Birds of a Feather Surf Together: Using Clustering Methods to Improve Navigation Prediction from Internet Log Files

Martin Halvey, Mark T. Keane, and Barry Smyth

Adaptive Information Cluster, Smart Media Institute, Department of Computer Science,
University College Dublin, Belfield, Dublin 4, Ireland
{martin.halvey, mark.keane, barry.smyth}@ucd.ie

Abstract. Many systems attempt to forecast user navigation in the Internet through the use of past behavior, preferences and environmental factors. Most of these models overlook the possibility that users may have many diverse sets of preferences. For example, the same person may search for information in different ways at night (when they are pursuing their hobbies and interests) as opposed to during the day (when they are at work). Thus, most users may well have different sets of preferences at different times of the day and behave differently in accordance with those preferences. In this paper, we present clustering methods for creating time dependent models to predict user navigation patterns; these methods allow us to segment log files into appropriate groups of navigation behaviour. The benefits of these methods over more established methods are highlighted. An empirical analysis is carried out on a sample of usage logs for Wireless Application Protocol (WAP) browsing as empirical support for the technique.

1 Introduction

One of the key challenges in adaptive hypermedia and personalization is to properly capture user preferences based on their past behavior, explicit and implicit preferences and other environmental factors. When these and other factors are known it becomes more possible to predict what a user is looking for, and for a system to automatically adapt to the user using such predictions. In this paper, we examine navigation in the context of the mobile-Internet with a view to predicting user preferences for certain sites based on environmental factors (i.e. time of the week and paths followed). Cotter and Smyth [6] have proposed that users should not have just one set of preferences but rather groups of preferences, characterizing their browsing *personality* in different contexts. The fundamental insight behind the present work is that a user navigating during office hours would have different preferences than that same user on the weekend. As such, the research challenge is to determine the boundaries between people's distinct personas and to be able to use these categories to predict preferences. In this paper, we outline methods for clustering together user sessions based on the paths followed in those sessions, where a path is a series of URLs selected within a session.

We advance these methods as a means for automatically determining different sets of user preferences, thus providing an important component for any system that wishes to be truly adaptive.

2 Task Context: Navigation on the mobile-Internet

WAP is the basis for the wireless Internet. It is an open global specification that provides Internet browsing functionality for small hand-held devices such as mobile phones to easily access and interact with information and services instantly. WAP users face additional problems that PC users do not, the screen real estate on a mobile phone is several orders of magnitude smaller than that of PCs. The mobile phones' capabilities are much more diverse than the more standardized PCs (e.g., in display resolution, color capability, operating system features, browser functionality). Mobile phones also have very limited input capabilities, featuring numeric keypads with minimal text entry, unlike the mouse and keyboard options available to PC users. Finally, the content-base of WAP is considerably less diverse and when this content *is* accessed, users face slow download times and incremental billing costs [16]. While some of the problems, for example slow download times and phone capabilities have been addressed, there is still no definitive solution to aid user surfing via WAP. Here we provide one possible solution.

Intuitively, users' navigation on the weekend when they have leisure days should differ from their navigation on weekdays when they are at work. Following this intuition, our hypothesis is that users surf differently during different time periods, and that these differences can be used to make predictions about user navigation. The main idea behind using clustering, to attempt to determine this split, is to group together sets of paths that users have followed that contain similar types of pages (for example user sessions involving sports sessions should be grouped together, in the hope that these types of pages will display some distinct patterns with relation to time for example and allow the segmentation of the log files to improve navigation prediction). Halvey et al [9] have previously shown that predictive models that are time dependent can greatly improve accuracy in navigation prediction. However the segmentation step in their time-based models were handcrafted rather than automated. Here we propose a more formal method of grouping together user sessions and segmenting log files. The purpose of this work is to aid users' navigation on the Internet. By predicting user navigation with a high degree of accuracy, you can for example aggressively promote a URL so that the URL that a user would like to select is always at the top of a series of menu selections or the desired URL is highlighted in order to help a user find it. To test this hypothesis we analyzed a data set from a mobile Internet portal. By exploiting the uneven distributions in WAP surfing pattern we endeavored to determine whether distinct navigation patterns arose and, if found, whether these patterns could be accurately predicted using Markov models.

