INTEGRATING DATA WAREHOUSES WITH WEB DATA FOR OLAP USING SEMANTIC DATA CLUSTERING TECHNIQUES

1. Muthukumar.R
Research scholar,
Dept. of Computer Science,
Periyar University, Salem,
TamilNadu - India.

2. C. Chandrasekar
Associate Professor,
Dept. of Computer Science,
Periyar University,Salem,
TamilNadu - India.

ABSTRACT

Nowadays, Information retrieval plays an important role in the web. Many researches presented techniques for information retrieval process from databases. The previous work presented extended tree pattern clustering process for XML massive storages. This paper presents a new technique termed semantic data clustering (SDC) technique for combining the Data warehouse and web data for OLAP by retrieving the semantic data from DW (Data Warehouse). Usually, XML technologies are used to store, retrieve, integrate and combine the web data and the applications in Data warehouse. Using semantic data clustering technique, the semantic data repositories are retrieved from DW, which is the devise of multidimensional databases for XML data sources, and the XML additions of OnLine Analytical Processing (OLAP) techniques. SDC will efficiently tackle the information retrieval process in a DW to utilize text-rich document collections. For the XML data sources, the SDC will build the tree pattern for a clustered XML schema to retrieve the massive storage of data for OLAP. SDC uses clustering technique for building tree-pattern framework in order to use massive XML databases to data warehouse for OLAP. We also show the advantages of using semantic data clustering for building the tree-pattern in handling large amounts of XML documents for OLAP in data warehouse. A reliable performance improvement is achieved with the proposed SDC in XML database to data warehouse, compared to an existing ETC technique for XML storages, in terms of building time, query execution time for deriving the semantic data from DW, effectiveness of clustering process.

Keywords: Data warehouse, Repository, XML, OLAP, Tree pattern, clustering
1. INTRODUCTION

Nowadays, Web is the world’s largest source of information. It provides interoperability to a wide range of different applications and provides a means of information transaction between applications, as well as a semi structured data model for integrating information and knowledge. XML is a promising standard for representing the extensible environment. XML provides flexibility in recounting variable structure of data. It is becoming a typical data format for data exchange and data incorporation. XML is also known as an easy-to write and easy-to-parse language, which presents a method to replace data between varieties of applications on the Internet.

The formation of an XML document is done by the use of matching tag pairs (termed elements), and the information between matching tags is referred to as a content element. Besides, an element is allowed to have further attributes, where values are dispersed to the attributes in the start tag of the element. The two forms of XML document are document centric and data centric XML document.

Example for data centric xml
document  Example for document centric xml

XML documents can be connected with and authenticated against a schema, e.g., a Document Type Definition (DTD). The DTD of an XML document identifies the diverse elements that can be integrated in the document, how these elements can be nested, and the attributes they may contain.

The DW is defined as an integrated, time variant set of data based on management’s decisions. The information stored in a DW is typically demoralized by OLAP tools. OLAP
tools theoretically form the information as multidimensional cubes. In the cubes, data is separated into facts, the vital entities/events for the preferred study (e.g., a sale), and dimensions, which offer relative information for the facts (e.g., the products sold). Regularly, the dimensions are hierarchically planned into levels. XML documents are used to represent the data (metadata).

The main contribution of this work is to integrate the DW and web data using semantic data clustering technique by clustering the retrieved semantic data from DW and build the tree pattern framework to those clustered semantic data for OLAP.

2. LITERATURE REVIEW

The world’s largest source of information is web from which easy to access number of information we want. It has fetched interoperability to a broad range of diverse applications. This achievement has been probable to XML-based technology [3], which presents a means of information exchange between applications, as well as a partly structured data model for combining information and knowledge [1].

Web document repositories [4] faced the work on constructing warehouses for partially structured XML data to provide an efficient storage, query, acquisition, change control, and schema incorporation of data collected from Web sources. A number of approaches are developing around XML. These technologies comprises of XML Schemas [10], a substitute to DTDs that develops data typing and constrict capabilities and XQuery [5], the ordinary query language for XML documents, which presents dominant constructs for restructuring XML data.

XML technologies permit us to handle the heterogeneity variances that emerge at the physical and the syntactic levels. Many of the sources send their data in XML format in standard APIs like JDBC, ODBC, and SOAP [6] present platform-independent boundaries for querying the data sources.

The authors of [7] develop a different grouping operator for XML, where the grouping dimensions are précised in tree patterns [2]. This work [8] learns the summarizability troubles that happen when combined XML data and presents diverse algorithms for evaluating the data cubes well. To
extract the documents information related to (but not present in) the facts of the warehouse, the author [9] presented a framework. To enhance the interoperability of web data and DW, we introduce SDC technique by clustering the retrieved semantic data and build the tree pattern framework for those data in an efficient manner.

3. SEMANTIC DATA CLUSTERING FOR INTEGRATING THE DW AND WB DATA

The proposed SDC technique for integrating the data warehouse and web data is designed to build the tree pattern framework for the semantic data retrieved from DW in XML databases. The proposed SDC technique comprises of three processes. The first process is to identify and retrieve the semantic data from data warehouse. The second process describes the clustering process of retrieved semantic data. The third process is to build the framework for the clustered semantic data. These three processes are used to integrate the data warehouse and web data for OLAP techniques. The architecture diagram describing the three processes is described in fig. 3.1.

