

# Analysing log files

Yue Mao ([mxxvue002@uct.ac.za](mailto:mxxvue002@uct.ac.za))

Supervisor: Dr Hussein Suleman,

Kyle Williams,

Gina Paihama

University of Cape Town

## ABSTRACT

A digital repository stores a collection of digital objects that can be accessed from other computers via networks and has become widely used as academic storage libraries nowadays. This project explores the usefulness of analyzing Web log files for improving the use of digital repositories. A log analysis tool collects information about visitors from the Web log files, summaries them and produces a broader overview of useful statistic. Evaluation suggests that the analysis of log files can give the user overviews and characteristics of the repository and information about the background of visitors, and thus improve the use of the digital repository.

## KEYWORDS

Repository, log analysis, statistics, Dspace, Eprints, digital libraries.

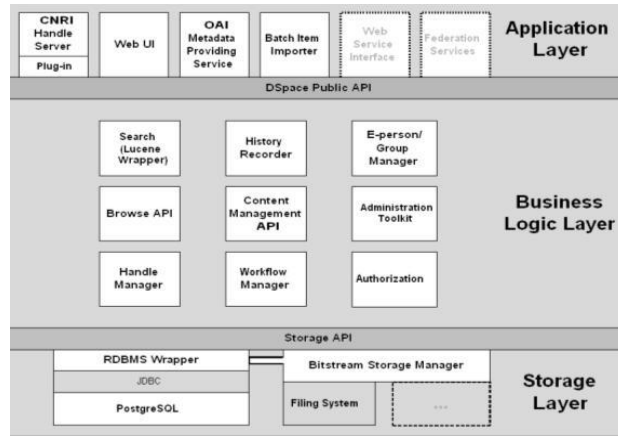
## 1. INTRODUCTION

A digital repository contains collections of organised electronic data, and can be accessed by users on network. The aim for this project is to analyse Web log files and enhance the use of digital repositories. Web servers maintain log files with all the requests made by visitors. A generic Web log contains a list of actions of the visitor on the website, including the resources accessed. A generic log analysis tool analyses and summarises visitor activity on the website. The result usually shows hits, access time, bandwidth, and indicates who, when and how the website is visited. The tool is useful for professionals who want to improve their websites and grow them rapidly. This project emphasises the analysis of Web log files to improve the use of digital repositories. The tool is written in Java, and the analysis cases are designed specifically based on how the resources in the repository are accessed. The tool was written to test the usability, reach and popularity of a digital repository. Section 2 will discuss the background to this project followed by a description of the design in section 3. Section 4 will discuss the experiment designed for testing and section 5 shows the result of evaluation from users.

## 2. BACKGROUND

As Greenstein said in *The Digital Library: A Biography*, 'A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers[1, 2002].' Repository software has been very popular recently especially in the academic field. A digital repository contains original digital documents, audio and video documents. Resources in the repository can be stored on any computers in the network and accessed locally and remotely within the network [2, 19-12-2009]. Some digital libraries enable search engines to check among all their resources by setting up special pages and sitemaps. As Koehler explained, digital libraries usually use OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) to reveal the metadata to other digital libraries. Alternatively, search engines can also use OAI-PMH to discover these embedded Web resources[3, 2006]. The advantage of a digital repository is that it accepts different file formats with high speed support, multiple access, low space and low cost (free usually).

Examples of repository software are Dspace and Eprints. They are open source software packages which aim at building and managing digital repositories. The following DSpace architecture (see Figure 1) shows the structure of a digital repository. The top layer, which is the application layer covers the interfaces to the system: the web UI and batch loader[4, 01-2003]. Other layers supports a wide range of digital data, which includes books, theses, 3D digital scans of objects, photographs, film, video, research data sets and other forms of content.[6, 09-11-2009] Eprints, on the other hand, is an open source software with flexible platform for building high quality, high value repositories. 'It is recognised as the easiest and fastest way to set up repositories of research outputs of literature, scientific data, theses and reports or multimedia artefacts from collections, exhibitions and performances[7, 09-11-2009].'



