Web Usage Mining on Proxy Servers: A Case Study

Jan Kerkhofs Prof. Dr. Koen Vanhoof Danny Pannemans
Limburg University Centre

July 30, 2001

Abstract

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Commercial companies as well as academic researchers have developed an extensive array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behaviour on a particular web site. Performing this kind of investigation on your web site can provide information that can be used to better accommodate the user's needs. An area that has received much less attention is the investigation of user behaviour on proxy servers. Servers of Internet Service Providers (ISP's) log traffic from thousands of users to thousands of web sites. No doubt that web server administrators are interested in comparing the performance of their own site with those of competitors. Moreover, this kind of research can give a general overview of user behaviour on the Internet or an overview of behaviour within a specific sector.

It was a Belgian ISP that showed interest in the subject and consequently provided data from one of their proxy servers for a thesis. This paper is a summary of that thesis and lays an emphasis on the attained results. The ISP chose to remain anonymous because of privacy-issues.

1 Introduction

The Internet is generally said to have become available to a large public around 1994–1995. Since that time a great number of companies have thrown themselves on this new medium. In the beginning many entrepreneurs saw great new opportunities to make money by setting up an internet company. Later on, some of the so-called *brick-and-mortar* companies began to see a need to go online. Some of those even changed their business so drastically that not much of the original company was left. Every large company has spent a lot of effort and money to develop a well-established web site. The ones that have not spent enough effort may find themselves faced with strategic disadvantages in years to come.

In order to have a successful web site (and especially a successful e-commerce site) it is crucial to know the users of that site. This need has given rise to a whole new field in research, called Web Usage Mining. It is commonly seen as a subdivision of Web Mining, which implies that data mining techniques are applied to data from the World Wide Web. When the data under consideration emerges from web servers log files, we enter the field of web usage mining. It is therefore the "automatic discovery of user access patterns from Web servers" [8].

Because it is so important to know one's customers in order to better suit their needs, companies are willing to spend money on the analysis of their log files. As a consequence, apart from tools that were developed by academic researchers, there is simultaneously a significant number of commercial tools that have been developed to meet these needs. Examples of academic tools include WebSIFT [1] and Web Utilization Miner [10]. For a more extensive overview, see [3]. An example of a commercial web usage mining tools is EasyMiner, developed by MINEit Software Ltd. All of these tools are designed to understand the most common log file formats so that the

process requires very little preprocessing. Unfortunately, when analyzing a log file from a Web server, one can only analyze browsing behaviour on a single site.

To perform research on a sector or even on general browsing behaviour, the log file data of a proxy server are a lot more appropriate because of the many-to-many relationship between sites and users. This topic will be further elaborated in the next sections. Section 2 will give a short introduction to data collection, preprocessing and data mining techniques in web usage mining. Section 3 will introduce the concept of e-metrics. These are metrics applied to web usage data that attempt to quantify the performance of web sites. Section 4 will introduce the data that research was executed on. After that, section 5 will describe the results of that research. Finally, in section 6 we will present a conclusion and some hints for future research.

2 Aspects of Web Usage Mining

As in other forms of data mining, in web usage mining it is equally important to pay attention to a proper data collection, a thorough preprocessing phase and the data mining techniques themselves. For a more general and extensive explanation of these topics, there are several other articles commonly available. We will only introduce them, bearing in mind that our research will be on data from a proxy server.

2.1 Data collection

Data for web usage mining can be collected at several levels. We may be faced with data from a single user or a multitude of them on one hand and a single site or a multitude of sites. Combining both factors offers four possibilities, illustrated in figure 1.

	1 Site	**Multiple sites**
1 user	Java applets or Javascripts	Modified browser
Multiple users	Server level	Proxy server level

Figure 1: Segments of web traffic

Data about behaviour of a single user on a single site can be collected by means of Javascripts or Java applets. Both methods require user participation in the sense that the user has to enable their functionality. An applet has the additional problem that it may take some time to load the first time. However, it has the advantage that it can capture all clicks, including pressing the back or reload buttons. A script can be loaded faster, but cannot capture all clicks.

A modified browser is situated in the second segment. It can capture behaviour of a single user over all visited web sites. It's advantages over Java applets and Javascripts are that it is much more versatile and will allow data collection about a single user over multiple Web sites [3]. That is why this kind of data collection is used requlary by market research groups, e.g. Nielsen//Netratings, in order to collect information on how certain user groups behave online.

The third way of data collection is on the Web server level. These servers explicitly log all user behaviour in a more or less standardized fashion. It generates a chronological stream of requests that come from multiple users visiting a specific site. Since Web servers keep record of these requests anyhow, this information is readily available. Sometimes an analyst will use some additional information to better identify users, such as information from *cookies* or socio-demographic information about the users that may have been collected. This kind of data collection also has a number of drawbacks. Like Javascripts, it cannot capture page views that were generated by pressing back or reload buttons. Apart from that, it also cannot log page views generated by a cache, either a local cache on the computer of the user, or a cache from an ISP's proxy server.

The fourth level of data collection logs behaviour of multiple users visiting multiple Web sites. This kind of information can be found in log files origination from proxy servers. These servers are used by ISP's to give customers access to the World Wide Web. They also function as a cache server. This means that they will keep pages that were recently requested on this server and, if the same request is made by another user shortly after that, they will send the cached page to that user, instead of requesting it once more on the Web server were that page is located.

2.2 Preprocessing

Preprocessing is an aspect of data mining of which the importance should not be underestimated. If this phase is not performed adequately, it is not possible for the mining algorithms to provide reliable results.

2.2.1 Data cleaning

First of all, irrelevant data should be removed to reduce the search space and to skew the result space. Since the intention is to identify user sessions, build up out of page views, not all hits in a log file are necessary. This is true for server logs as well as for proxy logs. A log file generates a hit for every requested file. Since a HTML-page may consist of several files (text, pictures, sounds, several frames) it would be useful if we could keep only a single hit for each page view. To get an idea of user behaviour, it is only necessary to keep track of the files that the user specifically requested. Very often, all hits with a suffix like .jpg, .gif, .wav, etc. are removed out of the log file. Even though this will also be done in the research that will be described later on, it also has a drawback. Sometimes users specifically want to see a picture on a separate page. This page view will be deleted while it shouldn't.

