605.744 Information Retrieval: Class Project

Goals

The class project allows you the opportunity to investigate a particular topic of information retrieval in greater depth than we could cover in the classroom. Projects are normally individual endeavors, and require advance planning and progressive effort to complete successfully. The clear majority of projects tend to involve writing or using software to conduct an experiment or analyze a textual dataset. However occasionally I have had students work on projects that are more theoretical vs. empirical. Some example projects[1] include: critically evaluating open source retrieval software using one or more IR test collections; analyzing a collection of police crime reports and performing data mining; or, developing a new algorithm for text classification and comparing it against a known baseline. For such projects, the relevant literature must be reviewed to give context to your work, a hypothesis must be developed, experiments must be designed, conducted, and analyzed, and finally, results must be presented in a report, and orally to the class. "Theory" projects are expected to include a much more extensive review and study of the literature; they should exhibit independent thinking, and the written report is expected to be longer and more comprehensive (~ 15 pages vs. 5-6 pages).

To earn a grade of A- or higher in the course, students must complete a project; however, completing a project does not guarantee receiving an A- or higher. Students are permitted to opt out of submitting a project, in which case the other coursework will determine the final grade, as discussed in the course description handout.

Grading Criteria

Project grades are based on the proposal (10%), the work performed and the written report (70%), and the class presentation (20%). A checklist I use for scoring oral presentations is attached.

Proposal

A written proposal must be submitted and approved. The proposal should identify a topic of interest, briefly motivate why this is an interesting or important problem, identify relevant scientific literature for the problem of interest, identify sources of data, and outline planned work for the project. Sufficient details about data, experimental design, and evaluation methodology should be provided. Proposals should be about 1 page in length. If you have a topic that interests you, but you have questions or are not sure how to proceed, you are encouraged to contact me informally for ideas, even before the proposal is due. I give feedback (and approval) on proposals, so it is important to think about an idea and submit the proposal on time. You may find some useful links or ideas from browsing the course web page: http://pmcnamee.net/ir.html

Written Report

The written report is the most significant project deliverable; it is where you document the work you have performed and it counts for most of the project grade. Reports should be scientifically-oriented and ought to include an abstract, an introduction to the problem being considered, a review of related work (extensive if a theory paper), discussion of ideas (extensive if a theory paper),

[1] A long list of ideas is attached below.

description of data (if relevant), experimental results with analysis (extensive if an empirical paper), conclusions supported by your work, and appropriate references. To compute page lengths you may assume 500 words per printed page. Generally the style of formatting is up to you; however, do include headings, and use a font between 10-12 points. Suitable tables and figures are highly encouraged. Written reports should be handed in during the last class meeting.

Presentation

A short oral presentation must be delivered to the class on the final day of the semester. Further guidance will be forthcoming, however, plan on giving a talk of around 10 to 12 minutes in length. Students should use prepared overheads or electronic (*e.g.,* PowerPoint) slides. As part of the presentation you may also demonstrate a software system, if that makes sense for your project. Presenters should clearly introduce the problem under discussion, briefly review prior work from the literature, explain in detail your contribution and results, and be able to intelligibly field questions. Experiments are not always successful, you can achieve a good score on the project, even with negative results; however, your design should be good and you need to articulate what *was* learned. If your work is theoretical you should illustrate how it can be evaluated and what applications should benefit from your ideas.

Schedule

By 3/13	Pick a topic and submit a written proposal (one page is enough, email is fine)
4/9-15	Send me a status report (one paragraph can be enough; email is fine)
5/1	Oral presentations & report is due

Literature

Numerous resources are available to you. Online papers can be found via Google Scholar, CiteSeer, or various websites for conferences; pointers to the TREC conferences and the ACL Anthology are on the course web page. I believe that the JHU libraries can provide no-cost access to the ACM and IEEE digital libraries. Finally, I may be able help in obtaining copies of difficult to find papers using my private collection.

Sample projects of former students

- Indexing and Computing Document Similarity using Hadoop
- Evaluating indexing and retrieval of Hindi song titles
- Analysis of online police crime reports and classification of crime narratives.
- Parallelization of machine learning algorithms and cloud-based classification.
- Exploring methods to compress indexes using document identifier reassignment
- Extracting obituary information from news sources for genealogical purposes
- Distributed indexing using Bloom filters
- Predicting author gender, time of authorship, or identify of author
- Extraction of apartment rental information from Craigslist ads
- Proposing a theoretical model for detecting click-through fraud
- Attempting the NetFlix challenge
- Collaborative filtering using beer recommendation reviews (from pintley.com)
- Attempting to predict the market using publicly disclosed financial documents (SEC filings, transcripts)

- Exploitation of open source information for maritime domain awareness - matching pictures of ocean-going vessels (from Flickr) to Coast Guard databases.

Empirical Ideas

- Can POS-tagging be used to improve IR performance?
- Can a given NLP technique improve performance (*e.g.,* keyword phrases or stemming)?
- Using electronic thesauri to automatically augment user queries
- Learning to spell correct or phrasify (add quotes to) user's web queries.
- Apply a machine learning algorithm (*e.g.,* SVMs/NNs/Decision Trees) for text filtering
- Develop and test a method for spam filtering.
- Implement an algorithm for document similarity that we did not cover extensively in class, such as Cover Density Ranking or Latent-Semantic Indexing. Compare your results to some baseline method such as vector-cosine or an out-of-the-box IR package.
- Experiment with cross-language retrieval by manually translating some of the HW#3 queries into another language and using a bilingual dictionary or on-line MT system to translate queries back to English prior to search.
- Write a mini intranet search engine with at least 5000 documents, analyze the contents and demonstrate an engine to search it
- Implement phrase-indexing efficiently (see work by Bahle et al.)
- Build a system for detecting when online reviews (Yelp, Travelocity) are likely to be genuine vs. fake.
- Build and evaluate a translation resource (*e.g.,* dictionary or parallel corpus) obtained from the Web
- Obtain a speech recognition package and run it on web-available audio files to support retrieval.
- Help users visualize textual information (such as a retrieved document set)
- Develop a system for retrieval of music
- Build a Web collection of 100k HTML documents. Analyze it in significant detail.
- Attempt retrieval of stored images (this is hard).
- Develop an information extraction system that learns a particular kind of fact from unstructured documents (*e.g.,* crimes: perpetrator, victim, date, officers involved)
- Build a system for collaborative filtering, to match people with similar interests, or to suggest movies, books, wines, etc... to an individual based comparing their profile with other’s
- Question answering for one particular type or question (e.g., how many or who).

Theory / Application Ideas

- What problems are current large-scale evaluations (like TREC) susceptible to?
- Develop a framework for retrieval against scanned document images
- Investigate retrieval of a specialized type of document (*e.g.,* a collection of source code or job openings).
- How can the relative efficacy of two *web* search engines be established?
- How can spam filtering software adapt to new changes (trends) from commercial spammers?
- How can commercial engines counter attempts to falsify click-throughs?
- Extensively compare desktop search tools (Google's Desktop Search vs. Apple's Spotlight)

Oral Presentation Scoring Sheet

Student:

Date:

Topic:

1. Were the project’s goals and motivation sufficiently explained? (1-10)

2. Was suitable and meaningful background information presented (e.g., prior work)? (1-10)

3. Did the talk provide sufficient technical detail (1-5) and articulate a novel contribution? (1-5)

4. Clarity of the oral presentation and argument (1-5) and quality of AV/materials. Finish on time? (1-5)

5. Was the work well thought out? Did conclusions follow from the argument or experiments? (1-10)

6. Other comments.