**Figure 1. DSpace technical architecture [5, 01-2003]**

The common log file format contains information for every individual interaction of visitors. A log analysis tool collects all the data, summaries it and gives a broader statistic overview for relevant cases. Different log analysis tools focus on different aspects of analysis. Generic log analysis tools focus more on the location of the visitors, number of visitors, and bandwidth statistics. For example, AWStats is a free web analysis software that works with a CGI script on a Web server or using the command line. It can also analyse ftp or mail server statistics, and produce a graphical output [8, 09-07-2009]. This software gives excessive test cases but most of them relate to visitors, not on how resources are accessed. For e-commerce websites, tracking how visitors are moving through the site is more useful for improving the service. This project emphasises the analysis of Web log files to improve the use of digital repositories. It produces some valuable analysis cases which are used to enhance the use of a digital repository. A log analysis program for a digital repository will demonstrate activity statistics about users visiting the digital repository based on log files.

### 3. DESIGN OF SOLUTION

#### 3.1 Research Question

Can we infer useful digital repository statistics largely based on analysis of log files?

The analysis should be directed to digital repositories specifically. Common log analysis focuses more on visitor clicks than the usefulness of the actual resources. A digital repository has more academic focused background and has a stricter data layout. The analysis is meant to focus not only the visitor clicks, but also user groups and whether the resources provided are indeed beneficial for the users. The analysis could thus include the location distribution of users, ranking of most frequently visited files, ranking of searching keywords and information about type of errors. The following are the cases that were included in the project:

1. Address/ DNS - This category gives the distribution of the number of visitors over time and shows the web growth pattern. It also gives the distribution for locations of the visitors which suggests the target of the repository.

- 1-1. Number of unique IP addresses
- 1-2. Rank of most visited countries
- 1-3. Number of unique new visitors per month
- 1-4. Number of visitors per month versus total hits per month

2. Date – This category shows the distribution of time spent and active duration of users using the repository. It suggests if the visitors found the content useful or not, and how the website is being used over different periods.

- 2-1. Average time spent (Average time spent per unique user and average time spent for a visit)
- 2-2. Duration spent distribution (0-1min, 1-10min, 10-30min, 30-60min, >1hour)
- 2-3. Rank of most active duration according to hits (hourly, daily, monthly, yearly)

3. Request – This category gives the most information about how the resources are accessed by the visitors. Number of accesses to the documents shows the usefulness of the repository to the visitors. Checking if visitors download documents after seeing the metadata shows whether they are using the introduction or if they find the metadata is useful.

- 3-1. Average page and file/ PDF views per visitor
- 3-2. Rank of most accessed page and files/ PDF

- 3-2. Distribution of number of accesses to PDFs (i.e. 1-10 times, 10-50 times, 50-150 times, 150-300 times, >300 times)
- 3-4. Percentage of users downloading documents after seeing the metadata
- 3-5. Percentage of monthly document (PDF) downloads after seeing the metadata

The tool analyses the following cases by linking the request with the repository log database. It checks if all the documents in the repository are accessed by the visitors, what keywords visitors are interested in, and checks the factors that can influence the popularity of a file.

- 3-6. From Repository database: Percentage of PDFs accessed in digital repository
  - 3-7. From Repository database: Comparison between number of PDFs submitted and downloads per month
  - 3-8. From Repository database: Relationship between PDFs creation time and their popularity
  - 3-9. From Repository database: Rank of accessed PDFs' keywords
  - 3-10. From Repository database: Rank of accessed PDFs' subjects
4. Status – This category shows the percentage of problem occurred in the website compared to the successes.
- 4-1. Rank of error types
  - 4-2. Comparison between percentage of errors versus successes
5. Volume – This category shows the volume downloaded by all visitors over time.
- 5-1. Average volume downloaded per visitor
  - 5-2. Average monthly volume downloaded per visitor (kilobytes)
6. Referral URL – This category shows how visitors access the website and their referral links, and search engines.
- 6-1. Comparison between number of referral and direct accessing of pages and files
  - 6-2. Comparison between monthly referrals and direct accessing of pages and files
  - 6-3. Rank of most searched phrases and words
  - 6-4. Rank of search engines and Google search engines
  - 6-5. Rank of most referral websites (home pages of the URLs)
  - 6-6. Rank of most referral URL
7. User Agent – This category shows the background system of the visitors so that the website can be improved according to their environments.
- 7-1. Rank of most used browsers
  - 7-2. Rank of most used operating systems
8. Comparison of analysis includes bots (i.e. Googlebot) and excludes bots – It shows the differences between bot's activity and human beings.

