The ACademic DireCtory - AC/DC

Technical Report 8-96
Computing Laboratory, University of Kent at Canterbury

Dave Beckett, *D.J.Beckett@ukc.ac.uk*
University of Kent at Canterbury[*]
<URL:http://www.hensa.ac.uk/parallel/www/djb1.html>

Neil G. Smith, *N.G.Smith@unix.hensa.ac.uk*
HENSA Unix[†] University of Kent at Canterbury[*]
<URL:http://www.hensa.ac.uk/team/N.G.Smith.html>

Abstract

AC/DC[1] is an experimental, collaborative research project that indexes and allows searches over all public academic WWW servers in the UK. This report describes why AC/DC was created, how it is built from existing software, the collaborative process used to collect and index the data and future activities.

HENSA Unix[2] and part of the authors research work is funded by JISC[3], the Joint Information Systems Committee of the Higher Education Funding Councils.

1 WWW Crawlers - Why A New One?

All the major WWW crawling programs such as Alta Vista[4] (Digital), InfoSeek[5], Lycos[6], Webcrawler[7], Excite[8] etc. are based in the USA and collect their pages across the transatlantic link. There are two problems with the USA based services:

1. They index the whole world and can return resources that are not very relevant for the UK. A UK based system will index only UK sites, and should return local answers to the queries which may be more relevant to UK academics on JANET sites.

2. The UK-USA connection is very busy and will remain so and the use of the bandwidth by such services isn't likely to help. A UK based system will be faster and can be more up to date.

---

[*] See URL http://www.ukc.ac.uk/
[†] See URL http://www.hensa.ac.uk/unix.html
[1] See URL http://www.hensa.ac.uk/search/acdc.html
[2] See URL http://www.hensa.ac.uk/unix.html
[3] See URL http://www.niss.ac.uk/education/jasper/intro.html
[4] See URL http://www.altavista.digital.com/
[5] See URL http://www.infoseek.com/
[6] See URL http://www.lycos.com/
[7] See URL http://www.webcrawler.com/
[8] See URL http://www.excite.com/

We decided that an experimental WWW crawling project, initially covering UK academic sites, would be a good starting point to test the viability of such a system. It should be collaborative and distributed to spread the work rather than depending on a single system.

2 Getting A WWW Crawler

The commercial services mentioned above have in the most part evolved from US University research projects and the software they use to fetch, index and search WWW pages is commercial and therefore not available to a research project such as this.

The alternatives to using an existing commercial server are to either write your own (a considerable project in its own right) or to find an alternative free software version. We chose to make use of the existing *Harvest*[9] system developed at the University of Colorado.

3 Harvest

The *Harvest* system runs on standard UNIX platforms and has two major components:

Gatherers These fetch or *gather* the data from WWW, ftp and gopher sites as well as some USENET newsgroups. The gathered data is then made available to the world.

Brokers Brokers are the indexing part of the system and they collect data from the gatherers, index them and provide a query interface via a standard WWW query form.

Harvest is not designed to operate as a large multi-site webcrawler. It is meant for individual sites to gather and serve their indices *locally* via the brokers. The collected data is then meant to be shared in a hierarchical fashion. In this sense, it was appropriate for the collaborative project that we envisaged – the gatherering of all the WWW sites could be split among different institutions and the results collected at our site for the top level index.

4 Running A WWW crawler

The WWW crawler makes use of earlier research on fetching site specific details for all the sites under `.uk` - the top level domain for the UK. This reads the DNS regularly and updates a database of domains. These domains are then used to look for a WWW host in the domain using the common host name suffixes such as `www.`*sitename*`.ac.uk`. If one is found, the top level WWW page of that host is retrieved and the `TITLE` tag used to get a short description of the organisation. This work is the *UK Internet Sites*[10] project and the URLs from it were used to provide a list of active `.ac.uk` WWW sites for the WWW crawler.

