SEBHA UNIVERSITY JOURNAL OF PURE & APPLIED SCIENCES VOL.20 NO. 2 2021
DOI: 10.51984/JOPAS.V20I2.1243

مجلة جامعة سبها للعلوم البحتة والتطبيقية
Sebha University Journal of Pure & Applied Sciences

Journal homepage: www.sebhau.edu.ly/journal/index.php/jopas

XML Datasets and Benchmarks for Performance Testing of the CLS Labelling Scheme

Alhadi A. Klaib

Department of Computer Science, Faculty of Information Technology, Elmergib University, Libya

Keywords:
CLS Labelling Scheme
Testing XML Labelling Schemes
XML Data
XML Datasets
XML Benchmarks

ABSTRACT

Extensible Markup Language (XML) has become a significant technology for transferring data through the world of the Internet. XML labelling schemes are an essential technique used to handle XML data effectively. Labelling XML data is performed by assigning labels to all nodes in that XML document. CLS labelling scheme is a hybrid labelling scheme that was developed to address some limitations of indexing XML data. Moreover, datasets are used to test XML labelling schemes. There are many XML datasets available nowadays. Some of them are from real life datasets and others are from artificial datasets. These datasets and benchmarks are used for testing the XML labelling schemes. This paper discusses and considers these datasets and benchmarks and their specifications in order to determine the most appropriate one for testing the CLS labelling scheme. This research found out that the XMark benchmark is the most appropriate choice for the testing performance of the CLS labelling scheme.

مجموعات البيانات لاختبار الأداء لتقنية CLS لفهرسة بيانات XML

الهادي علي كليب

قسم علوم الحاسوب، كلية تقنية المعلومات، جامعة المرقب، ليبيا

الكلمات المفتاحية:
تقنية CLS لفهرسة البيانات
اختبار تقنية CLS لفهرسة البيانات
بيانات XML
مجموعات بيانات XML
معايير XML

الملخص

أصبحت لغة الترميز الموسعة (XML) تقنية مهمة لنقل البيانات عبر عالم الإنترنت. تقنيات فهرسة XML هي تقنية أساسية تستخدم للتعامل مع بيانات XML بشكل فعال. يتم تنفيذ تسمية بيانات XML عن طريق تعيين تسميات لجميع العقد في مستند XML. تقنية CLS لفهرسة بيانات هو مخطط وسم مهجن تم تطويره لمعالجة بعض قيود فهرسة بيانات XML. علاوة على ذلك ، تُستخدم مجموعات البيانات لاختبار تقنيات فهرسة XML. هناك العديد من مجموعات بيانات XML المتاحة في الوقت الحاضر. بعضها من مجموعات بيانات حقيقية والبعض الآخر من مجموعات بيانات اصطناعية. تُستخدم مجموعات البيانات والمعايير هذه لاختبار تقنيات فهرسة XML. تناقش هذه الورقة وتدرس مجموعات هذه البيانات والمعايير ومواصفاتها من أجل تحديد أنسبها لاختبار تقنية CLS لفهرسة بيانات. وجد هذا البحث أن معيار XMark هو الخيار الأكثر ملاءمة لأداء اختبار تقنية CLS.

Introduction

XML was recommended in 1998 by the World Wide Web Consortium (W3C). Therefore, XML has become the dominant technology for transferring data across the internet. Indexing XML is a very important technique that used to improve XML data queries. The efficiency of the performance of any query in a database is based on indexing [1, 2]. labelling XML data is the technique used to index XML data efficiently. Labelling XML data is implemented by allocating labels to all nodes in that XML document. Every node is provided with a unique label that can be used to build the relationship among nodes in that XML tree [3, 4]. Many labelling schemes have been proposed [5-9]. However, none of these schemes meets all users' requirements, Therefore, they are only suitable for specific cases. An effective XML labelling scheme should give efficient query performance. These XML labelling are tested by using the XML benchmarks and XML datasets[10]. Various existing XML datasets and benchmarks are used for performance testing of XML labelling schemes[11]. Some of them are from real life datasets and others are from artificial datasets. These two types of datasets are

*Corresponding author:
E-mail addresses: alhadi.klaib@elmergib.edu.ly

Article History : Received 03 June 2021 - Received in revised form 07 July 2021 - Accepted 15 July 2021

used to assess the XML labelling schemes [12-18]. The CLS labelling scheme is a hybrid labelling scheme was proposed to improving the indexing XML data[11, 19, 20]. An investigation into the most used XML datasets is performed in this paper to pick up the most suitable dataset(s) for testing this scheme. The remainder of the paper is structured as follows: section two illustrates and describes the research method. Section three demonstrates the results. Section four demonstrates the analysis and discussion. Finally, section five discusses the conclusion.

