http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (22): 6710-6716, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.6710.6716

A Domain Web Data Standardization Organization Method

[1]Cao Rui, [2,3]Wang Rui,[1]Hao Li-Yun and [1]Wu Ling-Da
[1]Science and Technology on Complex Electronic System Simulation Laboratory,
Academy of Equipment, Beijing, 101416, China
[2]Department of automatic control and systems engineering, University of Sheffield,
Sheffield, S1 4DT, UK
[3]Department of Systems Engineering, College of Information Systems and Management,
National University of Defense Technology, Changsha, 410073, China

Abstract: In order to enable the Web data to be applied in a non-Internet environment, overcoming the timeliness of Web data, this study proposes a domain Web data standardization organization method. The domain Web standardization organization framework is established and the acquired data are divided into three categories: Structured, un-structured and semi-structured. With respect to the semi-structured data within an implicit scheme, we use a rational code design to transform semi-structured data into structured data. Combining file system and relational database, a standardization organization method is established for the three types of data. Experimental results show that this method is effective and efficient.

Key words: Web data, domain, semi-structured data, unstructured data, organization

INTRODUCTION

With the development of Internet technology, the information amount from the Web is more than hundreds of thousands of TB. Although the electronic commerce is the main driving force of the Web, the Web is not limited to this and oriented to various subject domains in the real world (Hicks *et al*., 2012; He *et al*., 2007), who has gradually developed into a huge, global information center (Marin-Castro *et al*., 2011). The Web, as an Internet information distribution platform, mainly distributes data in the form of Webpage and the Web data is very valuable in many fields. Due to the website in the internet is developed by different individuals around the world, it leads to that the Web data is heterogeneous, autonomous, distributed and domain diverse; which increases the difficulty of Web data application. Thus, extensive research of the Web is carried out in order to acquire useful information (Liu *et al*., 2007).

Due to the individual differences between Web data providers, there is a lot of timeliness in Web data. Once the Web data providers terminated the service, the users could not acquire the required data. On the other hand, to the Web data users who are not in the Internet environment, it is needed to organize web data reasonably so as to meet the applications requirements in non-Internet environment. In view of above reasons, a domain Web data standardization organization method is proposed. According to the main existence format of web data, the valuable data acquired from the Web data providers are divided into three categories; in the light of each data category characteristics, the standardization organization method is established. The study is a preliminary exploration of domain Web data standardization organizations and provides strong supports for the wider application of Web data.

DOMAIN WEB DATA STANDARDIZATION ORGANIZATION FRAMEWORK

The information in the Web can be divided into two parts which are the Surface Web and Deep Web, according to the "depth" of the information (Bergman, 2001). They focus on contents of different levels and forms (Tripathy *et al*., 2012): The Surface Web data mainly exist in the format of semi-structured and Deep Web data mainly exist in the format of structured and unstructured (Zheng *et al*., 2005). Therefore, in order to standardizing organization of the acquired valuable data, the domain Web data can be divided into different categories to dispose according to the Web data main existence format. Domain Web data standardization organization framework is shown in Fig. 1.

Corresponding Author: Cao Rui. Science and Technology on Complex Electronic System Simulation Laboratory, AcademyofEquipment, Beijing, 101416, China

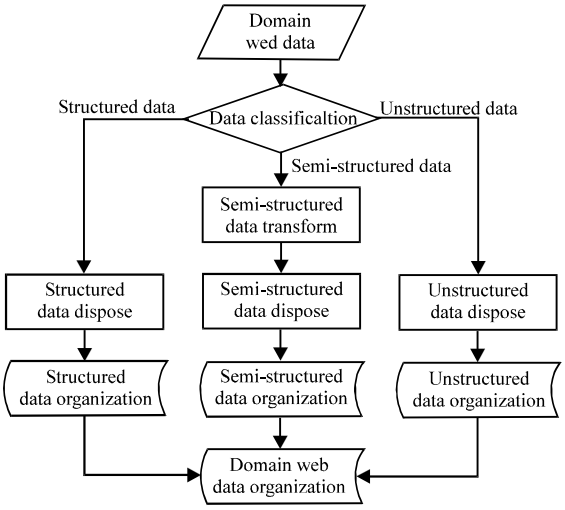

Fig. 1: Domain web data standardization organization framework

First, domain Web data classification. According to the existence format, the Web data are divided into structured data, unstructured data and semi-structured data three categories.

