Fall 2010

December 9, 2010

Jeppe Rishede Thomsen
Department of Computing
Hong Kong Polytechnic University

Abstract

All the papers either solve the problem of a more efficient cache in a specific domain, or use the network domain, which are both relevant, but not really useful when looking at shortest path caching.

The papers show some interesting ways to use cache, but ultimately their approaches are very domain or query specific so their approaces to caching and cache replacement/invalidation can not be applied directly.

1. Introduction

2. Problem

2.1. Definitions and problem setting

We assume a setting where owners of mobile, positioning enabled, devices want route planning assistance. We assume users prefer online route planning services over offline solutions. we expect users to use network enabled capable of determining and visualizing users location and route. Users want fast response times from online services, comparable to using an offline application [**?**] Using a cache reduces the computational burden [**?**] on an online service, providing faster end-user response time [**?**] by both freeing up computational resources to calculate new routes, as well as being able to immediately provide the shortest path result from the cache. We assume a scenario using only server side caching.

2.2. model

TODO: ***add more examples on advantages/disadvantages of*** $\{LRU, ImpBaseline, OSC\}$
TODO: ***Fig.4- add about text: space per edge***
TODO: ***Fig.4- add about text: probability cache item is useful*** We use a uniform random model(URM)

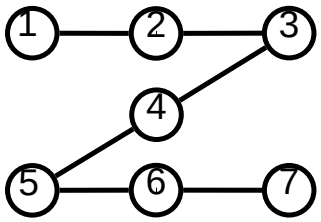

Figure 1. Simple map

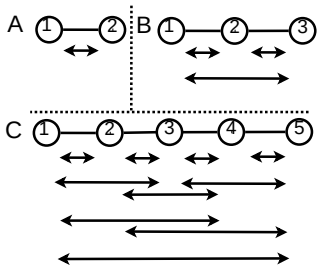

Figure 2. Number of queries possible

together with a simple 'map',the graph in figure 1, enable us to clearly argue about the advantages expected when using server side caching of shortest-path queries. By using the simple map (fig.1) and a URM together with we can reason about the probabilities that a specific query will occur. Figure 2 illustrates the number of queries possible for a map with 2,3, and 5 locations for figure 2A, 2B, and 2C respectively. Each line underneeth each of the simple graphs represents two possible queries (A->B, B<-A). The number of shortest-path queries possible on a tree-graph with n vertices is $n * (n-1)$. The probability of seeing any one query, q_i is then $P(q_i|n) = (n * n - 1)^{-1}$

2.3. methods

For the sake of simplicity the methods presented in this section will all be based on figure 1 and 3. Figure

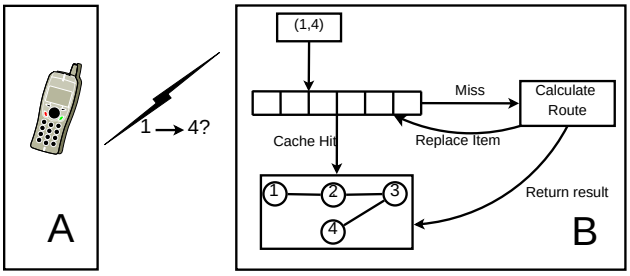

Figure 3. simple graph

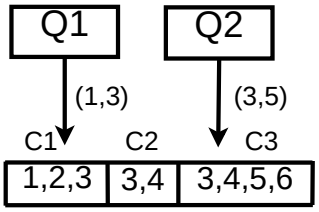

Figure 4. Queries

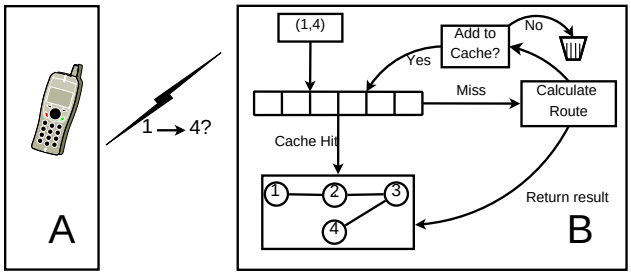

Figure 5. Advanced graph

1 shows a simple graph which we will use as our map and figure 3 shows the simple scenario in which a user (fig. 3A) issues a route-planning query from 1 to 4 (fig. 3) to an online route-planning server with a build in cache (fig. 3B).

2.3.1. Baseline. The strait forward baseline solution is illustrated in figure 3B. The idea is a server side shortest path cache which will store each query result in the cache and only consider exact query matches as cache hits, and to only use a simple cache policy such as LRU or FIFO. The advantage of this solution is clear: it is simple and easily implemented. This simplicity is however obviously also it's main disadvantage, as it is too simple and very inefficient in terms of the utility the cache provides. Using items in the cache only when there is an exact match makes it exceedingly unlikely to get a cache hit due to the nature of route planning (many people share parts of routes, but few the same start and end points) and the sheer number of start-/end-point combinations possible.

