ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Building a Terabyte-scale Web Data Collection "NW1000G-04" in the NTCIR-5 WEB Task

Masao Takaku, Keizo Oyama, Akiko Aizawa, Haruko Ishikawa, Kengo Minamide, Shin Kato, Hayato Yamana, and Junya Hayashi

NII-2006-012E
Sept. 2006

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Building a Terabyte-scale Web Data Collection "NW1000G-04" in the NTCIR-5 WEB Task

Masao Takaku[1]*, Keizo Oyama[2], Akiko Aizawa[2], Haruko Ishikawa[2]**, Kengo Minamide[3], Shin Kato[3], Hayato Yamana[3,2], and Junya Hayashi[4]

[1] Research Organization of Information and Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
[2] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
[3] Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, Japan
[4] TriAx Inc.
4F Nishi-Sando Bld., 4-29-3 Yoyogi, Shibuya-ku, Tokyo, Japan

Abstract. We built a terabyte-scale web data collection, NW1000G-04, which was used in the NTCIR-5 WEB task. This report describes the process of building the collection and some statistics of it in detail.

1 Introduction

The authors conducted the Navigational Retrieval Subtask (Navi-2) [1] in the WEB task at the NTCIR-5 Workshop, which was held from August 2004 to December 2005. In this report, we describe about a dataset which were built for Navi-2. The new terabyte-scale dataset, called "NW1000G-04", consists of text data of web pages more than one terabyte crawled from .jp and other domains from January 2004 to January 2005.

In the WEB Task at the NTCIR-3 Workshop, we built a web data collection, NW100G-01 [2], and re-used it at the NTCIR-4 Workshop. It consists of about 100GB text data of web pages crawled from .jp domains in 2001.

Since the Web has been exponentially increasing its amount, more realistic web data collection has been needed for Web retrieval experiments such as NTCIR. In fact, Saeki et al. [3] estimated that the total amount of the Web in .jp domain was around 292 million pages and 13.3TB, as of February 2004. The National Diet Library of Japan [4] also estimated that the Web in .jp domain was around 450 million pages and 18.4TB, as of March 2005. These estimates limit their scopes only for the ".jp" domains, i.e. they do not include other domains such as ".com" domain even though many Japanese companies and others use them.

We decided to make a new terabyte-scale collection for the NTCIR-5 Web Task. We intended that this new terabyte-scale dataset would represent major parts of the "Web of Japan". Since the Web has open and decentralized nature, it is difficult to simply define the scope of the document dataset, namely the Web of Japan. The Web of Japan means neither the Web pages written in Japanese language, nor the Web pages hosted on the servers located in Japan. We use "Web of Japan" instead of "Japanese Web", as the latter could be mistaken as the pages written in Japanese, which is not the case. Similar analysis for other countries was given by Baeza-Yates et al [5].

In Section 2, we describe how the Web data have been crawled and processed. We describe about the data files of overall web data collection in Section 3. Some statistics about NW1000G-04 are shown in Section 4.

* E-mail: masao@nii.ac.jp

** She was working for NTCIR-5 Workshop at National Institute of Informatics until September 2005. She is currently a fellow at The Energy and Resources Institute, New Delhi.

2 Crawling

In Navi-2, we planned to make a terabyte-scale data collection from the public web pages related to Japan in some ways, e.g. provided by Japanese organizations/people, providing information of Japanese matters. So we have crawled web pages from .jp domain mainly and other domains complementarily from January 2004 through January 2005 with the following steps:

1. About 450,000 start-up sites in "*.jp" domain were gathered from previous crawls and other measures.
2. The sites gathered at step (1) were crawled starting with the top page up to 15 hyperlink hops.
3. URLs in "*.jp" domain found at step (2) and recognized as new sites not included in the above mentioned start-up sites were collected. Then the sites were crawled starting with the top page up to 8 hyperlink hops.
4. URLs out of "*.jp" domain found at steps (2) and (3) and recognized as new sites were collected. Then the sites were crawled starting with the top page up to 10 pages.
5. Language identification was performed on the pages fetched at step (4). Then those sites that included at least one page judged as Japanese were selected.
6. The sites selected at step (5) were crawled starting with the top page up to 8 hyperlink hops.
7. Web pages that were judged to be written in other languages than Japanese or English by a language identifier were removed from the pages crawled at steps (2), (3), and (6) to make the data set.

At the step (7), we used a language identifier produced by Basis Technology Corp. However, a considerable number of pages in other languages than Japanese or English are still included in the data set.

