CS341: Project in Mining Massive Datasets

Infosession

Jure Leskovec
Anand Rajaraman
Jeff Ullman
Chris Re
Rok Sosic
Andreas Paepcke

CS341: Project in Data Mining

- **Data mining research project on real data**
  - Teams of **3 students** **(Use Piazza on CS246 to form teams)**
  - We have room for **10-15 teams**
- **We provide:**
  - **Data**
  - **Computers** (Amazon EC2, ~~**3k$ per team**)
  - **Mentoring:** Each group will have an assigned mentor that they meet on a weekly basis
- **You provide:**
  - Project proposals
  - Effort

CS341: Schedule

- Today (3/3): Info session.
- Friday 3/18: Project proposals due.
- Friday 3/25: Admission results.
  - 10 to 15 projects will be admitted.
- Mon 3/28: First class meeting in Herrin 195.
- Mon 5/2 and Weds 5/4: Midterm presentations.
- Week of May 30: Final presentations.

Projects: Proposal Must Address

- **(1) What is the problem/question your team is solving?**
  - **Give a brief but precise description or definition of the problem or question**
    - **Examples:**
      - **(a)** Analyze the data to understand why editors are leaving Wikipedia
      - **(b)** Build a social recommender engine for movies
      - **(c)** Design a better MapReduce algorithm for finding clusters in graphs
- **(2) What data will you use?**
  - **Why is the data you plan to use appropriate? Does it have the right labels/information?**
  - **It is ok to use your own data** (give detailed description)!
    - **Examples:**
      - **(a)** Wikipedia edit history where every action of every user is recorded
      - **(b)** We **crawled** Yelp and obtained X million reviews from Y million users
      - **(c)** We will use the Altavista web graph on X million nodes.

Projects: Proposal Must Address

- **(3) How will you solve the problem? What is your plan of action?**
  - **Describe and think about your approach!**
    - What method, algorithm, technique? How will you scale it up?
    - **Be as specific as you can!**
    - **Examples:**
      - **(a)** We will create edit histories of every article. We will then compare article edit histories and argue that users are leaving since all the "easy/obvious" articles have already been written
      - **(b)** Our hypothesis is that friends have similar tastes. We will include a regularization term to a Latent Factor Rec. Sys. which will encourage neighboring users to have similar parameters
      - **(c)** We will implement a scalable Frequent-itemset-based approach to identify cluster seeds (complete bipartite subgraphs). In the second pass we will then use a random walk based approach to expand around the seed and extract the clusters

Projects: Proposal Must Address

- **(4) How will you evaluate your method?**
  - **How will you measure performance or success of your method? What baselines will you use?**
    - **Examples:**
      - **(a)** Using insights from our analysis we will build a model that will predict how complete is the article (much the article will change in the future). We will evaluate predictive accuracy of the model
      - **(b)** We will measure **RMSE** of our system. As a baseline for comparison will use traditional latent-factor recommender
      - **(c)** We will measure resource usage and execution time of our algorithm and compare it to open source algs. Metis and Graclus
- **(5) What do you expect to submit/accomplish by the end of the quarter?**

Projects: Proposals

- **Submit** to cs341-spr1516-staff@lists.stanford.edu
  - **PDF should include**
    - **Project title**
    - **Project narrative addressing the 5 questions**
    - **Information about team members:**
      - **For each team member: 5 line CV/Bio** about prior experience, and why you are prepared to take this course
  - **No page limit** (but we don't promise to read past page 3)
  - **Due Friday 3/18 11:59pm Pacific time**
- **We will let you know whether you got in by Friday March 25**

SNAP Datasets

Collection of over 70 web and social network datasets: http://snap.stanford.edu/data

- **Social networks:** online social networks, edges represent interactions between people
- **Twitter and Memetracker :** Memetracker phrases, links and 467 million Tweets
- **Citation networks:** nodes represent papers, edges represent citations
- **Collaboration networks:** nodes represent scientists, edges represent collaborations (co-authoring a paper)
- **Amazon networks :** nodes represent products and edges link commonly co-purchased products

News Media

- **Online media**
  - **Collection of over 6B news documents and 300M short textual phrases that appear in them**
    - **Think of this as a complete trace of Internet news media space for the last 6 years!**
- **Goal:**
  - Detect trending topics and explores the dynamics of online news
    - Based on time, named entities, mutation of information

Microsoft Academic Graph

- **Exhaustive dataset of scientific papers**
  - 123M authors, 123M papers, 757M references
  - Affiliations, keywords, conferences, journals
  - 1.9 billion items, ~100GB
- **Problem**
  - Many duplicate entities
    - donald knuth appears 158 times
- **Goal**
  - Use textual and network structure features to identify duplicate entries

Online Reviews: Amazon

- **18 years of Amazon reviews up to March 2013**
  - Product and user information, ratings, review text

Dataset statistics	
Number of reviews	34,686,770
Number of users	6,643,669
Number of products	2,441,053
Users with > 50 reviews	56,772
Median no. of words per review	82
Timespan	Jun 1995 - Mar 2013

http://snap.stanford.edu/data/web-Amazon.html

Generic Places to Find Problems

- Kaggle (www.kaggle.com) runs competitions.
  - You can get both data + ideas + possibly win.
- Yahoo (http://webscope.sandbox.yahoo.com/)
  - Interesting datasets, no problem suggestions, but some ideas should be obvious.
- TREC (http://trec.nist.gov/).
  - Current and historical competitions.
  - May take a week or more to get authorization for data.

Send in Your Proposals

- For more detail on a dataset or problem, please contact the appropriate instructor
  - Andreas Paepcke (paepcke@cs.stanford.edu)
  - Anand Rajaraman (datawocky@gmail.com)
  - Chris Re (chrismre@cs.stanford.edu)
  - Rok Sosic (rok@cs.stanford.edu)
  - Jeff Ullman (ullman@gmail.com)
- Emails for outside contacts are provided at i.stanford.edu/~ullman/cs341slides.html