3 Background Research

3.1 Clustering

Clustering algorithms have been used in a broad range of applications. In particular we are interested in hierarchical agglomerative clustering algorithms as they are the type of clustering algorithm that we are using in this work. In the area of image segmentation an agglomerative clustering algorithm was applied in Silverman and Cooper [17] to the problem of unsupervised learning of clusters of coefficient vectors for two image models that correspond to image segments. Jain and Dubes have applied the complete-link hierarchical clustering scheme to the problem of object and character recognition [11]. They have also applied the complete link agglomerative clustering algorithm in the area of document retrieval to create a proximity dendogram for a collection of books [11]. Etzioni has applied hierarchical clustering methods in the area of data mining [7]. For a more complete review of clustering methods and their applications please refer to Jain et al [12].

3.2 Time-Based Analysis

In recent times, there has been an increasing interest in the use of time in conjunction with predictive models. Analyses of time-based patterns of work in office environments are an example. Begole et al [2] attempt to detect and model rhythms of work patterns in an office. Horvitz et al [10] use Bayesian networks built over log data to model time-based regularities in work patterns in order to predict meeting attendance and interruptability.

Time-based analyses of web searching have also been carried out. With the aim of supporting users, Lau and Horvitz [13] have constructed probabilistic models centering on temporal patterns of query refinement to predict how a user would continue their search. Beitzel et al [3] have analyzed search engine queries with respect to time and found that some topical categories vary substantially more in popularity throughout the day; they also found that query sets for different categories have differing similarity over time.

Halvey et al [9] have also recently conducted an analysis of mobile Internet browsing patterns with respect to time. They discovered that users browsing patterns had temporal variations and exploited these patterns to improve navigation prediction accuracy. However much of this work was done by hand and required many steps.

3.3 Web Navigation

Lieberman [14], Cooley et al [5] and Spilioulou [19] have all also presented solutions that take advantage of earlier user experiences to create adaptive Internet websites. Of particular interest to this current work is that other researchers have used Markov models to create predictive models of web navigation. Pirolli [15] has shown that k-means Markov models can be used to forecast user navigation patterns. Zhu et al [21] have used Markov chains, based on past navigation patterns, to predict web page accesses. As was discussed in the previous section navigation prediction in the mobile Internet presents additional problems, Billsus et al [4], Anderson et al [1] and Smyth and Cotter [18] all offer solutions to the problem of navigation in the mobile Internet.

4 Clustering Sessions based on Paths

The aim of this work is to segment web logs in such a way that making predictions about user navigation becomes simpler and more accurate. The WAP portal that is used can be represented as a tree. It is hoped that the clustering will result in paths from similar sections of the tree being grouped together to reveal information about users who have an interest in that section of the portal. Also paths of similar length may be grouped together, and may also reveal that users who favour longer sessions may have distinct features from users who favour shorter sessions.

4.1 Distance Metrics

To begin with each URL that was selected by a user was represented by a symbol, and a path was then represented by a sequence of these symbols. The first task was to calculate a distance between these paths that were traversed. The first distance metric that we used was a simple Euclidean distance $(\sqrt{(\sum_{i=1}^{N} (p_i - q_i)^2)})$ where if $p_i = q_i$ then a value zero is returned otherwise one is returned. So for example for two strings where the first string is “Mark” and the second string is “Martin” the Euclidean distance is $\sqrt{(0+0+0+1+1+1)} = 1.732051$. The second method that was used was Levenshtein Distance or Edit Distance, which is for two strings, s1 and s2, the minimum number of point mutations required to change s1 into s2, where a point mutation is one of either change a character, delete a character or insert a character. For example “Mark” and “Martin” have a Levenshtein Distance of 3, however “Martin” and “Barry” have a Levenshtein Distance of 4. The third and final method that we use is a derivation of Euclidean Distance that for the purposes of this work has been called *Total Euclidean*. As the WAP site in which the navigation took place is a tree we performed a depth first search on the tree and assigned each node incrementally an integer value. The Euclidean distance equation is then applied. As stated previously, if the two nodes are the same then the distance between them is zero. However if two nodes are different then the difference between them is the difference between the integer values assigned

