

Data Mining of Web Access Logs From an Academic Web Site

Vic Ciesielski and Anand Lalani
Department of Computer Science and Information Technology
RMIT University
{vc,alalani}@cs.rmit.edu.au

Abstract.

We have used a general purpose data mining tool to determine whether we can find any 'golden nuggets' in the web access logs of a large academic web site. Our goal was to use general purpose data mining algorithms to analyse visitors to the website and somehow characterise or distinguish them in some way. We used two web access logs, one from 2001 and one from 2003. We extracted 4 different feature sets from the web logs and used algorithms for classification (1R, J48/C4.5), clustering (EM), association finding (apriori) and feature selection (correlation based subset evaluation with best first search). We discovered several nuggets, the most significant being that a major difference between visitors from within Australia and visitors from outside Australia is that visitors from outside Australia generally arrive via search engines and are interested in information about postgraduate courses.

1 Introduction

Data mining is often described as the process of finding "golden nuggets" of information in vast quantities of data. Large web sites which record accesses by visitors can generate very large web access logs. It is only natural to ask whether data mining techniques can be used to find any golden nuggets in the web access logs and quite a number of researchers have done this, [1, 2, 3, 4, 5, 7], for example. Some of the major problems associated with extracting user transactions from web log files are described in [1] and a number of heuristics are suggested. Modification and personalization of web sites based on data mining of web logs have been investigated by [2, 5, 7]. Experiments in which keywords on the web pages visited are integrated with the web access data in the search for visitor access patterns are described in [3]. Some investigations have attempted to find new algorithms that will work effectively on the basic web log data or vistration graph [4, 6] while other investigations have been concerned with ways of doing effective transaction and feature extraction for subsequent use in standard data mining algorithms [2, 9]. Our work falls into the second category.

The computer science web site at RMIT currently contains over 5,500 files and services an average of 170,000 requests per day. Around 800MB of web access data are generated in a month. A wide range of information is provided. Some pages provide general information about the department and the various programs and course offered. Some information, such as programs and courses, is geared towards prospective students, other information, such as timetables, exam results and assignment details, is geared towards current students.

1.1 Goals

We have two major goals: (1) We would like to determine whether we can find any golden nuggets in web access log data using a number of existing data mining techniques, rather than developing special purpose algorithms for web logs. (2) We would like to analyse the visitors to the site and determine whether there are any significant patterns.

We have focused on visitor data that is readily available from the web logs and focused on distinguishing (1) Visitors from within Australia and visitors from outside Australia (2) Visitors from within RMIT university and visitors from outside RMIT university (3) Visitors from within RMIT university and visitors from outside RMIT university but within Australia (4) Visitors from educational institutions other than RMIT university and other visitors.

We have used the WEKA package [8] which provides a number algorithms for classification, clustering, association finding and feature selection. For classification we have used the 1R and J48 algorithms. The 1R classifier [8, p78] builds a one rule classifier based on a single attribute. All attributes are tried and the one that gives the highest classification accuracy is chosen as the classifier. J48 is an implementation of the c4.5 decision tree algorithm [8, p269]. We have chosen these algorithms as the learnt classifiers are easily understood. In particular, if the difference in accuracy between the rule from 1R and the decision tree from J48 is small, then most of the accuracy is due to the single attribute. Such an attribute is a potential golden nugget. We have used the EM algorithm [8, p221] for clustering since it performs a reasonably extensive search and generates the number of clusters. We have used the apriori algorithm [8, p294] for association finding. For attribute selection we have chosen best first search through the subsets and correlation based evaluation of the subsets [8, p232]. In our experience this combination gives useful attributes with reasonable computation time.

2 Preprocessing and Feature Extraction

Figure 1 shows 2 entries extracted from a web access log. The fields are separated by spaces and are (1) the IP address of the visitor (2) the userid (only known if the user has been required to log in) (3) the date and time (4) the file requested (5) the protocol used (6) the status (basically success or failure) (7) The number of bytes sent (8) the referrer, that is the site from which the visitor came (9) The browser used. For our work we have used items 1,3,4,6 and 8.

