Journal of Information Science
http://jis.sagepub.com/

Web robot detection based on pattern-matching technique

Shinil Kwon, Young-Gab Kim and Sungdeok Cha

Journal of Information Science 2012 38: 118 originally published online 27 February 2012
DOI: 10.1177/0165551511435969

The online version of this article can be found at:
http://jis.sagepub.com/content/38/2/118

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
cilip
Chartered Institute of Library and Information Professionals

Article

JIS

Journal of Information Science
38(2) 118–126

DOI: 10.1177/0165551511435969
jis.sagepub.com
SAGE

Web robot detection based on pattern-matching technique

Shinil Kwon, Young-Gab Kim and Sungdeok Cha
Korea University, Korea

Abstract

In web robot detection it is important is to find features that are common characteristics of diverse robots, in order to differentiate between them and humans. Existing approaches employ fairly simple features (e.g. empty referrer field, interval between successive requests), which often fail to reflect web robots' behaviour accurately. False alarms may therefore occur unacceptably often. In this paper we propose a fresh approach that expresses the behaviour of interactive users and various web robots in terms of a sequence of request types, called request patterns. Previous proposals have primarily targeted the detection of text crawlers, but our approach works well on many other web robots, such as image crawlers, email collectors and link checkers. In empirical evaluation of more than 1 billion requests collected at www.microsoft.com, our approach achieved 94% accuracy in web robot detection, estimated by *F*-measure. A decision tree algorithm proposed by Tan and Kumar was also applied to the same data. A comparison shows that the proposed approach is more accurate, and that real-time detection of web robots is feasible.

Keywords

web robot detection; web robot pattern; human pattern; pattern analysis

1. Introduction

With the explosive growth of the web-based economy, the accurate and rapid detection of various web robots has become an important research topic. There are a variety of web robots, such as text crawlers, image collectors, link checkers, download tools, and email harvesters. Web robots generally run many beneficial functions (e.g. powerful search engine and shopping robots in e-commerce sites). However, web robots can also be developed for malicious purposes, such as autonomous signing-up to send spam. Security engineers therefore need to monitor requests made to web servers, and identify sessions that have been initiated by web robots.

Various web robot detection techniques have been proposed. Some (e.g. [1–5]) can detect two or three web robots simultaneously, but the detection accuracy varies from one web robot to another. In other words, the features were not sufficiently representative to reliably distinguish the behaviour of interactive users from that of various web robots. Some techniques (e.g. [6–8]) use rather simplified assumptions – for example, that web robots would generate periodic requests in a shorter interval over an extended period, compared with interactive users. Such conventional wisdom does not always hold, according to more than 1 billion www.microsoft.com logs that we analysed. For example, image collectors and email collectors often issued requests only when they located image files or email addresses on web pages. In addition, most web robots do not use a referrer field, which enables one to see where the request originated in the HTTP request (e.g. Girafa, an image collector, and Arachmo, a download tool, only use a referrer field), despite the considerable research that has studied the relationships between web robots and the unassigned referrer field in the request.

One of the important activities for web robot detection is to find features that are common characteristics of diverse web robots, as this should be able to show definite differences between web robots and humans. Data-mining algorithms are usually used to detect web robots, but it is difficult to detect various web robots, because they do not use common features of diverse web robots in the web log sample. Although most web logs used by many researchers include various web robots, the text crawler constitutes a large proportion of IPs in web robot samples. For example, about 94% of MS

Corresponding author:
Young-Gab Kim, College of Information and Communication, Korea University, 1, 5-ga, Anam-dong, Sungbuk-gu, 136-701, Seoul, Korea.
Email: always@korea.ac.kr

web logs comprise three text crawlers (Google, Microsoft and Yahoo) although we found 29 types of web robot. Each text crawler is composed of several types of IP group, reflecting their primary purpose. About 55.7% of text crawlers collect only image files, and about 29% collect only web pages. This means that these two types of group have different characteristics (e.g. they have different goals), and both types are equally important features in detecting web robots.

