Mobility Tracking and Traffic Characterization for Efficient Wireless Internet Access

Shun-Zheng Yu, Brian L. Mark* and Hisashi Kobayashi
*Dept. of Electrical Eng., Princeton University, Princeton, NJ 08544; *Elect. and Comp. Eng. Dept., George Mason University, Fairfax, VA 22030*

Abstract: In a wireless communications network, the movement of mobile users presents a scenario of access to the Internet that is substantially different from the wired network. Requests for content issued by a mobile user depend on its mobile state (e.g., location, velocity and direction). We employ a semi-Markov process representation to construct a model that characterizes mobile user behavior in a general state-space. The states of a mobile user can then be estimated and tracked by using an algorithm for parameter estimation of a general Hidden Semi-Markov Model (HSMM). Dynamic behavior of the aggregate request rate can also be characterized. Finally, we show how the tracking model and the request model can be applied to prefetch Web content for each mobile user for efficient wireless Internet access.

Key words: Wireless Networks, Wireless Internet, Mobility, Traffic Modeling, Hidden Markov Model

1. INTRODUCTION

In a wireless communications network, the movement of mobile users presents a scenario of access to the Internet that is substantially different from the wired network. For an individual mobile user, the point of contact to the wired network changes with time. It is therefore imperative to be able to track dynamic mobile behavior and to take into account the request traffic when providing content to the mobile users.

The construction of mobility patterns for analysis and simulation has attracted considerable attention in recent years [1]-[3]. In [1], a cellular-based

location tracking system is developed that utilizes the estimated distance between the mobile and the referenced base station. In [3] mobile behavior is modeled as a random walk or Brownian motion on two-dimensional or three-dimensional grids. A stochastic model for mobility called Markovian highway Poisson arrival location model is proposed in [4].

In [5], a new mobility tracking model is introduced that characterizes mobile user behavior in a general state-space using a semi-Markov process representation. This model differs from the earlier work in that it allows us to exploit recent results in queuing and loss network theory [6] and to characterize the macroscopic mobility and traffic behavior in the wireless network. The mobility tracking can be implemented in real-time using a computationally efficient parameter estimation algorithm that has been proposed recently [7]. The mobility model is augmented in [8] by introducing a new request model that characterizes mobile user behavior of access to the Internet. Based on information extracted from this model, a resource allocation and an admission control scheme for wireless networks is proposed in [9].

In the present paper, we propose a predictive prefetching scheme for each individual mobile user, based on the estimation of the mobile user state. Our objective is to reduce the average latency that a mobile user experiences in accessing the wireless Internet. The remainder of the paper is organized as follows. Sections 2 and 3 present the mobility model and mobility tracking model, respectively. Section 4 discusses the characterization of web content traffic generated from mobile user requests to access the wireless Internet. Section 5 discusses the application of the mobility and request model to prefetching of Web content at proxy servers. Simulation results illustrating this approach to efficient wireless Internet access are presented in Section 6. Finally, Section 7 concludes the paper.

2. MOBILITY MODEL

Following [5][8][9], we define the state of a mobile user in terms of a vector $(x_1, \ldots, x_n)$, where the i th component, x_i, represents a value from a finite *attribute* space A_i. The attribute spaces represent properties of the mobile user such as location, moving direction, speed, etc. The set of possible states for a mobile user is an n -dimensional vector space given by

$$S = A_1 \times \cdots \times A_n, \tag{1}$$

where $\times$ denotes the Cartesian product. The abstract space S can be made as rich as desired by including the appropriate attributes as components in the state vector. The dynamic motion of a user, as defined by its time-varying attribute values, can then be described by its trajectory in this space.

We enumerate all possible states in S and label them as 1, …, M such that the state space S can more simply be represented as follows:

$$S = \{ 1, \ldots, M \}. \tag{2}$$

We introduce two *inactive* states in addition to the set of *active* states S: the *source* state 0 and the *destination* state d. A user enters the system by assuming the state 0. A user exits the system by assuming the state d. Thus, the user can assume states in the augmented state-space $S' = S \cup \{0, d\}$.

