INTERNATIONAL JOURNAL OF SCIENTIFIC PROGRESS AND RESEARCH (IJSPR) ISSN: 2349-4689
Volume-10, Number - 02, 2015

An Improved Hybrid Web Search Approach for Distributed Information Extraction

Prachi Mhatre[1], Seema Ladhe[2]
[1]*PG Scholar, Computer Engineering, MGMCET Kamothe, Navi Mumbai (India)*
[2]*Professor, Computer Engineering, MGMCET Kamothe, Navi Mumbai (India)*

Abstract* – Information Extraction (IE) from the databases is a critical issue. Web Service Asymmetry (WSA) is a major challenge in front of database users. We have proposed an algorithm which aims to retrieve critical data from the database. Existing Web Services proposed by various authors are black box, hard-coded and cannot be changed. The traditional IE usually takes advantages of NLP techniques such as lexicons and grammars, whereas Web IE adapts machine learning and pattern mining techniques. A major challenge in efficient evaluation of queries over such Web Services is to detect relevant data for the query execution and to avoid materialization of irrelevant information. So, considering the above issue we have proposed an algorithm, Web Information Extraction Using Service Calls (WIESC), a scheme which tries to overcome the problem of WSA. WIESC algorithm results in faster execution of query since it dynamically executes the services.

***Keywords:* Information Extraction(IE), Web Service Asymmetry(WSA), Web Information Extraction Using Service Calls(WIESC), NLP.**

I. INTRODUCTION

The Web is the biggest online Data Source and it’s still expanding. The Web contains a wealth of information about different fields, in different dimensions. The information stored in the databases is in some structured form which cannot be directly operated by users, so even if the data is present it cannot be fetched. This tends to be highly inconvenient. To utilize the available information we need to extract it from the database. The need for Information Extraction emerged from the vast growth of the web and increasing data sources. The task of Information retrieval is tough as the web services limits the types of queries that can be implied on the databases. This limitation of the service which restricts the retrieval of data from the database is known as Web Service Asymmetry. The web service may be available to access one type of data but it may fail to access the inverse of that data. For example, a We Service may provide functions that return the songs of a given singer, but it might not present a function that returns the singers of a given song. To overcome this limitation, our approach implies some binding patterns on the Web Service functions, by presenting some values as input guideline prior to the function call.

II. SYSTEM MODEL

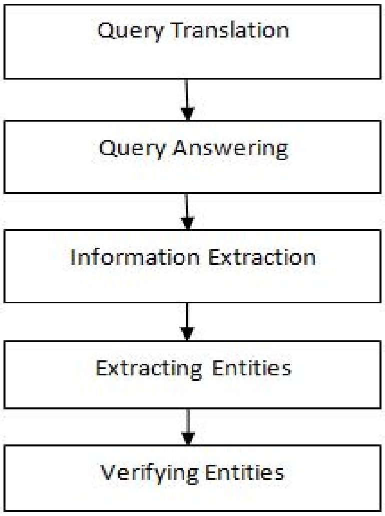


Fig 2.1 Information Extraction System

The existing system also provides a solution to the problem of Web Service Asymmetry. The information is been extracted from the Web data sources to be provided as input to the Web Services. Here, keyword based search is used. A keyword query of the element to be searched is issued and the search engine extracts promising entities from the resulted pages. Further, we use the existing system to verify and validate the entities. This will lead to the selection of the correct entity from the resulted data, thus discarding the other entities. In this way, we overcome the limitation of Web Service Asymmetry here by retrieving the information for which the Web Service does not provide a direct access. In this system, functions are used and prioritized over infinite chains of calls.

III. PREVIOUS WORK

In 2010, Wei Liu, Xiaofeng Meng and Weiyi Meng proposed a technique which allowed them to access the Deep Web contents submitted to the Web databases and the query results were displayed in the form of Web pages [1].

In 2010, Nicoleta Preda, Fabian Suchanek, Wenjun Yuan, Gerhard Weikum proposed a system which uses views to answer the Web Services and it also limits the number of function calls. This approach focuses on prioritizing the function calls that deliver crisp answers to queries [4].

In 2012, Michael Benedikt, Pierre Bourhis and Clemens Ley introduced a protocol which reduced the number of accesses by imposing restrictions on the queries [6].

In 2011, Xuan Liu, Xin Luna Dong, Beng Chin Ooi, Divesh Shrivastava proposed a Data Fusion Technique that overcomes the conflicts of data coming from various Data Sources and the result we get is a consistent set of data [7].

In 2010, Luciano Barbosa, Hoa Nguyen, Ramesh Pinnamaneni, Juliana Freire worked on the probing form fields and prepared the schema of a single form, so that the functions can situate correctly typed values into particular fields, preventing unnecessary requests [8].

IV. PROPOSED METHODOLOGY

A chief challenge in the systematic assessment of functions over Web Services is to detect which function calls may bring data that is admissible for the query execution and to avoid the manifestation of irrelevant information. The difficulty is complex, as service calls may be rooted anywhere in the services and service invocations possibly retrieve data consisting calls to new services. So the recognition of relevant calls becomes a uninterrupted process.