![Architecture Diagram of SDC in XML databases](image)
3.1 Retrieving semantic data from DW:

The first process in the proposed SDC technique is to retrieve the semantic data from the data warehouse. A semantic data is a data, which can be inferred significantly without human intervention. Semantic data is organized in groups of three parts, i.e., two objects and its relationships. The main goal of semantic data is to characterize the actual world as precisely as possible within the datasets. Using semantic data, it is easy to interact with the database without any interruption. The procedure for retrieving the data from DW is described below:

Step 1: Data Warehouse (large collection of data)
Step 2: Identify the semantic data (Entities and its relationships)
Step 3: Retrieve the semantic data
Step 4: Output: Retrieved semantic data

3.2 Process of SDC for integrating the web data and DW:

The next process is to cluster the retrieved semantic data and to build the tree pattern framework. The procedure describes the process of SDC technique for integrating the web data and DW. The sample data briefly describe the process of SDC technique.

Step 1: Input: (section 3.1 output)
Consider a sample set of XML schema as:

```xml
<xs:schema>
  <xs:element name="book">
    <xs:element name="title" type="xs:string"/>
    <xs:element name="author" type="xs:string" minOccurs="0" maxOccurs="2"/>
  </xs:element>
</xs:schema>
```

Step 2: Apply SDC technique,
Step 2.1: Cluster the retrieved semantic data as C1, C2,……,Cn.

Cluster 1

```xml
<book>
  <title> The Republic
</title>
</book>
```

Cluster 2

```xml
<book>
  <title> The Warrior</title>
  <author> Millman
</author>
</book>
```

Cluster 3

```xml
<book>
```
After retrieving the semantic data from DW, the clustering process is taken place to cluster the semantic data and SDC built the tree structure efficiently for every clustered semantic data for integrating the web data and DW in XML databases for OLAP process. Using SDC technique, it is easy to access the data on web applications for OLAP.

4. EXPERIMENTAL EVALUATION

Extensive experimental studies have been conducted to examine the proposed SDC tree-pattern clustering approach. We have implemented the SDC tree-pattern clustering approach in Java, and approved out a series of performance experiments in order to monitor the effectiveness of the approaches. The experiments were run on an Intel P-IV machine with 2 GB memory and 3 GHz dual processor CPU. The data sets were accumulated on the
local disk. We ran our experiments using different sets of XML data and queries. The proposed SDC tree-pattern clustering approach efficiently designed for building the tree pattern framework for the clustered semantic data from DW. This SDC technique successfully integrated the web data with DW for OLAP in xml databases. The performance of the proposed SDC technique for integrating the web data with DW by building the tree pattern framework for clustered semantic data is measured in terms of

i) Building time
ii) Clustering effectiveness
iii) Query execution time for retrieving semantic data

The effectiveness of clustering process is evaluated by how the semantic data be clustered in a particular interval of time. The efficiency of cluster is evaluated based on the data that has been clustered effectively using the eqn 4.1.

\[ X = \frac{1}{n} \sum_{a=1}^{n} \max_{a=b} \left( \mu_a + \mu_b / d(D_a, D_b) \right) \]  

Where \( n \) is the number of clusters, \( D_a \) is the centroid of cluster \( a \), \( \mu_a \) is the average distance of clustered data of \( a \) to centroid \( D_a \), and \( d(D_a, D_b) \) is the distance between centroids \( D_a \) and \( D_b \).

Eqn 4.1 described the efficiency of clustering process. Lower the value \( X \), higher the efficiency would be. The building time of the tree pattern framework (eqn 4.2) is defined as the time taken to cluster the retrieved semantic data and time taken to build the tree structure framework.

\[ T_{total} = \sum_{n} T_{retrieving data (n)} + T_{clustering process (n)} + T_{building time for tree pattern (n)} \]  

The query execution time is evaluated by the starting time of the query being processed and the ending time of the query after processing. An ending time is the time before the outcome of the query is displayed. It is calculated as,

\[ t_1(q_1) - t_2(q_1) \]  

Eqn 4.3, at a particular interval of time \( t \), let \( t_1 \) is the starting time of the query and \( t_2 \) be the ending time of the query and assume the query be \( q_1 \). The difference between the \( t_1(q_1) \) and \( t_2(q_1) \) is evaluated and compared with an existing EDC technique.
5. RESULTS AND DISCUSSION

In this work, we have seen how the tree pattern framework is efficiently built with the proposed SDC technique for integrating the DW and web data using semantic data with an existing ETC technique for XML storages written in mainstream languages such as Java. We used different sets of semantic data for comparing the results of the proposed SDC technique for integrating the DW and web data using semantic data with an existing ETC technique for XML storages.

