MINING USER BEHAVIOR FOR RESOURCE PREDICTION IN INTERACTIVE ELECTRONIC MALLS

Silvia Hollfelder

GMD - IPSI
Dolivostr. 15
64293 Darmstadt, Germany
hollfeld@darmstadt.gmd.de

Vincent Oria and M. Tamer Özsu*

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2H1
{oria, ozsu}@cs.ualberta.ca

ABSTRACT

Applications in virtual multimedia catalogs are highly interactive. Thus, it is difficult to estimate resource demands required for presentation of catalog contents. In this paper, we propose a method to predict presentation resource demands in interactive multimedia catalogs. The prediction is based on the results of mining the virtual mall action log file. The log file typically contains information about previous user interests and browsing behavior. These data are used for modeling users' future behavior within a session. We define heuristics to generate a start-up user behavior model as a Continuous Time Markov Chain and adapt this model during a running session to the current user[1].

1. INTRODUCTION

An electronic mall can be seen as a virtual catalog that integrates distributed electronic catalogs [6]. Such catalogs enable users to transparently search or browse items from various providers. Catalog providers can increase the probability of selling their items by using short multimedia shots for product descriptions.

Such distributed multimedia systems need continuous delivery of time-dependent data. Resource management mechanisms (e.g., admission control) have to consider limited system capacities in order to ensure a certain Quality of Service (QoS).

Virtual multimedia catalog applications are highly interactive, as users browse through a catalog, briefly inspect items of interest and switch to new items. These applications need a low start-up latency for subsequent multimedia shot presentations. Thus, for virtual catalog applications we manage resources for whole multimedia sessions, i.e., a sequence of short media presentations, containing discrete media presentations (e.g., images and text), as well as the playback of continuous data (e.g., video, audio, and VRML). One problem of resource management for whole sessions is to predict the resource needs, which we define as those required for presenting item descriptions, but not the resources required by the business processes involved in the completion of the purchase transaction.

These resource needs are determined by interactive user behaviors that are difficult to predict as user behaviors are not constant. A user who is looking for clothing is likely to have a different behavior than when s/he wants to buy a car. Hence, the user behavior is more likely to be related to what s/he is looking for.

Existing work on user behavior modeling in multimedia applications has focused on particular applications. In [5] the authors study user behavior for accessing movies. They observe which movies will be rented at which time at a video store. Popularity patterns of news articles are explored in [3] and [7]. In [4], the user search behaviors in browsing applications are analyzed, and suggestions with respect to items of interest are given by means of a learning mechanism. Heuristics for modeling user behavior in conceptual video browsing are introduced in [1]. One drawback is that no general assumptions can be deduced, since the user behavior is typically application-dependent. Another problem is the lack of empirical work that answers the questions "*how long* does a user present single media shots?" or "*which* objects will be preferred by a user?" when s/he browses through a digital multimedia library.

Observing user behaviors for commercial purposes is not unusual. Nowadays, companies like super-markets and banks collect user profiles and keep track of customer transactions. To make sure that customers give personal information, some supermarkets distribute member cards with some discounts for card holders. The company keeps a transaction file that tells which items the customers bought when,

---

*Current address: Computer and Information Science Department, New Jersey Institute of Technology, University Heights, Newark, NJ, 07102-1982, oria@cis.njit.edu.

[1]This research is supported by grants from the Canadian Institute for Telecommunications Research (CITR) under the Network of Centres of Excellence (NCE) program of the Government of Canada and from the German Academic Exchange Service (nr. D/99 25683).

where, etc. The same kind of information is available for electronic commerce through the log file.

Thus, in our approach, we log user past behaviors for resource prediction. We first describe, in Section 2, how log files can be mined and define appropriate user behaviors and types from that. For an individual session, we employ that information and heuristically build a user behavior model. From that model the resource demands can be calculated. Section 3 concludes with a summary and ongoing work.

2. APPROACH

2.1. Electronic Catalogs

Electronic catalogs are interactive and can contain multimedia data. In this sense, an electronic catalog can be seen as a set of hypermedia documents. Each document is defined by a structure, a content and a presentation. Therefore, standardized document definition languages, such as XML (eXtensible Mark-up Language), can be used. In this case, the structure of the document can be defined by a DTD (Document Type Definition), the content itself can be represented in an XML document, and the presentation of single documents can be specified by structures such as XSL (Extensible Stylesheet Language). Standards such as SMIL (Synchronized Multimedia Integration Language) that expresses multimedia presentations by a language specified in a DTD, can be used for the specification of the presentation of a sequence of hypermedia documents.