Domain structured data. If the users have a clear and definite understanding of data structure and types, at the time of submitting the Web data request, the acquired Web data will possess good relation schemata and these data which are made up of the basic type data (such as integer, float, string, etc.), are classified as domain structured data category.

Domain unstructured data. If the users have a clear and definite understanding of data types, at the time of submitting the Web data request; some data are too large (such as audio, video), or their structure are uncertain (such as the file who possess clear input and output but unclear internal structure), or their structure are too complex to split (such as multi-dimensional data files) and these data are classified as domain unstructured data category.

Domain semi-structured data. The data, without mandatory schema constraints and fixed structure, are defined as semi-structured data. It is an intermediate category between structured and unstructured (Buneman, 1997). Although it possesses some structure in a certain degree extent, it owns implicit schema, dynamic structure and relaxed type constraint. These acquired data are classified as domain semi-structured data.

Second, transforming the semi-structured data to structured data. Due to the flexible schema, semi-structured data can meet the needs of the Web data exchange in the distributed heterogeneous environment. But disposing those indeterminate schema data brings great difficulties to the traditional database and reduces the efficiency of data dispose. So, it is necessary to transform the unfixed structure data to the structured data who owns good schema constrains; the relational database, whose theory and technology are mature, is taken advantage of to organize semi-structured Web data and plays an obvious role in the data organization and management.

Third, taking the different data dispose approach, according to the domain Web data category.

Domain structured data dispose. The structured data owns good relationship schema and users are clear about data structure and types, so the relational database, whose tables are established according to the known relation schema, is made sufficient use of to store and manage these Web data since the basic storage unit of relational database is highly structured.

Domain unstructured data dispose. These unstructured data which owns explicit types and is not suitable to be stored in relational database table, can be stored in the file system. Their explicit schema information is used to establish the corresponding index mechanism and these Web data are managed by the file system and the relational database together.

Domain semi-structured data dispose. The semi-structured data can be stored in the definite schema relational database tables, after transforming the semi-structured data to structured data; these Web data are managed by the relational database.

Finally, Synthesizing data dispose results of the three categories to complete the Web data organization. Although the Web data amount is very huge, the amount of data which users really concerned about is limited with regard to specific domain, Through the correlation analysis between dispose results of structured data, unstructured and semi-structured data, the structure of relational database tables is optimized and the domain Web data are organized by relational data base and file system standard and orderly.

Based on the above data standardization organizing framework, domain Web data standardization storage is realized and the advantages of the mature relational database, such as integrity, security, concurrency control and recovery technology, can be take use to manage data effectively. Due to the structured data is highly compatible with the basic storage unit of relational database and the method, that using relational data table

to organize structured data, is relatively mature; the organization methods of unstructured data and semi-structured data in domain Web data standardization organization framework are the main research content.

DOMAIN UNSTRUCTURED DATA ORGANIZATION

For the domain unstructured data, storing them in the file system is not enough and standardization organizing these data for serving the applications is also need to be considered. Based on the domain web data standardization organization framework, the unstructured data is stored in the file system and, according to a certain standard; the index of unstructured data is established by the relational database which takes use of unstructured data basic information. The unstructured data is organized combining with the file system and relational database. As shown in the Fig. 2.

In the relational database, the basic information table is the core of standardization managing the domain unstructured data. It mainly consists of two parts, path information and description information. Path information includes data acquisition date, type and name. Description information is mainly used to explain the basic information of the unstructured data, such as the space and the application direction. The content of describe information should be established in line with specific application requirements. In addition, with the basic information table and other information description table (as the table of types shown in Fig. 2), the relational database can overall describe all kinds of information included in the unstructured data, in order to meet specific application requirements.

As to organizing domain unstructured data, the first step is to organize the basic information of the unstructured data in the relational database. And then, establishing multi-level index of unstructured data through the path information of the basic information table and building the map between basic information of unstructured data in the relational database and the storage location in the file system. At last, storing the unstructured data at the corresponding position in the file system where maps to the path information. As shown in Fig. 2, if the path information of unstructured data is "Date: March 2nd, 2014; Type: avi; Name: Name", then its location in the file system is" (File System Root Path)\2014\Mar\3nd\avi". According to this way, on the one hand, the file system and the relational database are combined together, to build the standard of organizing domain unstructured data. On the other hand,

