Camtology/iLexR Spider Plans

16th April, 2009

Mark Slater

Motivation

- *iLexR would like to generate an N-Gram corpus of English language words to rival that generated by Google*
- *For those of you who don't know (which included me until recently!) an N-Gram corpus basically consists of the frequency of word combinations (from 1 to N – 6 in this case I believe) in a scanned text*
- *The idea is to use spidering software in combination with the grid resources to scan the internet and reach a total of ~1 trillion words*
- *This is similar in scale to the scan done by Google (the only people capable of doing this at present). However there are some drawbacks to their corpus:*
  - *Cutoff at 40 – any frequencies below 40 are removed*
  - *No domain info – Difficult to tell where 'sub' corpus's have originated*
- *This corpus would provide an unrivalled academic resource as well as showing that Google are not the only ones who can take on the whole internet*

The Web Spider

- *On behalf of Camtology, I have been developing a web spider over the last few months that:*
  - *Tracks both visited and queued URLs based on domain name*
  - *Runs using many parallel jobs*
  - *Avoids any accidental Denial Of Service attacks*
  - *Obeys robots.txt restrictions*
  - *Allows limits on generated traffic*
  - *Contains a user-configurable payload that can be run on every page visited*
- *To this end, I have written a new plugin application for the Ganga Job Management Tool that uses the web parser Beautiful Soup:*

http://ganga.web.cern.ch/ganga/
http://www.crummy.com/software/BeautifulSoup/

- *This takes care of all the grid submission, job management and actual html parsing*
- *This left me with writing wrapper/control scripts, etc. to meet the above requirements*

Web Spider Testing

- *I have performed a number of tests using the spider:*
  - *Oxford university sites (PDF files)*
  - *UK university sites (Image files and PDF files)*
  - *US, UK and AU university sites (Image files and PDF files)*
- *In the largest test, the spider has successfully run over ~2 million links and ~30000 domains in a few days*
- *Assuming ~200 words of English per link, we would need to visit ~5 billion links to complete the corpus*
- *I have spent the last few weeks optimising the code to improve scalability:*
  - *It now takes advantage of GLITE bulk submission*
  - *To keep track of URL lists, md5 sums of the domain names are used and stored in subdirectories based on this: xx/xx/xx/<full_md5>/*
  - *We have optimised the jobs to maximise the CPU time*
  - *It is fully automated and reports back via a web page (in development)*

Requirements

- *The timescale over which we can generate the corpus is dependant on the number of CPUs we have available to us*
- *It is very dependant on the site, but generally each link takes ~1s (both downloading, scanning for links, etc. and running the payload)*
- *We therefore estimate that 5 billion links with the current camont allowance of ~300 CPUs would take ~200 days of continuous running*
- *This scales with number of CPUs however and so if we could run 10000 simultaneous job (~2/3 of the UK grid), it would take ~6 days*
- *This is of course just an estimate and will depend on the number of domains we can scan simultaneously. However in the previous large scale tests, we were certainly limited by the CPU restriction*
- *In conclusion, any additional CE resources that can be made available would be very greatly received!*