# Data Mining of Web Access Logs from an Academic Web Site

Vic Ciesielski

Anand Lalani

RMIT University

# Summary

- Our Goals
- Web logs
- Preprocessing, sessionalisation and feature extraction
- Data mining algorithms and data files
- Golden Nuggets found
- Conclusions

# Goals

- Data Mining is the process of finding “Golden Nuggets” of knowledge in vast amounts of data
- OUR QUESTION: What can we learn about the visitors to the CS site from the weblogs
- Are there any Nuggets?
- Use existing algorithms (WEKA)

# Three Weblog Entries

Visitor IP, userid (if passwd), date-time, file, protocol, status, bytes, referrer, browser

202.161.108.167 - - [01/Feb/2003:00:00:03 +1100] "GET /timetables/city/2003s1/logo.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 "

202.161.108.167 - - [01/Feb/2003:00:01:32 +1100] "GET /courses/academic_program_files/academic_program_2003.shtml HTTP/1.1" 200 213562 "http://www.cs.rmit.edu.au/timetables/" "Scooter-ARS-1.1"

212.113.164.99 - - [01/Feb/2003:00:09:16 +1100] "GET /cats03/ HTTP/1.1" 200 7406 http://www.google.com/search? q=the+cats+2003&btnG=Pesquisa+Google&hl=pt&ie=UTF-&oe=UTF-8" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

# Preprocessing

- Remove images (gif, jpeg)
- Proxy server entries
- Web crawler entries
- Bad requests
- Can’t get IP name for IP number
- Need to use some heuristics

# Extracting Sessions (Transactions)

- WANT: All of the entries by one user in a visit
- Need to use heuristics (ie guess)
- Transaction = all requests from 1 IP number separated by less than 30 mins
- Clearly there will be some errors here

# Data Mining Task

- Who are the visitors and what are they doing?

1. Australia vs Outside Australia
2. Inside RMIT vs Outside RMIT
3. Inside RMIT vs Outside RMIT, in Aus
4. .edu vs other visitors

- Use IP name (More heuristics)

# Classification

- Telling the difference between 2 or more classes
- Many algorithms
- We use OneR and Decision Tree
- Algorithms need a relational table
- Transactions are of varying length

# Feature Extraction

- First3-Last2
- First5-Last5
- Most-frequent-25-TF
- Most-frequent-25-time
- [4 arff files]

# First3-Last2

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

Also First5-Last5

# Most-frequent25-TF

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

* Most-frequent 6 shown for brevity

* Most-frequent25-time: T/F replaced by the time spent on the page

# Data Used

| Begin | End | Entries | Transactions |
|---|---|---|---|
| 29/05/01 | 03/06/01 | 1,000,000 | 4,591 |
| 04/02/03 | 23/04/03 | 11,390,257 | 55,602 |

# Experimental Plan

| | 2001 Data | 2003 Data |
|---|---|---|
| Aus vs Not Aus | Experiment 1 | Experiment 2 |
| Inside RMIT vs Outside RMIT | Experiment 3 | Experiment 4 |
| Inside RMIT vs not RMIT, but Aus | Experiment 5 | Experiment 6 |
| Edu vs Not edu | Experiment 7 | Experiment 8 |

Each experiment

* Classifiction
* Clustering
* Association Finding
* Attribute Selection

# OneR Algorithm

- Consider each attribute in turn
- Find a rule based on the single attribute which most accurately classifies the data

# First3-Last2 and First5-Last5

- OneR did not give any rules with 70% or greater accuracy
- [Why 70%?]
- Nor did any other classifier

# OneR Algorithm
# 20-Most-Frequent-TF 2001

/:

T -- Aus

F -- NotAus

=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 3226 | 70.2755 % |
| Incorrectly Classified Instances | 1365 | 29.7245 % |
| Total Number of Instances | 4591 | |