We use an object-oriented approach to design the content of a catalog. We distinguish two hierarchies: the *Item* hierarchy and the *Catalog* hierarchy which can be subtyped by the application designer. An item I is an object that can have some related items RI (for example trousers and shoes that match a shirt). Each item has a set of properties $A^I = \{A^I_j\}$, (j is an integer) among which at least one is the *item description* and another is the item identifier (iid). An item description can be a text, an image, a video, an audio, or a VRML document. An electronic catalog, then, is an object composed of *catalog pages*. Each catalog page refers to one item. Users access catalog pages through queries.

A *catalog session* is a sequence of discrete and continuous media presentations of catalog pages. A query initiates a session and, within a session, queries are used to refine the result. A query result can be represented by a hypergraph. A session related query returns a new graph, containing a subset of the nodes (i.e., items) of the initial graph. A running session terminates when a user sends a new query that is not related to the query that created the session.

We assume that a user U has a set of properties $A^U = \{A^U_i\}$ with i an integer, among which uid is the user identifier. The values of these properties are kept in a user profile. User profiles contain information about individual user preferences such as media preferences, and user specific meta data (e.g., age, sex, address, profession, family status).

2.2. Mining Log Files

The log file we are considering in this paper records *user actions* in browsing virtual catalogs. We call it the *action log file*. By action we mean (1) what the user is interested in (not necessarily what s/he buys), (2) the pages s/he visits, (3) the time s/he spends on a page, and (4) the time s/he spends on each media in a page. The actions are recorded even if the user does not buy any item. Each line in the action log file is called an action line. The action log records for each session k the user (uid_k), the number of pages np_k (items) found by a query Q, and the number na_k of pages accessed within a session. From np_k and na_k, we compute Pna_k that gives the percentage of pages accessed.

For each page l accessed (note that $l = iid$ since we define a page for an item in our model), we monitor the overall time (t_{kl}) that a user spends on the page, the time t^m_{kl} that is spent on each media type m (text [t^{tx}_{kl}], image [t^{im}_{kl}], video [t^{vd}_{kl}], sound [t^{sd}_{kl}] and VRML [t^{vr}_{kl}]). When text and images are mixed in a catalog page, only the total time spent on the page is recorded. For continuous media, the time recorded is the ratio of play time to the original presentation duration. We assume that audio, video, and VRML objects in a page are represented by hyperlinks which open up new windows allowing us to record these times separately. We also record the ratio $Pnri_{kl}$ (number of related item pages accessed to number of related item pages referenced in the page). Thus, a session sample space in the action log file can be represented as $\Omega = \{uid_k, Pna_k\{t_{kl}, t^{vd}_{kl}, t^{sd}_{kl}, t^{vr}_{kl}, t^{im}_{kl}, t^{tx}_{kl}, Pnri_{kl}\}\}$.

2.2.1. Mining For Item-Related User Behaviors

From the action log file we extract a subset referring to the same item or the same category of items to define the behavior of users accessing this item or category of items. For each of the variables $\{Pna_k, t_{kl}, t^{vd}_{kl}, t^{sd}_{kl}, t^{vr}_{kl}, t^{im}_{kl}, t^{tx}_{kl}, Pnri_{kl}\}$, we generate a discrete distribution function representing user behaviors in all sessions. To ignore outliers we reduce the behavior model, thereby removing unlikely future actions from our predictions. Thus, we aggregate the user behavior for a critical mass of users (for example 80 percent), using the cumulative distribution function derived from the distribution functions for sampled users.

We finally derive the *average aggregated behavior B* of all the users who access particular items. As a result, the so-called *item-related behavior* of a critical mass of users can be specified as a tuple $B = < iid$, $BPna$, Bt, Bt^{vd}, Bt^{sd}, Bt^{vr}, Bt^{d}, $BPnri >$. Note that in addition, the behaviors can be aggregated at different levels of the item category hierarchy. All the values are averages of the accumulated values with $BPna$ (average of accumulated Pna) percentage of pages accessed, the time (Bt) spent on a page, the time (Bt^m) spent on each continuous media, the percentage

($BPnri$) of related item pages accessed, and the time (Bt^d) spent on the discrete media (text and image) all together.

