

Paper 076-31

## Data Mining Techniques to Analyze a Library Database

Chakib Battioui, University Of Louisville, Louisville, KY

### ABSTRACT

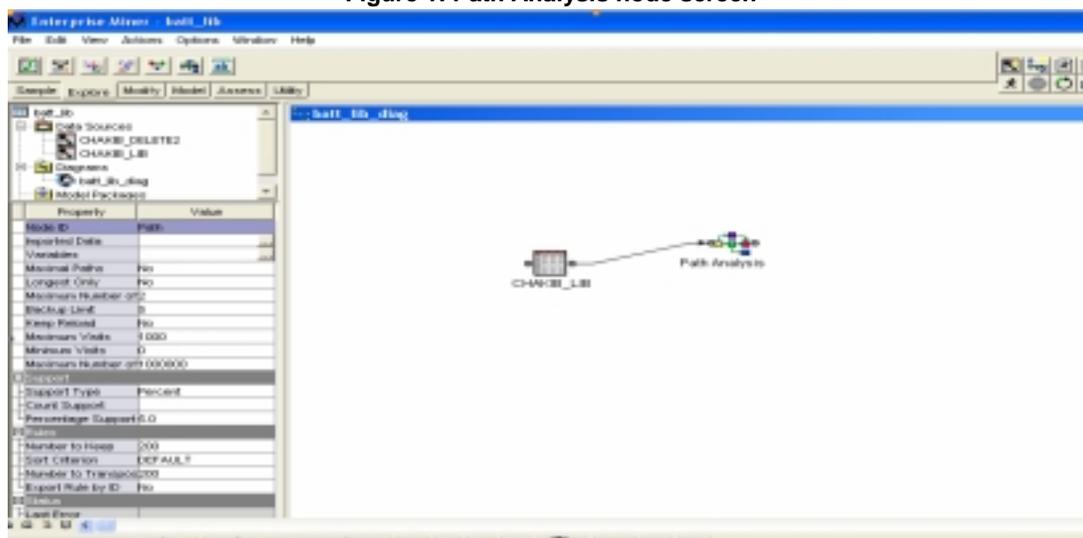
The paper focuses on using data-mining techniques to investigate a library database of URL information regarding access to electronic sources. The specific objective is to determine the different rules that describe associations between sets of items in the Ekstrom Library database at the University of Louisville, by using SAS® Enterprise Miner™ software. We obtained the data from the Ekstrom Library, which regularly saves the information on a monthly basis. Path Analysis was used to deal with the URL links. The results show how the Library can identify the patrons who log in to their web sites the most often, and where they log in from. Also, the results provide the Ekstrom Library with a count of the number of times each link occurred in the dataset, and a list of association rules that can be used for better organization of the library web site. It is a way of determining whether the databases purchased are of value to library patrons.

### INTRODUCTION

The main focus of this paper is to use data mining techniques to investigate a library database of URL information concerning access to electronic sources. The specific objective is to determine the different rules that describe associations between sets of items within the Ekstrom Library Database (University of Louisville) using SAS Enterprise Miner Software. The specific method developed to deal with URL links is Path Analysis. Path analysis is used to analyze preprocessed web log data. It allows the user to determine the paths that visitors take as they navigate through a web site. Also, it performs association analysis between web links and allows the user to extract sequential association rules among large sets of URL links.

The goal of association rules is to find interesting association and correlation relationships among large sets of data items where the presence of one set of items in a transaction implies the presence of other items. The rule is defined by  $A \rightarrow B$  where A is the set of antecedent items and B is the consequent item. An association rule has two numbers that express the degree of uncertainty about the rule: the Support and Confidence. The support is the number of transactions that include all items in the antecedent A and consequent B divided by the total number of transactions. It is the probability that a transaction contains both A and B. The Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent. It is the Conditional Probability that if a transaction contains A, it also contains B. Association rules help organizations to develop effective marketing strategies and arrange their websites for better use. The Enterprise Miner screen is given in Figure 1 with the Path Analysis node connected to the dataset used in this study. The left-hand side window shows the name of datasets used, the properties of Path Analysis node and their values. The values can be changed easily from this window.

Figure 1: Path Analysis node screen



We have obtained data from Ekstrom Library that contain URL information concerning access to electronic sources. The dataset was very large so we had to choose only one file (one month) out of these data. The file chosen contained URL information running from October 1, 2003 until October 31, 2003 with 482,766 observations.

#### METHOD:

SAS software 9.1 was used to import the data and create 4 variables: ip\_address, user\_id, Date\_time and URL.. The first ten observations of our dataset are given in table 1.