Our crawling target was only text pages, mainly *text/html* and *text/plain*, accessible with HTTP protocol at default port number 80. So we excluded image and binary files based on their file extensions, e.g. *.jpg*, *.gif*, *.png*, *.pdf*, *.doc*, *.xls*, *.swf*, etc. After and during the crawling process, we excluded URLs having a question mark "?", e.g. `http://example.jp/index.cgi?foo=1&bar=2`, for excluding dynamically generated pages. Note that this does not exclude all the dynamic pages, because some dynamic pages have parameters in their paths but not in query-parts of their URLs, e.g. *"Main page" on the Japanese Wikipedia site*[5] have no query-part, but the server semantically gets a parameter by using a path-style parameter and returns contents generated dynamically. Thus the resulting dataset includes many dynamically generated pages.

The crawling engine unified duplicate directory entry pages. If the URL of a page ends with a filename either of "index.html", "index.htm", or "default.htm" and its content is identical to the page whose URL ends with "/" (without a filename), the crawler saved only the latter page, and the former page was discarded from the collection [6]. The "linklist" distributed with NW1000G-04 is built considering this unification; however care should be taken when the users process the links by themselves. We will report the statistics of link information based on this "linklist" in Section 4.

[5] `http://ja.wikipedia.org/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8`

[6] This causes troubles on the link information ("linklist") especially for the sites that explicitly use filenames, such as "index.html" etc., as link targets. The "linklist" used at the Navi-2 experiments lacked a lot of link information for such sites. After the end of the NTCIR-5 workshop, we rebuilt the "linklist" in order to correct the side-effects of unifying the directory entry pages, and we will deliver this corrected version for the research-purpose distribution of the collection. Because the Navi-2 experiments were going with this problem, the users should be careful when comparing performance of systems that use the linklist.

3 Data Files

The document dataset NW1000G-04 consists of about 100 million web pages, approximately 1.36TB (1.5×10^{12} bytes) in total. We provide four variants of text data of web pages as follows in order to help its users making index for them.

- Raw data: web page data just as were crawled from the web; the total size is 1.36TB.
- Euc data: web page data processed from "raw data" converting Japanese character encodings to EUC-JP by NKF (Network Kanji Filter) [6].
- Cooked data: text data processed from "euc data" removing unnecessary HTML tags, comments, scripts etc., and normalizing characters and numeral entity reference code to EUC-JP correspondents.
- Segmented data: segmented word data generated from "cooked data" using a Japanese morphological analyzer Mecab [7].

Four kinds of lists were also provided (see also Figure 1):

- Sitelist: a list of crawled web sites consisting of site identifiers and site names.
- Doclist: a list of documents in the data set consisting of document identifiers and documents' URLs.
- Linklist: a list of link data consisting of document identifiers and documents' URLs of link source documents and link target documents respectively, both contained in the "doclist", and link types (i.e. `a,href`; `area,href`; `frame,src`; `meta,refresh`).
- Anchorlist [7]: a list of anchor text data consisting of document identifiers of link source documents and link target documents respectively, both contained in the "doclist", and anchor texts.

The site and document identifiers are given in dictionary order, and the lines in each list are sorted by the same order. Both linklist and anchorlist are available in two orders sorted by link source document identifiers (outlink) and link target document identifiers (inlink) respectively.

4 Statistics

The document collection was crawled from 389,875 sites, and contains 95,870,352 pages. The total size of all the pages in the collection is approximately 1.36TB (1,504,753,729,713 bytes). Table 1 compares NW1000G-04 and existing Web collections, namely NW100G-01 [2], TREC GOV [8] and GOV2 [9], SPIRIT [10], and EuroGOV [11].

Table 1. Comparing Web data collections,

Collection	Number of pages	Size (GB)	Year of crawling
EuroGOV	3,589,501	11	2004
TREC GOV	1,247,753	18	early 2002
NW100G-01	11,038,720	100	2001–2002
TREC GOV2	25,205,179	426	early 2004
SPIRIT	94,552,870	1000	mid 2001
NW1000G-04	95,870,352	1401	2004–Jan 2005

The collection contains the pages from 161 top level domains (TLDs). The distribution of the 20 most frequent TLDs in NW1000G-04 is given in Table 2. Each column indicates the domain

[7] Anchorlist was not provided at the NTCIR-5 Workshop.