Details of the web log data used for the experiments are shown in Table 1. For various reasons a considerable number of entries in the log files were not relevant to the goals of the data mining exercise and were removed. These included fetches of images (gif and jpeg files),

```
202.161.108.167 - - [01/Feb/2003:00:00:03 +1100] "GET
/timetables/city/2003s1/cc 4logo.gif HTTP/1.1" 206 14102
"http://www.cs.rmit.edu.au/timetables/city/2003s1/ cover.html"
"Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)"

213.183.13.65 - - [01/Feb/2003:00:00:16 +1100] "GET
/~winikoff/palm/dev.html HTTP/1.1" 302 244
"http://www.google.de/search?q=sources+onboardc+examples&ie=UTF-8
&oe=UTF-8&hl=de&meta=" "Scooter/3.3"
```

Figure 1: *Sample Web Access Log*

```
Host,Link0,Link1,Link2,Link2last,Linklast,Location
i-gate.abz.nl,/,/,/employment,/,/,/students,NotRMIT
vail.cs.ucsb.edu,/,/,/timetable,/,/timetable/city,/,/timetable,/,/NotRMIT
knu.cs.rmit.edu.au,/,/,/course,/,/course/pgrad,/,/course/pgrad/mit,/,/RMIT
csse.monash.edu.au,/,/,/staff,/,/,/general/contact/phone.html,/,/NotRMIT
```

Figure 2: A Sample of Instances Using First3-Last2 Feature Set

entries by proxy servers, entries by web crawlers, entries from IP numbers for which a host name could not be found and entries containing bad requests. It is important to note that this process involved a number of heuristics and while we are confident that we have removed most of these unwanted entries, we are aware that not all of the unwanted entries have been detected and removed.

2.1 Transaction Extraction

After the irrelevant entries were removed from the web log data, transactions were extracted based on the IP addresses. To identify transactions we used the heuristic suggested in [1]. All accesses from a single IP address separated by an access times of less than 30 minutes were considered to be part of the same transaction. The number of transactions extracted from each log file are shown in Table 1.

2.2 Feature Sets

The final step of preprocessing is to extract features from the transactions available. Four feature sets were extracted:

First3-Last2: In this feature set, an instance consists of the first 3 and the last 2 pages accessed by a visitor in a transaction. If fewer than 5 pages were visited in a transaction there will be missing values. This feature set is based on the conjecture that the first 3 and last 2 pages visited will contain information critical to distinguish classes of visitors. A sample of instances represented using this feature set is shown in Figure 2. The instances are represented in comma separated format. An attribute name in this feature set indicates the order of the pages visited. For example, the “link0” attribute corresponds to the first page visited in a transaction, “link1” the second, “link2last” the second last and “linklast” the last.

Web Access Log File	Number of Entries	Number of Transactions	Time Period
access2001	1000000	4591	29/05/2001 - 03/06/2001
access2003	11390257	55602	04/02/2003 - 23/04/2003

Table 1: Details of Web Access Log Files

First5-Last5: This feature set is the same as the First3-Last2 feature set except that in this feature set, an instance consists of the first 5 and the last 5 pages visited by the visitor in a transaction. This feature set is chosen to use more information in the data mining than the First5-Last2 feature set.

20-Most-Frequent-TF The web access log files were analysed to find the most frequently visited pages. The 20 most frequently visited pages were selected as attributes in this feature set. An attribute value in an instance that can be either “T” if that particular page was visited in that transaction or “F” otherwise. A sample of instances using, for brevity of presentation, a 6-Most-Frequent-TF feature set is shown in Figure 3. It was thought that it would be possible for categorising visitors browsing behaviours based on whether they visited a frequently visited page or not. Hence, this feature set was used. The question of whether infrequently visited pages carry useful discriminatory information is left for further work.