**Table 1: First ten observations of the dataset used**

	ip_address	User_ID	Date_Time	URL
1		-	[01/Oct/2003:00:00:08	POST http://echo.louisville.edu:2048/lo HTTP/1.0
2		-	[01/Oct/2003:00:00:09	GET http://echo.louisville.edu:2048/co HTTP/1.0
3		-	[01/Oct/2003:00:00:09	GET http://echo.louisville.edu:2048/m HTTP/1.0
4		-	[01/Oct/2003:00:00:12	GET http://echo.louisville.edu:2048/lo HTTP/1.0
5		-	[01/Oct/2003:00:00:13	GET http://library.louisville.edu:80/ HTTP/1.1
6		-	[01/Oct/2003:00:00:18	GET http://library.louisville.edu:80/main HTTP/1.1
7		-	[01/Oct/2003:00:00:22	GET http://library.louisville.edu:80/rese HTTP/1.1
8		-	[01/Oct/2003:00:00:26	GET http://database.library.louisville.ed HTTP/1.1
9		-	[01/Oct/2003:00:00:27	GET http://search.epnet.com:80/login. HTTP/1.1
10		-	[01/Oct/2003:00:00:28	GET http://bill.epnet.com:80/start.asp? HTTP/1.0

Since the goal of Path Analysis is to determine the paths that visitors take as they navigate through a website, we had to pre-process the dataset by keeping only the off-campus users with user id information. Also, in order to create sequential information, we had to adjust the dataset by adding a sequence number variable so that it can be used in path analysis. SAS code was written to create a sequence number.

```

/* Sort the data on USER_ID */
proc sort data= my lib.librarysubset out=work.ls;
by user_id;
run;
/* Create a SAS Datetime value from the character DATETIME field */
data ls2;
Set ls;
date_time=compress(date_time,'[ / ');
sas_date_time=input(date_time,datetime.);
format sas_date_time datetime18.;
run;
/* Sort the data on both USER_ID and SAS_DATE_TIME to prepare for sequencing */
proc sort data=ls2;
by user_id sas_date_time;
run;
/* Assign the SEQUENCE_NUM */
data ls3;
set ls2;
by user_id sas_date_time;
if first.user_id then sequence_num=0;
sequence_num+1;
run;

```

## RESULTS

SAS Enterprise Miner 5.2 was used to explore and analyze the data using the StatExplore and Path Analysis nodes. The number of URLs was increased in the path analysis from the default of 100,000 to 1 million because of the size of our dataset.

In order to obtain an effective and interesting result, we had to work with a member of the library staff with domain knowledge who can interpret the results obtained and suggest what links need to be filtered from the dataset. His first suggestion was to remove all the images and style sheet pages as they are a routine part of the initial access. These URLs were filtered from the original dataset. When re-running the path analysis, results were improved. The next suggestion was to remove the login pages and the users who could not be identified from our dataset since our interest was to investigate the path that patrons take when they log in to the library website. We had to filter the data on many steps. Every time we remove the unnecessary websites, we had better results. This process was done thru three main steps:

### STEP 1:

The original dataset was analyzed using Path analysis. The items report shows the number of times each item occurred in the dataset. The first ten items of the items report that have the highest count are given in table 2. Notice that the item with the highest count is <http://echo.louisville.edu:2048/menu/http/1.1>. This is the first page that a user will view when he logs in to the Library website. Also, note that the second highest count is [GET http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1](http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1), which is not important in our analysis since it is only an image link and not a library link. The path report indicates the number of times each rule has occurred in the dataset with the support and the confidence numbers. The first ten lines of the path report are given in table 3. Note that most rules are about images links, which is not an interesting result. It is clear that we have to remove these images and style sheets from our original dataset.

**Table 2: First ten lines of the items report**

Target Item	Count	Support
GET <a href="http://echo.louisville.edu:2048/menu/http/1.1">http://echo.louisville.edu:2048/menu/http/1.1</a>	1218	76.7970
GET <a href="http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1">http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1</a>	1058	66.7087
GET <a href="http://library.louisville.edu:80/css/master.css HTTP/1.1">http://library.louisville.edu:80/css/master.css HTTP/1.1</a>	992	62.5473
GET <a href="http://library.louisville.edu:80/ HTTP/1.1">http://library.louisville.edu:80/ HTTP/1.1</a>	981	61.8537
GET <a href="http://echo.louisville.edu:2048/login?url=http://library.louisville.edu/ HTTP/1.1">http://echo.louisville.edu:2048/login?url=http://library.louisville.edu/ HTTP/1.1</a>	531	33.4805
GET <a href="http://imagesrvr.epnet.com:80/bllimages/css/bodystyle_ehost.css HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/css/bodystyle_ehost.css HTTP/1.1</a>	522	32.9130
GET <a href="http://imagesrvr.epnet.com:80/bllimages/script/setFormFocus.js HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/script/setFormFocus.js HTTP/1.1</a>	485	30.5801
GET <a href="http://imagesrvr.epnet.com:80/bllimages/script/sessionNavigate.js HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/script/sessionNavigate.js HTTP/1.1</a>	476	30.0126
GET <a href="http://imagesrvr.epnet.com:80/bllimages/script/updateFolder.js HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/script/updateFolder.js HTTP/1.1</a>	470	29.6343
GET <a href="http://imagesrvr.epnet.com:80/bllimages/script/popupWin.js HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/script/popupWin.js HTTP/1.1</a>	464	29.2560