No transitions occur from states $j \in S$ to the source state, i.e., $a_{j0} = 0$. From any such state j, the user next assumes the destination state d with probability a_{jd}. No transitions are allowed from the destination state. Hence, the state d is considered to be the *absorbing* state of the Markov chain. Further, no transitions occur from state 0 to state d, i.e., $a_{0d} = 0$. The state transitions of a user are characterized by a Markov chain with transition probability matrix:

$$\boldsymbol{A}' = \begin{array}{c} \\ d \\ 0 \\ 1 \\ : \\ : \\ M \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 0 & 0 & 0 & \ldots & 0 \end{array} \\ \begin{bmatrix} 0 & 0 & a_{01} & a_{02} & \ldots & a_{0,M} \\ a_{1d} & 0 & a_{11} & a_{12} & \ldots & a_{1,M} \\ a_{2d} & 0 & a_{21} & a_{22} & \ldots & a_{2,M} \\ : & : & : & : & : & : \\ a_{M,d} & 0 & a_{M,1} & a_{M,2} & \ldots & a_{M,M} \end{bmatrix} \end{array}. \tag{3}$$

In practical applications transitions among the states are limited due to physical constraints (e.g., the street layout). We assume that from a given state, transitions can occur to on the order of ten neighboring states, such that the transition probability matrix is highly sparse.

We assume the dwell time of a user in state $m \in S$ to be generally distributed with mean d_m. Hence, the state process of a user is a semi-Markov chain. The transition probability matrix and the state duration distributions can be estimated by means of a parameter estimation algorithm discussed in [7][10].

The aggregate behavior of the system of mobile users can be represented by the vector process

$$\boldsymbol{N}(t) = (N_1(t), ..., N_M(t)), \tag{4}$$

where $N_m(t)$ represents the number of mobile users in state m at time t. We observe that the above system is equivalent to an open queuing network with M infinite-server stations corresponding to the states in S. Clearly, the source

and destination stations of the queuing network correspond to 0 and d, respectively. Results from the theory of queuing and loss networks [6] show that the steady-state distribution of $\boldsymbol{N}(t)$ is insensitive to the distributions of the dwell times at each station.

From

$$e_m = a_{0m} + \sum_{n \in S} e_n a_{nm}, \qquad m \in S, \tag{5}$$

we get the value e_m, which can be interpreted as the average number of visits that a user makes to state m during its sojourn in the system starting from the source state 0 until reaching the destination state d. Let N_m denote the expected number of users in state m in equilibrium (m=1, ..., M). The mean departure rate from state m is given by

$$\lambda_m = N_m / d_m = \lambda_0 e_m, \qquad m=1, 2, \ldots, M, \tag{6}$$

where λ_0 is the total rate at which mobile users transit from the inactive state 0 to an active state, i.e., the total rate of entry to the system, and d_m is the mean dwell time in state m.

3. MOBILITY TRACKING MODEL

The general mobility model was discussed in the context of a continuous-time parameter t. In practice, tracking of the system parameters must be based on measured observations sampled at discrete time instances. Therefore, we shall represent the user dynamics by a discrete-time semi-Markov chain, where the parameter t is now discrete, taking values in {0, 1, 2, ... }. Furthermore, the system states cannot, in general, be observed directly, i.e., the states are *hidden*. Hence, an appropriate model for the system is a discrete-time *Hidden Semi-Markov Model* (HSMM).

As in the continuous-time model, the evolution of the user state in the active state-space S is characterized by a state transition probability matrix denoted by $\boldsymbol{A} = [a_{ij}: i,j \in S]$. We shall assume that the mobile user dwell time in a given state is a random variable taking values in the set {1, ..., D}, with probability distribution function denoted by $p_m(d)$, d=1,..., D. We introduce the $M \times D$ matrix

$$\boldsymbol{P} = [p_m(d): m \in S, d = 1, ..., D]. \tag{7}$$

In order to track user mobility, the parameters of the semi-Markov model must be estimated based on observations of the user state. This leads to a Hidden Semi-Markov Model (HSMM) described as follows. Let $s_t \in$ {1, ...,

M} denote the state of the user at time t, t = 0, 1, 2, ... Let us denote the initial state probability distribution vector by

$$\boldsymbol{\pi} = (a_{0m} : m=1,..., M) , \quad (8)$$

where a_{0m} is the probability that the initial state of the user is state m.

