IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Data Fusion Using Conflict Identification Methods for Knowledge Mining Based Repository Creation

I. Carol[1], Dr. S. Britto Ramesh Kumar[2]

Research Scholar, Department of Computer Science, St. Joseph's College, Trichy [1]

Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy[2]

Abstract: Accessing heterogeneous metadata can provide deep and meaningful insights. However integrating such data poses a serious problem in terms of duplicates and conflicts. Eliminating those inconsistencies will lead to effective data integration and hence better mining. This paper presents a knowledge repository based data fusion technique that not only eliminated duplicates and conflicts, but also identifies the user's requirements to provide effective and faster results. The knowledge repository is built based on the user's feedback and consequent retrievals are made from both the repository and the web in-order to effectively increase the hit ratio. As the repository becomes more mature, retrievals are confined to the repository alone. Experiments conducted depict better accuracies and faster retrieval rates, hence providing an overall high quality of experience for the user.

Keywords: Conflict Identification; Conflict Resolution; Duplicate Identification; Knowledge Repository; Reinforcement Learning.

I. INTRODUCTION

The recent times have shown increased attention towards managing and handling large volumes of data from heterogeneous data sources. Data integration is one of the mostly used mechanisms for managing heterogeneous data and utilizing them effectively. However, data integration mechanisms have several problems associated with them [5, 6, 7]. One such major problem faced by data integration mechanisms is the varied data formats. Data source poses data in their native formats, such as; database specific formats, web pages, XML, deep web, JSON and files. Apart from these formats, there can also exist unstructured or semi-structured formats. Hence a unified format was required for processing such heterogeneous data. Retrieving heterogeneous data and converting it to a unified format can enable enhanced data mining in them. XML was considered as a mediator that can be effectively used for this purpose [17]. Metadata based data fusion is one of the major contributors to effective data integration. Utilizing metadata effectively can reduce a huge amount of aftermath processing. This also tends to reduce the conflicts that occur in the later phases due to similar data. The basic requirement is to improve the Quality of Experience (QoE) [19] provided to the user. This paper presents a knowledge repository based data fusion technique that builds the data repository according to the customer's needs and can effectively reduce the time consumed for the retrieval process consequently increasing the accuracy by increasing the number of positive hits.

II. RELATED WORK

Data integration from heterogeneous data sources has become a mandatory scheme due to the explosion of the availability of digital data. Since their mode of availability and structure varies, several mechanisms have been proposed to facilitate the process of integration. A rule based algorithm used for decomposing the data into triples was presented by Singh et al. in [8]. This method proposes a set of transformation rules for the existing data models. These rules are categorized individually for structured, semi-structured and unstructured data models. They operate on the data models to decompose them into standard triples format. A data space approach that integrates data sources in 'pay-as-you-go' manner was presented by Halevy et al. and Franklin et al. in [9,10]. This method improved the existing data management approach [11, 12] by providing uniform query formats for querying and managing data. Some data space systems include Personal Information Management (PIM) [13, 14], Scientific Data Management [15, 16]. Recent studies also deal with context awareness in the process of query retrieval in order to increase the probability of providing a hit. A data fusion approach that has its basis on identifying the context of the query when retrieving the results is presented was presented by TalebiFard et al. in [18]. This method utilizes the context similarity of the query and also incorporates fuzzy MADM in the process of retrieving results. A metadata based scheme that identifies the mappings between the schemas to identify a global schema is presented by Rebai et al. in [20]. The major downsides of metadata based approaches are that they return same results, irrespective of the user profile. The method proposed in [20] eliminates this problem and considers user's profile during the retrieval process. Some metadata based data integration approaches include MUSE [21], Clip [22], Clio [23], CUPID [24] and MuMIE [25].

