Journal of Theoretical and Applied Information Technology
15th December 2011. Vol. 34 No.1

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

AN OVERVIEW OF PREPROCESSING OF WEB LOG FILES FOR WEB USAGE MINING

[1]C.P. SUMATHI, [2]R. PADMAJA VALLI ,[3]T. SANTHANAM
[1]Department of Computer Science, SDNB Vaishnav College for Women, Chennai, Tamil Nadu, India
[2]Department of Computer Science, Mother Teresa Women's University Kodaikanal, Tamil Nadu, India
[3]Department of Computer Science,DG Vaishnav College,Chennai, Tamil Nadu, India
E-mail: santsum@hotmail.com, padmahari2002@yahoo.com

ABSTRACT

With the Internet usage gaining popularity and the steady growth of users, the World Wide Web has become a huge repository of data and serves as an important platform for the dissemination of information. The users' accesses to Web sites are stored inWeb server logs.However, the data stored in the log files do not present an accurate picture of the users' accesses to the Web site. Hence, preprocessing of the Web log data is an essential and pre-requisite phase before it can be used for knowledge-discovery or mining tasks. The preprocessed Web data can then be suitable for thediscovery and analysis of useful information referred to as Web mining. Web usage mining, a classification of Web mining, is the application of data mining techniques to discover usage patterns from clickstream and associated data stored in one or more Web servers.This paper presents an overview of the various steps involved in the preprocessing stage.

Keywords:*Web Server, Data Cleaning, User Identification, Session Identification, Path Completion*

1. INTRODUCTION

With the continuous growth and abundance of information available on the Internet, the World Wide Web(WWW) has become a huge repository of information. The Web has become an important medium to communicate ideas, transact business and promote entertainment. The discovery and analysis of useful information from the Web documents is referred to as Web mining[1].However, the data stored in the Web is heterogeneous. Web pages may not only contain textual content but also other types of Web data which may include audio files, video files and images. In addition, the Web data is primarily semi-structured.The huge, heterogeneous, dynamic and semi-structured characteristics of Web data raise a great challenge for the Web userssince it is available in an unsuitable format to extract useful information from the WWW.The preparation of a suitable data set is a pre-requisite and significant task for the mining of Web data[2],[3].Hence, Web mining is divided into 4 major sub-tasks[4]:

1) Data Collection
2) Data Preprocessing
3) Pattern Discovery
4) Pattern Analysis

The Web data is stored in Web servers, client machines, proxy servers or organizational databases.The primary data sources used in Web usage mining are the server log files which include Web server access logs, referrer logs and agent logs. Additional data sources that are also essential include the site files and meta-data, operational databases and domain knowledge. In some cases and for some users, additional data may be available in the client-side and proxy-server. Referrer logs contain information about the referring pages for each page reference. There are various types of Web data such as content data, structure data and usage data. Based on the type of data to be mined for analysis, Web mining can be further classified into Web content mining, Web structure mining and Web usage mining. Specifically, Web usage mining is the application of data mining techniques to discover usage patterns from clickstream and associated data stored in one or more Web servers to cater to the needs of Web-based applications[5],[6].However,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the data stored in the various data sources do not present an accurate picture of the pages requested or the identification of the user. Hence data preparation techniques are necessary to transform the raw server logs into a suitable data file for Web usage mining.

Data preprocessingis predominantly significant phase in Web usage mining due to the characteristics of Web data and its association to other related data collected from multiple sources. This phase is often the most time-consuming and computationally intensive step in Web usage mining. This process is critical to the success of Pattern discovery and Pattern Analysis. In short, the whole process deals with the conversion of raw Web server logs into a formatted user session file in order to perform Web usage mining. This paper focuses on the following:

i) Preprocessing of Web usage data from Web log files resulting in a user session file

ii) Formatting of the user session file suitable for mining tasks.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 discusses the various steps involved in preprocessing Web data for usage mining. Section 4 deals with the Experimental Design. Finally, Section 5 provides Conclusion.