2.2.2 User and session identification

After the log file has been cleaned, the next step is to identify users. This very often poses a serious problem. If every computer in the world had it's own unique IP-address, there wouldn't be a problem. However, most ISP's make use of *dynamic IP-addresses*. This means that every time a user logs on to the Internet, he will be given a different address. This makes it impossible to distinguish returning users. As a consequence, it is usually simply assumed that every new IP-address represents a new user. Sometimes it is possible to identify a new user even when the IP-address it the same. This occurs when the agent log shows a change in browser software or operating system [6]. Some sites try to solve the problem of user identification through the use of cookies that contain an identification number. However, users very often delete cookies or disable their use, which makes that this technique is not always reliable either. Other sites try to identify users by asking them for a login and password. It is clear, however, that not every site can do this since it very often scares users away.

Assuming that users have been identified, the next step is to identify sessions. The goal of session identification is to divide the page accesses of each user into individual sessions [9]. A rule of thumb that is commonly used is that when there is an interval of 30 minutes between two page views, the click stream should be divided in two sessions. This rule has been applied since Pitkow [7] established a timeout of 25.5 minutes, based on empirical data. This is why, in the further research in this paper, a timeout of 30 minutes has been adopted. After users and sessions have been identified, the file under consideration will have additional fields that mention for each line the number of the user and of the session.

2.3 Data mining techniques

For the actual pattern discovery in web usage mining, mostly the same techniques are employed as in other forms of data mining. The most common ones will be briefly described.

2.3.1 Log file analysis

Even though this is not a data mining technique as such, it is probably the most widely used technique to obtain structured information out of server logs. There is a large number of tools on the market that will accept the most common log file formats as an input to answer some basic questions that every Web site administrator has. It will provide information such as: the number of hits and page views, the number of unique and returning users, the average length of a page view, an overview of the browsers and operating systems that were used, an overview of keywords that were used in search engines and that led to the Web site, etc. Despite lacking in the depth of its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions [3].

2.3.2 Association rules

In web usage mining, association rules are used to find out which pages are frequently visited together. In the particular research carried out in this work, they will be used to find out which Web sites and which sectors are frequently visited together. An association rule is usually presented in the following syntax:

```
KetnetBe <= VtmBe & Tv1Be (15:2.788%, 0.27)
```

This rule means that out of the 15 instances (representing 2.788% of the database) that visited the sites of `www.vtm.be` and `www.tv1.be` together, 27% also visited `www.ketnet.be`. The *support* is 15, the *confidence* 27%.

2.3.3 Sequential patterns

This technique has some similarities with the association rules and will be used in this paper in addition to those association rules. The difference is that it takes the time dimension into account. The algorithm tries to find sequences in which a certain page (or Web site) usually comes before of after another page (or Web site). In other words, it "attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes" [3].

2.3.4 Clustering

In general, clustering is a process of creating a partition so that all the members of each set of the partition are similar according to some metric [2]. In web usage mining, we can narrow the definition to *a technique to group users in clusters based on their common characteristics.* Clustering algorithms learn in an unsupervised way. They discover their own classes and subsets of related objects in the training set. Then it has to find descriptions that describe each of these subjects.

2.3.5 Classification

Contrary to clustering, classification is a supervised way of learning. The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes whereas the remaining attributes are called predicting attributes. A combination of the predicted attributes defines a class [2]. In the Web domain one is interested in developing a profile of users belonging to a particular class or category. For example, 45% of users who visit two or more sites of television stations in a single session, are younger than 21. The algorithms that perform classification include decision tree classifiers, Bayesian classifiers, k-nearest neighbour classifiers, etc.

3 E-metrics

E-metrics are based on statistics, which are one data mining technique. Therefore, they can be considered to be a web usage mining method like any other. Moreover, they also try to gain insight into browsing behaviour of users and performance of Web sites.

E-metrics are measures with which Web sites can be evaluated. They can be compared with regular metrics and ratios as these are used in traditional industry, such as return on investment, net profit, market share, rentability, etc. As Web sites gain a more important position in companies, there emerges a need to evaluate these Web sites—that consume more and more money—and quantify their performance. The intention is to give indications of how well the Web site performs in order to investigate to what extent these measures change over time and how well they perform compared to those of competitors.

Two kinds of e-metrics can be identified, those that can be applied to every Web site and those that were designed for a specific kind of Web site, very often e-commerce sites.

3.1 General e-metrics

In this section we will briefly describe a few general e-metrics. However, both *stickiness* and *average duration* will be explained more in detail because of their importance in the undertaken research.

3.1.1 Stickiness

This is probably one of the most widely used e-metrics. It is a composite metric that indicates the effectiveness with which the content of the page or the Web site can keep the attention of the user. In general it is assumed that sticky sites are better than sites that are less sticky. A possible formula is as follows:

$$\text{Stickiness} = \text{Frequency} * \text{Duration} * \text{Total site reach}$$

Where

$$\text{Frequency} = \frac{\text{Number of visits in time period T}}{\text{Number of unique users who visited in T}}$$

$$\text{Duration} = \frac{\text{Total amount of time spent viewing all pages}}{\text{Number of visits in time period T}}$$

$$\text{Total site reach} = \frac{\text{Number of unique users who visited in T}}{\text{Total number of unique users}}$$

This formula can be reduced to:

$$\text{Stickiness} = \frac{\text{Total amount of time spent viewing all pages}}{\text{Total number of unique users}}$$

so that one doesn't need to have all the data for the complete formula to calculate stickiness. Usually stickiness is expressed in minutes per user.

3.1.2 Average duration

This is quite a simple e-metric with which several pages of a Web site (or complete Web sites) can be compared to each other. The metric expresses how long users view a certain page or site on average. In this work the following formula has been used:

$$\text{Average duration} = \frac{\text{Total duration of site (or page) X}}{\text{Total number of page views}}$$

It is impossible to suggest an ideal value for this metric. For entire sites, the value should usually be as high as possible. For individual pages, it depends on the nature of that page. A navigation page should have a low value, which means that users easily find their way, while content pages should have a higher value.

3.2 Specific e-metrics

Apart from a vast amount of possible general e-metrics, there is also a great number of metrics specifically for e-commerce sites. These include:

Personalization index: This expresses to what extent data that were asked from the user to fill in on a form, are used to offer a personalized service. This value should be greater than 0.75.

Acquisition cost: This divides promotion costs (in the form of banners) by the number of click-throughs so that the marketing team can discover to what extent the marketing efforts are effective to acquire users.

Cost per conversion: This divides the promotion costs by the number of sales. It is the number that marketing people use to determine the best investment of their promotional budget.

