Chapter 34
GeoCache:
A Cache for GML Geographical Data

Lionel Savary
PRiSM Laboratory, France

Georges Gardarin
PRiSM Laboratory, France

Karine Zeitouni
PRiSM Laboratory, France

ABSTRACT

GML is a promising model for integrating geodata within data warehouses. The resulting databases are generally large and require spatial operators to be handled. Depending on the size of the target geographical data and the number and complexity of operators in a query, the processing time may quickly become prohibitive. To optimize spatial queries over GML encoded data, this paper introduces a novel cache-based architecture. A new cache replacement policy is then proposed. It takes into account the containment properties of geographical data and predicates, and allows evicting the most irrelevant values from the cache. Experiences with the GeoCache prototype show the effectiveness of the proposed architecture with the associated replacement policy, compared to existing works.

INTRODUCTION

The increasing accumulation of geographical data and the heterogeneity of Geographical Information Systems (GISs) make difficult efficient query processing in distributed GIS. Novel architectures (Zhang, 2001) (Gupta, 1999) (Leclercq, 1999) (Chen, 2000) (Paolucci, 2001) (Corocoles, 2003) (Boucelma, 2002) (Stoimenov, 2000) (Voisard, 1999) are based on XML, which becomes a standard for exchanging data between heterogeneous sources. Proposed by OpenGIS (OpenGIS, 2003), GML is an XML encoding for the modeling, transport, and storage of geographical information including both the spatial and non-spatial fragments of geographical data (called features). As stressed in (Savary, 2003), we believe that GML is a promising model for geographical data mediating and warehousing purpose.

DOI: 10.4018/978-1-4666-2038-4.ch034

By their nature, geographical data are large. Thus GML documents are often of important size. The processing time of geographical queries over such documents in a data warehouse can become too large for several reasons:

1. The query evaluator needs to parse entire documents to find and extract query relevant data.
2. Spatial operators are not cost effective, especially if the query contains complex selections and joins on large GML documents.

Moreover, computational costs of spatial operators are generally more expensive than those of standard relational operators. Thus, geographical queries on GML documents raise the problem of memory and CPU consumption. To solve this problem, we propose to exploit the specificities of a semantic cache (Dar, 1996) with an optimized data structure. The proposed structure aims at considerably reducing memory space by avoiding storing redundant values. Furthermore, a new cache replacement policy is proposed. It keeps in cache the most relevant data for better efficiency.

Related works generally focus on spatial data stored in object-relational databases (Beckmann, 1990). The proposed cache organizations are better suitable for tuple-oriented data structures (Brinkhoff, 2002). Most cache replacement policies are based on Least Recently Used (LRU) and its variants. Other cache replacement policies proposed in the literature (Lorenzetti, 1996) (Cao, 1997) (Arlitt, 1999) deal with relational or XML databases, but have not yet investigated the area of XML spatial databases.

The rest of the paper is organized as follows: Section 2 gives an overview of related works. Section 3 presents our cache architecture adapted for GML geographical data. Section 4 discusses about the inference rules of spatial operators and presents an efficient replacement policy for geographical data considering inference between spatial operators. Section 5 shows some results of the proposed cache implementation and replacement policy. Finally, the conclusion summarizes our contributions and points out the main advantages of the proposed GML cache-based architecture.

RELATED WORKS

Cache Replacement Policy

In the literature, several approaches have been proposed for cache replacement policy. The most well-known is the Least Recently Used (LRU, Tannenbaum, 1992). This algorithm replaces the document requested the least recently. Rather at the opposite, the Least Frequently Used (LFU) algorithm evicts the document accessed the least frequently. A lot of extensions or variations have been proposed in the context of WWW proxy caching algorithms. We review some in the sequel.

The LRU-Threshold (Chou, 1985) is a simple extension of LRU in which documents larger than a given threshold size are never cached. The LRU-K (O'Neil, 1993) considers the time of the last K references to a page and uses such information to make page-replacement decisions. The page to be dropped is the one with a maximum backward K-distance for all pages in the buffer. The Log(size)+LRU (Abrams, 1995) evicts the document with the largest log(size), and apply LRU in case of equality. The Size algorithm evicts the largest document. The Hybrid algorithm aims at reducing the total latency time by computing a function that estimates the value of keeping a page in cache. This function takes into account the time to connect with a server, the network bandwidth, the use frequency of the cache result, and the size of the document. The document with the smallest function value is then evicted. The Lowest Relative Value (LRV) algorithm includes the cost and the size of a document in estimating the utility of keeping it in cache (Lorenzetti, 1996). LRV evicts the document with the lowest utility value.

Related Content

Improving Accessibility Through VGI and Crowdsourcing
Igor Gomes Cruzand Claudio E.C. Campelo (2017). *Volunteered Geographic Information and the Future of Geospatial Data (pp. 208-226).*
www.irma-international.org/chapter/improving-accessibility-through-vgi-crowdsourcing/178806

Linking Effective Whole Life Cycle Cost Data Requirements to Parametric Building Information Models Using BIM Technologies
Dermot Kehily, Trevor Woodsand Fiacra McDonnell (2013). *International Journal of 3-D Information Modeling (pp. 1-11).*
www.irma-international.org/article/linking-effective-whole-life-cycle-cost-data-requirements-to-parametric-building-information-models-using-bim-technologies/105902

Land Governance and the Information Society
Mohamed Timoulali (2019). *Geospatial Technologies for Effective Land Governance (pp. 15-27).*
www.irma-international.org/chapter/land-governance-and-the-information-society/214477

Creating an Interactive Web Map: A Service-Learning Project Aligned to the Geospatial Technology Competency Model
Lesli M. Rawlings (2015). *International Journal of Applied Geospatial Research (pp. 110-125).*
www.irma-international.org/article/creating-an-interactive-web-map/129811

Rapid Evaluation of Arid Lands (REAL): A Methodology
Daniel P. Dugas, Michael N. DeMers, Janet C. Greenlee, Walter G. Whitfordand Anna Klimaszewski-Patterson (2013). *Geographic Information Systems: Concepts, Methodologies, Tools, and Applications (pp. 1541-1558).*
www.irma-international.org/chapter/rapid-evaluation-arid-lands-real/70521