III.DATA FUSION USING CONFLICT IDENTIFICATION METHODS FOR KNOWLEDGE MINING BASED REPOSITORY CREATION

The data fusion using conflict identification methods for creating a knowledge mining repository is a conflict identification and resolution technique that operates on the basis of retrieving content from several available data sources to provide the type of content required by the user. This technique works in two major phases; the first phase deals with identifying the best data related to the user query and eliminate conflicts to obtain clean data and the second phase is to maintain and reinforce a knowledge repository such that the data retrieved in filtered again using the user feedback and added to the knowledge repository. This knowledge based feedback mechanism tends to build a domain based knowledge repository that can be used to retrieve results effectively.

A. Conflict/ Duplicate Identification and Resolution

The first phase involved in the process of conflict identification and resolution is to obtain and process the input query. Input query provided by the user can be of any size, hence the user tends to use complete sentences rather than specific keywords. General sentences contain stop words, usually used as connectors. They do not convey any meaning, but their presence tends to attract several text that are not related to the actual query. These stop words need to be eliminated from the text in order to converge the result retrieval process to specific and mandatory details. The next phase builds a basic query to be applied on the knowledge repository. This query retrieves any information available in the repository corresponding to the user's query. The results are checked and outdated results are filtered to obtain the final set of results. If the final result set contains the specified threshold count of data, these results alone are passed to the subsequent phases. If the threshold limit is not reached, the universal wrapper is used for building the actual query as described in our previous contribution [1]. The universal wrappers utilizes the basic query to provide queries corresponding to the data source being operated upon. Since the proposed approach deals with multiple data sources, the process of creating a universal wrapper becomes mandatory. The resultant queries are executed in their corresponding data sources to obtain the results relating to the query. These results are then integrated to obtain the final result set. Due to the presence of multiple data sources, the results obtained at this section would contain duplicates [2] and might be outdated.

The major goal of the next phase is to eliminate the inconsistencies, conflicts and duplicates. The results are prioritized on the basis of their properties. The properties considered for the current approach includes timestamp of the data, source, author and association level with the query. Each property is given a weight and the aggregated weights correspond to the final importance level of the data. The results are then ranked in decreasing order of the aggregated weights to obtain the prioritized list of the results. Similarity detection [3, 4] is performed on the data to eliminate duplicate entries. A pairwise comparison is carried out and directional similarity of the pairs is obtained. Pairs exhibiting similarity levels below 0.8 are retained, while the ones exhibiting similarity levels equal to and above 0.8 are marked for elimination. This phase follows the duplicate analysis phase and the actual elimination of duplicates. All pairs marked as duplicates are initially analyzed with their ranks. If their ranks differences are high, the data exhibiting the lower rank is eliminated. If the differences are moderate, metadata of the involved entries are analyzed and the most recent data is retained. If their difference is low, both the entries are retained and are marked as conflicts. This marks the beginning of the conflict identification phase. If the pair of data exhibits similarity between 0.5 and 0.7, the entries are said to be in conflict. Resolution is carried out on the basis of their metadata properties with major significance provided to the timestamp.

B. Reinforcement based Knowledge Repository Maintenance

The concept of reinforcement based knowledge repository has been introduced in this section in order to accumulate domain based knowledge to speed up operations (Figure 1). This repository is maintained corresponding to each client. Hence the queries provided by the clients and the corresponding results are maintained in this repository. At the initial stages, the repository is empty, as the user starts providing the queries, the repository starts building up. The results from conflict/ duplicate identification and resolution are displayed to the user. These results are intermediate, as they still do not directly provide the final results, i.e. the results expected by the user. The user examines these results and provides feedback on the results appropriate to their query and the results that are not. These results, along with the query and the metadata are stored in the knowledge repository. Every time the user provides the same query, results are retrieved from the knowledge repository along with the results from the external data sources. The results from knowledge repository have a huge probability of providing positive results when compared to the results obtained from the external data sources. On future queries, results are retrieved from the repository and external retrieval is slowly reduced with increase in data in the knowledge repository corresponding to the given query. After a threshold limit, data retrieval is limited to the knowledge repository alone and external queries are completely eliminated. This process is carried out on the basis of the query text, hence if a query has minimal or low representations in the repository, it still relies on the external sources for its results. This method works on the assumption that a user poses most of their queries corresponding to their domain and hence building up a domain based repository can benefit the users on most of their queries.