<table>
<thead>
<tr>
<th>Collection of data in DW</th>
<th>Proposed SDC technique</th>
<th>Existing ETC technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>200</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>300</td>
<td>20</td>
<td>39</td>
</tr>
<tr>
<td>400</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>500</td>
<td>24</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 5.1 Comparison table describes the time taken to retrieve semantic data

The above table described the time taken to retrieve the semantic data from a collection of data available in DW. The time for retrieving is evaluated using the proposed SDC technique and compare the results with an existing ETC technique.

![Time Taken to retrieve semantic data](image_url)

**Fig 5.1 set of data in DW vs. time taken to retrieve semantic data**
From the fig 5.1, described the time consumed to retrieve the semantic data from DW. In the proposed SDC technique, the process of retrieving the semantic data from DW consumed less time compared to an existing ETC technique for massive xml storages. Comparison result of SDC with an existing ETC based on time consumption variance, measured in terms of milliseconds(m/s). The semantic data which has been retrieved from the DW is used for the process of integrating the DW and web data for OLAP. The performance graph of the proposed SDC in tree pattern building framework is shown in the fig 5.1. The variance in the time consumption for building the tree pattern would be 12-20% low in the proposed SDC.

<table>
<thead>
<tr>
<th>Retrieved data from DW</th>
<th>Proposed SDC technique</th>
<th>Existing ETC technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>300</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>400</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>500</td>
<td>42</td>
<td>25</td>
</tr>
</tbody>
</table>

**Table 5.2 Comparison table describes the efficiency of clustering process**

The above table described the efficiency of clustering process for a retrieved data in DW. The efficiency of cluster using the proposed SDC technique is compared with an existing ETC technique.

**Fig 5.2 Retrieved data vs. clustering efficiency**
From the fig 5.2, described the efficiency of clustering process for retrieved semantic data from DW. In the proposed SDC technique, the clustering process is efficient for the semantic data in DW compared to an existing ETC technique for massive xml storages. Comparison result of SDC with an existing ETC based on clustering efficiency shows that the proposed SDC technique perform well in the process of integrating the DW and web data for OLAP. The performance graph of the proposed SDC in clustering process is shown in the fig 5.2. The variance in the clustering effectiveness for building the tree pattern would be 20-25% high in the proposed SDC.

<table>
<thead>
<tr>
<th>Clustered data</th>
<th>Time taken to build tree pattern framework (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed SDC technique</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>150</td>
<td>10</td>
</tr>
<tr>
<td>200</td>
<td>15</td>
</tr>
<tr>
<td>250</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 5.3 Comparison table describes the time taken to build tree pattern framework

The above table described the time taken to build the tree pattern framework for the retrieved semantic data from DW. The time for building the tree pattern framework in the proposed SDC technique is compared with the results obtained in an existing ETC technique.
Fig 5.3, described the tree pattern building time for clustered semantic data from DW. In the proposed SDC technique, the building time is low for the clustered semantic data in DW compared to an existing ETC technique for massive xml storages. Comparison result of SDC with an existing ETC based on time consumption variance for tree pattern building time framework, measured in terms of milliseconds(m/s). The performance graph of the proposed SDC in tree pattern building time framework is shown in the fig 5.3. The variance in the for building the tree pattern would be 20-25% low in the proposed SDC.

Finally, we concluded that the proposed SDC technique for integrating the web data and DW is the best approach by retrieving the semantic data in less time and cluster the data efficiently and build the tree pattern framework for clustered data in less time. So, the proposed SDC technique achieved successfully based on the services done by the proposed SDC technique.

5. CONCLUSION

In this paper, we proposed SDC technique, which are simply modified to data models for efficiently accessing semantic data from DW. In summary, the proposed SDC approach has several desirable features: 1) it presented an easy method to cluster the semantic data from DW, 2) it efficiently built the tree structure for clustered semantic data, 3) it consumes less time for clustering process and tree pattern framework, 4) Integrated the web data and DW for OLAP in xml databases. The experimental results showed that the proposed SDC technique presented an efficient way for tree-pattern framework in XML semantic data from DW. Compared to an existing ETC technique for massive xml storages, the proposed SDC for integrating the web data and DW for OLAP in xml databases outperforms well and the effectiveness is 70-80% high.
6. AUTHORS PROFILE

Mr. R. Muthukumar obtained M.Sc., Computer Science from P.S.G College, Bharathiyar University, Coimbatore, Tamil Nadu, India, in 2000, and M.Phil., Computer Science from Manonmaniam Sundaranar University, Thirunelveli, Tamil Nadu, India in 2002. He was working as IT analyst, in Department of Software development, TATA Consultancy Services, Bangalore, Karnataka, India. Currently he is working as Tech Lead in Department of Software development, Infosys, Bangalore, Karnataka, India. He is pursing Ph.D in Data warehouse.

Dr. C. Chandrasekar completed his Ph.D in Periyar University, Salem at 2006. He worked as Head, Department of Computer Applications at K.S.R. College of Engineering, Tiruchengode, Tamil Nadu, India. Currently he is working as Associate Professor in the Department of Computer Science at Periyar University, Salem, Tamil Nadu. His research interest includes Mobile computing, Networks, Image processing and Data mining. He is a senior member of ISTE and CSI.

REFERENCES