to their nodes. In this way paths in branches of the tree closer together will have a distance that is shorter than those in branches of the three that are further apart.

4.2 Clustering Paths

To cluster the paths we used hierarchical agglomerative clustering methods, single link, complete link and average link algorithms were implemented. A fixed number of clusters were not set for these methods; instead these algorithms were given a stopping parameter. The parameter chosen was that when d1, the average distance between clusters, is less than half of d2, the average distance between elements in the clusters, then the algorithm should stop i.e. when d1 > (d2)/2 stop. Initially the parameter was when d1 > d2, however due to some outlying nodes the majority of the clustering methods did not halt until all of the elements were members of one large cluster.

4.3 Finding Time Related Segments

As stated previously Halvey et al [9] have also recently conducted an analysis of mobile Internet browsing patterns and discovered that users browsing patterns had temporal variations. Accordingly each of the clusters formed was analysed to see if there is a distinct or dramatic rise or fall in the number of hits that the clusters have received with respect to time (either days or hours). If such an anomaly occurs sessions for that time period are extracted from the log files and a distinct predictive model is created for that time period.

5 Predicting User Navigation Using Log Data

5.1 Predicting User Navigation

To test whether these approaches could be used successfully to automatically segment web logs (and later be used to make predictions about user navigation and ultimately aid that navigation) we analyzed a data set from a mobile Internet portal. This data, gathered over four weeks in September 2002, involved 1,168 users and 147,700 individual user sessions using WAP phones to access a major European mobile Internet portal. Using the distance metrics and clustering algorithms outlined earlier the paths followed by users in the WAP portal were clustered. However not all nine possible combinations of distance metrics and clustering methods formed clusters, the Euclidean and Levenshtein distance methods formed clusters using both the average and complete link algorithms, the *Total Euclidean* distance method formed clusters using the complete link algorithm. The log files were then segmented according to the clusters of paths followed.

To evaluate the success of different clustering methods we constructed Markov models, similar to Zhu [21], for each of the segmented data sets, as well as models for all of the data. Each of these models was then used to predict the next URL clicked in a path given the current path for each of the segmented data subsets and for values of k between 1 and 5. Five was chosen as an upper limit as some of the segments contained only a small number of sessions and also some of the segments contained sessions where only short paths were followed. In each WAP menu the user has approximately seven selections (including going back to the previous page) from which they can choose, therefore the result of random recommendations should be approximately one in six a baseline 0.167%. However for these experiments a fully connected graph was assumed, to take into account instances where users used a bookmark or typed in a URL in the mid-session for example. As this theoretically gives users a one in 256 choice, there is a baseline of approximately 0.004% accuracy. The models created were then tested on a sample of random sessions from the log files to calculate the accuracy of the models, for these experiments we consider accuracy to be the proportion of URL's that are correctly predicted by the Markov models. The results of these experiments are shown in figure 1. The accuracy for the models created using the data from the segmented log files is contrasted with the accuracy for the predictive model built using all of the log file data. Overall, one major conclusion can be drawn from these models about our ability to predict navigation behavior and the success of our model for segmenting the log files. That is, if one tries to predict navigation during a particular set of sessions using the full data set the predictive accuracy of the model is not as accurate as the model that corresponds to that set of sessions.