This scheme of working reduces time and data transfer to a considerable extent, however consistency of the system and contemporary nature of information are two of the major problems experienced in this method. These problems can be effectively solved by validating the

DOI 10.17148/IJARCCE.2016.5295

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

information in the domain frequently. This is carried out after every query hit in the knowledge repository. If the repository returns results after a query hit, the results are validated. If the results were found obsolete, they are removed both from the repository and from the results set. Performing this process during query hits is advantageous, as the repository need not be completely updated and updating a small part of the repository is usually simple and less time consuming. Further, if the user does not use particular information frequently, updating it periodically becomes a major overhead. Hence this method of performing data fusion and conflict identification using a knowledge repository tends to improve the speed and accuracy of the data retrieval and also reduce the presence of conflicts in a domain specific manner.

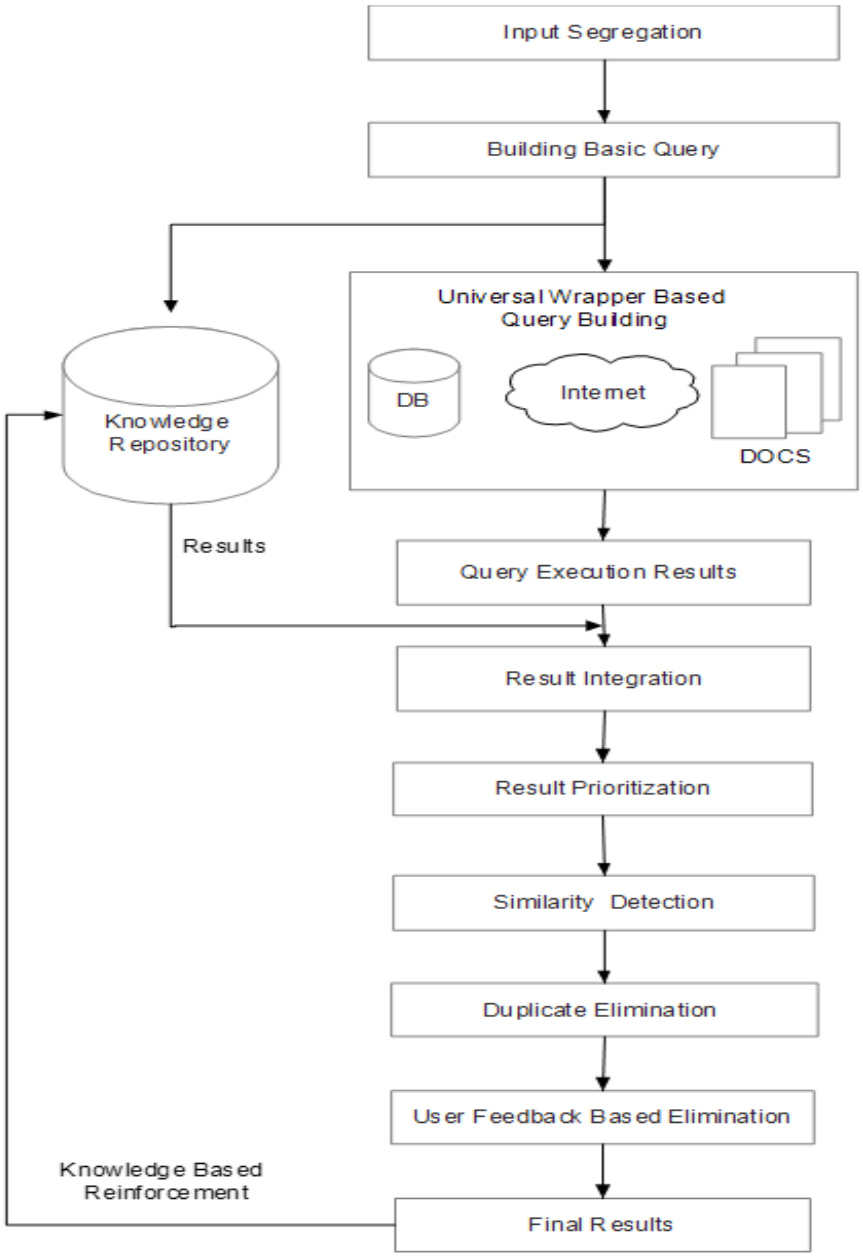

Fig 1: Data Fusion Using Conflict Identification Methods for Knowledge Mining Based Repository Creation