RMF-analysis: This is a special analysis in which customers are evaluated on three aspects: Recency (when was the last time they purchased something?), Monetary Value (how much money has he spent on our site already?) and Frequency (how frequent does the user purchase a good on our site?). All users are then placed in this three-dimensional model to find possible user segments.

An excellent white paper about e-metrics can be found on the Web site of Netgenesis [5].

4 The Data

The data used for research purposes in this work, were offered to us by a Belgian ISP, that chose to remain anonymous because of privacy issues. The data used here come from a proxy server that handles request of users that have a broadband internet connection.

4.1 The log file

To offer Internet access, this ISP makes use of six proxy servers, all located in Belgium. Every user is assigned to one of those. As long as this user doesn't change his installation settings, he will always log on to the same server. This means that the log history of every user can always be found on the same server. This is a clear advantage for data mining research.

Unfortunately, the ISP doesn't make use of fixed IP-addresses. Like most other ISP's, it uses a pool of dynamic IP-addresses, which makes it more complicated to identify users. Fortunately, *on-the-fly* IP-addresses—meaning that another address may be used for each file request—are not used here, since it would make this research quite impossible.

At the end of a session, the user can explicitly release his IP-address so that it can be used for another user. However, most people don't do this and turn off their computer without doing so. In this case, their address is automatically taken back after 30 minutes, which is convenient because it is also the time interval that will be used to distinguish sessions with the same IP-address.

Luckily, there are no time zones in Belgium, because if this was so, it would have to be solved somehow. Also, the server in this research is not load-balanced. This, too, would make matters

more complicated. As mentioned before, IPs that change within sessions would make it very difficult, if not impossible, to conduct a research. All these problems could arise with other data sets. In the framework of this article, there has not been looked for solutions for these problems. It may be an opportunity for future research.

4.2 Preprocessing

Before any actual data mining algorithms can be applied on the data, the data needs to be preprocessed so that it can serve as input of several algorithms. As mentioned in the introduction, most tools that were designed for conventional Web Usage Mining on Web servers, perform this preprocessing automatically. In this case, however, this part of the process will have to be executed manually. The advantage of this is that we have more control over the way it is done.

4.2.1 Data Cleaning

First of all, all files with an extension .jpg, .gif, .wav, etc. have been removed from the log file, an action which drastically reduced the size of the file. Secondly, all irrelevant fields were removed. Only three fields were retained: IP-address, time stamp and URL. Finally, only the DNS-name in the URL's was kept, in the form of *www.websitename.com* (other prefixes and suffixes are possible instead of www and com). This was done in order to facilitate the comparison of different lines in the database, and because the exact pages that were visited are irrelevant in this research. After this process of data cleaning, the original log file that comprised 600MB was reduced to 185MB. The file contains 1,714,813 hits.

4.2.2 User and session identification

The only data that are at our disposal to identify users and sessions, are the time stamp and the IP-address. To start with, every IP-address is seen as a different user. However, a different user can represent several sessions. Within the series of lines that are linked to a certain IP-address, different sessions are identified by using a *time out* of thirty minutes. This means that if there is a time interval of more than thirty minutes between two consecutive clicks, two sessions will be identified. The problem that we face by doing this, is that it is impossible to know whether it was the same person who simply didn't touch his computer for half an hour, or effectively a different user that was assigned an IP-address because the previous user turned of his computer. However, even if it was the same person, it is still useful to regard his click stream as two (or more) different sessions.

A problem here is that it is impossible to correctly identify the number of unique users, which is needed to calculate stickiness. However, the most important aspect of stickiness will be to compare the results of different sectors with each other. So, if the absolute values of stickiness are somewhat incorrect, that will be be crucial. They should be seen relatively to one another.

Our file contained 7,287 different IP-addresses and 10,548 sessions were identified.

5 Research

To give an illustration of what is possible by examining ISP log files, we have decided to examine browsing behaviour within a certain sector. The sector that was chosen to do this, is a collection of several sub-sectors. In this sector, five different sub-sectors are distinguished: Newspapers, Banks, Finance, Television and Radio. By *Finance*, we mean Web sites on which financial content (about certain stocks and shares) can be found. Altogether this is a list of fifty Web sites: 7 newspaper sites, 9 bank sites, 17 financial sites, 13 television sites and 4 radio sites. The Web sites are listed per sector in figure 2. Only sites that produced a minimum of 120 hits in the original file, were admitted in this list. If Web sites with a lower amount of total hits than 120 would be admitted,

this would probably have a negative influence on the reliability of the results. The reasons why this sector was chosen are firstly, because it is diverse and secondly, because it is frequently visited.

Financial	Newspaper	Bank	TV
biz.yahoo.com	www.destandaard.be	netbanking.dexia.be	www.bbc.co.uk
cure.vms-keytrade.com:443	www.gva.be	w.be.fortisbanking.com	www.bigbrother.be
dynamic.nasdaq.com	www.hbvl.be	www.aslk.be	www.canvas.be
dynamic.nasdaq-amex.com	www.standaard.be	www.bacob.be	www.cnn.com
finance.yahoo.com	www.tijd.be	www.bbl.be	www.kanaal2.be
forum.beleggers.net	www.vacature.be	www.cera.be	www.ketnet.be
het.beleggers.net	www.vacature.com	www.dexia.be	www.mtv.com
quote.yahoo.com		www.fortisbank.be	www.mtve.com
quotes.nasdaq.com		www.kbc.be	www.tmf.be
quotes.nasdaq-amex.com			www.tv1.be
www.aex.nl	**Radio**		www.vrt.be
www.beurs.be	www.donna.be		www.vt4.be
www.easdaq.be	www.radio2.be		www.vtm.be
www.easdaq.com	www.radiocontact.be		
www.nasdaq.com	www.topradio.be		
www.vms-keytrade.com			
www.wallstreetweb.nl			

Figure 2: Web sites per sector

Several data mining techniques will be used to discover interesting browsing patterns within this sector. The actual data mining techniques are association rules, sequence analysis and clustering. Apart from those we will start with a visualization technique and end with a few e-metrics that will prove to be highly interesting.

Some of those techniques required some additional preprocessing. Therefore, we have created a database in which each line represents a session. For each of the 50 Web sites, a column was created. Every cell in this database accordingly expresses whether or not a specific Web site was visited in a specific session. The other file (out of which the new file has been extracted and which will be used for sequence analysis) lists a sequence of hits in which a certain Web site may appear for a number of times, with only the time stamp changing.

