Identification of Web User Traffic Composition using Multi-Modal Clustering and Information Scent

Jeffrey Heer[1], **Ed H. Chi**
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304 USA
echi@parc.xerox.com

ABSTRACT

As computer scientists, we are constantly seeking ways to understand user behaviors so that we can build better information environments and applications. Therefore, Web site designers, producers, and maintainers often ask the questions: What are my users trying to do on my Web site? What's the mixture of my user traffic? In this paper, we introduce and describe a method to discover major types of information goals of Web surfers automatically. As part of this technique, we use Multi-Modal Clustering (MMC), the Longest Repeating Subsequences (LRS), and Inferring User Need by Information Scent (IUNIS) algorithms to extract significant user paths from the Web server logs, and then we represent these user path profiles using multi-modal vectors that encompass various sources of information, including Content, Topology, and URL. To confirm our method's utility in the real-world, we apply this technique to three different Web sites of varying sizes and purposes, and present the results.

Keywords

Multi-Modal Clustering, Identification of Web Tasks, User Profiles, Information Scent, World Wide Web

Word Count: ~5300

INTRODUCTION

As the vast information ecology of the Web evolves, the ability to quickly assess and comprehend the interests and behaviors of Web users holds a place of ever increasing importance. Current research has made some powerful contributions toward this goal. The Law of Surfing [15] has shown that stable and universal laws govern surfing behavior, while Information Foraging theory [19] has shown that information-seeking Web users can be modeled using the biological metaphor of animals foraging for food. A result of this research is Chi, et. al's Information

[1] Work done as a 2000 summer intern. Current address: University of California, Berkeley, EECS Department, Berkeley, CA 94720 USA, jheer@torus.cs.berkeley.edu

Scent algorithms [8,9], which can (a) infer a user's information need given a user's path and (b) predict user paths given an information need. While there are a wide variety of different surfing behaviors on a Web site, many users have the same information goal. Researchers are seeking ways to inform us of similarities and patterns in Web surfing behavior, so that the Web can be more successful in meeting user needs. Therefore, user interface professionals often ask the questions: What are my users trying to do on my Web site? What's the mixture of my user traffic?

In this paper, we present an automatic system to qualitatively assess site-wide Web usage by providing categorization of significant user types and the percentage and composition of these Web user types. We introduce both a form of clustering, called Multi-Modal Clustering (MMC), which utilizes multiple sources of information (modalities) to generate user groupings and an interface for rapid exploration of these groupings. The clusters generated by MMC can then be used to reveal site usage patterns, identify categories of user information needs, and inform future Web site design. Most importantly, our technique easily and automatically discovers these Web user types and the composition of user traffic.

We organize this paper as follows. First, we discuss important work related to understanding Web user behaviors. Second, we describe our approach to automatically discovering user types. We then present real-world scenarios and case studies with Web sites of varying size and purpose. Last, we discuss some future challenges.

RELATED WORK

Obviously, clustering is an information retrieval technique that has been applied to the Web domain. Most early efforts have concentrated on topic distillation for enhancing surfing on the Web, which is a hot research topic. Most notably, Dumais and Chen described a recent effort on achieving good hierarchical clustering of Web search results using a technique called Support Vector Machines [12].

Most relevant to our project is research on the clustering of usage of a Web site [23,11]. Cooley describes an algorithm that clusters users using a hypergraph partitioning technique [11]. The system is used successfully to identify particularly interesting and similar path histories. It does not come up with significant category groupings and describe the composition of every user profile. Thus, that system will not be able to gain an overall picture of all usage of a Web site. SurfAid [23], on the other hand, gives percentages and counts of user paths, as well as assigning each user path to a user path category. In the literature, there exists very limited information on how SurfAid works, other than that it uses On-Line Analytical Processing (OLAP) methods from the database field. However, we are certain that this system does not use the various sources of information (modalities) that are available to cluster the user profiles. Instead, the user paths themselves are clustered directly. The use of multiple modalities to cluster user needs is novel in our research.

A common, but not entirely adequate solution, to this problem is to use descriptive statistics that are provided by many software packages and services, such as Accrue [1], and NetGenesis [18]. These packages are extremely useful for analyzing events such as products bought and ad click-through rates. However, these solutions fail to

provide an adequate answer to the above questions, because they do not automatically identify tasks and user categories based on user information need.