IV. RESULT AND DISCUSSION

Experiments were conducted by querying data from Google [26] and the New York Times API [27]. Result retrieval using API, conflict identification and resolution and the knowledge repository maintenance were programmed using Python.

Improved efficiency in terms of results were observed.

Figure 2presents the duplicate analysis performed on the data retrieved from the query. It was observed that the reduction rate increases with increase in data retrieval.

DOI 10.17148/IJARCCE.2016.5295

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

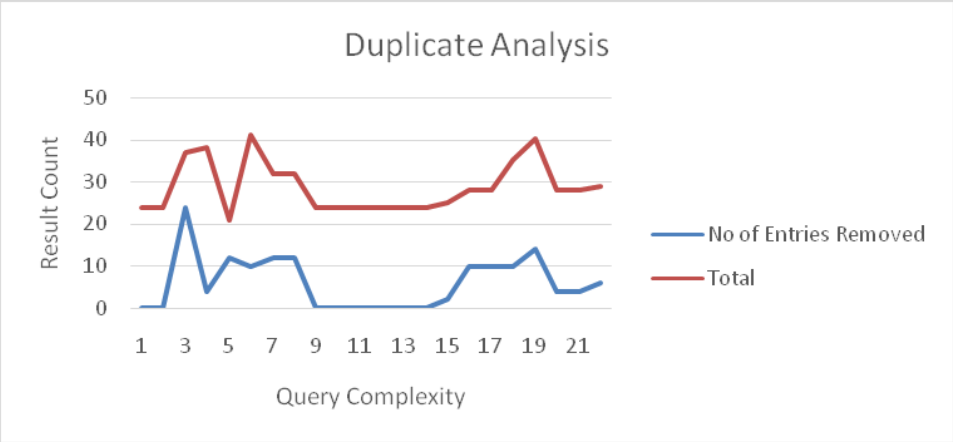

Fig 2: Duplicate analysis

Figure 3 shows the elimination percentage. It could be observed that our algorithm performs eliminations to maximum levels of 65%.

Figure 4 presents the time taken for result retrieval. The results exhibit several spikes and depressions in an irregular manner. Experiments were conducted by applying the same query multiple times to identify the retrievals. Due to the usage of the feedback mechanisms, subsequent queries with the same query term results in lesser retrieval time.

