

Schema Summarization *

Cong Yu
Department of EECS
University of Michigan
congy@eecs.umich.edu

H. V. Jagadish
Department of EECS
University of Michigan
jag@eecs.umich.edu

ABSTRACT

Real database systems can often be very complex. A person wishing to access data from an unfamiliar database has the daunting task of understanding its schema before being able to pose a correct query against it. A schema summary can be of great help, providing a succinct overview of the entire schema, and making it possible to explore in depth only the relevant schema components.

In this paper we formally define a schema summary and two desirable properties (in addition to minimizing size) of a summary: presenting important schema elements and achieving broad information coverage. We develop algorithms that allow us to automatically generate schema summaries based on these two goals. We further develop an objective metric for assessing the quality of a schema summary using query information. Experimental evaluation using this metric demonstrates that the summaries produced by our algorithms can significantly reduce the amount of user effort required to formulate a query through schema exploration.

1. INTRODUCTION

Real databases often have extremely complex schemas. However, a complex schema is difficult to comprehend, limiting the database accessibility (in terms of both querying and data exchange) to a small number of people, who have spent a significant amount of time understanding the schema. Consider the example schema based on the XMark [12] benchmark in Figure 1. The schema is small compared to that of most production databases, and a significant portion of the schema has in fact been suppressed. Even so, a user unfamiliar with the XMark dataset will take time to figure out the major themes of the schema.

Typically, a user has a query need that depends on a portion of the schema. But to be able to express this query need, the user has to study the entire complex schema and discover the schema elements of interest. For example, a user who wishes to find out the end time for an open auction in the XMark database, has to study the schema and filter

*Supported in part by NSF under grant IIS-0438909, and NIH under grant 1-U54-DA021519.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the publisher, ACM.

VLDB '06, September 12-15, 2006, Seoul, Korea.

Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09

away irrelevant information about items and persons. These problems become much worse for more complex schemas, especially when the schema can no longer be presented to the user in its entirety at a reasonable information density [15].

In this paper, we propose the notion of *schema summary* to address the above problems. As shown in Figure 2(A), a summary utilizes *abstract elements* and *abstract links* to summarize a complex schema and provide the users with a concise overview for better understanding. Each abstract element in the summary corresponds to a cluster of original schema elements (and other lower level abstract elements in the case of a multi-level summary), and each abstract link represents one or more links between the schema elements within those abstract elements. A user presented with the summary in Figure 2(A) can immediately understand that the schema is about auctions, along with the items and persons related to the auctions. Furthermore, if the user is interested in only information about open auctions, she can selectively expand that abstract element, as shown in Figure 2(C). She will then get more detailed information for that particular part of the schema alone, without being exposed to other unrelated details.

While schema summaries are useful, creating a good summary is a non-trivial task. A schema summary should be concise enough for users to comprehend, yet it needs to convey enough information for users to obtain a decent understanding of the underlying schema and data. Consider the two schema summaries in Figure 2(A & B). While both have the same number of abstract elements, A is intuitively a better summary than B because A informs the user about the *bidder* element in the schema, which corresponds to much more information (i.e., the bidders) in the database than the *region* element in B. In this paper, we capture this intuition formally through the notions of *summary importance* and *summary coverage*. Along with the intuitive notion of *summary size*, they describe the effectiveness of a summary for a complex schema and the database associated with it.

Humans can be good at summarization, and the database designer could generate a schema summary at the time the schema is being specified. In fact, a design summary (usually expressed in ER diagram) is sometimes created as part of the design process. However, such internal documents are unlikely to be made available, in a heterogeneous environment, to external users who may be permitted database access. On the other hand, it is exactly such users who would benefit most from having a schema summary available. Therefore, we have little choice but to generate summaries from existing databases. One possible approach is to generate summaries manually, but this is labor-intensive, and is impractical in