5.1 Visualization

Several data mining tools offer the possibility to represent data in a visual manner. To get a visual indication of strong relationships between several Web sites, this is a very good method. In this work, we have made use of the visualization node in SPSS Clementine, the so-called web node. It simply counts the number of times two Web sites appear together in the sessions. Then, it generates a graphic that draws links between all the sites that appear together. The more two sites appear together, the thicker the line between both. The overall picture looks like figure 3.

Not all the links are shown however. This is merely to lay the emphasis on the strongest links.

A number of these links can be identified as being not relevant. Visitors of `www.aslk.be` and `www.cera.be` are automatically redirected to respectively `www.fortisbank.com` and `www.kbc.be` because the two first banks have merged with the latter. Also the links between several Yahoo! sites and several Nasdaq sites are predictable. A way of deleting these links is to replace one with the other at the data cleansing stage. However, in the section about sequence analysis we will see that some users also try to browse in the direction opposite of the redirect, which can be interesting. For example, some users want to visit CERA after having visited KBC.

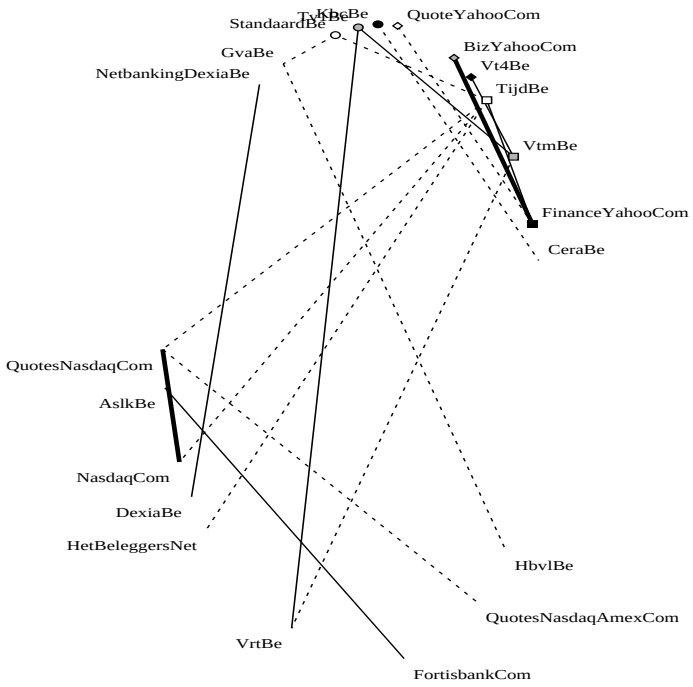

Figure 3: The overall picture

Apart from those, some other links are indeed quite interesting. There are quite strong links between the Web sites of VTM, TV1, VRT and VT4. All of those are television stations. Apparently, visitors often visit several TV-sites together. The same goes for the newspaper sector. `www.gva.be` is frequently visited together with `www.standaard.be` and `www.hbvl.be`. `Www.standaard.be` also shows a link with `www.tijd.be`. This last newspaper is a special case. It has more links with financial Web sites than with newspapers. The reason for this is that this newspaper (De *Financieel Economische Tijd* is comparable to the Financial Times) mainly offers financial news. Nevertheless, this is quite interesting information for this Web site. It should consider financial sites such as Nasdaq and Easdaq as its competitors, rather than newspaper sites. Figure 4 focuses on the links of this individual Web site and makes the observation even clearer.

As a conclusion, it can be stated that the visualization technique is a handy way of quickly discerning strong links between Web sites.

5.2 Association rules

In this context, association rule algorithms will discover which Web sites are frequently visited together. In the previous paragraph, an indication of this has already been offered by merely counting the number of times sites were visited together. These algorithms however, will generate rules such as the one in 2.3.2 with a level of support and confidence. Two algorithms were used: *Generalized Rule Induction (GRI)* and the APRIORI algorithm. Figure 5 shows the results of the

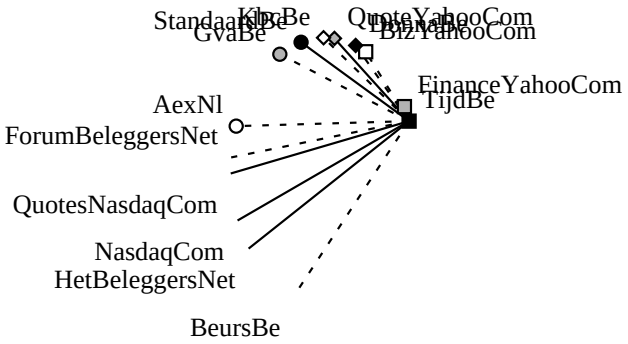

Figure 4: Financieel Economische Tijd

GRI algorithm. The results of APRIORI were mostly the same.

First of all, some obvious and less relevant rules were identified. Those are identical to the links found by the web node in the previous paragraph, so they will not be explained further.

Contrary to the results of the web node, the association rule algorithms didn't find a link between *De Tijd* and other newspaper Web sites. There are a number of rules, though, in which *De Tijd* appears with financial sites. The suspicion that *De Tijd* should be compared to financial sites rather than to newspaper sites, which has already been mentioned before, is hereby strongly confirmed.

Another important conclusion that can be drawn from these results is that they are all situated within a single subsector, with the exception of the radio sector that mingles with the TV sector. It should be noticed that a great number of the rules found deal with the TV and radio sector. Apparently it occurs very frequently that users visit several TV Web sites in one session.

5.3 Sequence analysis

The algorithm that was used to perform this analysis is the Capri algorithm, developed by MINEit Software. This algorithm is similar to the APRIORI algorithm but it has the extra feature of integrating the time variable into the analysis. It will therefore be able to detect which Web sites usually occur in the beginning or at the end of a session.

The syntax of a rule found by this algorithm is slightly different from the association rules. An example here may be:

3 (2, 12, 2.23, 92.31) "www.cera.be" , "www.kbc.be"

The sequence number is 3 and it is 2 items long. The sequence occurred 12 times, which represents 2.23% of the data set. In these 12 instances Cera was followed in 92.31% of the cases by KBC. (see also [4]) The sequences in figure 6 should be interpreted accordingly.

A first remark is that all of the items in a sequence usually belong to the same subsector[1]. This is the third time that this conclusion is made.

