International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 11, November 2014 **ISSN 2319 - 4847**

Logical Framing of Query Interface to refine Divulging Deep Web Data

Dr. Brijesh Khandelwal[1], Dr. S. Q. Abbas[2]

[1]Research Scholar, Shri Venkateshwara University, Merut, UP., India

[2]Research Supervisor, Shri Vinkateshwara University, Merut, U.P. India

Director, Ambalika Institute of Management & Technology, Lucknow, U.P.

ABSTRACT

This work is an extension to research works, being carried out entitled "Study of Information Retrieval Performance Refinement in Deep Web Mining". It is an effort to study overcoming the limitations of the existing Deep Web Information Retrieval Systems prevailing lack of logics, in my belief, regarding the retrieval process. Addressing this problem from the information providing services perspective indicates the significant potential of additional Logical Framings provided by websites. Web query interfaces, the interfaces to the majority of available information on the Deep Web, are interpreted as Semantic Deep Web Services [1]. Our introduction of a Logical Deep Web Data Retrieval Services (LDWDRS) framing carries to great potential for Information Retrieval services is based on the large variety of information available on the Deep Web. In this paper we report some of our analysis in framing our system and outline the main challenges we see in the further exploration and divulging deep-web content.

Keywords:- Logical Framing, Deep Web, Logical Deep Web Data Retrieval Service (LDWDRS), Web Query Interface, Structured Interface

1. INTRODUCTION

As we know Deep Web refers to content hidden behind HTML forms, any attempt to get to such content, a user has to perform a form submission with valid input values. The name Deep Web arises from the fact that such content was believed to be beyond the reach of search engines. The Deep Web is also believed to be the biggest source of structured data on the Web and hence accessing its contents has been a long standing challenge in the data management community [2]. An incessant increasing amount of content on the deep web is not directly accessible and/ or indexable by search engines. The content might, for example, be hidden in non-public, inaccessible areas or might be stored in backend databases and therefore only accessible through structured web query interfaces. This part of the web is known as the Deep Web (or Hidden Web) in contrast to the Surface Web which can be easily accessed and indexed by common search engines [3]. The Surface Web consists of mostly static content, which is directly inter-linked with static hyperlinks. "Search engines rely on hyperlinks to discover new web pages [11], but static websites are outnumbered by dynamic websites on an extremely large scale and the web has been rapidly deepened [8]. The content as part of dynamic websites are mostly not accessible through static hyperlinks, as these content are dynamically enwrapped into web pages as the response to a structured query submitted through a web query interface. These are intended to be framed by human users to retrieve content from a background database often containing highly relevant content of a specific domain. Common search engines do not reach this part of the web. This is caused by the fact that search engines "typically lack the ability to perform form submissions" [1]. The benefit from the automatic discovery of new knowledge from existing information on the web is depending on an excellent Information Retrieval. As the retrieval of information from the Deep Web is still limited, Knowledge Discovery services are also still limited in their potential. Therefore, more efficient and targeted retrieval mechanisms for the Deep Web are needed to achieve full potential of Knowledge Discovery services [1]. The conceptualization of Logical Framings for information on the web may play a significant role "to absorb information from multiple knowledge sources". This hypothesis can be worked upon, and may be resulting in standards like Resource Description Framework in attributes (RDFa) and Micro data markups like schema.org initiated by the search engine big players Bing, Google, Yahoo, etc.. Therefore, this paper addresses the improvement of accessing this logically framed content on the Deep Web.

2. REVIEW LITERATURE

The retrieval and indexing of Deep Web content have been addressed from different perspectives in the past. The effort has mostly focused specific applications to discover, retrieve and index structured data from the Deep Web. We can always describe the different types of structured data in the context of the varying search tasks that we can strive to support over them. This includes special emphasis on the automatic web query interface interpretation. Common

