The SPIRIT collection: an overview of a large web collection

Hideo Joho and Mark Sanderson

Department of Information Studies
University of Sheffield

H.Joho|M.Sanderson@sheffield.ac.uk

Abstract

A large scale collection of web pages has been essential for research in information retrieval and related areas. This paper provides an overview of a large web collection used in the SPIRIT project for the design and testing of spatially-aware retrieval systems. Several statistics are derived and presented to show the characteristics of the collection.

Introduction

SPIRIT[i] - Spatially-Aware Information Retrieval on the Internet - is a research project funded through the EC Fifth Framework Programme. The main aim of the project is to design and implement a search engine to find documents and datasets on the Web relating to places or regions referred to in a query. (Jones, et al., 2002). One of the resources the project is exploiting is a large scale static web collection. While the Internet is dynamic in nature, a static collection enables researchers to analyse retrieval systems with consistent data.

The web collection used in the SPIRIT project was originally crawled by Dr. Clarke and his colleagues at the University of Waterloo in Mid-2001 (Clarke, et al., 2002), and distributed to Prof. Fox at Virginia Tech. We obtained a version of this crawl from Prof. Fox which constitutes the data in our collection. It is our understanding that the crawl was bootstrapped by a set of educational web sites.

At Sheffield, we processed the data set so that every web page is tagged in a standard format based on the Web Track of TREC (Craswell, et al., 2003). Our version of the collection has been used not only by the SPIRIT project but also by the researchers at Dublin City University, the University of Glasgow, and the University of Brighton (e.g., Gurrin and Smeaton, 2004, Cacheda, et al., 2004). Our version is sometimes referred to the SPIRIT web collection.

This paper will show a sample document along with the tagging information, followed by a set of statistics derived from the collection.

Sample document

A sample document of our collection can be found in Figure 1. Every document starts with a <DOC> tag and ends with a </DOC> tag. <DOCNO> is a unique document ID, <DOCURL>[ii] is the original URL of the document, and <DOCHDR> contains the URL and the length of the original data in bytes followed by the HTTP response of the crawled web server. The rest of the data is the document contents.

```
<DOC>
<DOCNO>SPRT-022-021-086-0042537</DOCNO>
<DOCURL>http://www.shef.ac.uk/~is/</DOCURL>
<DOCHDR>
http://www.shef.ac.uk/~is/       3875
HTTP/1.1 200 OK
Date: Wed, 23 May 2001 22:03:00 GMT
Server: [removed]
Cache-Control: max-age=7776000, public, must-revalidate
Connection: close
Content-Type: text/html
</DOCHDR>

<!DOCTYPE HTML PUBLIC "-//SoftQuad//DTD HoTMetaL PRO 5.0::19980907::extensions to HTML 4.0//EN"
"hmpro5.dtd">

<HTML>

<HEAD>
<TITLE>Sheffield - Department of Information Studies</TITLE>
</HEAD>

<BODY BACKGROUND="gfx/g_back01.jpg" VLINK="#CC0000" LINK="#000099">

[omitted]

</BODY>
</HTML>
</DOC>
```

Figure 1. Sample document (DOCNO: SPRT-022-021-086-0042537)

Collection statistics[iii]

The total number of documents in the SPIRIT collection is 94,552,870 on an approximate size of 1 TB. A comparison with existing document collections is given in Table 1. The period in the first three collections is the date of the documents published, while the period for the last four collections is the date of the crawling.

Collection	Period	No. of docs	Size (MB)
Cranfield	1922-1963	1400	1.6
CACM	1958-1979	3204	2.2
TREC Disk 1-5	1987-1994	1,634,234	5,321
TREC .GOV	Early 2002	1,247,753	18,000
NTCIR-3 Web	2001	11,038,720	100,000
TREC .GOV2	Early 2004	25,205,179	426,000
SPIRIT	Mid 2001	94,552,870	1,000,000

Table 1. Comparison of the collection size

The distribution of document size is given in Table 2. The size in bytes is of the original web pages, while the number of words is based on the actual text displayed by a browser (`lynx` was used here). Words were counted as character strings separated by whitespace. In our collection, the size of a displayed text tends to be approximately a third of the original HTML page.