Research Method

The research approach is based on two parts. First a review of the available and most common datasets and benchmarks in order to clarify their specifications and usabilities. Second, a review on the testing of labelling schemes to identify the suitable dataset.

Overview of the Existing Datasets and Benchmarks:

There are plenty of XML datasets and benchmarks are available for scientific purposes such as testing XML labelling schemes. Some of them are from real life datasets and others are from artificial datasets. Both these kinds of datasets are used for testing of the XML labelling schemes [12]. Further detail about these datasets and benchmarks as follows:

1 Actual XML Datasets

These datasets are based on real data. They are also called production/ existing/ experimental XML datasets. Examples of them are SwissPort, DBLP Computer Science Bibliography, University Courses, Auction Data, NASA, Treebank, Protein Sequence Database, SIGMOD Record, Mondial, and TPC-H Relational Database Benchmark [12, 21].

2 Benchmark/Standard Datasets

Also called artificial datasets. These benchmarks were developed for the purpose of evaluating queries. XML benchmarks are categorised into two groups, namely, micro benchmarks and application benchmarks. Micro benchmarks are used to evaluate specific parts of a system whereas application benchmarks are used to evaluate the performance of an XML database in general [12, 15, 18, 22]. Examples of them are: XMark Benchmark, TPox benchmark, Michigan Benchmark, XBench benchmark, and XMack-1 Benchmark. These XML benchmarks are intended for both query processing and storing data [12, 23]. More details about the most used datasets and benchmarks as follows:

[1] XMark Benchmark

Tthis was proposed in 2002 by Schmidt et al. [24] and is mainly used to evaluate XML applications. It is one of the most used XML benchmarks nowadays. The XMark has a data generator called *xmlgen*. This generator can produce an artificial XML document that is based on the DTD of an internet database. This generator is available free of charge at the XMark project website. XMark can recreate an XML database in different sizes. Thus, users can generate their own datasets that are appropriate for their requirements. In addition, the XMark datasets are used to evaluate the system performance and efficiently. Twenty queries are included in XMark and they are used to assess different aspects of searching in databases. These queries do not include the update processes [24].

[2] XOO7 Benchmark

Tthis was introduced by Carey et al. [17] and is called Object Oriented RDBMS benchmark (OO7). It was, subsequently, implemented on XML data by Li et al. [25]. The XML dataset that XOO7 creates is a separate XML file. This file is created in three different sizes – small, medium and large. This XML dataset has only up to five levels and offers twenty-three queries. These queries only handle search operations [17]. The XOO7 benchmark can be downloaded free of charge from its website [16].

[3] Michigan Benchmark

Also called MBench. Introduced by Runapongsa et al., this benchmark was designed as a micro benchmark in order to evaluate specific system components [15, 26]. This benchmark's dataset has forty-six queries and seven update processes [18]. It comes as an XML file that includes a number of nodes starting from 728,000 nodes and up to ten times more. In addition, this dataset has a limited depth which is sixteen levels, whereas the width is changeable. With regard to the queries, this benchmark has thirty-one queries that handle and evaluate many different features of databases containing update operations [26].

[4] XML Data Management Benchmark (XMach-1)

This was introduced by [14] and supports multi-users. The dataset of the XMach-1 includes a large number of XML files with sizes between 2 KB and 100 KB. The number of levels is limited up to six levels. The query set has eleven queries, three of them for update processes and the other eight queries for search processes. This benchmark is supported by web applications and is comprised of four parts, namely XML database, server, loader and client. The application servers provide XML document handling. The loaders handle the processes of detecting and loading the XML data from the database. The clients query and retrieve XML data [13]. Figure 1 shows the components of the XMach-1.

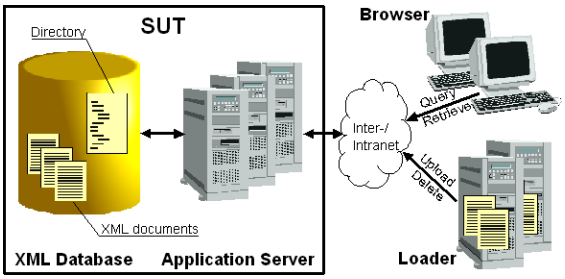

Fig. 1: XMach-1 components

[5] XBench Benchmark:

This creates a very wide range of XML files such as data centric and text centric. The database of this benchmark could be a single or multi XML file. The XBench is provided with a generator to generate XML data. This generator is built on ToXgene data generator and can generate different sizes of files from 10 MB to 10 GB. This benchmark has twenty queries that support search only without update [23, 27].

