https://doi.org/10.36375/prepare_u.iei.a203

Engineering collaborations for accessing hidden web resources

Manpreet Singh Sehgal.
Department of Computer Science and Engineering
Apeejay Stya University
Sohna, Gurugram – 122103
manpreet.sehgal@asu.apeejay.edu

Abstract - The world wide web has always been a research platform for information and data scientists. The value of information is generally seen as a proportional to the depth of the world wide web. While the surface web is being taken good care of by the search engines at a great length, It is the deep web that posed challenges to the information and data scientists. A lot of research has been expended in the direction of extracting information from the deep web resources. Many researchers have studied the problem statement in their perspective and devise solutions to address their formulated problem statements. Individually these efforts showed promising results. However, there has been a gap of the joint effort to bring these individual efforts together to reap the collective fruits to address the much larger and more sensible problem statement. This paper highlights the different problem statements of different researchers and suggests the collaborative approach to bring out one common problem statement to envisage one big problem and its probable solution.

Keywords - Collaborative approach; Hidden Web; Deep Web; Information retrieval.

Introduction

World Wide Web (WWW) scores high when it comes to the ranking of information resources to satisfy the information need. There is a plethora of information (organized, unorganized, validated and invalidated) available in the WWW that needs to get tapped and validated before it gets consumed. Various validating organizations certify the correctness and validity of the information prevalent on the WWW. However, such validated information resides in the databases of the validating organizations which can only be accessed by presenting queries to these databases. For automating queries, organizations design search interfaces which need to be filled up to form queries and fetch information from the databases [1]. The information residing in databases is categorized as hidden web resource since it is hidden behind search interfaces into the databases. Traditional search engines can extract the information present in the documents but they are not able to extract data from the databases of validating organizations. However, many researchers have extended their resources and efforts to tap this area of information extraction and have been successful as well. But there is a research gap of collaborative efforts to completely address the problem in a way it should have been addressed. This paper firstly presents the related work in the direction of hidden web data extraction followed by the examination of the efforts which are potential candidates for collaboration. The anticipated results that will be probably achieved in the light of collaboration are discussed at the end of the paper.

Related Work

There are two parts to the World Wide Web (WWW). One part is exposed to the search engines so is easy to navigate, while the other is hidden in the databases, unlinked, non-textual form and hence out of reach of the traditional search engines thereby is hard to tap. The irony of the matter is that the later part of the WWW is 99% of the WWW implying the fact that what traditional search engines are capable to tap is just 1% of the entire WWW [2-5].

A lot of work has been done to study the techniques to extract the hidden data. Khurana

and Chandak [6] surveyed methodologies to select deep web source to extract the data hidden underneath. The problem statement of their research is to highlight the fact that "among multitudes of information available in the deep web, the quality of data should derive its selection measure to keep less relevant and redundant information from getting unnecessarily processed". Singh and Anuradha [7] proposed a methodology of sponger and squeezer to extract (Sponge) data from the hidden web page and after analyzing and processing it, pouring it (Squeeze) to the local repository for the future analysis by knowledge experts. Their problem statement is to "detect the type of data in the hidden web pages to guide the correct extraction of tabular information presented in HTML page onto the local repository".

Singh and Prasad in [8], suggested an ontology-based approach to extract the hidden web data for automatically filling up the query interface form after fetching the domain specific values from the built ontology of the specific domain. In their research, they also mentioned the need of ontology builder, in case there is no ontology for the subject being investigated. Their problem statement is to "design a unique system to uncover hidden web using existing ontologies". In the related research [9], Singh and Prasad proposed the replacement of ontology with that of database indexed by search engines, to gain the advantage of the existing WWW repositories containing virtually any domain information, thereby eliminating the need of Ontology Builder asked for in the earlier research [9]. Their problem statement is "overcoming the need to design ontology builder by replacing ontologies with the indexed databases of traditional search engines as they are there for virtually any domain". In [10] Y. Wang and J.Hu proposed a machine learning approach to sense the presence of tables on the hidden web pages. Their problem statement is to investigate the web pages for the presence of relational data so as to categorize the pages as relevant/irrelevant." B. Liu et. al in [11] proposed the effective strategy to mine the contiguous and non-contiguous data records to extract important information out of them. Their problem statement is to "address the issues present in exiting data mining approaches to bring out more accurate data mining approach". On the lines of extracting and processing data at the large scale in a collaborative manner Felix, Biswanath and Mirek introduced Web Lab Collaboration Server [12]. The underlying problem statement is to "recognize the fact that despite latest advances in data extraction and processing technologies, the users from non-technical background find it difficult to deal with the tools to fetch the relevant data for various applications and build an abstract model that presents easy to use interface for inputs to extract and process data to the non-technical users".

