IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4916

Discovering Behavioral Traversal Patterns of Users of Streaming Information Systems

Ajumobi Udechukwu
Advanced Database Systems and Applications Laboratory, Department of Computer Science, University of Calgary, 2500 University Dr. N.W. Calgary, AB T2N IN4, Canada, ajumobiu@cpsc.ucalgary.ca

Ken Barker
Advanced Database Systems and Applications Laboratory, Department of Computer Science, University of Calgary, 2500 University Dr. N.W. Calgary, AB T2N IN4, Canada, barker@cpsc.ucalgary.ca

Reda Alhajj
Advanced Database Systems and Applications Laboratory, Department of Computer Science, University of Calgary, 2500 University Dr. N.W. Calgary, AB T2N IN4, Canada, alhajj@cpsc.ubalgary.ca

ABSTRACT

Several research efforts have been directed towards mining the navigational patterns of users of information systems. Previous research on identifying navigational patterns was directed at browsing patterns of web users. The contents of web pages do not change frequently, thus, existing approaches for discovering navigational patterns are guided by this philosophy.

In this work we motivate a new domain for data mining, which involves mining the navigational patterns of users in environments with streaming content (for example, cable-TV viewers). The access sites we study in our framework display continuously changing content unlike web pages that are relatively static. This fundamental difference in site characteristics calls for new algorithms for identifying interesting navigational patterns. The algorithms proposed in this work are generalized and can be applied to any system with similar characteristics. The approach used in this paper is based on behavioral predicates, and incorporates the rich temporal semantics existing in sites with streaming content.

INTRODUCTION

Historically, technological advances that improve the collection of data have led to new domains for data mining. For example, the widespread use of the World Wide Web and the ability to collect and store web logs of user sessions has driven research interest in Web usage mining [CMS97, SCDT00]. An interesting problem in Web usage mining that has attracted the attention of several researchers is the discovery of traversal patterns of Web users [CPY98, NM00]. Mining path traversal patterns involves identifying how users access information of interest to them and travel from one object to another using the navigational facilities provided. Tracking user-browsing habits provides useful information for service providers and businesses, and ultimately should help to improve the effectiveness of the service provided. Previous research on identifying path-traversal patterns have been directed at traversals between relatively static objects (e.g., web pages). By "static", we mean information that can be regenerated by the user as required. Thus, dynamic web pages fall under our definition of static objects because the user may regenerate the dynamic web pages on each visit.

In this research, our focus is on navigational patterns in environments where the objects are continuously changing in time (i.e., streaming content). An example of such a system is cable-TV where the program sequence is continuously changing. The viewers of cable-TV are able to navigate from one object (channel/station) to another. However, if viewers navigate away from a station/channel and later return to that channel, the content being displayed may have changed. Thus, there is a strong temporal component in the systems studied in this research. The temporal component in our framework motivates new techniques to capture navigational patterns, as existing techniques in the literature do not take temporal semantic information into consideration. Our framework can be applied to identifying navigational patterns in any environment with streaming content. However, the discussions in this paper will be motivated by cable-TV viewing patterns. The choice of cable-TV viewing patterns is due to recent technological innovations that enable the collection of anonymous logs of viewing data through digital video recorders attached to cable-TV receivers. The logs are kept anonymous to protect the privacy of the viewers. This is similar to the ethical standards that have long been adopted in analyzing web and transaction logs. In the past, there had been very limited ways to collect data on the viewing patterns of cable subscribers. The advent of digital video recorders and the ability to track and report logs on the channels being viewed by subscribers (on a second by second basis) opens up several interesting areas for data mining. Digital video recorders are growing in popularity (with Tivo Inc reporting over 700,000 subscribers in the U.S.A in 2003, and a projection of over a million subscribers by the year-end) and a massive deployment is expected in the near future [Tiv03]. Digital video recorders keep track of the channels viewed through the cable receiver. The view logs are uploaded to the service provider daily. The challenge is to extract interesting patterns from all the view logs submitted to the service provider.

There are several interesting questions that can be addressed by analyzing the view logs. For example, an advertiser may be interested in knowing if more viewers stay tuned during the commercial breaks of prime-time programs than for regular programming. It may also be of interest to know the advertising slot that is most effective, i.e., is it more likely for an advert to be viewed if it is the first ad during the commercial break or if it has the last slot, or middle slot, *etc*. It may also be interesting to discover the percentage of viewers that return to a program once they tune off during a commercial break. Several other interesting patterns may also be discovered. In our framework, we propose a novel technique that categorizes the dynamic content of sites into distinct event sequences, and then explores the navigational patterns of users relative to the distinct event sequences.