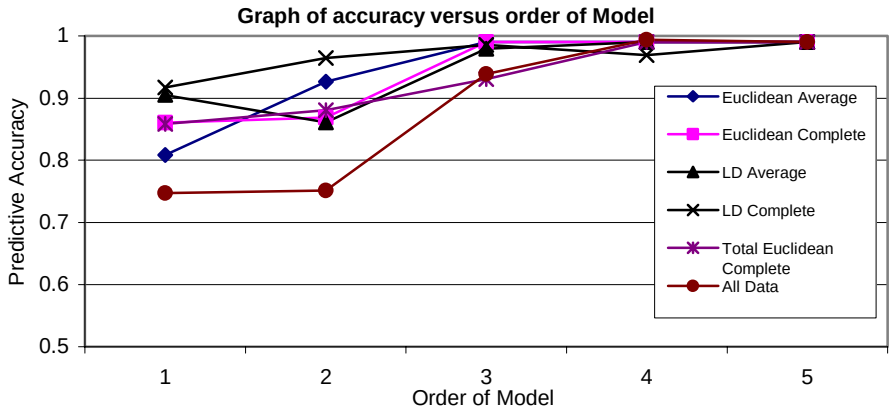

Fig 1. Graph illustrating the predictive accuracy for each of the segmented data sets with respect to k

As outlined in section 4.2 the clusters formed were then analyzed to see if there were any time dependencies in the clusters. Three of the five sets of clusters had the same

time dependencies, the other two clusters found slight variations of the first time relationship. Once again the log files were segmented, however on this occasion the segmentations were based on the time dependencies discovered. Also as was done previously Markov Models were formed, however for these segments the maximum value of k was 10 as there were no significantly small segments and all of the segments contained sessions with various path lengths. As before the models created were then tested on a sample of random sessions from the log files to calculate the accuracy of the models, and once again for these experiments we consider accuracy to be the proportion of URL's that are correctly predicted by the Markov models. The results of these experiments are shown in figure 2. The same conclusions can be drawn from the results in figure 2 as were concluded in figure 1. However, it may be noted that after a certain order of model the accuracy begins to tail off. However, this is not really a concern as most WAP users favour shorter sessions over longer sessions according to the Universal Law of Web Surfing [8].

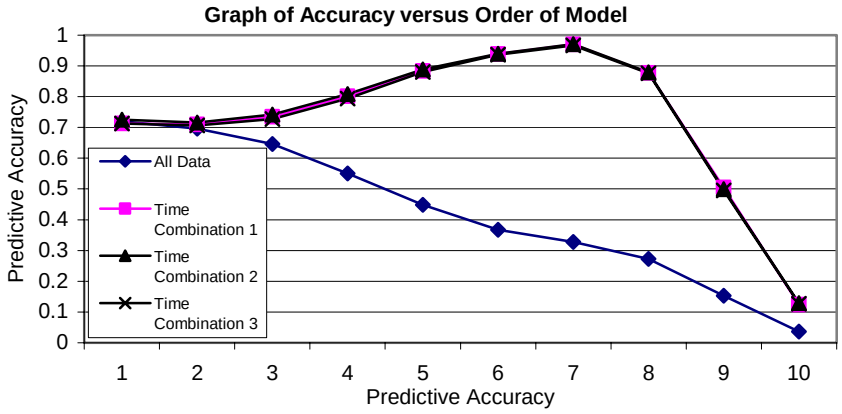

Fig 2. Graph illustrating the predictive accuracy for each of the segmented data sets with respect to k.

5.2 Empirical Evaluation of Predictive Models

We can also put the knowledge we have gained from our Markov modeling to work in assessing how it would impact personalization of a mobile portal. Smyth & Cotter [18] have developed the Click-Distance Model to estimate the likely navigation gains in a given mobile-portal when personalization is used for menu re-ordering. Figure 3 illustrates the results of the click distance analysis for the cluster based models and Figure 4 illustrates the results of the click distance analysis for the time-based models. These results can be summarised in a few points. The use of any navigation data reduces mean click-distance significantly in comparison with when it is not used; therefore personalisation using navigation patterns helps. Also the effectiveness of these models improves with the order of the Markov models. Finally, in nearly all of the

models, the models based on the clustered data results resulted in shorter click distances than the models based on the whole data set.

