CACHE REPLACEMENT STRATEGIES FOR MOBILE DATA CACHING

Preetha Theresa Joy[1] and K. Polouse Jacob[2]

[1,2] Department of Computer Science, Cochin University of Science and Technology, Kochi, Kerala, India.

ABSTRACT

Data caching is an important technique in mobile computing environments for improving data availability and access latencies particularly because these computing environments are characterized by narrow bandwidth wireless links and frequent disconnections. Cache replacement policy plays a vital role to improve the performance in a cached mobile environment, since the amount of data stored in a client cache is small. In this paper we reviewed some of the well known cache replacement policies proposed for mobile data caches. We made a comparison between these policies after classifying them based on the criteria used for evicting documents. In addition, this paper suggests some alternative techniques for cache replacement.

KEYWORDS

Cache Replacement, MANET, Cooperative caching ,Wireless Mobile Network.

1. INTRODUCTION

Mobile wireless networks are predominantly of two types, architecture based and architecture-less. In both types of networks, the wireless links may be of low bandwidth and subject to frequent disconnections, leading to weakly connected mobile clients. Consequently, mobile clients often can be disconnected from their data servers. Another characteristic of mobile computing environment is the severe constraint on the availability of resources at the mobile node. A typical node in such environments has limited power and processing resources. In spite of these limitations a mobile user would like to have some connection transparency –in the sense that he would like to have easy access to data vital to the application. The major challenges to ensure high data availability in a mobile computing environment is to reduce bandwidth and resource utilization.

Caching and prefetching is an effective technique to reduce the impact of low bandwidth and intermittent wireless links in a mobile environment. By caching frequently requested data items, bandwidth can be conserved as it eliminates repetitive data transfers for the same data item to different mobile nodes. The data management schemes developed for architecture based wireless network cannot be use directly to solve the data management problems in a MANET as they are inherently peer to peer networks with high node mobility. Cooperative caching has been used in this type of networks to provide more cache space and faster speeds.

1.1 Cache Replacement Policies

When the cache is full, an object has to be removed from the cache to make room for the data that has to be brought in .While it would be possible to pick a random object to replace when cache is full, system performance will be better if we choose an object that is not heavily used.

DOI : 10.5121/ijasuc.2012.3410

If a heavily used data item is removed it will probably have to be brought back quickly, resulting in extra overhead. So much work has been done on the subject of cache replacement.
Caching in wireless environment has unique constraints like scarce bandwidth, limited power supply, high mobility and limited cache space. Due to the space limitation, the mobile nodes can store only a subset of the frequently accessed data. The availability of the data in local cache can significantly improve the performance since it overcomes the constraints in wireless environment. A good replacement mechanism is needed to distinguish between the items to be kept in cache and that is to be removed when the cache is full. The extensive research on caching for wired networks can be adapted for the wireless environment with modifications to account for MT limitations and the dynamics of the wireless channel. These limitations include the MT's limited battery life and its small cache size.
This paper provides a general comparison of the cache replacement policies in wireless mobile networks based on the criteria used for evicting documents. We reviewed the various replacement policies for wireless networks with more focus on function based and location based policies. The different policies used in ad hoc networks are also reviewed. The topic of caching in ad hoc networks is rather new, and not much work has been done in this area. We classified the replacement policies for MANETs in to two groups local and coordinated. In coordinated replacement policy the mobile nodes which forms cooperative cache collectively takes the replacement decision. In the later case the data item to be evicted is determined independently by each node based on its local access information. Alternative techniques for cache replacement are also proposed.

2. CACHE REPLACEMENT POLICIES IN WIRELESS NETWORKS

Efficient replacement schemes for wireless mobile environments should consider different parameters like data access pattern, access costs, mobility pattern, connectivity, bandwidth, update rates, location dependence of the data. Most of the replacement algorithms form a value function by combining these parameters and evicts the data with minimum value. This section discusses some of the function based replacement policies in wireless environment.

Yin and Cao [1] proposed a generalized cache replacement policy for mobile environment. The value function they proposed can be used for different performance metrics and they considered minimum query delay and minimum download traffic as the target. The value function was based on parameters like probability of reference, cost of fetching data item, cost of validation, probability of invalidating cached data item and cost of getting updated data item to the cache. Based on these parameters the algorithm replaces a data item with min Value (i)/S_i, where S_i is the size of the data item. Here a strong consistency model is assumed. Xu and Lee [2] proposed a gain based replacement policy SAIU, for on demand broadcasts. The gain function for each data item is calculated as gain (i) = L_i. A_i/S_i. U_i where L_i is data retrieval delay, A_i is the access rate, S_i is the size of the data item and U_i is the update frequency.

