MOBILITY-AWARE WEB PREFETCHING OVER HETEROGENEOUS WIRELESS NETWORKS

Stephen Drew and Ben Liang

Department of Electrical & Computer Engineering, University of Toronto, Toronto, Canada
{drews, liang} @comm.utoronto.ca

Abstract - This paper presents a predictive framework for mobility-aware prefetching to enhance the experience of a mobile Web user roaming between heterogeneous wireless access networks. We consider a heterogeneous two-tier wireless access network system, composed of smaller, but faster and cheaper wireless local area networks (WLANs) placed within a much larger, but slower and more expensive cellular wireless data network. An optimal prefetch threshold algorithm is proposed, which takes into consideration the user mobility pattern, the relative characteristics of the networks, and the user perceived value of time. Using current industry network parameters, we study the performance of the proposed prefetching algorithm. Our numerical results show how user mobility and the heterogeneous network configuration significantly alter the prefetching threshold. The performance of this algorithm is compared with that of a static prefetching algorithm, demonstrating that mobility-aware prefetching can significantly improve the performance of future-generation heterogeneous wireless networks.

I. Introduction

In the next-generation wireless networking paradigm, wide-area cellular networks and wireless local area networks (WLANs) will co-exist to offer Internet access to end users. These technologies have characteristics that perfectly complement each other. Cellular networks built under the 3GPP or 3GPP2 standards offer global-scale coverage, providing data rate from 64 Kbps to 2 Mbps. They require careful planning and significant capital investment in terms of cell-site hardware and spectrum licensing, but enjoy a level of universal access unmatched by any other accessing technologies. WLANs operate within the free, unlicensed spectrum, typically support data rates from 10 Mbps to 100 Mbps, but are limited to a coverage range of a few hundred meters from the wireless access point. This makes them suitable for high-speed Internet access at home, office, and public hot-spots. Since no single system meets the ideal of high bandwidth, universal availability, and low cost, it is envisioned that these wireless access technologies will be strategically integrated in order to provide network services efficiently to end users[1]. Due to the different coverage sizes, levels of Quality-of-Service (QoS), and operation costs of these wireless access networks, a mobile user roaming within a heterogeneously integrated system should employ a dynamic strategy to adapt and to ensure that the optimal trade-offs are made.

In this work, we consider the scenario of wireless Web access by mobile users in such a system. Because of the limited bandwidth in a cellular wireless data network, Web users in this network may experience a large amount of delay. To alleviate this problem, we propose a predictive, mobility-aware Web page prefetching scheme for a mobile device to optimally reduce the time delay experienced by a mobile user in heterogeneous wireless networks.

Prefetching is a technique used to pro-actively increase system performance. Ideally, if a document is predicted to be accessed in the near future, the system should use this information and access the document at a strategic time. The item, once prefetched, is stored in a cache. The benefit of using prefetching is that it can reduce the total perceived time at the user by turning a long data-access time into a near-zero cache-access time. However, the cost of incorrectly predicting a document and its toll on the available resources must be taken into consideration. The trade-offs need to be carefully balanced for optimal system performance. For users of the Web, prefetching has been shown to be useful because web pages can be fetched in the downtime between user page requests. The cost of Web prefetching is consumption of bandwidth, the impact on total system performance, and the storing of unused pages.

The mobile Web user can also benefit from using prefetching, but there are added issues to consider with mobility. The power of transmission adds additional cost to each item and is examined in [2] and [3]. A mobile user may also move between networks with different quality of service. In [4], a data access scheme for hot-spot networks is proposed but employs user-initiated requests, instead of the automated approach proposed in this paper. Reference [5] presents a study on Web prefetching in different mobile environments, but it does not consider user movement between networks.