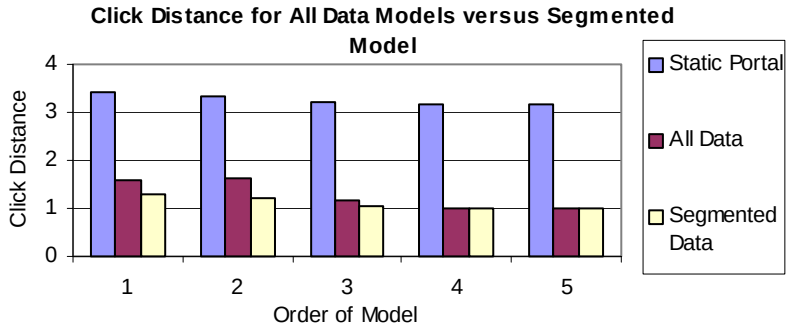

Fig 3. Results of click distance analysis for static portal, Markov models using all data and Markov models created using clustered data.

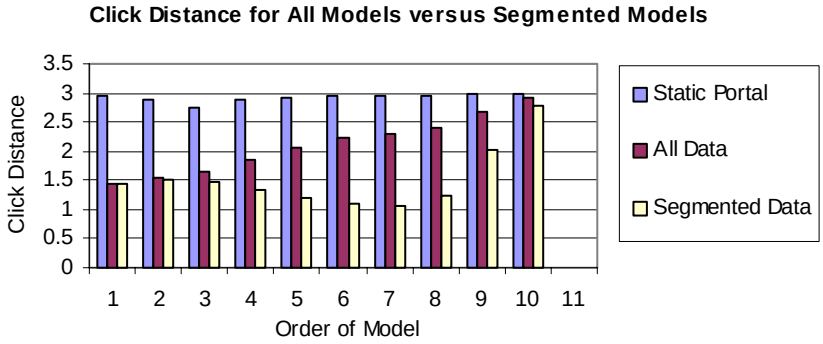

Fig 4. Results of click distance analysis for static portal, Markov models using all data and Markov models created using data segmented based on time.

6 Conclusions and Future Work

Users have different needs and goals at different times and as such user navigation patterns in the Internet are time dependent. In this paper we have presented a method that takes advantage of this phenomenon by automatically segmenting log file data, based different time periods and different goals. This has been confirmed by an analysis of a WAP usage log file. Clusters of usage patterns were created, and from these clusters Markov models were learnt. These predictive models were compared with Markov models built over all of the log data. The predictive accuracy for the Markov models for the explicit time periods and clusters was far greater than the accuracy of

the other models constructed. These more accurate models can also be used reorganize the structure of the portal to reduce the click distance and thus reduce the amount of effort for users. These results support our hypothesis as well as highlighting the potential of such data segmentation for aiding user navigation creating truly adaptive systems. Consequently, there is a huge potential benefit to Internet users for usage of such techniques, in particular mobile Internet and WAP users who encounter additional problems that desktop users do not encounter [16], which we have highlighted previously.

Additionally predicting user navigation could be used to improve download times. While a server is not busy, predicted pages could be pre-fetched for users to reduce download times, this would be particularly useful for mobile users for whom download times are a particular problem [16]. Also predicting and pre-fetching pages could also reduce the load on servers. As this is an initial attempt at segmenting log data according to time there are, of course, other extensions that can be made to this work. Firstly this segmentation of the data could quite easily be used in conjunction with some other predictive model, for example the ClixSmart navigator [18] to make more accurate predictions about user navigation and adapting portal structure to the needs of users. With the integration of some of these techniques we may be able to discover other temporal segmentations and make even more accurate recommendations.

