Optimizing Information Agents by Selectively Materializing Data *

Naveen Ashish
Information Sciences Institute, Integrated Media Systems Center and
Department of Computer Science
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
ashish@isi.edu

There is currently great interest in building information gathering agents that provide integrated access to data from a number of distributed Web sources. Some of the prominent research projects in this area include The Internet Softbot, Information Manifold, InfoMaster and InfoSleuth. A critical problem faced by such agents is a high query response time, due to the fact that a lot of data from Web sources has to be retrieved, extracted, and integrated to answer queries. The aim of this work is to develop an approach for improving query response time in information agents. The resulting system is being developed as part of the Ariadne project (Knoblock *et al.* 1998).

Our approach to optimizing an agent is to selectively materialize data from the different sources being integrated. In order to speed up query response time we materialize information that is frequently queried or used to construct answers to user queries. A key problem is to determine the classes of information that must be materialized. There are several factors that must be taken into account in order to determine such classes, including frequency of access by users, cost of getting the information from the actual Web sources, space required to store the data, and the frequency of updates at the original Web sources.

We have developed an algorithm that identifies classes of information to materialize by analyzing patterns in user queries. An important constraint is that the number of information classes materialized must be kept small. This is because each class materialized is defined as another information source that the agent can access. Having a large number of such sources will create performance problems for the query planner that must generate plans to retrieve information requested by the user from the various sources being integrated. Our strategy is to replace whenever possible, a number of fragmented information classes in which all the data is frequently accessed, by a single class that covers the data in all these classes and possibly some extra data not accessed at all. The advantage is that in this manner we are able to keep the number of materialized classes small. This algorithm thus attempts to cover frequently accessed data using the minimum number of information classes with the constraint that the extra data (data not frequently accessed by user queries but which is part of the classes identified by the algorithm) is within a certain threshold value. The algorithm utilizes the KR system LOOM to maintain a dynamically constructed ontology of subclasses of information frequently queried. The ontology guides the merging of a set of classes into a single class.

There are several other important issues that we will address as part of this dissertation. First, although the materialized data can be of great help in speeding up query processing time, we have to take into account the fact that the data in the original Web sources might change. Developing solutions to this problem requires considering several factors such as at what frequency the data in a Web source changes, is the time of change known, is the data stored affected by changes in the source, and also how critical it is for the users to have the absolute current data. Second, another important feature that must be present in an integration environment is to be able to materialize data in terms of classes that are an integrated or aggregated view over classes of information in individual sources. This is complicated because the integrated classes and the individual source classes cannot be analyzed independently for materialization.

Acknowledgments

I am grateful to my research advisor Craig Knoblock for guidance with this work. I also wish to thank Dennis McLeod and Cyrus Shahabi of USC, and Richard Hull and Gang Zhou of Bell Laboratories. This work has been supported by USC's Integrated Media Systems Center (IMSC), an NSF Engineering Research Center.

References

Knoblock, C.; Minton, S.; Ambite, J.-L.; Ashish, N.; Modi, P. J.; Muslea, I.; Philpot, A.; and Tejada, S. 1998. Modeling web sources for information integration. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence.*