approaches focusing on exposing Deep Web content can be classified to surfacing and virtual integration approaches. oth of these approaches were found significant in discovering, retrieving and indexing structured data from the deep web. The surfacing approach focuses a search engine initiated process to index the search result pages for pre-computed (randomized) queries to discover Deep Web content on large scale [2]. The virtual integration approach follows the data integration paradigm, using a mediator system to map queries to relevant web query interfaces [2]. The content, that is retrieved, is brought to the user by the virtual integration to the search result page. Both of these approaches have been approved as useful in some cases. But in general the virtual integration approach is related to a lot of manual effort setting up query mapping rules for each Deep Web query interface in the mediator system. Furthermore, the surfacing approach is too imprecise and additionally not scalable regarding the pre-computation of queries for domain independent sets of deep web contents. Regarding the discovery and cataloging of Deep Web sources Hicks et al. [9] highlight the challenges and demonstrate via prototype implementation, that their Deep Web discovery framework can achieve high precision using domain dependent knowledge for probing web query interfaces. Wenye et al. [13] focus "Manufacturing Deep Web Service Management [...] [by] Exploring Logical Web Technologies" by logically framing the Deep Web Services to reflect their hidden, dynamic, and heterogeneous contents while the relevance of Logical Framings for the Deep Web has already been identified in 2003 by Handschuh et al. [7]. Whereas these publications as well as Chun et al. [5] discuss theses challenges from the information retrieving services. The structured data we find on the Web and find that it comes mainly in two forms: one where the data itself is already structured in the form of tables, and the second where the query interface to the data is structured. The first type of data was analyzed by the WebTables Project [14] that collected all the HTML tables on the Web, extracted the ones that offer high-quality data, and offered a search interface over them. The second kind of structured data is available through structured query interfaces. HTML-form interfaces enable users to pose specific queries in particular domains and obtain results embedded in HTML pages. However, these results are not as structured as the HTML tables mentioned in the first type [2]. Furche et al. [6] introduced a promising automated form understanding ontology based approach, which is far beyond heuristics to fill out search forms [12], combining signals from the text, structure, and visual rendering of a web page. But according to Li, Xian et al. in "Truth Finding on the Deep Web: Is the Problem Solved?" [10], the challenges arising from the Deep Web are regarded as not yet solved. In general, current approaches are still limited either in being domain specific or limited in their efficiency. Deep Web Information Retrieval problem especially applies to the retrieval of dynamic content. Therefore, it seems to be unlikely to improve retrieval and indexing mechanisms towards reaching cent percent coverage of all available Deep Web content. At the same time, this is not the focus of our research too. Alternatively, we can focus on the reduction of manual effort regarding the query mapping on the one hand and more precise logical and structured query interface generation or pre-computation for the targeted retrieval from deep web; naming as Logical Deep Web Data Retrieval System (LDWDRS). Therefore, this paper is intended to improve access to Deep Web content by providing substantial analysis for refined Information Retrieval mechanisms and for the significant improvement of previously existing mechanisms.

3. OUR PROPOSAL

3.1. Theme

With Deep Web Content Retrieval we can always have little use of logics of the forms (structured interface), we are crawling. In fact semantics (or logics) plays a significant role in the latter virtual-integration approach to retrieve data from deep web. Our theme is to work around the logical analysis which can extend the coverage of deep web surfacing. With step forward towards a Logical Deep Web, which is the hypothetical long-term objective, it is essential to focus on additional research questions resulting from previously identified limitations. For the targeted data retrieval especially of dynamic Deep Web content, the need of an efficient and automatic approach is mandatory. Therefore, the attention needs to be set to these challenges: content providing service Detection, Incantation & Implementation and the Composition. By meeting these challenges we will ensure the identification of appropriate web query interfaces providing access to relevant content (Detection), the appropriate query mapping and query submission (Incantation & Implementation) and the service interoperability (Composition) as described in [1]. Common approaches for Deep Web Data Retrieval focus these challenges from the data retrieving services perspective. The conceptual idea being introduced in this section focuses these challenges from the information providing services viewpoint.

3.2. Logical Framing

Common Logical Framing standards like RDFa and schema.org micro data address particularly the framing of web content and do not have means for the existing lack of logics at the crucial point of the Deep Web Data Retrieval process. This crucial point is regarding the structured web query interfaces. To improve common crawling, indexing and content retrieval mechanisms and to ensure refined mechanisms, a logical framing for structured query interfaces can be recommended (Figure-1). This will reprocess the query interfaces originally proposed for human users in a combined computer and human readable arrangement. The abstract concept, to describe recommended structured query interfaces in a computer readable format only, is derived from the Logical Framing of Deep Web Data Retrieval Services. Standards like Logical Framings for WSDL and XML Schema (SAWSDL) can always provide a machine

readable Web Data Retrieval Service, as in the case of Semantic Deep Web Services, Framing describing the functionality and retrievable data. A Logical Framing for structured query interfaces may provide machine readable information for henceforth called Logical Deep Web Data Retrieval Services (LDWDRS).