[6] TPoX Benchmark

TPoX stands for Transaction Processing over XML. This benchmark was designed to evaluate the whole system. The schema of the benchmark controls the size of the XML files. Regarding the XML database, it includes many small XML files with sizes between 2 KB and 20 KB. This benchmark has seventeen queries that mainly focus on updates [21].

A review on the testing of most relevant and common labelling schemes:

The details below discusses the performance testing of most common labelling schemes:

1 Testing the LLS scheme: the (*Beg*, *End*) was used in the experiments to show the improvement in the LLS. Two datasets were used to test the LLS which are as follows: the DBLP computer science Bibliography dataset and the XMark dataset. Furthermore, four kinds of XML queries were used to evaluate these schemes. These kinds of queries are the most common ones

used. The technique of mapping to relational database tables was used to evaluate XML queries. Each query was run twenty times and then the average was taken.

2 Testing the DDE scheme: three datasets, namely, XMark, Treebank and Nasa were used to test this scheme against two other labelling schemes. The two labelling schemes tested against the DDE were ORDPATH and Compact DDE (CDDE). The executed experiments are as follows: initial labelling, querying static document, querying dynamic document, and processing updates. The processing updates included uniform insertions and skewed insertions which were classified into order skewed insertions and random skewed insertions.

3 Testing the LSDX scheme: the XMark was used as a dataset for generating XML data. This scheme was tested against the GRP scheme (by Lu and Ling, 2004) and the SP scheme (by Cohen, Kaplan and Milo, 2002). The experiments that were performed are the following: length of labels, time used to generate labels, insertion and deletion time.

4 Testing the ORDPATH scheme: as for the dataset, the XMark and XMach-1 were used to test this scheme. Different measures were used such as arbitrary insertions, insert-friendly IDs, ORDPATH length. This scheme was compared with the Dewey order and others for evaluation purposes.

Existing labelling schemes have been tested in different techniques based on the aspects that will be assessed such as performance, scalability, and efficiency. Most of the labelling schemes are compared using different measures such as initial labelling, label size, creating labelling time, and the cost of updating. Usually, the proposed scheme is compared with one or more existing labelling schemes to show the improvements that this proposed scheme can deliver. Different experiments are designed according to the testing aims for testing. In order to compare a labelling scheme with others, the queries that these schemes support should be considered.

Results

Having studied the datasets and benchmarks, it's found out that the XMark benchmark is the most appropriate dataset for the testing experiments. The XMark benchmark helps both implementers and users to obtain insights into XML storage. XMark was chosen to test the CLS scheme for the following reasons: first and most importantly, the XMark was used to evaluate the performance of the LLS scheme, which is one of the schemes used to build the CLS scheme, by the founder. Thus, it would be appropriate to use the same dataset to evaluate the CLS scheme and compare the results [28]. Secondly, this benchmark is widely used to test XML queries and XML database performance [29]. Moreover, XMark is a good choice since it has many features such as providing a document generator to create documents in different sizes. Thus, users can generate datasets that are appropriate for their requirements [24]. Also, this benchmark provides a binary version of the XMark that can be run as an independent platform on any operating system. Moreover, XMark provides a broad range of queries – twenty-one in total. These queries are designed to evaluate different aspects of the datasets. They are divided into groups based on their goals and purposes. Table 1 shows these groups [24]. Query number 10 of the XMark queries was eliminated since it is irrelevant to the CLS scheme as this query is used to translate the results into another language. To conclude, the above XMark features offer a great choice for evaluating the CLS scheme.

