Revisiting “Fragmentation of XML Documents”

Hui Ma[1], Klaus-Dieter Schewe[2]

[1] Victoria University of Wellington, School of Engineering and Computer Science, Wellington, New Zealand
hui.ma@ecs.vuw.ac.nz
[2] Software Competence Center Hagenberg, Hagenberg, Austria
kdschewe@acm.org

Categories and Subject Descriptors: Information Systems [**Miscellaneous**]: Databases

Keywords: Distribution, Fragmentation, XML

The article “Fragmentation of XML Documents” in this issue is a reprint of our original contribution to the 2003 Brazilian Symposium on Databases [Ma and Schewe 2003]. It is one of our early publications devoted to the development of a sound theory of distribution design for complex-value databases, a notion we use as an umbrella for all kinds of databases with tree-structured objects such as nested relational, object oriented and also XML databases. Our goal was to extend and generalise the known approach by fragmentation and allocation – replication was not considered – for relational databases.

We started looking at fragmentation operations focusing on horizontal and vertical fragmentation, first for object oriented databases [Schewe 2002], then also for XML [Ma and Schewe 2003]. As fragmentation is coupled with algebra operations selection and projection, while defragmentation uses union and join, we added a third kind of fragmentation operation based on splitting with dereferencing needed for defragmentation.

In addition, we considered a cost-based approach in order to determine an optimal fragmentation. This lead to defining heuristics for determining when further fragmentation may not be helpful [Ma 2003]. Our first steps only considered horizontal fragmentation [Ma and Schewe 2005], which was later generalised to the other cases [Ma et al. 2007; Ma et al. 2007b].

When looking at cost-based fragmentation and allocation we discovered an unpleasant circularity in the problem: we can only obtain a good estimate for cost efficiency if we assume to deal with optimised queries, but optimising queries requires knowledge of the fragmentation and fragment allocation [Ma and Schewe 2006; Ma et al. 2007]. This lead us to try an approach, which we now call the *variational approach to fragmentation and allocation* [Ma et al. 2007a].

The approach starts from the assumption of a given set of optimised query trees on an optimal allocation of database fragments, and asks how a single elementary fragmentation operation would impact on the allocation of the new fragments and on the query trees by means of another round of query optimisation. We could show that the effect on the query trees is small, as only projection-selection-subqueries would be affected. As a consequence the problem of query optimisation including optimal allocation of intermediate results could be largely separated from the fragmentation and allocation problem. With the completion of Hui Ma’s Ph.D. [Ma 2007] we did not continue this line of research, though we felt that the approach still requires further investigation, in particular with respect to its mathematical rigour.

Throughout the research programme we frequently encountered two major counter-arguments: Firstly, distribution design is a largely solved problem, and secondly, distribution of XML documents is the "wrong" problem, as the inter-operation of XML documents that are de facto distributed over the web is the much more relevant problem. As for the first of these arguments we discovered through our research that distribution design is to the contrary a largely unsolved problem, even for relational databases. In the literature one of the motivations for distribution is an expected performance gain, but no proof for this is given. In particular, connections with query optimisation are rare and the circularity of the problem is widely ignored.

As for the second argument we did indeed redirect our interest to data-intensive services and cloud computing. However, when many tenants address services offered by a cloud, the problem of distribution (and replication) resurfaces naturally. Furthermore, as many services offered via the web are XML-based, the distribution of XML documents (or better: XML databases) becomes unavoidable. We therefore believe that in due time our new line of research on web-based data-intensive services will bring us back to picking up the thread on distribution design that we abandoned at the end of 2007.

REFERENCES

Ma, H. *Distribution Design in Object Oriented Databases.* M.S. thesis, Massey University, 2003.

Ma, H. *Distribution Design in Complex-Value Databases.* Ph.D. thesis, Massey University, 2007.

Ma, H. and Schewe, K.-D. Fragmentation of XML documents. In *Proceedings of the Brazilian Symposium on Databases.* Manaus, Brazil, pp. 200–214, 2003.

Ma, H. and Schewe, K.-D. A Heuristic Approach to Horizontal Fragmentation in Object Oriented Databases. In J. Barzdins and A. Caplinskas (Eds.), *Selected Papers from the Sixth International Baltic Conference DB&IS2004.* Frontiers in Artificial Intelligence and Applications, vol. 118. IOS-Press, pp. 20–33, 2005.

Ma, H. and Schewe, K.-D. Query optimisation as part of distribution design for complex value databases. In *Proceeding of the Conference on Information Modelling and Knowledge Bases.* Trojanovice, Czech Republic, pp. 289–296, 2006.

Ma, H., Schewe, K.-D., and Kirchberg, M. A Heuristic Approach to Fragmentation Incorporating Query Information. In O. Vasilecas, J. Eder, and A. Caplinskas (Eds.), *Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference DB&IS'2006.* Frontiers in Artificial Intelligence and Applications, vol. 155. IOS-Press, pp. 103 – 116, 2007.

Ma, H., Schewe, K.-D., and Wang, Q. Distribution design for higher-order data models. *Data and Knowledge Engineering* 62 (2): 400–434, 2007a.

Ma, H., Schewe, K.-D., and Wang, Q. A heuristic approach to cost-efficient derived horizontal fragmentation of complex value databases. In *Proceedings of the Eighteenth Australasian Database Conference.* Ballarat, Australia, pp. 103–112, 2007b.

Schewe, K.-D. Fragmentation of object oriented and semi-structured data. In *Proceedings of the 5th International Baltic Conference on Databases and Information Systems.* Tallinn, Estonia, pp. 1–14, 2002.