costa JúniorZhiguo, Web Structure Mining: An Introduction, Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.

9. SILA: a Spatial Instance Learning Approach for Deep Web Pages Oro Ermelinda, Massimo Ruffolo Technical Report ICAR-CNR, WOODSTOCK ’97 El Paso, Texas USA Copyright 20XX ACM X- XXXXX-XX-X/XX/XX.

10. A Suresh Babu, Dr. P. Premchand and Dr. A. Govardhan, Record-Level Information Extraction from a Web Page based on Visual Features, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2, pp0 99-105, ISSN 2249-6343.

11. Mr. Vinod Kumar Raavi, Satya P Kumar Somayajula , Automatic Template Extraction from Heterogeneous Web Pages, International Journal of Advanced Research in Computer Science and Software Engineering,Volume 2, Issue 8, August 2012, pp. 408-418, ISSN: 2277 128X,.

12. 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, pp0 1186-1200, JULY 2012.

13. WEIFENG SU, JIYING WANG, FREDERICK H. LOCHOVSKY, ODE: Ontology-Assisted Data Extraction, ACM Transactions on Database Systems, Vol. 34, No. 2, Article 12, Publication date: June 2009.