What we propose in this paper is a novel Web prefetching algorithm that operates seamlessly over a heterogeneous wireless system, where a mobile device can dynamically adjust its prefetching decision based on mobility pattern, network topology, and Web page accessing probability. Our algorithm uses the knowledge of mobility to predict the user's future location, capturing the likelihood of the user moving into different parts of the network. Since QoS can dramatically change as the user moves between networks, our algorithm uses the location information to dynamically adjust the amount of prefetching performed. The fundamental idea is that if the user is predicted to leave a high speed,

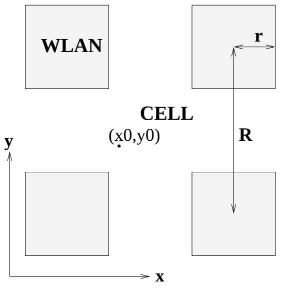

Fig. 1
A 2-Dimensional model of the network topology. The pattern continues infinitely in all directions.

low cost network in the near future, it should access more aggressively than it otherwise would.

The rest of this paper is organized as follows. An overview of our model and the assumptions used for our algorithm are outlined in Section II. Section III provides the details on how to compute the optimal mobile-aware prefetching threshold. In Section IV, we study how the prefetching performance gain can be affected by user mobility, network topology, and the costs associated with data accessing and time delay. The concluding remarks are given in Section V.

II. System Model

We model a heterogeneous wireless networking environment, composed of two types of networks: multiple WLANs of arbitrary size and separation, and a cellular wireless data network (e.g. CDMA2000 or GPRS) which covers the remaining area. The mobile user is free to physically roam anywhere within the system. Although the user may use either of the two networks, the mobile user prefers to use the WLAN given the choice since it is faster and less expensive. The locations of the networks are known by the mobile, perhaps provided by the service provider. As well, the current location and velocity of the mobile are assumed to be known by means of some mobility management scheme[6]. We use a simple two-dimensional model as shown in Figure 1 as our example topology. In our example topology, each side of a WLAN measures at $2r$, and the distance between the centers of WLAN is fixed at size R. We refer to r as the radius of each WLAN. The topology can therefore be represented by the ratio of r/R.

The mobile user accesses a sequence of Web pages over time as it roams the two-tier network. We use the technique of prefetching to access Web documents prior to the user requesting them. Prefetching helps reduce the total user time by retrieving and storing the highest access probability Web destinations while the user is reading the current page. Mobility-awareness helps predict the next location of the user. This information aids us in evaluating the cost of the access, and hence the cost of prefetching.

To reduce unnecessary waste of system resources, not all potential Web pages are cached. Instead, a prefetching threshold is used. The threshold represents the break-even cost point of the algorithm. If the probability of accessing a specific item is higher than the threshold, then using prefetching will be beneficial to the user. Items that have an access probability lower than the threshold should not be prefetched.

To build the mobility aware prefetching algorithm there are three steps : Derive the probability of being in each network for the next web access, compute the threshold level based upon the network probabilities and current network parameters, and determine the files, if any, that exceed the threshold level. All steps should be calculated at discrete time intervals, since the location and velocity of the user is assumed to be dynamic. The document access probabilities can be determined either from examining user history or from the server or from a shared proxy cache, but the exact details lay outside the scope of this paper. An example method is found in [7]. The sum of all page access probabilities should equal to one.

III. The Mobility-Aware Threshold Algorithm

In this section, we introduce the formulation for the first two steps as discussed above: deriving the network probabilities, and then looking at the costs to create our mobility-aware threshold.

A. Determining the Network Probabilities

To estimate the future location, and to incorporate the flexibility of different mobility patterns, a discrete time Gauss-Markov velocity model, as in [6], is used. Mobile movement is decomposed into any one direction, and the velocity at the current time step, v_n, is related to the velocity at the previous time step, v_{n-1}, by

$$v_n = e^{-\beta\Delta t} v_{n-1} + (1 - e^{-\beta\Delta t})\mu + \sigma\sqrt{1 - e^{-2\beta\Delta t}} w_{n-1} \ , \tag{1}$$

where μ is the mean velocity, σ is the root-variance, Δt is the discrete time interval, and w_{n-1} is a Normal random variable. The memory-parameter, β, controls to which degree the mobility pattern behaves. A near-zero value of β behaves like a constant velocity pattern, whereas a large value of β behaves like a random walk.