Employing multiple modalities of information has proven to be a useful methodology for our goal of identifying significant user types. However, the multi-modal methodology is by no means limited to our work. For example, Fass successfully used multi-modal clustering in a system for image retrieval [13], and Allan et al utilized multi-modal features in a system for multi-modal image retrieval [3]. In the field of Pattern Recognition and Computer Vision, there also has been a technique of combining multiple classifiers, e.g. [24]. However, to our knowledge, our specific approach of combining the features into a multi-modal vector is unique. In either case, as long as one is careful not to introduce unnecessary complexity, we believe the practice of utilizing all available information in clustering can significantly improve on more traditional uni-modal techniques.

One area of our own work has focused on recognizing significant user paths from hypertext collections [20], and we use these methods to identify interesting user paths. Cooley has also systematically examined this area [11]. Another area of our work is the visualization of information foraging patterns on the Web [7,8]. Even though the visualizations are useful in finding repeated patterns of navigation, the visualizations do not automatically categorize the paths into similar groupings.

Several Collaborative Filtering [14] systems use Pearson's correlation and other similarity metrics to identify and group similar user profiles for the purpose of recommending informational items to users. Alexa Internet's Web page recommendation system [4] is a notable example of social filtering recommendation applied to the Web. However, these systems' purpose is not to identify significant user groups for analysis, nor to identify significant Web surfing patterns. They are designed for applications that recommend products or Web pages.

A number of efforts have concentrated on characterizing Web surfing behaviors through surveys or protocol analysis. Particularly notable are the early Catledge and Pitkow characterization of browsing strategies on the Web [6] and the more recent Choo et. al. study of 34 participant's browsing behaviors [10]. While these works explicitly studied Web visitation patterns, they did not try to group Web users into specific task types. Instead, they manually analyzed specific browsing methods and strategies used by Web surfers.

METHOD

In this section, we will describe the basic ideas in our approach. As an overview of our approach, the data flow of the method is presented in Figure 1. As depicted in this Figure, we first collect the Content, Usage, and Topology (CUT) data of the Web site to be analyzed. To identify Web user types, we create a representation (or profile) of user interests based on the documents that lie on each user's surfing history, because we assume that, implicitly, each document that a user sees is a part of that user's information interest. We create a multi-modal vector space to describe the features of the Web pages. We then model user profiles as multi-modal vectors that are combinations of the pages they have accessed. We cluster the user profile vectors to obtain significant user type

categories. The resulting clusters are analyzed using the Cluster Viewer, an interface for exploring Multi-Modal Clustering results.

Multi-Modal Clustering (MMC) is a new way to create groupings of items. The idea, which we initially conceived and described in an internal report in Schuetze et al [21], is to use as much information as we have on each item to cluster the items. To briefly describe the algorithm, first, each source of information (or modality) is expressed as a feature vector. Then each modality's feature vector is combined into a single multi-modal feature vector. For example, the content keywords for a Web page can have their own feature vector (e.g., the frequency of each keyword's occurrences on that page), which can be combined with feature vectors that describe the images on that page (e.g., the color of each pixel). All available information on items is embedded into this single large multi-modal vector space, where each modality occupies a sub-space. We then define a similarity metric for the combined multi-modal feature vectors (e.g., a linear combination of the similarity metrics for each individual modality), and then apply traditional clustering algorithms.

We will now turn to a detailed description of our approach outlined above. First, we will describe how we extract significant surfing paths using the Longest Repeating Subsequence (LRS) method [20]. We then extract information needs using the IUNIS (Inferring User Need by Information Scent) algorithm [9]. We will describe how we embed each Web page feature vector as a multi-modal vector in a vector space model. Next, we will discuss how we represent each user path profile as a combination of multi-modal feature vectors describing each Web page. Then we will apply the clustering technique to these user profile vectors.

Extracting Significant Web User Paths

Pitkow and Pirolli [20] systematically investigated the utility of a Web-mining technique that extracts significant surfing paths by the identification of longest repeating subsequences (LRS). A longest repeating subsequence (LRS) is a sequence of items where (1) subsequence means a set of consecutive items, (2) repeated means the item occurs more often than some threshold *T*, where *T* typically equals one, and (3) longest means that although a subsequence may be part of another repeated subsequence, there is at least one occurrence of this subsequence where this is the longest repeating.

