Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 1249–1254

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Web Search and Data Mining (WSDM)

April 29 - May 2, 2019, Leuven, Belgium

Recovery of scientific data using Intelligent Distributed Data Warehouse

Amelec Viloria[a]*, Dionicio Neira Rodado[b], Omar Bonerge Pineda Lezama[c]

[a,b] *Universidad de la Costa (CUC), Barranquilla 080003, Colombia*

[c] *Universidad Tecnológica Centroamericana (UNITEC), Tegucigalpa 11101, Honduras*

Abstract

A Retrieval System requires several components that define its functionality and behavior. In the case of a meta-search engine for the retrieval of scientific data, a schema that defines the way to store such data is considered a necessary element for its evolution. Unified profiles have been developed for the data storage of the entities involved in the scientific data management, generated from the fact of publishing a scientific paper. Such profiles are considered the beginning of the generation of new components for the meta-search engine that, using the proprietary information, can deliver information relevant for the user of the tool. To this end, the use of an intelligent distributed data warehouse is proposed.

Keywords: scientific data; meta-data; meta-search engine; recovery of information; intelligent distributed data warehouse.

1. Introduction

The scientific data retrieval, as an activity framed in the discipline of information retrieval, has gained increasing interest in the last times Garciarena U. et al (2015)[1]. The impact of the Internet and related technologies has led to

* Corresponding author. Tel.: +573046238313.
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2019.04.180

the generation of large data sets derived from the actions mentioned above Bose, R. et al (2005)[2]. Likewise, these data led to the development of tools for the management, maintenance, publishing and processing. In this way, different publishers and associations, recognized by part of the scientific community, have published scientific data repositories in websites which constitute important consultation tools for the researcher-user. In addition, there are several tools of this type with a greater or lesser number of features that differ in data source, organization, and the processing performed on them Bhaduri K. et al (2008)[3]. In this context, a meta-search engine has arisen as an alternative Duan L. et al. (2009)[4] to operate on the area of computer science, to have access to a variety of sources for retrieving and sorting scientific data using an algorithm considering their impact on the scientific community Abhay K. et al (2017)[5]. In this tool, the search for integration of new functionalities and for improvements in both performance and efficiency requires the definition of a homogeneous structure for the storage of the retrieved scientific data, besides the consideration of issues related to the technology to be used, in an operating environment increasingly related to Big Data. In this document, the data retrieval approach is proposed from the Intelligent Distributed Data Warehouse (IDDW), which is a hierarchical distributed data store of N levels. The approach of data retrieval begins when the user enters the UIN, corresponding to the data store located in IDDW. Once the data store is located, the desired data are retrieved.

2. Related Works

Initially, the data retrieval techniques were restricted only to the centralized processing, as discussed by Agrawal R. et al. (1994)[6], Chiang D. et al. (2001)[7], Duan L. et al. (2009)[4]. According to Abhay K. et al (2015)[8], the retrieval of data from the distributed data store refers to the implementation of the classic procedure of retrieving data in a distributed computing environment that seeks to maximize the use of available resources (communication network, computers, and databases).Some algorithms and systems used for the distributed retrieval of databases are the following: the partition algorithm of Savasere A. et al. (1995)[9]; Multiagent system based on JAVA JAM by Stolfo S. et al. (1997)[10], Prodromidis A. et al. (2000)[11]; Grossman R. L. et al. (1999)[12] proposed the Papyrus, a JAVA-based system which aims to wide-area distributed data on clusters and meta-clusters; and the system based on Java for distributed enterprises by Chattratichat J. et al. (1999)[13].

The data retrieval in a highly parallel environment on multiple processors was explained by Wang L. et.al. (2013)[14]. There are two parallel programming models commonly used: Subprocesses (POSIX subprocesses by Butenhof D. R. (1997)[15]) and message passing (OpenMP by Duan L. et al. (2009)[4]). Modern programming languages are also structured to efficiently use innovative architectures. There are paradigms of parallel programming dedicated to parallelizing the algorithms on multiprocessor systems and in networks. OPENMP and MPI are used to achieve the parallelization of shared and distributed memory. CUDA is a programming language that is designed for parallel programming used by Garciarena U. et al (2015)[1]. In CUDA, the threads access different memories of the GPU. CUDA offers a model of data parallel programming. Parallel programming is incomplete without discussing the more recent approach called MapReduce, which can process large volumes of data in a highly parallel way, as shown by Bhaduri K. et al. (2008)[3]. Several data recovery algorithms have been modified for parallel processing architectures as discussed in [8].

