Status of the African Web

Rizza Camus Caminero
Language Observatory
Nagaoka University of Technology
Nagaoka, Niigata, 940-2188 Japan
+81-258-47-9355
055933@mis.nagaokaut.ac.jp

Pavol Zavarsky
Language Observatory
Concordia University
Edmonton, AB, T6A 1W1, Canada
+1-780-413-7810
pavol.zavarsky@concordia.ab.ca

Yoshiki Mikami
Language Observatory
Nagaoka University of Technology
Nagaoka, Niigata, 940-2188 Japan
+81-258-47-9355
mikami@kjs.nagaokaut.ac.jp

ABSTRACT

As part of the Language Observatory Project [4], we have been crawling all the web space since 2004. We have collected terabytes of data mostly from Asian and African ccTLDs. In this paper, we present results of the current status of the African web and compare it with its status in 2004 and 2002. This paper focuses on the accessibility of the web pages, the web tree growth, web technology, privacy protection, and web interconnection.

Categories and Subject Descriptors

K.1 [**Computing Milieux**]: The Computer Industry - Statistics

General Terms: Experimentation, Measurement, Security.

Keywords: ccTLD, Africa, interconnection, Internet statistics, web tree, web graph, web accessibility, privacy protection

1. INTRODUCTION

The Language Observatory was planned primarily to provide means for assessing the usage level of each language in cyberspace. Using UbiCrawler [2], a scalable fully distributed web crawler developed for use in the project, we have collected a total of 84.6 million web pages from more than 142 thousand sub-domains of 60 country code Top Level Domains (ccTLDs) of Africa. The information downloaded by the crawler is enormous, and only a small portion of it is presented in this paper. This endeavor has proven to be beneficial to individuals and organizations conducting research and developing solutions on bridging the digital divide, such as UNESCO and ACALAN (African Academy of Languages).

2. SURVEY STATISTICS

During our most recent crawling, in December 2005, we have collected 84.6 million pages from the whole of Africa, among them South Africa alone occupies more than 80%. The growth of the pages during the last three years reached as high as 251.17% per year. Due to the space limitation, the paper picks up only a few findings from the survey. For more detailed survey statistics data, refer to the project website [5].

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

Table 1. Collected pages

	February 2002	December 2005	Annual Growth
All of Africa	1955524	84689415	251.17%
South Africa (.za)	1609722	69815327	251.34%
The rest of Africa	345802	14874088	250.38%

Note: 2002 data are taken from [1]

2.1 Accessibility of the Web Pages

We have analyzed responses to HTTP requests to get an insight into African web servers' accessibility and possible network-related problems. Top Level Domains with network-related problems are Somalia and Guinea-Bissau, from which all pages were unavailable during the experiment. On the other hand, pages stored on the web servers registered in Benin, Gambia, and Libyan Arab Jamahiriya maintained a high success rate of 95-99% even as their number of pages stored in the web server increased by almost 50% in a year. Coincidentally, responses to requests of web pages from Mali and Djibouti web spaces had similar high success rates, their number of pages with even more increases at 75% and 137%, respectively. There was almost no problem in downloading pages from the web space of the following African countries: Democratic Republic of Congo, Reunion Island, Eritrea, Burundi, Seychelles, and Rwanda.

2.2 Web Tree Growth

There is a complete freedom in how Africa's ccTLDs allocate their sub domains. We are analyzing how the African web tree is organized. As an example, during our experiment, Morocco topped the list with 801 second-level sub-domains, followed by Sao Tome & Principe and Senegal with 353 and 315 second-level sub-domains, respectively. On the other hand, Mozambique and Nigeria had no change in their number of second-level sub-domains from 2004 to 2005.

Out of the 142 thousand domains which our crawler has visited in the African web space, Libyan Arab Jamahiriya's number of domains multiplied 5 times in a period of 1 year, which indicates an exponential growth of the internet in the region. At the same time, the current domain count for Gambia almost doubled compared to the previous year's count. Almost all countries had increases in their number of domains, except for some special cases like Somalia and Guinea Bissau.

2.3 Open Source vs Proprietary Technology

A report about the African web in 2002 [1] showed that 56% of the websites use Microsoft technology. However, in the current survey, Microsoft served in only 22.82% of the pages. With the decrease, it became second only to Apache (71.47%), which formerly occupied 37.95% of the African market.