In this paper we have outlined new methods to aid users of both the mobile-Internet and Internet. This study is a new direction in Internet navigation prediction and will hopefully lead the way in finding the solution to what is a very difficult problem.

7 Acknowledgements

This material is based on works supported by the Science Foundation Ireland under Grant No. 03/IN.3/I361 to the second and third authors.

References

1. Anderson, C.R., Domingos, P. & Weld, D.S., Adaptive Web Navigation for Wireless Devices, Proceedings of Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001.
2. Begole J., Tang J.C. & Hill B., "Rhythm Modeling, Visualizations and Applications", Proceedings of the 2003 Symposium on User Interface Software and Technology (UIST 2003), pp. 11-20, 2003.
3. Beitzel S., Jensen E., Chowdhury A., Grossman D. & Frieder O. HoURLy analysis of a very large topically categorized web query log. Proceedings of the 27th annual international conference on Research and development in information retrieval, pp 321 – 328, 2004.

4. Billsus D., Brunk C., Evans C., Gladish B. & Pazzani M. Adaptive Interfaces for Ubiquitous Web Access, Communications of the ACM, Vol 45, No 5, 2002.
5. Cooley R., Mobasher B. & Srivatava J. Web Mining : Information and Pattern Discovery on the World Wide Web, Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA, November 1997.
6. Cotter P., & Smyth B. "PTV: Intelligent Personalised TV Guides". Proceedings of the 12th Innovative Applications of Artificial Intelligence (IAAI-2000) Conference. AAAI Press, 2000.
7. Etzioni, O. The World-Wide Web: quagmire or gold mine? Communications of the. ACM 39, 11, 65–68
8. Halvey M, Keane M.T. & Smyth B. "Mobile Web Surfing is the same as Web Surfing", Communications of the ACM, 2005. (Accepted; In Press).
9. Halvey M., Keane M.T., & Smyth B., "Predicting Navigation Patterns on the Mobile-Internet using Time of the Week", World Wide Web 2005, (Accepted; In Press)
10. Horvitz E., Koch P., Kadie C.M., & Jacobs A. Coordinate: Probabilistic Forecasting of Presence and Availability In: Proceedings of the Eighteenth Conference on Uncertainty and Artificial Intelligence, Edmonton, Alberta. Morgan Kaufmann Publishers, pp. 224-233, 2002.
11. Jain A. K., & Dubes R. C. Algorithms for Clustering Data. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.
12. Jain A., Murty M.N., & Flynn P. Data clustering: A review. ACM Computing Surveys, 31(3):264–323, September 1999.
13. Lau T. & Horvitz E. Patterns of Search: Analyzing and Modeling Web Query Refinement. Proceedings of the Seventh International Conference on User Modeling, 1999.
14. Lieberman H. Letizia: An Agent That Assists Web Browsing, International Joint Conference on Artificial Intelligence, Montreal, August 1995.
15. Pirolli P. "Distributions of Surfers' Paths through the World Wide Web: Empirical Characterizations." The Web Journal, 2 : 29-45, 1998.
16. Ramsay M., Nielsen J. Nielsen Report, "WAP Usability Deja Vu: 1994 All Over Again", 2000.
17. Silverman J. F., & Cooper D. B. Bayesian clustering for unsupervised estimation of surface and texture models. IEEE Trans. Pattern Anal. Mach. Intell. 10, 482–495, 1998.
18. Smyth B. & Cotter P. "The Plight of the Navigator: Solving the Navigation Problem for Wireless Portals". Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Systems. Malaga, Spain, pp 328-337, 2002.
19. Spiliopoulou M. The laborious way from data mining to Web log mining. International Journal of Computer Systems Science and Engineering, 14(2):113–125, 1999.
20. Wu Z. & Leahy R. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 15, 1101–1113, 1993.
21. Zhu J., Hong J., & Hughes, J.G. Using Markov models for web site link prediction. ACM Conference on Hypertext/Hypermedia, 2002.