Sitelist

```
2110014 http://www.visit-germany.jp
2110032 http://www.visit-oahu.jp
2110048 http://www.visit.busan.kr
```

Doclist

```
2110014_0000001 http://www.visit-germany.jp/
2110014_0000002 http://www.visit-germany.jp/Default.asp
2110014_0000003 http://www.visit-germany.jp/Stylesheet/Stylesheet32.css
2110032_0000001 http://www.visit-oahu.jp/
2110032_0000002 http://www.visit-oahu.jp/about.html
2110032_0000003 http://www.visit-oahu.jp/about_oahu.html
2110032_0000004 http://www.visit-oahu.jp/accom_bb.html
```

Linklist (outlink)

```
2110014_0000001 http://www.visit-germany.jp/    1146533_0003696 http://www.goethe.de/os/tok/jpindex.htm a,href
2110014_0000001 http://www.visit-germany.jp/    1451557_0000001 http://www.lufthansa.co.jp/     a,href
2110014_0000001 http://www.visit-germany.jp/    1760718_0000002 http://www.raileurope.jp/contest/       a,href
2110014_0000001 http://www.visit-germany.jp/    2110014_0000002 http://www.visit-germany.jp/Default.asp a,href
2110014_0000002 http://www.visit-germany.jp/Default.asp 1146533_0003696 http://www.goethe.de/os/tok/jpindex.htm a,href
2110014_0000002 http://www.visit-germany.jp/Default.asp 1451557_0000001 http://www.lufthansa.co.jp/     a,href
2110014_0000002 http://www.visit-germany.jp/Default.asp 1760718_0000002 http://www.raileurope.jp/contest/       a,href
2110014_0000002 http://www.visit-germany.jp/Default.asp 2110014_0000002 http://www.visit-germany.jp/Default.asp a,href
2110032_0000001 http://www.visit-oahu.jp/       1146896_0000001 http://www.gohawaii.jp/ a,href
2110032_0000001 http://www.visit-oahu.jp/       2110032_0000002 http://www.visit-oahu.jp/about.html     a,href
2110032_0000001 http://www.visit-oahu.jp/       2110032_0000008 http://www.visit-oahu.jp/accommodation.html     a,href
2110032_0000001 http://www.visit-oahu.jp/       2110032_0000009 http://www.visit-oahu.jp/activity.html  a,href
```

Anchorlist (outlink)

```
2110014_0000001 1760718_0000002 こちら
2110014_0000002 1760718_0000002 こちら
2110032_0000001 1146896_0000001 ハワイ州観光局
2110032_0000001 2110032_0000002 オアフ島について
2110032_0000001 2110032_0000008 宿泊施設
2110032_0000001 2110032_0000009 アクティビティ
2110032_0000001 2110032_0000018 オアフ観光局メディア・業界サイト
2110032_0000001 2110032_0000055 9/8 NEW !
2110032_0000001 2110032_0000076 バリアフリーやさしいオアフ
2110032_0000001 2110032_0000088 オアフ年間イベント
```

Fig. 1. Excerpts from sitelist, doclist, and linklist

names, the number of pages, the number of sites, and the total page size of the domain, respectively. As shown in Table 2, ".jp" domain is the highest population and followed by ".com" domain in all columns. About 50% of the documents in NW1000G-04 are from ".jp" domain, and 27% are from ".com" domain. These two TLDs occupy over 75% of all the pages in the collection. Table 3 shows the distribution of the second level ".jp" domain in NW1000G-04. In the ".jp" domain, ".co.jp" domain is the most dominant second level domain; it occupies approximately 40% of the documents from ".jp" domain.

Figure 2 shows the distribution of the size of raw pages and cooked pages. The average of raw page size in NW1000G-04 is 15,696 bytes, and the median is 6,922 bytes. Removing HTML tags, scripts, comments, etc. ("cooked" data) reduces these values to the average of 4,216 bytes and the median of 1,479 bytes. This means HTML tags and the others occupy around 75–80% of the pages. As can be seen, these sizes follow the power-law distribution. Figure 3 shows the distribution of the number of pages per site. The average number of pages per site is 245.9, and the median is 25.

