Sequential Pattern Mining in Web Log Data using Generalize Sequential Pattern Mining(GSP) Algorithm

Khin Su Hlaing, Ei Ei Moe Tun
Computer University, Magway
suhlaing8708@gmail.com, eieimoetun@gmail.com

Abstract

The rising popularity of electronic commerce makes data mining an indispensable technology for several applications, especially online business competitiveness. The World Wide Web provides abundant raw data in the form of web access logs. However, without data mining techniques, it is difficult to make any sense out of such massive data. In this paper focus on the mining of web access log, commonly known as Web usage mining. Frequent pattern mining is a heavily researched area in the field of data mining with wide range of applications. One of them is to use frequent pattern discovery methods in Web log data. Discovering hidden information from Web log data is called Web usage mining. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users. This can be used for advertising purposes, for creating dynamic user profiles etc.In this paper, ***GSP*** *algorithm is used for sequential pattern mining in web log data.*

1. Introduction

Information dominates the world more than anytime before. In the information world, internet is the roads and the highways, the content providers are the road workers, and the visitors are the drivers. As in the real world, there can be traffic jams, wrong signs, blind alleys, and so on. The content providers, as the road workers, need information about their users to make possible Web site adjustments. However, the content providers have a big advantage in comparison with the road workers. Web logs store every motion on the provider's Web site. So the providers need only a tool to analyze these logs. This tool is called Web Usage Mining. Web Usage Mining is a part of Web Mining which is knowledge discovery techniques focused on Web analysis. Web Usage Mining methods analyze moves on Web sites and give the content providers welcomed feedback. There are various application domains like web site design, business and marketing decisions support, personalization, usability studies, etc. For example, e-shop providers can alter the recommended merchandise or restructure the Web site to be more users' friendly [1].

The rate of group of the World Wide Web (Web) may be slowly down, but online library center researchers concluded that the web would continue to grow rapidly in their annual review of the web. The continue popularity of electronic commerce through its prediction that global on-line trade would expand. So, to stay competitive and profitable in a fast placed environment like the Web, companies must be able to extract knowledge from their web access logs, web transaction logs and web user profiles to ensure the success of Customer Relationship Management (CRM) [2].

Web Usage Mining (WUM) is the extraction of meaningful user patterns from web server access logs using data mining techniques. The term log file refers to a web server access log.WUM is fast gaining importance because of the wide availability of log files as well as its applicability in CRM. In addition, it has diversified applications such as web personalization; web structuring, marketing, user profiling, caching and perfecting [2].

In this system, the sequential pattern mining in web log data is presented to process the pattern discovery. In this paper, it is used the GSP algorithm for searching the frequent pattern in web log data. The frequent pattern in Web log data is to obtain information about the navigational behavior of the users.

The rest of the paper is organized as follows. In Section 2, the background theory of the presented system is described. Section 3 shows the overview of the system. In section 4 some evaluation analysis can be described. In this section, sample evaluation is shown with example logs data set. The presented sequential pattern mining in web log data is concluded in Section 5.

2. Background Theory

2.1 Web Mining

Web servers use the log files to record an entry for every single access they get. As the complexity of the web site or application increases, simple statistics give no meaningful hints on how the web site is being used. Moreover, the log files of popular web sites may grow of several hundreds of megabytes per day, making analysis tasks award. Web mining refers to the application of such techniques to web data repositories, to enhance the analytical capabilities of the known statistical tools [3].

Web mining involves three tasks: (1) structure mining, (2) content mining, and (3) usage mining. Web structure mining refers to the process of information extraction from the topology of the Web and aims at extracting data which describes the organization of the content.

Web content mining is the process of extracting useful information from the web sites and the pages they are composed of. One of the main challenges is the definition of what the web content is. Web content is composed of multiple data types: text, images, audio, video, metadata and hyperlinks and multimedia data mining has become a specific instance of web content mining [3].

Web usage mining is devoted to the investigation on how people use web pages they access and on recognizing their navigational behavior. It involves automatic discovery of user access patterns from one or more Web servers or clients. Web usage mining is based on the analysis of secondary data describing interactions between users and the Web. Web usage data include data recorded in Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions and transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data derived from the above interactions. As it can be easily understood, boundaries between these three categories are blurred. Our main concern is with user access behavior, in this paper we mainly concentrate on the usage mining as a mean to track the behavioral patterns of users surfing either a web site or a page.