2. RELATED WORK

Data preprocessing phase presents a number of challenges that has been discussed in various research papers. Various steps in data preprocessing steps have been discussed in detail in [3],[5],[6].The research carried out by Pirolli etal. [7] and Pitkow[8] h**a**ve discussed the difficulty in the identification of users and sessions from Web server logs. A study to assess heuristics for session identification from Web log data has been presented by Spiliopoulou et al. [9].In the research carried out by Borges and Levene[10], data cleaning was performed by removing the erroneous requests and image requests. A session was defined as a sequence of requests from the same IP address with a timelimit of 30 minutes between consecutive requests. After session identification, sessions with a single request were removed.The data preprocessing techniques detailed by Srivatsava et al.[6] and Mobasher[11] have been experimented with the Web server log of the library of South-Central University for Nationalities[12].As part of data preprocessing, data conversion was performed on the CTI data set byMustapha et al. [13],assigning a numerical value to each URL and bitmap algorithm was used to group a set ofattributes into one attribute. Navin Kumar Tyagi etal. have outlined the various data preprocessing activities and have presented algorithms for data cleaning and data reduction [14]. Edmond H. Wu etal have presented an efficient multidimensional data model for aggregating user access sessions for Web usage analysis[15]. The model focuses on the page, user and timeattributes to form a multidimensional cubeto effectively support differentdata mining applications.Yang.Q. et. alhaveintroduced a data-cubemodel to contain the original access sessions for data mining from Weblogs [16].

3. MATERIALS AND METHODS

Usage Preprocessing consists of the following steps [3]:

- Data Fusion
- Data cleaning
- User identification
- Session identification
- Path completion
- Formatting

The Usage preprocessing phase results in the creation of a suitable user session file suitable for input into specific data mining operations. Figure 1 shows the various steps in data preprocessing for Web Usage Mining.

3.1.USAGE DATA PREPROCESSING

A Web server log is an important source for performing Web usage mining since it explicitly records the browsing behavior of users to the site[6]. These log files are stored in various formats such as Common Log Format (CLF) or Extended Log Format (ELF). Every entry in the log file in the ELF stores the following fields:

- Client IP address or host name
- User Id(‘ -’, if anonymous),
- Access time
- HTTP request method(GET, POST, HEAD)
- Path of the resource on the Web server
- Protocol used for transmission
- Status code
- Number of bytes transmitted.
- User agent(browser, operating system type and version)
- Referrer

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

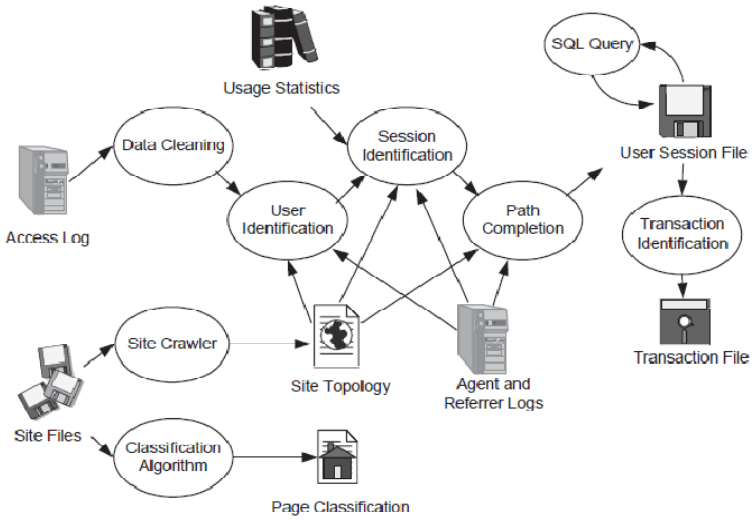

Figure 1: Steps in Data Preprocessing for Web Usage Mining

The CLF does not contain the user agent and referrer fields. A portion of the raw Web server access log files on the server(maya.cs.depaul.edu) before data cleaning is shown in Figure.2.

2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1
maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)
http://dataminingresources.blogspot.com/

2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1
maya.cs.depaul.eduMozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.507
27) http://maya.cs.depaul.edu/~classes/cs589/papers.html

2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1
maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)
http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey

---

2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1
maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/

2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1
maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/announce.html

2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
http://maya.cs.depaul.edu/~classes/cs480/announce.html

Figure 2: A portion of server log

3.1.1. DATA FUSION

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In extensiveWeb sites, data is stored in various Web logs such as access log, referrer log and agent logs. Data fusion refers to the use oftechniques that combine data from multiple sources to retrieve additional informationwith respect to identification of users and sessions, than if they are retrieved from a single data source.

3.1.2. DATA CLEANING

The purpose of data cleaning is to remove irrelevant items stored in the log filesthat may not be useful for analysis purposes[3],[5]. When a user accesses a HTML document, the embedded images, if any, are also automatically downloaded and stored in the server log. For example, log entries with file name suffixes such as gif, jpeg, GIF, JPEG, jpg and JPG can be removed. Since the main objective of data preprocessing is to obtain only the usage data, file requests that the user did not explicitly request can be eliminated. This can be done by checking the suffix of the URL name. In addition to this, erroneous files can be removed by checking the status of the request (such as status code 404). Data cleaning also involves the removal of references resulting from spider navigations which can be done by maintaining a list of spiders or through heuristic identification of spiders and Web robots [17].The cleaned log represents the user's accesses to the Web site. Sample information from access, referrer and agent logs [6] is shown in Table1.