IF visit home page THEN from Australia ELSE from outside Aus

# Decision Tree
# 20-Most-Frequent-TF

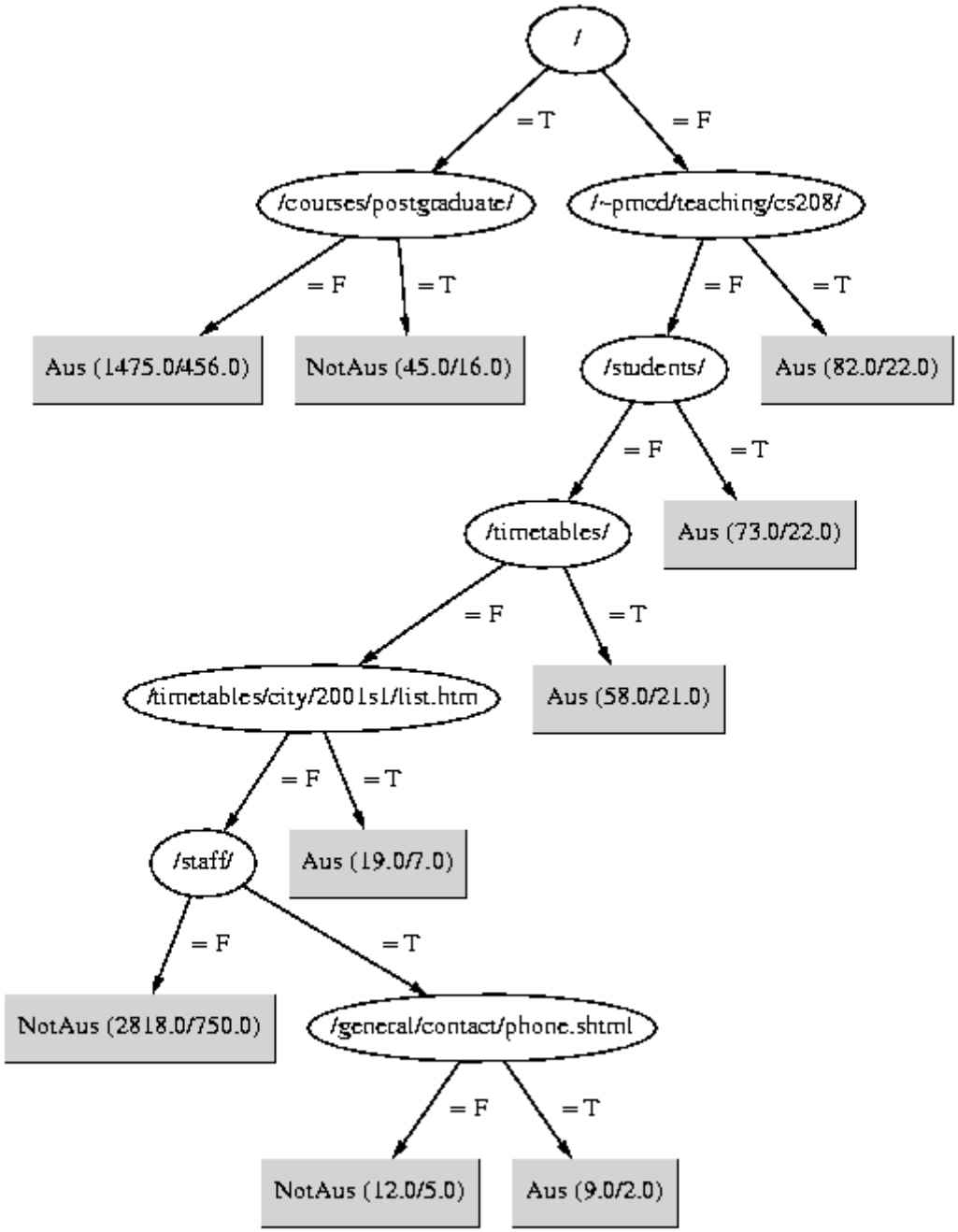
Number of Leaves : 9

Size of the tree : 17

Time taken to build model: 4.05 seconds