2.2 Web Usage Analysis

Web usage analysis relates to the development of techniques for discovering and/or predicting the behavior of a user interacting with the web. The data to be analyzed are (see figure 1) data logged by the user client (e.g. web browser) or server side (web server, proxy server, application server). Data logged at different locations represent different navigation patterns: client sides data describe, in general, single user multi-site navigation, server side logs describe multi-user single-site interaction and proxy server logs describe multi-user multi-site interaction.

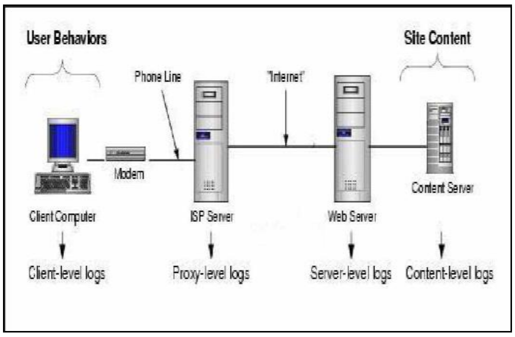


Figure 1.Various data sources

Server side logged data include client IP address (machine originating the request, maybe a proxy), user ID (if authentication is needed), time/date, request (URI, method and protocol), status (action performed by the server), bytes transferred, referrer and user agent (operative system and browser used by the client). Since this information is incomplete (e.g., not showing hidden request parameters automatically downloaded using the POST command) and not entirely reliable, the information should be integrated using packet snifters and, where available, application server log files. Client side data collection has been implemented using remote agents (Java applets, HTML embedding Java script code) or ad-hoc browsers. Recently, especially tailored browsers have been used to implicitly and automatically capture user’s interests in a particular page and provide an overall rating for that page [3]. Proxy server data describe, on one side, access to cached pages and, on the other, access to sites from actual clients seen as a single anonymous entity from the web server. Three different phases, as shown in figure 2, are identified: pre-processing, pattern discovery and pattern analysis.

2.3. Pre-Processing

In this phase, abstraction data may be built representing, as examples, users (single actor using a browser to access files served by a web server), page views (the set of files served to the browser in response to a user action such as a mouse click), click streams (a sequential series of page views requests), user sessions (click stream for a single user across the Web), and server sessions (set of page views for a user session on a single web site). A particular stress is given to data preparation and preprocessing. As stated earlier, data may be incomplete (especially when client side logs are unavailable) leading to difficulties in user identification and difficulty to detect the user session termination (a 15 and 30 minute default time is normally assumed for session termination). Depending on the data actually available for the analysis, typically known problems in user behavior reconstruction include: multiple server sessions associated to a single IP client address (as in the case of users accessing a site through a proxy); multiple IP addresses associated to multiple server sessions (the so called mega-proxy problem); user accessing the web from multiple machines; and a user using multiple agents to access the Web. Assuming that the user has been identified, the associated click-stream has to be divided into user sessions. As expected, the first relevant problem is the identification of the session termination. Other relevant issues are the need to access application and content server information and the integration with proxy server logged information relative to cached resources access. Related to the usage preprocessing is the content preprocessing. Page views might be classified or clustered depending on their intended use and the results of this process be used to limit discovered usage patterns [3].

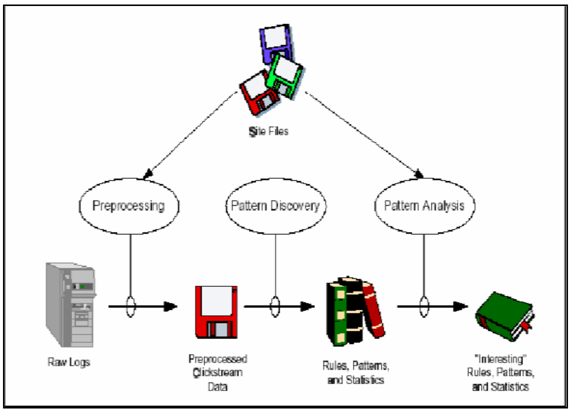

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 2. Phases of web usage mining

2.4 Pattern Discovery

Techniques adopted for this phase, strictly depend on the aim of the analysis. Methods available draw upon several fields such as statistics, data mining, machine learning, and pattern recognition. Statistical analysis is in general applied to discover information such as most accessed pages or average length of a navigation path through a web site [3]. Use of association rules to find correlation between pages most often referenced together in a server session with a support value exceeding a given threshold.

The results may find application in developing marketing strategies for e-business sites as well as for providing hints for restructuring a web site. Clustering is used to group items having similar characteristics. In this case clustering may be used to group users exhibiting similar navigation behavior (usage clusters) or groups of pages having a related content (page clusters). In the first case, the information is again relevant to marketing scopes while, in the second one, it might be used by search engines. Classification techniques are often used to associate navigation behaviors to groups of users (or profiles). On the application of sequential pattern discovery techniques to identify set of items followed by further items in a time ordered sequence. This is relevant for marketing purposes, i.e. for placing advertisements along the navigation path of certain users.