3.1.3. PAGEVIEW IDENTIFICATION

In Web usage mining, the most basic data construct is a pageview. A pageview can be defined as an aggregate representation of a collection of Web objects as a result of a user action. Each pageview is assigned certain attributes that include pageview id, a URL which identifies the pageview, duration and other metadata such as keywords [11].

3.1.4. USER IDENTIFICATION

Though Web usage analysis does not require the identity of the users, it is essential to differentiate among users. This step requires the identification of unique users. User identification deals with associating page references with different users. As already mentioned, local caching and proxy servers pose some of the problems for obtaining reliable usage data. To reduce network traffic and improve performance, the pages that are requested are cached by most Web browsers. Hence, when the user navigates backwards by using the "back" button, the repeat page access is not recorded in Web server log. Proxy servers provide an intermediary solution but the difficulty of user identification still persists. All requests coming from a proxy server have the same identifier even though the requests are put forth by multiple users. Two solutions for this problem are user registration data and use of cookies. One method to identify users is by means of the user id field in the server log files. The user registration data helps in capturing additional demographic information in addition to the data which is automatically collected in the server log. However, due to privacy reasons, many users prefer not to browse sites that require registration and logins. Sometimes user registration data is not compulsory and users may often provide incorrect information. Hence, user identification becomes a complex task unless an exact user id is provided. In the absence of authentication mechanisms, the most well-known approach is the use of cookies. However, not all sites utilize cookiesand due to privacy concerns, client-side cookies are sometimes disabled or deleted by users. The difficulty of identifying users from Web server logs has been addressed in research[8].

In the absence of meaningful information in the user id field, various heuristics are employed using the other data available in the server log files such as the IP address, User agent and Referrer for user identification.

3.1.4.1. USER AGENT

User agent plays an important role in user identification. It refers to the browser used by the client. A change in the browser or the operating system under the same IP address represents a different user heuristically. From Table 1, rows 5 and 6 show that the pages have been accessed using a different browser than the other entries. These imply that the log represents at least two different users.

3.1.4.2. REFERRING URL

IP addresses, alone, are generally not sufficient for user identification. This is due to the fact that proxy servers may assign the same IP address to multiple users or the same user may be assigned multiple IP addresses by the proxy server. Another heuristic for user identification is to make use of theaccess log, referrer log and site topology to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

#	IP Address	User Id	Time	Method of URL / Protocol	Referred	Agent
1.	192.168.1.1	-	19/Feb/1998:03:04:41-0500	"GET A.html HTTP/1.0"	-	Mozilla3.04(Win95,1)
2.	192.168.1.1	-	19/Feb/1998:03:05:34-0500	"GET B.html HTTP/1.0"	A.html	Mozilla3.04(Win95,1)
3.	192.168.1.1	-	19/Feb/1998:03:05:39-0500	"GET L.html HTTP/1.0"	-	Mozilla3.04(Win95,1)
4.	192.168.1.1	-	19/Feb/1998:03:06:03-0500	"GET R.html HTTP/1.0"	L.html	Mozilla3.04(Win95,1)
5.	192.168.1.1	-	19/Feb/1998:03:06:58-0500	"GET A.html HTTP/1.0"	-	Mozilla3.01(Win95,1)
6.	192.168.1.1	-	19/Feb/1998:03:07:42-0500	"GET B.html HTTP/1.0"	A.html	Mozilla3.01(Win95,1)
7.	192.168.1.1	-	19/Feb/1998:03:07:55-0500	"GET F.html HTTP/1.0"	B.html	Mozilla3.04(Win95,1)
8.	192.168.1.1	-	19/Feb/1998:03:09:58-0500	"GET O.html HTTP/1.0"	F.html	Mozilla3.04(Win95,1)
9.	192.168.1.1	-	19/Feb/1998:03:10:28-0500	"GET G.html HTTP/1.0"	B.html	Mozilla3.04(Win95,1)
10.	192.168.1.1	-	19/Feb/1998:05:05:58-0500	"GET A.html HTTP/1.0"	-	Mozilla3.04(Win95,1)
11.	192.168.1.1	-	19/Feb/1998:05:06:58-0500	"GET D.html HTTP/1.0"	A.html	Mozilla3.04(Win95,1)

Table 1:*Sample information from access, referrer and agent logs*

determine the navigation paths for each user.Each HTTP request method is checkedif it is directly linked to the previous pages already visited[7]. If a page that is requested is not linked to the previous pages, it is assumed that multiple users are assumed to exist under the same IP address.From Table 1, page L is not directly reachable from page A or page B. This results in the identification of a third user.