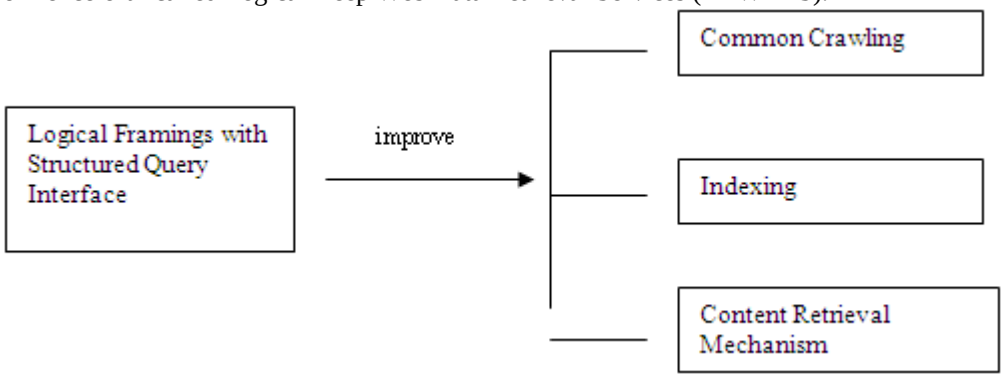

Figure 1 Impact in Deep Web from Logical Framings with Structured Query Interface

3.3. Logical Deep Web Data Retrieval System– Refinement

Logics for forming automatic input driven interface lies with a mediated schema with lists of values associated with different elements. [2] In this concept, an input in a form matches closely with an element in the mediated schema automatically. Thereafter, the values for the element can be used to frame the queries in the shape of logical structured interface form. An implementation of the LDWDRS Framing should comply with formation of structured interface logics, if possible, providing a generalization, as in the case of SDWS interfaces with the ability to include own vocabularies to enhance the automation in formation of structured query interface and avoid human intervention. Furthermore, a LDWDRS Framing prototype should automatically facilitate with information about general properties regarding the content that is provided by the LDWDRS (content details) and finally the desired interface field properties to describe the Logical and Internal structure dependencies of the LDWDRS structured interfaces (i.e., details of the attributes). The content type attribute may be described based on schema.org micro data and the supplementary usage of other vocabularies. The prototypes of refined LDWDRS content properties describe the content domain of the retrievable information, the content language, as well as the content data type. Introduction of an additional content property might provide information about the counts of available data only. Such properties of content are just the extendable basis for above referred prototype providing general information about the retrievable content. The prototype of proposed LDWDRS has attribute properties describing the attribute type, as well as the input domain and output range of each particular LDWDRS interface attribute. The input domain attribute describes valid input values of a specified LDWDRS attribute. Subsequently, it will become a starting point for the output range attribute, as its input value defines the restriction set for the retrieval process at time of structured form submission. Another dimension of such Retrieval System with respect to structured interface contains groups of related attributes that may affect each other. The first select attribute as part of the designated group defines the relation to the other groups. The second select attribute as part of this group defines the input attribute domain and may restrict the input attribute range of the input attribute that is part of the referred groups (Figure-2). More complex examples may establish that Logical meaning behind a LDWDRS interface might be quite complex and automated form understanding approaches may quickly reach their limits. Especially the automated detection of related attributes and the detection of complex relations within groups might be the most difficult but uncommon successful part.

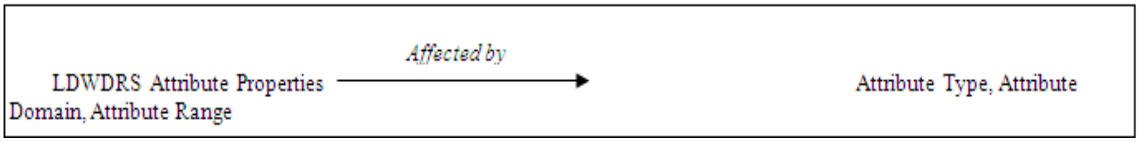

Figure 2 Groups of related attributes properties

The proposed LDWDRS structured interface automatic framing has led to great potential for new information retrieval mechanisms and plays a significant role for the improvement of current mechanisms. Queries from interfaces will automatically be mapped to further structured interface design with proposed model of LDWDRS. As suggested in the case of SDWS interface framing, it may here again be advisable to link every LDWDRS structured interface from the websites root index, which will determine a targeted LDWDRS detection and can be realized by using XML Sitemaps to define a LDWDRS retrieval index. Additionally, the LDWDRS structured interface framing will also be used for purposes, not directly focusing on the retrieval itself but on client side structured query interface input validations. This process of developing such interface and framing requires additional effort on the one hand, but on the other hand it also enables the webmasters to control the information content that may be retrieved by various retrieving services like search engines. Of-course, the webmasters may only use common HTML attributes to control the crawling behavior on their websites, hence is a limitation too.