Dependency modeling is also used with the goal of developing a model to represent significant dependencies among various variables in the web (for instance, modeling the stages a user undergoes during a visit to an on-line store). This is useful not only in predicting the user behavior but also in predicting web resource consumption. With Semantic Web, pattern discovery benefits from semantics included into web pages. This enables mapping Http requests to meaningful units of application events. Ontology, Resource Description Framework (RDF) repository and user profile can be updated with new information.

In addition to the relationships between concepts, ontology also contains logical axioms that enable inferring new knowledge. The meaning that is constituted by the set of web pages accessed can be captured and taken into account. Hence, usage pattern can reveal semantic relationships that can help in learning the ontology itself or ontology instances [3].

2.5 Pattern Analysis

The last phase deals with pattern analysis. The aim is to filter out irrelevant information and extract only interesting information from the output of pattern discovery phase. One of the approaches used in analyzing patterns is the use of visualization techniques, such as graphing patterns, charts, diagrams, coloring schemes, and coordinated views. The goal is to highlight overall patterns and trends in the data [3]. Relational databases have also been used for this phase.

3. Overview of the System

This system is presented for discovering hidden information from large amount of Web log data collected by Internet Information Servers and introduced the process of web log mining. This system used **GSP** algorithm for analyzing the web log data in order to obtain useful information about the user's navigation behavior. Figure 3 shows the system flow diagram of presented system.

3.1. System Flow Diagram of the presented System

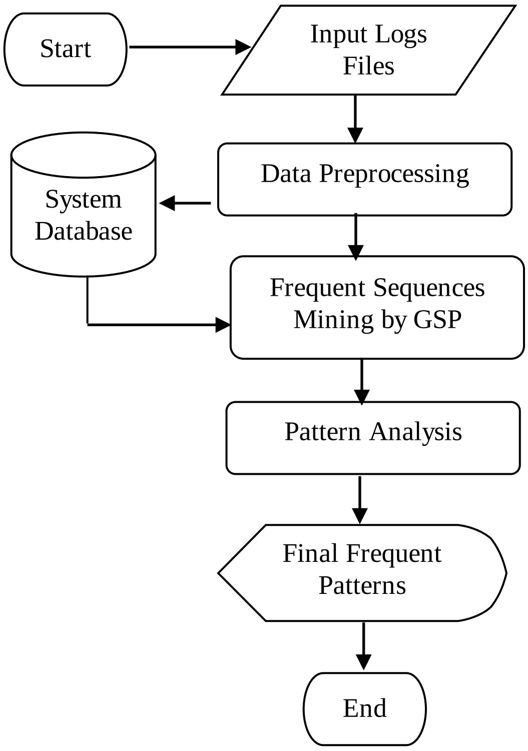


Figure3. System flow diagram

3.2. Algorithm for GSP

In this section shows the GSP algorithm for analyzing the web log data about the user's navigation behavior of the web sever. This algorithm is very similar to the Apriori algorithm.

Algorithm GSP(*S*)

```
C1← init-pass(S); // the first pass over S
F1  ← { {f}f∊ C1 ,f.count /n  ≥ minsup };//n is
       the number of sequence in S
for (k=2;Fk-1≠Φ;k++) do // subsequence pass
       over S
  Ck   ← candidate-gen-SPM(Fk-1);
for each data sequence  s ∊ S do // scan the data
       once
for each candidate c ∊ Ck do
    if c is contained in s then
c. count++; // increment the support  count
end
end
   F k   ←  {c ∊ Ck / c.count/n ≥ min –sup}
   end
    return  Ck  Fk;
```

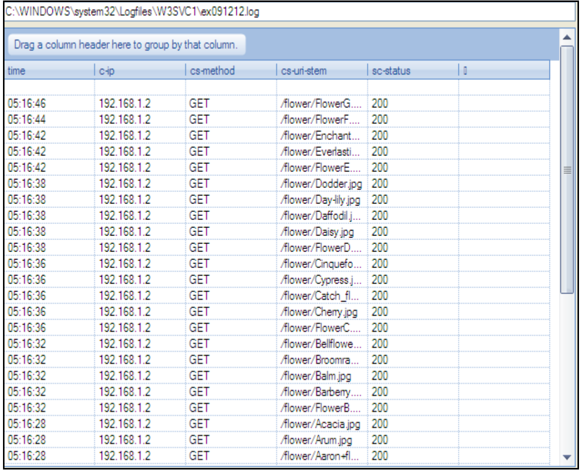
Candidate-gen-SPM (F_k-1)

1. Join Step: Candidate sequences are generated by joining Fk-1 with Fk-1.

2. Prune Step: A candidate sequence is pruned if any one of its (k-1) subsequence is infrequent (without minimum support)

4. Evaluation Analysis

In this section, we show evaluation analysis with sample web log data set. There are three sever site log format. They are W3C, IIS and NCSA formats. In this system, use the W3C log data format to find the frequent patterns by using GSP algorithm. Sample W3C log data format are shown in figure 4.



