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.citeseeer.com/
- Highwire [Stanford]
  - http://highwire.stanford.edu/
- There are various other online archives, but this service isn’t available for most disciplines
1. Transferring Cooperative Machine Learning Strategies to Human Groups
   
   Dan C. Regan, Luc A. Bechet
   Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road
   
   Abstract: This paper proposes a model of cross-collaborative learning research involving learning and human learning the importance of collaborative learning strategies into human learning. The underlying research question is the role of cooperation in support of individual learning. The use of computer-based tools prompts the development of a new approach to learning and teaching known as the "cooperative learning" approach. It is proposed that humans could be taught to learn cooperatively. The paper investigates artificial learning strategies with human agents. The results of the experiments show situations in which machine learning agents using reactive learning strategies are superior to individual machine learning agents.

   Keywords: Machine learning agents, machine learning, cooperative learning, human learning, computer-based tools.

   References: Please see the original paper for detailed information.

2. Training and Intelligent Agents
   
   Arara Masada
   Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260
   
   Abstract: This paper presents a model of the research in intelligent agents, describing the use of computer-based agents for cooperative learning and collaborative strategies. The model of computer-based agents has been developed to support learning by integrating computer-based agents with students' experiences. The work was carried out in support of a project on learning and teaching, various students' experiences, and the use of computer-based tools in learning. The major hypothesis of this research is that computer-based agents can support learning and teaching, and the use of computer-based tools in learning and teaching can support learning.

   Keywords: Machine learning agents, machine learning, cooperative learning, human learning, computer-based tools.

   References: Please see the original paper for detailed information.

3. Mobile Intelligent Agents for Dynamic Classification and Retrieval of Machine Learning Approaches
   
   Pierre Vang, Parthasarathy, and Vassilis R. Tsotras
   
   Abstract: This paper presents an experimental study of mobile, computer-based intelligent agents for dynamic classification and retrieval of machine learning approaches. The mobile agents are autonomous, intelligent, and can interact with other mobile agents to obtain information about machine learning approaches. The mobile agents are designed to provide users with access to a wide range of machine learning approaches, and to assist users in their decision-making process. The mobile agents are also designed to be able to learn from their experiences and adapt to new situations.

   Keywords: Machine learning agents, machine learning, cooperative learning, human learning, computer-based tools.

   References: Please see the original paper for detailed information.
Learning Many Related Tasks at the Same Time With Backpropagation

Rich Caruana
caruana@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, PA 15213

1994
Abstract

Hinton [5] proposed that generalization in artificial neural nets should improve as nets learn to represent the data's underlying regularities. Alu-Mustafa's latest work [1] shows that the inputs of a backprop net can be used, as in a single agent, to learn general information. We extend these ideas by showing that in a net with many related tasks, the same input can create different patterns in each task. The net has been trained to identify five mechanisms by which a single input can affect the output of a network in all tasks.

Note: From Working Notes at the AAAI-94 Workshop on Case-Based Reasoning, Seattle, WA. AAAI Press

References:

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!]
- **Thu Jan 30:** Phase 1b starts
- **Tue Feb 11:** Phase 1b due
- **Thu Feb 18:** Phase 2 project plan due
- **Tue Mar 4:** Phase 2 project check-in point
- **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. Focussed 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...