Hence, user identification step results in the identification of three unique users with navigation paths such as A-B-F-O-G-A-D (row1,row2,rows7-11), A-B(row5,row6), L-R(row3,row4).

3.1.5SESSION IDENTIFICATION

Session identification splits all the pages accessed by a user into different sessions. Users may have visited the pages for long periods of time. It is necessary to divide the log entries of a user into multiple sessions through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user has started a new session. In general, a 30-minute default timeout is considered. Hence the log file, after user identification, may be further divided into sessions for every user. Hence each user's page visits will be split into one or more sessions. Using a 30-minute timeout, the path for user 1 is broken into two separate sessions;

User1: A-B-F-O-G
A-D

User 2: A-B

User 3: L-R

Hence the session identification step results in four sessions.

3.1.6 PATH COMPLETION

It is necessary to determine the existence of important accesses that are not recorded in the access log. Path completion refers to the inclusion of important page accesses that are missing in the access log due to browser and proxy server caching. Similar to user identification, the heuristic assumes that if a page is requested that is not directly linked to the previous page accessed by the same user, the referrer log can be referred to see from which page the request came. If the page is in the user's recent click history, it is assumed that the user browsed back with the "back" button, using cached sessions of the pages. Hence each session reflects the full path, including the pages that have been backtracked.

After the path completion phase, the user session file results in paths consisting of a collection of page references including repeat page accesses made by a user.

User 1: A-B-F-O-F-B-G
A-D

User 2: A-B

User 3: L-R

Data preprocessing results in the creation of a user session file consisting of a set of users, each user associated with one or more sessions. A user session is a collection of page references made by a user during a single visit to a site. In certain cases, user sessions may be further divided into semantically meaningful groups of page references referred to as transaction. A transaction can be identified using one of the approaches such as maximal forward reference, reference length and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

time window leading to the creation of a user transaction file.

3.1.5 FORMATTING

However, the resultant filemay not be in a suitable format to be used for any mining tasks. Hence, the data should be appropriately formatted according to the type of mining tasks undertaken. Information which is viewed irrelevant or unnecessary for the analysis may not be included in the resultant session file.

4. EXPERIMENTAL DESIGN

The data set available in http://www/cs.depaul.educonsists of a random sample of users visiting the site for a 2-week period. There are 5 files: cti.cod, cti.std, cti.tra, cti.stat, cti.nav. The cti.nav file has been used for experimental purposes. It contains unfiltered, preprocessed, sessionized data. The original unfiltered data contains a total of 20,950 sessions. It is assumed that the various phases of data preprocessing such as data cleaning, user identification and session identification have been implemented in the data set resulting in 13,745 sessions and 5446 users. Multiple consecutive sessions corresponding to the same user id are likely to exist. Each line consists of three fields: timestamp (number of seconds relative to January 1, 2002), pageview accessed and referrer.

Each session begins with a line of the form: Session #n (User_id = k) where 'n' is the session number and 'k' is the user id. For example, a sample of the file is shown in Figure 3.

This user session file is in an unstructured format which is not suitable for specific mining tasks to be performed.Web usage mining mainly deals with a sequence of pageview requests of users in sessions. The user session file should represent a collection of page references, which may include repeat page accesses, made by a user. One of the possible formats that can be adopted is to format the above file as a sequence of pagereferences suitable for Web usage mining tasks. The first column represents the session id, the second column represents the user id and the subsequent columns represent the page references. The resulting formatted user session file has been obtained usingJava as shown in Figure 4. Additionally, appropriate page identification can be assigned to each page reference. Subsequently, post-processing of the formatted user session file can be done. The entire set of sessions can be represented as a usage matrix, referred to as session-pageview matrix, where the i^{th} row in the matrix represents the pages visited by the user in the session and the j^{th} column in the matrix represents the sessions in which the page *j* has been visited. The cell value *(i.j)* in the session-pageview matrix corresponds to the frequency of the pageview*j* in session *i.*

The above steps in data preprocessing can be applied to any data set before mining tasks are performed.

5. CONCLUSION

The data collected in the Web server and other associated data sources do not reflect precisely about the pages visited by the user during his interactions with the Web. Due to the presence of superfluous items, in addition to the inability to identify users and sessions, it is essential that the log files need to be preprocessed initially before the mining tasks can be undertaken.