They found that the LRS technique serves to reduce the complexity of the surfing path model required to represent

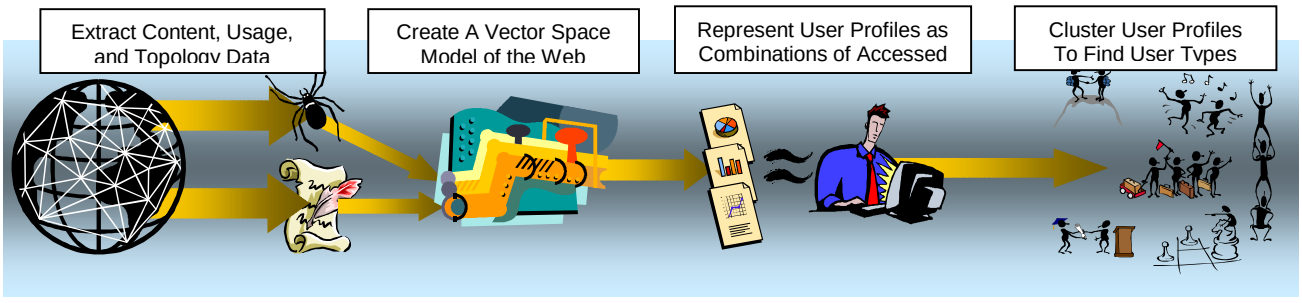

Figure 1: Architectural Data Flow of the Multi-Modal Clustering System for Identifying Web User Types.

a set of raw surfing data, while maintaining an accurate profile of usage patterns. In previous work, we have fruitfully applied this technique to extract significant surfing paths [8]. In essence, the LRS technique extracts surfing paths that are likely to re-occur and reduces noise in the usage data. We use the LRS data mining technique to identify significant surfing paths in Web usage data.

Each significant surfing path is treated as a user profile. Thus, each user profile is essentially a list of documents that represent a significant user history through the Web site. To represent this profile, we build up a feature vector of each Web page, and then construct the profile as a weighted combination of the feature vectors.

Vector Space Embedding of Web Page Features

We must first develop a way to represent each page as a feature vector in order to represent each user profile as a combination of these page feature vectors. To represent each Web page as a feature vector, we need to represent the data within a uniform model. Not only must this model accurately represent the information at hand, it must also provide a readily calculable similarity metric, without which any clustering is impossible. To this end, the common practice is to embed data into n-dimensional vector space. Vector space embeddings provide a rich mathematical framework including a number of possible distance metrics [22]. This includes the Euclidean distance between vectors, and, more commonly, the cosine measure (the cosine of the angle between any two given vectors).

So after extracting the CUT data of a Web site, we represent a single page as a multi-modal vector that is comprised of four modalities in the current implementation: page content, URLs, inlinks, and outlinks.

- The content modality consists of each unique content word that appears in the body of the Web pages.
- URLs are broken up into tokens delimited by forward slashes (‘/’) and each of these tokens is treated as a separate term.
- Inlinks for a given page consist of all the hyperlinks on other pages within the Web site that link to that page.
- Outlinks are all hyperlinks on a given page which link out to other pages, whether within the Web site or not.

We can use more than just these four modalities. Most importantly, each modality here is weighted using TF.IDF.

Term Frequency by Inverse Document Frequency

In this vector space, we often need to weight elements according to their importance in a collection. A particular way to do this is by ***TF.IDF*** (Term Frequency by Inverse Document Frequency) weighting schemes [22, p.542]. The TF.IDF value is a real number indicating the relative importance of a term in a given document. This value is determined by the number of times the term appears in the document (term frequency) weighted by the ratio of the number of all documents to the number of documents that contain the term (inverse document frequency). The insight behind this scheme is that an often-used term may hold high relevance on a given page, but this relevance should be reduced if this term appears regularly throughout the entire document collection. Similarly, if a term only appears on a single page, it should be weighted higher due to its uniqueness. While TF.IDF weights are

traditionally used with content words as the terms, we have expanded this approach to other forms of document information such as hyperlinks, and URL tokens.

By treating each of the modalities separately, we can easily use TF.IDF schemes to create the needed weighted numerical representation.