According to our analysis of 1 billion requests made to www.microsoft.com, most web robots use many IPs, and the IPs for each web robot can be grouped by their function. That is, web robots run their tasks with many IPs to achieve faster performance, which is also more efficient. The IPs belonging to one web robot can be divided into many different groups, according to the requested file types. Some IPs collect only web pages to build the structure of the web server over the internet. Others gather various file types (e.g. images, multimedia, or documents) to provide services, such as previews or image searching. For example, text crawlers have two typical IP groups: one collects web pages, and the other gathers other file types. The former group records constant intervals, and generates three times as many session requests as interactive humans do. In the latter group, which especially collects image files, the number of requests belonging to one IP recorded only one of 30 requests compared with the first group. Many web robots follow this pattern for better performances. Consequently, this feature is useful to distinguish between web robots and humans, because humans usually use only one IP.

In this paper, we propose an effective web robot detection method, based on the web robot's patterns that keep track of which types of file are requested by a target session. We can detect five web robots – Google, MS, Yahoo!, CFNetwork and Powermarks – using the proposed pattern approach with about 94% accuracy.

The rest of our paper is organized as follows. Section 2 surveys existing approaches to web robot detection, and their limitations. In Section 3, we explain the Microsoft web log used in this study. We also describe patterns of typical web robots and humans in detail, with data obtained in our analysis of Microsoft web logs. In Section 4, we report our experimental results. Finally, Section 5 concludes the paper.

2. Related work

Doran and Gokhale [9] classified existing robot detection techniques into four categories: syntactical log analysis, analytical learning techniques, traffic pattern analysis, and Turing test systems. Syntactical log analysis techniques can find only robots that are well known in advance, because they are based on prior knowledge of specific keywords, and the addresses from which the robots originate. Analytical learning techniques estimate the likely connection between observed sessions and web robot sessions using a formal machine-learning model. Traffic-based analysis techniques seek to find common characteristics of the web robot traffic that contrast with the features of human traffic. Turing test systems try to classify a conversation reliably as being produced by a computer or a human in real time. Traffic-based analysis classifies web robots and finds characteristics of target sessions, whereas the Turing test classifies them only as human or computer. Our proposed approach – web robot detection based on a pattern-matching technique – is similar to traffic pattern analysis.

Most research [1–5] with analytical learning techniques does not find common features until data-mining algorithms are applied to web robots, because they analyse the characteristics of web robots based on the results of data mining. Therefore the features proposed by analytical learning techniques are not common characteristics of the various web robots, and these approaches are unable to show high performance on web robot detection against a small number of IPs.

Tan and Kumar [1] examined web robot detection in terms of 26 features, such as the ratio of empty strings in the referrer field, and the ratio of the head method with the C4.5 decision tree algorithm. They also evaluated each feature using correlation analysis. Lourenco and Belo [2] chose roughly the same features as Tan and Kumar, and the C4.5 decision tree algorithm. These constitute the overall platform that detects web robots in real time, while the crawler visit is still in progress. Bomhart et al. [3] added eight features (e.g. the minimum, maximum, average and standard deviation of the number of HTML links contained in each page), and used a neural network to detect web robots with a web log preprocessing tool called RDT (robot detection tool). They compared the performance of the neural network with the performance of a decision tree. Stassopoulou and Dikaiakos [4] constructed a Bayesian network using six features: maximum sustained click rate, duration of session, percentage of image requests, percentage of pdf/ps requests, percentage of 4xx error responses, and robots.txt file requests. Lu and Yu [5] detected web robots using a hidden Markov model (HMM) and the inter-arrival time. They partitioned time into discrete intervals of the same length. Each interval with the same length is considered as an observation for the HMM. However, these studies did not show the individual performance for each web robot.

Several other studies [6–8, 10] have proposed new features based on traffic pattern analysis of web robots. They focused on common characteristics of web robots. Some studies stated that web robots showed a consistent time

difference among requests. Almeida et al. [6] presented a characterization of the workload generated by web robots. This characterization focused on the statistical properties of the arrival process, and on the robot behaviour graph model. Almeida et al. showed that web robot sessions generated requests more regularly, and maintained their presence on a website longer, than did users. Huntington et al. [7] proposed a multi-step log analysis technique with the online scientific journal *Glycobiology*. They presented web robots that variously generated a high usage count, viewed pages very rapidly, and viewed over a long time. Duskin and Feitelson [8] studied the interaction between the query submittal rate and the minimum time interval between different queries. They focused particularly on the smallest interval between different queries in a session. Ye et al. [10] measured a server workload using the difference between HTTP response time and ICMP response time. They presented web robot navigational patterns, and traffic analysis techniques.