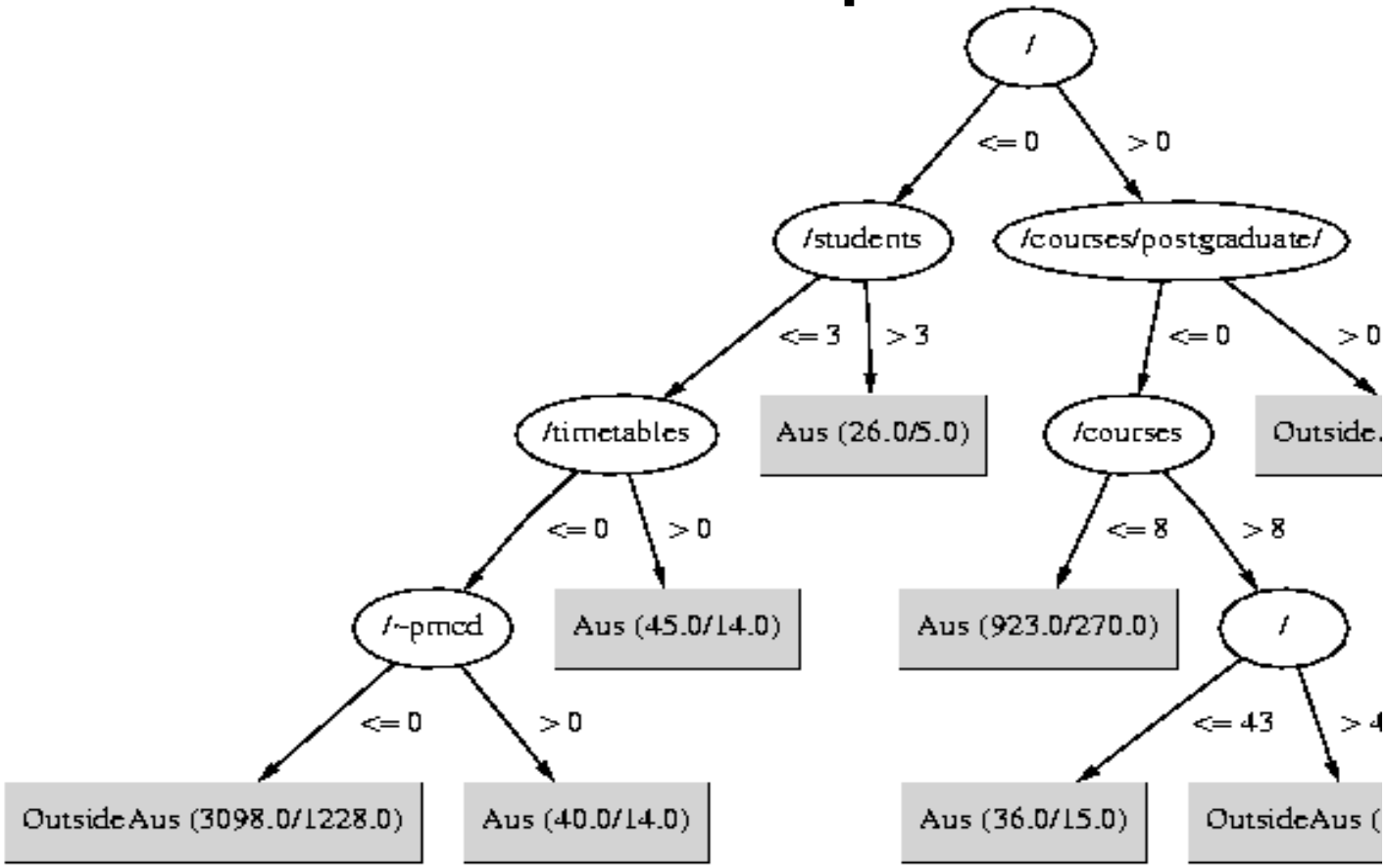
=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 3274 | 71.3134 % |
| Incorrectly Classified Instances | 1317 | 28.6866 % |
| Total Number of Instances | 4591 | |