Data preprocessing is a significant and pre-requisite phase in Web mining. Various heuristics are employed in each step so as to remove irrelevant items and identify users and sessions along with the browsing information. The output of this phase results in the creation of a user session file. Nevertheless, the user session file may not exist in a suitable format as input data for mining tasks to be performed. This paper has focused on a design that can be adopted for preliminary formatting of a user session file so as to be suited for various mining tasks in the subsequent pattern discovery phase.

In addition to the above mentioned preprocessing and formatting tasks, the future work involves various data transformation tasksthat are likely to influence the quality of the discovered patterns resulting from the mining algorithms. The discovered patterns can then be used for various Web usage applications such as site improvement, business intelligence and recommendations.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
SESSION #9  (USER_ID = 9)
9483628     /news/default.asp/              -
9483687     /people/                       /news/default.asp
9483792     /people/search.asp?sort=ft     /people/


SESSION #10  (USER_ID = 10)
9469488     /news/default.asp       -
9469503     /courses/               /news/default.asp
9469522     /courses/schedule.asp   /courses/
9469662     /courses/default.asp    /courses/schedule.asp
9469688     /courses/syllabilist.asp /courses/default.asp


SESSION #11  (USER_ID = 10)
10262065    /news/default.asp       -
10262083    /courses/               /news/default.asp
```

Figure 3: A portion of preprocessed session file

```
9 9 /news/default.asp /people/ /people/search.asp?sort=ft

10  10/news/default.asp  /courses/ /courses/schedule.asp    /courses/default.asp   /courses/syllabilist.asp

11 10 /news/default.asp  /courses/
```

Figure 4: A portion of formatted user session file

REFERENCES:

[1] O.Etzioni, "The world wide web: Quagmire or gold mine." Communications of the ACM, :39(11):65–68, 1996.

[2] U.Fayyad , G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery: An overview." In Proc. ACM KDD,1996,1-34.

[3] R.Cooley, Bamshad Mobasherand Jaideep Srivastava, "DataPreparation for Mining World Wide Web Browsing Patterns." Knowledge and Information Systems,1(1),1999,5-32

[4] R.Kosala and H. Blockeel, "Web Mining Research : A Survey." ACM SIGKDD Explorations, 2000, 1-15.

[5] R.Cooley, B. Mobasher and J. Srivatsava, "Web mining: Information and pattern discovery on the World Wide Web." 9th IEEE Inernational Conference on Tools with Artificial Intelligence, CA, 1997, 558-567.

[6] J.Srivatsava, R.Cooley, M.Deshpande and P.N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter,2000,12-23.

[7] P.Pirolli,J.Pitkowand R. Rao, "Silk from a sow's ear: Extracting usable structures from the Web." In Proc. of 1996 Conference on Human Factors in Computing Systems (CHI-96). Vancouver, British Columbia, Canada, 1996,118-125

[8] J.Pitkow, "In search of reliable usage data on the WWW." In Sixth International World Wide Web Conference. Santa Clara, CA, 199, 451-463.

[9] M.Spiliopoulou,B.Mobasher, B.Berendt and M.Nakagawa, "A framework for the evaluation of session reconstruction heuristics in web usage analysis." IN-FORMS Journal on Computing,2003, 171-190

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[10] J.Borges and M. Levene, "Mining users' web navigation patterns and predicting their next step." NATO Secur. Sci. Ser. D-Inform. Commun. Secur. 2008,45-55.

[11] B.Mobasher, "Web Usage Mining and Personalization." Practical Handbook of Internet Computing. Ed. Editor, CRC Press M.P. Singh. 2004,1-37

[12] L.Chaofeng, "Research on Web Session Clustering."Journal of Software,2009, 460-468.

[13] N.Mustapha, M. Jalali and M. Jalali, "Expectation maximization clustering algorithm for user modeling in web usage mining systems." Eur.J. Sci. Res.,2009,467-476.

[14] Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi, "An Algorithmic Approach To Data Preprocessing in Web Usage Mining." International Journal of Information Technology and Knowledge Management ,2010, 279-283.

[15] Edmond H.Wu, Michael.K.Ng and Joshua Z.Huang,"A Datawarehousing and Data Mining Framework for Web Usage Management." Communications in Information and Systems, 2004,301-324.

[16] Q.Yang, J. Huang, and M. Ng, "A data cube model for prediction-based Web prefetching." Journal of Intelligent Information Systems, 2003,11-30.

[17] P.Tan, and V.Kumar, "Discovery of Web Robot Sessions Based on Their Navigational Patterns." Data Mining and Knowledge Discovery, 6:9-35, 2002.