CS6784 - Advanced Topics in Machine Learning
Understanding Archives
Spring 2010
Thorsten Joachims
Cornell University

Archives
Motivation: We now have more then >10 years of online
- Newspaper archives
- Conference proceeding
- Personal email and photos
- Blogs, Wikipedia(?), etc.

- Archival, self-referential process of corpus development

Possible Research Questions
- How did the topics in the corpus change over time?
- What are the influential articles?
- Who were the most influential authors?
- How did the topics change over time?
- What are the related articles?
- Did one article influence another article?
- Who were the most influential authors?
- Who are the bloggers that are ahead of the curve?
- News: New stories identification. Remove redundancy
- Reflection: how do you spend your time/money.
- Social influence, how do stories travel.
- Photos to stories, reduce information. Your year in photos.
- Speed up desktop search, make interactive.
- Temporal representations as a way of organizing search.
- Collaborative search, use the many humans.
- Time-aware search, consistency across corpora
- Self-organizing encyclopedia, multimedia
- Predicting trends, life cycle
- What blogs are hot, personal interest.
- Visualizing social network
- Categorizing images, use the many images on the internet.
- Handling ambiguity in search
- Graphs separated
- Evolution of information, obligations
- Visual and multimedia features, trends
- Relative time as they may be part of query?
- Search as a means of a collection
- Why are we storing archives? Events, personalities, Change of personality

Change in Topics over Time:

Finding Related Articles
- Problem: Impose structure on set of documents
  - Approach 1: Content similarity
  - Approach 2: Usage data
- Experiment: Find related papers on ArXiv
  - Use http-logfiles

Usage Data on ArXiv
ArXiv HTTP access log data (14 years), 156GB
Co-Access as a Measure of Relatedness

- Co-Access
- Session: "Uninterrupted sequence of accesses of same user"
  - More detailed measures possible
  - Further restrictions on sessions to use
  - Ordering of accesses (abstract view ➔ download)
  - Usage of time intervals (between accesses, …)
  - …

Citation-based Measures of Relatedness

- Co-citation:
- Co-reference (Bibliographic Coupling)

Experiment: Coverage

- “Today” = December 31, 2004
  - Age of paper is #months published before today.
  - Compute Co-Access and Co-Citation as of “today”.

How many papers have non-zero Co-Access / Co-Citation?

Experiment: Accuracy

- “Today” = December 31, 2004
  - Age of paper is #months published before today.
  - Compute Co-Access and Co-Citation as of “today”.

How accurately do Co-Access / Co-Citation predict “related”?

- Related = both paper in a reference list in 2005.
- Pick one paper from reference list, rank all <= 2004 papers by co-access/co-citation, measure rank of other papers in reference list.