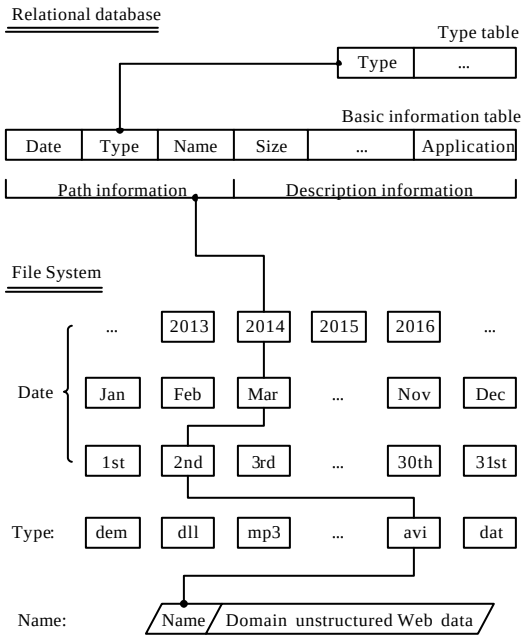

Fig. 2: The domain unstructured data organization

organizing structured, unstructured and semi-structured data in the same relation database is achieved.

DOMAIN SEMI-STRUCTURED DATA ORGANIZATION

As the latest standard of internet data organization and exchange, XML has become the most important source Fof semi-structured data. As it were, XML is semi-structured data of the internet. So, for the domain semi-structured data XML, designing code rational and transforming the schema indeterminate XML to structured data which is organized and managed by relational database, is the key of standardization organization.

The concept of vector in mechanics and analytic geometry is introduced to the semi-structured XML data coding. Coding scheme identifies the order of nodes with vector gradients and records the nodes path information to support the string pattern matching.

Mathematical model: All the used vectors are located in the fourth quadrant of two-dimensional Descartes coordinates and let v = (x, y) denote vector.

Definition 1: Gradient, the gradient of vector v is:

$$G(v) = -y/x \quad (1)$$

Theorem 1: Suppose that a = (x_1, y_1), b = (x_2, y_2) denote two vectors and their gradients are G (a) and G (b). If $y_1/x_1 > y_2/x_2$, then G (a)<G (b).

Theorem 2: Suppose vector c = (x_3, y_3), for the vector a and b in the theorem 1, if c = a+ b, as shown in Fig .3, G (c) = $-y_3/x_3 = (y_1+y_2)/(x_1+x_2)$, then G (a)<G (c)<G (b).

Proof: Theorem 1 states that $y_1/x_1 > y_2/x_2$, so:

$$G(c) = -\frac{y_1 + y_2}{x_1 + x_2} = -\frac{(y_1 + y_2)/x_1}{(x_1 + x_2)/x_1} < -\frac{y_2/x_2 + y_2/x_1}{(x_1 + x_2)/x_1} = -\frac{y_2}{x_2} = G(B) \quad (2)$$

G (a)<G (c) is similarity and then G (a)<G (c)<G (b).

Corollary 1: The vectors *a* and *b*, via linear operating, can get the vector 2a+b, a+b, a+2b. According to theorem 2, their gradients are in the gradient interval constituted by G (a) and G (b).

Corollary 2: If the gradient interval of G (c) is (G (a), G (b), according to the corollary 1, the gradients of 2a+b, a+b, a+2b are also in the gradient interval of G (c).

Corollary 3: For the gradients in corollary 2, the gradients difference between 2a+b, a+b is G (a), while the gradients difference between a+b, a+2b is G (b). Thus the gradients difference between 2a+b, a+b is smaller and it is in the gradient interval of G (c).

Data model: The basic idea of modeling is to convert semi- structured XML data to tree graphs, making the basic elements of XML data as nodes and the relationship between the elements as edges in the tree graph and establishing the node data model. Without loss of generality, four types of nodes are considered and they are element, attribute, text and empty element. The modeling principles in the study can apply for the other types of elements, such as annotation, namespace. For each node in the data model, adopting the non schema mapping method and the fixed relation schema, the basic information is recorded and structural information is reflected in the relation schema.

