Alexa rankings

PAUL WARING
School of Computer Science, University of Manchester

1. INTRODUCTION

When analysing sites for events, we want to ensure that we cover a significant number of sites which users are likely to visit. The most obvious way to achieve this is by selecting sites which are 'popular' amongst the majority of Web users – i.e. those sites which receive more traffic than the rest. Fortunately, the basic data pertaining to popular sites is freely available from several sources on the Web.

2. THE ALEXA RANKING SYSTEM

One site which offers ranking data for popular sites is Alexa,[1] which has been tracking traffic to web sites for several years and is now owned by Amazon.com. The statistics are obtained by users downloading a toolbar for their browser, which sends data such as the URL accessed and the amount of time spent on a page to Alexa, where it is then processed to obtain rankings for individual websites.

One reason why Alexa is a useful source for ranking websites is that its data is freely available, both in terms of ease of access and cost. The list of the top 500 sites globally is available on the Alexa site without registration or a fee, as are lists of the top 100 sites for each country and language. Although Alexa does charge a fee for some data services, such as customised reports and its API, all of the sources which we will be using are free of charge. This factor will allow other researchers to verify our results and extend our work at a later date.

In addition to the Alexa data being freely available, it also accessible in a machine-readable format. Each site listing in the top 500 results shares a common format, so it is a trivial matter to extract this data programatically. The same applies to the top 100 sites listed by country or language, which means that we can always obtain the latest rankings by running a simple script to extract this information from the relevant page on Alexa's website.

Alexa also offers global statistics, whereas some alternative sites, such as Compete,[2] only track statistics for visitors in the US. As a result, Alexa largely avoids any bias which might be introduced as a result of country-specific interests, and also includes sites which are popular to a global audience, such as BBC News.[3]

Finally, the Alexa ranking system has been utilised by a number of scholars in the existing literature in a wide variety of areas, including transforming Web pages to become standards-compliant [Chen and Shen 2006], segmenting Web pages for mobile devices [Hattori et al. 2007], measuring privacy loss and protection [Krishnamurthy et al. 2007] and testing the reliability of the Domain Name System

[1] `http://www.alexa.com`
[2] `http://www.compete.com`
[3] `http://news.bbc.co.uk`

[Ramasubramanian and Sirer 2004].

2.1 Potential issues with Alexa

Whilst the Alexa ranking system offers a number of benefits and, as has already been discussed, is pften used in the literature as a way of selecting sample websites, it is not without its problems. One key fact to bear in mind is that Alexa does not differentiate between subdomains, so it will treat `news.google.com` and `blogsearch.google.com` under the general domain of `google.com`. For most sites this is not a problem, as either subdomains are not used (other than the standard `www`), or the site is intended to be treated as a whole anyway, despite the use of subdomains. However, for large sites which use . For example, it is possible that `yahoo.com` as a domain receives very little traffic, as the majority of links from the site point to pages which are hosted on `news.yahoo.com`, `sports.yahoo.com` and other Yahoo! sites. As a result, `yahoo.com` has an inflated popularity ranking, which is not representative of the amount of traffic it receives. This is not necessarily a problem in itself, but it is something which we need to be aware of when analysing the results.

3. FILTERING ALEXA RANKINGS

In order to obtain a selection of popular sites from which to choose pages to extract event information from, we will extract the top 100 English sites from Alexa[4] and apply a number of filters to remove sites which are not relevant to our research.

Firstly, we will be removing all sites which are solely search engines (e.g. Windows Live Search[5]) or web directories (e.g. Open Directory Project[6]). Our reasoning for doing so is two-fold. First of all, such sites contain only a small amount of content pages – the majority of pages encountered by users are either dynamically generated search results in response to a query, or a hierachy of links to other sites, and therefore these sites are unlikely to contain a significant amount of information about events. Furthermore, sites which are predominantly search engines or web directories are used not as information points in themselves, but to find sites which may provide the information which a user is looking for [Lindemann and Littig 2006].

The second filter which we shall apply is to remove sites which predominantly consist of videos, images and other non-textual content, such as YouTube[7] and Flickr.[8] In a similar vein to search engines and web directories, video and image sites contain little in the form of textual information which we can analyse to extract data related to events. Whilst the videos and images offered on these sites may well contain a significant amount of information about events, the extraction of this data is beyond the scope of our work and could form a separate project in itself.

Thirdly, we will remove all sites which are using any language other than English, as attempting to support multiple languages would complicate the project to an

[4] `http://www.alexa.com/site/ds/top_sites?ts_mode=lang&lang=en`
[5] `http://www.live.com`
[6] `http://www.dmoz.org`
[7] `http://www.youtube.com`
[8] `http://www.flickr.com`

Paul Waring – pwaring@cs.man.ac.uk

extent to which it would not be completed within three years. However, support for languages other than English is an interesting potential area for future work, which could be built upon at a later date.

Finally, any sites which require users to login in order to access the majority of content will be excluded from our rankings. Our reasoning for this is down to two factors – firstly, it may not be possible to fetch a particular page and extract event information for it if some form of authentication (which we cannot guarantee to have) is required. Secondly, we do not wish to redirect users to pages containing potentially related events if they cannot access this information without logging in.

4. MODIFIED ALEXA RANKINGS

We have taken the Alexa top 100 sites in English and applied our filters described previously.[9] The results are shown in Table I.