## 3.2 Input and Output

### *General Web log file:*

The program is based on console interaction. It reads in a log file name and derives statistics about the visitors from that file. A line in the log file has the following 9 fields (see Figure 2):

```
address  rfc931  authenticatedUser  [date]  "request"  status  bytes  "refererURL"  "userAgent"
```

```
84.251.116.204 - - [24/Nov/2009:22:19:13 +0200] "GET /wp-content/themes/img/header2.jpg HTTP/1.1" 200 3563
"http://kylewilliams.co.za/style.css" "Mozilla/5.0 (X11; U; Linux x86_64; en-US; rv:1.9.0.15) Gecko/2009102815
Ubuntu/9.04 (iamtv) Firefox/3.0.15"
```

**Figure 2. Example of general Web log format**

### Repository log file:

The program asks the user if they want to analyse a repository log file. This is optional so it does not matter if no repository log files exists. The functionality of the analysis tool can be enhanced by including the log files of digital repository systems. The example log file used is extracted from an Eprints database with the following fields (see Figure 3):

eprintid	datestamp	authors	keywords	subjects	title
18	2003-7-11	Suleman,,Hussein	NDLTD, ETD, protocols	H.3 INFORMATION STORAGE AND RETRIEVAL	Leveraging OAI Harvesting to Disseminate Theses

Figure 3. Example of repository database file format

### Output

The output is a report in HTML that includes a collection of numbers, percentages and tables of the summarised result from the analysis of log files. See Appendix for details.

### 3.3 System Development

The program was written in Java. It consists of a main file and several other class files for log fields interpretation. One file works specifically for analysis and one works with exporting to HTML. When the program obtains the data file, fields are separated by spaces. The program is designed to rearrange the data as it reads because in this way sorting is faster. Binary search is used for inserting and rearranging fields. The program first finds the approximate positions of the variable in a sorted list using binary search, and then inserts the new variable into the list.

Algorithm (see Figure 3): Read the log file → separate lines by space → get all the tokens → validation of each field → sort according to different fields → interpret → store result in separate array → analyse repository data → write to HTML file.