Applying the mathematical model established, starting from the origin of coordinates O, a series of vector are used to identify the elements order information of XML data, as shown in Fig. 4; then every element of XML data can find it`s unique identifier vector from mathematical model established. It is noticed that elements and their attributes are in the same tag, while they are two types of nodes in the mathematical model which is needed to be identified by two vectors. As

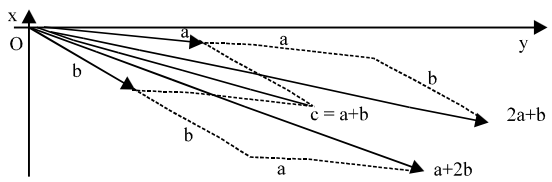

Fig. 3: The linear operation of plane vectors

Table 1: Start-end gradients of nodes

Order	Start vector	End vector	Identification gradient	End gradient
1	(1,0)	(0,1)	0	+8
2	(5,1)	(3,2)	0.2	0.6667
3	(4,1)	(3,1)	0.25	0.3333
4	(5,2)	(5,3)	0.4	0.6
5	(2,1)	(7,4)	0.5	0.5714
6	(4,3)	(1,4)	0.75	4
7	(1,1)	(5,8)	1	1.6
8	(3,4)	(2,3)	1.3333	1.5
9	(3,5)	(1,3)	1.6667	3
10	(1,2)	(2,5)	2	2.5

shown in Fig. 4, No. 3 and No. 10 are attribute nodes and the dotted arrows are their marking vector sketch.

Through the preorder traversal, the node model of XML data is established and the start-end code of every node is recorded (Zhang *et al*., 2001). The method of reference (Xu *et al*., 2007) applies vector encoding into start-end code and gets the start-end vector of each node, as shown in Fig. 5. The start vector is taken as the identification vector, while the gradients of each start vector are calculated by definition 1. Then, the mixed code of each node are make up by identification gradient (start gradient), termination gradient and path information. For the path information of each node, as shown in Table 1, the path code of node is separated by "#/" instead of "/", in order to avoid the query error of the direct use of SQL pattern matching function. For example, in the query of ancestor-child nodes that "AF" and "name", regular expression "#/AF#%/name" can match the right results; if "/" as the separator, the regular expression "/AF% /name" will mismatch "/AFS/name" as the query results.

Relationship schema: According to node data model, relationship schema is designed to store nodes and structural information between nodes and the data of each node is mapped and recorded. Relationship schemas is constituted as follows>

Identification gradient (Gradient), end gradient (End), path code (Path), node type (Type), extension (Extend), node value (Value).

Among them, the extension (Extend) is mainly used for recording attributes, namespace and such similar types

Table 2: Hybrid code tables

Gradient	End	Path	Type	Extend	Value
0	99999	#/site	1		site
0.2	0.6667	#/site#/AS	1		AS
0.25	0.3333	#/site#/AS	3	id	a1
0.4	0.6	#/site#/AS#/name	1		name
0.5	0.5714	#/site#/AS#/name	2		China
0.75	4	#/site#/AF	1		AF
1	1.6	#/site#/AF#/name	1		name
1.3333	1.5	#/site#/AF#/name	2		Congo
1.6667	3	#/site#/AF#/supply	4		supply
2	2.5	#/site#/AF#/supply	3	id	f1

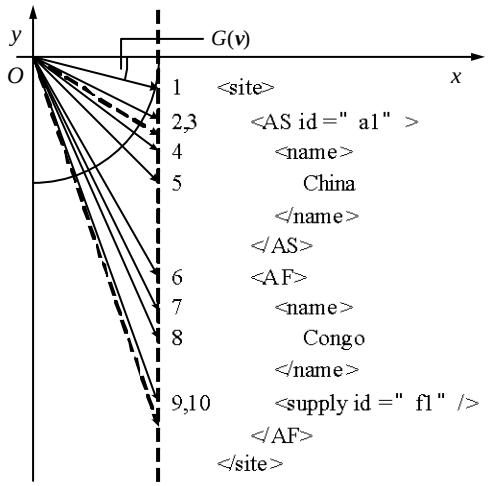

Fig. 4: Mappings between nodes order and identification vectors

of element node. For example, when the node Type (Type) is an attribute node, extension record the node attribute name and attribute values are recorded by node Value (Value).

The relationship schema ensures the integrity of the data, that the original XML data information will not be lost in data reconstruction; supports the query based on the content and structure and is stored by one basic table, to a certain extent, it avoids multi-table connection operations. Especially the design relationship schema is not depended on DTD of XML data which is universal applicability.

According to standardization organization methods, semi-structured XML data can be stored in definite schema relationship database tables. For example, the dispose results of semi-structured XML data in Fig. 4 is shown in the Table 2.