Other studies [6, 11] have considered the ranking of web pages by popularity. Lee et al. [11] developed an effective characterization metric, based on workload characteristics and resource types, in detecting and classifying various web robots. They proposed that the popularity of web pages referenced by human beings is highly concentrated. Lin et al. [12] analysed the relationship between 'partial response' (e.g. some clients are requesting isolated URLs pointing to popular files) and web robots. They showed that, since web robots usually requested different segments of the target object from multiple sites, individual web robot sessions requested far fewer objects than human sessions.

Few of the approaches based on behavioural analysis were conducted to detect web robots with their proposed features in a real environment. Geens et al. [13] evaluated different robot detection methods, such as IP address list, robots.txt, robotic user-agent, the HEAD method, unassigned referrer, no image requests, night, standard deviation of arrival time, and average arrival time. They said that these features (time interval, for example) were not effective in the classification of web robots, because the features extracted from most IP groups collect web pages with numerous requests, and do not include the characteristics of all IP groups collecting image files, document files and multimedia files.

The approach proposed in this paper extracts typical patterns by analysing the cooperative relationship between IP groups in a web robot. Cooperative relationships are common features. The web robot has to use the cooperative relationship between IP groups to guarantee time efficiency (e.g. if a text crawler uses only one IP, the crawling operation for collecting web pages would stop frequently to download image files). We successively detected web robots with 94% accuracy using the typical patterns.

3. Analysis of typical patterns on web robots

3.1. Microsoft's web server log

The web log is an important application possibility for various websites. Much research on empirical experiments has used a relatively small dataset, collected at universities. The MS web log is excellent in size and generality. The web log used in this study is the same as that used in [11]. It is about 250 GB. More than 1 billion requests are captured in 24 hours. We divided the web log into four segments, each containing 6 hours, to evaluate web robot detection accuracy. Care was taken to ensure that 'peak' hours in different continents (e.g. North America, Asia, and Europe) were appropriately reflected in each segment. Data belonging to each segment was further divided into training and test data in a 2:1 ratio. That is, logs collected during the first four hours were used as training data, and the rest were used as test data.

IPs initiated by various web robots were identified, based on information included in the user agent field, to extract common characteristics of various web robots. The database compiled at www.user-agents.org lists user-agent values found in all known web robots. While we are aware of the possibility that the user-agent value can be faked (e.g. an interactive user issuing requests as though it is a Google crawler), we concluded that such a chance is sufficiently low, and therefore could be ignored. We thus built a database containing more than 5 million requests generated by three different web robots, as shown in Table 1. All requests from the same IP containing the same user-agent were grouped into the

Table 1. Sessions initiated by web robots

Web robot	Type	Sessions	Requests	Ratio of sessions*
Google	Text crawler	585	837,221	
Microsoft		6,106	1,073,705	93.5%
Yahoo		22,372	3,600,438	
CFNetwork	Email collector	520	391,050	1.7%
Powermarks	Link checker	1,486	61,093	4.8%

*Ratio of sessions $= \frac{\text{Sum of sessions belonging to IPs that have the same web robot}}{\text{Total sessions}}$.

same session during session construction. The proposed approach used three types of web robot – text crawlers, email collectors and link checkers – to create patterns and evaluate our approach.

As depicted in Table 1, the text crawler had about 15 times the number of sessions compared with the sum of those of the email collector and link checker. Therefore it is necessary to take common features between the three types of web robots into consideration.

3.2. Pattern analysis of web robots and human

We found that web robots had typical patterns created by the sequences of request file types. For example, about 61.5% of text crawler sessions request image files, especially in the bootstrapping step, for which the sequence of requests does not include web pages containing link information for image files, whereas humans first request a main page or other web pages.

We analysed the pattern of web robots based on the sequence of request file types. Patterns were created by analysing the cooperative relationship among IP groups in a web robot. We used the categorization criteria proposed by Tan and Kumar [1] to identify file types. In addition, we added a 'main page' file type. Table 2 shows the types of file used, and their meanings.

The main page is an important feature when detecting web robots, because humans usually start web surfing from a main page, such as www.microsoft.com/en-us/default.aspx or www.msn.com. According to our experiment, humans chose the main page to start web surfing 33.6% of the time. Web robots, however, chose the main page only 0.5% of the time.