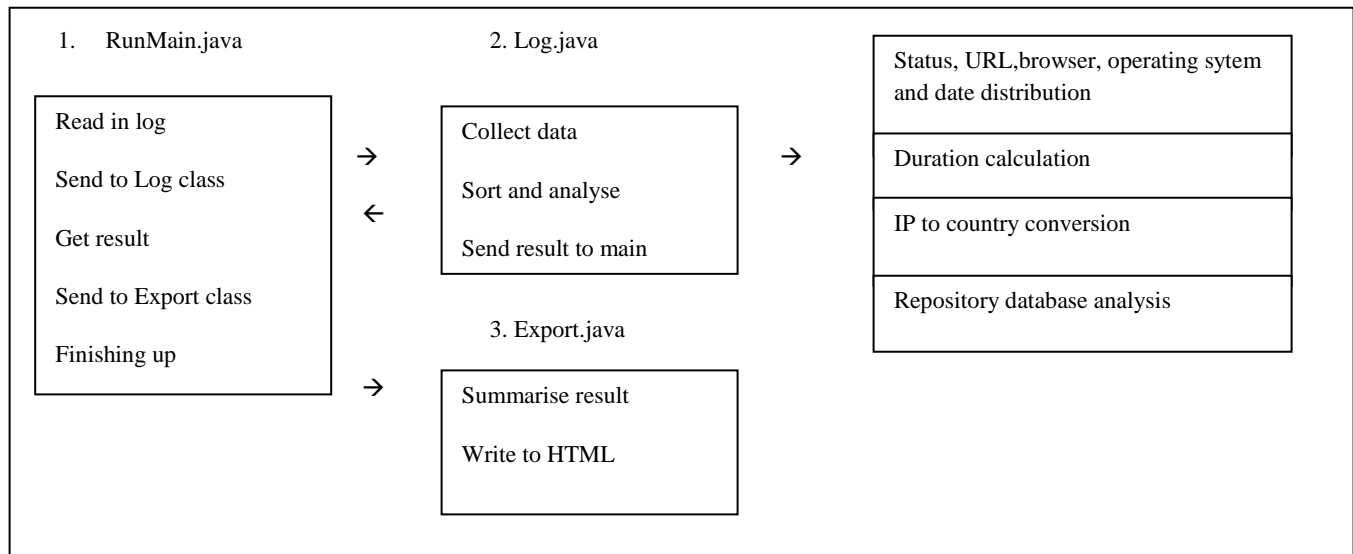
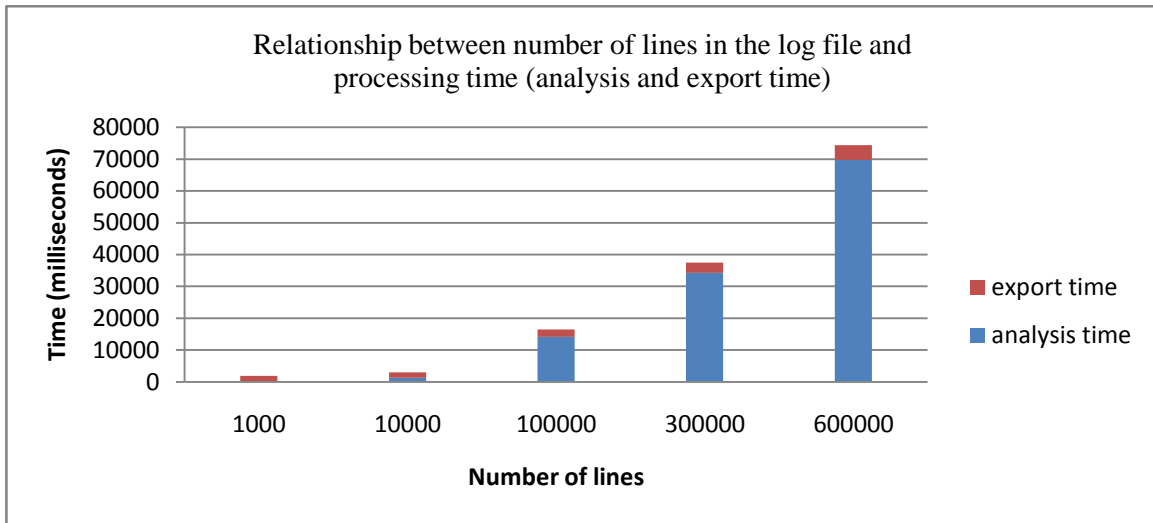


Figure 4. Algorithm diagram for the program

## 4. EXPERIMENTS

The program was tested using different general Web log files with various size.

For extreme cases, the program was tested on different size log files. A summary is given below (see Figure 4) of the relationship between the number of lines in the log file and processing time. The time takes for the algorithm is  $O(n^2)$ .



**Figure 5. Relationship between number of lines in the log file and processing time (analysis and export time)**

For reliability, format of the file should be consistent. It is checked for every hit by the visitor. If the format is incorrect the system tries to correct it or if it still cannot re-order, just skips to the next line.

For evaluation of presentation purposes, the output was checked for readability by different readers. Ten typical analysis cases were included in the evaluation, and a summary (in HTML format) from the output was attached for users to determine if the analysis is useful for the repository.