A second remark is that a lot of the sequences that consist only out of two items, are present in both directions. When this is the case, it is important to identify the strongest sequence (The

[1]It is here assumed that the Web site of De Tijd is indeed a financial site, rather than a newspaper site.

Minimal confidence 50
Minimal support 0
Maximal number of rules 50
Maximal number of rule conditions 5

NetbankingDexiaBe ⇐ DexiaBe (30:5.576%, 0.97)
DexiaBe ⇐ NetbankingDexiaBe (33:6.134%, 0.88)
QuotesNasdaqCom ⇐ NasdaqCom (53:9.851%, 0.75)
NasdaqCom ⇐ QuotesNasdaqCom (58:10.781%, 0.69)
FinanceYahooCom ⇐ BizYahooCom (51:9.48%, 0.84)
AslkBe ⇐ FortisbankCom (20:3.717%, 0.9)
FortisbankCom ⇐ AslkBe (23:4.275%, 0.78)
Tv1Be ⇐ VrtBe (32:5.948%, 0.75)
BizYahooCom ⇐ FinanceYahooCom (98:18.216%, 0.44)
VrtBe ⇐ Tv1Be (47:8.736%, 0.51)
KbcBe ⇐ CeraBe (13:2.416%, 0.92)
CeraBe ⇐ KbcBe (31:5.762%, 0.39)
Tv1Be ⇐ VtmBe & VrtBe (11:2.045%, 0.91)
VrtBe ⇐ VtmBe & Tv1Be (15:2.788%, 0.67)
GvaBe ⇐ HbvlBe (19:3.532%, 0.68)
VtmBe ⇐ Vt4Be (43:7.993%, 0.4)
Vt4Be ⇐ VtmBe (46:8.55%, 0.37)
HbvlBe ⇐ GvaBe (55:10.223%, 0.24)
CanvasBe ⇐ Vt4Be & Tv1Be (10:1.859%, 0.5)
Tv1Be ⇐ Vt4Be & CanvasBe (5:0.929%, 1.0)
VrtBe ⇐ VtmBe & Vt4Be & Tv1Be (4:0.743%, 1.0)
FinanceYahooCom ⇐ TijdBe & BizYahooCom (9:1.673%, 0.89)
NasdaqCom ⇐ TijdBe & QuotesNasdaqCom (13:2.416%, 0.62)
MtvCom ⇐ MtveCom (4:0.743%, 0.75)
Tv1Be ⇐ VtmBe (46:8.55%, 0.33)
TopradioBe ⇐ VtmBe & Kanaal2Be (7:1.301%, 0.57)
ForumBeleggersNet ⇐ QuoteYahooCom & HetBeleggersNet (3:0.558%, 1.0)
MtveCom ⇐ MtvCom (8:1.487%, 0.38)
CanvasBe ⇐ VtmBe & Vt4Be & VrtBe (4:0.743%, 0.75)
CanvasBe ⇐ VtmBe & Vt4Be & Tv1Be & VrtBe (4:0.743%, 0.75)
CanvasBe ⇐ VtmBe & Vt4Be & Tv1Be (4:0.743%, 0.75)
QuotesNasdaqCom ⇐ TijdBe & NasdaqCom (14:2.602%, 0.57)
KetnetBe ⇐ VtmBe & Tv1Be (15:2.788%, 0.27)
GvaBe ⇐ StandaardBe (41:7.621%, 0.34)
QuotesNasdaqAmexCom ⇐ QuotesNasdaqCom & NasdaqCom (40:7.435%, 0.23)
StandaardBe ⇐ GvaBe (55:10.223%, 0.25)
Kanaal2Be ⇐ TopradioBe (15:2.788%, 0.33)
TopradioBe ⇐ Kanaal2Be (17:3.16%, 0.29)
VrtBe ⇐ VtmBe (46:8.55%, 0.24)
Kanaal2Be ⇐ VtmBe & TopradioBe (9:1.673%, 0.44)

Figure 5: Results GRI algorithm

1 (2, 7, 1.30, 63.64) dynamic.nasdaq.com , quotes.nasdaq.com
2 (2, 9, 1.67, 81.82) dynamic.nasdaq.com , www.nasdaq.com
3 (2, 12, 2.23, 92.31) www.cera.be , www.kbc.be
4 (2, 9, 1.67, 32.14) het.beleggers.net , www.tijd.be
5 (2, 6, 1.11, 35.29) www.aex.nl , www.tijd.be
6 (2, 6, 1.11, 35.29) www.aex.nl , finance.yahoo.com
7 (2, 6, 1.11, 11.54) biz.yahoo.com , quote.yahoo.com
8 (2, 42, 7.79, 80.77) biz.yahoo.com , finance.yahoo.com
9 (2, 8, 1.48, 14.55) www.gva.be , www.hbvl.be
10 (2, 8, 1.48, 14.55) www.gva.be , www.tijd.be
11 (2, 7, 1.30, 12.73) www.gva.be , www.standaard.be
12 (2, 6, 1.11, 18.75) www.vrt.be , www.tv1.be
13 (2, 6, 1.11, 12.77) www.tv1.be , www.kanaal2.be
14 (2, 7, 1.30, 14.89) www.tv1.be , www.vt4.be
15 (2, 10, 1.86, 21.28) www.tv1.be , www.vtm.be
16 (2, 6, 1.11, 27.27) forum.beleggers.net , quote.yahoo.com
17 (2, 6, 1.11, 27.27) forum.beleggers.net , het.beleggers.net
18 (2, 8, 1.48, 36.36) forum.beleggers.net , finance.yahoo.com
19 (2, 6, 1.11, 6.00) finance.yahoo.com , www.aex.nl
20 (2, 9, 1.67, 9.00) finance.yahoo.com , quote.yahoo.com
21 (2, 6, 1.11, 16.67) quote.yahoo.com , biz.yahoo.com
22 (2, 6, 1.11, 16.67) quote.yahoo.com , www.tijd.be
23 (2, 9, 1.67, 25.00) quote.yahoo.com , finance.yahoo.com
24 (2, 12, 2.23, 38.71) www.kbc.be , www.cera.be
25 (2, 7, 1.30, 22.58) www.kbc.be , www.tijd.be
26 (2, 13, 2.41, 30.23) www.vt4.be , www.vtm.be
27 (2, 6, 1.11, 13.33) www.vtm.be , www.bigbrother.be
28 (2, 6, 1.11, 13.33) www.vtm.be , www.topradio.be
29 (2, 11, 2.04, 24.44) www.vtm.be , www.vt4.be
30 (2, 11, 2.04, 57.89) www.hbvl.be , www.gva.be
31 (2, 6, 1.11, 35.29) www.kanaal2.be , www.vtm.be
32 (2, 7, 1.30, 46.67) www.topradio.be , www.vtm.be
33 (2, 6, 1.11, 17.65) www.donna.be , www.tijd.be
34 (2, 9, 1.67, 21.95) www.standaard.be , www.tijd.be
35 (2, 10, 1.86, 24.39) www.standaard.be , www.gva.be
36 (2, 6, 1.11, 5.77) www.tijd.be , quote.yahoo.com
37 (2, 6, 1.11, 5.77) www.tijd.be , www.beurs.be
38 (2, 6, 1.11, 5.77) www.tijd.be , www.standaard.be
39 (2, 17, 3.15, 85.00) www.fortisbank.com , www.aslk.be
40 (2, 15, 2.78, 65.22) www.aslk.be , www.fortisbank.com
41 (2, 29, 5.38, 96.67) www.dexia.be , netbanking.dexia.be
42 (2, 26, 4.82, 78.79) netbanking.dexia.be , www.dexia.be
43 (3, 8, 1.48, 80.00) quotes.nasdaq-amex.com , quotes.nasdaq.com, www.nasdaq.com
44 (3, 7, 1.30, 87.50) quotes.nasdaq-amex.com , www.nasdaq.com , quotes.nasdaq.com
45 (3, 7, 1.30, 87.50) www.nasdaq.com , quotes.nasdaq-amex.com , quotes.nasdaq.com
46 (3, 6, 1.11, 66.67) quotes.nasdaq.com , quotes.nasdaq-amex.com , www.nasdaq.com
47 (3, 7, 1.30, 18.92) quotes.nasdaq.com , www.nasdaq.com , www.tijd.be
48 (3, 7, 1.30, 18.92) quotes.nasdaq.com , www.nasdaq.com , quotes.nasdaq-amex.com
49 (3, 7, 1.30, 17.07) finance.yahoo.com , biz.yahoo.com , www.tijd.be
50 (3, 7, 1.30, 87.50) www.vtm.be , www.tv1.be , www.vrt.be
51 (3, 6, 1.11, 66.67) www.tijd.be , quotes.nasdaq.com , www.nasdaq.com
52 (3, 6, 1.11, 46.15) www.tijd.be , finance.yahoo.com , biz.yahoo.com