Figures 4 and 5 present the distribution of in-degree links within NW1000G-04 and that of out-degree links, respectively. We regarded the following elements in a page as a link: anchor-tags (<a href=...), client-side image maps (<area href=...), frames (<frame src= and <iframe

Table 2. Distribution of top level domains in NW1000G-04

Domain	Number of pages	(%)	Number of sites	(%)	Total size (bytes)	(%)
.jp	48,246,587	50.3	261,159	67.0	622,644,571,083	41.4
.com	25,598,807	26.7	87,514	22.4	529,480,787,801	35.2
.net	5,740,568	6.0	18,204	4.7	81,823,078,477	5.4
.org	5,054,396	5.3	7,684	2.0	77,936,967,823	5.2
.edu	2,621,211	2.7	652	0.2	41,024,077,793	2.7
.uk	1,145,478	1.2	881	0.2	35,330,193,582	2.3
.to	643,416	0.7	2,524	0.6	7,272,218,763	0.5
.au	489,524	0.5	333	0.1	7,185,766,286	0.5
.de	452,288	0.5	1,146	0.3	7,574,778,682	0.5
.gov	405,562	0.4	101	0.0	8,645,321,942	0.6
.nl	392,240	0.4	364	0.1	7,415,860,673	0.5
.info	338,834	0.4	1,304	0.3	4,814,742,180	0.3
.ca	308,671	0.3	226	0.1	4,918,137,087	0.3
.tv	264,421	0.3	1,218	0.3	3,897,464,659	0.3
.se	247,600	0.3	153	0.0	2,748,915,514	0.2
.cz	231,225	0.2	186	0.0	2,847,299,435	0.2
.us	224,333	0.2	103	0.0	3,520,449,601	0.2
.pl	177,279	0.2	294	0.1	3,219,510,666	0.2
.fi	166,570	0.2	78	0.0	1,928,863,494	0.1
.ch	165,011	0.2	211	0.1	3,580,024,464	0.2

Table 3. Distribution of the second level ".jp" domains in NW1000G-04

Domain	Number of pages	(%)	Number of sites	(%)	Total size (bytes)	(%)
co.jp	19,132,953	39.7	140,622	53.8	257,939,542,629	41.4
ac.jp	5,849,830	12.1	16,748	6.4	63,903,217,845	10.3
ne.jp	4,929,226	10.2	20,038	7.7	68,836,127,165	11.1
or.jp	2,837,991	5.9	14,534	5.6	25,552,150,399	4.1
go.jp	1,825,843	3.8	1,957	0.7	19,143,831,144	3.1
gr.jp	926,432	1.9	5,243	2.0	12,628,373,651	2.0
ed.jp	628,059	1.3	5,595	2.1	4,220,986,073	0.7
thebbs.jp	388,504	0.8	83	0.0	12,501,049,060	2.0
ad.jp	375,811	0.8	318	0.1	4,304,314,210	0.7
tokyo.jp	285,182	0.6	554	0.2	2,416,340,883	0.4
yamaguchi.jp	168,550	0.3	60	0.0	669,561,409	0.1
cpan.jp	152,525	0.3	3	0.0	938,836,297	0.2
saitama.jp	143,608	0.3	163	0.1	1,249,200,264	0.2
hokkaido.jp	141,550	0.3	243	0.1	1,331,304,266	0.2
lg.jp	140,647	0.3	45	0.0	1,303,811,708	0.2
chiba.jp	137,057	0.3	155	0.1	1,045,311,722	0.2

src=...), and redirections (<meta http-equiv="refresh" content="*n*;URL=...). The total number of the links among pages within NW1000G-04 is 1,805,160,288. The average number of links per page is 18.83, the median of out-degree links per page is 5, and the median of in-degree is 2. As reported by many prior works [12] [13], our Web graph basically follows the power law too. Obviously, the plot in Figure 5 (out-degree plot) has anomalies at pages which have around 300–500 out-degrees. This might be derived from some "spam-sites" which consists of dynamically generated pages and very strongly connected links with each other.

Tables 4 and 5 show the distribution of URL length and URL depth, respectively. The average URL length in NW1000G-04 is 58.89 bytes. URL depth means the depth of the path component in the URL. For example, both http://www.example.com/ and http://www.example.com/