Let o_t denote the value of an observation of the user state at time t. We assume that there are K distinct state observation values, 1, ..., K. Note that the observation value o_t is generally different from the true state s_t, due to geolocation and estimation errors. We define the following observation probability distribution matrix:

$$\boldsymbol{B} = [b_m(k): m \in S, k=1, ..., K], \quad (9)$$

where $b_m(k)$ denotes the probability that the observed value at an arbitrary time t is $o_t = k$, given that the actual user state is $s_t = m$. The 4-tuple ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{P}$, $\boldsymbol{\pi}$) provides a complete specification of the discrete-time Hidden Semi-Markov Model for the system.

To track the state of a mobile user, we apply the forward-backward and re-estimation algorithms for HSMM parameter estimation discussed in [7][10]. The main steps of the tracking algorithm are summarized as follows:

1. Apply the *HSMM re-estimation algorithm* to obtain initial estimates $(\hat{\boldsymbol{A}}_0 , \hat{\boldsymbol{B}}_0 , \hat{\boldsymbol{P}}_0 , \hat{\boldsymbol{\pi}}_0)$ of the HSMM model parameters by using training data.
2. Apply the *HSMM forward-backward estimation algorithm* to estimate the state s_t of the mobile user at time t, based on the geolocation observation sequences o_1^t.
3. Obtain refined estimates, $(\hat{\boldsymbol{A}}_i , \hat{\boldsymbol{B}}_i , \hat{\boldsymbol{P}}_i , \hat{\boldsymbol{\pi}}_i)$, by applying the HSMM re-estimation algorithm to the given observation sequences.

Estimation of the mobility model parameters must in general be made based on missing data. Due to physical constraints, geolocation measurement and/or transmission of geolocation data may not take place frequently enough to allow precise tracking of the user's state at all times. We consider four different cases [11]:

1. *Deterministic observation pattern:* The geolocation observations are generated periodically but some mobile states may be missing if the observations are not made frequently enough.
2. *Random observation pattern:* The geolocation observation are generated at random times. Again, some mobile states may be missing due to insufficient observation frequency.
3. *State-dependent missing observation*: In some states, there may be a finite probability that a null output is generated. For example, in a certain state,

a mobile user may not request any Web content. In this case, the system is not able to log any requests from the user.

4. *Output-dependent missing observation.* Even when a non-null output is generated by a given state, the corresponding observation could still be missing, e.g., if the signal received by a base station is too weak or is corrupted noise, or if the state duration is too short.

A detailed development of the main elements of the HSMM parameter estimation algorithm and its validation by simulation are reported in [7][11]. The algorithm has a computational complexity proportional to D, where D is the maximum value of the dwell time for all states. The more general forward-backward algorithm reduces to the Baum-Welch algorithm when D=1. We note that the algorithm offers a significant improvement over an earlier algorithm by Ferguson (1980) [10] which has computational complexity proportional to D^2.

We define one of the *forward variables* [10] [7] as follows:

$$\alpha_t^*(m) = \Pr[o_1^t \text{ ; state } m \text{ begins at } t+1] / \Pr[o_1^t] , \quad (10)$$

where o_1^t is the sequence of observations from time 1 to time t, and $\alpha_t^*(m)$ is the probability that a mobile user is entering its next state m at time t+1 for given observations o_1^t. The forward variables are then computed inductively for t = 1, 2, …, T [7][10]. Similarly, the backward variables can be defined and computed inductively for $t = T, T-1, \ldots, 1$. After computing the forward and backward variables, the maximum a posterior (MAP) state estimate can be found.

A simple iterative procedure for re-estimating the HSMM parameters is reported in [7]. By applying the well-known EM (Expectation / Maximization) algorithm, it can be shown that this iterative procedure is increasing in likelihood. The overall computational complexity of the re-estimation algorithm is essentially proportional to T. Thus, the parameters for the HSMM model can be estimated efficiently within the framework of dynamic mobility model tracking.