2.2.2. *Rules on Users*

The action log file is also mined to find some rules: *behavior-related rules* and *item-related rules*. A *behavior-related* rule correlates the presence of a group of users and a group of behaviors. A *behavior-related rule* assign a behavior to a group of users if their behaviors are similar (within a certain threshold) no matter which items they are looking for. An example of such a rule is "customers less than 20 year old have the particular behavior B_{20}". For example, B_{20} can be a behavior in which video- and VRML-documents are preferred. *Item-related rules* specify which items a group of users are interested in, an example is "women around 30 like VW-beetles".

Finding rules in a log file is the classical association rule mining problem [2]. In our case, we are interested in finding what type(s) of customers look for what type(s) of items and which users have the "same" behaviors. A rule has some confidence (a rule $X \implies Y$ has confidence c if $c\%$ of the action lines that contain X also contain Y) and support (a rule $X \implies Y$ has support s if $s\%$ of the action lines contain X or Y). Item- and behavior-related rules with support and confidence greater than a minimum user-specified confidence level and user-specified support level are turned into *user types*.

2.2.3. *User Types*

User types are equivalence classes over users. Each user type T has a set of properties $A^T = \{A_n^T : cond_n^T, B\}$ where $A^T \subseteq A^U \cup A^I$ with n as integer, $cond_n^T$ is a condition attached to property A_n^T, and B is a behavior. The behavior is not always required (user types derived from item-related rules do not have any behavior). In the same way, some user types representing behavior-related rules might not have any item property. The rule "customers older than 25 looking for a car have behavior B_{25}" can be represented as $T_{25} = (U.Age \geq 25, I.Item\_type =' car', B_{25})$ and "women around 30 like VW-beetles" can be represented as $T_W = (28 \leq U.Age \leq 35, U.Sex =' F', I.Item\_type =' car', I.Item\_model =' VW - beetle')$. In the above examples, $U.Age$ and $U.Sex$ are user properties, and $I.Item\_type$ and $I.Item\_model$ are item properties.

2.3. Resource Prediction

We use the term *role* to designate a user behavior within a session. The user identifies himself at the beginning of a session and sends a query. A query Q results in a limited number of items I and related items RI. Thereby, we are able to map the user role to one or several user types if the user has common properties with the user type, and/or if the items s/he is looking for have common properties with the items in the user type. Furthermore, we assume that the meta data, including data rates, for all these item presentations in the catalog are available. A session can be divided into subsessions if the query refers to several item types.

Using the information on the user and the found items, a start-up user behavior model can be generated at the beginning of a new catalog session. The idea here is to generate the start-up model by focusing on the actions a user is likely to do. This model represents the access probabilities of the items, the item descriptions, and the time a user spends on single item presentations. We employ the information, recorded in item-related user behavior and/or user types, for the specification of the model. From that model we predict the resource demands of a catalog session.

2.3.1. *Start-up Model*

We use Continuous Time Markov Chains (CTMC) to represent user behaviors. CTMC consists of a set of states $\{S_i\}$ with exponentially distributed holding times of the states (mean $\frac{1}{v_i}$) and transition probabilities $p_{i,j}$ between the states. The user behaviors within a state are memoryless. Thus, the probability of stopping the presentation of an item is independent of the time a user has had it presented. The assumption of having exponentially distributed holding times is realistic for multimedia catalog applications since we have short media presentations.

The structure of a CTMC is generated top down, starting with the catalog pages and refining the model by the item descriptions of a catalog page: the top level of Figure 1 represents the user browsing behavior among the catalog pages I^Q that refer to a query Q and the related items RI^Q of the hits. Users can only browse from one item to related items, or jump back to the root page (i.e., result hits of a query).

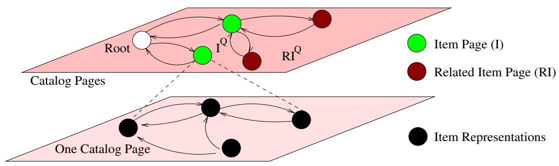

Figure 1: The CTMC of a Start-up Model

Now we specify the parameter setting of the model, namely the time a user will spend within one catalog page (i.e., holding times) and the transition probabilities in between the catalog pages. To generate the transition probabilities from the root to the single catalog pages, we proceed as follows: (1) If the user types contain no information about the hits I^Q, we simply take, for each catalog page, the same transition probability value. Thus, each item of the set I^Q has the transition probability $\frac{1}{np_k}$ from the root, with np_k as number of hits. (2) Otherwise, we employ the rules of the user types satisfying the items a user queried as follows: weight the item pages that are not relevant for the user types with the factor w^- and weight the prefered items pages with a factor w^+. In case of conflicts among multiple user types, the more general rules (of the item hierarchy) have higher priority with respect to the weighting. (3) To ensure that the sum of all transition probabilities from the root to the catalog pages is equal to 1, the transition probabilities have to be normalized. For related items, we specify the transition probabilities by interpreting user types, too.