The velocity model is recursively expanded to yield a formula for the displacement s_k along an axis in k time intervals. The distribution is Gaussian with mean

$$E[\boldsymbol{s_k}] = \frac{1 - e^{-k\beta\Delta t}}{1 - e^{-\beta\Delta t}} v_0 + \mu(k - \frac{1 - e^{-k\beta\Delta t}}{1 - e^{-\beta\Delta t}}) \tag{2}$$

and variance

$$var[\boldsymbol{s_k}] = \sigma^2(1 - e^{-2\beta\Delta t}) var[\boldsymbol{w}] \tag{3}$$

where

$$var[\boldsymbol{w}] = E\Big\{\big(\sum_{i=1}^{k-1}\sum_{j=0}^{i-1} e^{-\beta\Delta t + i - j - 1}\boldsymbol{w_j}\big)^2\Big\} \ . \tag{4}$$

The computation of $var[\boldsymbol{w}]$ follows a similar approach as in [6].

A distribution for k, which is the number of time intervals between web page accesses, can be estimated from the history of web usage. The probability distribution of the location along any axis for the next web page access is then simply

$$P\{\boldsymbol{x} \leq x\} = \sum_{k=1}^{\infty} P[k] P\{\boldsymbol{s_k} \leq x \mid k\} \ . \tag{5}$$

Therefore, the initial location along the axis, the future location probability distribution, and the bounds of the network together determine the probability of the next web access lying within a specific network. For the two-dimensional model, we compute a similar model along both the x and y axes, breaking the initial velocity, velocity mean, and velocity variance into their corresponding axis components, $\{v_x, v_y\}$,$\{\mu_x, \mu_y\}$,$\{\sigma_x, \sigma_y\}$. The network probability of being in the WLAN, p_W, is then

$$p_W = \int_A \frac{d^2}{dx\,dy} P\{\boldsymbol{x} \leq x, \boldsymbol{y} \leq y\} \mathrm{d}A \ , \tag{6}$$

where A represents the total area coverage of all WLANs.

B. Pre-fetching Threshold

With a model for the network probabilities, we can add mobility-awareness to our prefetching model. We consider a single-user system where we need not account for the effects of system load. To compute the prefetching threshold, we need to assess the costs of prefetching and not prefetching, and minimize the average cost. This cost reflects the expected price paid for wireless data access and the penalty due to delay in data reception. A delay cost factor, α_T, is defined as in [7], which represents the cost per unit time as viewed by the user.

Then, the cost of not prefetching a document is

$$\begin{aligned} c_1 = p_a \cdot \Big\{ p_W \Big(\alpha_{B_W} s + \alpha_T \big(\frac{s + s_0}{b_W} \big) \Big) + \\ (1 - p_W) \Big(\alpha_{B_C} s + \alpha_T \big(\frac{s + s_0}{b_C} \big) \Big) \Big\} \ , \end{aligned} \tag{7}$$

and the cost of prefetching a document is simply

$$c_2 = \alpha_{B_{curr}} s \ , \tag{8}$$

where p_a is the access probability of the page, α_{B_W} is the cost per byte of the WLAN, α_{B_C} is the cost per byte of the cellular data network, b_W is the bandwidth capacity in bytes for the document using the WLAN network, b_C is the bandwidth capacity in bytes for the document using network 2, p_W is the probability the next access will be in the WLAN network, and $\alpha_{B_{curr}}$ is the cost per byte of the current network. The size of each document is s, and its setup size is s_0.

The cost for prefetching a document does not include any parameters for time because the document is assumed to be prefetched entirely before the user chooses to load the next page. Its cost is solely determined by the size of the document and the current network since it will be prefetched immediately. The average cost of not prefetching is broken into the cost of access in WLAN, and the cost of access in cellular, based on the probabilities of the user roaming into either network. It is again multiplied by the access probability of the document, but in the case the document is prefetched, it is accessed with probability 1.

The average total cost of all documents is minimized if and only if a page is prefetched when its access probability is greater than the threshold. The threshold value, H, is equal to p_a when we set c_1 equal to c_2 from above. This yields a threshold value of

$$H = \frac{\alpha_{B_{curr}}}{\alpha_T (\frac{p_W}{b_W} + \frac{1 - p_W}{b_C})(1 + \frac{s_0}{s}) + \alpha_{B_W} p_W + (1 - p_W) \alpha_{B_C}} \ . \tag{9}$$

By applying this formula, we now can use mobility-awareness to determine which items to prefetch.

