Country Domain Governance: An analysis by Datamining of Country Domains

Katsuko T. Nakahira, Hiroyuki Namba, Minehiro Takeshita, Shigeaki Kodama, and Yoshiki Mikami

Nagaoka University of Technology, Nagaoka, Japan
(Tel : +81-258-47-9847)
(Email katsuko@vos.nagaokaut.ac.jp)

Abstract: Along with the expansion of opportunities to use the internet, the system of managing it has becom e a serious problem. Even from the global perspective, questions concerning internet governance have been ra ised at various places, starting with the WSIS (World Summit on the Information Society). However, the mai n discussion has been from the viewpoint of administrators and service providers, placing emphasis on ensuri ng safety and equal opportunities of use, while the discussion from the user's point of view is limited. This i s thought to be due to the acquisition of statistical data which the discussion is based on being largely deriv ed from statistics related to the physical means of communication, and to the difficulty of comprehending all user activity in the internet space. In this paper, in response to the challenge of internet governance, we prop ose the country domain governance index. This is an index for comprehensive discussion of governance of in ternet resources from the point of view of both administrators and users by carrying out data mining by regi on by merging various pieces of Web information obtained from results of observations in the Language Obse rvatory with various statistical data.

Keywords: Less than 6 words**.**

I. INTRODUCTION

Along with the progress of the IT revolution that started with the use of the internet, it has been pointed out since the 1990s that a gap in enjoyment of the benefits continues to exist between countries[1]. This gap is termed the "digital divide," but the digital divide is classified into several layers according to the object of research. For example, in [2] the difference in the "place" in which the IT technology is used is identified. Among these, in particular, the digital divide between countries caused by economic disparities and various other gaps, starting with the language barrier (education gap), are addressed, and the need for the elimination of the digital divide is discussed. However, these are all based on statistical values related to the physical means of communication, and do not afford a view of the big picture of the overall disparity.

Mikami et. al. established the Language Observatory (LOP)[3] with the purpose of carrying out observations of the digital divide between languages on the internet. LOP combines Web crawling technology and a language identification engine[4] to retrieve up to 10^9 multi-lingual Web pages, and automatically retrieve data included in a Web page such as URL information, tags, the main body, and Web server response time. By combining the results of these observations and existing statistical data, it is possible to obtain a large amount of information such as the number of Web pages per capita and the number of Web pages by language, the ratio of native language pages, the degree of mixing-up of character codes, and basic research concerning the infrastructure.Some of the results have been published by Nakahira and Mikami[5], but this does not extend to the interpretation of the results obtained nor to the development of an index to systematically handle them. In this paper, we consider country domain governance based on a distribution/growth chart of outdegree, which is one of the data mined LOP observation results, and on an investigation of numbers of links.

II. COUNTRY DOM AIN GOV ERNANCE

Country domain governance (CDG) indicates appropriate governance of a country code domain, but in this paper, it indicates in particular the domain administration policy. The creation of an index is required for investigating CDG. We set up what serves as a CDG index as in Table 1. By carrying out general mining of widely available statistical data and Web crawling data gathered in LOP, these indices can be shown for the first time.