Now, we specify the user preferences for the various item descriptions within a catalog page. The second level of Figure 1 displays the browsing behavior within a catalog page. Note that this layer is a refined subgraph of a node representing an element of I^Q or RI^Q of the top level. We specify a user behavior model for each catalog page. This is necessary since the presentation times of the item descriptions differ. The transition probabilities within a catalog page are generated similarly as for the top level. These probabilities can be overwritten if the user profile specifies some media preferences. We assume that the holding time of a catalog page is determined by the item-related browsing behavior. The mean holding times of the item descriptions are adapted from the time the critical mass of users spends for each media m. The heuristic rule for all media m is: $\frac{1}{v_m} = Bt^m$. This means that we take the average past presentation time ratio for a media m of the critical mass of users and map it to an exponentially distributed function with mean $\frac{1}{v_m}$.

The CTMC can be analyzed, using transient analysis, to calculate the access probability of each state of the chain. Based on this analysis, the resource demands of single time intervals of a look-ahead window can be computed using data rates of the media involved. For more details on the mathematical analysis we refer the reader to [8]. The predicted resource demands are used for the management of resources of that session.

2.3.2. Dynamic Adaptations on the User Role

Since users do not behave the same all the time, we adapt the prediction to a user's current role. Therefore, we observe a user within a session k for a period of Bt (time the critical mass of users spend on one page). This means that we assume the user behavior at the beginning of a session to be representative of his current role.

Then we adapt the general start-up model to the user current behavior (e.g., ratio of presentation time, media selected, etc) for a more precise prediction. The heuristic rule is to decrease the probabilities and the mean holding times on the media types the user spends less time on.

This adapted model is used for resource management of running sessions. For example, for admission control, the future resource needs of admitted clients can be calculated from the adapted model (if a session is longer than Bt active). For the pending clients, the start-up model is used. We predict the resources for an estimated session duration. This duration is dependent on the number of hits of the user's query, the values $BPna$ and Bt. Thus, we employ here item-related behavior, too.

3. CONCLUSION

In this paper we specified parameters to observe user behaviors in electronic multimedia catalogs. We developed heuristic methods to predict resource demands of virtual catalog sessions. These methods employ predefined user types and behaviors obtained by mining the log file, and are adapted to the current user session.

In the future, we want to consider the session history, i.e., the order in which users select catalog pages and item descriptions. One property of CTMC is that it can be very complex when a large number of items and item representations are found by a query. For that case, we want to investigate how to reduce the probability on the paths of the model the user will likely not follow. Furthermore, we want to make empirical studies on user behavior in multimedia electronic mall applications.

4. REFERENCES

[1] K. Aberer and S. Hollfelder. Resource prediction and admission control for interactive video browsing scenarios using application semantics. In *Proc. of Semantic Issues in Multimedia Systems (DS-8)*, pages 27–46, Jan. 1999.

[2] R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In *Proc. of VLDB*, pages 487–499, Sept. 1994.

[3] T. Choi, Y.-J. Kim, and K.-D. Chung. A prefetching scheme based on the analysis of user access patterns in news-on-demand systems. In *Proc. of ACM Multimedia*, pages 145–148, Oct./Nov. 1999.

[4] C. Drummond, D. Ionescu, and R. Holte. Intelligent browsing for multimedia applications. In *Proc. of ICMCS*, pages 386–389, June 1996.

[5] C. Griwodz, M. Bar, and L. C. Wolf. Long-term movie popularity models in video-on-demand systems. In *Proc. of ACM Multimedia*, pages 349–357, Oct. 1997.

[6] A. Keller. Smart catalogs and virtual catalogs. In R. Kalakota and A. Whinson, editors, *Readings in Electronic Commerce*, pages 259–274, 1997.

[7] T. D. Little and D. Venkatesh. Popularity-based assignment of movies to storage devices in a video-on-demand system. *Multimedia Systems*, 2(6):280–287, Jan. 1995.

[8] H. C. Tijms. *Stochastic Models. An Algorithmic Approach.* Wiley, 1994.