## 5. RESULTS

The evaluation survey about the log analysis tool (see Figure 6) was answered by three people. The table shows the average scores for the understandability (if the user can understand the case and result), usefulness (if the result is useful for showing the target market and how the resources are accessed which related to the repository), and interpretability (if the user can relate the result for improving the digital repository). One or two test cases from each category are selected here to avoid bias. Most people agreed that the analysed log data is very useful for revealing the details of the repository, which implies that the purpose of implementing the log analysis tool has been achieved. At the same time, there are also some suggestions about the tool:

- Some test cases still need more clearer clarifications.
- A brief summary of the result would be useful.
- Separated result for bots and people is useful. Some people are interested in not only human beings' behaviour but also tracking the action of bots. According to the result (see Appendices, test case 2-2), bots take more time on the web than human beings.
- More information about directories under errors occurred can be useful.
- For presentation, some people suggested an easier navigation of the heading is needed: a sidebar or separated pages for different sections of results are examples of how this can be achieved

Scores: 0 - 5 (means disagree – agree)			
Test cases	Understandable?	Useful?	Interpretable?
1-3. Number of unique new visitors per month This shows if the resources are useful and the website is attracting new people.	4.67	4.67	3.33
1-4. Number of visitors per month versus total hits per month This shows the period when visitors access the most resources on the website.	4.33	4.67	5
2-2. Duration spent distribution This suggests how visitors use the website.	5	4	4.67
2-3. Rank of most active duration according to hits This shows the busiest period over day, month, and year thus suggests the target	4.67	5	4.33

visitors and whether the website has grown.			
3-3. Distribution of number of access to PDFs This shows how resources are being used over time and suggests if the resources in the repository are useful overall.	2.67	2.5	2.5
3-5. Percentage of monthly accesses to PDFs after seeing the introduction versus just accessing from index This suggests if the visitor is using metadata efficiently and the comparison between the two cases.	4	3	3.67
4-2. Rank of error status types This shows the most problematic error statuses of the website.	4	4.33	4.67
5-2. Average monthly downloading volume per visitor (kilobytes) This suggests the maximum bandwidth needed for the website and the busiest times.	5	3.33	4.67
6-4. Rank of search engines and referral websites This shows the most used search engines and referral websites.	5	3	4.67
7-1. Rank of most used browsers This shows the working environment of visitors.	5	4.67	5

**Figure 6. Table of evaluation result for the log analysis tool**

According to the evaluation, distribution of number of access to PDFs is not very useful (2.5) and hard to interpret (2.5). The reason might be the user could not make any improvement according to the result. Every website has some resources gaining more hits than the others. There is no comparison between other website so this test case is not very useful here.

Percentage of monthly accesses PDFs after seeing the introduction versus just accessing from index is not very useful either (3). According to the result, the different does not depend on the quality of resources, but the period. So there is no related work needs to be done here.

Users found number of visitors per month versus total hits per month, duration spent distribution, rank of most active duration according to hits, rank of error status types, and rank of most used browsers are useful, understandable and interpretable suggesting that they added the most value to the log analysis tool. The evaluation survey gains positive responses proved that the statistics can provide the developers and the administrators more in-depth understanding about the user groups, data usability, and also the weaknesses of the data provided, so as to improve the data quality and layout of the digital repositories.

## 6. CONCLUSIONS

This project created a log analysis tool for improving the use of digital repositories. It determined the area of weaknesses of the repository based on statistical analysis of web logs. The evaluation proved that analysis of log files is effective in inferring useful digital repository statistics based on the output of this tool. Management can improve the service using data on status and system requirements of the visitors. Administrators can know the busy time of websites, country distribution and duration of each visit. Researchers can check the usefulness of the service by popularity of resources, type of resources accessed and search phrases. The finding in this paper suggest that the use of digital repositories can thus be improved by indicators from an analysis tool.