IV. Numerical Analysis and Discussions

Using the threshold algorithm from the above section, we substitute numerical values to study the optimal prefetching threshold and the performance gain of our scheme.

A. Optimal Prefetching Threshold

For our model of WLAN and cellular networks, we use approximate industry values and unless otherwise indicated, the graphs will use $b_W = 100KB/s$, $b_C = 5KB/s$, $\alpha_{B_W} = \$0.5/MB$, and $\alpha_{B_C} = \$0.05/KB$. For α_T, the value, which is measured in \$/hr, will vary among users based on how valuable they perceive delay on web access to be. The affect of this variable will be examined more closely in the discussion about cost savings. The size of the documents is assumed to $s = 10KB$ with a setup size of $s_0 = 1KB$. A geometric distribution was used for modeling the time intervals between web page accesses, using an expected value of k=12 as found in [8].

We plot the threshold as it varies with initial position, using the origin as the center point of a WLAN. For the example 2-dimensional topology, a contour graph in Figure 2 illustrates how the threshold varies with position. For this figure, the cost parameters are set to $\alpha_{B_W} = \$1/MB$, $\alpha_{B_C} = \$0.02/KB$, and $\alpha_T = 100$. The velocity parameters used in this case assume a moderate memory velocity pattern $\beta = 0.1$, with a mean and initial velocity heading north-east. The prefetching threshold is significantly lower in the WLAN network. If the current position is in the WLAN coverage area, the threshold is lower if the mobile is predicted to leave into the cellular region. Conversely, if a mobile resides in the cellular network, then unless the next access network probability for WLAN is very low, prefetching should for

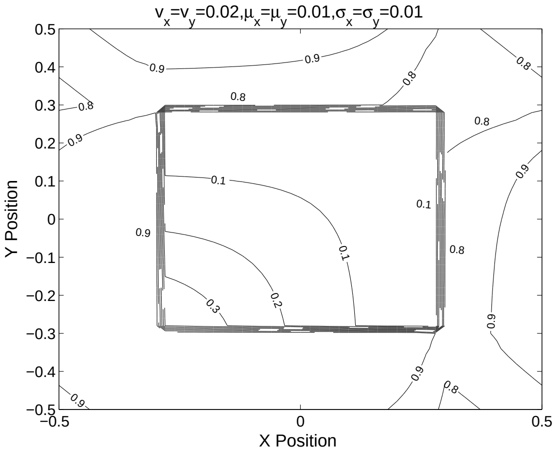

Fig. 2
Threshold contours for the topology by x-y position.

the most part not be used. These results demonstrate the rationale behind mobility-aware, predictive prefetching. Next we examine the performance of our algorithm.

B. *Cost Savings from Mobility-Aware Prefetching*

In this subsection we look at the average performance gain that is achieved by our mobility-aware prefetching scheme. To measure the gain from our algorithm, we have computed the average cost over a range of different document probabilities. We assume 10 possible documents to choose from, together having access probabilities from a truncated geometric distribution with parameter q. Twenty uniformly distributed values of q between 0 and 1 are used to represent a sample set of different types of document probabilities. The two extreme values thus are the case where all 10 documents have equal probabilities, up to the case where one document has 100% access probability.

For each document, the cost formulas (7) and (8) determine the cost of a document based on it being prefetched or not. We total the individual costs for all documents in the above set of document probabilities to yield an average cost per position, and then over all positions in the network to yield a system-wide average cost. This value is compared to the accessing scenario where no prefetching is used, where the cost of each document is calculated using (7).

The values being used for α_B in the previous models represent a good approximation of current industry practice, but are subject to change. We have plotted the average performance gain over varying values of $\alpha_{B_{WLAN}}$ while keeping $\alpha_{B_{CELL}}$ constant as before, and the results are displayed in Figure 3. The performance gain decreases as the cost per byte of WLAN approaches that of the cellular network. As the costs of the network approach each other, the threshold values do the same, thus decreasing the amount of gain available from location prediction.