The rest of this paper is organized as follows. In Section 2 we discuss related work. Section 3 presents the framework for the approach

proposed in this paper. Section 4 presents the proposed approach for discovering behavioral patterns in streaming environments. Conclusions and future work are discussed in Section 5.

RELATED WORK

There are strong similarities between the work discussed in this paper and web usage mining. Web usage mining is the application of data mining techniques to discover usage patterns from Web data [SCDT00]. The objects in our framework (e.g., channels) may be viewed as web pages. Also, a viewer can jump to any object/channel just like a web user may navigate to any URL. However, mining viewing patterns in our framework has a stricter temporal component. It is not sufficient to know the order in which the objects are viewed. There is a need to know the information content of the objects at the periods the viewer navigates to, or away from, the object. The work by Yang *et al.* [YWZ02] proposes an event prediction algorithm for web usage mining. Their approach is aimed at predicting when web accesses would occur. This is an extension of earlier works that only identify the order in which web accesses would occur. The problem they address is different from the problem addressed in this paper since we are interested in the information content of the objects at the times they are visited. Furthermore, the objects we study have streaming information content. Several other researchers have proposed techniques for identifying frequent path traversal patterns [e.g., CPY98, BL00, NM00, PHMZ00, HC02, *etc.*]. However, these approaches do not incorporate the temporal semantics we introduce in our framework. Tivo Inc. [Tiv03] has developed audience measurement tools that are able to report viewing statistics. However, their tools (just like tools for measuring web hits) do not explore navigational patterns of users.

FRAMEWORK

The general framework of the class of information systems covered in our work consists of independent sites with links connecting all sites. Unlike web pages that are grouped together into websites with internal navigational ordering, our framework is made up of stand-alone sites that are interlinked to each other. Using our example of a cable-TV system, each channel/station represents a site in our framework. A viewer is able to navigate from one channel to another either by following the ordering of the cable channels or by specifying the desired channel.

We define a *user session* as the complete set of activities by a user from the time the system is entered until departure. In our example, a *user session* starts when the user turns on the cable-TV and ends when the system is switched off. The system consists of sites with streaming content that can be divided up into categorical episodes. An episode is an event sequence that makes sense in its domain of application. In our example, we may identify three broad categories for the episodes, these are: programming, commercials, and shutdown. The programming category can be further divided into specific types of programs (e.g., movies, sitcom, sports, news, *etc*), and the commercials can be further divided into slots (i.e., first commercial slot, second, *etc*.). The categories may be abstracted further so that individual programs and commercials are identified. The choice of abstraction is determined by the data-mining analyst.

The information displayed by each site in the system can be broken into event sequences that fall into one of the episode categories defined. Thus, for each site, its streaming content (for 24 hours a day) can be categorized into definite episodes with the associated start and end times for each episode. Further, for each user of the system, the viewing patterns must be categorized for every user session during a given day.

IDENTIFYING NAVIGATIONAL PATTERNS

The first step in our framework is data preprocessing. The content/program information for each of the sites has to be preprocessed into a format suitable for mining. Similarly, the user logs have to be preprocessed. Each user-session is counted independently, i.e., one subscriber may have multiple user sessions in a day and each of the sessions would be independently considered in the framework. For example, given that time is represented in the 24 hour format *hh:mm:ss*, and that the numbers *4 – 62* represent channels/stations available to the user; a typical user log would specify the channels/stations the user viewed from the start of a session to its end. Table 1 gives an example of a typical user-log for one subscriber for a given day.

Table 1: An Example of a User Log

Time of Day	Channel Viewed
00:00:00	-
00:00:01	-
012:15:00	10
13:05:15	10
13:05:16	32
13:30:00	32
13:30:01	-
18:05:05	31
18:30:00	31
18:30:01	-
23:59:59	-

The logs record the viewing activity for each second of the day. The broken lines in Table 1 represent periods when there is no change in the channel viewed. From Table 1, we can identify two user sessions; the first starting at 12:15:00 and ending at 13:30:00, while the second session starts at 18:05:05 and ends at 18:30:00. Preprocessing the user log involves identifying all the user sessions, and breaking each session into time-brackets for the channels/stations viewed. The result of preprocessing the user log in Table 1 is shown below:

User session 1:
Channel 10: 12:15:00 - 13:05:15
Channel 32: 13:05:16 - 13:30:00
User session 2:
Channel 31: 18:05:05 - 18:30:00