## 7. FUTURE WORK

- Adding more vertical bar graph analysis. For example: monthly volume downloaded compared with monthly visitors.
- More clear indication on ranking of operating systems.
- Adding navigation sidebar or separate pages for different sections of results.
- Including analysis for repository log files for EPrints

## 8. REFERENCES

- [1] Greenstein, Daniel I., Thorin, Suzanne Elizabeth. The Digital Library: A Biography.[2002] [Accessed 02-01- 2009]:. DOI=  
<http://www.clir.org/PUBS/reports/pub109/pub109.pdf>
- [2] Digital library - Wikipedia, the free encyclopedia [Last updated 19-12-1009] [Accessed 02-01- 2009]. DOI=  
[http://en.wikipedia.org/wiki/Digital\\_library](http://en.wikipedia.org/wiki/Digital_library)
- [3] Koehler AEC. Some Thoughts on the Meaning of Open Access for University Library Technical Services Serials Review Vol. 32, 1, 2006, p.17
- [4] Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, Julie Harford Walker, DSpace An Open Source Dynamic Digital Repository. D-Lib Magazine, January 2003, Volume 9 Number 1, ISSN 1082-9873, [Accessed 29-01-2010]. DOI= <http://www.dlib.org/dlib/january03/smith/01smith.html>
- [5] Figure 1: DSpace technical architecture, [Last updated 10-01-2003] [Accessed 29-01-2010]. DOI=  
<http://www.dlib.org/dlib/january03/smith/fig-2.gif>
- [6] DSpace - Wikipedia, the free encyclopedia, [Last updated 09-11-2009] [Accessed 28-12-2009]. DOI=  
<http://en.wikipedia.org/wiki/DSpace>
- [7] Open Access and Institutional Repositories with Eprints [Last updated 09-11-2009] [Accessed 28-12-2009]. DOI=  
<http://www.eprints.org/>
- [8]. Partho. Top 10 Web Log Analysis Software [Last updated 09-07-2009] [Accessed 07-12-2009]. DOI=  
<http://blog.taragana.com/index.php/archive/top-10-web-log-analysis-software/>

## 9. BIBLIOGRAPHY

- [1]. Logging in W3C httpd, httpd@w3.org, [Last updated July 1995]. DOI=  
<http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format>
- [2]. IP to country database [Last updated 05-12-2009 02.15]. DOI=  
<http://www.iporigin.net/>
- [3]. Demo for web server log files (AWStats), [Last updated 05-12-2009]. DOI=  
<http://www.nltechno.com/awstats/awstats.pl?config=destailleur.fr>
- [4]. Stevan Harnad. EPrints, DSpace or Espace. at ecs.soton.ac.uk [Last updated 20-03-2004] [Accessed Jan 2, 2009]. DOI=  
<http://www.bio.net/bionet/mm/jrnlnote/2004-March/002230.html>

## 10. APPENDICES

### 10.1 Sample Input

Please type in the log file name: httpd-access.log

Do you want to analysis with repository database? (Y/N) Y

Fields in log file has to be separated by tab, in the following format:

eprintid    timestamp            authors    keywords            subjects

Please type in the repository log file name:

logfile-tabDelimited.txt

Type 1 for including Bots only

Type 2 for excluding Bots (human action only)

Type 3 for including all cases

## 10.2 Sample output:

Processing...

Unique visitors = 7899, referURL = 4172, browser = 244, os = 1819 , total visitors = 44735

skipped (format error) = 0

Analysed result has been written to 'Result-all.htm'






## 10.3 Output file:

### 1. Address/ DNS

#### 1-1. Number of unique IP addresses






Total unique IP addresses is **7408**.

#### 1-2. Rank of most visited countries

Rank	Value	Percentage of most visited countries	Counter
1	UNITED STATES		2721
2	SOUTH AFRICA		1763
3	CHINA		460
4	UNITED KINGDOM		372
5	INDIA		212

### Date Result

#### 2-2. Duration spent distribution

Rank	Value	Distribution of time spend	Counter
1	0-1min		5187
2	1-10min		917
3	10-30min		396
4	30-60min		223
5	>60min		685

### Comparison with bots activity summary

#### Date Result

#### 2-1. Average time spent

Average time spent per unique user is **968 minutes**

(It shows the average accumulative time spent per unique visitor.)






Average time spent for a visiting is **33 minutes**

(It shows the average time spent per visitor.)

**Note:** When the time gap for two hits of a visitor is bigger than 2 hours counts as a new visit.

(The time would be different because some users visited repository several times.)