4. CHARACTERIZATION OF TRAFFIC

We can augment the above mobility model by introducing state-dependent information. Let {0, 1, ..., J} represent a set of user requirements for web content, where content type $j=0$ specially represents no requirement, as shown in *Figure 1*. We suppose that a mobile user entering state m requires web content of type j from the network with probability $c_m(j)$, with:

$$\sum_{j=0}^{J} c_m(j) = 1, \qquad m=1, \ldots, M. \tag{11}$$

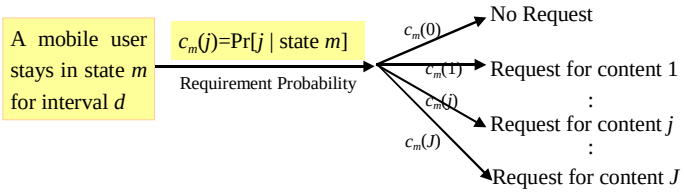

Figure 1. Request model.

The mobile users' requests are logged in a wireless Internet Web server, forming an observation sequence that can be obtained independently from the geolocation observation sequence. As mentioned in Section 3, based on the geolocation observations, the distribution $\{b_m(k)\}$ defined in (9) can be determined by the model parameter re-estimation algorithms for HSMM. Similarly, the probability distribution $\{c_m(j)\}$ can be treated as a model parameter and be determined using the parameter re-estimation algorithms based on the observations of requests.

Traffic characterization is a necessary step in determining the amount of system resource that should be allocated for each user in order to meet their quality-of-service (QoS) requirements. The wireless Internet Web servers should allocate sufficient computational resources to process user requests. The network should also allocate sufficient bandwidth and buffer resources to provide QoS for transmissions from the mobile user. Using the mean departure rate λ_m given in (6), the average request rate for content j can be determined by [8]:

$$R_j = \sum_{m=1}^{M} \lambda_m c_m(j) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} N_m, \qquad j=1, \ldots, J. \tag{12}$$

The instantaneous request rate for content j can be defined by

$$R_j(t) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} N_m(t), \qquad j=1, \ldots, J. \tag{13}$$

If the dwell time distributions of the user states are assumed exponential, then $\boldsymbol{N}(t)$ is a Markov process. Hence, the request rate process $R_j(t)$ defined here can be viewed as a Markov modulated rate process (MMRP) as studied

in [12]. If we allow the dwell times to have general distributions, $R_j(t)$ becomes what we may term as a semi-Markov modulated rate process. Let $\boldsymbol{X}(t)$ be an M-dimensional diffusion process that approximates the M-dimensional semi-Markov process $\boldsymbol{N}(t)$. Under a set of reasonable assumptions [12], $\boldsymbol{X}(t)$ can be expressed as an M-dimensional Ornstein-Uhlenbeck (O-U) process. Hence, the process $R_j(t)$ can be approximated by a Gaussian process

$$\tilde{R}_j(t) = \sum_{m=1}^{M} \frac{c_m(j)}{d_m} X_m(t), \qquad j=1, \ldots, J. \tag{14}$$

5. WEB PREFETCHING

Proxy Web servers have been introduced to the Internet in order to prefetch or cache frequently requested web content, thus improving the web access speed perceived by the end user [13]. Fast access to the Internet is especially important in the wireless environment, where the bandwidth and other system resources are expensive commodities. Under conventional prefetching schemes, the hit ratio is typically less than 50%, even when the storage capacity of the proxy server is relatively large [13]. This implies that more than half of the web content requested by a typical user must be obtained directly from the origin servers. Consequently, under conventional prefetching schemes, users may still experience relatively large average latencies and highly variable delays in accessing web content. In the wireless network, this results in a considerable waste of the wireless resources.

In [5], a static prefetch scheme for wireless Internet services based on the statistical data collected from user requests and server responses is proposed. In the following, we apply the integrated mobility/traffic tracking model to develop a predictive prefetch scheme for each mobile user based on the estimation of the user's mobility and the web access probabilities. Our objective is to improve the access latency performance over conventional prefetching schemes.