**Table 3: First ten lines of the path report**

Rule #	Rule	Support	Confidence
1	<a href="http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1">http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/css/red.css HTTP/1.1">http://library.louisville.edu:80/css/red.css HTTP/1.1</a>	50.2522	75.3308
2	<a href="http://library.louisville.edu:80/css/master.css HTTP/1.1">http://library.louisville.edu:80/css/master.css HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1">http://library.louisville.edu:80/css/headerBlue.css HTTP/1.1</a>	47.9823	76.7137
3	GET <a href="http://library.louisville.edu:80/css/red.css HTTP/1.1">http://library.louisville.edu:80/css/red.css HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/ HTTP/1.1">http://library.louisville.edu:80/ HTTP/1.1</a>	35.8134	58.1967
4	<a href="http://library.louisville.edu:80/css/master4f.css HTTP/1.1">http://library.louisville.edu:80/css/master4f.css HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/css/red3.css HTTP/1.1">http://library.louisville.edu:80/css/red3.css HTTP/1.1</a>	34.7415	81.2684
5	GET <a href="http://library.louisville.edu:80/css/red3.css HTTP/1.1">http://library.louisville.edu:80/css/red3.css HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/css/path.css HTTP/1.1">http://library.louisville.edu:80/css/path.css HTTP/1.1</a>	33.6066	73.1139
6	GET <a href="http://echo.louisville.edu:2048/menu HTTP/1.1">http://echo.louisville.edu:2048/menu HTTP/1.1</a> ==> GET <a href="http://echo.louisville.edu:2048/login?url=http://library.louisville.edu/ HTTP/1.1">http://echo.louisville.edu:2048/login?url=http://library.louisville.edu/ HTTP/1.1</a>	31.2106	40.6404
7	<a href="http://imagesrvr.epnet.com:80/bllimages/script/popupWin.js HTTP/1.1">http://imagesrvr.epnet.com:80/bllimages/script/popupWin.js HTTP/1.1</a> ==> GET <a href="http://imagesrvr.epnet.com:80/bllimages/script/sessionNavigate.js">http://imagesrvr.epnet.com:80/bllimages/script/sessionNavigate.js</a>	29.1929	99.7845

Rule #	Rule	Support	Confidence
	HTTP/1.1		
8	http://imagesrvr.epnet.com:80/bllimages/script/updateFolder.js HTTP/1.1 ==> GET http://imagesrvr.epnet.com:80/bllimages/script/popupWin.js HTTP/1.1	28.8777	97.4468
9	http://imagesrvr.epnet.com:80/bllimages/script/sessionNavigate.js HTTP/1.1 ==> GET http://imagesrvr.epnet.com:80/bllimages/css/bodystyle_ehost.css HTTP/1.1	28.8146	96.0084
10	http://imagesrvr.epnet.com:80/bllimages/script/KeyEvent.js HTTP/1.1 ==> GET http://imagesrvr.epnet.com:80/bllimages/script/queryBuilder.js HTTP/1.1	27.5536	94.5887

The statistics plot and the items plot are very helpful to understand the distribution of the rules obtained from the analysis associated with the support and the confidence numbers. The statistics plot is given in Figure 2. Every small square in this graph represents one association rule with two items, antecedent and consequent. Each rule is identified by the support and confidence number. Figure 3 shows the items plot. The higher the support of the small squares, the darker their color is. Notice that most of the small squares are building a straight line that shows the high number of website noises.

