T-122.101 Mining and Modeling the Web
Practical Assignment

Fall 2003

Jaakko Hollmén, Anne Patrikainen

1 Background

The main aim of this practical assignment is to construct a search engine for ranking scientific papers related to high energy particle physics. These SPIRES-HEP papers are collected from the e-print archive arXiv. The data set was originally used in the data mining competition KDD Cup 2003 (see http://www.cs.cornell.edu/projects/kddcup/). The following introctory paragraphs are copied from the KDD Cup web pages.

The e-print arXiv, initiated in Aug 1991, has become the primary mode of research communication in multiple fields of physics, and some related disciplines. It currently contains over 225,000 full text articles and is growing at a rate of 40,000 new submissions per year. It provides nearly comprehensive coverage of large areas of physics, and serves as an on-line seminar system for those areas. It serves 10 million requests per month, including tens of thousands of search queries per day. Its collections are a unique resource for algorithmic experiments and model building. Usage data has been collected since 1991, including Web usage logs beginning in 1993. On average, the full text of each paper was downloaded over 300 times since 1996, and some were downloaded tens of thousands of times.

The Stanford Linear Accelerator Center SPIRES-HEP database has been comprehensively cataloguing the High Energy Particle Physics (HEP) literature online since 1974, and indexes more than 500,000 high-energy physics related articles including their full citation tree.

2 The task

A. Construct a search engine (20 points)

You are required to construct a search engine for ranking scientific papers in the SPIRES-HEP database. As data, you will use the abstracts of the papers (there is too much data; whole papers cannot be used) and the citation structure of the papers.

You are free to use any method of your choosing to rank the papers according to a query. It is desirable that you use both text analysis and link analysis methods to judge the relevance of the papers.

B. Analyze the citation graph (5 bonus points) (optional!)

You do not need to complete this task, it is only for the purpose of acquiring some bonus points. You may freely choose one aspect of the citation graph to analyze and apply whatever methods you choose in your analysis. Some possibilities are the analysis of the cluster structure of the citation graph, or the various power law distributions in the graph, or examining the importance of individual authors, etc. You're invited to use your imagination here.

3 Timetable

The first deadline is *Thursday 20.11, 9:00.* By this time, you have to choose two query strings (for instance "particle accelerator CERN" or "pink elephant ice cream") and send these to the course assistant (annep@james.hut.fi). You might pick queries for which your engine performs well or queries which you think are difficult for the other search engines.

Shortly afterwards, the assistant will mail the list of all suggested query strings to all course participants. Your task is to run your search engine for every query string. You must save your results as ascii text files and send these files to the assistant by *Thursday 27.11, 9:00*.

Then the assistant will send all results to everybody. This allows you to compare your search engine's performance with everybody else's.

The absolute deadline for the written report is *Thursday 4.12, 9:00*. The report might be some 5 pages long and contain the descriptions of the methods you have used, the justification for using them, and the analysis (qualitative or quantitative) of comparing your search engine against the other search engines. At this point, also provide the assistant with the source code of your search engine. Furthermore, if you have completed part B, include the results in your report.

There will be a feedback session for the practical assignment on *Wednesday 10.12*, normal seminar time and place. In this session, we will see why everybody thinks their search engine is better than everybody else's :-).

4 Practical details

You can find the data sets in http://www.cs.cornell.edu/projects/kddcup/datasets.html. There is more data than you need; the relevant files are numbers 2 and 4.

Number 2: The abstracts for all the hep-th papers as a hep-th abstracts tarball. For each paper the abstract file contains: arXiv submission date, revised date(s), title, authors, abstract.

Number 4: The citation graph of the hep-th portion of the arXiv as a hep-th citations tarball. The format for citations is a sorted 2 column vector where the left column is the cited from paper arxiv id and the right column is the cited to paper arxiv id.

When you mail your results to the assistant, save the results of each query in a separate file. The ascii file format is as follows:

`ranking; arXiv id; name of the paper`

Hopefully, your output is sorted by the ranking. The ranking is an integer (1, 2, 3, ...), NOT the relevance score. You should use file names of the form "searchenginename_querynumber.txt".

You do not need any fancy query handling; considering the query words to have an equal importance should be fine.

You should come up with a funny name for your search engine in order not to confuse it with the other search engines.

Perhaps you wish to start by selecting a subset of all possible words to work with (eliminate stop words, etc). Stemming might also be needed. You may use text analysis methods as preprocessing for the link analysis methods, or use both at the same time to rank the papers. Perhaps you might also consider things such as emphasizing the importance of the words in the titles of the articles.

5 Evaluation

You need to obtain one half of the points (10) in order to pass the course and 3/4 of the points (15) in order to pass with distinction. (As you notice, the bonus points are not included in the point total; they are purely extra.) You will receive full credit for part A if

- you manage to construct a search engine that gives reasonable results,
- you are able to convincingly justify the methods you have used in designing the search engine,
- your engine is at least as good as the other engines (according to your own judgement).

You will receive full credit for part B if you manage to present clever in-depth analysis of some aspect of the citation graph structure. You may use the bonus points of part B to substitute for e.g. some missing exercise points. Note that completing part B is purely optional!