BLM5121: Introduction to Web Mining (3 CR)

Some Ideas for Future Projects

- We know that social networks tend to become polarized around controversial topics in social media, for example political discourse on Twitter. It has been suggested that social media increase such polarization, by making it easy for people to communicate only with those who think like themselves. This is the phenomenon of so-called "echo chambers." Can you think of a way to find evidence for such a **causal** relation?
- Wikipedia's publicly-available data set provides a wealth of information about the structure and evolution of a dynamic socially-edited forum. (i) Possible research includes automated fact-checking and user authority measures (date of registration, number of posts, average size and longevity of posts, etc). (ii) Another avenue of investigation would be correlating IP addresses / usernames with changes to specific hot-topic issues. An automated method to detect suspicious editing / revisions could aid in the identification of biased or self-serving modifications leading to the identification of the offending individuals / organizations.
- Create a meta-search engine that finds potentially "embarrassing" personal material (photos, videos, text, etc) by mining various sources such as search engines, social network sites, photo and video sites, etc. This could play an educational service by highlighting the dangers of posting private information on the Web.
- Develop an application (eg for the Google Desktop, the Yahoo Desktop, Windows Desktop, browser extension, Mac Dashboard Widget, or mobile platform) that implements some (simplified/extended) version of the HITS algorithm, discussed in class. This would be a client-based solution for the query-time analysis.
- The economics of Google Ads are inducing (creating incentives for) a "pollution" of information on the Web. People create fake "original content" with popular query terms to attract traffic and make a profit through advertising. This is being done both manually (by underpaid hired writers) and with automatic text generation scripts. Can we devise techniques to clean the Web from such pollution? UPDATE: Google recently announced changes to its ranking algorithm to demote so-called content farms.
- Try to come up with a list of top-X sites frequented by spam harvesters. For example I created a gmail account to submit a script to CPAN, which posts the email of authors on their site. That account has quickly become a honeypot with thousands of spam messages. So clearly spammers harvest emails from CPAN. What other sites? What are the worst? You could write a crawler that automatically posts email addresses to sites it encounters (message boards, etc) and then monitors which sites generate the most spam.
- A machine learning method to classify an arbitrary Web page as blog or not blog, for crawling purposes.
- A text mining algorithm to find a huge set of triples (email address, name, address) from crawls of personal websites, blogs, bios, etc. and cross-reference with structured databases such as ip-to-zip converter, phone book, and other online resources.

Project Proposal Format

The project proposal is free-format but cannot be longer than a page in length. Use your best judgement about margins and font size (fitting too much on a page would be a bad idea!). The

proposal should be concise, concrete, focused and to the point. It should answer a few basic questions:

1. Why? (Motivate your idea; is it interesting, important, relevant?)
2. What? (Exactly what do you propose?)
3. How? (State your hypothesis and evaluation procedure)
4. When? (You need a realistic timetable and deliverable; is it doable?)

Top

Project Wiki

Each group should maintain a wiki page about their project. The Oncourse site wiki tool is available for this purpose. The function the wiki is that of an open lab notebook, to note planned research, problems and solutions, delays, preliminary results and analyses, and to track progress along the proposed timeline and toward the stated project objectives, as well as any deviations from the approved proposal. The instructors will monitor the wiki on a weekly basis to check for progress, and the wiki will be used as a component of the project evaluation toward the course final grade.

Top

Sample of past projects

Spring 2006

1. Usage Statistics of Robots Exclusion Standard
2. Mining for Blog communities
3. KidsCrawler
4. Using Page History to Rank Search Results
5. Web Topology of the Indiana University Domain

Top