Representing Each User Profile as a Multi-Modal Vector

Once we complete the construction of the multi-modal vector space model of the pages, we construct user profile vectors as weighted combinations of accessed pages. To do this, we use the IUNIS algorithm [8,9] as follows[2]:

Before summing the related page vectors that make up a user surfing path, each vector is scaled by two different terms according to the IUNIS algorithm. The first is a page access TF.IDF weighting. In this case the term frequency corresponds to the access frequency of the page by the given user and the inverse document frequency corresponds to the ratio of total users to the number of distinct users who have accessed the given page. This helps to reduce the weight of pages that are accessed by many users and may not be very relevant to the user's information need (e.g. a site's splash page).

The second term is a path position weighting, in which we weight pages in favor of recency of access. Here we are assuming that the further along a page is in a surfing path, the more likely it is to be representative of the user's information goal.

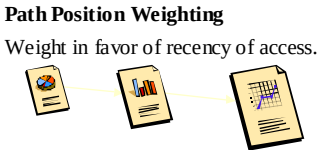

Figure 1b: Path Position Weighting

We produce the user profile vectors as a weighted linear combination of the page vectors using the above two weighting terms. These user profile vectors are representations of user surfing activities.

Multi-Modal Clustering

Once we have these representations of the user surfing activities, we can cluster them into user type categories.

Clustering is a form of statistical data analysis that organizes a data set into individual *clusters* – element groupings whose membership is determined by a shared similarity. This similarity is measured using a computable *distance metric*, a function whose value indicates the relative distance of two elements. In general, clustering algorithms appear in one of two forms: agglomerative or partitioning. *Agglomerative* algorithms work hierarchically, first merging individual elements of highest similarity, and then recursively merging clusters until the desired level is reached. Thus, Agglomerative algorithms are often called *Hierarchical* algorithms.

Partitioning algorithms, on the other hand, start with a set of seed vectors and then undergo a series of iterations in which elements are assigned to clusters and then cluster centers are recomputed.

One of the best-known clustering algorithms is *K-Means*, a partitioning method originally formulated by MacQueen [16]. K-Means begins by choosing *k* random vectors as initial cluster centers. Then each vector is assigned to the cluster to which it is most similar, as determined by a distance metric comparison with the cluster center. Cluster centers are then recomputed as the average (or mean) of the cluster members. Then the process repeats, ending either when the clusters converge or a specified number of iterations have passed.

Another closely related partitioning clustering algorithm is *Wavefront*, a new variant of the K-Means algorithm that features an advanced form of cluster seeding, also first described in the internal report [21]. At first a number of random vectors are chosen and their average is computed to produce a global centroid *c*. Initial cluster centers are then computed as points in between the centroid *c* and one of *k* randomly selected vectors. This is determined by the formula $x_i = \alpha c + (1-\alpha)k_i$, where x_i are the initial cluster centers, *c* is the global centroid, k_i are random seed vectors, and α is a parameter less than or equal to 1. By using a more developed seeding, Wavefront helps to reduce the number of necessary K-Means iterations. The name 'Wavefront' comes from visualizing the seeding as a wave spreading out from the global centroid towards the seed vectors.

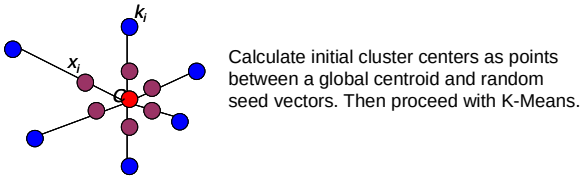

Figure 1c: Wavefront Clustering

Traditionally, clustering approaches to data analysis only use one source – or *modality* – of information. By utilizing multi-modal clustering, we are able to include within our representation not only data on Web page content, but a myriad of other information. The possibilities include page URL, topology, image statistics, and user demographics (if available). In our approach we use information modalities of content, URL, inlink, and outlink, as mentioned above.

On this last step, we feed the collection of user profile multi-modal vectors into the Wavefront clustering algorithm to generate clusters representing Web user types. We can also use a hierarchical clustering algorithm. These clusters can then be refined and analyzed to reveal the different Web user types.

[2] In the interest of saving space, here we omit the description of the spreading activation step originally in the IUNIS algorithm, but it can certainly be applied here. For details, see [9].