Figure 2: The Statistics Plot

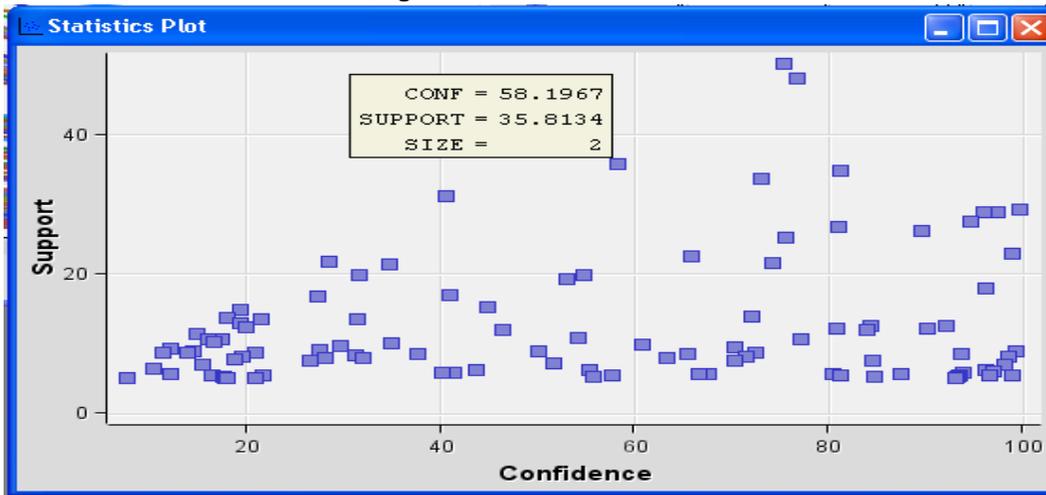
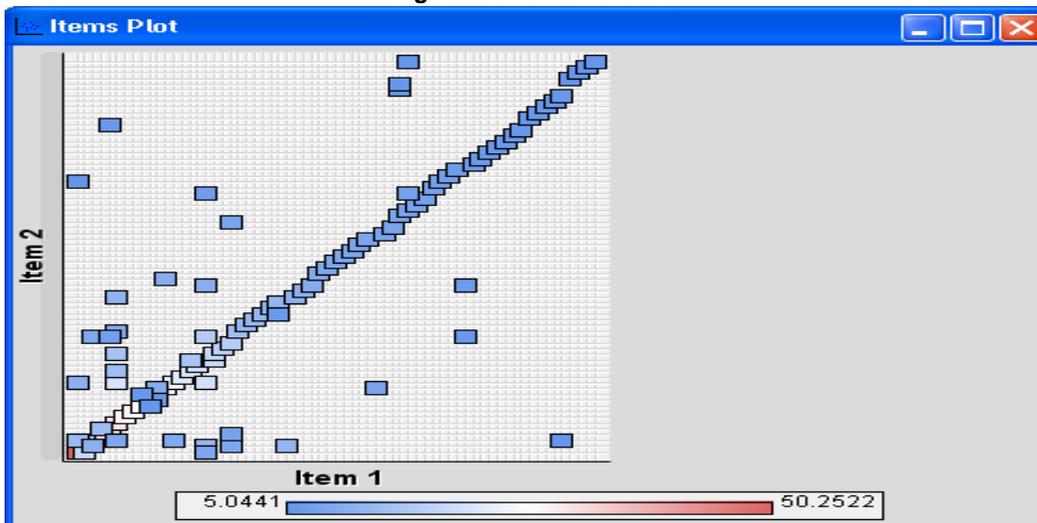
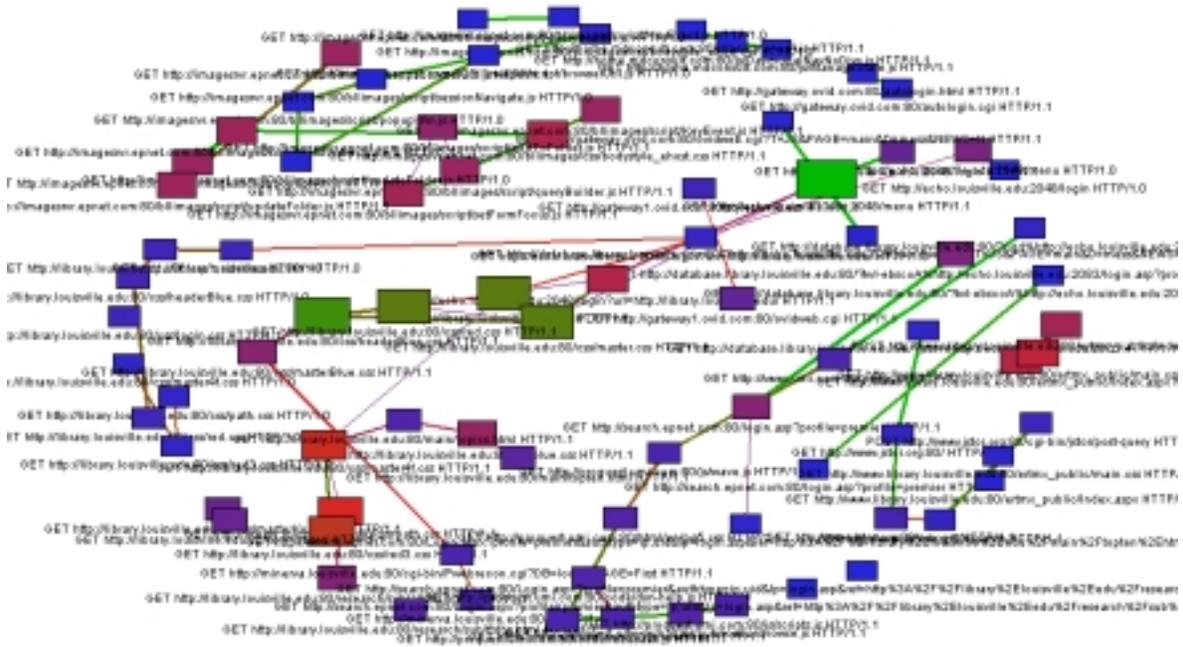


Figure 3: The Items Plot



The graph link shows nodes of items within the dataset that are connected to each other. The link represents a connection between two items in a rule. The higher the confidence, the thicker the link between nodes. The Graph Link is given in Figure 4. Notice that the number of links is too high due to the unnecessary image links.

Figure 4: The Link Graph



The user\_id and ip\_address variables were investigated using the StatExplore node. The resulting analysis was very helpful to identify patrons who log in the most often to the Library website. Figures 5 and 6 provide us with the distribution of the user\_id and ip\_address respectively. Figure 5 shows that user A logs in the most often to the library website. S/He has used the site five times more than any other user while user E has the least frequency of using the library website. Other users exist in the graph and can be explored using the ActiveX capabilities of the Java language. On the other hand, Figure 6 indicates that some machines are used most often to log in to the library website while others are not.