TABLE 1: Xmark Benchmark Queries

No	Query number	Group name	Description
1	Q1	Exact match	Return the name of the person with ID 'person0'.
2	Q2	Ordered access	Return the initial increases of all open auctions
3	Q3		Return the first and current increases of all open auctions whose current increase is at least twice as high as the initial increase
4	Q4		List the reserves of those open auctions where a certain person issued a bid before another person.
5	Q5	Casting	How many sold items cost more than 40?
6	Q6	Regular path expression	How many items are listed on all continents?
7	Q7		How many pieces of prose are in our database?
8	Q8	Chasing references	List the names of persons and the number of items they bought. (joins person, closed auction)
9	Q9		List the names of persons and the names of the items they bought in Europe. (joins person, closed auction, item)
10	Q11	Joins on values	For each person, list the number of items currently on sale whose price does not exceed 0.02% of the person's income.
11	Q12		For each person with an income of more than 50,000, list the number of items currently on sale whose price does not exceed 0.02% of the person's income.
12	Q13	Reconstruct portions of the original XML document	List the names of items registered in Australia along with their descriptions.
13	Q14	Full text	Return the names of all items whose description contains the word 'gold'.
14	Q15	Path traversals	Print the keywords with a emphasis in annotations of closed auctions.
15	Q16		Return the IDs of the sellers of those auctions that have one or more keywords emphasised.
16	Q17	Finding missing elements	Which persons don't have a homepage?
17	Q18	Function application	Convert the currency of the reserves of all open auctions to another currency.
18	Q19	Sorting	Give an alphabetically ordered list of all items along with their location.
19	Q20	Aggregation	Group customers by their income and output the cardinality of each group.

Discussion

The XML dataset and benchmark used for testing the targeted schemes, there are many XML datasets available nowadays. Some of them are from real life datasets (a. k. a. Actual XML datasets) and others are from artificial datasets (a. k. a. Benchmark/standard datasets). These two types of datasets are used to evaluate the XML labelling schemes [12]. A review into the most commonly used XML datasets is performed in order to select the most suitable dataset(s) for testing the CLS scheme. Regarding real life datasets, these datasets are based on real data. Examples of them are: SwissPort, DBLP Computer Science Bibliography, University Courses, Auction Data, NASA, Treebank, Protein Sequence Database, SIGMOD Record, Mondial, and TPC-H Relational Database Benchmark [12]. Concerning the artificial datasets, these benchmarks were designed for the purpose of evaluating queries. XML benchmarks are categorised into two groups which are, micro benchmarks and application benchmarks. Micro benchmarks are used to evaluate specific parts of a system whereas application benchmarks are used to evaluate the performance of an XML database in general [12, 15, 18, 22]. Examples of these benchmarks are: XMark Benchmark, TPox benchmark, Michigan Benchmark, XBench benchmark, and XMack-1 Benchmark. These XML benchmarks are intended for both

query processing and storing data [12].

Conclusion

The aim of testing the CLS XML labelling scheme is to ensure it achieved the objectives. Furthermore, there are particular features that need to be tested, namely the query performance, efficiency of labelling XML documents, efficiency of scalability, and functionality of the proposed scheme. Thus, these features were used as critiria for selecting the suitable XML datasets. Therefore, the most common datasets and benchmarks were studied. This research found out that the XMark benchmark is the best choice for the testing experiments. The XMark features offer a great choice for evaluating the CLS scheme.

References

[1]- T. M. Connolly and C. E. Begg, Database systems: a practical approach to design, implementation, and management (no. Book, Whole). Boston, Mass: Addison-Wesley, 2010.

[2]- A. Klaib and J. Lu, "Investigation into Indexing XML Data Techniques," in Proceedings on the International Conference on Internet Computing (ICOMP), 2014: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 1.

[3]- J. Bosak and T. Bray, "XML and the second-generation Web," Scientific American, vol. 280, no. 5, pp. 89-93, 1999.

[4]- R. Elmasri and S. Navathe, Fundamentals of database systems (no. Book, Whole). Upper Saddle River: Pearson, 2016.

[5]- T. Amagasa, M. Yoshikawa, and S. Uemura, "QRS: A robust numbering scheme for XML documents," in Data Engineering, 2003. Proceedings. 19th International Conference on, 2003: IEEE, pp. 705-707.

[6]- E. Cohen, H. Kaplan, and T. Milo, "Labeling dynamic XML trees," SIAM Journal on Computing, vol. 39, no. 5, pp. 2048-2074, 2010.

[7]- T. Eda, Y. Sakurai, T. Amagasa, M. Yoshikawa, S. Uemura, and T. Honishi, "Dynamic range labeling for XML trees," in International Conference on Extending Database Technology, 2004: Springer, pp. 230-239.

[8]- P. O'Neil, E. O'Neil, S. Pal, I. Cseri, G. Schaller, and N. Westbury, "ORDPATHs: insert-friendly XML node labels," in Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 2004: ACM, pp. 903-908.