3. Data recovery approach of IDDW

3.1. Flowchart of the approach

Flowchart of the data retrieval approach of IDDW shown in Fig. 1. It is the modified flowchart of the approach to store user data into the common table in the most suitable data store in IDDW presented in Abhay K. et al (2017)[5]. In the flowchart shown in Fig. 1, the content in bold is the new/modified content, the content that is not highlighted is not required to retrieve data from IDDW, and the rest of the part is the same as in the corresponding flowchart presented in Abhay K. et al (2017)[5].

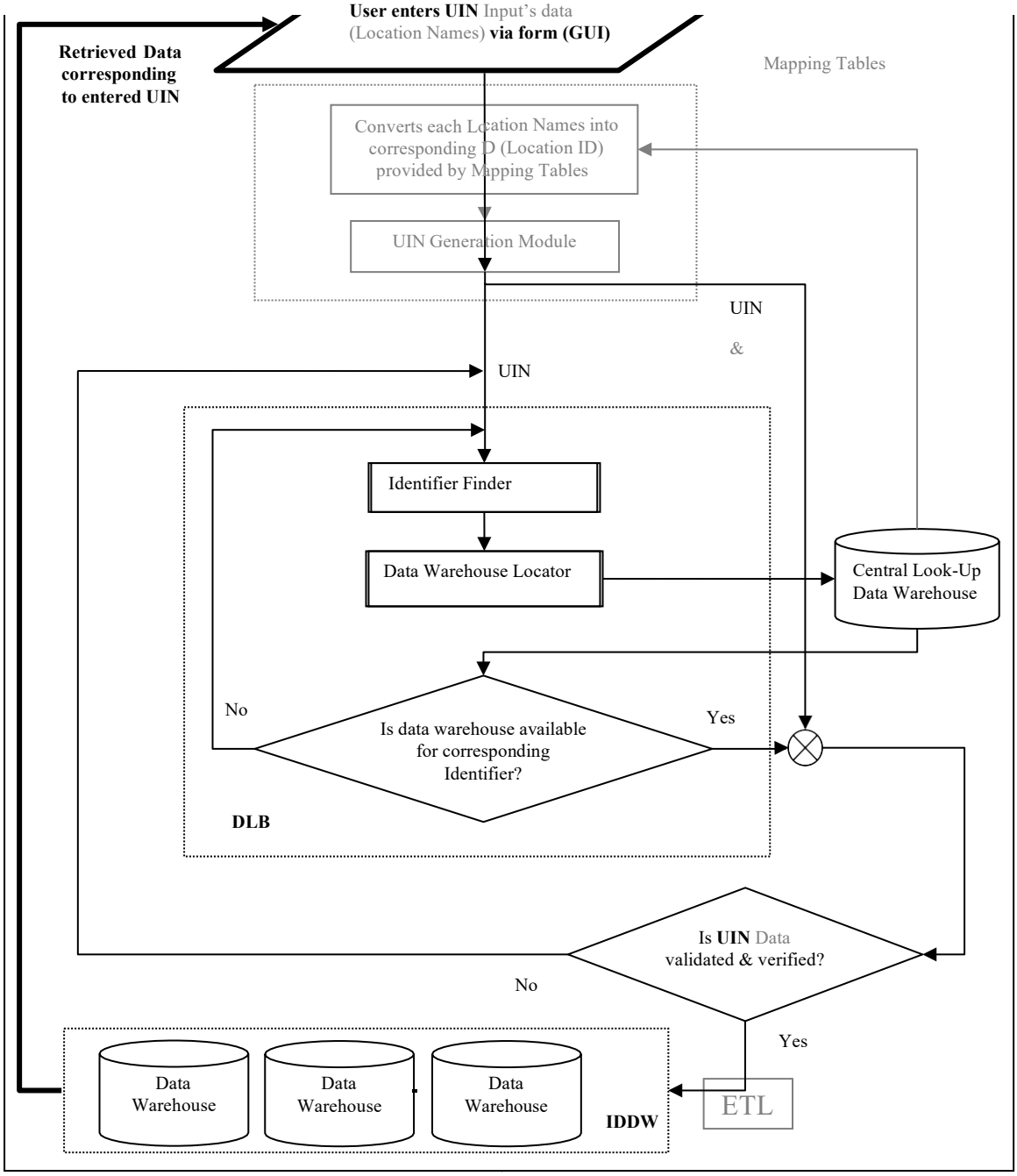

Fig. 1. Flowchart of the data retrieval approach from the IDDW Abhay K. et al (2017)[5].

