On Business Activity Modeling using Grammars

Savitha Srinivasan, Arnon Amir, Prasad Deshpande, and Vladimir Zbarsky
IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120
1-408-927-1430
savitha@almaden.ibm.com

ABSTRACT

Web based applications offer a mainstream channel for businesses to manage their activities. We model such business activity in a grammar-based framework. The Backus Naur form notation is used to represent the syntax of a regular grammar corresponding to Web log patterns of interest. Then, a deterministic finite state machine is used to parse Web logs against the grammar. Detected tasks are associated with metadata such as time taken to perform the activity, and aggregated along relevant corporate dimensions.

Categories and Subject Descriptors

H.4.1 [Office Automation]: Desktop Publishing, Groupware, Time Management, Workflow Management,.

General Terms: Design, Experimentation, Measurement.

Keywords: Web log analysis, data mining.

1. INTRODUCTION

Web based applications are used within and across enterprises to manage and run an increasing number of business processes like contract management, procurement, asset management, directory services, benefits administration etc. Web usage mining has emerged as a means of usage characterization, web site performance improvement, personalization, adaptive site modification and market intelligence for such applications.

Path analysis is usually the basis of many web usage mining tools – its goal is to help understand visitor's navigation of a website. Path analysis can be simply defined as the list of pages, in order, that a visitor traverses in one visit. While this provides the exact, complete path for each visitor, it may not provide useful insights in terms of visitor behaviors. Therefore, various modifications of path analysis have been proposed such as a focused path analysis, where ultimately visits are classified as "success" or "failure" against certain business objectives of making a sale. Free-form discovery of popular visitor paths is not necessarily insightful in evaluating the efficacy of web applications that support an enterprise business process. Instead, the web site is typically designed with a set of features to meet a set of requirements of the process it serves. Metrics that are relevant to evaluating such Web sites are *business activity-oriented* as in: how effective was the site in getting the task accomplished, how long it takes to complete a specific activity, what are the trends over time across different user populations in the corporation etc.

In this context, we present a grammar-based framework to model and detect business process activity from web log patterns. The framework has two components to it, a declarative stage and a processing stage. The Backus Naur notation is used to represent a regular grammar corresponding to the patterns of interest. A deterministic finite state machine is then used to parse web logs against the grammar that encodes the activity of interest. After detection, tasks are associated with the actual users and their organizational and geographical association. The labeled tasks are then aggregated along relevant corporate dimensions such as geography and division, and is analyzed along a period of time. This methodology has the following advantages over traditional URL and path analysis:

1. A grammar-based framework, flexible enough to define many different types of business activities
2. Deeper analysis of user actions, using meaningful units of user interaction, i.e. tasks and business cost/value metrics
3. Fine grained metrics on a per activity basis
4. Measure of the effectiveness of the web site based on the *time* it takes to perform an activity, activity frequency, etc.

2. RELATED WORK

All material on each page should fit within a rectangle of 18 x Pattern discovery from Web logs draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition [1][2][3][5]. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as age views, viewing time and length of a navigational path. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site [4]. The discovery of typical frequent usage patterns seems to be exactly what sequence miners [1] are built for. However, in corporate services, regular and rare sequences might be of equal importance, depending on the activities they represent. Only the designer of the service is aware of the larger tasks within which all detected patterns must be analyzed and evaluated.

In this paper, we present an approach to address this problem: i.e. the ability to define the patterns that represent the entire activity of interest using a formal grammar. The grammar serves to extract knowledge from data, where the knowledge has been expressed in relatively abstract terms.

3. GRAMMAR_BASED ACTIVITY

We use a grammar to define relevant activity supported by a web site. In this context, our alphabet symbols correspond to URLs and the language "words" represent activities as defined by the regular language. Regular languages are commonly used to define search patterns and the lexical structure of programming languages. For convenience, we define our activity detection regular grammar using the standard Backus-Naur form (BNF). The overall process for the activity detection is as follows:

1. Preprocess daily Web logs to extract URLs of interest
2. Extract URL logs per user into separate session files
3. Tokenize the session files

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

4. Detect patterns using the activity grammar
5. Compute time taken to perform activity
6. Associate each activity with its metadata, such as task duration and user information
7. Analyze and plot trends over time

We use this approach to define a grammar to detect the activities performed in an enhanced employee directory for the IBM intranet. It allows employees to create and modify a detailed personal profile, listing of his/her technical skills, work experience, areas of expertise, teams and assigned projects. Leveraging this multitude of information enables other employees to search for and locate an expert with desired knowledge and skills quickly and efficiently, bridging across multiple organizations and geographic locations, through the report-to structure, etc. Users may also maintain their personal directory of contacts using the My BluePages option.