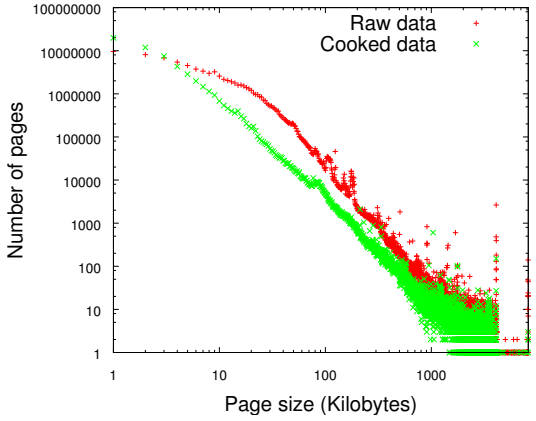

Fig. 2. Distribution of page size in NW1000G-04

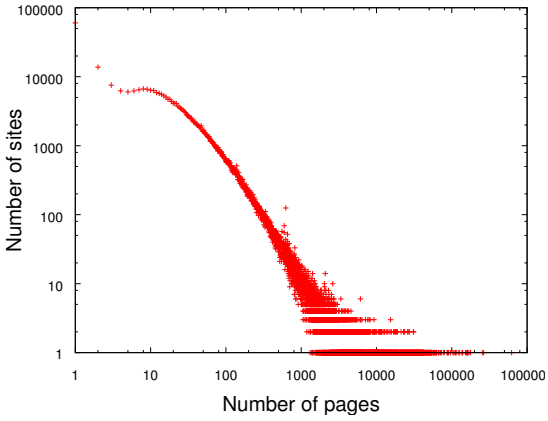

Fig. 3. Distribution of site size in NW1000G-04

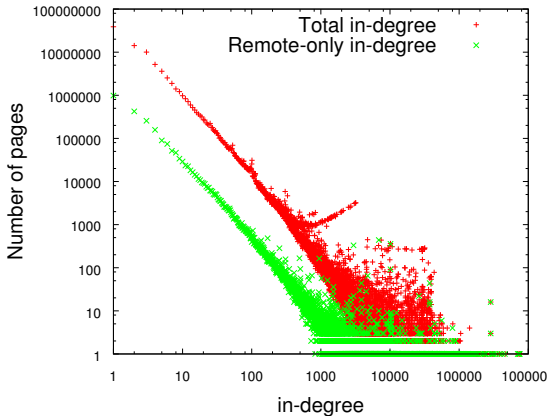

Fig. 4. Distribution of in-degree links within NW1000G-04

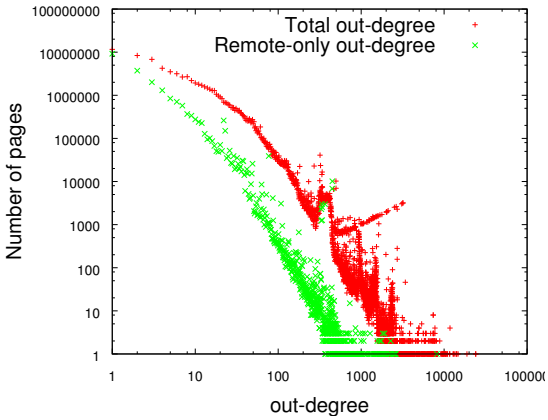

Fig. 5. Distribution of out-degree links within NW1000G-04

`index.htm` have URL depth of 1, and `http://www.exmaple.com/foo/index.htm` has URL depth of 2.

Table 6 shows the distribution of the extension of URL paths in NW1000G-04. Note that we excluded non-textual files based on extensions of URL. The most frequent extensions are ".html" (50%) and ".htm" (20%). In the table, "(*dir*)" (12%) represents the directory-index type pages, whose URLs end with "/", and "(*empty*)" (9%) represents the pages without extensions. Typical extensions for dynamically generated pages, ".shtml", ".php", ".asp", ".cgi", ".cfm", ".php3", ".aspx", ".jsp", and ".jhtml", are also found.

Table 7 shows the distribution of the number of IMG tags in a page, and that of IMG tags with ALT attributes. About 23% of pages in NW1000G-04 have no IMG tags in content, and 45% has no IMG tags with ALT attribute. The number of pages embeded with the Flash animation (application/x-shockwave-flash), that is not shown in the Tables, is 1,306,286 (1.36%).

5 Summary

In this report, we have described a terabyte-scale web data collection, NW1000G-04, in details about the process of crawling, data files and some statistics. The authors believe that this dataset will be a useful testbed for various Web researches, including information access technologies, social and academic activity analyses, etc.