Figure 5: The distribution of User\_id

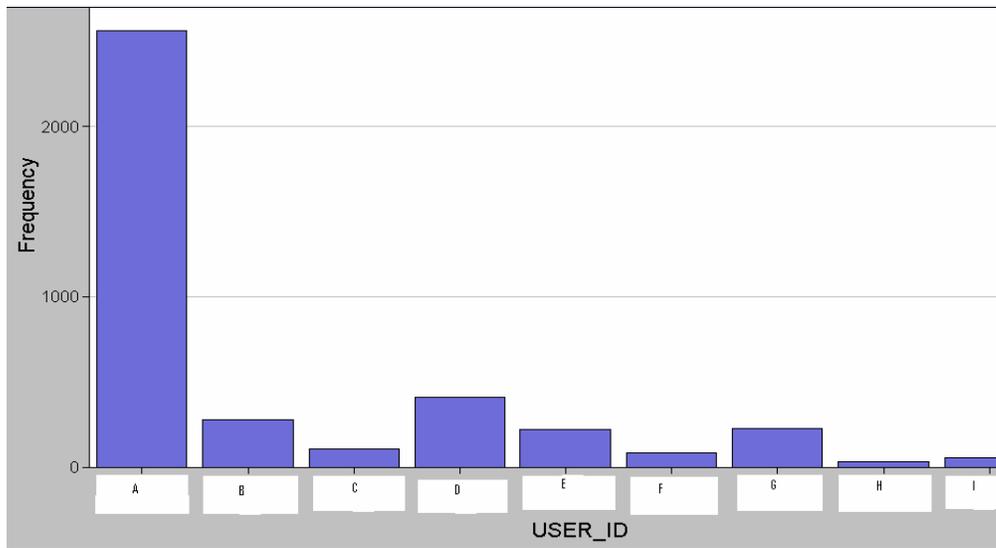
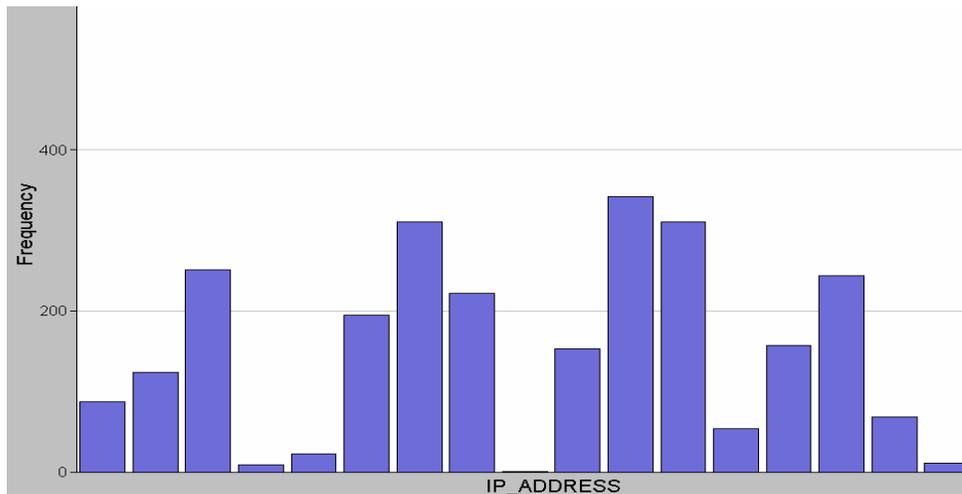


Figure 6: The distribution of the ip\_address

**STEP 2:**

The images and style sheet pages were removed from the original dataset as they are a routine part of the initial access. Their effect on the analysis output was very clear as step 1 showed. When re-running the path analysis, results were improved. The ten lines of the items report are given in table 4. Notice that the item with the highest count is <http://echo.louisville.edu:2048/menu/http/1.1>. This is the first page that a user will view when s/he logs in to the Library website. The second item with the highest count is <GET http://library.louisville.edu:80/ HTTP/1.1>. This link is the library web page. On the other hand, the link <GET http://echo.louisville.edu:2048/login HTTP/1.1> has a high count and support but it is not useful in our analysis since this is a log in page only. The path report indicates the number of times each rule has occurred in the dataset with the support and the confidence numbers. The first ten lines of the path report are given in table 5. Note that the path (rule) that has the highest support was <GET http://echo.louisville.edu:2048/menu HTTP/1.1 ==> GET http://library.louisville.edu:80/ HTTP/1.1>. This rule also has high confidence, which indicates that there is a 63.1806% chance that when a user clicks on the first link, s/he will then click on the second. Some rules have a high support and confidence, but since they contain the login pages, they are not important results.