As illustrated in Figure 1, the proposed algorithm is very simple, but its effectiveness is powerful. First, we randomly choose the same number of IPs from both web robots and humans to extract typical patterns for each web robot, and then extract the patterns in the form of the sequence of file request types from all sessions (e.g. pattern III represents a specific IP requesting image files continuously three times). Second, we create a pattern table containing patterns and the number of matched sessions for each pattern, as depicted in Figure 1.

The proposed method detects the web robot as it requests the same sequence of file types defined in the pattern table, and the number of web robot sessions exceeds that of human sessions. For example, if one detects session PPP, the proposed algorithm finds Human: 111, Robot: 434 in a pattern table. This means that the session that requests pattern PPP will be classified a web robot.

The proposed method, which is based on the typical patterns of each web robot, detects various web robots irrespective of their request size (e.g. the ratio of requests in Table 1). As mentioned previously, data-mining algorithms, which do not consider the common features of all IP groups, may not detect small groups in a web robot, or even web robots that have small numbers of IPs. Therefore, if someone finds common characteristics of entire IP groups and web robots, web robots can be easily distinguished from humans. Thus the pattern-based web robot detection method can distinguish effectively between humans and web robots.

3.2.1 Typical patterns of text crawlers. The text crawler has contributed to the development of internet search engines. We found that a small number of IPs in text crawlers only collected web pages at short, constant intervals, and had a longer connection. In contrast, a large number of IP groups requested image files or other file types at irregular time intervals, because they performed on demand. In addition, the IP groups collecting image files showed a high variation of request frequency. Lee et al. [11] pointed out that, urban legend notwithstanding, web robots had a greater variation of request intervals than existing research based on behavioural analysis. Lee et al. calculated the variation of all the IPs that belonged to the target text crawler. Their result indicated that the characteristic of IP groups collecting image files was

Table 2. File types requested in web server

Type	Remark
P (web page)	Web pages including dynamic web pages (e.g. html, htm, shtm, js, php, asp)
M (main page)	Main page of Microsoft website
I (image)	Image files (e.g. jpg, bmp, ico, pic)
Z (compressed)	Compressed files (e.g. zip, rar, tar, bz2)
U (multimedia)	Multimedia files (e.g. avi, mpg, mpeg, mp3)
E (exec)	Binary executable files (e.g. cgi, exe)
D (document)	Binary documents (e.g. ps, pdf)
A (ASCII)	ASCII files (e.g. txt, c, java)
O (others)	Other files (e.g. eml)