Size (byte)	%	Size (byte)	%	No of words	%	No of words	%
10	2.7309	10,000	19.3494	10<	0.0000	5,000	3.4789
20	0.0038	20,000	18.9109	10	6.6160	10,000	0.9379
50	0.0189	50,000	14.5617	20	8.4310	20,000	0.3178
100	0.1532	100,000	2.2913	50	9.1817	50,000	0.1221
200	1.4795	200,000	0.4989	100	9.3134	100,000	0.0254
500	9.6051	500,000	0.1751	200	16.0338	200,000	0.0104
1,000	4.1247	1,000,000	0.0389	500	25.1390	500,000	0.0030
2,000	7.3258	2,000,000	0.0166	1,000	13.7170	1,000,000	0.0003
5,000	18.7087	5,000,000	0.0065	2,000	6.6723	>1,000,000	0.0000

Table 2. Size of documents in the SPIRIT collection

The distribution of the 20 most frequent top-level and second-level domains is shown in Table 3, along with the data provided by Internet Software Consortium (ISC) which are based on 169 million pages as of July 2001 (a similar period to our collection). Although the rate of portions varies, the top-level consists of a reasonably similar set of domains in both cases, while the second-levels are more scattered.

SPIRIT Mid-01		ISC Jul 01		SPIRIT Mid-01		ISC Jul 01	
Top Domain	%	Top Domain	%	Second Domain	%	Second Domain	%
com	42.29	com	36.12	co.uk	1.44	lucent.com	3.20
edu	17.05	net	24.76	ac.uk	0.95	aol.com	2.65
org	10.17	edu	4.38	co.jp	0.70	uu.net	2.61
net	4.82	jp	3.54	edu.au	0.47	ne.jp	1.97
us	3.05	ca	1.87	rootsweb.com	0.45	home.com	1.86
uk	2.75	uk	1.68	ne.jp	0.44	rr.com	1.36
gov	2.42	de	1.46	com.au	0.43	pacbell.net	0.84
de	2.35	us	1.32	amazon.com	0.42	t-dialin.net	0.81
jp	2.15	it	1.21	yahoo.com	0.40	avaya.com	0.80
ca	2.00	mil	1.21	lycos.com	0.38	qwest.net	0.73
au	1.30	au	1.13	ac.jp	0.37	co.uk	0.68
fr	0.78	nl	1.07	or.jp	0.36	genuity2.net	0.64
it	0.60	org	0.86	sun.com	0.32	dialsprint.net	0.60
nl	0.47	fr	0.84	geocities.com	0.32	hinet.net	0.60
ru	0.46	tw	0.77	tripod.com	0.32	ac.uk	0.58
se	0.43	br	0.63	nasa.gov	0.27	att.net	0.53
ch	0.42	se	0.62	amazon.de	0.27	splitrock.net	0.51
es	0.33	es	0.56	ca.us	0.25	net.tw	0.48
mil	0.32	gov	0.52	aol.com	0.24	ad.jp	0.45
cn	0.32	fi	0.52	uiuc.edu	0.22	com.br	0.45

Table 3. Top and second level domains, compared with the data from ISC in Jul 2001 (Source: Internet Software Consortium (http://www.isc.org/))

European domains are of most interest for the SPIRIT project. A total of 22 European domains were found in the 50 most frequent domains. The European domains consist of over 9.6 million web pages which is approximately 10.24% of the entire collection. The distribution of European domains is shown in Table 4. As can be seen, nearly half of European pages are of uk or de. However, there are between 34,000 and 730,000 pages for the rest of 20 European domains.

Country	Domain	%	Country	Domain	%
United Kingdom	uk	26.878	Denmark	dk	2.187
Germany	de	22.980	Poland	pl	1.838
France	fr	7.635	Belgium	be	1.711
Italy	it	5.819	Czech Republic	cz	1.398
Russian Federation	ru	4.487	Ireland	ie	1.292
Sweden	se	4.238	Portugal	pt	0.829
Switzerland	ch	4.060	Hungary	hu	0.828
Spain	es	3.176	Greece	gr	0.778
Norway	no	3.078	Turkey	tr	0.604
Finland	fi	2.917	Estonia	ee	0.525
Austria	at	2.388	Slovak Republic	sk	0.353

Table 4. European domains (N=9,682,678)

Table 5 shows the 20 most frequent character sets defined in a meta tag of web page. As can be seen, most pages do not define the character set. Of those defined, iso-8859-1 and windows-1252 are the dominant ones.

Charset	%	Charset	%
(Empty)	71.876	windows-1250	0.137
iso-8859-1	18.558	iso8859-1	0.125
windows-1252	6.221	euc-kr	0.108
gb2312	0.609	iso-2022-jp	0.077
shift_jis	0.506	euc-jp	0.059
x-sjis	0.459	windows-1256	0.045
windows-1251	0.194	us-ascii	0.037
big5	0.182	windows-1254	0.033
utf-8	0.172	windows-874	0.032
iso-8859-2	0.158	x-euc-jp	0.031

Table 5. Charsets defined in a meta tag

Table 6 shows the distribution of URL length and depth. URL length is the number of characters without URIs (e.g. http://). For example, the length of `x.com` is 5 and `www.google.com` is 14. URL depth is the number of hierarchies where the web page is stored. For example, `www.shef.ac.uk`, `www.shef.ac.uk/`, and `www.shef.ac.uk/index.html` are counted as 1, while `www.shef.ac.uk/~is/index.html` is counted as 2.