The information obtained from the mobility tracking and the request model is used to estimate the access probability that a mobile user requests a Web document. By using a forward-backward algorithm, we can obtain the probabilities that the mobile user enters its next state, $\{\alpha_t^*(m) : m \in S\}$, given in (10). Therefore, the access probability that the mobile user requests content j at time t is given by

$$\gamma_j(t) = \sum_{m \in S} \alpha_t^*(m) c_m(j). \tag{15}$$

Conventional prefetching schemes are based on the access probability of a web document. Therefore, this probability can be used straightforwardly to design a prefetch scheme. We use the prefetch criterion proposed in [14] to reduce the average access latency. Define

$$\eta_j(t) = \gamma_j(t)(1-h_j)\Delta T_j. \quad (16)$$

where h_j is the average hit ratio for the requests for content j, and ΔT_j is the average response delay for content j imposed by the Internet:

$$\Delta T_j = \mathrm{E}\{response\_time_j - request\_time_j\}. \quad (17)$$

where $request\_time_j$ is the time when the proxy server sends out the request for content j, and $response\_time_j$ is the time when the proxy server receives the response for content j. If the proxy server cache can store up to r documents, then the r documents of highest value $\eta_j(t)$ are prefetched [14].

6. SIMULATION RESULTS

We consider an example scenario of a serving area (about 1 km by 1 km) consisting of 128 street segments in a rectangular mesh layout. Each street segment is about 100 meters long. We assume that for each street segment, there are two walking states (in two directions), two driving states (in two directions) and one shopping state. There are a total of 640 active states plus one inactive source state and one absorbing state. Each active state has about ten neighbor states. Transitions can occur from the inactive source state to any active state and from any active state to the absorbing state. The mean dwell time for a walking state is about 3 minutes, while that for a driving state is 16 seconds and that for a shopping state is 12 minutes. There are 50 mobile users involved in the wireless Internet services. There are 20 categories associated with each street segment and each category has 20 distinct web contents. Therefore, there are a total of 51,200 contents for the serving area. We assume that the average response time delay imposed by the Internet is ΔT = 500 ms.

We denote l_m as the real location (i.e., street segment) of a mobile user when it is in state m. Then we assume that the access probability in state m for contents associated with location l is inversely proportional to the square of the distance between l and l_m, i.e.,

$$c_m(l) \propto 1/(l-l_m)^2, \qquad \text{for } l \neq l_m. \quad (18)$$

Specifically, we let

$$c_m(l_m)=0.9 \text{ and } \sum_{l \neq l_m} c_m(l) = 1 - c_m(l_m) = 0.1 \quad (19)$$

when m corresponds to a shopping state. In other words, when the mobile user is in a shopping state m, it requests, with probability 0.9, the contents that are associated with the location l_m. In a similar way, we assign $c_m(l_m)=0.8$ and $c_m(l_m)=0.6$ when m corresponds, respectively, to the walking and driving states. When the user is in a given state m, it accesses the content from the associated categories according to a uniform distribution. We also assume that the distribution of content access probabilities for a given state, location and category follows a Zipf's law-like distribution [15][16], where the probability of requests for the ith most popular content is proportional to $1/i^{\alpha}$, with $\alpha=1$.

The simulation results for the prefetch scheme are shown in *Figure 2*. From this figure, we see that if the proxy server prefetches five documents for each mobile user, the average latency can be reduced by about 54 ms for each access request to the Internet. If the proxy server prefetches 100 documents for each mobile user, the average latency can be reduced by about 156 ms. Note that prefetching a selected document means that whenever there is no fresh copy of the document in the cache, the proxy server fetches the document from the origin server.