Table 2. Server types (as of December 2005)

	Apache (%)	Microsoft (%)
All of Africa	71.47	22.82
South Africa (.za)	71.76	23.83
The rest of Africa	70.18	17.80

Table 2 shows that the global trend of preference to Apache-like open source server software applies to whole Africa. A large number of pages were created by freely available PHP technology (55.96%) as compared to Microsoft's Active Server Pages (ASP, 15.36%). On the contrary, the 2002 report indicated ASP as the most common file type.

Table 3. File extension types (as of December 2005)

Extension	%	Extension	%
.php	55.96	.pl	1.18
.asp	15.36	.cfm	0.51
.html	14.54	.jsp	0.47
.htm	6.03	.php3	0.41
.aspx	2.19	.shtml	0.25
.cgi	1.30	.py	0.03

2.4 Persistent and Session Cookie Usage

Monitoring the use of cookies and pointing to their inappropriate use can help protect the privacy of users from abuse [3]. The survey showed that Africa had a low rate of 18.73% of persistent cookies and a relatively high rate of session cookies (42.91%). One possible interpretation of this contradictory fact could be that there are not many e-commerce servers that tend to intensively use persistent cookies to monitor web surfing patterns of the users (typically shown in the case of Saint Helena, one of web hosting server domain). There could also be a few commercial servers that don't use session cookies.

Table 4. Cookie usage (as of December 2005)

	No cookie (%)	Session cookie (%)	Persistent cookie (%)
All of Africa	38.36	42.91	18.73
Saint Helena (.sh)	39.75	2.52	57.73
South Africa (.za)	34.05	45.63	20.32

2.5 Interconnection

The interconnection of the countries of Africa was analyzed through the number of outgoing links from each downloaded web page, where the total number of outgoing links is 3.9 billion. We are interested not only in the internal links (links to the web pages within the African ccTLDs), but also the external links (links to web pages outside the African ccTLDs). All the countries, except British Indian Ocean Territories and Zambia, had a high rate of internal interconnection. British Indian Ocean Territories had more links to .com (53.04%) domain, while Zambia had 42.62% of its links distributed among the generic TLDs. For all TLDs of Africa, internal links is 95.19%, while external links is 4.81%.

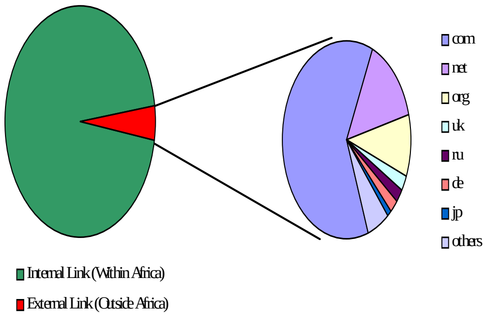

Figure 1. External and internal links of web pages

Figure 1 shows the distribution of the external links of all the African ccTLDs. Interested readers can find much more detailed web graph data on our website [5].

3. CONCLUSION

In this paper, we outlined the status of the African web, how it is growing and adapting to the WWW bandwagon, and which countries are experiencing difficulties. In terms of server technologies, the global trend from proprietary technologies to open source technologies is observed in Africa too. Also, we have indicated the profile of servers and users through the analysis of cookie usage of the servers. The analysis of interconnection of African web has shown that African web is mostly connected within its country domain, and only a few outgoing links goes outside of Africa. The African web is not well connected with outside world.

On our website [5] any interested reader can find additional data which were not possible to include in this paper due to the space limitations.

4. ACKNOWLEDGMENTS

The study was made possible by the financial support of the Japan Science and Technology Agency (JST) under the RISTEX program. We also thank UNESCO for giving official support to the project since its inception.

5. REFERENCES

[1] P. Boldi, B. Codenotti, M. Santini and S. Vigna, Structural Properties of the African Web, in poster proceedings of WWW2002 (Honolulu, Hawaii, USA, May 2002), http://www2002.org/CDROM/poster/164/index.html

[2] P. Boldi, B. Codenotti, M. Santini and S. Vigna, UbiCrawler: A scalable fully distributed web crawler, Software: Practice & Experience, 34(8):711-726, 2004.

[3] Privacy slip on official US sites, http://news.bbc.co.uk/1/hi/technology/4569184.stm

[4] Y. Mikami, P. Zavarsky, et. al., The Language Observatory Project, in poster proceedings of WWW2005 (Chiba, Japan, May 2005), 990-991.

[5] http://gii.nagaokaut.ac.jp/Status_Of_The_African_Web