Figure 6: Results Capri algorithm

one with the highest support and the highest confidence).

In the paragraph about visualization, it was argued that a few strong links exist merely because of *redirects* from old Web sites that don't exist anymore, to their new address. This analysis confirms this observation. Let's take the example of `www.cera.be` and `www.kbc.be`. There are two sequences that contain those two sites, namely sequences number 3 and number 24, both containing 12 instances. However, the first sequence (from Cera to KBC) has a confidence level of 92.31% against 38.71% for the opposite sequence. Apparently, a number of people that have already visited `www.kbc.be` still try to go to `www.cera.be`, even though the site doesn't exist any longer. The same reasoning can be made for the strong sequence from `biz.yahoo.com` to `finance.yahoo.com` (number 8) since there exists an automatic redirect from the former to the latter.

The Dutch Web site of AEX usually comes first in a sequence (see sequence number 5 and 6). Only once it appears at the end (number 19), but this sequence has such a low level of support and confidence, that it can be neglected.

Also in the TV sector, a few observations can be made. When TV1 is visited with one of its direct competitors, it is usually visited first. It should be noted, however, that the levels of support and confidence are relatively low. The cache server that placed these data at our disposal, could perform the same research with more data to verify the reliability of these results.

Next to the TV sector, there are also some interesting sequences in the newspaper sector, mainly between `www.gva.be`, `www.standaard.be` and `www.hbvl.be`. Interesting here is that there are sequences between *Gazet van Antwerpen* en *Het Belang Van Limburg* at one hand and between *Gazet van Antwerpen* and *De Standaard* on the other hand, but not between *De Standaard* and *Het Belang Van Limburg*. Possibly, the readers of both newspapers differ too much from one another while the readers of *Gazet van Antwerpen* are a bit in between the other two groups.

In the previous section, the Web site of *De Tijd* received some special attention because of its special nature[2]. Here, we find De Tijd in 10 different sequences, of which we don't take the last three (36–38) into account because of their low level of support and confidence. In the other seven sequences, De Tijd appears invariably at the end of the sequence. It may be interesting for the management of this Web site to know that visitors who are looking for financial information usually end their session by visiting this site.

We can conclude this paragraph by stating that a sequence analysis can result in interesting observations, especially when combined with the results of an association rule algorithm. Even though the results suffered a bit from low levels of support, we were nevertheless able to distinguish a number of interesting trends.

5.4 Clustering

The fourth technique is the one that tries to group sessions into clusters in such a way that the clusters have as much intra-cluster similarities, and as little inter-cluster similarities as possible. To do this, we made use of the Kohonen algorithm in SPSS Clementine. The best results were generated when both the X and the Y axis were set to 3, so that the algorithm identified 9 clusters.

The file that was used as an input for the algorithm, is the same file that was used for the visualization technique and the association rule algorithms. Based on their X and Y coordinates, the results were divided in nine tables to get an overview of the clusters. After that, the tables were transformed into tables that only indicate for each session how many Web sites of a certain subsector were visited. For each cluster, column totals were calculated to easily see how many times a certain sector was visited in a cluster. This made it possible to compare each cluster with the others. Because a full list of each cluster is too extensive, we have included in figure 7 the totals of each cluster.

We will now briefly describe each cluster (not in the same order as in figure 7 though). Cluster 0–2 is quite large and the emphasis is evidently on the TV sector. Hits in other sectors are merely

[2]It is a newspaper, but the Web site has more similarities with financial Web sites.

Cluster	Financial	Bank	TV	Radio	News	# of sessions
0–0	30	143	25	5	93	157
0–1	1	5	25	18	9	32
0–2	1	3	223	23	8	108
1–0	13	0	1	0	0	7
1–1	23	0	0	0	0	11
1–2	9	2	6	1	3	9
2–0	164	1	9	2	17	80
2–1	38	3	1	0	19	19
2–2	170	14	9	8	100	114

Figure 7: Results Clustering

present because some users visit other sectors apart from this one. *We will therefore call this cluster the TV cluster.*

Cluster 0–1 is a lot smaller than the previous cluster but is also very much focused on the TV sector. However, in contrast to 0–2, this cluster has a relatively high number of radio hits since the radio sector is by far the smallest in the original file. Here, however, it is almost as important as the TV sector. *We will therefore call this cluster the Radio cluster.* The reason for the large amount of TV-hits in this cluster is that many radio sites are visited in sessions in which also one or more TV stations are visited.