4. Study Case

The 8-Level hierarchy structure presented in the study case taken in Abhay K. et al (2017)[5] is used here to assess the approach presented in this paper.

4.1. Design of the profiles

In order to generate the structure for the internal management of the scientific data, the researchers proceeded to examine a set of sources such as search engines and BD4 of a scientific nature, obtaining, from each of them, a list of meta-data used for the registry of the different entities with which they operate in Wang L., et. Al (2013)[14].

Tables 1, 2, and 3 show the attributes description of the main entities of the model: article, author and author profile, and publication source with their specializations for journals and scientific events (Gaitán-Angulo M. et al; 2018)[16], (Torres-Samuel M. et al; 2018[a])[17], (Torres-Samuel M. et al; 2018[b])[18], (Torres-Samuel M. et al; 2018[c])[19], (Vasquez C. et al; 2018)[20].

Table 1. Description of attributes of the entity Article

Attribute	Description
Title	Document title
Snippet	Description of the document
QuantityQuotations	Number of citations according to the consulted source
Keywords	Related key terms
Place Retrieval	Source from where the article was retrieved
TypeDocument	Identification if it is an article, book or pre-print
YearPublication	Year of publication of the document

Table 2. Description of the attributes of the entities Author and Author Profile

Attribute	Description
Surname	Author's last name
First name	Author's names
Alias	Set of names retrieved from the author
OriginProfile	Source from where the profile was retrieved
URL	Profile URL

Table 3. Description of the attributes of the entity Source Publication and Instances

Attribute	Description
Full name	Name without abbreviations of the source
Short name	Name with abbreviations
Initials	Acronym of the source
YearPublication	Year in which the magazine was published, or the event was held
NumbEdition	Issue number of the magazine / event
Volume / Location	Identification of the volume of the journal / location of the event
ISSN / ISBN	Identifier of the publication

4.2 Experimental configuration

The implementation of the presented approach for data retrieval from IDDW is carried out by writing the programs in JAVA with Net Beans: 6.9 as IDE, using the Apache Tomcat 6.0.26 web server. All the required tables are

integrated into the MySQL 5.0.45 database. The details of implementation are the same as those presented in Abhay K. et al (2017)[5].

4.3 Experimentation

In order to validate the operation of the IDDW when integrating the generated profiles, 50 queries were made, with a limit of 40 articles to be retrieved for each, executing the data retrieval processes mentioned in section 3. The effectiveness of the IDDW was evaluated in three aspects: a) in the storage of the links to the profiles, b) in the retrieval of the entity data, and c) in the registry of the relationships between the entities retrieved from the same document. The results obtained can be seen in Table 4.

Table 4. Results of the validation

Metrics	Value
Number of profiles to generate	900
Effectiveness of persistence	92%
Effectiveness of retrieval	78%
Effectiveness of the generation of relationships between entities	98%

Initially, the UIN "13302010410520017" is entered through the developed form Abhay K. et al (2017)[5]. The first identifier calculated by the identifier search engine for this UIN is "1330201041052001". The data store locator searches for the address of the machine corresponding to this identifier in the Central Look-Up data store tables. Levels of hierarchy H1 (see Table 5).

Table 5. The percentage of data correctly retrieved from the Common table of the data store located at different levels of hierarchy H1.

Level in hierarchy	Percentage (%)
1	90
2	91
3	93
4	95
5	96
6	98
7	99

5. Conclusions

In this work, the experimentation shows that the data from the data stores, available at various levels of IDDW, can be retrieved by using the user's UIN. It is concluded from an experiment on academic data that the average time to retrieve data is reduced from 6.4 milliseconds to 3.8 milliseconds when the data store is located at the lowest level of the hierarchy H1 compared to the data store located in the highest level of the hierarchy H1. Since the H1 hierarchy is formed from the hierarchical structure of 8 levels that is analogous to IDDW, it is concluded that, as the data store is built at a lower level of IDDW, the time elapsed to retrieve the data decreases. It is also observed from the results obtained when performing the experiment that the correct data retrieved also increase from 90% to 99% when the data store is located at the lowest level of IDDW compared to the data store located on top.

References

[1] Garciarena Ucelay, M.J., Villegas, M.P., Cagnina, L., Errecalde, M.L.: Cross domain author profiling task in spanish language: an experimental study. J. Comput. Sci. Technol. 15, no. 2, (2015).

[2] Bose, R., Frew, J.: Lineage retrieval for scientific data processing: a survey. ACM Computing. Surveys. CSUR. 37, 1–28 (2005).