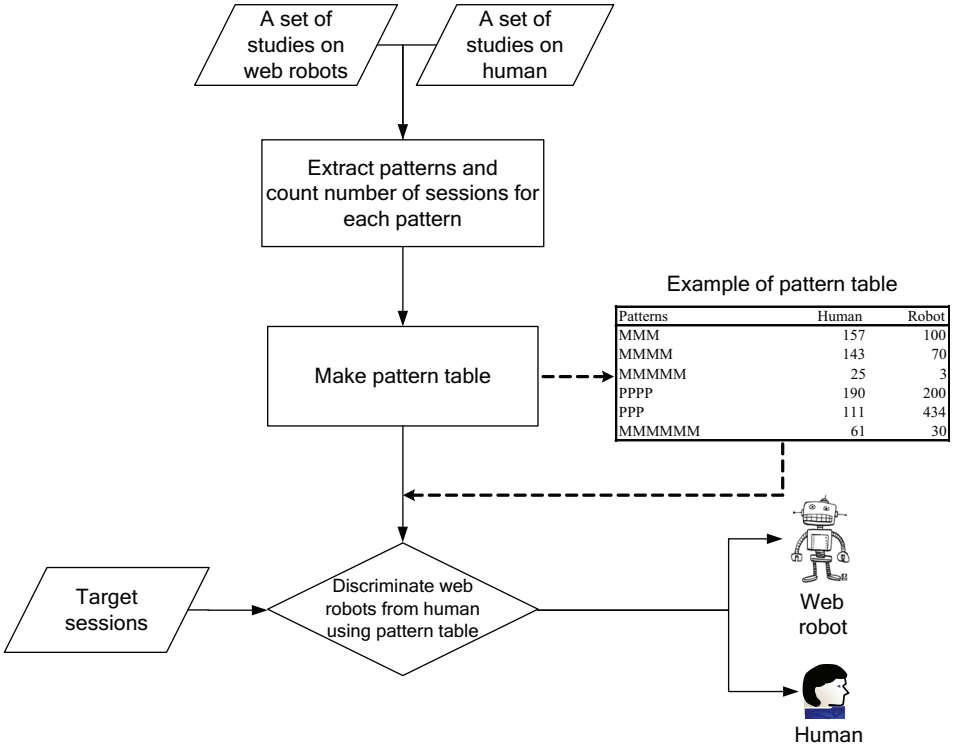

Figure 1. Overview of proposed approach.

Table 3. Typical patterns of text crawlers

File types requested	Pattern	Ratio of IPs*	Ratio of requests**
Image files only	I	4.3%	0.1%
	II	4.1	0.1
	III	6.1%	0.4%
	IIII	15.2%	1.3%
	IIIII	20.7%	2.2%
	Other patterns	5.3%	1.1%
	Total	55.7%	5.2%
Web pages only	P	4.2%	0.1%
	PP	4.1%	0.2%
	PPP	1.3%	0.1%
	PPPP	4.1%	0.3%
	over 10 thousand	0.1%	36.4%
	Other patterns	15.2%	24.9%
	Total	29.0%	62.0%
Main pages only	M	1.4%	0.1%
	Other patterns	0.4%	0.0 %
	Total	1.8%	0.1%

*Ratio of IPs $= \frac{\text{IPs that have the same pattern}}{\text{Total IPs}}$

**Ratio of requests $= \frac{\text{Sum of requests belonging to IPs that have the same pattern}}{\text{Total requests}}$

strongly reflected in the final variation value, because the group collecting image files contained three times the number of IPs as the group collecting web pages. If other methods, such as the Bayesian approach, statistical analysis and data-mining algorithms, are applied to real web servers without considering the request size differences between the two groups, which have very different characteristics, these methods have limitations in detecting both groups.

Pattern analysis is an effective method to detect various IP groups. Table 3 illustrates typical patterns of text crawlers analysed by the proposed method depicted in Figure 1.

The first IP group, which requested only image files, and the second IP group, which requested only web pages, were 56 and 29% respectively of the total IPs belonging to the text crawler (see more details in Table 3). Although the second IP group is 29% of the total IPs, these IPs comprised 62% of the total requests of text crawlers: for example, the average number of requests belonging to the IP group for collecting web pages is about 12 times the number of requests created by the IP group for collecting image files. This means that, even if an IP group is small, but has a large number of requests, it can still be the most important IP group. Thus, if we try to detect a web robot according to the number of IPs (e.g. 56% of total IPs requesting image files) without considering common features, it is possible to ignore important IP groups with small numbers of IPs (e.g. 29% of total IPs requesting web pages). The detection method based on the pattern-matching technique is effective, because IP groups have a common feature, which is that each IP group requests only one file type, considering the characteristics of all IPs in a web robot.

3.2.2 Typical patterns of email collectors. The email collector gathers email files, image files and web pages from the internet. Email collectors (e.g. CFNetwork) have entirely different characteristics from those of text crawlers, although they gather the same type of file as text crawlers. The text crawler uses about 29% of IPs to crawl web pages, whereas approximately 72% of IPs belong to the email collector crawl web pages (see more details in Table 4). That is, the text crawler uses a small IP group with an average of 94 requests for one IP, and the email collector uses a larger IP group with an average of three requests for one IP to crawl web pages. Table 4 illustrates typical patterns of the email collector in the MS web log.

The most important characteristic of the email collector is that it uses only one IP to gather email files. Furthermore, about 0.2% of IPs belonging to the email collector make 97.8% of the total requests. If data-mining algorithms are used to detect email collectors without considering the size of IP groups, it is impossible to guarantee that only one IP collecting email files can be detected. However, web robot detection based on the request pattern is also effective when detecting email collectors. We can classify one IP collecting email files easily with typical patterns of the email collector, because humans rarely request only email files continuously with a pattern such as OOOO.

3.2.3 Typical patterns of link checkers. Link checkers have a different purpose, compared with text crawlers and email collectors: they only check their registered links, and do not collect files. That is, they request only web pages corresponding to their registered links with about 97.2% of IPs. Table 5 illustrates typical patterns of the link checker in the MS web log.