20-Most-Frequent-Time: This feature set has the same attributes as those in 20-Most-Frequent-TF. In this feature set, an attribute value is the amount of time spent in seconds by the visitor on that particular frequently visited page. The duration was calculated by taking the time difference between two consecutive page requests. A value of 0 was used if a page was not visited in a transaction and the last page visited is given a missing value for this attribute. This feature set is based on the conjecture that time spent on frequently visited pages might be a very distinguishing factor to categorize visitors.

3 Experiments

3.1 Experiment 1: AusVsOutsideAus2001

In this experiment our goal was to determine whether there were any differences in access patterns between visitors to the CS website from inside Australia and those from outside Australia. We prepared the 4 feature files described above from the access2001 web log in which the class variable was constructed from the IP address. If the IP address ended in .au then the class was set to *Aus*, otherwise it was set to *NotAus*. We then ran the 1R, J48, EM, apriori and attribute selection techniques from the WEKA package on these files.

3.1.1 Classification

We decided that the classification accuracy needed to exceed some threshold before the results could be considered as significant. An accuracy of 70% was chosen as this threshold. This

```
Host,/,/,course,/,student,/,timetable,/,course/pgrad,/,staff,Location
i-gate.abz.nl,F,F,T,F,F,F,NotAus
vail.cs.ucsb.edu,F,T,F,F,F,T,NotAus
knu.cs.rmit.edu.au,T,F,F,F,T,T,Aus
csse.monash.edu.au,T,T,T,F,T,F,Aus
```

Figure 3: A Sample of Instances Using 6-Most-Frequent-TF Feature Set

value was chosen as high accuracy can not be expected because of the known inaccuracies involved in the preprocessing. An accuracy of 50% on a two class problem can be achieved by guessing. An accuracy of 70% is a considerable improvement on guessing.

First3-Last2 and First5-Last5 The classification accuracy did not reach 70%. Analysis of the feature files revealed that each of the *linki* attributes had a large number of values. This makes it impossible for an algorithm like J48 to find significant groupings of attributes to guide the splitting process resulting in trees that are the results of random choices. Thus classification accuracy akin to guessing would be expected.

20-Most-Frequent-TF The result of the 1R algorithm is shown in Figure 4. The output can be interpreted as: If visitors visit the root page then they are from within Australia and if they do not then they are from outside Australia. Initially we were surprised by this result, however a little reflection suggested that this result is probably dominated by students of RMIT university who know the home page and probably have it set as their browser home page. Also, it suggests that international visitors arrive via search engines which direct them to specific pages found by a search rather than coming through the home page.

The decision tree generated by the J48 algorithm had an accuracy of 71%. The top level of the tree is shown in Figure 5. As can be seen from the figure, the home attribute has been chosen as the root of the tree indicating that it is the major discriminatory attribute. This is consistent with the 1R result. The small difference in accuracy between 1R and J48 indicates that the other attributes might not contribute much to the decision. The decision tree reveals one more pattern: If the visitors visit the root page and also visit pages related to postgraduate study then they are mostly from outside Australia. This can be considered as a golden nugget as the school is very keen to attract international postgraduate students and these visitors can be regarded as potential students seeking information about courses.

The decision tree also shows a pattern of access to a specific course page. This result is obtained because there may be some activity, such as an assignment, running for this course during the time period of this log file and hence, many students have accessed this page. This may be a temporary situation for this log file and this pattern is not expected in a log file for another time period.

20-Most-Frequent-Time: The 1R rule and the J48 decision tree were consistent with those obtained using the 20-Most-Frequent-TF feature set. The decision tree also indicated that if

```

/:
  T      -- Aus
  F      -- NotAus
(3226/4591 instances correct)
=== Summary ===
Correctly Classified Instances    3226    70.2755 %
Incorrectly Classified Instances  1365    29.7245 %
Total Number of Instances        4591

```

Figure 4: A Sample Output from the WEKA 1R Program in Experiment 1

visitors do not visit the root page and they visit the “/students” page then they are from within Australia. This result is probably dominated by the students of RMIT university.

3.1.2 Association Rules

First3-Last2 and First5-Last: The Apriori algorithm found only 2 rules. These are shown in Figure 6. The first rule can be interpreted as: If the first page visited is the root page then the visitor was from inside Australia. The second rule can be interpreted as: If visitors are from inside Australia then the first page they visited was the root page. This result is consistent with the classification results presented above.