C:\WINDOWS\system32\Logfiles\W3SVC1\ex091212.log

time	c-ip	cs-method	cs-uri-stem	sc-status	
05:16:46	192.168.1.2	GET	/flower/FlowerG....	200	
05:16:44	192.168.1.2	GET	/flower/FlowerF....	200	
05:16:42	192.168.1.2	GET	/flower/Enchant...	200	
05:16:42	192.168.1.2	GET	/flower/Everlasti...	200	
05:16:42	192.168.1.2	GET	/flower/FlowerE....	200	
05:16:38	192.168.1.2	GET	/flower/Dodder.jpg	200	
05:16:38	192.168.1.2	GET	/flower/Day-lily.jpg	200	
05:16:38	192.168.1.2	GET	/flower/Daffodil.j...	200	
05:16:38	192.168.1.2	GET	/flower/Daisy.jpg	200	
05:16:38	192.168.1.2	GET	/flower/FlowerD....	200	
05:16:36	192.168.1.2	GET	/flower/Cinquefo...	200	
05:16:36	192.168.1.2	GET	/flower/Cypress.j...	200	
05:16:36	192.168.1.2	GET	/flower/Catch_fl...	200	
05:16:36	192.168.1.2	GET	/flower/Cherry.jpg	200	
05:16:36	192.168.1.2	GET	/flower/FlowerC....	200	
05:16:32	192.168.1.2	GET	/flower/Bellflowe...	200	
05:16:32	192.168.1.2	GET	/flower/Broomra...	200	
05:16:32	192.168.1.2	GET	/flower/Balm.jpg	200	
05:16:32	192.168.1.2	GET	/flower/Barberry....	200	
05:16:32	192.168.1.2	GET	/flower/FlowerB....	200	
05:16:28	192.168.1.2	GET	/flower/Acacia.jpg	200	
05:16:28	192.168.1.2	GET	/flower/Arum.jpg	200	
05:16:28	192.168.1.2	GET	/flower/Aaron+fl...	200	

Figure 4. Sample W3C log format

4.1 Preprocessing the raw logs files

The above log files are changed in pattern preprocessing for pattern mining in web log data. These changed patterns are shown in table 1.In this table, Client IP and pages are shown. These two fields are required for frequent pattern mining. So the rest fields of W3C log files in figure 4 are cleaned at pattern preprocessing phase.

Table 1. Sequential patterns table for frequent sequent mining

Client IP	Page View by Client (Pages)
192.168.1.2	<Page A, Page B>
192.168.1.3	<Page C, Page D>, <Page A>, <Page E,Page F,Page G>
192.168.1.4	<Page A,Page H,Page G>
192.168.1.5	<PageA>,<PageE,PageG>,<PageB>
192.168.1.7	<Page B>

4.2 Candidate and Frequent Sequence Generation by GSP

A preprocessed log files consist of a file where each record represents information about the navigational behavior of the users in data stored on the Web. A file typically includes a unique client ip (Client IP), and sequential patterns of navigational behavior of the users (Pages). After preprocessed the log files, the process of frequent pattern miming is preceded. The following example shows the frequent pattern mining by using GSP.

Finding Length -1 sequence, sup_Count=25%

C1

Itemsets	Sup-count
Page A	80%
Page B	60%
Page E	40%
Page G	60%

L1

Itemsets	Sup-count
<Page A>	80%
<Page B>	60%
<Page C>	20%
<Page D>	20%
<Page E>	40%
<Page F>	20%
<Page G>	60%
<Page H>	60%

Generating Length_2 candidates

This is Join Step.

	Page A	Page B	Page E	Page G
Page A	{A,A}	{A,B}	{A,E}	{A,G}
Page B	{B,A}	{B,B}	{B,E}	{B,G}
Page E	{E,A}	{E,B}	{E,E}	{E,G}
Page G	{G,A}	{G,B}	{G,E}	{G,G}

This is Prune Step.