The algorithm proposed in our system is dynamic, in the sense that it gets adapted to the state of Web Services as the state gets customized by service invocations and decides at each point which services should be invoked next.

The difficulty we tackle is related to the mediation paradigm. Here the data sources are invoked to answer queries on a mediated scheme where, Service calls may emerge anywhere in the data. They may also appear dynamically in the results of previously invoked calls.

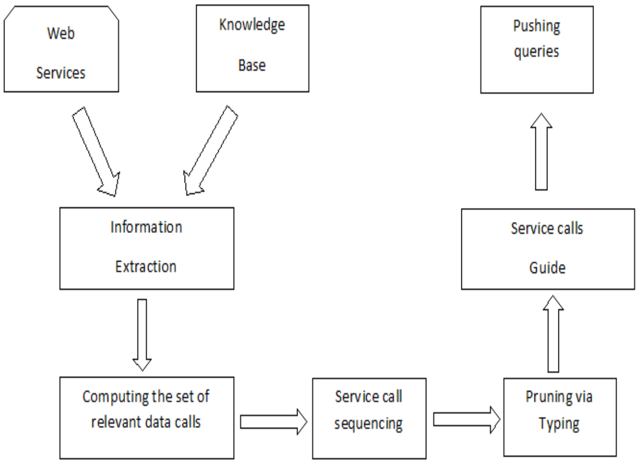


Fig. 3.1 WIESC Model

Furthermore, the significant results of calls depends on one another and the return types of the service calls must be taken into consideration. To understand the system properly, we need to get better perception of the probable relationships among the service calls embedded in various Web Services and their manipulation on the relevance of function calls to queries.

Based on our above concepts, our essential contribution is an efficient algorithm for lazy query evaluation, whose key facets are as follows.

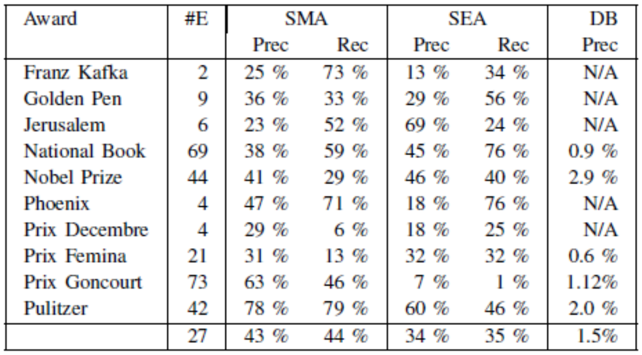
- The algorithm generates a set of queries that fetch all service calls relevant because of their position which helps in computing the set of relevant function calls.
- Relationships among the calls are scrutinized, to deprive a sequence of call invocations suitable to answer a query which comes under service calls sequencing.
- Return types of services are used to rule out more function calls which leads to pruning using typing.
- A specific access structure is used to speed up the recognition of relevant calls which defines the service calls guide.
- Accurate knowledge of the interface between the query and each service call permit pushing queries to be proficient Web services, like do with data sources.

Our approach show that, compared to the naive approach, the pruning of extraneous service calls may shrink the overall query evaluation time by orders of magnitude.

V. EVALUATION

In our system, the Information Extraction algorithms are used to generate candidate entities which can be used as inputs for Web Services. Here, we first evaluate the performance of the IE algorithms. To evaluate the IE algorithms, various queries are implemented. The procedure goes as mentioned here. For every query type some property values are chosen. Then further, for each property value, a keyword query is generated, it is then sent to Google and the data is fetched from there. The fetched pages many be varied in characteristics like different structures of entities. Next step is to manually extract the targeted entities from these pages. Further, we run the IE algorithms and measure the performance according to our set standard.

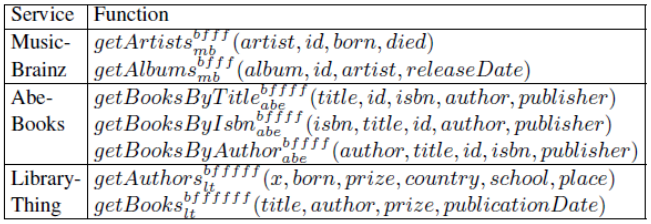
The following table shows the results of the information extracted from the web pages. #E is the average number of candidate entities per page. SMA is the String Matching Algorithm and SEA is the Structured Extraction Algorithm. And DB is the raw algorithm. Thus, the algorithms find too many entities from the web pages.



Award	#E	SMA		SEA		DB
		Prec	Rec	Prec	Rec	Prec
Franz Kafka	2	25 %	73 %	13 %	34 %	N/A
Golden Pen	9	36 %	33 %	29 %	56 %	N/A
Jerusalem	6	23 %	52 %	69 %	24 %	N/A
National Book	69	38 %	59 %	45 %	76 %	0.9 %
Nobel Prize	44	41 %	29 %	46 %	40 %	2.9 %
Phoenix	4	47 %	71 %	18 %	76 %	N/A
Prix Decembre	4	29 %	6 %	18 %	25 %	N/A
Prix Femina	21	31 %	13 %	32 %	32 %	0.6 %
Prix Goncourt	73	63 %	46 %	7 %	1 %	1.12%
Pulitzer	42	78 %	79 %	60 %	46 %	2.0 %
	27	43 %	44 %	34 %	35 %	1.5%

Fig. 5.1 Results for "Authors who won prize X".