It is necessary to state that, the main benefits for taking the start-end vector gradient as start-end code is that: First, it is data reconstruction oriented and can obtain tuples according to the semi-structured XML data reconstruction nodes order by simply ascending sort identification gradients. Second, infinite number of new

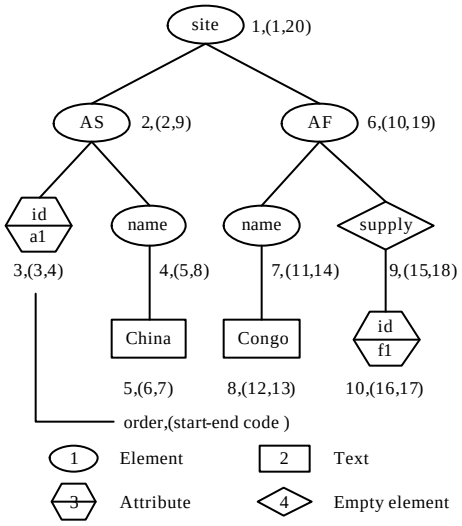

Fig. 5: Node data model

vectors can be inserted in vectors, so the dynamic performance of hybrid code is good and supporting node update operations; Third, it is more reasonable to allocate each node code in theory: on the one hand, start and end gradients offer different gradient interval for different depth nodes substantially, on the other hand, according to the Corollary 3, the update of insert node occupies less gradient range which can effectively avoid the fast increase of start-end code range. As shown in Table 2, the end gradient of No.6 node is 4 which has a higher utilization of gradient interval. In a word, this semi-structured organization method combines identity vector and path information, meets the two basic requirements that keeping structure relations and order; meanwhile it can support data query, update and reconstruction well.

EXPERIMENT RESULT AND ANALYSIS

The prototype system was implemented with C++ QT 4.0, the experiment environment was: CPU Intel Pentium IV3.0 GHz; 1 GB memory; Windows XP operating system; Microsoft SQL Server 2005 relational database.

Due to the space limited, here semi-structured XML data was taken as the experiment object and two kinds of experiment was represented, to analysis the standardization organization method feasibility and validity. QtXml module was taken as XML parsing interface, the experimental data adopted the XMark test

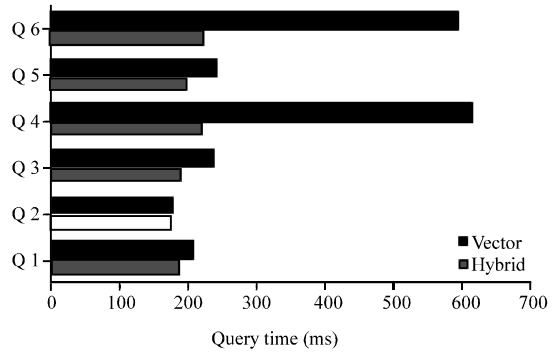

Fig. 6: Comparison of query times

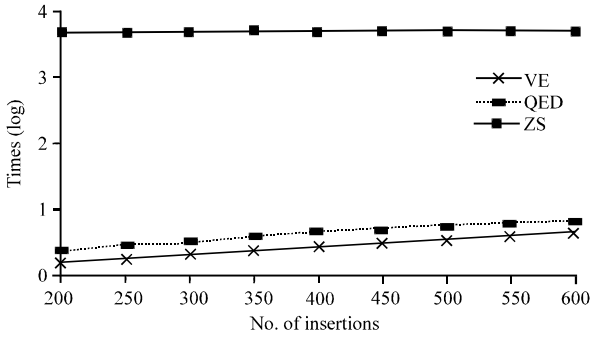

Fig. 7: Comparison of update times

Table 3: The query set

Query ID	Query
Q1	/site//mailbox//mail
Q2	//item//mail
Q3	//item/mailbox/mail
Q4	//item/mailbox/mail/text
Q5	//item[./mailbox]//mail
Q6	//item[./mailbox]//mail/text

sets (CWI, 2002) and the parameter was set to 0.005 which generated 567 KB test data and the maximum depth was 12.

Data query: Six groups of path query experiment, as shown in Table 3, were carried out with the original vector code (Xu *et al*., 2007) and our hybrid code, respectively. The results are shown in Fig. 6,

All the experiments were repeated ten times. The result is the average value excludes the maximum and the minimum.

It is obviously that our code query time is less than the original vector code except the query time is the same for the second query set. It is because that the hybrid code recorded the path of node information, effectively avoids the influences of time-consuming operation, such as join query, intermediate result sets generating.

The second set of experiment belongs to the structure relation judgment of ancestor axis or descendant axis. The vector code query, using principle of start-end code, consumed 175 ms; if the hybrid code using path information to query, the time-consuming was 183 ms which was longer than the time consuming of vector code. As the start-end code was suitable to this kind of path query, so the hybrid code adopted the start-end gradient to query which was the same principle with start-end code and abandoned querying with path information. So the time- consuming was the same with vector original code.

