Xaggregation: Flexible Aggregation of XML Data

Hongzhi Wang, Jianzhong Li, Zhenying He, and Hong Gao

Abstract. XML is an important format of information exchange and representation. One of its features is that it has tag representing semantics. Based on this feature, an extensive aggregation of operation of XML data, Xaggregation, is represented in this paper. Xaggregation permits XPath expression decorating dimension property and measure property. With Xaggregation, statistics of XML data becomes more flexible with function of aggregating heterogeneous data and hierarchy data along some path of XML. Xaggregation could be embedded in query language of XML such as XQuery. In this paper, the definition of Xaggregation is presented, as well as the semantics and application of it. Implementation of Xaggregation based on native XML database with XML as tree structure is also designed.

1 Introduction

XML is an important information representation format. On one hand, because of its extendability, it is used widely as information exchange format. On the other hand, model of XML could be considered as semi-structured [1], information representation ability of which is stronger than that of tradition relational database. Therefore, XML database has its special meaning. XML warehouse is often used as cache of information integration system based on XML is an. XML database is also used as web database.

Today, one usage of database especial massive database is for decision support. Aggregation is a basic operation of decision support. There are many aggregations definition and implementation methods of relational database. But in XML database, the work of aggregation is quite little. [6, 7, 8] is a serious of work of a kind of aggregation related to XML. It uses XML data as decoration of dimension properties. In [12], an operation to implementation aggregation of XML stream is presented. This operation is to aggregate a series of XML documents into one based on their same parts, and do not touch the real statistics of the information contained in XML document.
Aggregation of data in XML format should be different from that of relational database in semantics. With stronger representation ability of tag representing semantics, aggregation of XML should be extended. Possible extensions of aggregation of XML data are:

- The aggregation of the same property should permit a special set of paths, such as a path set defined with XPath[2]. That is to say, the dimension property and measure could be a set of object in XML document. The object in various positions of XML document with various paths could be aggregated together.

- The structure of XML document should be considered, representing the complex way of aggregation. This process is quite like roll-up operation in common aggregation with the difference of multiple roll-up paths that make up of a complex structure.

With above extensions, the aggregation of information in XML format becomes flexible.

In aggregation on relational database, the property to be aggregated up is measure property and the property the aggregation based on is dimension property. This paper continues to use these definitions to describe new aggregation. In this paper, path of a node n in XML tree refers to the path from root of the tree to n. Node a nearer to root than node b is defined as a is higher than b.

In this paper, a special aggregation operation based on XML, Xaggregation, is presented. This operation permits XPath decorating dimension properties and measure properties. We believe this paper to be the first to consider this instance of aggregation in XML data.

However, the implementation of this kind of aggregation is more complex, because of not only complex process of aggregation, but also various logic and storage structures of XML data.

XML database has several storage structures. Mainly, four kind storage of XML databases are used, common file system, OO database, relational database and native XML database[11]. Native XML database is used more and more because its implementation is optimized specially for XML data. Relational database [9, 10] is also widely used to store and process query of XML data. But because of the core idea of storing XML data in relational database depends on decomposing of schema of XML data into tables. Xaggregation needs travel of path, as will bring out many join operations. The efficiency is affected. Therefore, the implementation of Xaggregation in this paper is based on native XML database with tree structure.

This paper focuses on the definition of aggregation on XML data as well as implementation of Xaggregation on native XML database. The contribution of this paper includes:

- Operation Xaggregation are defined. The usage of Xaggregation is presented by example.
- Implementation algorithm of Xaggregation is presented. The implementations are based on native XML database.

This paper is structured as follows. Section 2 presents the motivation of defining new aggregation operation. Xaggregations are defined in section 3. Implementation of