Table 4: First ten lines of the items report

Target Item	Count	Support
GET <a href="http://echo.louisville.edu:2048/menu HTTP/1.1">http://echo.louisville.edu:2048/menu HTTP/1.1</a>	1855	74.0519
GET <a href="http://library.louisville.edu:80/ HTTP/1.1">http://library.louisville.edu:80/ HTTP/1.1</a>	1535	61.2774
GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1">http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1</a>	947	37.8044
POST <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1">http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1</a>	852	34.0120
GET <a href="http://echo.louisville.edu:2048/menu HTTP/1.0">http://echo.louisville.edu:2048/menu HTTP/1.0</a>	619	24.7106
GET <a href="http://library.louisville.edu:80/research/sub/all dbs.html HTTP/1.1">http://library.louisville.edu:80/research/sub/all dbs.html HTTP/1.1</a>	593	23.6727
GET <a href="http://echo.louisville.edu:2048/login HTTP/1.1">http://echo.louisville.edu:2048/login HTTP/1.1</a>	497	19.8403
GET <a href="http://library.louisville.edu:80/kornhauser/ HTTP/1.1">http://library.louisville.edu:80/kornhauser/ HTTP/1.1</a>	495	19.7605
GET <a href="http://database.library.louisville.edu:80/?kvl-ebascoA%http://echo.louisville.edu:2083/login.asp?profile=premier HTTP/1.1">http://database.library.louisville.edu:80/?kvl-ebascoA%http://echo.louisville.edu:2083/login.asp?profile=premier HTTP/1.1</a>	478	19.0818
POST <a href="http://gateway2.ovid.com:80/ovidweb.cgi HTTP/1.1">http://gateway2.ovid.com:80/ovidweb.cgi HTTP/1.1</a>	478	19.0818

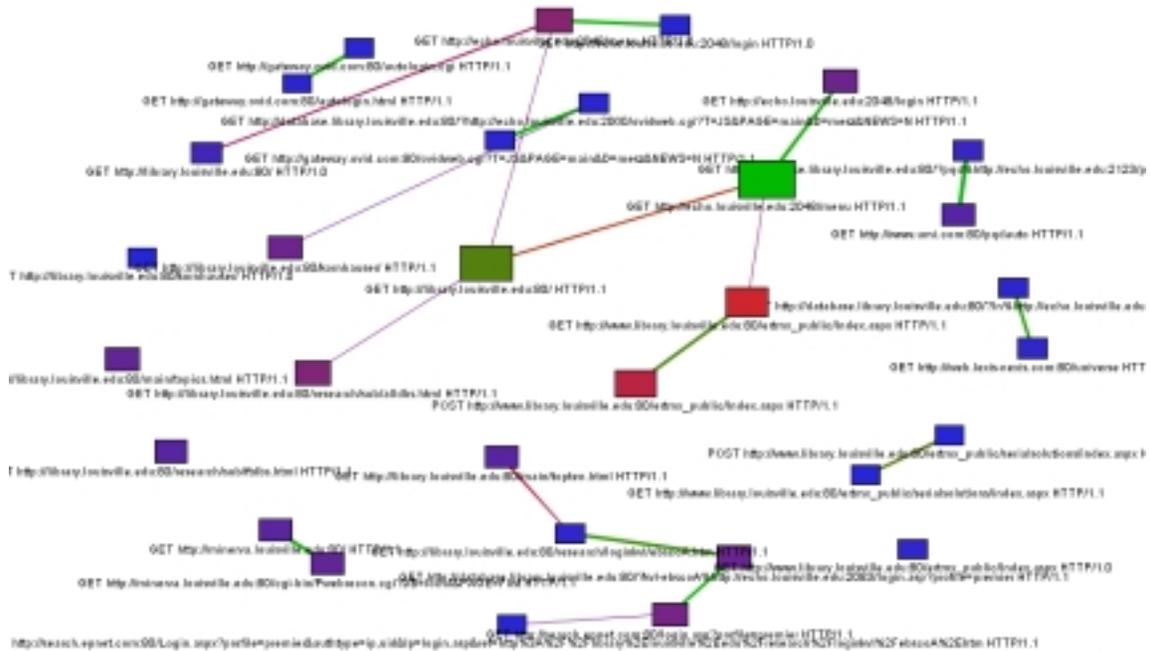
Table 5: First ten lines of the path report

Rule #	Rule	Support	Confidence
1	GET <a href="http://echo.louisville.edu:2048/menu HTTP/1.1">http://echo.louisville.edu:2048/menu HTTP/1.1</a> ==> GET <a href="http://library.louisville.edu:80/ HTTP/1.1">http://library.louisville.edu:80/ HTTP/1.1</a>	46.7864	63.1806
2	GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1">http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1</a> ==> POST <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1">http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1</a>	31.6168	83.6325