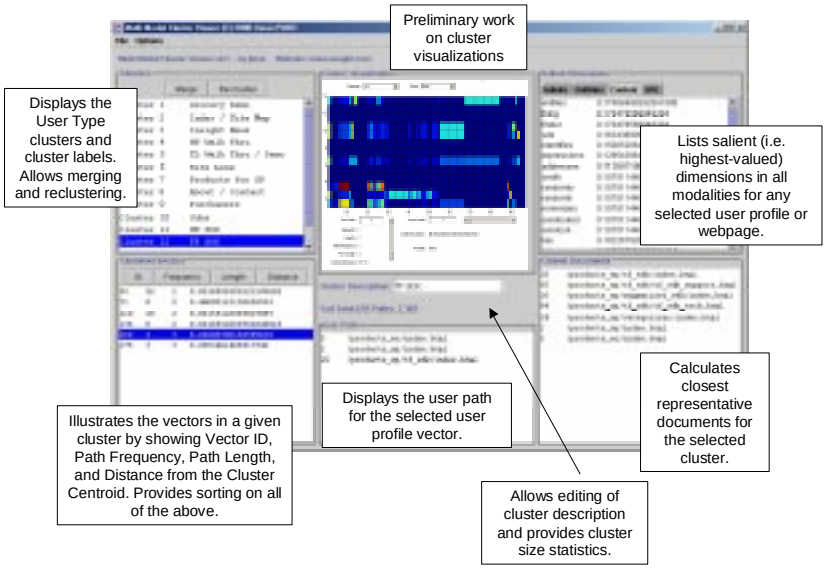

Figure 2: ClusterViewer enables users to browse the clustering results, and analyze for interesting patterns in each cluster.

Cluster Analysis

To analyze the results, we have developed the Cluster Viewer shown in Figure 2, a browser interface that provides quick access to cluster data as well as cluster refinement operations. The interface enables browsing of each cluster data and the vectors in each cluster: (a) The viewer supports user-defined cluster labeling and presents size percentage statistics for each user grouping. (b) It provides a view of the salient (i.e. highest-valued) dimensions for all modalities, highlighting the most important content words or links for the selected cluster or vector. (c) The viewer also computes the most-closely-related Web pages for a given cluster. (d) Paths in a cluster can be sorted by vector ID, path frequency, path length, or distance from the cluster centroid. (e) Drilling down to individual user paths can determine cluster exemplars and check clustering accuracy. (f) Finally, merging and reclustering operators can further refine the clusters. The reclustering supports a full range of clustering options, including changing the modality weights, choosing the clustering algorithm, and specifying different clustering parameters. In short, the Cluster Viewer allows rapid exploration and labeling of the user type clusters and provides mechanisms for improving upon the initial clustering results.

Implementation

In our implementation of this method, we constructed an integrated toolkit that supports Web data extraction, information processing, clustering, and cluster analysis. It functions as a component of ScentBench, a Web analysis suite currently under development at Xerox PARC. ScentBench is built on top of PIPes, an information processing platform for researchers [2]. The system is written primarily in Java, with the exception of the LRS methods, which are written in Perl.

CASE STUDIES

To test the efficacy of our system, we applied our methods to three websites of varying scale and diversity. Our first case was Inxight (www.inxight.com), a provider of Web site enhancement and visualization products. This site is small and well-organized, providing a great testing ground for our methods as well as a good representative of a burgeoning company's Web site. For our second case we chose the Web site for the Computer Science department of the University of Minnesota (www.cs.umn.edu). This site serves as a good example of a large, diverse site with a great variety of focused visitors. Finally, we focused on the Xerox corporation home page (www.xerox.com), testing our system on a large, heavily used site. As the case studies show, Multi-Modal Cluster provided an automated way to identify major user types and percentages of the particular mixture of user traffic on each of the sites.

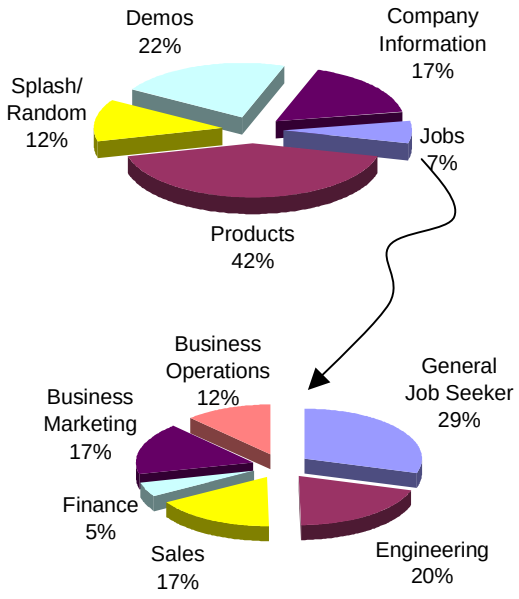

Figure 3: Inxight.com – top (general distribution), bottom (Jobs subcategory distribution).

www.inxight.com

Data were collected and analyzed for the dates of July 17-23, 2000. The significant paths generated from the usage logs consist of 2126 distinct significant surfing paths, reduced from 125,429 total paths using LRS!

