Available online at www.sciencedirect.com
ScienceDirect
The Journal of Systems and Software 80 (2007) 1102–1119
The Journal of Systems and Software
www.elsevier.com/locate/jss

A context-aware cache structure for mobile computing environments

Stylianos Drakatos [a,*], Niki Pissinou [a], Kia Makki [a], Christos Douligeris [b]

[a] *Telecommunication and Information Technology Institute, Department of Electrical and Computer Engineering, Florida International University Miami, FL 33174, USA*
[b] *Department of Informatics, University of Piraeus, Piraeus 18534, Greece*

Received 14 April 2006; received in revised form 5 October 2006; accepted 10 October 2006
Available online 28 November 2006

Abstract

This paper proposes a cache management method that maintains a mobile terminal's cache content by prefetching data items with maximum benefit and evicting cache data entries with minimum benefit. The data item benefit is evaluated based on the user's query context which is defined as a set of constraints (predicates) that define both the movement pattern and the information context requested by the mobile user. A context-aware cache is formed and maintained using a set of neighboring locations (called the *prime list*) that are restricted by the validity of the data fetched from the server. Simulation results show that the proposed strategy, using different levels of granularity, can greatly improve system performance in terms of the cache hit ratio.
Published by Elsevier Inc.

Keywords: Context-aware information service; Mobile computing; Location-awareness; Caching; Prefetching

1. Introduction

Data caching as a means to achieve higher performance has been a perpetual quest in the computer industry. Most recently and due to the significantly increased size of memory available for small devices, caches are used as small data base systems that hold content most likely to be used in the near future (this is called prefetching or hoarding). However, due to the size limitation of the cache, cache management strategies exist to efficiently manage the cache content. The location information is used as a key field of the *user's query context*, but not enough attention has been given to the other query fields (predicates) which define the user's information context. Emerging location-based service (LBS) providers use the location information of mobile users to provide them with relevant information based on their geographical positions. Information disseminated to mobile users potentially can be context-sensitive and highly personalized. Therefore, an effective cache management scheme needs to adapt dynamically to the user's query context. Additionally, both the cached data items and the prefetched ones should be determined and adjusted according to the user's movement pattern and information context.

In evaluating the data item's benefit as far as the cache content is concerned, we propose a scheme that uses the query context as an information filtering mechanism to limit the amount of prefetched information to the data items with maximum benefit. A main aspect of this work involves predicting the future context that will be required by the user. In some situations forecasting may be impossible, but in situations where the content is changing gradually and continuously i.e., in continuous type of queries, this may be possible and very effective. Forecasting may be done, for example, by analyzing the user's current query context. The purpose of trying to predict future contexts is to anticipate the user's future retrieval needs, and to perform retrievals in advance of the need. Assuming the

* Corresponding author. Tel.: +1 305 8541009; fax: +1 305 3484126.
E-mail addresses: drakatos@fiu.edu (S. Drakatos), pissinou@fiu.edu (N. Pissinou), makkik@fiu.edu (K. Makki), cdoulig@unipi.edu (C. Douligeris).

0164-1212/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jss.2006.10.027

prediction is correct, the response to retrieval requests will then be very fast, since the necessary retrieval will have been done in advance. When a cache-miss happens, the mobile terminal (MT) asks for several other items and not just the cache-missed data item, with little additional cost. This action will prevent future cache misses and will reduce the number of uplink requests.

A mobility-based semantic cache structure and query processing was first proposed in Dar et al. (1996). Ren and Dunham (2000) and Ren et al. (2003) have extended this work to use the location information attached to each segment, making it more efficient and they have also proposed a cache management replacement policy. This work is based on this previous body of research on semantic cache management, however it focuses on the cache management prefetching strategy. To design an effective cache management strategy we consider the neighboring cells with valid information and the current query information context factors. Based on these two factors, we outline the major contributions of this paper below:

- The key contribution of this study is a *context-aware* (CA) cache management prefetching strategy, which first uses the validity of the data (*valid scope distribution*) based on their location to derive a set of most likely future cells called the "prime" list of cells. Next, in order to identify data items with high benefit as far as the cache content is concerned the user's query context is exploited to limit the amount of prefetched information within the predicted set of future cells (*the prefetching zone*).
- A direct result of the proposed strategy is the formation and maintenance of the context-aware cache of data items with a high cache value which are included at a low cost. The context-aware cache is then updated if the mobile user subsequently strays out of the predefined prime list of cells.
- The performance of the cache management strategy is examined through a number of experiments which show that the context-aware prefetching produces significant improvements over the standard direction velocity based prefetching strategies.

The rest of the paper is organized as follows: Section 2 gives a brief description of location-dependent data caching and prefetching strategies-related research. Section 3 presents the mobile architecture together with the mobility and query models. This section also explains the semantic cache description at the mobile client. Section 4 presents the proposed cache management strategy and explains its associated components, i.e. the presentation of the query-context prefetching method and prefetching algorithm. Section 5 presents a prefetching cost analysis. Section 6 describes the simulation model and performance comparison is discussed in Section 7. Finally, Section 8 provides conclusive remarks and future plans regarding the presented research.

2. Previous work

Research on cache management has been active over the past few years. The least recently used (LRU) replacement policy, which evicts the object that has not been accessed for the longest time, works well when the most recently referenced objects are most likely to be referenced again in the near future. Ren and Dunham (2000) have proposed a mobility model to represent the moving behavior of mobile users and have formally defined location-dependent queries. Based on their mobility model, they developed a Location Dependent Data (LDD) semantic caching scheme called Furthest Away Replacement (FAR), which implies future location prediction based on tangent velocity. A body of previous research (Lee et al., 2002; Zheng and Lee, 2002; Xu et al., 2003), incorporated the valid scope as an important factor in data caching replacement policy using the geographical mobility models. Akyildiz et al. (1996) and Levine et al. (1997) proposed the use of the *shadow cluster* model for future location prediction and used it for resource allocation in asynchronous transfer mode (ATM)-based wireless networks. Akyildiz and Wang (2004) have enhanced their approach by including user profiling. They proposed a matrix of transition probabilities using the shadow cluster model, considering all possible cell locations, for a total of $3k^2 + 3k + 1$ cells, each assigned to one state, where k is the number of rings. However, this number of states explodes quickly as k increases, sometimes making analysis and simulation difficult and costly. A good amount of redundancy is built into this approach because not all of these cells are of interest if they have no valid answers for the query. Mao and Douligeris (2000) have used the random movement model with cell location granularity in their location-based mobility management research.

In Liu et al. (1998), a two-level user mobility model is used to represent the movement behavior at global and local levels. The next cell is predicted by considering speed and direction of a user's trajectory. Through estimation of mobile users' trajectory and arrival/departure times in Aljadhai and Znati (2001), a group of future cells are determined, which constitute the most likely cluster into which a terminal will move. Most recently, Park et al. (2004) have proposed prefetching policies based on the current position and the velocity of the MT and used the value of the tangent velocity to predict the future location of the MT. This method uses the *geographical mobility models* and is effective only within a short time interval Δt. In order to limit the amount of prefetched information, current research has used geographical mobility models to focus solely on the mobile user's movement pattern (Ren and Dunham, 2000; Duhham and Kumar, 1998; Zheng and Lee, 2002; Akyildiz et al., 1996). The *geographical mobility models* inherently use continuous calculations of the tangent velocity, which is proven to have considerable high processing overhead (Satyanarayana, 2002; Datta et al., 2003). In summary, most of the existing methods are aimed at