The activities supported by this directory service are first defined using a grammar. Figure 1 lists a few of the 50 rules of a grammar for detecting nine different tasks. The alphabet for our grammar is represented by tokens *A* through *X*. Any feasible sequence of symbols build a word in the language - a task in our case. Hence the language defines all the activities we are interested in detecting. Broken sequences, not accepted by the grammar language, represent uncompleted tasks.

```
<Tasks> = <SearchTasks> | <BrowseTasks> | <ProfileTasks> .
<SearchTasks> =  <SimpleContactLookup> | <SimpleSearchbyContactI> |
   <SimpleSearchbyExpertise> | <AdvancedSearchbyContact> |
   <AdvancedSearchbyExpertise>
<BrowseTasks> = <LearnAbtAPerson> | <BrowseReport2Chain> .
<ProfileTasks> = <UpdateProfileUsingForm>  | <AddtoMyBluePages> .
<SearchTaskBeginners> = <A> | <T> | <U> .
<NonTaskBeginners> =  <F> | <G> | <H> | <I> | <J> | <K> | <L> | <N> | <O> | <P> | <Q>
| <R> | <S> |  <V> | <X> .
<SimpleContactLookup> = <A> .
<SimpleSearchbyContact> = <A> <F> .
<SimpleSearchbyExpertise> = <A> <BrowseMultipleResults> <BrowseMultipleResults>+
.<I> = "GET /pilot/bluepages/searchByName\.wss\?\.\*tab\=00_experience_skills\.wss" .
<J> = "GET /pilot/bluepages/searchByName\.wss\?\.\*tab\=00_project_team\.wss" .
<O> = "POST /pilot/bluepages/editExperience\.wss" .
<P> = "POST /pilot/bluepages/editProject\.wss" .
```

Figure 1. Example for some of the ~50 rules used to define nine activities in corporate directory service application

A specific implementation of the FSM is based on regular expressions. Modification to the rules is much easier to handle than direct editing of these regular expressions. The use of grammatical rules allows intuitive explanation, and ensures consistency and correctness of the automatically derived expressions. The result of detecting the task supported by the directory service application and tracking them temporally is shown in Figure 2. It shows the counts for Simple Expertise Search (red) being low, but the task times are high. This shows that time spent per Simple Expertise Search is quite high compared to other tasks. It takes a long time for users to find people with a particular expertise since they have to browse their profiles and assess if it is the person they are seeking.

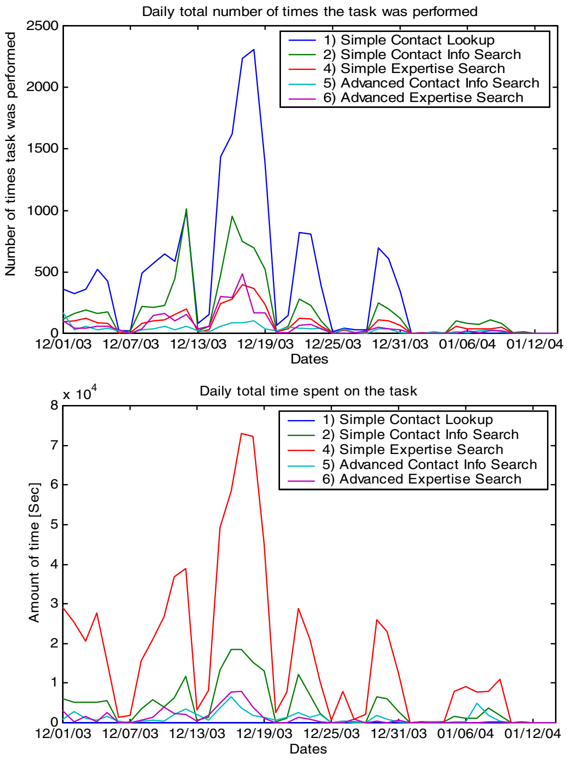

Figure 2. Daily totals of tasks performed and the time they consumed. Notice a peak before the year end and a decline during weekends and the first week of the new year.

4. CONCLUSION

We have presented a grammar based framework to model activities served by web applications. Gathering statistics on an activity basis rather than URL basis provides for a deeper analysis and more meaningful numbers. We applied task analysis on a corporate intranet portal application over a 45 days period to get some valuable insights into the usage pattern for different users. Comparison with standard web log analysis tools show that task analysis can counts patterns of interest rather than simple URLs thus providing more useful metrics.

5. REFERENCES

[1] Agrawal, R. and Srikant, R. Mining sequential patterns. In Proceedings of the International Conference on Data Engineering (Taipei, Taiwan, Mar. 1995).

[2] M.S. Chen, J. Hart, and P.S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering,* 8(6):866-883, 1996.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Proc. ACM KDD,* 1994.

[4] Surfaid analytics, http://surfald.dfw.ibm.com.

[5] T. Yah, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Fifth International World Wide Web Conference,* Paris, France, 1996.