[3] Bhaduri K., Wolf R., Giannella C., and Kargupta H., "Distributed decision-tree induction in peer-to-peer systems.", Statistical Analysis and Data Mining, Vol. 1, Issue 2, pp. 85–103, 2008.

[4] Duan L., Xu L., Liu Y. and Lee J., "Cluster-based outlier detection.", Annals of Operations Research 168, pp. 151–168, 2009.

[5] Abhay Kumar Agarwal and Neelendra Badal "Data Storing in Intelligent and Distributed Data Warehouse using Unique Identification Number" published in International Journal of Grid and Distributed Computing, Publisher: SERSC Australia, (ISSN: 2005-4262 (Print) ISSN: 2207-6379 (Online)), Volume 10, No. 9, pp. 13-32, September 2017.

[6] Agrawal R. and Srikant R., "Fast algorithms for mining association rules in large databases.", In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, VLDB, Chile, pp. 487–499, 1994.

[7] Chiang D., Lin C. and Chen M., "The adaptive approach for storage assignment by mining data of warehouse management system for distribution centre's.", Enterp. Inf. Syst, Vol. 5, Issue 2, pp. 219–234, 2001.

[8] Abhay Kumar Agarwal and N. Badal "A Novel Approach for Intelligent Distribution of Data Warehouses" published in Egyptian Informatics Journal-Elsevier, Egypt, (ISSN: 1110-8665), http://dx.doi.org/10.1016/j.eij.2015.10.002, Volume 17, pp. 147-159, October, 2015.

[9] Savasere A., Omiecinski E. and Navathe S., "An efficient algorithm for data mining association rules in large databases", In Proceedings of 21st Very Large Data Base Conference, pp. 432- 444, 1995.

[10] Stolfo S., Prodromidis A. L., Tselepis S., Lee W. and Fan D. W., "Jam: Java agents for meta- learning over distributed databases.", In Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining., pp. 74-81, 1997.

[11] Prodromidis A., Chan P. K., Stolfo S. J., "Meta learning in distributed data mining systems: Issues and approaches.", In Kargupta H., Chan P. (eds) Book on Advances in Distributed and Parallel Knowledge Discovery, AAAI/MIT Press, 2000.

[12] Grossman R. l., Bailey S. M., Sivakumar H. and Turinsky A. L., "papyrus: A system for data mining over local and wide area clusters and super-clusters.", In Proceedings of ACM/IEEE Conference on Supercomputing, Article No. 63, 1999.

[13] Chattratichat J., Darlington J., Guo Y., Hedvall S., Kohler M. and Syed J."An architecture for distributed enterprise data mining.", In Proceedings of 7th International Conference on High- Performance Computing and Networking, Netherlands, pp. 573-582, 1999.

[14] Wang L., et. al., "G-Hadoop: MapReduce across Distributed Data Centers for Data-Intensive Computing.", Future Generation Computer Systems, Vol. 29, Issue 3, pp. 739-750, 2013.

[15] Butenhof D. R., "Programming with POSIX threads.", Addison-Wesley Longman Publishing Company, USA, 1997.

[16] Gaitán-Angulo M., Cubillos Díaz J., Viloria A., Lis-Gutiérrez JP., Rodríguez-Garnica P.A. (2018) Bibliometric Analysis of Social Innovation and Complexity (Databases Scopus and Dialnet 2007–2017). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[17] Torres-Samuel M., Vásquez C.L., Viloria A., Varela N., Hernández-Fernandez L., Portillo-Medina R. (2018)[a] Analysis of Patterns in the University World Rankings Webometrics, Shanghai, QS and SIR-SCimago: Case Latin America. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[18] Torres-Samuel M, Carmen Vásquez, Amelec Viloria, Tito Crissien Borrero, Noel Varela, Danelys Cabrera, Mercedes Gaitán-Angulo, Jenny-Paola Lis-Gutiérrez. (2018)[b] Efficiency Analysis of the Visibility of Latin American Universities and Their Impact on the Ranking Web. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[19] Torres-Samuel M., Vásquez C., Viloria A., Lis-Gutiérrez JP., Borrero T.C., Varela N. (2018)[c] Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[20] Vásquez C, Maritza Torres-Samuel, Amelec Viloria, Tito Crissien Borrero, Noel Varela, Jenny-Paola Lis-Gutiérrez, Mercedes Gaitán-Angulo. (2018) Visibility of Research in Universities: The Triad Product-Researcher-Institution. Case: Latin American Countries. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham