

CS276B
Text Information Retrieval, Mining, and Exploitation

Practical 1
Jan 14, 2003

The course project

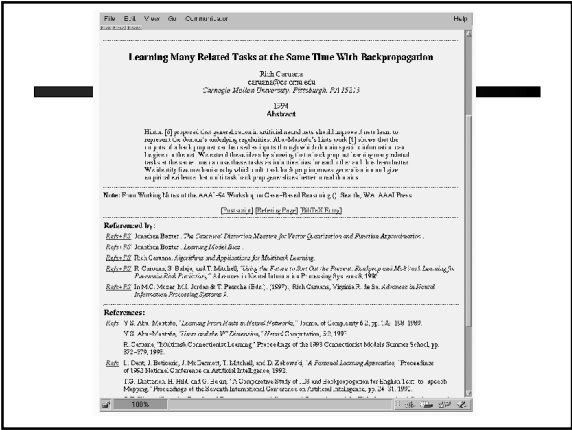
- Building a digital library of academic papers, from those freely available on the web
- This is a great learning context in which to investigate IR, classification and clustering, information extraction, link analysis, various forms of text-mining, textbase visualization, collaborative filtering ... really everything we cover in this course (and the last one)
- Project name?
We're looking for a good one!

Organization & scope

- This is a project of reasonable scope; our plan is to have people taking the class work together to implement the components of it
- So a secondary benefit should be some exposure to software engineering issues...
- But it's not an impossibly large project
 - The main components are a series of stages that map between clearly defined data representations
- Several groups of two did things like components of this as components of their projects last quarter

Motivations/Predecessors

- Machine Learning Papers [Andrew Ng, defunct]
- Cora [Just Research, Andrew McCallum, defunct]
- CiteSeer/ResearchIndex [NEC Research]
 - <http://www.citeseer.com/>
- Highwire [Stanford]
 - <http://highwire.stanford.edu/>
- There are various other online archives, but this service isn't available for most disciplines



Organization

- Two halves:
 - In first half, people will build basic components, infrastructure, and data sets/databases for project. In two phases:
 - First steps
 - Further development (extensions, needed fixes)
 - Second half: student-designed project, which will focus in on a particular issue of interest related to goals of this project
- In general, work in groups of 2 on projects

Timeline

- Tue Jan 14 [today]: Phase 1a starts
- Mon Jan 27: Phase 1a due [and name suggestion!]
Course staff integrates, debugs, evaluates
- Thu Jan 30: Phase 1b starts
- Tue Feb 11: Phase 1b due
Course staff integrates, debugs, evaluates
- Tue Feb 18: Phase 2 project plan due
- Tue Mar 4: Phase 2 project check-in point
Course staff integrates, debugs, evaluates
- Wed Mar 12: Phase 2 due
- Thu Mar 13: Presentation of projects in class

Grading

- Project will be 40% of the grade distributed over phases:
 - Phase 1a: 8%
 - Phase 1b: 8%
 - Phase 2: 24% (4% hand out for check-in point)
- Phase 1 will mainly involve getting parts of a system working and well-implemented. In evaluating it, we'll value good systems engineering as well as course-related stuff
- Phase 2 is meant to be a research project: you'll write up a research report/paper, and it'll be evaluated largely based on its quality.

Opportunities for improvement

- Much of citation search is fielded search, and a text search interface is awkward
- Citations are not very well parsed
- Duplicates are poorly detected
- Lots of things that you could do with link analysis (important conferences, cliques)
- Getting reference information from HTML pages as well as papers
- Subject classification (esp. if broad domain)
- Visualization
- Using collaborative filtering

Take initiative and ask questions

- You should look to acquire information relevant to solving these problems well
 - There are lots of relevant papers on many of these problems
- We're here to help!
 - We'd like this project to succeed, and would be eager to answer questions and give advice on how to do things
 - There may also be things in our rough specification that actually *need* correcting
 - Talk to Teg (and other staff)

Basic processing stages

1. Crawler downloads HTML pages that contain links to papers, and papers
2. Focused crawler does this intelligently
3. Extract links and context from HTML
4. Convert papers to (marked up) text
5. Decide if they're really research papers
6. Extract header (author, title, abstract) and references sections
7. Separate citation block into individual citations

Basic processing stages

8. Do information extraction of author, title, etc. information in citations
9. Find context(s) of each citation in body of paper
10. Work out sets of variant forms for each person name, conference, paper (de-duping)
11. Normalize citations to unique full form
12. Map citations to papers to which they refer
13. Build Lucene IR system index (with fields)
14. Provide UI for querying, browsing (and visualization)

Tools

- We don't need to reinvent the wheel. There are lots of tools that you can and should use for various stages:
 - Lucene IR engine
 - MySQL database
 - PS/PDF to text engines
- We'll do the project in Java
 - Good URL handling, multithreading, etc.
 - various packages for all sorts of things (e.g., touchgraph for visualization)

Computers etc.

- We're going to start off doing development on Leland systems, with a CVS repository there
 - We've got some small data sets, and you may make others
 - At this stage, keep small: just download a couple of hundred papers, restrict yourself to the Stanford domain, etc.
- Later in the quarter we'll transition things to a dedicated Linux machine (under construction) and attempt to run it on a larger scale...

Questions?

- Ok, concrete organization time...