As part of the exercises on Web Dynamics, everybody is required to do a practical project. The projects can be done in groups up to three students. There are three topics to choose from: Web Structure, Web Size Estimation and Web Crawling. The first two projects will use the ClueWeb09 Dataset, while the third one will focus on the MPII web site. In the end summarize your results, describe the techniques and the algorithms you have used and provide argumentation.

## ClueWeb09 Dataset

The ClueWeb09 `http://boston.lti.cs.cmu.edu/Data/clueweb09/` dataset was created by the Language Technologies Institute at Carnegie Mellon University to support research on information retrieval and related human language technologies. The dataset consists of 1 billion web pages, in ten languages, collected in January and February 2009. The dataset is used by several tracks of the TREC conference.

Web Pages:

- 1,040,809,705 web pages, in 10 languages
- 5 TB, compressed. (25 TB, uncompressed.)

Web Graph:

- Entire Dataset
  - Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
  - Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)
- TREC Category B (first 50 million English pages)
  - Unique URLs: 428,136,613 (30 GB uncompressed, 10 GB compressed)
  - Total Outlinks: 454,075,638 (3 GB uncompressed, 1 GB compressed)

Details about how to access the dataset will be given later.

# 1 Web Structure

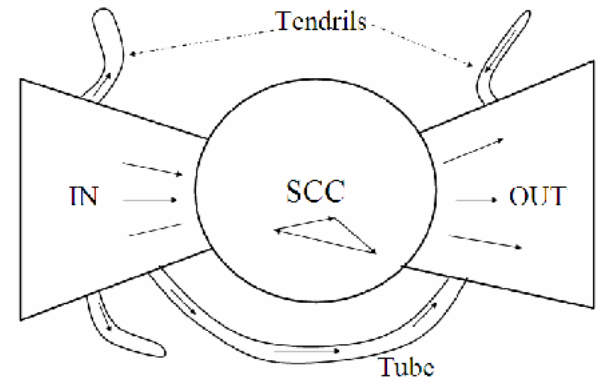


Figure 1: Web Graph Model

Analyze the ClueWeb09 dataset according to the model of the Web Graph (Figure 1) introduced by Broder et al.. Implement the algorithms for finding strongly and weakly connected components in a graph. Apply the algorithms to identify the URLs which belong to the strongly connected component *(SCC)*, to the *IN* and *OUT* subgraphs, to the tendrils, and to the tubes. Compute the diameters of the whole graph and the *SCC*.

We recommend to use *Walrus* (`http://www.caida.org/tools/visualization/walrus/`) or *JUNG* (`http://jung.sourceforge.net/`) for graph processing and analysis.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener: *Graph structure in the Web*, Computer Networks, Vol. 33, No. 1. (June 2000), pp. 309-320.

## 2 Web Size Estimation

Use the ClueWeb09 dataset to estimate the size of the web with the method developed by Bharat and Broder. Implement a sampling procedure for picking pages uniformly at random pages from a search engine and from the data set. Implement a checking procedure or determining whether a particular page is indexed by the search engine or is part of the dataset. Analyze the results and using the figures about ClueWeb09 give an estimation for the size of web as indexed by the search engine.

Use the public search API for the communication with a search engine of your choice.

K. Bharat, A. Broder: *A technique for measuring the relative size and overlap of public Web search engines*, Computer Networks, Vol. 30, No. 1-7. (1998), pp. 379-388.

## 3 Web Crawling

Crawl the MPII `http://mpi-inf.mpg.de` site between 14.07.2009 and 17.07.2009 with no more than 100000 queries for the whole period. Cover as many different page versions as possible. Report how many changed pages you have detected. Crawl regularly the web site several weeks before the experiment in order to get insights about the behaviour of the different pages. Try to estimate the change rates of the pages with the approach developed by Cho and Garcia-Molina. Use the change rate information to design an optimal crawling strategy for the experiment.

We recommend to use the open source web crawler *Heritrix* (`http://crawler.archive.org/`).

J. Cho, H. Garcia-Molina: *Estimating frequency of change* ACM Transactions on Internet Technology, Vol. 3, No.3, pp. 256-290

J. Cho, A. Ntoulas: *Effective change detection using sampling* Proceedings of the 28th international conference on Very Large Data Bases, Hong Kong, China, pp. 514-525