4. CONCLUSION

The research work focused the lack of logical information regarding structured web query interfaces in the process of efficient data retrieval from the Deep Web. Transferring the concepts of logical web content framings on the one hand and logical web service descriptions on the other hand confirm to the great potential of Logical Framings for proposed LDWDRS structured interfaces. A variety of current data retrieval mechanisms and prescribed structured form understanding systems endeavor to study LDWDRS structured interfaces automatically by concentrating the Deep Web Data Retrieval challenge from the retrieving services angle. The proposed system follows the open knowledge sharing model as part of the Logical Web vision from Berners-Lee et al. [4]. This is based on the assumption, that the information provided on websites is planned to be retrieved by various services, LDWDRS is one of them but not the last. Any additional licensing issues restricting the retrieval and further usage of the retrievable information may also be considered in future. Our contribution to automatic Data Retrieval mechanisms is based on the introduced LDWDRS. Human intervened effort for Deep Web Data Retrieval mechanisms will have to be reduced if not stopped.

5. CHALLENGES AHEAD

Introduction of an additional automated layer of intelligent structured interface can never be conceived perfect in all senses for all possible kind of deep web data. Hence extension to LDWDRS is an open field to work upon by researcher times to come. The understanding of proposed LDWDRS structured interface framing based on the usage of existing framing standards will concern the challenge about retrieving services. Reduction of human intervention or effort for the said framing process also requires further research effort.

REFERENCES

[1] Arne Martin klemenz, Klaus Tochtermann, Semantification of Query Interface to Improve Access to Deep Web Content, SDA 2013, 3rd International Workshop on Semantic Digital Archives.
[2] Madhavan, J., Afanasiev, L., Antova, L., and Halevy, A. Harnessing the deep web: Present and future. 4th Biennial Conference on Innovative Data Systems Research (CIDR) (Jan. 2009).
[3] Bergman, M. K. White paper: The deep web: Surfacing hidden value. the journal of electronic publishing 7, 1 (2001).
[4] Berners-Lee, T., Hendler, J., Lassila, O., et al. The Logical web. Scienti_c American 284, 5 (2001), 28{37.
[5] Chun, S. A., and Warner, J. Logical Framing and search for deep web services. In E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on (2008), IEEE, pp. 389{395.
[6] Furche, T., Gottlob, G., Grasso, G., Guo, X., Orsi, G., and Schallhart, C. Opal: Automated form understanding for the deep web. In Proceedings of the 21st international conference on World Wide Web (2012), ACM, pp. 829{838.
[7] Handschuh, S., and Staab, S. Framing for the Logical web, vol. 96. IOS Press, 2003.
[8] He, B., Patel, M., Zhang, Z., and Chang, K. C.-C. Accessing the deep web. Communications of the ACM 50, 5 (2007), 94{101.
[9] Hicks, C., Scheffer, M., Ngu, A. H., and Sheng, Q. Z. Discovery and cataloging of deep web sources. In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on (2012), IEEE, pp. 224{230.
[10] Li, X., Dong, X. L., Lyons, K., Meng, W., and Srivastava, D. Truth _nding on the deep web: Is the problem solved? In Proceedings of the 39th international conference on Very Large Data Bases (2012), VLDB Endowment, pp. 97{108.
[11] Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., and Halevy, A. Google's deep web crawl. Proceedings of the VLDB Endowment 1, 2 (2008), 1241{1252.
[12] Masan_es, J. Archiving the hidden web. In Web Archiving. Springer, 2006, pp. 115{ 129.
[13] Wenyu, Z., Jianwei, Y., Ming, C., Jian, W., and Lanfen, L. Manufacturing deep web service management: Exploring Logical web technologies. Industrial Electronics Magazine, IEEE 6, 2 (2012), 38 {51}.
[14] M.J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. WebTables: Exploring the Power of Tables on the Web. In VLDB, 2008

AUTHOR

Dr. Brijesh Khandelwal did MCA from Lucknow University in year 1994. In 2001, he became Sun Certified Programmer. In 2007 he did PhD (Appllied Economics) from Lucknow University. He did MBA in 2010 from Punjab Technical University. In 2010 he also became licentiate in Life Insurance from Insurance Institute of India, Mumbai.

Dr. S. Q. Abbas has more than 23 years of experience in Academic and Administration. Currently, he is DG at Ambalika Institute of Management& technology, Lucknow. Dr. Abbas has rich & diverse experience in academia and is Visiting Professor at various universities & colleges. He has several publications in International/ National Journals & Conferences. He supervised many candidates of Ph.D & M.Phil. He is in advisory board and reviewer of various Int. & National Journals.