20-Most-Frequent-TF Surprisingly association finding on this attribute set did not produce any interesting rules. This is because the number of visited pages was very much less than the number of pages which were not visited in a transaction and the apriori algorithm generated rules in which all the attribute values had “F” values, that is, associations of the kind “if the user did not visit page X then they did not visit page Y ”, which are not very interesting. Restricting the search to transactions which had a significant number of ‘T’ values gave too few transactions for meaningful rules to be generated.

20-Most-Frequent-Time: The apriori algorithm does not work with numeric attributes. Hence, this feature set was not used for association rule mining.

3.2 Clustering

The EM algorithm was applied to all of the feature sets to determine whether any interesting clusters were generated.

First3-Last2 and First5-Last5: The generated clusters had very small numbers of associated instances and allocation of instances to clusters appeared to be arbitrary. This is due to the same reason that classification was not successful on this data set, that is, that each attribute has a very large number of values and there are not many groupings of common values.

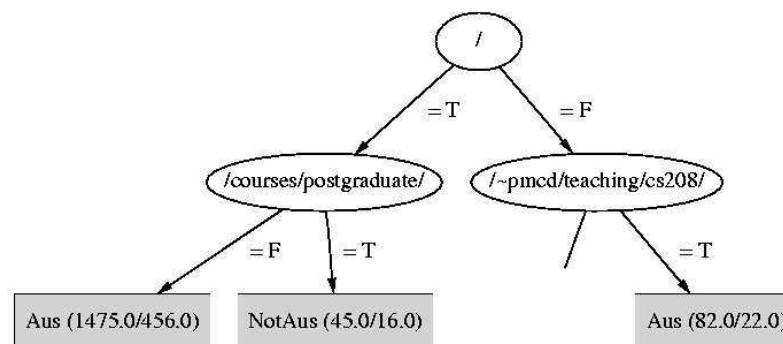


Figure 5: Partial Decision Tree from the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment1

```

Apriori
=====
Minimum support: 0.05
Minimum metric (confidence): 0.4

Best rules found:
1. link0=/ 339 ==> Location=Aus 301    conf:(0.89)
2. Location=Aus 531 ==> link0=/ 301    conf:(0.57)

```

Figure 6: *Partial Output from the WEKA Apriori Program Using the First3-Last2 Feature Set in Experiment1*

20-Most-Frequent-TF: The output of the EM algorithm revealed two interesting clusters. One cluster consisted of visitors who browsed pages related to staff members and their contact information. A sample output of this cluster is shown in Figure 7. As can be seen from this figure, this cluster of visitors tend to visit both pages “/staff/” and “/general/contact/phone.shtml” pages since their *Counts* for the “T” value is much higher than their *Counts* for “F” value. The output also shows that these 2 are the only pages that this cluster of visitors have mostly visited. Hence, it can be deduced that some visitors visit the RMIT computer science website to access information about staff members and their contact details. This can be considered as another nugget.

```

Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: log
Instances: 4591

                (F)    (T)
Attribute: /
Discrete Estimator. Counts = 73.29 12.71 (Total = 86)
Attribute: /students/
Discrete Estimator. Counts = 83.51 2.49 (Total = 86)
Attribute: /staff/
Discrete Estimator. Counts = 7.91 78.1 (Total = 86)
Attribute: /courses/
Discrete Estimator. Counts = 78.53 7.47 (Total = 86)
Attribute: /general/contact/phone.shtml
Discrete Estimator. Counts = 5.97 80.03 (Total = 86)

```

Figure 7: *Annotated Partial Output from the WEKA EM Program Using the 20-Most-Frequent-TF Feature Set*

The other interesting cluster consisted of visitors who browsed information about post-graduate courses. This cluster is consistent with the classification results for this feature set as described above.