As described above, the Harvest gatherer is not intended to be a full multi-site WWW crawler and it needed a little controlling software to bend it to this purpose. Since we would be crawling

[9] See URL http://harvest.cs.colorado.edu/
[10] See URL http://www.hensa.ac.uk/uksites/

several hundred WWW sites, it was essential to make this work efficiently - not index every site every time. This software invokes the Harvest gatherer regularly with a few top level site URLs to merge them into the database.

The initial WWW crawl of the UK consisted of fetches of up to 50 URLs from the root of each of around 300 sites. The crawling was further restricted in that it was prevented from going more than 2 levels "down" (the depth) from the root and was restricted to the single WWW site contacted. Around 15,000 WWW pages were indexed in the first run.

Experience from this showed that the gathering needed careful configuration changes to ignore certain WWW pages that were of no use - images, executable binaries, archive files, CGI scripts, proprietary document formats etc.

Once the configuration was more satisfactory a deeper and more extensive WWW crawl was started, taking up to 400 pages from each WWW site, up to a depth of 5 levels and up to 10 hosts at each WWW site. This is still the core part of the indexing data.

As more URLs were added it became clear that the Harvest interface to the network was rather crude, so the outgoing calls were done via a private proxy cache[11]. We initially used the Harvest Cache[12] system which is now a commercial product and have recently switched to the new free Squid[13] cache software which is a development of the Harvest cache. The cache system stores WWW pages so that they do not need to be re-fetched if they have not expired and can use the `If-Modified-Since` HTTP header which makes more efficient use of the network.

5 Querying the Index

The Harvest brokers provide a query service[14] using the *Glimpse*[15] indexer by default. This required a little configuration to make it return results in a sensible order, with *best* results, in some sense, first. The output of the query also needed some formatting to present the results better - this work is ongoing. Since many of the queries are repeated, future work could involve caching them for this reason and to allow more later results to be viewed without a re-query.

6 Collaboration

A posting was made on the uk.jips newsgroup about AC/DC inviting people to test AC/DC and collaborate with us. As mentioned above, Harvest is designed to handle remote updates easily. Within a few weeks several other sites had started using Harvest on their local and regional sites. The data from these are merged into the AC/DC broker regularly.

Our experience of the work was shared with the collaborating sites via a mailing list - mostly help in configuring the Harvest gatherers for this project.

[11] See URL http://wwwcache.lut.ac.uk/caching/
[12] See URL http://www.netcache.com/
[13] See URL http://www.nlanr.net/Squid/
[14] See URL http://www.hensa.ac.uk/search/acdc.html
[15] See URL http://glimpse.cs.arizona.edu:1994/

As additional sites are added, the work done by the large AC/DC WWW crawler is being reduced. This is done by removing the sites from the list of URLs used by AC/DC but also by forbidding the AC/DC gatherer from attempting access to the remote-collected sites.

7 Current state

At present AC/DC indexes 218,000 WWW, gopher and news documents covering over 1,130 UK `.ac.uk` sites using 13 gatherers and brokers around the country. The current organisation of the collaboration is shown in Figure 1:

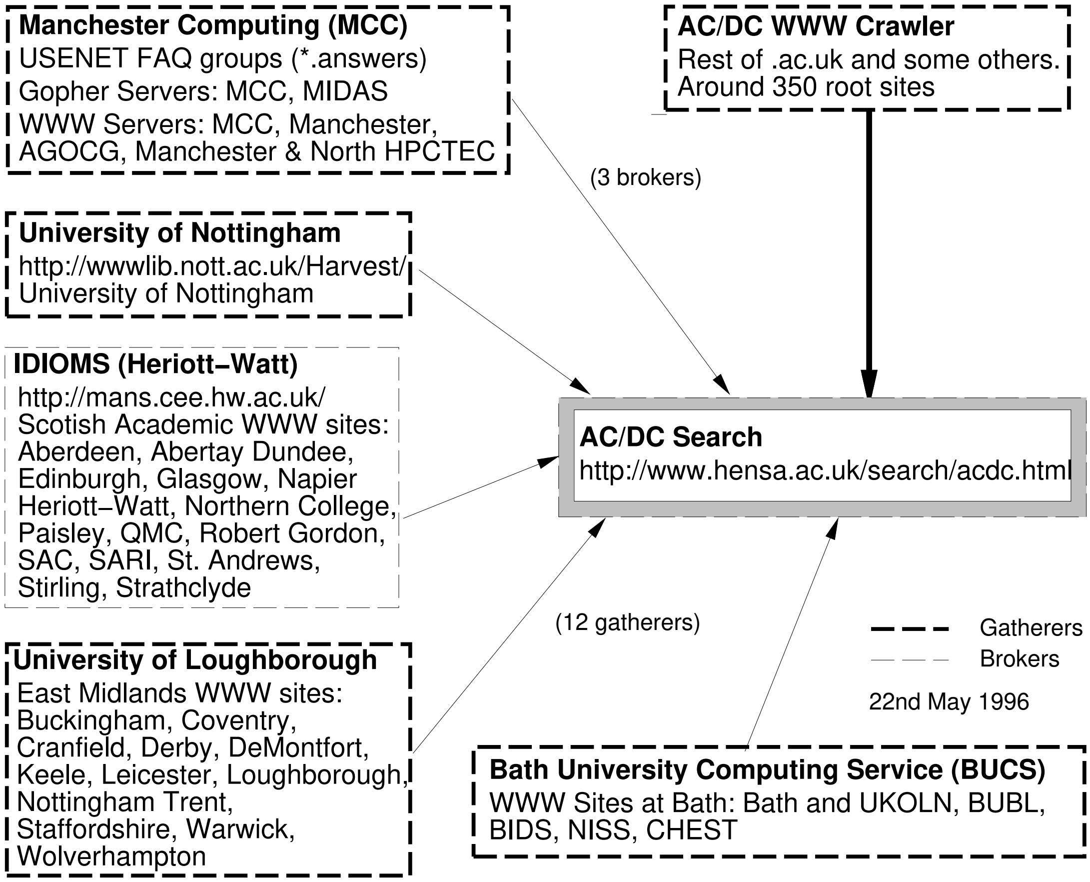

Figure 1: AC/DC Structure

8 Future

This is an *experimental*, *research* project and cannot be guaranteed to be running in the long term. If the administrative and resource requirements are low, it would be nice to continue if there is sufficient interest in this project.

The index for the gathered data is getting rather large, around 200 Mbytes. With the current indexer, this runs rather slowly and updating it too frequently causes problems. It isn't clear what the solution is to this - maybe multiple brokers indexing at different times or a better indexing system.

There were rumours on a UK newsgroup that *Alta Vista* was starting a UK based service - whether that was an outpost of the US service or a specific UK based one, we do not know.

For further information on AC/DC, please look at the AC/DC search page on HENSA/Unix at URL http://www.hensa.ac.uk/search/acdc.html.

9 Thanks

Thanks to Tim Hopkins and Maggie Bowman of HENSA Unix for their comments and support in this work.

Thanks to the software developers who wrote the freely-available or free systems used in this project: Harvest[16], Perl5[17], Apache[18], Squid[19], GNU[20] Emacs with PSGML[21] and W3[22] modes, W3C[23] Arena[24], ...

---

[16] See URL http://harvest.cs.colorado.edu/
[17] See URL http://www.perl.com/perl/index.html
[18] See URL http://www.apache.org/
[19] See URL http://www.nlanr.net/Squid/
[20] See URL http://www.cs.pdx.edu/.trent/gnu/
[21] See URL http://www.lysator.liu.se/projects/about.5Fpsgml.html
[22] See URL http://www.cs.indiana.edu/elisp/w3/docs.html
[23] See URL http://www.w3.org/
[24] See URL http://www.w3.org/pub/WWW/Arena/