hand, many applications will need higher accuracy.

Although predicting the time of the occurrence of next movement besides the next location is important for some applications, in this paper we focus on predicting the next location, since most domain-independent predictors do not predict time.

We found that the simple Markov predictors performed as well or better than the more complicated LZ predictors, with smaller data structures. We also found that many predictors often fail to make any prediction, but our simple fallback technique provides a prediction and improves overall accuracy. We conclude the paper with a more extensive summary of our conclusions.

II. Background

In this section, we discuss the nature of location prediction, and summarize several predictors from the literature. We define the metrics we used to evaluate the predictors, and how we collected the empirical data set we used in our evaluation.

A. Location

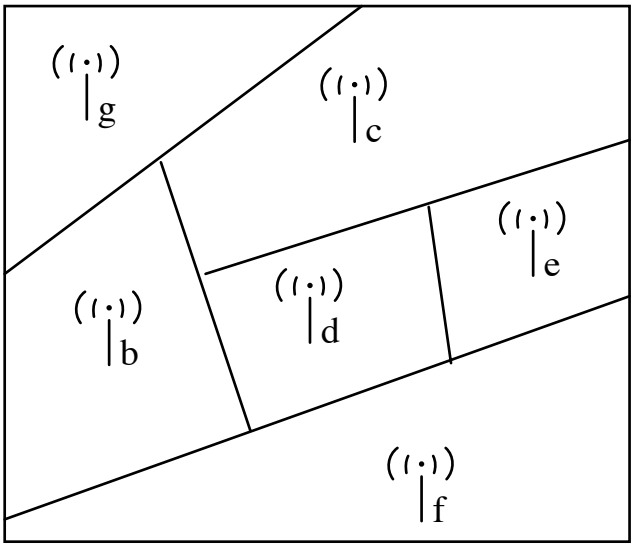
In the context of this work we assume that a user resides at a given discrete location at any given time; sample locations include "room 116" within a building, "Thayer Dining Hall" within a campus, "cell 24" within a cellular network, or "berry1-ap" access point within an 802.11 wireless network. (In our data, as we discuss below, each location is the name of an access point with which the user's device associated.) We list all possible such locations in a finite alphabet $\mathcal{A}$, and can identify any location as a symbol a drawn from that alphabet. For a given user we list the sequence of locations visited, its *location history* L, as a string of symbols. If the history has n locations, $L_n = a_1 a_2 \ldots a_n$.

The location history may be a sequence of location *observations*, for example, the user's location recorded once every five seconds, or a sequence of location *changes*. In the latter case, $a_i \neq a_{i+1}$ for all $0 < i < n$. All of the predictors we consider in this paper are agnostic to this issue. It happens that our data is a sequence of location changes.

All of the predictors we consider in this paper are domain-independent and operate on the string L as a sequence of abstract symbols. They do not place any interpretation on the symbols. For that reason, our location history does not include any timing information, or require any associated information relating the symbols such as geographic coordinates. As an example, though, consider the environment with six wireless cells (labeled b through g) diagrammed in Figure 1, accompanied by one possible location history.

B. Domain-independent predictors

In this paper we consider only domain-independent predictors; we will examine domain-dependent predictors in future work. We are interested in *on-line predictors*, which examine the history so far, extract the current context, and predict the next location. Once the next location is known, the history is now one symbol longer, and the predictor updates its internal tables in preparation for the next prediction.



Sample location history $L = gbdcbgcefbdbde$

Figure 1. Sample cell map and location history

During the rest of this section, we discuss two families of domain-independent predictors, Order-k Markov predictors and LZ-based predictors.

Markov family

The order-k (or "$O(k)$") Markov predictor assumes that the location can be predicted from the current *context*, that is, the sequence of the k most recent symbols in the location history $(a_{n-k+1}, \ldots, a_n)$. The underlying Markov model represents each state as a context, and transitions represent the possible locations that follow that context.

Consider a user whose location history is $L = a_1 a_2 \ldots a_n$. Let substring $L(i,j) = a_i a_{i+1} \ldots a_j$ for any $1 \leq i \leq j \leq n$. We think of the user's location as a random variable X. Let $X(i,j)$ be a string $X_i X_{i+1} \ldots X_j$ representing the sequence of random variates $X_i, X_{i+1}, \ldots X_j$ for any $1 \leq i \leq j \leq n$. Define the context $c = L(n-k+1, n)$. Let $\mathcal{A}$ be the set of all possible locations. The Markov assumption is that X behaves as follows, for all $a \in \mathcal{A}$ and $i \in \{1, 2, \ldots, n\}$.

$$\begin{aligned} &P(X_{n+1} = a | X(1,n) = L) \\ &\quad = P(X_{n+1} = a | X(n-k+1, n) = c) \\ &\quad = P(X_{i+k+1} = a | X(i+1, i+k) = c) \end{aligned}$$

where the notation $P(X_i = a_i | \ldots)$ denotes the probability that X_i takes the value a_i. The first two lines indicate the assumption that the probability depends only on the context of the k most recent locations. The latter two lines indicate the assumption of a stationary distribution, that is, that the probability is the same anywhere the context is the same.

We experimented with a simple alternative to the frequency-based approach to Markov predictors, using recency (probability 1 for most recent, 0 for all other) to define the transition matrix. Although this recency approach was best among $O(1)$ Markov predictors, it was worst among $O(2)$ Markov predictors, and we are still investigating the underlying reason.