The approaches mentioned can be categorized as the improvements of the existing methods to make the life of application users easy. The author has envisaged the need to engineer the collective efforts of some of the above-mentioned techniques to achieve a larger objective. The following section throws light on the potential candidates of collaboration.

PROPOSED WORK

In this paper, the different individual approaches to extract hidden web data are selected and viewed from a larger perspective for collaboration to extract hidden web data using the goodness of the collaborating approaches.

https://doi.org/10.36375/prepare_u.iei.a203

Fig. 1 represents the collaborated technologies proposed in the current research. D and SD in Figure 1 represent the domain and statement of domain as two inputs to the UDDWE [9]. UDDWE initially fetches the documents from WWW for the corresponding D (Domain) and SD (Statement of Domain). At the later stage of UDDWE processing the hidden web pages from the fetched documents are fetched. Two categories of hidden web pages (Structured and Unstructured) are identified by UDDWE [9]. White colored documents are the pages which represent unstructured information, whereas the red colored blocks the hidden documents with the structured information. This mixed set of documents is passed to the data mining approach suggested by [11] and Machine learning approach proposed by [10] at the same time. Both [10] and [11] are designed to identify the documents with and without table tag. The output of Machine learning approach [10] is the set of documents with the tabular structures whereas Data mining approach [11] provides the mined data out of non-tabular unstructured documents. This mined data is stored in the local repository LR1 The tabular structured documents are parsed by HWPDE [7] as a result of which the data types of the columns of the tables are detected, the data present in the data cells of the table is sponged (extracted) and filled (squeezed) in the local repository LR2. The data from LR1 and LR2 is fed to the agent-based system for Data Integration suggested by [14]. The agent-based system introduced at the end collects the data from LR1 and LR2 and stores in its local repository LR3. LR3 is the considered as the data rich repository containing structured relational and unstructured mined data. This source of knowledge paves the way of strategic planning for the organizations dealing with the data.