Table 4. Typical patterns of email collectors

File types requested	Patterns	Ratio of IPs	Ratio of requests
Web pages only	P	45.2%	0.2%
	PP	13.7%	0.1%
	Other patterns	12.8%	0.5%
	Total	71.7%	0.8%
Image files only	I	10.4%	0.1%
	II	2.1%	0.0%
	Other patterns	11.3%	1.1%
	Total	23.8%	1.2%
Other files only (especially .eml)	Over 60,000	0.2%	97.8%

Table 5. Typical patterns of link checkers

File types requested	Patterns	Ratio of IPs	Ratio of requests
Web pages only	P	8.2%	0.3%
	PP	10.9%	0.8%
	PPP	8.5%	0.9%
	Over 20 times	34.6%	77.4%
	Other patterns	35.0%	11.9%
	Total	97.2%	91.3%

Table 6. Typical human patterns

File types requested	Patterns	Ratio of IPs	Ratio of requests
Main pages only	M	35.9%	0.9%
	MM	12.5%	0.7%
	MMM	4.5%	0.4%
	Other patterns	4.1%	0.6%
	Total	57.0%	2.6%
Image files only	I	3.0%	0.1%
	Other patterns	1.1%	0.2%
	Total	4.1%	0.3%
Web pages only	P	4.8%	0.1%
	PP	3.7%	0.2%
	Other patterns	2.9%	1.1%
	Total	11.4%	1.4%
Various files	Various patterns	27.5%	95.7%

IPs requesting web pages over 20 times are 34.6% of IPs, and the ratio of their requests is 77.4%. The proposed approach using these patterns is effective when detecting robots, because humans rarely request only web pages over three times continuously.

3.2.4 Typical patterns of humans. Humans have two dominant characteristics compared with web robots: first, they initially connect to a website with a main page, whereas web robots do not distinguish between a main page and other web pages. Only 2.6% of web robots' IPs accessed the main page at least once. Conversely, 33.6% of human sessions accessed the main page to start web surfing, and 64.0% of human sessions requested the main page at least once. Second, humans also requested more than two types of file within one session with one IP. Thereforeall types of files, including image files, multimedia files, document files and web page files, generally were discovered in one user session. Unlike text crawlers, IPs collecting only one file type constituted a small proportion of user IPs (e.g. 4.1% of IPs collect only image files and 11.4% collect only web pages). Table 6 illustrates typical human patterns.

As illustrated in Table 6, 20% of human IPs generate 95% of total requests. That is, 20% of IPs have their own patterns and unusual sequences of file request types, distinguishing them from web robot patterns.

4. Experimental evaluation

The *F*-measure, computed from the recall and precision values shown below, is widely used in machine-learning research. It provides an impartial viewpoint, in that one can increase the value of one, at the expense of the other. At one extreme, an algorithm may achieve 100% recall value if it simply declares all the sessions as those initiated by web robots. Such an algorithm is obviously useless, and filled with numerous false alarms. The *F*-measure definition is:

$$F\text{-measure} = \frac{2RP}{R+P} \tag{1}$$

where

$$\text{Recall, } R = \frac{\text{Number of web robots found correctly}}{\text{Number of actual web robots}} \tag{2}$$

$$\text{Precision, } P = \frac{\text{Number of web robots found correctly}}{\text{Number of predicted web robots}} \tag{3}$$

Figure 2 illustrates the accuracy of the proposed approach. Figure 2(a) shows that the proposed web robot detection method provides high accuracy with about a 94% *F*-measure. However, the proposed method for the first and second requests, especially those generated by the link checker and email collector (see Figures 2(c) and 2(d)), shows low performance, because all web robots have a predictable, similar pattern compared with humans, as they requested only web pages in the first and second requests. These requests are easily found in patterns of the email collector (or link checker) and humans. For example, as illustrated in Table 7, the pattern P or PP will be classified as human, because the number of sessions for pattern P in humans was 2034 and for web robots was 1162; and for pattern PP the number of sessions for humans was 1409 and for web robots was 1047 for our analysed pattern table. From the third requests, the *F*-measure