The top of Figure 3 shows the generalized usage of the Inxight Web site. As expected, many users come to the site for demos and product information. More interestingly, 7% of users are job seekers. We suspected that Inxight might be interested in the composition of its job seekers, so we performed another round of clustering only on these job seekers, who are depicted in the bottom portion of Figure 3. Although Inxight is a technology company, most of the job seekers who looked at these Web pages were business development people (51%). This fact may be indicative of the tight labor market for software engineers, who primarily get sought out rather than seeking.

Our clustering program enables this capability to "zoom" into a sub-category of a particular user type and re-cluster. This enabled us to further understand the composition of users in that sub-category. We used this technique to further understand the other sub-categories as well, such as the relative usage of the different Inxight products contained within the 'Products' sub-category, and the various product demos in the 'Demos' sub-category.

Another interesting aspect of the clustering was that it seemed to distinguish between different foraging behaviors. At the high-level, many of the user paths whose activity was centered on the splash page and site map showed no

obvious user interest (12%). The users accessed random, often unrelated pages and backtracked frequently. We saw this behavior in the 'zoomed-in' views of the sub-categories as well. For example, both the Jobs and Products categories contained a sub-cluster featuring more general, undirected browsing (e.g., 29% of all job seekers looked at all of the job postings). These grazing and roaming behaviors may represent users with less defined information needs, or users who were having difficulty satisfying their need.

As we saw, for Inxight.com, Multi-Modal Clustering of the user profiles automatically identified these significant user types, and furthered our understanding of user behaviors.

www.cs.umn.edu

We collected and analyzed for the dates of July 19-25, 2000. The LRS paths generated from the usage logs consist of 1284 distinct significant surfing paths, representing 17,831 total paths. We show the general usage distribution in Figure 4. As we see, there are a wide variety of different user interests for this Web site. Not surprisely, Research (8 clusters, 28%), Dept Info (3 clusters, 26%), and Graduate Admissions (1 cluster, 15%) comprise a large percentage of users. Other clusters corresponded to systems help/support (5 clusters), faculty pages (3), employment (1), student society (1) and courses (3), Cisco certification classes (2), conferences (2), and the splash page / random browsing (1).

Many professors want to know how much attention their research is getting. Accordingly, the distribution of users interested in research is displayed in the bottom portion of Figure 4. The GroupLens collaborative filtering (19%), the Ajanta mobile agents (18%), and the GIMME multimedia user interface (26%) projects received the most attention.

Our approach of applying Multi-Modal methodology enabled us to find these different granularities of user behaviors. Most notably, the general grazing foraging behavior we saw in the Inxight and the Xerox data is nearly missing from this data set (only 1%). We believe this may indicate that Web surfers for this Web site have more goal-oriented tasks.

We noticed that users interested in courses made up only a small percentage of the total. Since our data comes from a period when school is not in session, this is not entirely surprising. However, we would like to analyze the site again when classes have resumed and contrast the results.