The below shown figure describes the signatures of some sample functions. The queries where chosen in such a manner that they have different alternative ways of composing unction initiations. And this results greater number of service calls.



Service	Function
Music-Brainz	$getArtists_{mb}^{bfff}(artist, id, born, died)$ $getAlbums_{mb}^{bfff}(album, id, artist, releaseDate)$
Abe-Books	$getBooksByTitle_{abe}^{bffff}(title, id, isbn, author, publisher)$ $getBooksByIsbn_{abe}^{bffff}(isbn, title, id, author, publisher)$ $getBooksByAuthor_{abe}^{bffff}(author, title, id, isbn, publisher)$
Library-Thing	$getAuthors_{lt}^{bfffff}(x, born, prize, country, school, place)$ $getBooks_{lt}^{bffffff}(title, author, prize, publicationDate)$

Fig. 5.2 Sample functions integrated in our system.

VI. CONCLUSION

In this system, we have introduced the problem of Asymmetric Web Services. We have shown that a substantial number of Web services allow requesting for only one parameter of a relationship, but not for the other. In our system, we have proposed a efficient algorithm named WIESC for lazy query evaluation. It uses information extraction to estimate binding pattern for the input values and then authenticate these bindings by the Web service. Through this approach, a entire new class of queries has become tractable. We have revealed that providing inverse functions alone is not enough. They also have to be prioritized consequently. We have implemented our system and showed the validity of our approach on real data sets.

VII. FUTURE SCOPES

Our current implementation uses lazy query evaluation techniques that serve mainly as a proof of concept. This approach leads better results compared to the naive approach, the pruning of irrelevant service calls may reduce the overall query evaluation time by orders of magnitude. Our system can be extended to some more algorithms in future.

VIII. ACKNOWLEDGEMENT

I am Prachi Mhatre and I am greatly thankful to Prof. Seema Ladhe for her inspiration and guidance. We also thank the college authorities, PG coordinator and Principal Sir for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to M.E. Staff and friends.

REFERENCES

[1] Wei Liu, Xiaofeng Meng, Weiyi Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 3, MARCH 2010

[2] Benjamin Habegger, Mohamed Quafafou,"Web Services for Information Extraction from the Web,"Proceedings of the IEEE International Conference on Web Services(ICWS'04)

[3] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan,"A Survey of Web Information Extraction Systems," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 10, OCTOBER 2006

[4] Nicoleta Preda , Gjergji Kasneci , Fabian M. Suchanek , Thomas Neumann , Wenjun Yuan , Gerhard Weikum ,” Active Knowledge: Dynamically Enriching RDF Knowledge Bases by Web Services, SIGMOD’10

[5] Xin Jin, Nan Zhang, Aditya Mone, Gautam Das, “Randomized Generalization for Aggregate Suppression Over Hidden Web Databases,” Proceedings of the VLDB Endowment, Vol. 4, No. 11

[6] Michael Benedikt, Pierre Bourhis, Clemens Ley, "Querying Schemas With Access Restrictions,“ Proceedings of the VLDB Endowment, Vol. 5, No. 7

[7] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, Divesh Srivastava “Online Data Fusion,” Proceedings of the VLDB Endowment, Vol. 4, No. 12

[8] Luciano Barbosa, Hoa Nguyen, Thanh Nguyen, Ramesh Pinnamaneni, Juliana Freire,” Creating and Exploring Web Form Repositories,” Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...

[9] Luis Galárraga1, Christina Teflioudi1, Katja Hose2, Fabian M. Suchanek,” AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases,” ACM 978-1-4503-2035-1/13/05.

[10] Antoine Amarilli, Luis Galaarraga, Nicoleta Preda, and Fabian M. Suchanek,” Recent Topics of Research around the YAGO Knowledge Base,” Telecom ParisTech, Paris, France University of Versailles, France 2014.

[11] Fabian M. Suchanek, Johannes Hoffart, Erdal Kuzey, Edwin Lewis-Kelham,” YAGO2s: Modular High-Quality Information Extraction with an Application to Flight Planning,” Max Planck Institute for Informatics, Germany 2013.

[12] Fabian Suchanek and Gerhard Weikum,” Knowledge Harvesting from Text and Web Sources,” Max Planck Institute for Informatics 66123 Saarbruecken, Germany 2013.

[13] Richi Nayak, Pierre Senellart, Fabian M. Suchanek, and Aparna S. Varde,” Discovering Interesting Information with Advances in Web Technology,”

[14] Information Extraction basic information available http://en.wikipedia.org/wiki/Information_extraction