caseStudy: decision tree application to Web Robot Detection
Tan, P.N., Steinbach, M., and Kumar, V. (2014). Introduction to Data Mining, Pearson

Web usage mining is the task of applying data mining techniques to extract useful patterns from Web access logs. These patterns may reveal interesting characteristics of site visitors — e.g., people who repeatedly visit a Web site and view the same product description page are more likely to buy the product if certain incentives such as rebates or free shipping are offered.

One potential difficulty with Web usage mining is the need to distinguish between accesses made by human users from accesses due to Web robots. A Web robot (also known as a Web crawler) is a software program that automatically locates and retrieves information from the Internet by following the hyperlinks embedded in Web pages. These programs are often deployed by search engine portals to gather the necessary documents for indexing the Web. Since Web robot accesses do not reflect the true browsing behavior of a human user, they must be discarded before applying Web mining techniques.

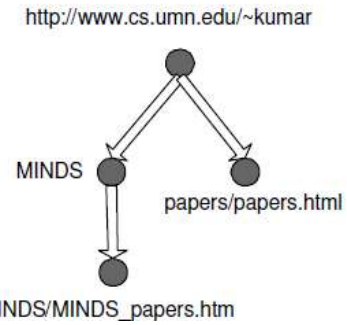
This case describes how decision tree classification can be used to distinguish between accesses by human users and Web robots. The input data was obtained from a Web server log, a sample of which is shown in Figure 4.16(a). Each line corresponds to a single page request made by a Web client (user or robot). The fields recorded in the Web log include the IP address of the client, timestamp of the request, Web address of the requested document, size of the document, and the client's identity (via the User Agent field). A Web session is a sequence of requests made by a client during a single visit to a Web site. Each Web session can be modeled as a directed graph, where the nodes correspond to the requested pages and the edges correspond to hyperlinks connecting one Web page to another. For example, Figure 4.16(b) is a graphical representation of the first Web session shown in the Web server log.

To classify the Web sessions into accesses by humans or robots, we must first derive new attributes describing the characteristics of each session. An example of the attribute set one may use for Web robot detection is shown in Figure 4.16(c). Among the notable attributes include the depth and breadth of the traversed Web pages. Depth measures the maximum distance of a requested page, where distance is determined based on the number of hyperlinks away from the entry point of the Web site. For example, the home page http://www.cs.umn.edu/¡«kumar is assumed to be at depth 0 while http://www.cs.umn.edu/kumar/MINDS/MINDS papers.htm is located at depth 2. Based on the Web graph shown in Figure 4.16(b), the Depth attribute for the first session is equal to two. In the meantime, the breadth attribute measures the width of the corresponding Web graph. With this definition, the breadth of the Web session shown in Figure 4.16(b) is equal to two.

The data set for classification contains 2916 records, with equal number of sessions due to Web robots (class 1) and human users (class 0). 10% of the data are reserved for training while the remaining 90% are used for testing. The fitted decision tree model is shown in Figure 4.17. The model has an error rate equal to 3.8% on the training set and 5.3% on the test set.
The model suggests that Web robots can be distinguished from human users in the following way:
1. Accesses by Web robots tend to be broad but shallow, whereas accesses by human users tend to be more focused (deep but narrow).
2. Unlike human users, Web robots seldom retrieve the image pages associated with a Web document.
3. Sessions due to Web robots tend to be long and contain a large number of requested pages.
4. Web robots are more likely to make repeated requests for the same document since the Web pages retrieved by human users are often cached by the browser.

Session	IP Address	Timestamp	Request Method	Requested Web page	Protocol	Status	Number of bytes	Referrer	User Agent
1	160.11.11.11	08/Aug/2004 10:15:21	GET	http://www.cs.umn.edu/~kumar	HTTP/1.1	200	6424		Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
1	160.11.11.11	08/Aug/2004 10:15:34	GET	http://www.cs.umn.edu/~kumar/MINDS	HTTP/1.1	200	41378	http://www.cs.umn.edu/~kumar	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
1	160.11.11.11	08/Aug/2004 10:15:41	GET	http://www.cs.umn.edu/~kumar/MINDS/MINDS_papers.htm	HTTP/1.1	200	1018516	http://www.cs.umn.edu/~kumar/MINDS	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
1	160.11.11.11	08/Aug/2004 10:16:11	GET	http://www.cs.umn.edu/~kumar/papers/papers.html	HTTP/1.1	200	7463	http://www.cs.umn.edu/~kumar	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
2	35.9.2.2	08/Aug/2004 10:16:15	GET	http://www.cs.umn.edu/~steinbac	HTTP/1.0	200	3149		Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7) Gecko/20040616

(a).Example of Web Server Log



(b).Graph of a Web session

Attribute Name	Description
totalPages	Total number of pages retrieved in a Web session
ImagePages	Total number of image pages retrieved in a Web session
TotalTime	Total amount of time spent by Web site visitor
RepeatedAccess	The same page requested more than once in a Web session
ErrorRequest	Errors in requesting for Web pages
GET	Percentage of requests made using GET method
POST	Percentage of requests made using POST method
HEAD	Percentage of requests made using HEAD method
Breadth	Breadth of Web traversal
Depth	Depth of Web traversal
MultiIP	Session with multiple IP addresses
MultiAgent	Session with multiple user agents

(c). Derived attributes from Web robot detection

Figure 4.16. Input data for Web robot detection.

Decision Tree:

```
depth = 1 :
|  breadth > 7 : class 1
|  breadth <= 7 :
|  |  breadth <= 3 :
|  |  |  ImagePages > 0.375 : class 0
|  |  |  ImagePages <= 0.375 :
|  |  |  |  totalPages <= 6 : class 1
|  |  |  |  totalPages > 6 :
|  |  |  |  |  breadth <= 1 : class 1
|  |  |  |  |  breadth > 1 : class 0
|  |  width > 3 :
|  |  |  MultiIP = 0:
|  |  |  |  ImagePages <= 0.1333 : class 1
|  |  |  |  ImagePages > 0.1333 :
|  |  |  |  |  breadth <= 6 : class 0
|  |  |  |  |  breadth > 6 : class 1
|  |  |  MultiIP = 1:
|  |  |  |  TotalTime <= 361 : class 0
|  |  |  |  TotalTime > 361 : class 1
depth > 1 :
|  MultiAgent = 0:
|  |  depth > 2 : class 0
|  |  depth <= 2 :
|  |  |  MultiIP = 1: class 0
|  |  |  MultiIP = 0:
|  |  |  |  breadth <= 6 : class 0
|  |  |  |  breadth > 6 :
|  |  |  |  |  RepeatedAccess <= 0.0322 : class 0
|  |  |  |  |  RepeatedAccess > 0.0322 : class 1
|  MultiAgent = 1:
|  |  totalPages <= 81 : class 0
|  |  totalPages > 81 : class 1
```

Figure 4.17. Decision tree model for Web robot detection