

Using the Penn State Search Engine

Jeffrey D'Angelo and James Leous
`root@aset.psu.edu`

`http://aset.its.psu.edu/`

ITS Academic Services and Emerging Technologies



How Does Search Work?

In general search engines deploy a piece of software called a “crawler” which begins searching at a given URL or URLs and follows the links contained in one page to others.

There may be boundaries on this search (*e.g.* only sites in this domain or only sites N levels deep in the current site).

Crawlers collect these pages and pass all or parts of it to a “catalog” or “index” of the site(s).

The index is passed to the actual search engine software where pages are ranked according to a numerical algorithm. Google’s numerical algorithm is called PageRank.

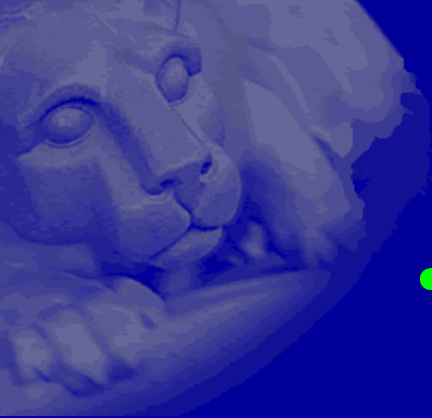
The usefulness of a crawler based search engine depends on all three of these parts.



By the Numbers

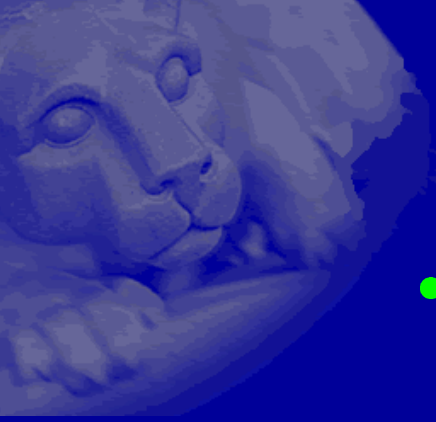
From the last crawl (Wednesday evening, October 27, 2004):

- Time:
 - took 7 hours 34 min 20 sec to query the Web servers,
 - 2 hours 55 min 29 sec to build the index
 - and another 50 min and 49 sec to replicate and test the index
 - Start to finish: 11 hr 20 min 38 sec to crawl/index Penn State



By the Numbers (cont'd)

- Amount:
 - 838,958 total URLs crawled/indexed
 - 466,077 URLs found that we excluded (non .psu.edu, containing ?, etc)
- Queries:
 - Total for September, 2004: 582,441
 - Average for September: 19,415/day
 - 1 every 5 seconds
 - about 1,200 in the time to give this seminar



Why Search?

- **Navigational** — The immediate intent is to reach a particular site.



Why Search?

- **Navigational** — The immediate intent is to reach a particular site.
- **Informational** — The intent is to acquire some information assumed to be present on one or more pages.



Why Search?

- **Navigational** — The immediate intent is to reach a particular site.
- **Informational** — The intent is to acquire some information assumed to be present on one or more pages.
- **Transactional** — The intent is to perform some Web-mediated activity.



Does Google Engine Scale?

Look at the numbers:

- www.Google.com → ~4.3 Billion pages
- search.psu.edu → ~0.83 Million pages

Look at the connections:

- www.Google.com → relies heavily on connections of “important” pages to other “important” pages
- search.psu.edu → with the exception of some central pages, very few departmental pages actually link across academic or administrative boundaries.



What Really Matters?

Although Google's PageRank algorithm is a secret, the following is known:

- Links from other pages are heavily weighted
- <TITLE> content is very important
- Keyword density within a page is important
- Keyword-laden links vs. "click here"
- Meta keyword tags are not that important



Are Tags Irrelevant?