2.3.2. Improved Baseline. One way to possible increase the utility of a naive cache as proposed in 2.3.1 would be to exploit the *optimal substructure property* [1] of the cache items. There is a significant increase in cache hits to be expected by utilizing the optimal substructure of shortest path cache items since it is unlikely many people will plan a route from/to the same place, but it is very likely that some sub-parts will be shared among users, and some users' full path laying within a longer path already calculated. The idea is illustrated in figure 4 where the baseline method would be able to answer query Q1 from the cache, but not Q2. It is this specific disadvange which ImpBaseline addresses and ImpBaseline can therefor answer both Q1 and Q2 from the cache since the result of Q2 now exist as a solution to a subpath of cache item C3. Doing this adds a need for aditional computational resources **TODO:** ***how many resources?*** required to examine the substructure of each cached shortest path search result. It is currently not known if doing this is worth the effort, compared to just calculating the route, possibly multiple times.

2.3.3. OSC - Optimal Substructure Cache. OSC is more advanced than the two previous proposed solutions and therefor the scenario has been updated in figure 5. To further improve upon Improved Baseline we will again utilize the optimal substructure, making it possible to have much fewer items in cache and still retain a high cache hit percentage [**?**]. Results with sub-paths shared by many users and longer, rather than short, paths are prefered to increase the utility of the cached shortest-path results.

By adding a more intuitive cache replacement policy which takes in to consideration both the usage of each cache item, as well as the coverage of previously often seen queries it is likely that the utility of the cache would be much higher. This addition is shown with the addition of the "'add to cache"' box in figure 5B, added to show a huristic[1] will be used instead of a very simple method like LRU.

3. Related work reference

reference support for related work section.

3.0.4. On effective presentation of graph patterns: a structural representative approach. They develop an approach that combine two focuses when mining

1. the actual huristic will ofcause only be defined later

patterns in graphs. 1. they introduce a method to relax the tightness of the pattern subgraph pattern matching, so they can have high support for subgraphs which are very similar, but not exact. 2. as many mining approaches return allot (often very similar) patterns, they propose a method to collapse similar patterns so the user is presented with something that is easier to get an overview of and gain an understanding of the data. [2]

3.1. Cache Invalidation and Replacement Strategies for Location-Dependent Data in Mobile Environments

They develop two cache replacement and invalidation techniques for mobile clients communicating with a LBS. They argue that in the setting of spatial data and LBS then it is important to consider more than just the access time when doing cache replacement. They look at the spatial area where an object in the cache is valid as well as the direction the user is moving. They do this besides calculating the probability that this object will be accessed again.

Assumes all POI objects are fixed size and no updates will be made. [3]

3.2. Nearest-Neighbor Caching for Content-Match Applications

[4]

3.3. Caching Content-based Queries for Robust and Efficient Image Retrieval

They study how to do caching with Content-based Image Retrieval, and they support range and kNN queries. They focus on how to do caching when many of the queries are similar, but not the same (e.g. picture cropped or color changes) without polluting the cache. Their approach works in metric space and they develop an approximate method to check if the result can be satisfied by the cache. They archive good results, getting few direct cache hits, but still satisfying many queries from similar queries in the cache.

[5]

3.4. Caching Complementary Space for Location-Based Services

They develop the notion of Complementary Space(CS) to help better use a cache on a mobile client. CS is different levels for representing the objects on a map within MBRs. At the lowest level they just show the object, and as the levels go up they include more and more objects within MBRs, looking at the trade of in communication up/down link from a mobile client. They always have the entire world represented within the clients cache, at different levels, and offer no solution to how they will handle server updates to the map.

This is very similar to [6], although the approach does not formally depended on an R-tree, they still use one and offer no viable alternative, which lessens the difference even more. Their results are better than their competitors, including [6], though it seems that they stop their graphs just before [6] beats them.

3.5. Proactive Caching for Spatial Queries in Mobile Environments

They develop an approach which uses the index of an R-tree to add context to a cache of spatial object on a mobile client. They develop several communication and space saving techniques by representing less important parts of the R-tree in more compact ways, or just not storing the lower nodes/leaves of the tree. They also formally prove the asymptotic bounds of their algorithms.