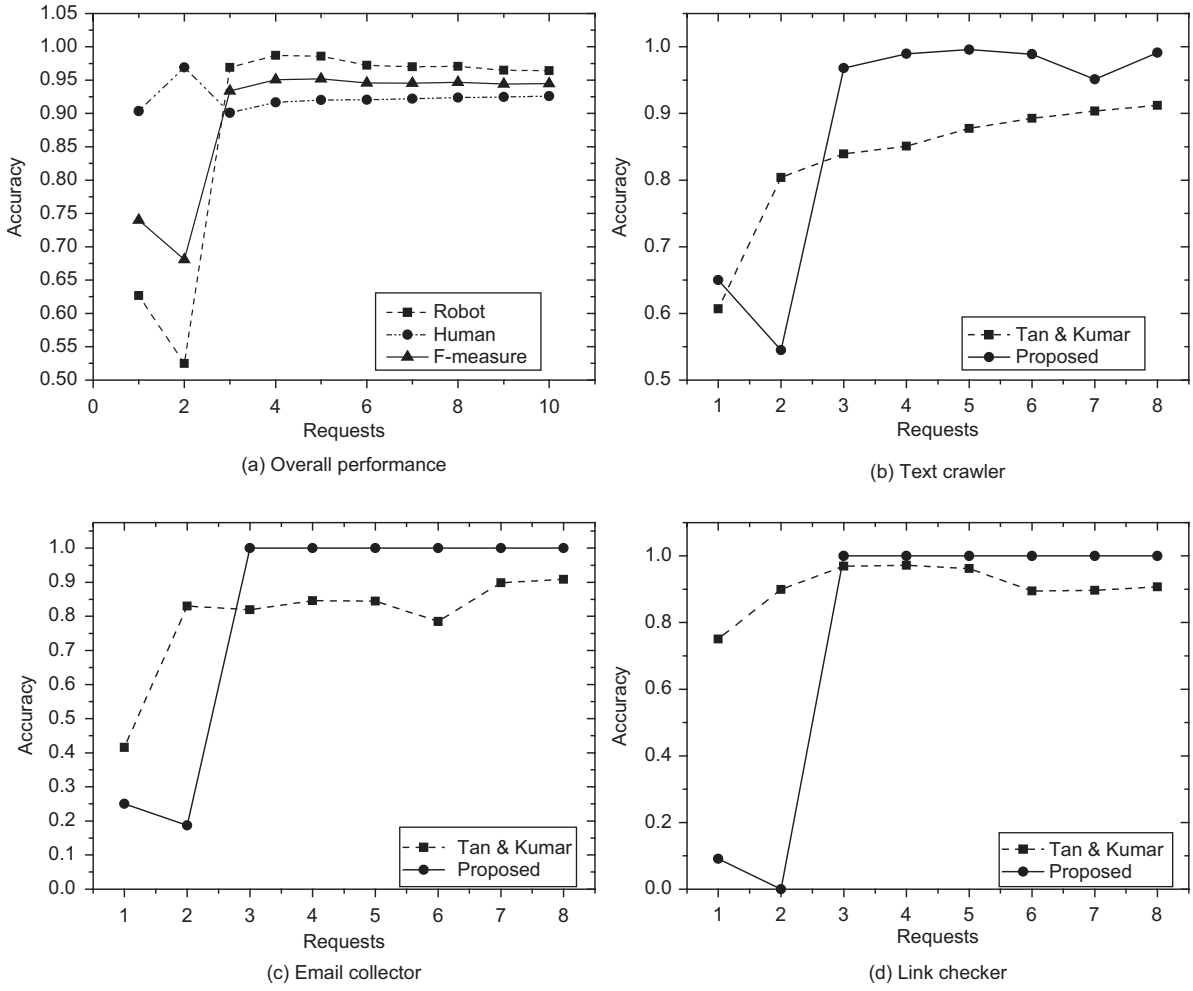

Figure 2. Accuracy of proposed approach: (a) overall performance; (b) text crawlers; (c) email collectors; (d) link checkers.

Table 7. Pattern table for web page file types

Pattern	Human	Web robot
P	2034	1162
PP	1409	1047
PPP	217	371
PPPP	227	998
PPPPP	85	158
PPPPPP	95	298

performance increased rapidly, mainly because humans perform fewer sessions than do web robots. For example, the number of sessions for pattern PPP in humans was 217, and for web robots was 371, because humans requested both web pages and image files included in web pages within one session. Figures 2(b–d) illustrates the web robot detection accuracy (text crawler, email collector and link checker respectively). The proposed approach achieved higher performance than Tan and Kumar's method.