20-Most-Frequent-Time: As described above, the attribute values in this feature set represent the time period a visitor spent on a page during the visit. The majority of the attribute values in an instance were “0”. This presented problems for the probability density estimation component of the EM algorithm and no meaningful clusters were produced.

3.2.1 Attribute Selection

The WEKA feature selection process was applied to all of the feature sets, using best first search and correlation based feature selection.

First3-Last2: The results produced using this feature set showed “link0” as the only attribute selected. This result indicates that the first page a visited is the only one at all correlated with the class variable and is consistent with the 1R and association finding results.

First5-Last5: Attribute selection using this feature set gave “link0”, “link2last” and “linklast” as the relevant attributes in this order of significance. The fact that these attributes are not the same as for First3-Last2 suggests that these results, while intuitively plausible, are fragile.

20-Most-Frequent-TF: The root page was the only attribute selected. This result is in accordance with the classification results for this feature set as described above.

20-Most-Frequent-Time As would be expected from the previous result, the root was the only significant attribute.

3.3 Experiment2: AusVsOutsideAus2003

This experiment was identical to experiment 1, except that the 2003 data were used. As can be seen from the decision tree in figure 8, whether the root page was visited or not is still the most significant factor. Analysis of the tree suggests that activities of RMIT students looking for timetabling information at the beginning of a semester are dominant in the period that the web log was captured. This conjecture was tested in experiments 3 and 4. An association rule found using the First5-Last5 feature set showed that if visitors visit the “/timetables” page then they are from within Australia. This rule is consistent with the pattern discovered using the classification technique in experiment 2. Attribute selection using the 20-Most-Frequent-TF feature set showed the root page and “/timetables” as the significant attributes

4 Other Experiments

We carried out 6 further experiments using the procedure described above for experiments 1 and 2, using different ways of setting the class value and selecting instances as appropriate.

Experiments 3 (2001 data) and 4 (2003 data) were conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university. The results were consistent with those obtained in experiments 1 and 2. One interesting pattern that differentiates access patterns is that some visitors from outside RMIT university tend to access information about the employment prospects and career and industry collaborations of the computer science department of RMIT university. It can be conjectured that these are prospective students and are concerned about their career and employment prospects as a result of doing a course at RMIT, another nugget.

Experiments 5 (2001 data) and 6 (2003 data) were designed to compare access patterns of visitors from within RMIT university and visitors from outside RMIT university but within

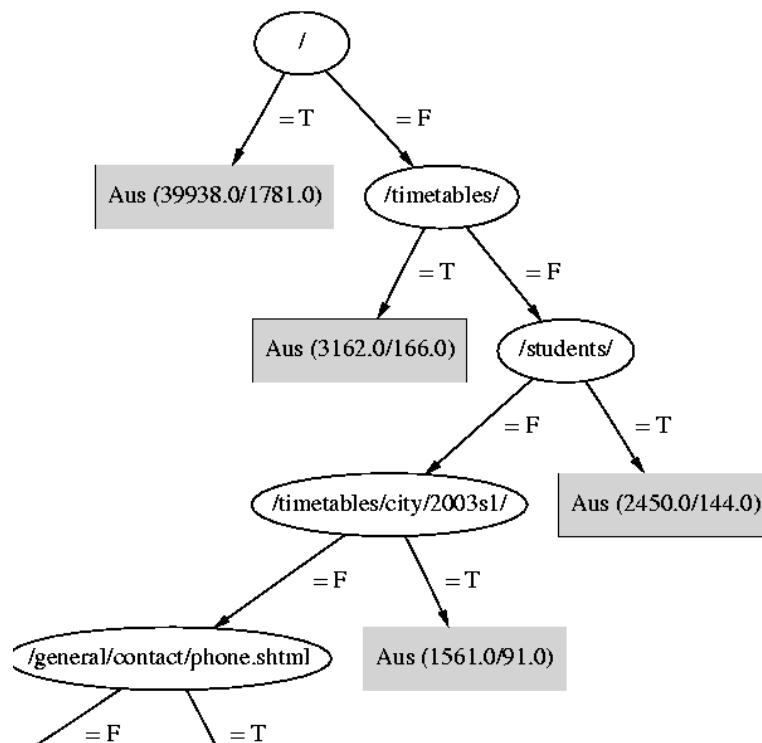


Figure 8: Partial Decision Tree from the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment 2

Australia. The major new finding was that visitors from outside RMIT university generally spend more time on the pages compared to visitors from within RMIT university. This is mostly because students at RMIT university are familiar with the website structure and hence quickly navigate to the page they require.