Table 1. CDG Index

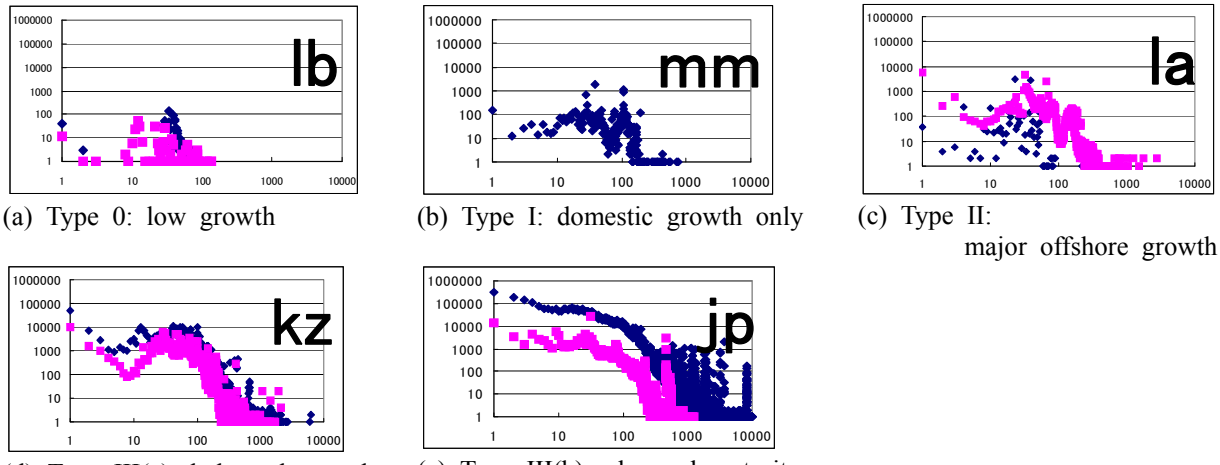
Level		Title	Index
System	Accessible/ Affordable	Accessible and Affordable	Relative Price of DNs to Monthly Income/the Global average price, Number of DNs (published) per Population/ GDP, Ratio of Servers Located Outside of the jurisdiction
	Linguistic Diversity	IDN service	Availability of IDN Service in local language, Number of Active IDNs
	Secure and Trusted	Security, Stability, and Resiliency	Ratio of DNs whose Registrant is Anonymous, …(other)
Content	Accessible/ Affordable	Openness of Network	Ratio of the number of total outgoing links to the number of Outgoing links to news Media
	Linguistic Diversity	Local language use	Number of Local language pages per population, Ratio of total pages to pages written in Local Languages, Linguistic Diversity measured by Lieberman Index/Entropy
	Secure and Trusted	Trustable content	Availability of Dispute Resolution Process, …(other)

III. WEB LINK ANALYSIS AND OUTDEGREE GROWTH CHART

On the other hand, for an investigation carried out of the usual Web space, outdegree distribution is used. Outdegree distribution plots the number of Web pages taking each outdegree value on the vertical axis against the outdegree value on the horizontal axis. The outdegree distribution follows a power law as indicated originally by Broder et. al.[7],but in cases where there are a lot of automatic Web page creation such as with shopping sites, which are coming to form the mainstream these days, or CMS, an outdegree distribution is shown in the form of a power law with a δ function superimposed, because things with particular outdegree values are distributed in greater numbers (Fukuda, [8]).

If this property is used, the state of the Web space that produced the outdegree distribution can be determined simply by looking at the outdegree distribution. Namely, if the outdegree distribution looks like white noise then the Web space is undeveloped, and if it is a power series, the Web space is developing with mutual links. However, with the increase in automatically created sites, as more in-depth services are being provided in the Web space with the formation of portal sites, shopping sites and the like, the Web can be said to be in the mature stage.

These kind of properties are ascertained and the outdegree distribution is produced for each country code top level domain (ccTLD), and the distribution indicated for each state of the Web space is the



(a) Type 0: low growth (b) Type I: domestic growth only (c) Type II: major offshore growth

(d) Type III(a): balanced growth (e) Type III(b) : beyond maturity

Figure 1. Example of an outdegree maturity diagram. The horizontal axis represents the number of links, and the vertical axis the number of pages which have that many links. Black squares represent domestic s ervers, and white squares the offshore servers.

outdegree growth chart. An example is shown in Figure 1.We have roughly divided them into 5 grades from type 0 to type III(b).

VI. EXAMPLE OF OBSERVED CDG I NDEX

Next, we show several actually observed Web space examples based on the CDG Index.