5. Conclusion

Sessions initiated by web robots can be separated into several groups according to their purposes. However, existing approaches, such as the Bayesian approach, statistical analysis and data-mining algorithms, have a limitation in detecting

small groups on a web robot. These approaches were specialized for classifying major web robots such as text crawlers, which constitute a large proportion of all web robot sessions. We therefore extracted and analysed common features and patterns of web robots, to detect various IP groups with high performance.

The pattern-based web robot detection method proposed in this paper was highly effective in distinguishing humans from web robots. IPs belonging to web robots had different patterns from those humans, owing to the cooperative relationship of web robots between groups that had different purposes. Empirical evaluation of the idea, based on the web server log collected at www.microsoft.com, which contained more than 1 billion requests, confirmed the effectiveness of the detection features.

Real-time detection of web robots seems feasible, because the proposed features are relatively simple to compute. A real-world demonstration remains a major task for future research. Furthermore, several factors were left unaddressed. We were unable to perform temporal analysis. If similar analysis is repeated periodically, further insight may be derived. IP-based analysis, rather than the company-based analysis used in this paper, would also be an interesting topic to research.

Acknowledgements

The authors would like to thank the Microsoft Corporation, and MSRA UR in particular, for its generous support, without which the research reported in this paper could not have been performed. MSRA provided us with raw data as well as a research grant. This research was supported by the National IT Industry Promotion Agency (NIPA) under the programme of Software Engineering Technologies Development.

References

[1] Tan P-N and Kumar V. Discovery of web robot sessions based on their navigational patterns. *Data Min Knowl Disc* 2002; 6: 9–35.

[2] Lourenco A and Belo O. Catching web crawlers in the act. In: *Proceedings of the International Conference on Web Engineering (ICWE)*, Palo Alto, California, USA, 11–14 July 2006, pp. 265–272.

[3] Bomhardt C, Gaul W and Schmidt-Thieme L. Web robot detection: preprocessing web logfiles for robot detection. In: Bock H-H, Gaul W, Vichi M, Arabie Ph, Baier D, Critchley F, Decker R, Diday E, Greenacre M and Lauro C et al (eds). *New Developments in Classification and Data Analysis*. Berlin: Springer, 2005, pp 113-124. (Studies in Classification, Data Analysis, and Knowledge Organization).

[4] Stassopoulou A and Dikaiakos M. Web robot detection: a probabilistic reasoning approach. *Comput Netw* 2009; 53(3): 265–278.

[5] Almeida V, Menasce D, Riedi R, Peligrinelli F, Fonseca R and Meira W Jr. Analyzing web robots and their impact on caching. *Proceedings of the Sixth International Workshop on Web Caching and Content Distribution*, Boston, USA, 20–22 June 2001, pp. 299–310.

[6] Huntington P, Nicholas D and Jamali H. Web robot detection in the scholarly information environment. *J Inf Sci* 2008; 34: 726–741.

[7] Duskin O and Feitelson D. Distinguishing human from robots in web search logs: preliminary results using query rates and intervals. *Proceedings of the Workshop on Web Search Click Data*, Barcelona, Spain, 9–12 February 2009, pp. 15–19.

[8] Doran D and Gokhale S. Web robot detection techniques: overview and limitations. *Data Min Knowl Disc* 2011; 22: 183–210.

[9] Ye S, Lu G and Li X. Workload-aware web crawling and server workload detection. *Proceedings of the Second Asia-Pacific Advanced Network Research Workshop*, 2004.

[10] Lee J, Cha S, Lee D and Lee H. Classification of web robots: an empirical study based on over one billion requests. *Comput Secur* 2009; 28(8): 795–802.

[11] Lin X, Quan L and Wu H. An automatic scheme to categorize user sessions in modern HTTP traffic. *Proceedings of the Global Communications Conference (GLOBECOM)*, Los Angeles, CA, USA, 30 November – 4 December 2008, pp. 1–6.

[12] Geens N, Huysmans J and Vanthienen J. Evaluation of web robot discovery techniques. *Lect Notes Artif Intell* 2006; 4065; 121–130.