Experiments 7 (2001 data) and 8 (2003 data) were conducted to compare access patterns between visitors from educational institutions and other visitors. The major difference is that visitors from educational institutions access staff details and their contact information, whereas other visitors access career and program related information. It would appear that academics use the website for getting contact information.

4.1 Long Transactions

During the course of the investigation we noticed that there was quite a large number of long transactions and thought that they were worth analysing. It turned out that these were mostly visitors who looked at a large number of programs of study before downloading a brochure. This can be regarded as a golden nugget as it is confirmation that the web site is providing the intended information to prospective students.

5 Conclusions

Our first goal was to determine whether we could find any golden nuggets in the web log data using general purpose data mining algorithms. In this we have been successful and the major discovered nuggets are listed below. A major issue in using general purpose data mining

algorithms is the preparation of the feature sets to be used. Finding the “right” feature set is a difficult problem and requires some intuition regarding the goal of data mining exercise. We are not convinced that we have used the best feature sets and we think that there is more work to be done in this area.

Our second goal was to analyse the visitors to the site and somehow characterise or distinguish them in some way. In this we found 9 nuggets, the most significant ones being (1) Visitors from outside Australia do not come via the root, but arrive at an internal page via a search engine. (2) Visitors from outside Australia go to the postgraduate course work programs page and are probably prospective students. (3) Visitors from outside RMIT generally spend more time on a page than visitors from inside RMIT. (4) Visitors from academic sites look for staff contact information while visitors from non academic sites look for program and career information.

The process of transforming the original web logs into transaction feature sets is not without error. It requires the use of heuristics at several steps. Thus high accuracy from the data mining algorithms cannot be expected. However, the evidence supporting the golden nuggets comes from a number different algorithms and feature sets and we believe it is compelling.

Most web sites only perform a rudimentary analysis of web logs based on hits in some time period. Our work has shown that, after an initial investment into reusable libraries for transaction identification and feature extraction, it is possible to use existing data mining packages to find golden nuggets in the web logs.

Acknowledgements: We thank Laura Thomson for providing the transaction extraction scripts and for a number of useful suggestions about the work.

References

- [1] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), February 1999.
- [2] Y. Fu, M. Creado, and C. Ju. Reorganizing websites based on user access patterns. In *Proc. of the ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA*, pages 583–585, 2001.
- [3] Jeffrey Heer and Ed Huai hsin Chi. Separating the swarm: categorization methods for user sessions on the web. In Loren Terveen, Dennis Wixon, Elizabeth Comstock, and Angela Sasse, editors, *Proceedings of the CHI 2002 Conference on Human Factors in Computing Systems (CHI-02)*, pages 243–250, New York, April 20–25 2002. ACM Press.
- [4] V. Uma Maheswari, A. Siromoney, and K. M. Mehata. Mining web usage graphs using example state space search. *International Journal of Computational Intelligence and Applications*, 2(2):209–220, 2002.
- [5] F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. In *ACM SigWeb Letters*, 8(3): pages 13-19, 1999.
- [6] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), At-lanta*, 2001.
- [7] M. Spiliopoulou, C. Pohle, and L.C. Faulstich. Improving the effectiveness of a website with web usage mining. In *Advances in Web Usage Analysis and User Profiling, Berlin, Springer, pp. 14162*, 2000.
- [8] I.H. Witten and E. Frank. Data mining - Practical machine learning tools and techniques using JAVA implementations. *Morgan Kauffmann Publishers*, 2000.
- [9] Osmar R. Zaiane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Proc. Advances in Digital Libraries, ADL*, pages 19–29, April 1998.