Cluster 1–0 is very limited in number of sessions and is completely focused on financial sites. If we take a look at the full table of this cluster, we can see that this cluster groups sessions in which the emphasis lays on `biz.yahoo.com` and `quotes.yahoo.com`. Each session contains at least one of both and 4 out of the 7 sessions even contain both. Cluster 1–1 is similar to this cluster and is concentrated on the financial Web sites `forum.beleggers.net` and `het.beleggers.net`. 10 out of 11 sessions contain both Web sites.

Cluster 2–0 is again quite large (80 sessions). It is very clear that this cluster is a cluster of sessions that were focused on financial Web sites. Apart from visiting several financial sites, some users also visited a Web site from another sector, which explains the other hits. *We can safely call this cluster the financial cluster.*

Cluster 2–1 is very special in the sense that it groups 19 sessions that all contain `www.tijd.be` and `finance.yahoo.com` (with only one exception). Apparently the relationship between those two sites is so strong that the algorithm found a separate cluster for them.

Cluster 2–2 is related to the previous one because here the Web site of *De Tijd* plays once more a special role. Each of the 114 session in this cluster contains this Web site, together with one or more financial sites. Clearly, the Kohonen algorithm has divided the sessions in which *De Tijd* is visited with another financial site into two separate clusters: one in which it occurs together with Yahoo! and one in which it occurs with other financial Web sites.

The reason why cluster 1–2 was created is not very clear. It contains 9 sessions in which all subsector are more or less proportionally represented. It is probably not even a real cluster, but a small group of sessions that the Kohonen algorithm didn't assign to any cluster.

Cluster 0–0 is the largest cluster and contains both a lot of bank-hits and newspaper-hits. While analyzing the full table of this cluster, we were able to discover that hardly any of these sessions contain Web sites of both sectors. It is therefore slightly strange that the algorithm hasn't created two clusters instead of one since there is hardly a link between the two sectors that are so strongly represented in this cluster.

Even though the last cluster groups sessions that were focused on newspapers, it doesn't contain even once the Web site of *De Tijd*, which confirms the previously made conclusion that this site is a financial, rather than a newspaper site.

It is remarkable that each cluster is focused on a specific subsector. With the exception of the

bank sector. Cluster 0–0 indeed groups sessions that contained two or more bank sites, but if we take a look at the details of these sessions, it becomes apparent that they were either generated by *redirects* from one bank site to another, or because a user wanted to perform some on-line transactions and entered a specific Web site for that purpose after having visited the general Web site of his bank. *We can conclude that several competing banks are virtually never visited together in a site.* Users stick to their own bank.

5.5 E-metrics

The emphasis in this section will be on the measurement of *stickiness* and *Average duration*. By computing both measures for each Web site, we will be able to give an indication of user behaviour on that particular Web site. Then, we will also find out whether or not there are differences in behaviour among the subsectors.

Stickiness was calculated by dividing the total number of seconds that a Web site has been visited by the total number of users. Average duration was calculated by dividing the total number of seconds that a Web site has been visited by the total number of page views of that site. After having computed these measures for each individual Web site, we calculated the average value for each sector. The results of this can be found in table 1.

Sector	**Stickiness**	**Average duration**
Newspaper	3' 9"	10.67 sec.
Financial	5' 16"	6.18 sec.
Bank	6' 01"	13.01 sec.
Television	2' 10"	7.91 sec.
Radio	2' 21"	5.78 sec.
Overall avg.	*4' 10"*	*8.66 sec.*

Table 1: Average Stickiness (in minutes and seconds) and Average duration per sector (in seconds)

As explained before, stickiness is a measure that indicates to what extent a Web site (or a whole sector) is able to keep the attention of its users. The higher the value, the better. Average duration is the average amount of time with which a page on that Web site is viewed. Not each page should have a high average duration, but we can safely assume that an overall average should be as high as possible as well.

To obtain a better overview of the meaning of these figures, a graphic (figure 8) was constructed that has Stickiness on its X-axis and Average Duration on its Y-axis.

The bank sector scores best on both e-metrics. One reason for this may be that among these Web sites there are a few with which clients can perform on-line transactions. Since users always do this with only one bank and because they take their time in doing so, both stickiness and average duration are high.

Newspapers have a high average duration but a relatively low stickiness. The results of financial sites are the opposite: a high stickiness but a low average duration. The differences in average duration are easy to explain. While a user will take his time to read an article on a newspaper site, on a financial site he will very often simply request a page to see the changes concerning certain shares, which takes very little time. On the other hand, a financial site can hold the attention of its users quite well. Even though they watch each page only for a limited amount of time, they stick to the site for a relatively long time. This is not so in the newspaper sector. Here, users watch a page for a long time, but in the end they only watch a few pages and leave early.

Both the TV and radio sector show low scores on both measures. Uses don't stay long on one site and quickly jump from one to the other. This was also shown in the great number of association rules in this sector.

It is useless for an individual Web site to compare itself with overall averages. It should take the averages of its own sector into account when evaluating its performance. Unfortunately, these