Another algorithm proposed by Zeitunlian and Haraty [3] uses a least unified value cache replacement for SACCS, scalable asynchronous cache constituency scheme. Here the replacement is based on the reference information of the object, fetch cost and size. They considered the complete reference history for finding the probability of reference in the future. The book keeping involved in this method is too high. Chem. and Xiao [4] presented a cache replacement policy called on bound selection which used both data access and update information for replacement decision. The above mentioned schemes uses a function based policy. Since the relative importance of these parameters can vary from one type of request to another, some policies are needed to adjust the weights dynamically to achieve the best performance. Table 1 gives the summary of function based replacement policies.

2.1. Location based cache replacement policies

In Location Dependent information services (LDIS) the value of the data item depends on the location and varies as the user changes his location. The factors that are considered in a location aware replacement policy are the valid scope area, distance and direction of client movement. The area under which the data item is valid is the valid scope area. Distance is the distance between mobile node's current location and the valid scope area. When the data is distant from the valid scope area, it will have a lower chance to become useful. Direction indicates the direction of data movement from the valid scope area. The data that are moving in the opposite direction of the valid scope area will be irrelevant after sometime.

The cache replacement policy that supports location dependent services was early proposed by [5] (Manhattan) .Here the replacement was based on the Manhattan distance, which is the distance between the location of each cached data item's origin location and a mobile client's current location .The data items having the highest Manhattan distance are replaced. The only parameter considered for replacement is the distance.

The FAR (Farther Away Replacement) [6] replacement policy considers the current location and direction of the mobile client to make the replacement decision. The replacement strategy is based on the fact that the data which are not in the moving direction and farthest away from the user won't be visited in the near future. Based on the direction of movement, the data is arranged as two sets, In –Direction and Out-Direction. Whenever we want to replace data the Out Direction set is considered first, when it is empty the furthest segment in the In Direction set will be replaced. FAR considers only the spatial properties for cache replacement and the temporal properties are not taken.

In [7] two cache replacement policies PA and PAID are proposed. In this replacement policy a cost function is formed by considering the parameters access probability, valid scope area and data distance. Valid scope area refers to the geometric area of the valid scope of a data value. When this area is broad there is a higher chance that the client will request the data. In PA the cost function is formed as the product of access probability and valid scope. In PAID in addition to the above mentioned parameters data distance is also considered. The data with low access probability, a small valid scope area, and a long distance is evicted first.

K .Lai at el designed and implemented [8] mobility aware replacement scheme (MARS) which uses a cost function which consists of a clients location, movement of direction and access probability. The data item with lowest value for cost function is removed first. They also proposed an extension to this, The MARS+ tries to keep the clients movement patterns and from this history the future location of the client can be predicted. This is incorporated in to the replacement cost function and more accurate replacement decisions are made.

A network distance based cache replacement policy (ND – CRP) introduced by [9] considers the network distance which is the shortest path from current location of the mobile client (P) to a point of interest Pi for data eviction. Access probability and network density are the other factors considered in the replacement policy. This algorithm assumes that when the network density is high there is more chance to remain in that area for a long time. Dijkstra's algorithm is used to find the shortest path from the single source to single destination. The policy would choose the data with less access probability, less network density and greater network distance for eviction.

Prioritized Predicted Region based Replacement Policy (PPRRP) [10] tried to get the benefit of both temporal and spatial property in one unified scheme. In their scheme the distance is calculated based on a predicted region, where the client can be in the near future. In this policy

instead of taking the direction of client's movement they predict an area in which the client will be in the near future. The data item cost is calculated based on the access probability, valid scope area, data size in cache and distance of data based on the predicted region. Table 2 summarizes the various location based replacement policies.