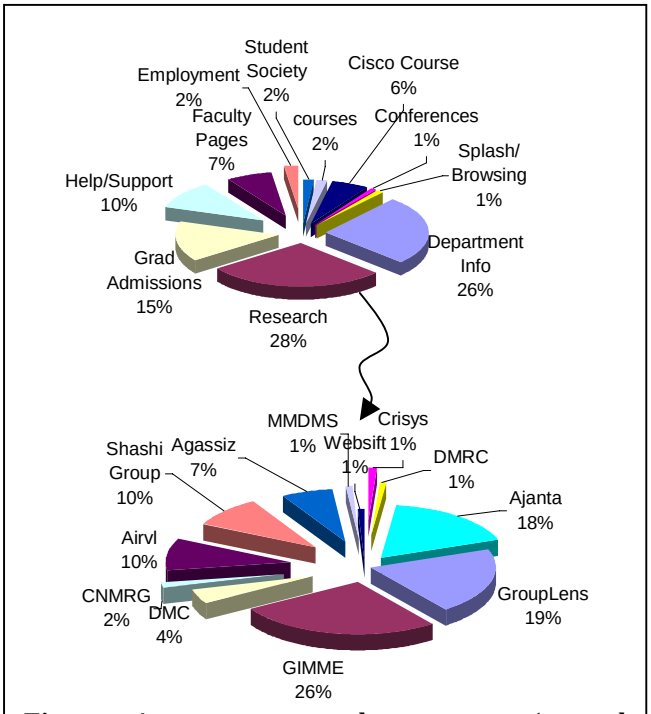

Figure 4: www.cs.umn.edu -- top (general distribution); bottom is Research subcategory.

www.xerox.com

We collected and analyzed data from analyzing May 16-19, 1998. While this data set is slightly dated, for comparison purposes, we used this same data set in previous related papers [7,8]. This is by far the largest data set we processed, with 17,831 LRS distinct significant surfing paths, representing 394,778 total paths.

The MMC results are shown in Figure 5. We see that various areas received differing amounts of attention from the visitors. While the product clusters (14 clusters, 41%) received a great amount of attention as expected, what was unexpected was the number of user profiles that were related to one specific product (TextBridge Pro98, 9 clusters, 16%). Investors comprised 6% of the significant traffic, while Company Info Seekers comprised 9%. There is a significant amount of general undirected random browsing (2 clusters, 14%), indicating that many users coming to the Xerox site has less defined information needs, or they were having difficulty satisfying their need.

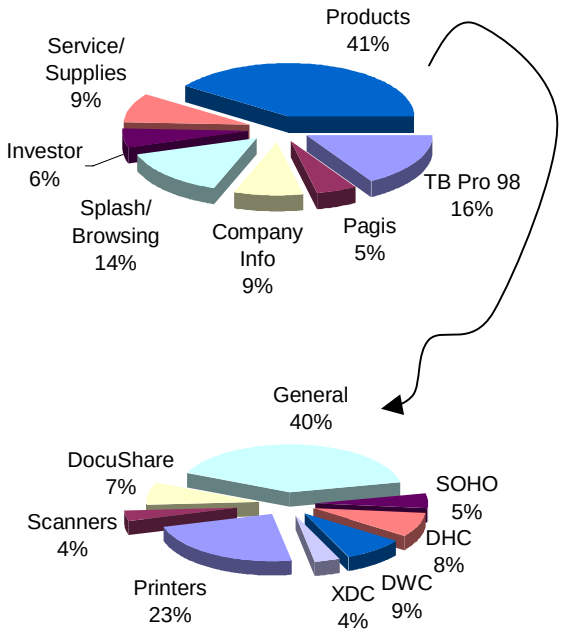

Figure 5: Xerox.com data – top general breakdown; bottom Product breakdown

Multi-Modal Clustering enabled us to further broke down the Products cluster automatically, and found that there are users looking for specific information on particular product series, such as Document WorkCenter (DWC) and Document HomeCenter (DHC). Within this Product Seekers category, we see that a large percentage of user profiles have undirected roaming behavior (40% of all Product-related traffic). This is probably because many users have little idea of the specific product category they should look in for their product needs, which is a fact that is useful to Xerox marketing department.

Summary

In this section, we showcased three real-world scenarios of analyzing user profiles and identifying significant user information needs. We showcased the scalability of our method by applying it to three Web sites of varying size and purpose. We accomplished this by extracting significant user profiles by utilizing the LRS, IUNIS, and Multi-Modal Clustering. In each case, we also showed further sub-clustering of a major interesting user profile category. This enabled us to understand user behaviors at several granularities. In two of the cases, we saw both directed and roaming foraging behaviors. We are interested in using this technique on these Web sites over a longer period of time, seeing how user composition and information goals change over time, identifying stable and dynamic user types, and then using this information to inform site design methodologies.

CONCLUSION

As the Web and its associated usage grow by leaps and bounds, the task of understanding how users are foraging will remain important. In the quest to understand the daunting chain of user interactions on the Web, researchers and analysts need tools for getting quick and accurate pictures of site-wide usage. In this paper, we have described a novel method that utilizes multiple modalities of information to group similar user profiles into significant user categories. The system enables researchers and analysts to extract and analyze significant user types at several granularities, and understand the mixture or composition of users visiting the site. While limited in some aspects, the tool provides accurate profiles of site usage, and helps inform Web site usability and design. This capability significantly improves our ability to understand the different foraging behaviors on the Web.