Table I: Modified Alexa rankings

Alexa rank	Domain	Include in our ranking?	Reason	Our rank
1	yahoo.com	Yes	n/a	1
2	google.com	Yes	n/a	2
3	youtube.com	No	Non-textual content	-
4	live.com	No	Search engine	-
5	msn.com	Yes	n/a	3
6	myspace.com	Yes	n/a	4
7	facebook.com	No	Login required	-
8	blogger.com	Yes	n/a	5
9	orkut.com	No	Login required	-
10	rapidshare.com	No	Non-textual content	-
11	microsoft.com	Yes	n/a	6
12	google.co.in	Yes	n/a	7
13	ebay.com	Yes	n/a	8
14	hi5.com	No	Login required	-
15	aol.com	Yes	n/a	9
16	google.co.uk	Yes	n/a	10
17	photobucket.com	No	Non-textual content	-
18	amazon.com	Yes	n/a	11
19	imdb.com	Yes	n/a	12

[9]By using the English list, sites in other languages have, for the most part, been automatically filtered out by Alexa.

20	imageshack.us	No	Non-textual content	-
21	youporn.com	No	Non-textual content	-
22	wordpress.com	Yes	n/a	13
23	flickr.com	No	Non-textual content	-
24	friendster.com	Yes	n/a	14
25	adultfriendfinder.com	No	Login required	-
26	go.com	Yes	n/a	15
27	bbc.co.uk	Yes	n/a	16
28	craigslist.org	Yes	n/a	17
29	dailymotion.com	No	Non-textual content	-
30	redtube.com	No	Non-textual content	-
31	cnn.com	Yes	n/a	18
32	mininova.org	No	Non-textual content	-
33	google.ca	Yes	n/a	19
34	fotolog.net	No	Non-textual content	-
35	imagevenue.com	No	Non-textual content	-
36	espn.go.com	Yes	n/a	20
37	rediff.com	Yes	n/a	21
38	adobe.com	Yes	n/a	22
39	apple.com	Yes	n/a	23
40	yourfilehost.com	No	Non-textual content	-
41	veoh.com	No	Non-textual content	-
42	perfspot.com	Yes	n/a	24
43	deviantart.com	No	Non-textual content	-
44	about.com	Yes	n/a	25
45	megaupload.com	No	Non-textual content	-
46	metroflog.com	Yes	n/a	26
47	fastclick.com	Yes	n/a	27
48	clicksor.com	Yes	n/a	28
49	geocities.com	Yes	n/a	29

50	google.co.id	No	Non-English content	-
51	ebay.co.uk	Yes	n/a	30
52	mediafire.com	No	Non-textual content	-
53	partypoker.com	Yes	n/a	31
54	gamespot.com	Yes	n/a	32
55	download.com	No	Non-textual content	-
56	nytimes.com	Yes	n/a	33
57	google.com.au	Yes	n/a	34
58	weather.com	Yes	n/a	35
59	thepiratebay.org	No	Non-textual content	-
60	ign.com	Yes	n/a	36
61	bebo.com	Yes	n/a	37
62	depositfiles.com	No	Non-textual content	-
63	adultadworld.com	Yes	n/a	38
64	nba.com	Yes	n/a	39
65	zshare.net	No	Non-textual content	-
66	digg.com	Yes	n/a	40
67	4shared.com	No	Non-textual content	-
68	aim.com	Yes	n/a	41
69	netlog.com	No	Login required	-
70	studiverzeichnis.com	No	Non-English content	-
71	isohunt.com	No	Non-textual content	-
72	comcast.net	Yes	n/a	42
73	doubleclick.com	Yes	n/a	43
74	sourceforge.net	Yes	n/a	44
75	usercash.com	No	Login required	-
76	badongo.com	No	Non-textual content	-
77	cnet.com	Yes	n/a	45
78	google.co.th	No	Non-English content	-
79	easy-share.com	No	Non-textual content	-

80	pornhub.com	No	Non-textual content	-
81	megarotic.com	No	Non-textual content	-
82	imeem.com	No	Non-textual content	-
83	gmx.net	No	Non-English content	-
84	metacafe.com	No	Non-textual content	-
85	reference.com	Yes	n/a	46
86	information.com	No	Search engine	-
87	multiply.com	No	Login required	-
88	888.com	No	Login required	-
89	livejasmin.com	No	Non-textual content	-
90	realitykings.com	No	Non-textual content	-
91	torrentz.com	No	Non-textual content	-
92	google.co.za	Yes	n/a	47
93	soso.com	No	Non-English content	-
94	mozilla.com	Yes	n/a	48
95	filefactory.com	No	Non-textual content	-
96	icq.com	Yes	n/a	49
97	brazzers.com	No	Non-textual content	-
98	tinypic.com	No	Non-textual content	-
99	vnexpress.net	No	Non-English content	-
100	hp.com	Yes	n/a	50

As can be seen in Table I, our rankings are a subset of the original Alexa rankings and the filters have removed approximately 50% of the sites.

REFERENCES

Chen, B. and Shen, V. Y. 2006. Transforming web pages to become standard-compliant through reverse engineering. In *W4A: Proceedings of the 2006 international cross-disciplinary workshop on Web accessibility (W4A)*. ACM, New York, 14–22.

Paul Waring – pwaring@cs.man.ac.uk

Hattori, G., Hoashi, K., Matsumoto, K., and Sugaya, F. 2007. Robust web page segmentation for mobile terminal using content-distances and page layout information. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, New York, 361–370.

Krishnamurthy, B., Malandrino, D., and Wills, C. E. 2007. Measuring privacy loss and the impact of privacy protection in web browsing. In *SOUPS '07: Proceedings of the 3rd symposium on Usable privacy and security*. ACM, 52–63.

Lindemann, C. and Littig, L. 2006. Coarse-grained classification of web sites by their structural properties. In *WIDM '06: Proceedings of the 8th annual ACM international workshop on Web information and data management*. ACM, New York, 35–42.

Ramasubramanian, V. and Sirer, E. G. 2004. The design and implementation of a next generation name service for the internet. In *SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM, New York, 331–342.