A Performance Analysis of Semantic Caching for XML Query Processing

Boris Novikov, Department of Computer Science, Saint-Petersburg University, Saint-Petersburg, Russia

Alice Pigul, Department of Computer Science, Saint-Petersburg University, Saint-Petersburg, Russia

Anna Yarygina, Department of Computer Science, Saint-Petersburg University, Saint-Petersburg, Russia

ABSTRACT

Caching is important for any system attempting to achieve high performance. The semantic caching is an approach trying to benefit from the certain knowledge of data semantics. The authors expect that this information might enable reuse of semantically close data rather than exactly equal to cached data in the traditional system. However, the major obstacle for extensive application of semantic caching for any data model or query language is the computational complexity of the query containment problem, which is, in general, undecidable. In this article the authors introduce and compare three approximate conservative query matching algorithms for semantic caching of semi-structured queries. The authors then analyze their applicability for distributed query processing. Based on this analysis, the authors outline few scenarios where semantic caching can be beneficial for query processing in a distributed system of heterogeneous semi-structured information resources.

Keywords: Distributed Systems, Extensible Markup Language (XML), Heterogeneous Resources, Information Systems, Performance, Query Processing, Semantic Caching, Semi-Structured Data

INTRODUCTION

The need to uniformly access heterogeneous information resources triggered a huge amount of research in the area of information integration performed in last decades. Several industrial strength integration software tools are currently available. However, the query evaluation performance delivered by these tools is still far from practically acceptable for any complex queries and is often unpredictable. Common techniques for performance improvement, such as indexing, query optimization, and caching, are well understood and developed for traditional databases but should be revised and re-evaluated in the integrated distributed heterogeneous environment.

The focus of this research is on advanced semantic caching techniques which are expected to be especially good in a distributed hetero-
In contrast with usual caching techniques, in addition to the usage patterns, semantic caching takes into account relationships and inter dependencies between data items to improve cache effectiveness and extend the ability of local query processing of cached data instead of distributed query processing.

The goal of this research is to evaluate the potential impact of different approaches to semantic caching and obtain quantitative estimations of semantic caching suitability for XML query processing.

The caching is known as an important technique for any system attempting to achieve high performance. The idea of any caching is to reduce the number of repetitive accesses to the data items located in relatively slow large storage. Instead, copies of data items are temporarily stored in a fast (and relatively small) memory.

In almost all implementations cache is completely transparent for the application or system that uses the cached data. That is, the presence of cache affects performance, but not the functionality of the application, and, moreover, the application cannot explicitly control the behavior of the cache.

In this research, the data items copied from the primary data source into a cache are called cache units. Depending on the nature of cache units, physical, logical, and semantic caching are distinguished (Figure 1).

In a common physical caching the cache units are pages, blocks, or other low-level data structures defined for a physical storage. Physical caching is hardly applicable for distributed environment because low level units are not related to the units transferred over a network, and page sizes may vary from site to site.

When a logical cache is used, high level objects, such as web pages, are cached instead of physical blocks. A potential disadvantage of a logical caching is that the cache unit is considered atomic and can be used only for exactly the same request.

The caching is called semantic caching if the cache units are meaningful items or sets in the high-level application data model. Thus, parts of a cache unit may be used to process queries other than original one, or data from several cache units may be used together to evaluate a new query.

The semantics of data in a semantic cache is usually represented with queries (Chidlovskii...
Tacit Knowledge: Conceptualizations and Operationalizations
www.igi-global.com/article/tacit-knowledge-conceptualizations-operationalizations/2708?camid=4v1a

Knowledge Identification and Acquisition in SMEs: Strategically Emergent or Just Ad Hoc?
www.igi-global.com/article/knowledge-identification-and-acquisition-in-smes/148315?camid=4v1a