[6]

3.6. Aggregate Aware Caching for Multi-dimensional Queries

They develop a method to re-use items in a cache for a data werehouse. They take advantage of the levels of aggregation which exist in OLAP query results and come up with a way where they can get aggregated results using full or partial more detailed data from the cache. The use most of the paper on showing and proving that their algorithms have a good running time and admit them selfes that the approach is not very mature, though they still manage to get resonable results. [7]

3.7. DynaMat: A Dynamic View Management System for Data Warehouses

abstract: Pre-computation and materialization of views with aggregate functions is a common technique in Data Warehouses. Due to the complex structure of the warehouse and the different profiles of the users who submit queries, there is need for tools that will automate the selection and management of the materialized data. In this paper we present DynaMat,

a system that dynamically materializes information at multiple levels of granularity in order to match the demand (workload) but also takes into account the maintenance restrictions for the warehouse, such as down time to update the views and space availability. DynaMat unifies the view selection and the view maintenance problems under a single framework using a novel ąğgoodnessą́ measure for the materialized views. DynaMat constantly monitors incoming queri and materializes the best set of views subject to the space constraints. During updates, DynaMat reconciles the current materialized view selection and refreshes the most beneficial subset of it within a given maintenance window. We compare DynaMat against a system that is given all queries in advance and the pre-computed optimal static view selection. The comparison is made based on a new metric, the Detailed Cost Savings Ratio introduced for quantifying the benefits of view materialization against incoming queries. These experiments show that DynaMat's dynamic view selection outperforms the optimal static view selection and thus, any sub-optimal static algorithm that has appeared in the literature. [8]

3.8. Cache-Oblivious Data Structures and Algorithms for Undirected Breadth-First Search and Shortest Paths

[9]

3.9. Cached Shortest-Path Tree: An Approach to Reduce the Influence of Intra-Domain Routing Instability

They assume a network setting and try to reduce the time and computational load it takes when network topology changes, as well as prevent any links from being unreachable if the topology changes often. The propose a cache with shortest-path trees, arguing that even if the topology changes often, then it is mostly between the same configurations (e.g. a computer/router is turned off/on) meaning that a cache with the most common seen configurations will be able to drastically reduce the amount of computation needed to recalculate routing tables.

[10]

3.10. On Designing a Shortest-Path-Based Cache Replacement in a Transcoding Proxy

[11]

3.11. Optimizing Graph Algorithms for Improved Cache Performance

[12]

References

[1] T. H. Cormen, C. E. Leiserson, and C. Stein, *Introduction to Algorithms*, 3rd ed. MIT Press, 2009.

[2] C. Chen, C. X. Lin, X. Yan, and J. Han, "On effective presentation of graph patterns: a structural representative approach," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 299–308.

[3] B. Zheng, J. Xu, and D. L. Lee, "Cache invalidation and replacement strategies for location-dependent data in mobile environments," *IEEE Trans. Comput.*, vol. 51, no. 10, pp. 1141–1153, 2002.

[4] S. Pandey, A. Broder, F. Chierichetti, V. Josifovski, R. Kumar, and S. Vassilvitskii, "Nearest-neighbor caching for content-match applications," in *WWW '09: Proceedings of the 18th international conference on World wide web*. New York, NY, USA: ACM, 2009, pp. 441–450.

[5] F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti, "Caching content-based queries for robust and efficient image retrieval," in *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*. New York, NY, USA: ACM, 2009, pp. 780–790.

[6] H. Hu, J. Xu, W. S. Wong, B. Zheng, D. L. Lee, and W.-C. Lee, "Proactive caching for spatial queries in mobile environments," in *ICDE '05: Proceedings of the 21st International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 403–414.

[7] P. Deshpande and J. F. Naughton, "Aggregate aware caching for multi-dimensional queries," in *Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology*, ser. EDBT '00. London, UK: Springer-Verlag, 2000, pp. 167–182. [Online]. Available: http://portal.acm.org/citation.cfm?id=645339.650140

[8] Y. Kotidis and N. Roussopoulos, "Dynamat: A dynamic view management system for data warehouses," in *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA*, A. Delis, C. Faloutsos, and S. Ghandeharizadeh, Eds. ACM Press, 1999, pp. 371–382.

[9] G. S. Brodal, R. Fagerberg, U. Meyer, and N. Zeh.

[10] S. ZHANG, K. IIDA, and S. YAMAGUCHI, "Cached shortest-path tree : An approach to reduce the influence of intra-domain routing instability," *IEICE transactions on communications*, vol. 86, no. 12, pp. 3590–3599, 2003-12-01. [Online]. Available: http://ci.nii.ac.jp/naid/110003221599/en/

[11] H.-P. Hung and M.-S. Chen, "On designing a shortest-path-based cache replacement in a transcoding proxy," *Multimedia Systems*, vol. 15, pp. 49–62, 2009.

[12] J.-S. Park, M. Penner, and V. K. Prasanna, "Optimizing graph algorithms for improved cache performance," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 9, pp. 769–782, 2004.