Under our current system:

- Meta keywords are not that heavily weighted
- Penn State sample differs from Google sample with regard to interconnections
- Use keywords in <TITLE> tags of page as well as “keyword laden” links to other pages within and outside of your site
- Remember Google is current state of the art; the future is probably a more federated search mechanism



Search Integration

Integrating the Penn State Search Engine Into Your Site

- Invoking



Search Integration

Integrating the Penn State Search Engine Into Your Site

- Invoking
- Restricting Search Results



Search Integration

Integrating the Penn State Search Engine Into Your Site

- Invoking
- Restricting Search Results
- Customizing Search Results Style * New Feature!



Invoking Search Engine

Invoking the Search Engine from Your Site

- Easy and convenient way for visitors to search from your site
- Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#invoke>

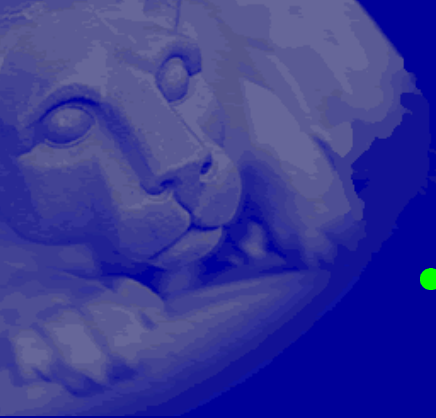


Restricting Search Results

- `as_sitesearch`

Directions:

[<http://aset.its.psu.edu/googledocs/instructions.html#restrict>](http://aset.its.psu.edu/googledocs/instructions.html#restrict)



Restricting Search Results

- `as_sitesearch` — Restrict results to only URLs that begin with the value you set for this parameter. Adds “`site:–your–url–`” to the search query.
- `sitesearch`

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>



Restricting Search Results

- `as_sitesearch` — Restrict results to only URLs that begin with the value you set for this parameter. Adds “`site:–your–url–`” to the search query.
- `sitesearch` — Restrict results to only URLs that begin with the value you set for this parameter. Does not add “`site:–your–url–`”
- `restrict`

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>



Restricting Search Results

- `as_sitesearch` — Restrict results to only URLs that begin with the value you set for this parameter. Adds “`site:–your–url–`” to the search query.
- `sitesearch` — Restrict results to only URLs that begin with the value you set for this parameter. Does not add “`site:–your–url–`”
- `restrict` — Uses a subcollection which administrators must set up for you before a crawl. Allows multiple sites to be grouped together; takes a few days to take effect.

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#restrict>



Customize Format/Style

Customize Search Results Format/Style

- Two choices for alternate format: XML and XSLT.
- XML formatted results — good for dynamic applications
- XSLT — translate HTML into your desired format — good for Web visitors



Using XML Results

- XML is the eXtensible Markup Language. It is a format containing data that can be easily passed between programs and systems.
- You can have a dynamic application query the search engine directly.
 - This application can do offline processing.
 - This application can serve as an intermediary between the Web visitor and the search engine. This allows you to provide your own formatting controls. – Downside: it requires some programming work.



Using XML Results (cont'd)

Directions on receiving XML:

<http://aset.its.psu.edu/googledocs/instructions.html#xml>



Customizing HTML results

- All search results begin in XML format. The Google Search Appliance “translates” them into another format, such as HTML.
- It uses Extensible Stylesheet Language Transformation (XSLT) documents to translate into a particular format or style. Use the proxystylesheet parameter to specify what XSLT you wish to use.
 - proxystylesheet=PennState means, use the XSLT for the collection called “PennState” which is the master index for our appliance.
 - proxystylesheet=<a URL> means, attempt to read an XSLT file at the address provided.



Customizing HTML results

Directions:

<http://aset.its.psu.edu/googledocs/instructions.html#format>



Self serve XSLT [NEW!]

The *NEW* Self serve XSLT file generator form.

- Basic knowledge of HTML required; knowledge of XSLT *not* required.
- You may design your site style in any program you wish (Dreamweaver, etc), and simply copy/paste the HTML into the Web form.

Generator URL with directions:

`<http://aset.its.psu.edu/cgi-bin/googledocs_custom_xslt.cgi>`



Final Notes

- Google Search Appliance Backup Strategy Proposal:

`<http://www.personal.psu.edu/staff/j/c/jcd/useless/google_backup_strategy/>`

- Questions?