The table shows that approximately half of the URLs fall into the length between 31 and 50 characters. As for the depth, the majority of URLs consist of up to four levels of the hierarchies with the second as the most frequent level.

Length	%	Length	%	Depth	%
5<	0.0000	55	8.9599	1	18.3809
5	0.0002	60	6.4870	2	28.3299
10	0.0732	65	4.4197	3	24.2538
15	0.9477	70	2.8489	4	14.8797
20	2.1855	75	1.9178	5	7.5039
25	4.2286	80	1.3666	6	3.5230
30	8.3106	85	1.0130	7	1.6065
35	12.6004	90	0.6561	8	0.8954
40	15.1254	95	0.4518	9	0.2351
45	14.5864	100	0.3209	10	0.1231
50	12.0536	>100	1.4463	>10	0.2086

Table 6. URL length and depth

Table 7 shows the number of `<IMG>` tags defined in web pages. Nearly two in three pages contains at least one image tag, and one in four pages contains more than 10 image tags. Of those pages with the image tags, approximately 70% define at least one `ALT` attribute for the alternative text information.

Number	IMG (%)	ALT (%)
0	33.28	30.57
5	29.74	40.29
10	10.81	11.49
20	11.09	9.46
50	10.42	6.40
100	3.70	1.45
>100	0.96	0.33

Table 7. Number of images and ALT attribute

Linkage analysis

Linkage information is an important factor when examining web document collections. The density of outgoing and incoming links from on-sites and off-sites is used to analyse the structure of web collections. Gurrin and Smeaton (2004) compared the linkage density of several web

collections including the SPIRIT collection. Their analysis showed that the density and distribution of outgoing links of SPIRIT documents are very close to an estimate of the level on the Web (i.e. 4.9 off-site links, 14.2 on-site). The density of incoming links from off-sites is 1.24 which is lower than .GOV collection (1.98). The ideal level is estimated to be 4.9. However, their experiment suggests that a subset collection of 1.5-2 million pages with the ideal linkage density can be generated from the SPIRIT collection.

Summary

In this paper we have presented the SPIRIT web collection and a number of statistics derived from our initial analysis. The collection appears to be a useful resource for those who require geographically more heterogeneous data than existing web collections. The sampling of the real web is not a trivial task. A realistic approach to understand the structure of the Web would be to gather information from a number of different samples. The authors hope that the data provided in this report will contribute to such an approach.

Acknowledgments

The authors wish to thank Dr. Charlie Clarke for the conducting original crawl, and Prof. Edward Fox for giving us an access to his copy. The authors also wish to thank members of the SPIRIT project for the ideas of “what to count”. Financial support for the work was provided by the EC 5th Framework RTD project SPIRIT: contract number IST-2001-35047.

References

Cacheda, F., Plachouras, V. & Ounis:, I. (2004). "Performance Analysis of Distributed Architectures to Index One Terabyte of Text". In: McDonald, S. & Tait, J. (eds.), *Advances in Information Retrieval, Proceedings of the 26th European Conference on IR Research*, Lecture Notes in Computer Science, Vol. 2997, Sunderland, UK. pp. 394-408. Springer.

Clarke, C. L. A., Cormack, G. V., Laszlo, M., Lynam, T. R., and Terra, E. L. (2002) “The impact of Corpus Size on Question Answering Performance”. In: Beaulieu, M., Baeza-Yates, R. & Myaeng, S.H. (eds.), *Proceedings of the 25th Annual ACM Conference on Research and Development in Information Retrieval*, 369-370, Tampere, Finland: ACM Press.

Craswell, N., Hawking, D., Wilkinson, R. & Wu, M. (2003). "Overview of the TREC 2003 Web Track". In: Voorheer, E. (ed.), *NIST Special Publication 500-255:The Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD. pp. 78-92. NIST.

Gurrin, C. & Smeaton, A. (2004). "Replicating Web Structure in Small-Scale Test Collections". *Journal of Information Retrieval*, 7 (3-4), 239-263.

Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Kreveld, M.v. & Weibel, R. (2002). "Spatial information retrieval and geographical ontologies an overview of the SPIRIT project". In: Beaulieu, M., Baeza-Yates, R. & Myaeng, S.H. (eds.), *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland. pp. 387-388. ACM Press.

---

[i] The project web site: http://www.geo-spirit.org/ [Last accessed: 27/09/2004].

[ii] This tag is not used in the Web Track.

[iii] Due to the diversity of document structures, the percentages are not always based on the total number of documents in the collection. The margin of error in the figures is estimated to be approximately 0.01%.