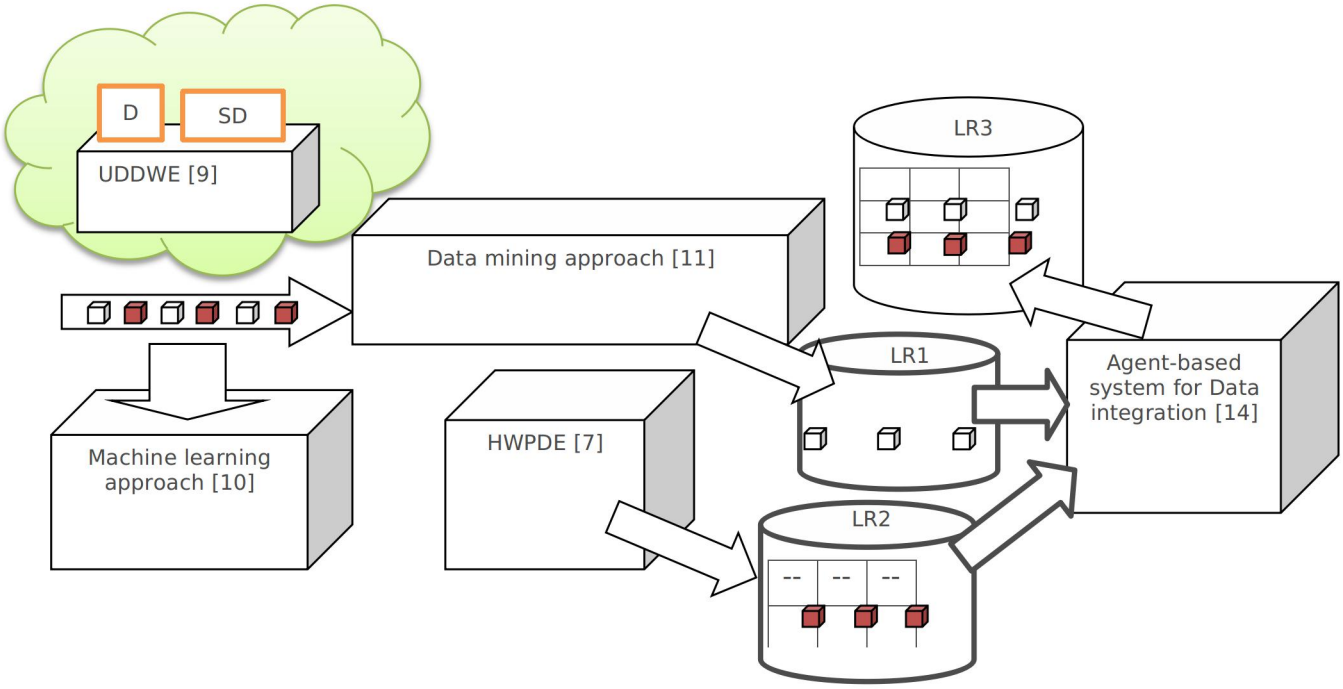

Fig. 1: Collaborative approach for effective data extraction from hidden web

Algorithmic Analysis And Results

The process of such collaboration needs a planning and planning leads to the algorithm. Following algorithm CollabH (coined for the collaboration for achieving Hidden web resources) is suggested to accomplish the stated objective.

Algorithm CollabH (D, SD)

HybridC = UDDWE (D, SD);
UnstructC = DMA (HybridC);
StructC = MLA (HybridC);
LR1=PrUSC(UnstructC);
LR2 = PrSC(HWPDE(StructC));
IntegratedC = DIA(LR1,LR2);
LR3=PrIR(IR)
return L3;

https://doi.org/10.36375/prepare_u.iei.a203

TABLE I
DOMAINS AND STATEMENT OF DOMAINS FOR COLLABORATIVE ANALYSIS.

Sr. No	Domain	Statement of Domain (Query)
1	Cricket	Orange Cap award in IPL 2019
		ICC World Cup history
2	Food	Order from Swiggy: One Masala Dosa, Two Cokes and one small French Fries
		I want to buy pedigree animal food for my Labrador dog.
3	Civilization	The Egyptian civilization
		The culture and civilization of Sindh

D (Domain) and SD (Statement of domain) are the only two arguments required for the proposed algorithm. In the beginning UDDWE is called to fetch the structured and unstructured collection (collectively refered here as hybrid collection and mentioned as HybridC in the algorithm) of hidden web pages. The hybrid collection is processed by DMA (Data Mining Approach) and MLA (Machine Learning Approach) to filter the unstructured and structured collections (refered in the algorithm as UnstructC and StructC). The unstructured collection (UnstructC) is processed and translated (by Processing function PrUSC) in to the format suitable for getting stored into the local repository (LR1 in the algorithm). The structured collection is processed by HWPDE (Hidden Web Page data Extractor) to automatically transfer the structured data on the webpage into the relational database (LR2) after suitable processing

by the processing function PrSC (identification of data types, checking if the repository is being built for the first
time etc). Finally, the DIA (Data Integrator Agent) picks up the contents of the local repositories LR1 and LR2 and integrates the repositories in the collection IntegratedC. This Integrated Collection is stored in Local repository LR3 which is returned as the output. The difference in HybricdC and IntegratedC is in their format. HybridC is the collection of Webpages both structured and unstructured, whereas IntegratedC is the collection of the relevant and the focused data extracted from the webpages.

The experiments were carried out on sample domains (D) and statement of domains (SD) depicted in Table I.

UDDWE brings out the hidden web entry points and consequently enable fetching of hidden web data (both unstructured and structured) corresponding to each statement of domain for the corresponding domain mention in Table I.

TABLE II
UDDWE Results

Sr. No	Domain	Statement of Domain	Fetched Pages (K)	Relevant Pages(R)	Identified Hidden Web Entry Points(H)
1	DM1: Cricket	Orange Cap award in IPL 2019	12	8	1 (Relevant)
		ICC World Cup history	27	23	17 (Relevant)
2	DM2: Food	Order from Swiggy: One Masala Dosa, Two Cokes and one small French Fries	15	11	0
		I want to buy pedigree animal food for my Labrador dog.	22	18	4 (Relevant)
3	DM3: Civilization	The Egyptian civilization	26	22	13 (Relevant)
		The culture and civilization of Sindh	17	13	9 (Relevant)

The result of UDDWE on the information present in Table I is represented in Table II. The fetched pages by UDDWE are from the WWW, which are processed further in UDDWE for the presence of hidden web entry point signature (Search query interface form). After the firing of queries by UDDWE and human intervention two categories of documents

(Structured and Unstructured) are fetched. Data mining approach identifies unstructured data and pour in in the local repository, whereas machine learning approach identifies structured data and passes it to HWPDE for sponging and squeezing processes. HWPDE automatically detects the data types of the columns of structured data stored in HTML table, and creates the local repository with the detected type information before storing the information in it. Finally, both structured and unstructured information repositories are merged together using agent-based approach for data integration.