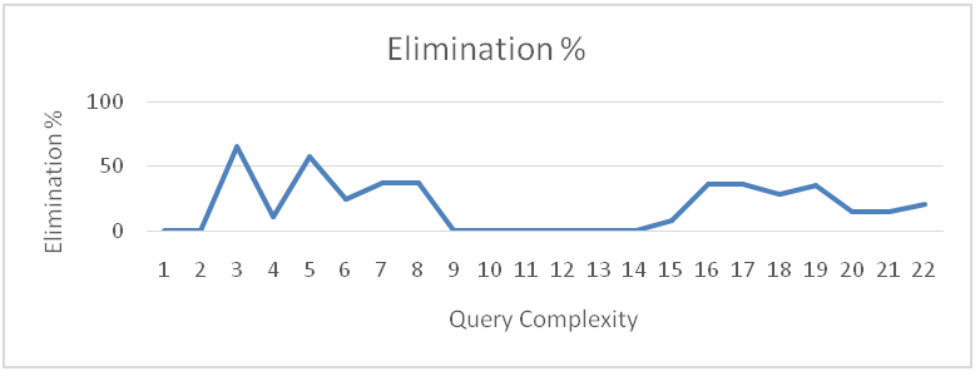

Fig 3: Elimination Percent

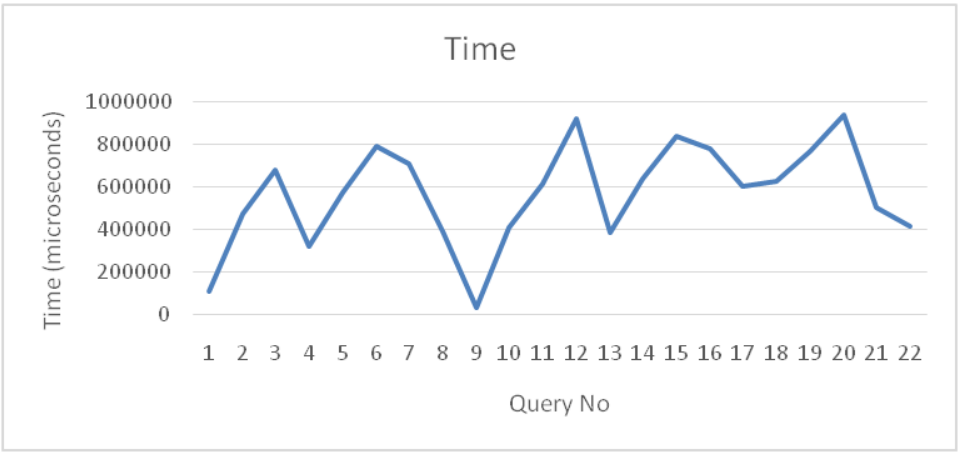

Fig 4: Time Taken

Figures 5, 6, 7 show a clear representation of the working of the feedback mechanism. It could be observed from Figures 5,6,7 that the initial query takes a higher amount of time, while in the subsequent queries, the time taken is reduced.

Due to the feedback mechanism, as the knowledge repository gets populated, the time taken reduces and after a threshold, the time taken becomes stable, as results are retrieved only from the repository and not from the web.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

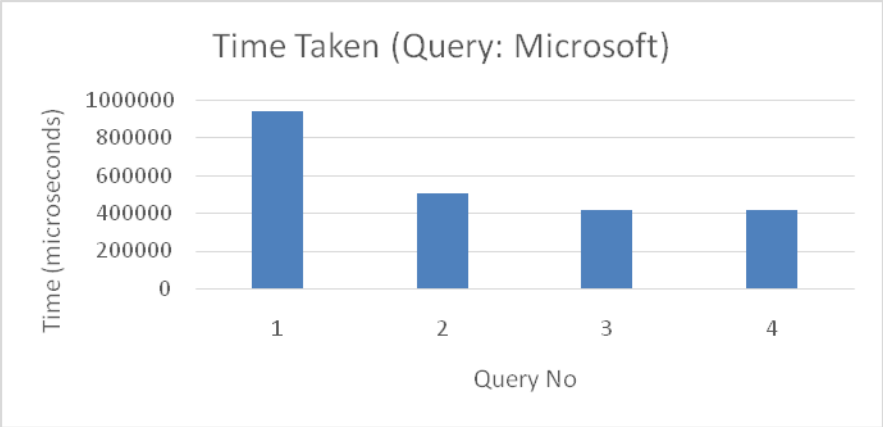

Fig 5: Time Taken (Query: Microsoft)

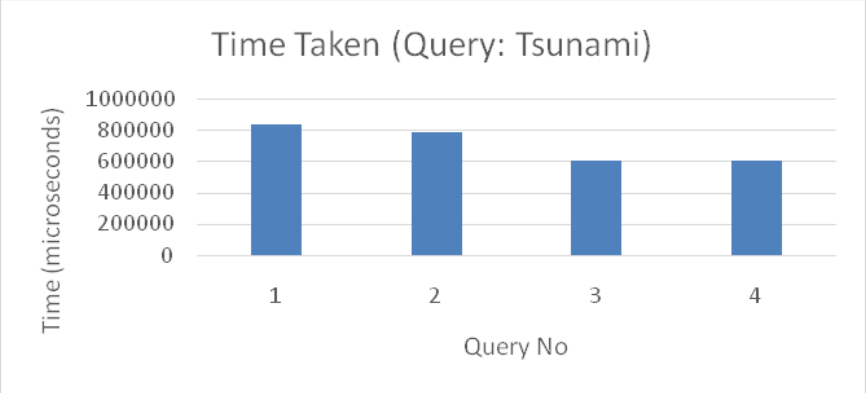

Fig 6: Time Taken (Query: Tsunami)