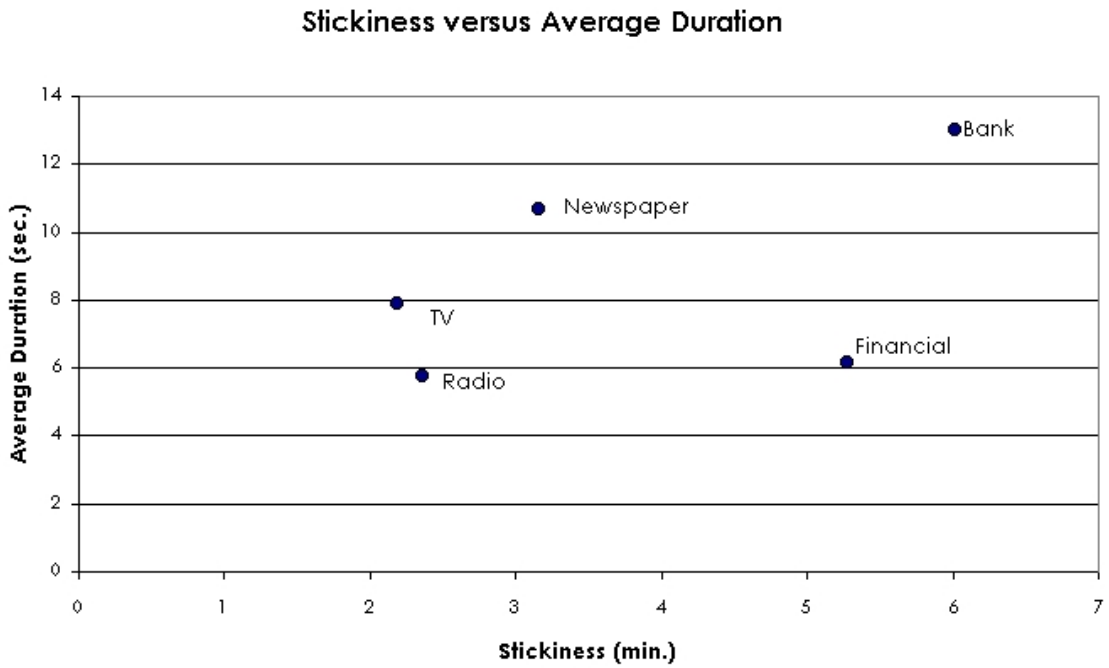

Figure 8: Stickiness — Average Duration

averages are not readily available and therefore, Web sites can usually only evaluate themselves by examining their e-metrics throughout time.

Knowing that previous data mining methods showed that the Web site of *De Tijd* is usually visited together with financial sites instead of newspapers, it would be interesting to know which sector its browsing behaviour resembles most. Its stickiness is 3.01 minutes and its average duration is 8.86 seconds. Remarkably, this resembles much more the averages of newspapers than those of financial Web sites. The conclusion we can make for *De Tijd* is: *It is almost invariably visited together with financial Web sites but shows the browsing behaviour of a newspaper Web site.*

6 Further research

Even though section 5 showed some of the possibilities of Web Usage Mining on proxy servers and offered some very interesting conclusions, much more is possible if we dispose over the necessary data.

First of all, the ISP that owns this data, could do this kind of research on a regular basis to investigate possible changes over time. The data that was researched here, was spread only over a time period of three hours. It was therefore impossible to search for any evolutions in time.

Proper user identification was very difficult in this work. However, this could be made much easier if these data are linked to customer data. Each ISP knows which customer used which IP-address at each given time. Using these data, users can be perfectly identified. Unfortunately, because of privacy matters, it was impossible for us to obtain these data.

Most ISP also have a more or less extensive database with socio-demographic data about their customers, for example the address of the customer, profession, age, number of family members, etc. This can add a whole new dimension to this research. We could then define classes of users based on one or more of these socio-demographic data and analyze differences in behaviour between those classes. We would also be able to find out which kind of users visit certain Web sites or sectors.

Another way of using an additional demographical attribute in the data without requiring access to the user information is using a specific set of IP numbers for certain dial-in numbers. This way, we can compare different areas (for example provinces) with one another. T-online in Germany is an ISP that uses this technique.

7 Conclusion

Very little research has been done about the possibilities of Web Usage Mining on proxy servers (or cache servers). The intention of this work was to give an indication of what kind of information can be extracted from the log files of these servers. Using several data mining and other techniques, we have been able to draw a number of conclusions that could not have been found on another level of data collection.

Every technique showed that it is useful to make a distinction between several sectors. Users who visit several Web sites in a single session, very often visit Web sites that can be considered direct competitors. This statement is especially applicable in the TV, radio, newspaper and financial sector. Not in the bank sector though. Users are only interested in the Web site of their own bank.

Despite the clear distinction between sectors, the sectors that the data mining techniques found, were slightly different from the ones that we defined before the actual research. First of all, it could be said that the TV and radio sectors are in fact only one sector. There is hardly a difference between the browsing behaviour of users who visit these sites. Also, the association rule algorithms made no difference between them. Only the clustering algorithm made a slight distinction. It created one cluster that was completely focused on the TV sector, and another one that contained a relatively high number of radio sites. Yet, even here the distinction was not very obvious.

Secondly, the Web site of *De Financieel Economische Tijd* was allocated at first to the newspaper sector. All techniques except e-metrics showed that this site should be considered to be a financial rather than a newspaper site. Nonetheless, the e-metrics (stickiness and average duration) demonstrated that user behaviour on this Web site strongly resembles that of other newspaper sites.

The combination of association rules and sequence analysis led to a number of interesting conclusions for specific Web sites. Knowing which sites are usually visited together with your own (and therefore are competitors) and whether the visit to your site comes first or last, can be interesting information for the designers of a specific Web site and is information that can only be deducted from this kind of log files.

Also, the comparison of sectors by means of e-metrics proved to be very useful. There are clear differences in browsing behaviour between several sectors.

We can conclude that proxy servers contain much information that can be of considerable interest to specific Web sites. Unfortunately, the administrators of these Web sites have no access to this information. Moreover, ISP's have little personal interest in this information and are therefore not very inclined to perform such (expensive) research. We feel that there must be a way so that individual companies can obtain this information from an ISP in such a way that the privacy of the customer is not violated. We hope that this paper may be an inducement for further discussion about this issue.

References

[1] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Discovery of interesting usage patterns from web data. Technisch rapport TR 99-022, University of Minnesota, 1999.

[2] R. Dilly. Data mining - an introduction. Student notes, December 1995.

[3] M. Deshpande en P-N. Tan J. Srivastava, R. Cooley. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12, Januari 1999.

[4] MineIt Software Ltd. Capri user guide 1.0. Handleiding, 2000.

[5] NetGenesis. E-metrics - business metrics for the new economy. White paper, 2000. www.netgenesis.com.

[6] R. Rao P. Pirolli, J. Pitkow. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*, 1996.

[7] J. Pitkow and L. Catledge. Characterizing browsing strategies in the world wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, April 1995.

[8] B. Mobasher R. Cooley, J. Srivastava. Web mining: Information and pattern discovery on the world wide web. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, 1997.

[9] Bamshad Mobasher en Jaideep Srivastava Robert Cooley. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.

[10] Myra Spiliopoulou and Lukas C. Faulstich. WUM: a Web Utilization Miner. In *Workshop on the Web and Data Bases (WebDB98)*, pages 109–115, 1998.