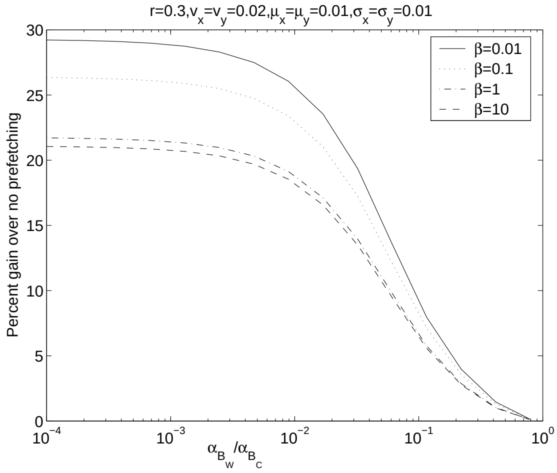

Fig. 3
Performance gain by varying costs per byte. Plotted for multiple values of velocity memory parameter β

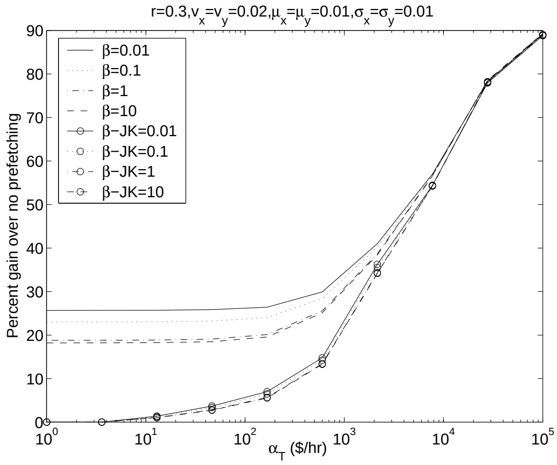

Fig. 4
Performance gain by varying α_T. Plotted against a mobility-unaware prefetching scheme from [7].

Figure 4 shows the average performance gain for the mobility-aware prefetching and compares it against the performance gain for the mobility-unaware prefetching as presented in [7], while varying the user-assessed value of time delay. At low α_T, our system achieves gains of up to 35%. As α_T increases, the thresholds in both systems decrease, causing the number of documents being prefetched to increase, resulting in a greater performance gain. Our system achieves higher gains than the mobility-unaware scheme, with the least difference occurring at extremely large, and unrealistic values of α_T.

We now change the initial and mean velocities to zero to represent a statistically stationary user. Figure 5 plots the percent gain for varying degree of velocity memory, and compares the location-aware and location-unaware prefetching models. At low values of β, adding network prediction results in little performance gain because the mobile is predicted to stay put with very high probability. This conforms with the observation in [6] that the predictability of a mobile's movement is not necessarily a monotonic function

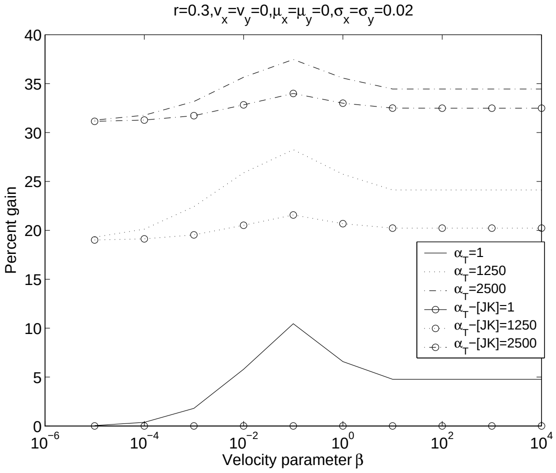

Fig. 5

Performance gain over varying velocity memory parameter β. Uses zero mean, zero velocity.

of β. The highest gains occur for a mid range value of β, at around 0.1. This is in contrast to the prior non-zero mean scenarios where the greatest gains occur for the smallest values of β.