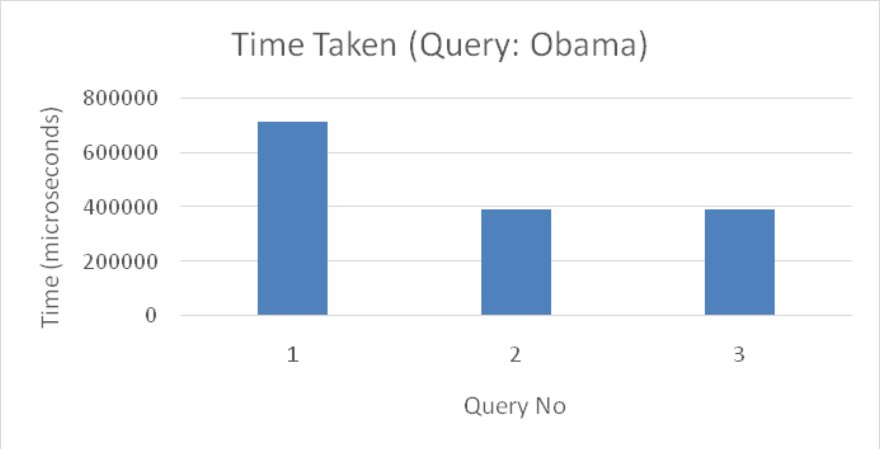

Fig 7: Time Taken (Query: Obama)

IV. CONCLUSION

The process of identification and resolution of conflicts is a major functionality being carried out in any data integration system. This paper presents a knowledge based data fusion mechanism that operates effectively on the heterogeneous data and provides effective results. A reinforcement learning mechanism is incorporated by integrating the user's feedback. This mechanism effectively reduces the time taken for the retrieval process by maintaining a local copy of the data that is most frequently required by the user. A maintenance mechanism that eliminates stale data is also incorporated to maintain the stability of the system. Experiments indicate effectiveness in the retrieval rates and accuracy hence improved quality of experience for the user. Future works will include enhancing the retrieval accuracy by incorporating context awareness into the current system.