ACKNOWLEDGMENTS

Jeff Heer was supported under a summer 2000 internship. Hinrich Schuetze, Jun Li, Jim Pitkow, Peter Pirolli, and Francine Chen contributed the initial formulation of the Multi-Modal Clustering algorithm, which we utilize here for clustering user profiles. We also especially thank Peter and Hinrich for initial conversations that led to this research. Rob Reeder and Pam Schraedley helped proofread this paper. This research was funded in part by Office of Naval Research Contract N00014-96-C-0097 to Peter Pirolli and Stuart Card.

REFERENCES

1. Accrue Insight. (1999) http://www.accrue.com
2. Adar, E., Larner, D. (1999) The PIPes Information Processing System. Intranet at Xerox PARC.
3. Allan, J., Hanson, A., Manmatha, R. (2000) Multimodal Indexing, Retrieval and Browsing: Combining content-based image retrieval with text retrieval. http://ciir.cs.umass.edu/projects/mmir.html
4. Alexa Internet (1999) http://www.alexa.com
5. Boshart, J., Pirolli, P. (1999) Multi-modal Clustering Produces Groupings Similar to Human-Produced Groupings. Xerox PARC User Interface Research Tech Report.
6. Catledge, L.D., Pitkow, J.E. (1995) Characterizing Browsing Strategies in the World Wide Web. In Proceedings of the Third International World Wide Web Conference. Darmstadt, Germany. http://www.igd.fhg.de/www/www95
7. Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., and Card, S. (1998) Visualizing the Evolution of Web Ecologies. *Proceedings of the Human Factors in Computing Systems, CHI '98.* (pp. 400-407). Los Angles, CA.
8. Chi, E.H., Pirolli, P., and Pitkow, J. (2000) The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a Web site. In *Proc. of the ACM Conference on Human Factors in Computing Systems, CHI 2000* (pp. 161-168), The Hague, Netherlands.
9. Chi, E.H., Pirolli, P., Chen, K., Pitkow, J. (2000) Using information scent to model user information needs and actions on the Web. *Proceedings of the Human Factors in Computing Systems, CHI '2001*. (to appear). Seattle, WA.
10. Choo, Chun Wei, Detlor, Brian, Turnbull, Don. (2000) Working The Web: An Empirical Model of Web Use. *33rd Hawaii International Conference on System Science (HICSS)*, Maui, Hawaii. Jan. 2000.
11. Cooley, R. (2000) *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. Ph.D. Thesis. University of Minnesota. May 2000.
12. Dumais, S., Chen, Hao. (2000) Bringing Order to the Web: Automatically Categorizing Search Results. *Proceedings of the ACM Conference on Human Factors in Computing System, CHI 2000* (pp. 145-152). The Hague, Netherlands.
13. Fass, Adam. (1999) PicturePiper: A user interface to image-finding services. Internal Xerox PARC report.
14. Herlocker, J. (2000) *Understanding and Improving Automated Collaborative Filtering systems*. Ph.D. Thesis, University of Minnesota, Sep. 2000.

15. Huberman, B.A., Pirolli, P., Pitkow, J.E., and Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. *Science, April 3, 1998, vol. 280, num. 5360* (pp. 95-97).
16. MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pages 281-297, Berkeley, University of California Press.
17. Medin, D. L., Lynch, E. B. Coley, J.D., Altran, S. A. (1997) Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*: 32, 49-96.
18. NetGenesis. (1999) http://www.netgenesis.com
19. Pirolli, P. and Card, S.K. (1999) Information Foraging. *Psychological Review 106(4)* (pp. 643-675).
20. Pitkow, J., Piroll, P. (1999) Mining longest repeated subsequences to predict World Wide Web surfing. *Proceedings of the USENIX Conference on Internet.*
21. Schuetze, H., Pitkow, J. E., Pirolli, P., Chi, E. H., Li, Jun. (1999) System and Method for Multi-Modal Clustering. Xerox PARC Technical Report.
22. Schuetze, H., Manning, C. (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
23. SurfAid. (1999) http://surfaid.dfw.ibm.com
24. Woods, K., W. P. Kegelmeyer Jr., and K. Bowyer, Combination of multiple classifiers using local accuracy estimates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 4, pp. 405-410, April 1997.