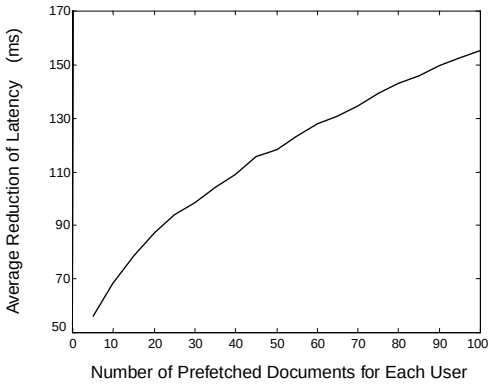

Figure 2. Reduction in average latency using the proposed prefetch scheme.

7. CONCLUSION

In this paper, we constructed a model to characterize mobile user behavior in a general state-space using a semi-Markov process

representation. We discussed how to build a user request model to characterize traffic patterns generated by web document requests. Based on the mobility tracking and request model, we proposed a prefetch scheme for each individual mobile user to reduce the average access latency incurred when a mobile user accesses wireless Internet Web content. Besides an improvement in the perceived QoS of the user, the reduction in access latency implies a significant savings in wireless resources. The simulation results for a representative scenario showed reductions ranging from 10% to 30%, depending on the number of documents prefetched for each user.

REFERENCES

[1] P. C. Chen, "A cellular based mobile location tracking system," in *Proc. IEEE VTC'99*, pp. 1979–1983, 1999.

[2] M. Hellebrandt and R. Mathar, "Location tracking of mobiles in cellular radio networks," *IEEE Trans. on Vehicular Tech.*, 48(5):1558–1562, Sept. 1999.

[3] S. Tekinay, "Modeling and analysis of cellular networks with highly mobile heterogeneous sources," Ph.D. dissertation, School of Information Technology and Engineering, George Mason University, 1994.

[4] K. K. Leung, W. A. Massey, and W. Whitt, "Traffic models for wireless communication networks," *IEEE J. Select. Areas in Comm.*, 12(8):1353–1364, Oct. 1994.

[5] S.-Z. Yu and H. Kobayashi, "A prefetch cache scheme for location dependent services," *submitted for publication.*

[6] H. Kobayashi and B. L. Mark, "Product-Form Loss Networks," in J. H. Dshalalow, editor, *Frontiers in Queueing: Models and Applications in Science and Engineering*, CRC Press, pp. 147–195, 1997.

[7] S.-Z. Yu and H. Kobayashi, "A Forward-Backward Algorithm for Hidden Semi-Markov Model and its Implementation," *submitted for publication.*

[8] H. Kobayashi and S-Z Yu, "Performance Models of Web Caching and Prefetching for Wireless Internet Access," in *Int. Conf. on Performance Evaluation: Theory, Techniques and Applications (PerETTA 2000)*, University of Aizu, Fukushima, Japan, Sept. 2000.

[9] H. Kobayashi, S-Z. Yu and B.L. Mark, "An Integrated Mobility and Traffic Model for Resource Allocation in Wireless Networks," in *Proc. 3rd ACM Int. Workshop on Wireless Mobile Multimedia (WoWMoM-2000)*, August 2000.

[10] J. D. Ferguson, "Variable duration models for speech," *Symp. on the Application of Hidden Markov Models to Text and Speech*, pp. 143–179, Oct. 1980.

[11] S.-Z. Yu and H. Kobayashi, "Extensions to Hidden Semi-Markov Model with Missing Observations," *submitted for publication*.

[12] Q. Ren and H. Kobayashi, "Diffusion process approximations of a statistical multiplexer with Markov modulated bursty traffic sources," *IEEE J. Select. Areas in Commun.*, 16(5):679–691, 1998.

[13] D. Wessels and K. Claffy, "ICP and the Squid Web cache," *IEEE J. Select. Areas in Commun.*, vol. 16, pp. 345–357, April 1998.

[14] S-Z. Yu and H. Kobayashi, "A New Prefetch Cache Scheme," in *Proc. IEEE Globecom 2000*, San Francisco, CA, Nov. 2000.

[15] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," in *Proc. IEEE INFOCOM'99*, pp.126–134,1999.

[16] G. Voelker *et al*, "On the Scale and Performance of Cooperative Web Proxy Caching," *Proc. 17th SOSP*, pp. 16–31, Kiawah Island, SC, Dec. 1999.