Data update: The test data was same with that in last section. The same group of test sets was coded by three different methods, that traditional region coding (Zhang *et al*., 2001), QED coding which supports dynamic update (Li and Ling, 2005) and our hybrid coding. 9 groups of test sets were randomly inserted nodes; inserted nodes sequence was 200, 250 … 600. All the tests were repeated three times, the result was the average value, as shown in Fig. 7.

The x-axis in Fig. 7 showed the inserted nodes number of each experimental group. The y-axis was the time for updating the node coding in logarithmic format and its base was 10. The basic unit of node update time was millisecond. It is obviously that, the update time of hybrid coding and QED coding were several order of magnitude less than traditional region coding; and the hybrid coding update efficiency was superior to that of QED. This is because that the hybrid coding and QED coding supported coding dynamic update, avoiding the time-consuming of whole data recoding. In addition, the essence of QED coding was supporting node updates by four numbers and Lexicographical Order and with the increasing of inserted nodes, calculation amount and length of the coding were increasing fast; the time-consuming was more than that of hybrid coding.

The two kinds of experimental results show that the hybrid coding absorbs the advantages of the vector coding and path coding; on the one hand, it provides various query methods for a same query path and owns query efficiency, adopting the less time-consuming query method according to the query path; on the other hand, due to the low coding complexity, it can support data dynamic update well. All these state that the standardization organization method in this study is effective and efficient.

CONCLUSION

According to the main existence format of web data, a domain Web data standardization organization method

is proposed. Experiment results and analysis show that, comparing with traditional methods, the method of this study can get good results and it is feasible. Web data standardization organization method was discussed in the study which can provide technology support to relevant research. The distributed system technology will be introduced to organize Web data next step.

ACKNOWLEDGMENT

The study thanks for the support by Natural Science Foundation of China under the Grant No. 61202129.

REFERENCES

Bergman, M.K., 2001. White paper: The deep web: Surfacing hidden value. J. Electron. Publishing, Vol. 7. 10.3998/3336451.0007.104

Buneman, P., 1997. Semistructured data. Proceeding of ACM Symposium on Principles of Database Systems, May 11-15, 1997, Tucson, AZ, USA., pp: 117-121.

CWI, 2002. XMark-An XML benchmark project. http://www.xml-benchmark.org/downloads.html.

He, B., M. Patel, Z. Zhang and C.K. Chang, 2007. Accessing the deep web: A survey. Commun. ACM, 50: 94-101.

Hicks, C., M. Scheffer, A.H.H. Ngu and Q.Z. Sheng, 2012. Discovery and cataloging of deep web sources. Proceedings of the 13th IEEE International Conference on Information Reuse and Integration, August 8-10, 2012, Las Vegas, NV., pp: 224-230.

Li, C.Q. and T.W. Ling, 2005. QED: A novel quaternary encoding to completely avoid re-labeling in XML updates. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 31 October-5 November, 2005, Bremen, Germany, pp: 501-508.

Liu, W., X.F. Meng and W.Y. Meng, 2007. A survey of deep web data integration. China J. Comput., 30: 1475-1489.

Marin-Castro, H.M., V.J. Sosa-Sosa and I. Lopez-Arevalo, 2011. A strategy for identi?cation of web query interfaces using supervised learning. Proceedings of the 7th International Conference on Next Generation Web Services Practices, October 19-21, 2011, Salamanca, pp: 233-237.

Tripathy, A.K., N. Joshi, S. Thomas, S. Shetty and N. Thomas, 2012. VEDD-a visual wrapper for extraction of data using DOM tree. Proceedings of the International Conference on Communication, Information and Computing Technology, October 19-20, 2012, Mumbai, pp: 1-6.

Xu, L., Z.F. Bao and T.W. Ling, 2007. A dynamic labeling scheme using vectors. Proceedings of the 18th International Conference on Database and Expert Systems Applications, September 3-7, 2007, Regensburg, Germany, pp: 130-140.

Zhang, C., J. Naughton, D. DeWitt, Q. Luo and G. Lohman, 2001. On supporting containment queries in relational database management systems. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, May 21-24, 2001, California, USA., pp: 425-436.

Zheng, D.D., P.P. Zhao and Z.M. Cui, 2005. On the research and design of deep web crawler. J. Tstingh Univ., 45: 1896-1902.