/
= T
= F
/courses/postgraduate/
/~pmcd/teaching/cs208/
= F
= T
= F
= T
Aus (1475.0/456.0)
NotAus (45.0/16.0)
/students/
Aus (82.0/22.0)
= F
= T
/timetables/
Aus (73.0/22.0)
= F
= T
/timetables/city/2001s1/list.htm
Aus (58.0/21.0)
= F
= T
/staff/
Aus (19.0/7.0)
= F
= T
NotAus (2818.0/750.0)
/general/contact/phone.shtml
= F
= T
NotAus (12.0/5.0)
Aus (9.0/2.0)

# Decision Tree

# 20-Most-Frequent-Time



# Golden Nuggets?

- OneR 70%,  J48 71%
  - Other variables not very important
- Visit the root, post grad page, not Aus
  - Potential international students looking for course information

# Association Finding First3-Last2

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)

# Association Finding 20-Most-Frequent-TF

/~caspar/turbo_mazda_323.htm=F 911 ==>
/~shyam/cs492meta.html=F

/~caspar/turbo_mazda_323.htm=F 911 ==>
/~winikoff/palm/dev.html=F

/conf/doa/2001/=F 907 ==>
/~winikoff/palm/dev.html=F /~shyam/cs492meta.html=F

/~winikoff/palm/dev.html=F /conf/doa/2001/=F 907
==> /~shyam/cs492meta.html=F

/~shyam/cs492meta.html=F /conf/doa/2001/=F 907
==> /~winikoff/palm/dev.html=F

# Association Finding

- Most of the associations found are not very useful

# Clustering

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

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: /timetables/
Discrete Estimator. Counts = 79.28 6.72 (Total = 86)
Attribute: /employment/
Discrete Estimator. Counts = 84.08 1.92 (Total = 86)
Attribute: /general/contact/phone.shtml
Discrete Estimator. Counts = 5.97 80.03 (Total = 86)
Attribute: /courses/postgraduate/
Discrete Estimator. Counts = 83.72 2.28 (Total = 86)

# Significant Clusters

- Visitors who browsed pages with staff contact details
- Visitors who browsed information about post graduate courses
- Nugget:  Why were there no clusters with undergraduate courses?  Website problem?

# Attribute Selection

- First3-Last2:  link0
- First5-Last5: link0, link2last,linklast
  - Fragile
- 20-Most-Frequent-TF: link0
- 20-Most-frequent-time: link0
- Consistent with OnerR, J48 results

# Aus vs Outside Aus

- Aus => Visit root page
- Outside  Aus => Don't visit root (Arrive via Search Engine
- Outside  Aus => visit pgrad courses
- Prospective Students?
- Why not ugrad courses?  A problem with the website?

# RMIT vs Outside RMIT

- Outside RMIT => visit employment prospects
- Prospective students?

# Long Transactions

- We noticed a number of very long transactions
- Visitors looked at a large number of programs and downloaded brochures
- Nugget: Prospective students getting information