Table 4. Distribution of URL length

Length	Number of pages	(%)
< 20	18,798	0.02
30	868,954	0.91
40	9,013,137	9.40
50	23,377,352	24.38
60	24,464,918	25.52
70	16,899,693	17.63
80	9,555,059	9.97
90	4,958,595	5.17
100	2,860,302	2.98
110	1,697,733	1.77
120	924,251	0.96
130	454,071	0.47
140	406,779	0.42
150	312,552	0.33
150 <	58,158	0.06

Table 5. Distribution of URL depth

Depth	Number of pages	(%)
1	6,295,831	6.57
2	19,313,480	20.15
3	22,897,812	23.88
4	19,124,779	19.95
5	13,879,387	14.48
6	6,992,827	7.29
7	3,504,585	3.66
8	1,684,911	1.76
9	969,553	1.01
10	449,228	0.47
11	188,972	0.20
12	139,308	0.15
13	74,773	0.08
14	36,427	0.04
15	54,405	0.06
15 <	264,074	0.28

Table 6. Distribution of extensions of URL path in NW1000G-04

Extension	Number of pages	(%)
.html	48,075,647	50.15
.htm	19,095,667	19.92
(*dir*)	11,527,183	12.02
(*empty*)	8,744,140	9.12
.shtml	1,174,767	1.23
.txt	1,034,239	1.08
.php	967,179	1.01
.css	607,115	0.63
.asp	547,961	0.57
.readme	521,329	0.54
.cgi	369,266	0.39
.xml	205,816	0.21
.cfm	170,753	0.18
.php3	128,383	0.13
.aspx	126,091	0.13
.jsp	122,724	0.13
.jhtml	115,551	0.12
.changes	86,374	0.09
.meta	75,662	0.08
.sign	56,855	0.06

Table 7. Distribution of the number of IMG tags within a page, and the number of IMG tags with ALT attribute

Number of IMGs	Number of pages	(%)	Number of pages (with ALT)	(%)
0	22,303,548	23.26	43,879,914	45.77
1	10,667,357	11.13	8,863,642	9.25
2	6,737,106	7.03	6,755,698	7.05
3	4,798,843	5.01	3,697,171	3.86
4	4,127,973	4.31	2,992,805	3.12
5	3,256,211	3.40	2,319,427	2.42
6–10	10,815,019	11.28	7,664,506	7.99
11–20	10,457,575	10.91	8,299,891	8.66
21–30	6,428,451	6.71	3,834,821	4.00
31–40	3,954,144	4.12	2,108,908	2.20
41–50	2,925,798	3.05	1,468,851	1.53
51–100	6,382,102	6.66	2,816,313	2.94
101–	9,181,282	9.58	3,898,704	4.07

References

1. K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, H. Yamana, Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2), in: N. Kando, M. Takaku (Eds.), Proceedings of NTCIR-5 Workshop Meeting, 2005, pp. 423–442.
   URL http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/WEB/NTCIR5-OV-WEB-OyamaK.pdf
2. K. Eguchi, K. Oyama, E. Ishida, N. Kando, K. Kuriyama, Overview of the web retrieval task at the third NTCIR workshop, Tech. Rep. NII-2003-002E, National Institute of Informatics (2003).
   URL http://research.nii.ac.jp/TechReports/03-002E.html
3. C. Saeki, H. Shimada, S. Tahata, Report on statistical survey of world wide web content: Web content amount as an indicator of internet development in Japan, Tech. rep., Institute for Information and Communications Policy, Japan (2004).
   URL http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2004/2004-1-02-3.pdf

4. Survey on comprehensive collection, storage, and archiving of Japanese web sites, Tech. rep., The National Diet Library of Japan (2006).
URL http://www.ndl.go.jp/en/aboutus/bulkresearch2005index_e.html
5. R. Baeza-Yates, C. Castillo, E. N. Efthimiadis, Characterization of national web domains, ACM Transactions on Internet Technology(To appear).
6. NKF (Network Kanji Filter).
URL http://nkf.sourceforge.jp/
7. T. Kudo, MeCab: Yet another part-of-speech and morphological analyzer.
URL http://mecab.sourceforge.jp/
8. The .gov test collection.
URL http://ir.dcs.gla.ac.uk/test_collections/govinfo.html
9. C. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2004 terabyte track, in: Proceedings of TREC 2004, 2004.
URL http://trec.nist.gov/pubs/trec13/papers/TERA.OVERVIEW.pdf
10. H. Joho, M. Sanderson, The SPRIT collection: an overview of a large web collection, ACM SIGIR Forum 38 (2) (2004) 57–61.
11. B. Sigurbjönsson, EuroGOV: Engineering a multilingual web corpus, in: Working Notes for the CLEF 2005 Workshop, 2005.
12. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, in: Proceedings of the 9th international World Wide Web conference, North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, 2000, pp. 309–320.
13. R. Albert, H. Jeong, A.-L. Barabasi, Diameter of the world-wide web, Nature 401 (1999) 130–131.