Rule #	Rule	Support	Confidence
3	GET http://echo.louisville.edu:2048/menu HTTP/1.1 ==> GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1">http://www.library.louisville.edu:80/ertmx_public/Index.aspx HTTP/1.1</a>	27.5449	37.1968
4	GET http://echo.louisville.edu:2048/login HTTP/1.1 ==> GET <a href="http://echo.louisville.edu:2048/menu HTTP/1.1">http://echo.louisville.edu:2048/menu HTTP/1.1</a>	19.0818	96.1771
5	GET http://database.library.louisville.edu:80/?kvl-ebSCO%A%http://echo.louisville.edu:2083/login.asp?profile=premier HTTP/1.1 ==> GET <a href="http://search.epnet.com:80/login.asp?profile=premier HTTP/1.1">http://search.epnet.com:80/login.asp?profile=premier HTTP/1.1</a>	18.7226	98.1172
6	GET http://library.louisville.edu:80/ HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/research/sub/all dbs.html HTTP/1.1">http://library.louisville.edu:80/research/sub/all dbs.html HTTP/1.1</a>	18.6427	30.4235
7	GET http://echo.louisville.edu:2048/menu HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/kornhauser/ HTTP/1.1">http://library.louisville.edu:80/kornhauser/ HTTP/1.1</a>	15.8882	21.4555
8	GET http://minerva.louisville.edu:80/ HTTP/1.1 ==> GET <a href="http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First HTTP/1.1">http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First HTTP/1.1</a>	15.8882	97.0732
9	GET http://library.louisville.edu:80/ HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/main/topten.html HTTP/1.1">http://library.louisville.edu:80/main/topten.html HTTP/1.1</a>	13.6926	22.3453
10	GET http://library.louisville.edu:80/ HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/main/topics.html HTTP/1.1">http://library.louisville.edu:80/main/topics.html HTTP/1.1</a>	13.3333	21.7590

The Link Graph looks much better than the one from step 1 since many image links were removed from the original dataset. Notice that there are now a lot of interesting links, for example, Proquest, Ebsco and Ovid databases. The Link Graph is given in Figure 7.

Figure 7: Link Graph



**STEP 3:**

The login pages and the users who could not be identified from our dataset were removed since our interest was to investigate the path that patrons take when they log in to the library website. Results were satisfactory. Table 6 shows the first ten items with the highest count. Note that all links are meaningful and the main library web page has the highest count. The first ten lines of the items report are given in table 7. When a patron visits the Ekstrom Library main page, the search for articles is ordered such that a choice for all databases is the first option, Resources by topics is the second, and the top 10 article databases is the last choice. This order was picked randomly according to the Assistant Director for the technology services of the library. The results from table 7 show that clicking on all databases after visiting the library website has support 19.1176 (rule 3), while rule number 5 indicates that visiting the top ten article databases after visiting the main library website has support 114.0114, which is greater than the support for rule 6 (13.7255). This rule is about going from the Library main page to resources by topics. The Library website can be rearranged for better use by putting the top ten article databases as a second option instead of the third, and the resource by topics option third instead of second

**Table 6: First ten lines of the items report**

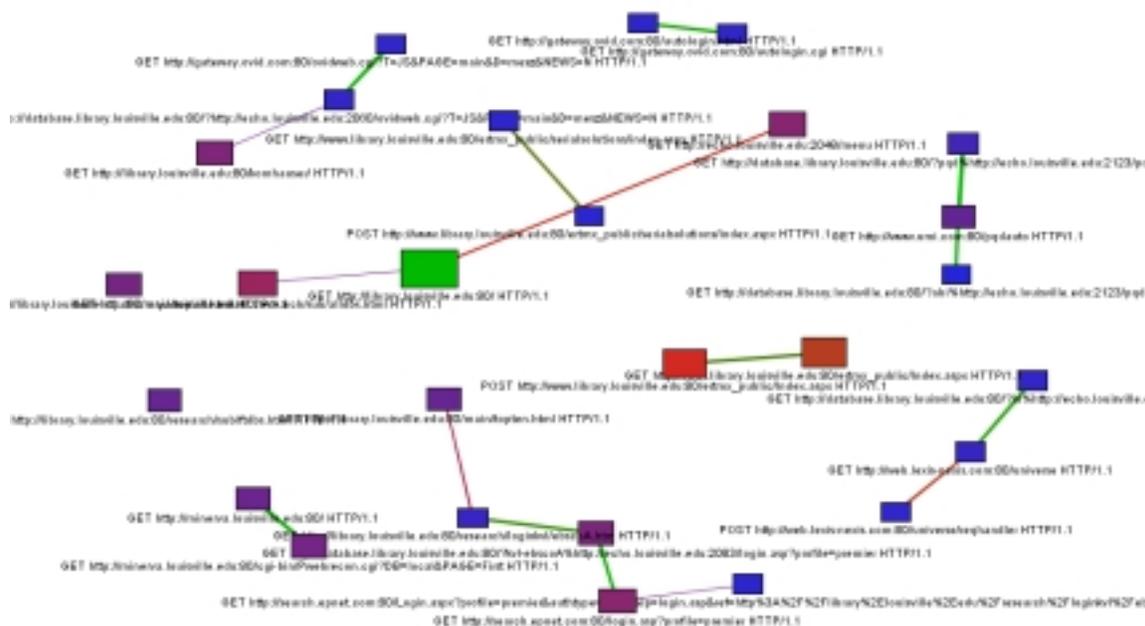
Target Item	Count	Support
GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1	1537	62.7859
GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx">http://www.library.louisville.edu:80/ertmx_public/Index.aspx</a> HTTP/1.1	949	38.7663
POST <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx">http://www.library.louisville.edu:80/ertmx_public/Index.aspx</a> HTTP/1.1	854	34.8856
<a href="http://library.louisville.edu:80/research/sub/all dbs.html">http://library.louisville.edu:80/research/sub/all dbs.html</a> HTTP/1.1	594	24.2647
<a href="http://search.epnet.com:80/login.asp?profile=premier">http://search.epnet.com:80/login.asp?profile=premier</a> HTTP/1.1	530	21.6503
<a href="http://echo.louisville.edu:2048/menu">http://echo.louisville.edu:2048/menu</a> HTTP/1.1	519	21.2010
<a href="http://library.louisville.edu:80/kornhauser/">http://library.louisville.edu:80/kornhauser/</a> HTTP/1.1	494	20.1797
<a href="http://database.library.louisville.edu:80/?ln%http://echo.louisville.edu:2157/universe">http://database.library.louisville.edu:80/?ln%http://echo.louisville.edu:2157/universe</a> HTTP/1.1	479	19.5670
GET <a href="http://library.louisville.edu:80/main/topics.html">http://library.louisville.edu:80/main/topics.html</a> HTTP/1.1	455	18.5866
GET <a href="http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First">http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First</a> HTTP/1.1	448	18.3007