Table 1. Summary of Function based Cache Replacement Policies

Algorithm	Parameters Considered	Eviction	Performance measure	Advantage	Disadvantage
Target Based	Reference Probability, cost of fetching data validation cost, probability of invalidating cached data item, cost of getting updated data.	Value is calculated using the parameters considered and replaces data with min value by size.	Average delay, Average downlink traffic	Can be used for multiple targets. Considered data updations.	Too many parameters to consider. How to select the target is not specified.
SAIU	Data retrieval delay, access probability, size,update frequency.	Low access rate, low delay and maximum sized data	Cache Hit Ratio, Strech	Uses a new performance metric	Parameters considered are not easily available
LUV - SACCS	Access frequency, recency, fetch cost, size	Smaller size, low access frequency, low cost	Cache hit ratio, Total Delay	Relates cache replacement with consistency	Book keeping is high, usage of a fixed parameter
On Bound Selection	Access frequency, update frequency	Low access frequency ,high update frequency	Cache Hit Ratio, Communication cost	Stale documents are evicted increases hit ratio	Not useful for short term access

Table.2. Summary of Location based Cache Replacement Policies

Algorithm	Parameters Considered	Eviction	Performance measure	Advantage	Disadvantage
Manhattan	Manhattan distance	Lowest Distance	Response time, Network traffic	Supports location dependent queries	Single parameter. Difficult to find estimated weights
FAR	Distance and movement direction of clients	Data in the out direction set is evicted first then the farthest in the indirection	Average Response time	Considers the direction of client motion and future movements	Not taken temporal properties. Ineffective when client changes its direction frequently.
PA	Access probability and valid scope area	Low access probability ,minimum valid scope area	Cache Hit Ratio	Considers temporal property	Objects close to the client are often replaced as their valid scope area is smaller
PAID	Distance between the current location and valid scope ,Access Probability, valid scope area	Low access probability ,minimum valid scope area ,maximum distance	Cache Hit ratio	Considers temporal and spatial property	Considers only the clients current movement direction
MARS	Client location, movement direction ,access probability ,update and query rate	Low temporal score and spatial score	Cache Hit ratio	Temporal and spatial properties are taken along with update frequency	Fails to recognize regular client movement patterns

3. CACHE REPLACEMENT POLICIES IN AD HOC NETWORKS

Data caching in MANET is mostly proposed as cooperative caching. In cooperative caching the local cache in each node is shared among the adjacent nodes and they form a large unified cache. So in a mobile cooperative caching environment, the mobile hosts can obtain data items not only from local cache but also from the cache of their neighboring nodes. This aims at maximizing the amount of data that can be served from the cache so that the server delays can reduced which in turn decreases the response time for the client. In many applications of MANET like automated highways and factories, smart homes and appliances, smart class rooms, mobile nodes share common interest. So sharing cache contents between mobile nodes offers significant benefits.

Cache replacement algorithm plays a central role in response time reduction by selecting suitable subset of data for caching. The available cache replacement mechanisms for ad hoc network can be categorized in to coordinated and uncoordinated depending on how replacement decision is made. In uncoordinated scheme the replacement decision is made by individual nodes. In order to cache the incoming data when the cache is full, replacement algorithm chooses the data items to be removed by making use of the local parameters in each node. Effective caching schemes in mobile environments should ideally consider proper cache admission control, consistency maintenance and replacement. Cache admission control decides whether the incoming data is cacheable or not. Substantial amount of cache space can be saved by proper admission control, which can be utilized to store more appropriate data, thereby reducing the number of evictions. If a node doesn't cache the data that adjacent nodes have it can cache more distinct data items which increase the data availability. Another feature of coordinated replacement is that the evicted data may be stored in neighboring nodes which have free space. In the following section we discuss various uncoordinated cache replacement policies for mobile ad hoc networks.

LRU

LRU (Least Recently Used) is based on the observation that data that have been heavily used recently will probably be heavily used again in the future. Conversely, data that have not been used for ages will probably remain unused for a long time. In LRU when cache is full the data item that has been unused for the longest time has been thrown out. It is a widely used algorithm in cache replacement. Logically, the cache consists of a list with most recently referenced data being in the front of the list. When a data is referenced it is moved from its existing position to the front of the list. When a new data comes in it is placed on the top of the list and the data at the back end is removed. LRU doesn't take in to account the non uniformity in the size of data, which is an important factor in mobile communication as the cost to fetch the data depends on size.

LRU Min

LRU Min [11] is a variant of LRU that tries to minimize the number of documents replaced. It is similar to LRU in implementation but will consider size of the data during replacement. In this scheme the data is arranged on the basis of access time and if a data item of size S needs to be cached it will search for items least recently accessed with size greater than S. If there is'nt any data in cache with size S, we start removing the items with size greater than S/2 and then objects of size S/4 until enough cache space is created. LRU Min policy will increase the hit ratio of smaller sized data items.