We found most of the literature defining these predictors to be remarkably insufficient at defining the predictors for implementation. In particular, none defined how the predictor should behave in the case of a tie, that is, when there was more than one location with the same most-likely probability. We investigated a variety of tie-breaking schemes within the Markov predictors, but found that the accuracy distribution was not sensitive to the choice.

Since some of our user traces extend over weeks or months, it is possible that the user's mobility patterns do change over time. All of our predictors assume the probability distributions are stable. We briefly experimented with extensions to the Markov predictors that "age" the probability tables so that more recent movements have more weight in computing the probability, but the accuracy distributions did not seem significantly affected. We need to study this issue further.

We examined the original LZ predictor as well as two extensions, prefix and PPM. LZ with both extensions is known as LeZi. We found that both extensions did improve the LZ predictor's accuracy, but that the simple addition of fallback to LZ did just as well, was much simpler, and had a much smaller data structure. To be fair, PPM tries to do more than we require, to predict the future path (not just the next move).

We stress that all of our conclusions are based on our observations of the predictors operating on over 2000 users, and in particular whether a given predictor's accuracy distribution seems better than another predictor's accuracy distribution. For an individual user the outcome may be quite different than in our conclusion. We plan to study the characteristics of individual users that lead some to be best served by one predictor and some to be best served by another predictor.

There was a large gap between the predictor's accuracy distribution and the "optimal" accuracy bound, indicating that there is substantial room for improvement in location predictors. On the other hand, our optimal bound may be overly optimistic for realistic predictors, since it assumes that a predictor will predict accurately whenever the device is at a location it has visited before. We suspect that domain-specific predictors will be necessary to come anywhere close to this bound.

Overall, the best predictors had an accuracy of about 65–72% for the median user. On the other hand, the accuracy varied widely around that median. Some applications may work well with such performance, but many applications will need more accurate predictors; we encourage further research into better predictors.

We continue to collect syslog data, extending and expanding our collection of user traces. We plan to evaluate domain-specific predictors, develop new predictors, and develop new accuracy metrics that better suit the way applications use location predictors. We also plan to predict the location along with the time in the future.

ACKNOWLEDGEMENTS

The Dartmouth authors thank DoCoMo USA Labs, for their funding of this research, and Cisco Systems, for their funding of the data-collection effort. We thank the staff of Computer Science and Computing Services at Dartmouth College for their assistance in collecting the data.

REFERENCES

[1] David A. Levine, Ian F. Akyildiz, and Mahmoud Naghshineh, "The shadow cluster concept for resource allocation and call admission in ATM-based wireless networks," in *Mobile Computing and Networking*, 1995, pp. 142–150.

[2] William Su and Mario Gerla, "Bandwidth allocation strategies for wireless ATM networks using predictive reservation," in *Proceedings of Global Telecommunications Conference (IEEE Globecom)*, November 1998, vol. 4, pp. 2245–2250.

[3] George Liu and Gerald Maguire Jr., "A class of mobile motion prediction algorithms for wireless mobile computing and communications," *ACM/Baltzer Mobile Networks and Applications (MONET)*, vol. 1, no. 2, pp. 113–121, 1996.

[4] Sajal K. Das and Sanjoy K. Sen, "Adaptive location prediction strategies based on a hierarchical network model in a cellular mobile environment," *The Computer Journal*, vol. 42, no. 6, pp. 473–486, 1999.

[5] Amiya Bhattacharya and Sajal K. Das, "LeZi-Update: An information-theoretic approach to track mobile users in PCS networks," *ACM/Kluwer Wireless Networks*, vol. 8, no. 2-3, pp. 121–135, March–May 2002.

[6] Fei Yu and Victor C. M. Leung, "Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks," *Computer Networks*, vol. 38, no. 5, pp. 577–589, 2002.

[7] Christine Cheng, Ravi Jain, and Eric van den Berg, "Location prediction algorithms for mobile wireless systems," in *Handbook of Wireless Internet*, M. Illyas and B. Furht, Eds. CRC Press, 2003.

[8] Sajal K. Das, Diane J. Cook, Amiya Bhattacharya, Edwin Heierman, and Tze-Yun Lin, "The role of prediction algorithms in the MavHome smart home architecture," *IEEE Wireless Communications*, vol. 9, no. 6, pp. 77–84, 2002.

[9] Jeffrey Scott Vitter and P. Krishnan, "Optimal prefetching via data compression," *Journal of the ACM*, vol. 43, no. 5, pp. 771–793, 1996.

[10] Jacob Ziv and Abraham Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, Sept. 1978.

[11] P. Krishnan and Jeffrey Scott Vitter, "Optimal prediction for prefetching in the worst case," *SIAM Journal on Computing*, vol. 27, no. 6, pp. 1617–1636, 1998.

[12] Meir Feder, Neri Merhav, and Michael Gutman, "Universal prediction of individual sequences," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1258–1270, July 1992.

[13] Timothy C. Bell, John G. Cleary, and Ian H. Witten, *Text Compression*, Prentice Hall, 1990.

[14] David Kotz and Kobby Essien, "Analysis of a campus-wide wireless network," in *Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking*, September 2002, pp. 107–118, Revised and corrected as Dartmouth CS Technical Report TR2002-432.

[15] David Kotz and Kobby Essien, "Analysis of a campus-wide wireless network," *ACM Mobile Networks and Applications (MONET)*, 2003, Accepted for publication.