The program content for each site (channel/station) in the system is also preprocessed. The analyst specifies categories for each program. For example, given that a station airs its programs between 08:00:00 and 18:30:00, and also given the complete program schedule of the station. If the categories specified are as follows: N – news; S – sitcom; C – commercials; and M – movies, a program episode can be represented by its category and an identifier. For example, the first news episode can be represented as N1, the second N2, *etc*. Similarly, the first sitcom may be represented as S1 and the second S2, *etc*. The identifiers are necessary if it is of interest to keep track of complete program episodes, since a program episode may be interleaved with another episode (e.g., several

Table 2: A partial listing of a sample program categorization for a channel / site

Program	Time Slot
N1	12:00:00 – 12:15:00
C1	12:15:01 – 12:17:00
C2	12:17:01 – 12:18:00
C3	12:18:01 – 12:20:00
N1	12:20:01 – 12:35:00
S1	12:35:01 – 12:55:00
C1	12:55:01 – 12:57:00
C2	12:57:01 – 12:58:00
C3	12:58:01 – 13:00:00
S1	13:00:01 – 13: 30:00
CO	13: 30:01 – 13: 33:00

commercial episodes may interleave a program episode). It may also be of interest to separate the program content into slots, (for example, the first commercial in a commercial break takes slot one - C1, the next commercial takes slot two - C2, *etc*. Commercials that are not embedded within other programs may also be separated into a category, e.g., CO in Table 2). Further, the analyst may choose to capture different segments of a program into separate categories. For example, it may be interesting to differentiate how users respond to the first segment of a program from how they respond to other segments, especially if they did not view the earlier segments. The salient point here is that categories may be defined for every program grouping of interest. Finally, the preprocessed program content for our example will be in a format similar to the one shown in Table 2. If the channel (or site) preprocessed in Table 2 is Channel 10 (from Table 1), then it is easy to extract the categories viewed during user session 1 (from Table 1).

Once the usage-logs have been transformed into user-sessions and the program schedules have been transformed into event categories, data mining procedures may then be performed on the processed data. The mining problem addressed in this work is formulated as follows:

- How do the users of the system navigate between sites in response to the contents displayed by the sites?

Details of the proposed techniques for discovering these behavioral and navigational patterns are discussed in the next section.

Discovering Event-Related Navigational Patterns

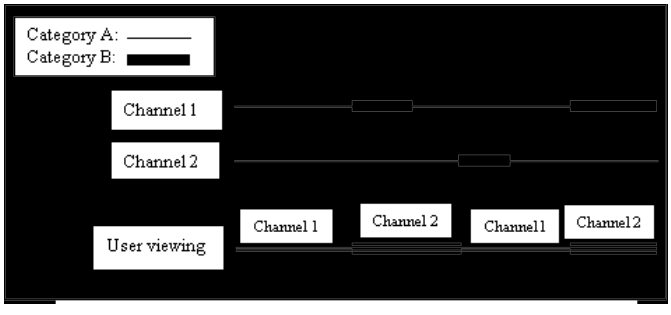
This section examines the problem of identifying frequent navigational patterns of users relative to specific event categories (or collections of categories). For example, it may be of interest to know if viewers navigate away from a channel/station immediately after a news event begins, or if they stay briefly before changing sites, or if they stay through the news event. It may also be of interest to discover if viewers navigate away at the commencement of a commercial break, and whether they return to the same program once they navigate away. Figure 1 shows an example user navigation between two sites (channels). From Figure 1, it is easy to see that the user navigates away from program content that belong to Category B irrespective of the site/channel being viewed.

The data mining problem here is thus to determine all behavioral navigation patterns, relative to program content, that are frequently exhibited by various users of the system. Given that the user logs and program schedules (content) have been preprocessed into user-sessions and categories respectively, the next step in the framework is to define the behavioral predicates that would capture the users' navigational patterns in response to dynamic content categories. The behavioral predicates chosen may include the following:

- Navigate away
- Stay through
- Stay briefly
- Return to the same program content (i.e., after navigating away)

The set of behavioral predicates considered in the framework depends on the interests of the analyst. Further, the quantitative time

Figure 1: An example of user navigation relative to event categories



units attached to some of the predicates (e.g., the definition of "briefly") are set by the analyst. Given a threshold confidence (e.g., 0.75), it is then possible to discover rules of the form: *"users of the system navigate away from program content in category B, with x confidence (where x is a user defined threshold)"*. It is also possible to capture navigational patterns of users in response to new program content in relation to the previous content they were viewing. An example of such a rule is: *"users of the system navigate away from content in category B given that they were previously viewing content in category C, with x confidence (where x is a user defined threshold"*. The details of the data mining process are given in the paragraphs that follow.