SXO

This is a local replacement policy[12] which considers the parameters data size and access frequency for replacement .Here larger sized data items are removed first as they occupy more cache space. More cache space can be made available by replacing bigger objects. The second parameter considered is order(di) which gives the frequency of access of data .Here replacement is done by combining the two parameters as value(di)=S*order(di).The advantage of this scheme is that the parameters used are easily available. But recently accessed data are not given any privilege.

LUV

A cache replacement policy based on least utility value (LUV) has been used in [13] For computing the LUV of a data item the access probability (Ai), size of the data item (Si), coherency which can be known by TTLi field and distance (δ) between the mobile client and data source were considered. Eq. for *utility*i function for a data item (di) is:

$utility_i = Ai \; . \; TT \; Li \; \delta i \; / Si$

3.1. Coordinated Cache replacement Policies

The **TDS**

The cache replacement [14] is based on two parameters distance (D) which is measured as the number of hops and access frequency. As the network is mobile the value of distance (d) may become obsolete .So the value is chosen based on the time at which it is last updated. The T value is obtained by the formula $1/t_{cur}- t_{update}$. Distance is updated by looking at the value of T. Based on how the distance and time is selected three different schemes are proposed TDS_D,TDS_T and TDS_N.TDS_D considers distance as the replacement criteria. If two data items have the same distance least value of (D+T) is replaced. In TDS_T the replacement decision is made by selecting the data with lowest T value. In the third scheme product of distance and access frequency is considered. In these algorithms TDS_D has the lower success rate and TDS_T has the higher hit ratio.

LUV Mi

This replacement scheme [15], has two parts replacement and migration. The replacement decision is based on a utility value formed by combining the parameters access probability, distance, Size and Coherency. In the migration part the replaced data is stored in the neighboring nodes which have sufficient space. For migration the data with highest utility value is given preference. Here even though the replacement decision is made locally migration is a coordinated operation. In order to save the cache space the data item is cached based on the location of the data source. If it is from the same cluster the data is not cached. The limitation of this scheme is that no checking is done whether the data is already present in the migrating node.

ECORP

Energy efficient Cooperative cache Replacement Problem (ECORP) [16] is an energy efficient cache replacement policy used in ad hoc networks. They considered the energy cost for each data access. For this, they considered the energy for in zone communication, energy for sending

the object, energy for receiving and energy cost for forwarding the object. Based on this they proposed a dynamic ECORP DP and ECOPR _greedy algorithms to replace data. The neighboring nodes will not cache the same data item in its local cache which reduces the redundancy and increases hit ratio.

Count Vector

In this scheme [17], each data item maintains a count which gives the number of nodes having the same data. Whenever the cache is full data item with maximum count is removed first as this will be available in the neighboring nodes. Whenever a data item is removed from the cache the access count will be decremented by one. Initially when the data is brought in to cache the count is set to zero.

4. DISCUSSION AND FUTURE WORK

Most of the replacement algorithms used in ad hoc networks are LRU based which uses the property of temporal locality. This is favorable for MANET which is formed for a short period of time with small memory capacity. Frequency based algorithms will be beneficial for long term accesses. It is better if the function based policies can adapt to different workload condition. In these schemes if we are using too many parameters for finding the value function, which are not easily available the performance can be degraded. Most of the replacement algorithms mentioned above uses cache hit ratio as the performance metric. In wireless network the cost to download data item from the server may vary. So in some cases this may not be the best performance metric. Schemes which improve cache hit ratio and reduce access latency should be devised. In cooperative caching coordinated cache replacement is more effective than local replacement since the replacement decision is made by considering the information available in the neighboring nodes. The area of cache replacement in cooperative caching has not received much attention. Lot of work needs to be done in this area to find better replacement policies.

Location dependent services are becoming popular in ad hoc networks. Replacement policies which consider location dependent parameters should be devised for cooperative caching in ad hoc networks. Another area of research in ad hoc networks is semantic caching in which the query is served from the cache based on the semantic description and results of previous queries. Cache admission control also plays role in improving the performance of cooperative cache. Value based admission control can be incorporated to minimize the number of replacements. Cache replacement based on Quality of Service (QOS) parameters can be explored. An alternative to cache replacement is that the data items that have their Time to Live (TTL) expired can be removed as the data becomes stale and cannot be used. So periodical checks can be done to delete the data items with TTL expired.