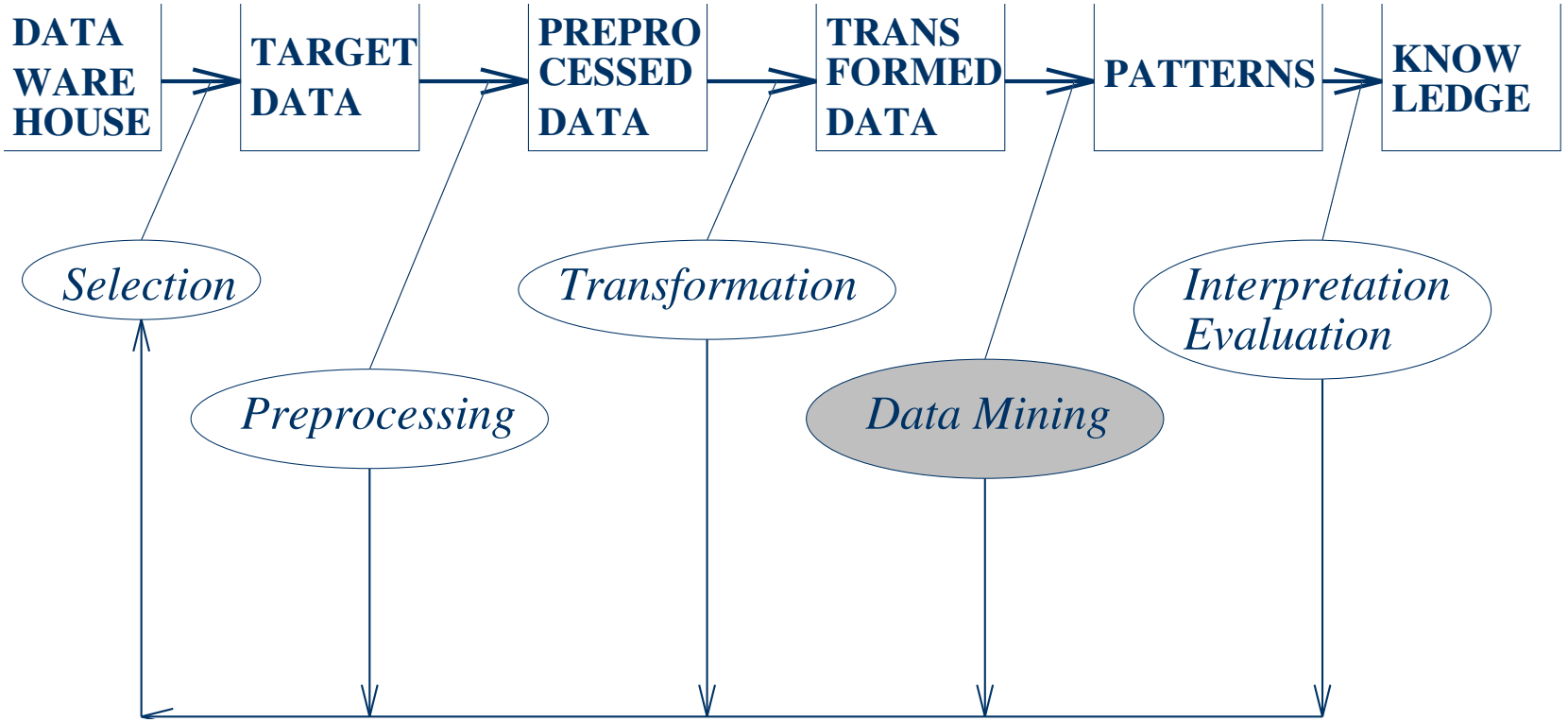
# 9 Nuggets Found

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
5. Long transactions indicate prospective students

# Conclusions about the Data Mining Task

- Found some nuggets.  How valid?
- Errors in preprocessing suggest high accuracy cannot be expected.  BUT
- Several paradigms give evidence for the same result
- Nuggets can be found in web logs without specialised algorithms

# KDD Process


DATA WARE HOUSE
TARGET DATA
PREPRO CESSED DATA
TRANS FORMED DATA
PATTERNS
KNOW LEDGE
Selection
Preprocessing
Transformation
Data Mining
Interpretation Evaluation

# Conclusions about the Process

- You need a methodology based on
  - Thinking up reasonable hypotheses (questions)
  - Preparing suitable data
  - Running a suitable algorithm
  - Interpreting the results
- Often in any kind of mining you dont find anything
- What else could have been done with the web log data?