[9]- X. Wu, M. L. Lee, and W. Hsu, "A prime number labeling scheme for dynamic ordered XML trees," in Data Engineering, 2004. Proceedings. 20th International Conference on, 2004: IEEE, pp. 66-78.

[10]- A. Klaib and J. Lu, "Development of Database Structure and Indexing Technique for the Wireless Response System," in INFOCOMP 2013, The Third International Conference on Advanced Communications and Computation, 2013, pp. 110-116.

[11]- A. Ali Klaib, "Clustering-based Labelling Scheme-A Hybrid Approach for Efficient Querying and Updating XML Documents," University of Huddersfield, 2018.

[12]- A. Schmidt et al., "Why and how to benchmark XML databases," ACM SIGMOD Record, vol. 30, no. 3, pp. 27-32, 2001, doi: 10.1145/603867.603872.

[13]- T. Böhme and E. Rahm, "XMach-1: A benchmark for XML data management," in Datenbanksysteme in Büro, Technik und Wissenschaft, 2001: Springer, pp. 264-273.

[14]- T. Böhme and E. Rahm, "Multi-user evaluation of XML data management systems with XMach-1," in Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web: Springer, 2003, pp. 148-159.

[15]- J. M. P. Kanda Runapongsa, H. V. Jagadish, Yun Chen, Shurug Al-Khalifa "The Michigan Benchmark." Michigan University http://dbgroup.eecs.umich.edu/mbench/description.html (accessed 12/6/2014, 2016).

[16]- S. Bressan et al., "XOO7: Applying OO7 Benchmark to XML Query Processing Tools," in Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), 2001.

[17]- M. J. Carey, D. J. DeWitt, C. Kant, and J. F. Naughton, "A status report on the OO7 OODBMS benchmarking effort," ACM SIGPLAN Notices, vol. 29, no. 10, pp. 414-426, 1994.

[18]- I. Mlýnková, "Xml benchmarking: Limitations and opportunities," Technical Report, Department of Software Engineering, Charles University, Czech Republic, 2008.

[19]- A. A. KLAIB, "A NEW METHOD FOR QUERYING XML DATA."

[20]- A. Klaib and J. Lu, "Development of Database Structure and Indexing Technique for the Wireless Response System," in Proceedings of the Third International Conference on Advanced Communications and Computation. Infocomp. IARIA, Lisbon, Portugal, 2013: Citeseer, pp. 110-116.

[21]- M. Nicola, I. Kogan, and B. Schiefer, "An XML transaction processing benchmark," in Proceedings of the 2007 ACM SIGMOD international conference on Management of data, 2007: ACM, pp. 937-948.

[22]- D. Barbosa, A. Mendelzon, J. Keenleyside, and K. Lyons, "ToXgene: a template-based data generator for XML," in Proceedings of the 2002 ACM SIGMOD international conference on Management of data, 2002: ACM, pp. 616-616.

[23]- B. B. Yao, M. T. Özsu, and J. Keenleyside, "Xbench-a family of benchmarks for xml dbmss," in Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web: Springer, 2003, pp. 162-164.

[24]- A. Schmidt, F. Waas, M. Kersten, M. J. Carey, I. Manolescu, and R. Busse, "XMark: A benchmark for XML data management," in Proceedings of the 28th international conference on Very Large Data Bases, 2002: VLDB Endowment, pp. 974-985.

[25]- Q. Li and B. Moon, "Indexing and querying XML data for regular path expressions," in VLDB, 2001, vol. 1, pp. 361-370.

[26]- K. Runapongsa, J. M. Patel, H. Jagadish, Y. Chen, and S. Al-Khalifa, "The Michigan benchmark: towards XML query performance diagnostics," Information Systems, vol. 31, no. 2, pp. 73-97, 2006.

[27]- B. B. Yao, M. T. Ozsu, and N. Khandelwal, "XBench benchmark and performance testing of XML DBMSs," in Data Engineering, 2004. Proceedings. 20th International Conference on, 2004: IEEE, pp. 621-632.

[28]- S. Mohammad and P. Martin, "LLS: level-based labeling scheme for XML databases," in Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research, 2010: IBM Corp., pp. 115-127.

[29]- A. A. Almelibari, "Labelling Dynamic XML Documents: A GroupBased Approach," PhD, Computer Science, Sheffield University, Sheffield 2015. [Online]. Available: http://etheses.whiterose.ac.uk/8729/1/FinalThesis_Almelibari_100203910.pdf