Finally, we have varied the radius r for our example topology, resulting in different ratios of r/R, displayed in Figure 6. This figure shows how the performance of our prefetching algorithm changes as the WLAN coverage rate changes. For lower α_T, a peak exists because most of the gain in our system comes from exploiting the fact that potential accesses in the cellular network result in more cost savings when accessed in the WLAN. When no cellular system is available, the performance of our algorithm will be the same as that of a mobility-unaware algorithm. At high values of α_T, much of the gain comes from lowering delay opposed to gain from network prediction, and thus no peak. Overall, location prediction can achieve considerable performance gain over a wide range of WLAN coverage rates.

V. Conclusions

Introducing mobility awareness and prediction to the mobile Web environment gives users a means to more fully optimize their time and operation costs. We have presented a novel threshold algorithm for Web page prefetching in heterogeneous wireless networks, one that dynamically computes the threshold value based on the user's current location and future predicted location. Our numerical results show how the prefetching decision can be altered by user mobility and the heterogeneous network configuration. The proposed prefetching algorithm provides significant performance gains over that of no prefetching, using current industry parameters. It also results in higher performance gain than mobility-unaware, non-predictive prefetching schemes such as that used in [7].

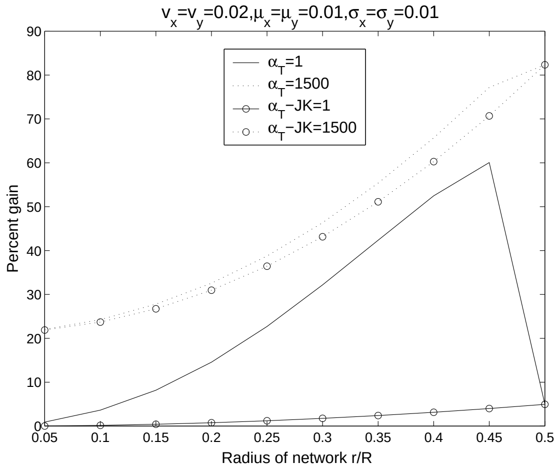

Fig. 6

Performance gain by varying radius r. Plotted against the mobility-unaware algorithm for various values α_T.

The model of square WLANs surrounded by a cellular area is only an example topology used for this paper. A similar approach can be applied to any topology and can be extended to three dimensions. Additional network models of varying QoS may also be added to the analysis for cases where more than two networks are available. Furthermore, the proposed mobility-aware framework of predictive prefetching not only can be used for Web applications, but also can be modified to suit any other type of application that obeys similar access patterns.

References

[1] R. Berezdivin, R. Breinig, and R. Topp, "Next-generation Wireless Communications Concepts and Technologies," *IEEE Communications Magazine*, March 2002.

[2] S. Gitzenis and N. Bambos, "Power-Controlled Data Prefetching/Caching in Wireless Packet Networks," *INFOCOM 2002.*, pp. 1405–1414, June 2002.

[3] L. Yin and G. Cao, "Adaptive Power-Aware Prefetch in Wireless Networks," *IEEE Transactions on Wireless Communication*, 2004 (to appear).

[4] N. Iami, H. Morikawa, and T. Aoyama, "Prefetching Architecture for Hot-Spotted Networks," *IEEE International Conference on Communictaions*, vol. 7, 2001.

[5] Z. Jiang and L. Kleinrock, "Web Prefetching in a Mobile Environment," *IEEE Personal Communications*, vol. 5, pp. 25–34, Oct. 1998.

[6] B. Liang and Z. J. Haas, "Predictive Distance-Based Mobility Management for Multidimensional PCS Networks," *IEEE/ACM Transactions on Networking*, vol. 11, Oct 2003.

[7] Z. Jiang and L. Kleinrock, "An Adaptive Network Prefetching Scheme," *IEEE JSAC*, vol. 16, 1998.

[8] E. Casilari, A. Reyes-Lecuona, F. Gonzalez, A. Daz-Estrella, and F. Sandoval, "Characterisation of Web Traffic," *GLOBECOM '01*, vol. 3, Nov 2001.