CONCLUSION AND FUTURE SCOPE

The collaborative approach depicted in the research promises the combined goodness of the individual research efforts. The final repository shown in the proposed work serves as the backbone of the data analysis and proves helpful in effective decision making for the stakeholders involved. Though the approaches discussed in the collaboration effort are separated apart in the time domain, however such efforts will always be welcome in future as by using collaborative approach, system designers don't need to re-invent the wheel for designing a subsystem if the similar subsystem is already in place by other researchers. The paradigm shift to collaborative efforts will ensure the greater emphasis on solving bigger problems with the solutions of the smaller sub problems which are already solved for specific applications.

ACKNOWLEDGMENT

Author would like to acknowledge the research facilities of J.C Bose University of Science and Technology, Faridabad, MVN University, Palwal and Apeejay Stya University, Sohna for the help in the proposed research.

REFERENCES

[1] A. Ntoulas, P. Zerfos, J.Cho., "Downloading Textual Hidden Web Content Through Keyword Queries," in Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'05, Denver, USA, Jun 2005 IEEE , pp. 100-109.
[2] The size of the World Wide Web (The Internet) [Online], Available: https://www.worldwidewebsize.com/, Accessed on 10 June 2019.
[3] Bergman and M.K, "The Deep Web: Surfacing hidden value," Journal of Electronic Publishing, vol. 7, no.1, pp. 1-17, 2001.
[4] S. Lawrence and C.L. Giles, "Searching the World Wide Web," Science, International Journal of Science, vol. 280 no.5360, pp. 98–100, 1998.
[5] S. Lawrence and C.L. Giles, "Accessibility of information on the web," Nature, International Journal of Science, vol. 400, no. 107. https://doi.org/10.1038/21987, 1999.
[6] K.Khurana and M.B Chandak, "Survey of Techniques for Deep Web Source Selection and Surfacing the Hidden Web Content", International Journal of Advanced Computer Science and Applications, Vol 7, No.5, pp. 409-418, 2016.
[7] M. Singh and Anuradha, "HWPDE: Novel approach for data extraction from structured web pages," International Journal of Computer Applications, vol. 50, no. 8, pp. 22-27, July 2012.
[8] M. Singh and J.S. Prasad, "All Domain Hidden Web Exposer Ontologies: A Unified approach for excavating the web to unhide deep web," in Proceedings of International Conference on Smart Innovations in Communication and Computational Sciences, Indore, India, Ed. by Springer, Singapore, pp. 423-431, 2018.
[9] M. Singh and J.S. Prasad. "UDDWE: Universal Domain Deep Web Exposer," International Journal of Engineering and Technology (UAE), vol. 7 no.4 pp 4398-4404, 2018.
[10]Y. Wang and J. Hu, "A machine learning based approach for table detection on the web," in Proceedings of the 11th international conference on World Wide Web, New York, NY, USA, ACM Press, pp. 242–250, 2002.
[11] B. Liu, R. Grossman, and Y. Zhai, "Mining data records in web pages," In Proceedings of the ninth ACM SIGKDD International Conference on Knowledge discovery and data mining, New York, NY, USA, ACM Press, pp. 601–606, 2003.
[12] Felix Weigel, Biswanath Panda, and Mirek Riedewald, "Large-Scale Collaborative Analysis and Extraction of Web Data", VLDB endowment, pp. 1476-1479, 2008
[13] Baker Ross Inspiring Creativity, Available Online at https://www.bakerross.co.uk/wooden-jigsaw-puzzles, Accessed on August 24, 2019.
[14] J.T. McDonald, M.L. Talbert, and S.A. DeLoach, "Heterogeneous database integration using agent-oriented information systems," in Proceedings of the International Conference on Artificial Intelligence (IC-AI' 2000), Monte Carlo Resort, Las Vegas, Nevada, CSREA Press, vol. 3, pp. 1359-1365, June 26–29, 2000