1. Openness of Network

Openness of the network is determined by verifying whether or not the environment allows users within the domain to freely view content outside the domain. Care must be taken over the fact that the guarantee of openness is roughly divided into (1) access restrictions due to censorship, and so on, and (2) social barriers to implementation, for example undeveloped infrastructure. Case (1) is connected to the restriction of the user's free use of the internet, and case (2) is seen as an infrastructure development problem, in advance of domain management problems. Whichever the case, investigation of the openness of the network can be analysed through investigating the number of Web pages belonging to the relevant ccTLD, and the link structure.

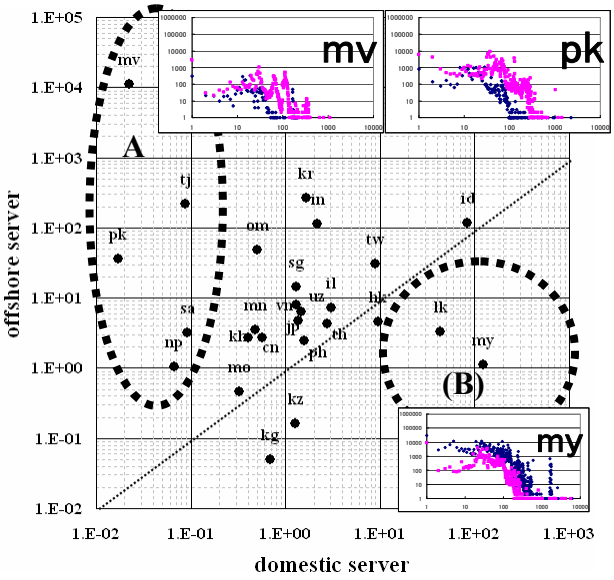


Figure 2. Ratio of the number of total outgoing l inks to the number of Outgoing links to news m edia and outdegree diagrams.

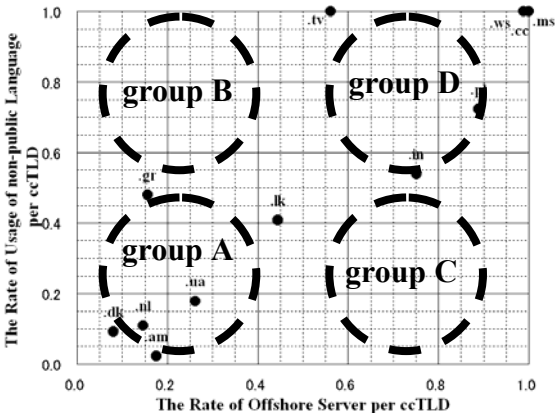


Figure 3. Estimated actual usage of ccTLD

Several results of this are shown. In Figure 2, the number of links to the global news media (GNM) site of each ccTLD to which Web pages belong are graphed according to the physical location of the Web server on

which the Web page resides. In producing this data, the UbiCrawler[6] that LOP uses was run in December, 2009 on approximately 2.3×10^7 Web pages used in 42 countries across Asia.

In Figure 2, the average number of links to the GNM per server from Web pages on domestic servers on the horizontal axis, are plotted against equivalent data for offshore servers on the vertical axis. Therefore, the point at the top left of chart A indicates that the links to the GNM are concentrated on offshore servers, and the point at the bottom right of B indicates that the links to the GNM are concentrated on domestic hosts. The 4 countries fj, np, pk, mv illustrate these possibilities for A, and the 2 countries lk and my for B. The outdegree distribution corresponding to these ccTLDs is also shown in Figure 2.

2. Relationship between Indicators

With the CDG Maturity Index, it is also possible to integrate each of the indicators and analyse them.

An example of this is shown in Figure 3. Figure 3 indicates the percentage of overseas installation of Web servers in each ccTLD and the percentage of non-official language usage in the regions belonging to the ccTLD. This corresponds to visualisation in an identical space of the index and indicated "percentage of servers located outside of the jurisdiction" in Accessibility and Affordable,and the index and indicated "ratio of total pages to pages written in local languages" in Local Language Use.