5. CONCLUSIONS

In this paper we made a general comparison of the major replacement policies in wireless networks and summarized the main points. Numerous replacement policies are proposed for wireless networks, but a few for cooperative caching in ad hoc networks. We also summarized the operation, strengths and drawbacks of these algorithms. Finally we provided some alternatives for cache replacement and identified topics for future research.

REFERENCES

[1].Yin, L., Cao, G., and Cai, Y. A Generalized Target-Driven Cache Replacement Policy for Mobile Environments. In Proceedings of SAINT. 2003, 14-21.

[2]. J.Xu,Q.L.Hu, L. Lee and W-C Lee, SAIU:" An efficient cache replacement policy for wireless on demand broadcasts, Proceedings of the 9th ACM International Conference on Information and Knowledge management (CKIM 200) pp 46-53,McLean, VA,USA, Nov 2000.

[3]. Aline Zeitunlian, Ramzi A. Haraty, "An Efficient Cache Replacement Strategy for the Hybrid Cache Consistency Approach", World Academy of Science, Engineering and Technology 63, 2010.

[4]. H. Chen and Y. Xiao, On-bound selection cache replacement policy for wireless data access. *IEEE Transactions on Computers*, 56 12 (2007), pp. 1597–1611.

[5].S.Dar,M. J. Franklin ,B.T.Jonsson,D.Srivatava and M. Tan Semantic Data Caching and Replacement,In Proceedingsofthe 22ndVLDBConference,pages 330-341,,India,1996.

[6] Q. Ren and M. Dhunham. Using semantic caching to manage location dependent data in mobile computing. Proc. Of ACM/IEEE MobiCom, 99:210–221, 2000.

[7] B. Zheng, J. Xu, and D. L. Lee. cache invalidation and re-placement strategies for location-dependent data in mobile environments. IEEE Trans. on Comp, 51(10):14–21, 2002.

[8]. K . Lai, Z .Tari, and P. Bertok . Mobility aware cache replacement for location dependent information services. In Technical Report T R- 04-04 (R MI T School of C S & I T), 2004.

[9] Mary Magdalene Jane.F, Yaser Nouh and R. Nadarajan, "Network Distance Based Cache Replacement Policy for Location-Dependent Data in Mobile Environment", Proceedings of the 2008 Ninth International Conference on Mobile Data Management Workshops ,IEEE Computer Society Washington,DC,USA,2008.

[10].Kumar, A., Sarje, A.K. and Misra, M. 'Prioritised Predicted Region based Cache Replacement Policy for location dependent data in mobile environment', Int. J. Ad Hoc and Ubiquitous Computing , Vol. 5, No. 1, (2010) pp.56–67.

[11]. Denko, M.K., Tian, J.,Cross-Layer Design for Cooperative Caching in Mobile Ad Hoc Networks, Proc .of IEEE Consumer Communications and Networking Conf(2008).

[12]. L. Yin, G. Cao: Supporting cooperative caching in ad hoc networks, IEEE Transactions on Mobile Computing, 5(1):77-89(2006).

[13]. Chand, N. Joshi R.C., and Misra, M., Efficient Cooperative Caching in Ad Hoc Networks Communication System Software and Middleware.(2006).

[14]S. Lim, W. C. Lee, G. Cao, C. R. Das: A novel caching scheme for internet based mobile ad hoc networks. Proc .12th Int. Conf. Computer Comm. Networks (ICCCN 2003), 38-43 (Oct. 2003).

[15]. Narottam Chand, R.C. Joshi and Manoj Misra, "Cooperative Caching Strategy in Mobile Ad Hoc Networks Based on Clusters," International Journal of Wireless Personal Communications special issue on Cooperation in Wireless Networks, Vol. 43, Issue 1, pp. 41-63, Oct 2007

[16]. Li, W., Chan, E., & Chen, D. (2007). Energy- efficient cache replacement policies for cooperative caching in mobile ad hoc network. In Proceedings of the IEEE WCNC (pp 3349–3354).

[17]. B. Z heng, J. Xu, and D. L ee. Cache invalidation and replacement strategies for location dependent data in mobile environments,. IEEE Transactions on Computers, 51(10) : 1141–1153, October 2002.

[18]. C. Aggarwal, J.L. Wolf, and P.S. Yu, "Caching on the World Wide Web," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 1, pp. 94-107, Jan./Feb. 1999