**Table 7: First ten lines of the path report**

Rule #	Rule	Support	Confidence
1	GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx">http://www.library.louisville.edu:80/ertmx_public/Index.aspx</a> HTTP/1.1 ==> POST <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx">http://www.library.louisville.edu:80/ertmx_public/Index.aspx</a> HTTP/1.1	32.4755	83.7724
2	GET <a href="http://database.library.louisville.edu:80/?kvl-ebSCO%http://echo.louisville.edu:2083/login.asp?profile=premier">http://database.library.louisville.edu:80/?kvl-ebSCO%http://echo.louisville.edu:2083/login.asp?profile=premier</a> HTTP/1.1 ==> GET <a href="http://search.epnet.com:80/login.asp?profile=premier">http://search.epnet.com:80/login.asp?profile=premier</a> HTTP/1.1	19.1993	98.1211
3	GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/research/sub/all dbs.html">http://library.louisville.edu:80/research/sub/all dbs.html</a> HTTP/1.1	19.1176	30.4489
4	GET <a href="http://minerva.louisville.edu:80/">http://minerva.louisville.edu:80/</a> HTTP/1.1 ==> GET <a href="http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First">http://minerva.louisville.edu:80/cgi-bin/Pwebrecon.cgi?DB=local&amp;PAGE=First</a> HTTP/1.1	16.3399	97.0874
5	<a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/main/topten.html">http://library.louisville.edu:80/main/topten.html</a> HTTP/1.1	14.0114	22.3162
6	GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/main/topics.html">http://library.louisville.edu:80/main/topics.html</a> HTTP/1.1	13.7255	21.8608
7	GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/research/sub/ftdbs.html">http://library.louisville.edu:80/research/sub/ftdbs.html</a> HTTP/1.1	13.0310	20.7547
8	GET <a href="http://echo.louisville.edu:2048/menu">http://echo.louisville.edu:2048/menu</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1	12.4592	58.7669
9	GET <a href="http://database.library.louisville.edu:80/?pqd%http://echo.louisville.edu:2123/pqdauto">http://database.library.louisville.edu:80/?pqd%http://echo.louisville.edu:2123/pqdauto</a> HTTP/1.1 ==> GET <a href="http://www.umi.com:80/pqdauto">http://www.umi.com:80/pqdauto</a> HTTP/1.1	10.0899	99.5968
10	GET <a href="http://www.library.louisville.edu:80/ertmx_public/Index.aspx">http://www.library.louisville.edu:80/ertmx_public/Index.aspx</a> HTTP/1.1 ==> GET <a href="http://library.louisville.edu:80/">http://library.louisville.edu:80/</a> HTTP/1.1	7.8840	20.3372

Without the images, style sheets and login pages, and without observations about users who could not be identified from our dataset. The Link Graph looks very meaningful, interesting and easy to interpret by a library staff person with the required domain knowledge. The Link Graph is given in Figure 8.

Figure 8: Link Graph



## DISCUSSION

The use of path analysis is very helpful to explore the library data about URL information concerning access to electronic sources. It allows the user to identify the distribution of patrons who log in to the library website and from where they login. Also, Path Analysis provides us with a count of the number of times each link has occurred in the dataset and a list of association rules. The graph link is very easy to interpret. It contains association rules that are very helpful in understanding the path that patrons take as they log in through the library website. It is helpful to use these results for better organization to the library website. Also, it is possible to take this study much further by investigating the library data on a continuing basis.

## CONTACT INFORMATION

Chakib Battioui  
 University Of Louisville  
 2610 Whitehall Ter # 120  
 Louisville, KY 40220  
 Work Phone: 852-6240  
 E-mail: c0batt01@louisville.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.