Figure 3 divides the whole region into 4 groups and gives each group a meaning. In group A, the domestic

infrastructure is basically stable, and a country is envisaged in which installation of a Web server is also possible if desired. For this reason, the ratio for overseas installation of Web servers is low, and also, as it is easy to make use of the internet in the mother tongue, usage of Web pages in the official language or mother tongue is common. In group B, although the domestic infrastructure is stable, for some reason such as users being restricted to overseas or to a minority living in that country, usage of Web pages in a language besides the official language or mother tongue is common. Group C represents cases where the domestic infrastructure is too unstable to be able to set up a Web server within the country, or where the domain for a local server is too expensive to purchase. Even in this kind of country, if mobile communication is developed, Web pages can be placed on overseas servers, and it is conceivable that Web pages in the country's official language or mother tongue could be used. Group D represents countries that do not even consider web page usage within the country, but want to profit by selling domains, or where a minority is able to obtain the country's ccTLD and make use of it.

Through these considerations, by looking in detail at the relationship between indicators, we believe that we can more precisely understand the reality of the digital divide using the CDG Maturity Index.

V. ANA LYSIS T OOL F OR CD G

Through the analysis carried out so far, by using the CDG index, we have shown that we have approached more closely the realities of the management policy for each ccTLD and the realities of the digital divide.

However, to gather the data on which these analyses are based, appropriate analysis tools are required. Accordingly, we have added modules to and carried out preparation for the stable operation of the analysis tool (iGalaksy) currently developed by Arai et. al. [9]. iGalaksy consists mainly of an information input part (including crawling data), an information management part (CDG Index analysis), and an output part, and in one crawl it is possible to gather roughly 10^7-10^8 Web pages (approximately 10GB of crawling data). If these data were, for example, gathered every quarter, the data size would be so huge as to require a database with special data striping. We have implemented this using pgpool-II[10].

VI. CONCLUSION

In this paper, the CDG index is proposed as an observation index by which to home in on the reality of country domain governance. This allows indexing of the the reality of the digital divide in a form that is closer to the conditions of use to be undergone, along with carrying out data mining by region in a form in which the various pieces of Web information obtained from the Language Observatory observation results are merged with the various statistical data. Several observation examples of this have been provided here. In the future, we plan to continue observation of Web space based on this index.

REFERENCES

[1] Ohashi Ikuo,Toward the construction of a global information society: international trends and issues surrounding the digital divide,and creation of intell ectual property 2009(3):6-36.

[2] NTT C&C Foundation, the Digital Divide(in Ja panese), NTT Foundation book: 2002.

[3] Y. Mikami et. al., The Language Observatory Project: Proceedings of the 14th International World Wide Web Conference 2005 :990-991

[4] S. T. Nandasara et. al., An Analysis of Asian Language Web Pages: The International Journal on Advances in ICT for Emerging Regions 2008 01 (01) : 12 - 23

[5] K. T. Nakahira and Y. Mikami, Measuring Lang uage Diversity in Cyberspce: International Conferen ce on Linguistic and Cultural Diversity in Cyberspa ce, Yakutsk, Russia, July 2-4, 2008.

[6] Boldi, P. et. al., UbiCrawler: a scalable full Dis tributed web crawler: Software-Practice & Experienc e, 34(8):711-726.

[7] Broder, A. et. al., Graph Structure in the Web, Computer Networks: The International Journal of Computer and Telecommunications Networking AR chive 33: 1-6

[8] Kensuke Fukuda (2004), Analysis of Statistical Properties of WWW (in Japanese), IPSJ SIG Tech nical Report, 2004(136): 17-22.

[9] Y. Arai et. al., A survey on language distrib ution on the internet : The 8th Proceedings of Forum on Information Technology 2008-4:529-532.

[10] http://pgpool.projects.postgresql.org/