#### 2-2. Duration spent distribution

Rank	Value	Distribution of time spend	Counter
1	0-1min		179
2	1-10min		64
3	10-30min		61
4	30-60min		24
5	>60min		254



## Request Result





### 3-1. Average page and file(PDF) views per visitors

Average page views is **9** per visitor

Average document (pdf,zip) views is **12** per visitor

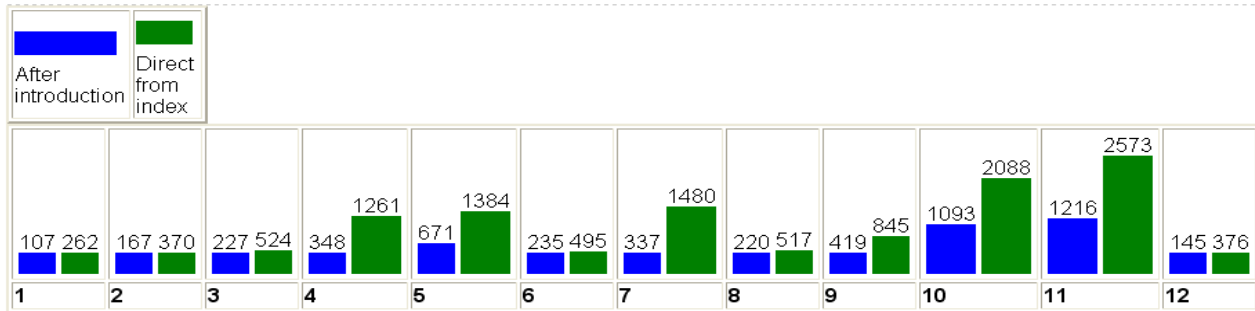
### 3-2. Rank of most accessed page and file(PDF)

Rank of most accessed page

Rank	Value	Percentage of access counter	Counter
1	/		234432
2	/~honsproj/		5185
3	/%7Ehonsproj/		4839
4	/~honsproj/2006/		2638

### 3-5. Percentage of monthly document(PDF) downloads after seeing the metadata






versus just downloading from the index



[Go back to index](#)

## 4. Status Result

### 4-1. Rank of error status types

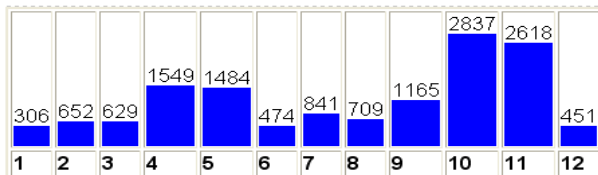
Rank	Value	Percentage of error types	Counter
1	404		26413
2	301		10932
3	304		8070
4	206		3048
5	400		1300

## 5. Volume Result

### 5-1. Average downloading volume per visitor

Average downloading volume per visitor is **5000 Kilobytes**






### 5-2. Average monthly downloading volume per visitor(kilobyte)



[Go back to index](#)






## 6. Referred URL

### 6-5. Rank of most referred websites (home page of the URLs)

Rank	Value	Percentage of referred websites	Counter
1	http://shenzi.cs.uct.ac.za		7850
2	http://pubs.cs.uct.ac.za		5293
3	http://www.cs.uct.ac.za		254
4	http://people.cs.uct.ac.za		157
5	http://www.husseinspace.com		122

## 7. User Agent Information

### 7-1. Rank of most used browsers used

Rank	Value	Percentage of browsers	Counter
1	"Mozilla/5.0		3681
2	"Mozilla/4.0		1997
3	"-		326
4	"msnbot-media/1.1		181
5	"Baiduspider+ (+http://www.baidu.com/search/spider.htm)		70

## 10.4 Details of experiment

Relationship between number of lines in the log file and processing time  
(analysis and writing time)

number of lines	analysis time	expot time	total time
10	31	1703	1734
100	78	1672	1750
1000	250	1657	1907
10000	1297	1672	2969
100000	14203	2313	16516
300000	34265	3266	37531
600000	69797	4625	74422