DOI 10.17148/IJARCCE.2016.5295

REFERENCES

[1] I. Carol, and S. B. R. Kumar, Conflict Resolution and Duplicate Elimination in Heterogeneous Datasets using Unified Data Retrieval Techniques. Indian Journal of Science and Technology, 8(22), p.1.2015.
[2] D. B. T. Zesch, and I. Gurevych, Text reuse detection using a composition of text similarity measures. In Proceedings of COLING (Vol. 1, pp. 167-184).2012.
[3] R.Mihalcea, C. Corley, andC. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity. In AAAI (Vol. 6, pp. 775-780),July2006.
[4] V. Hatzivassiloglou, J.L.Klavans, and E. Eskin, Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In Proceedings of the jointsigdat conference on empirical methods in natural language processing and very large corpora (pp. 203-212),June 1999.
[5] X. L. Dong, A. Halevy, and C. Yu,Data integration with uncertainty. The VLDB Journal—The International Journal on Very Large Data Bases, 18(2), pp.469-500, 2009.
[6] S. H. A. El-Sappagh, A. M. A.Hendawi, and A. H. El Bastawissy, A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), pp.91-104, 2011.
[7] M. Lenzerini,Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246). ACM, June 2002 .
[8] M. Singh, and S.K. Jain, Transformation rules for decomposing heterogeneous data into triples. Journal of King Saud University-Computer and Information Sciences, 27(2), pp.181-192, 2015.
[9] A. Halevy, M. Franklin, and D. Maier, Principles of dataspace systems. In Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 1-9). ACM, June 2006.
[10] M. Franklin, A. Halevy, and D. Maier, From databases to dataspaces: a new abstraction for information management. ACM Sigmod Record, 34(4), pp.27-33,2005.
[11] C. Hedeler, K.Belhajjame, A. A Fernandes, S. M.Embury, andN. W. Paton, Dimensions of dataspaces. In Dataspace: The Final Frontier (pp. 55-66). Springer Berlin Heidelberg, 2009.
[12] H. T. Mirza, L. Chen, and G. Chen, Practicability of Dataspace Systems. JDCTA, 4(3), pp.233-243, 2010.
[13] J. P. Dittrich, andM. A. V. Salles, iDM: A unified and versatile data model for personal dataspace management. In Proceedings of the 32nd international conference on Very large data bases (pp. 367-378). VLDB Endowment,September 2006.
[14] J. P. Dittrich, L.Blunschi,M.Färber, O. R. Girard,S. K. Karakashian, and M. A. V. Salles, From Personal Desktops to Personal Dataspaces: A Report on Building the iMeMex Personal Dataspace Management System. In BTW (pp. 292-308), 2007.
[15] N. Dessì,and B. Pes,Towards scientific dataspaces. In Web Intelligence and Intelligent Agent Technologies, WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on (Vol. 3, pp. 575-578). IET, September 2009.
[16] I. Elsayed, and P. Brezany, Towards large-scale scientific dataspaces for e-science applications. In Database Systems for Advanced Applications (pp. 69-80). Springer Berlin Heidelberg, April 2010.
[17] H. Li, and J. Liu, Research on Heterogeneous Data Exchange based on XML. Physics Procedia, 25, pp.1382-1387, 2012.
[18] P. TalebiFard, and V.C. Leung, A data fusion approach to context-aware service delivery in heterogeneous network environments. Procedia Computer Science, 5, pp.312-319, 2012.
[19] L. Alben, Defining the criteria for effective interaction design. interactions, 3(3), pp.11-15, 1996.
[20] R. Z. Rebaï, F. Mnif, C. A.Zayani, andI. Amous, Adaptive Global Schema Generation from Heterogeneous Metadata Schemas. Procedia Computer Science, 60, pp.197-205, 2015.
[21] B. Alexe, L. Chiticariu, R. J. Miller, D. Pepper, and W. C. Tan, Muse: a system for understanding and designing mappings. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1281-1284). ACM, June 2008.
[22] A. Raffio, D. Braga, S. Ceri, P. Papotti, andM. A.Hernández, Clip: a tool for mapping hierarchical schemas. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1271-1274). ACM, June 2008.
[23] R. J. Miller, M. A. Hernández, L. M. Haas, L. L. Yan, C. H. Ho, R. Fagin,and L. Popa, The Clio project: managing heterogeneity. SIgMOD Record, 30(1), pp.78-83, 2001.
[24] J. Madhavan, P. A. Bernstein, and E. Rahm, Generic schema matching with cupid. In VLDB (Vol. 1, pp. 49-58),September 2001.
[25] S. Amir, I. M.Bilasco, andC. Djeraba, Mumie: Multi-level metadata mapping system. Journal of Multimedia, 6(3), pp.225-235,2011
[26] http://developer.nytimes.com/docs
[27] https://developers.google.com/apis-explorer/#p/

BIOGRAPHY

I. Carol is pursuing doctor of philosophy in Department of Computer Science, St. Joseph's College, (Autonomous), Tiruchirappalli, Tamil Nadu, India. He received his M. Phildegree from St. Joseph's College, Tiruchirappalli. He received his MCA degree from St. Joseph's College, Tiruchirappalli. He has published many research articles in the International conferences and journals. His area of interest is Data mining and Web mining.

S. Britto Ramesh Kumar is working as Assistant Professorin the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has published many research articles in the National/ International conferences and journals. His research interests include DataMining, Web Mining, and Mobile Networks.

DOI 10.17148/IJARCCE.2016.5295