	Page A	Page B	Page E	Page G
Page A		{A,B}	{A,E}	{A,G}
Page B			{B,E}	{B,G}
Page E				{E,G}
Page G				

The set of Length -1 sequential patterns generates the set of 4*4+ 4*3/2=22 candidate sequences.

C2

Itemsets	Sup-count
{A,A}	0%
{A,B}	40%
{A,E}	40%
{A,G}	40%
{B,E}	0%
{B,G}	0%
{E,A}	0%
{E,B}	20%
{E,E}	0%
{E,G}	0%
{A,G}	20%
{B,E}	0%
{B,G}	0%
{E,G}	40%

L2

Itemsets	Sup-count
<{A}{B}>	40%
<{A}{E}>	40%
<{A}{G}>	40%
<{E,G}>	40%

Generating Length -3 candidate sequences

Join Step

	<{A}{B}>	<{A}{E}>	<{A}{G}>	<{E,G}>
{A}	<{A}{A}{B}>	<{A}{A}{E}>	<{A}{A}{G}>	<{A}{E,G}>
{B}	<{B}{A}{B}>	<{B}{A}{E}>	<{B}{A}{G}>	<{B}{E,G}>
{E}	<{E}{A}{B}>	<{E}{A}{E}>	<{E}{A}{G}>	<{E}{E,G}>
{G}	<{G}{A}{B}>	<{G}{A}{E}>	<{G}{A}{G}>	<{G}{E,G}>

Prune Step

	<{A}{B}>	<{A}{E}>	<{A}{G}>	<{E,G}>
{A}		<{A}{A}{E}>	<{A}{A}{G}>	<{A}{E,G}>
{B}			<{B}{A}{G}>	<{B}{E,G}>
{E}				<{E}{E,G}>
{G}				

C3

Item sets	Sup-Count
<{A}{A}{B}>	0%
<{A}{A}{E}>	0%
<{A}{A}{G}>	0%
<{A}{E,G}>	40%
<{B}{A}{E}>	0%
<{B}{B}{G}>	0%
<{B}{E}{G}>	0%
<{B}{E,G}>	0%
<{E}{A}{B}>	0%
<{E}{A}{E}>	0%
<{E}{A}{G}>	0%
<{E}{E,G}>	0%
<{G}{A}{B}>	0%
<{G}{A}{E}>	0%
<{G}{A}{G}>	0%
<{G}{E,G}>	0%
<{A,A,E}>	0%
<{A,A,G}>	0%
<{A,E,G}>	0%
<{B,A,G}>	0%
<{B,E,G}>	0%
<{E,E,G}>	0%

L3

Item sets	Sup-Count
<{A}{E,G}>	40%

Table 2. Final output sequences patterns

1-sequences	<{A}>,<{B}>,<{E}>,<{G}>
2-sequences	<{A}{B}>,<{A}{G}>,<{A}{E}>,<{E,G}>
3-sequences	<{A}{E,G}>

In table 2, the client view only one sequential pattern is < {Page A}, {Page B},{Page E}, {Page G}>.

Two sequential patterns are< {A} {B}>, < {A} {G}>, < {A} {E}>, < {E, G}>.Three sequential pattern is < {A} {E, G}>.

5. Conclusion and Further Extension

5.1. Conclusion

This system deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of the system is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior.

5.2 Further Extensions

The area of Web usage mining is still in its preliminary stages, as a consequence, related issues need further exploration and research. Some important points that need special attention include: (1) as the content served by a web server becomes more and more dynamic, the need of integration in the analysis process of the data coming from sources different from Web servers become mandatory. In particular, as multiple applicative services are able (and have to) to identify single user sessions, the user session reconstruction might become relatively straightforward, (2) development of advanced tools to deal with highly structured content such as XML, which requires more than just text mining, (3) development of techniques which would improve automatic Web navigation and page pre-fetching on the behalf of the users, (4) use of overall user interests discovered to design future theme based search engines as opposed to current key word based engines, and (5) establishment of more robust procedures for reconstructing user behavior patterns while visiting Web sites and pages.

This system will extend by using other mining methods. This system uses the server site log format for mining. So, other user can use the other lot format such as client site and proxy site log format for mining frequent pattern in web log data.

6. References

[1] Luk_a_s _Cenovsk_y "Web Usage Mining on is.muni.cz" Master's Thesis, 2003

[2] A. Scime "Web Mining: Application and Techniques" State University of New York College at Brockport .USA, 2004

[3] Haider Ramadhan, Muna Haten, Zuhoor Al- Khaljr "A Classification of Techniques for Web Usage Analysis"Computer Science Department Sultan Qaboos University PO Box 36 Muscat 123