Recall that all the user-logs are preprocessed into independent user-sessions. Each user session details the channels viewed and the viewing times. By examining each user-session against the program categories airing at the sites/channels viewed, it is possible to extract the program categories viewed during each user-session and the behavioral navigational patterns of the user during the session being examined. Given that X is the set of categories over which a rule R is defined, then the set of active user-sessions with respect to rule R, A, is made up of user sessions with events in each of the categories in X. For example, given the rule *"users of the system navigate away from content in category B given that they were previously viewing content in category C, with x confidence (where x is a user defined threshold)"*, only user-sessions with content in both categories B and C are active with respect to this rule (i.e., $X = \{B, C\}$). The confidence of a rule is calculated as the ratio of the support count of user-sessions that satisfy the rule to the number of active user-sessions with respect to that rule. The contribution of a user session to the support count of a rule is weighted and may range from 0 to 1. For example, if a user-session encounters three instances of program content in category B, and if in two of the three instances the user navigated away from the program content, then the contribution of this user session to the rule *"users of the system navigate away from program content in category B, with x confidence (where x is a user defined threshold)"* will be 0.67 (i.e., 2/3). The support count for the rule is then obtained by summing the support contributions of each user session for that rule. The outline of the algorithm is presented below:

INPUT:
- A set U of user sessions (obtained from pre-processing all the user logs)
- A set P of categorized program schedules for all the sites in the system
- An empty set R of all rules defined on the categories in P

PROCESSING:

FOR all $u \in U$ DO
  Associate u with content categories by comparing its contents with relevant elements of P
  Identify the rule set present in u. If any rule found in u is not in R, add the rule to R.
  Increment the count of active user sessions for all rules on program categories found in u

  FOR all rules $r \in R$ on categories found in u DO
    Calculate the support contribution of u to r
    Add the support contribution to the total support for rule r
  END FOR
END FOR

FOR all $r \in R$ DO
  Confidence of r = total support of r / number of active user sessions for r
END FOR

OUTPUT:

Set of rules with confidence $\geq$ user-specified threshold

Given that U is the number of user sessions identified in the logs, and $\bar{R}$ is the average number of rules defined on program categories in the user sessions, then the algorithm has a time complexity of $O(U\bar{R})$. However, $\bar{R} << U$, thus, the algorithm runs in $O(U)$ time.

CONCLUSIONS AND FUTURE WORK

This paper motivates a new domain for data mining that involves discovering user navigational patterns in information systems that disseminate dynamically changing (or streaming) content. The ap-

proach proposed in this work can be extended in several ways. For example, it may be of interest to separate ad-hoc and non ad-hoc user sessions (i.e., some viewers may target certain programs while others may not). It may also be of interest to study the navigational behaviour of users relative to the time of day the viewing occurred, or navigational patterns relative to outlying content (e.g., a movie aired in a music channel). Several other extensions to the framework are also possible.

In the future, we hope to liaise with an industry service provider to test our approach on real-world user logs.

REFERENCES

[BL00] Borges, J., Levene, M., A Heuristic to Capture Longer User Web Navigation Patterns, in Proceedings of the 1st International Conference on Electronic Commerce and Web Technologies (EC-Web), pp. 155-164, 2000.

[CPY98] Chen, M-S., Park, J.S., Yu, P.S., Efficient Data Mining for Path Traversal Patterns, IEEE Transactions on Knowledge and Data Engineering, vol. 10, no. 2, pp. 209-221, 1998.

[CMS97] Cooley, R., Mobasher, B., Srivastava, J., Web Mining: Information and Pattern Discovery on the World Wide Web, in IEEE International Conference on Tools with Artificial Intelligence, pp. 558-567, Newport Beach, 1997.

[HC02] Heer, J., Chi, E.H., Mining the Structure of User Activity using Cluster Stability, in Proceedings of the Workshop on Web Analytics, SIAM Conference on Data Mining, 2002.

[NM00] Nanopoulos, A., Manolopoulos, Y., Finding Generalized Path Patterns for Web Log Data Mining, in J. Stuller et al. (Eds.): ADBIS-DASFAA 2000, LNCS 1884, pp. 215-228, Springer-Verlag, Berlin Heidelberg, 2000.

[PHMZ00] Pei, J., Han, J., Mortazavi-asl, B., Zhu, H., Mining Access Patterns Efficiently from Web Logs, in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00), pp. 396-407, 2000.

[SCDT00] Srivastava, J., Cooley, R., Deshpande, M., Tan, P-N., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, vol. 1, no. 2, pp. 12 – 23, ACM SIGKDD, Jan 2000.

[Tiv03] Tivo Inc, http://www.tivo.com

[YWZ02] Yang, Q., Wang, H., Zhang, W., Web-log Mining for Quantitative Temporal-Event Prediction, IEEE Computational Intelligence Bulletin, vol. 1, no. 1, pp. 10-18, December 2002.