CS6784 - Advanced Topics in Machine Learning

Understanding Archives

Spring 2010

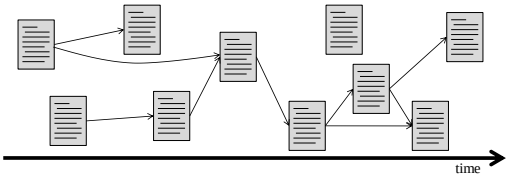
Thorsten Joachims
Cornell University

Archives

Motivation: We now have more then >10 years of online

- Newspaper archives
- Conference proceeding
- Personal email and photos
- Blogs, Wikipedia(?), etc.

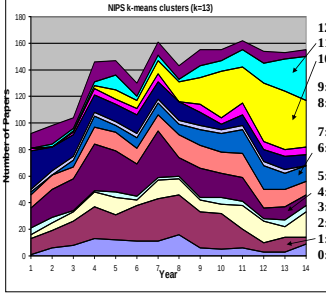
- **Archival, self-referential process of corpus development**



Possible Research Questions

- How did the topics in the corpus change over time?
- What are articles are related?
- Did one article influence another article?
- Who were the most influential authors?
- Who are the bloggers that are ahead of the curve?
- An automatic personal diary from email and photos.
- 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 other peoples traces.
- Time-aware search, consistency across corpora
- Self-organizing encyclopedia, mult-media
- Predicting trends, life-cycle
- What blogs are hot, personal interest.
- Visualizing social network
- Categorizing images, use the many images on the internet.
- Questions qnswering
- Handling analogy in search
- Google squared
- Evolution of information, wikipedia
- Trends and relationships between trends
- Changes of scene over time (time travel in mages)
- Relative time in time (use time as part of query)
- Search as a zoom of a collection
- Why are we storing archives? Events, personalities, Change of personality

Change in Topics over Time: Neural Information Processing Systems (NIPS) 1987 - 2000

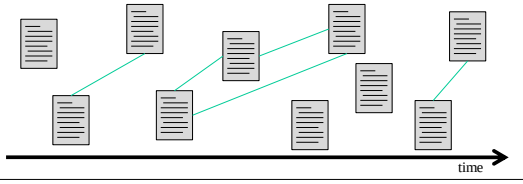


12: chip, circuit, analog, voltage, vlsi
11: kernel, margin, svm, vc, xi
10: bayesian, mixture, posterior, likelihood, em
9: spike, spikes, firing, neuron, neurons
8: neurons, neuron, synaptic, memory, firing
7: david, michael, john, richard, chair
6: policy, reinforcement, action, state, agent
5: visual, eye, cells, motion, orientation
4: units, node, training, nodes, tree
3: code, codes, decoding, message, hints
2: image, images, object, face, video
1: recurrent, hidden, training, units, error
0: speech, word, hmm, recognition, mlp

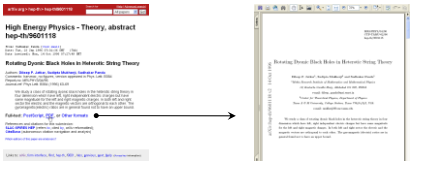
[Benyah Shaparenko]

Finding Related Articles

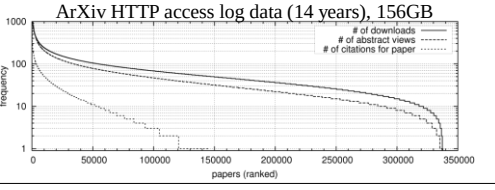
- **Problem: Impose structure on set of documents**
  - Approach 1: Content similarity
  - Approach 2: Usage data
- **Experiment: Find related papers on ArXiv**
  - Use http-logfiles



Usage Data on ArXiv

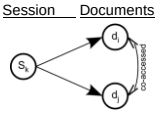


ArXiv HTTP access log data (14 years), 156GB



Co-Access as a Measure of Relatedness

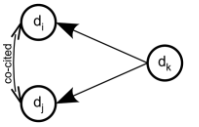
- **Co-Access**



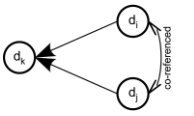
- **Session: "Uninterrupted sequence of accesses of same user"**
  - More detailed measures possible
    - Further restrictions on sessions to use
    - Ordering of accesses (abstract view ➔ download)
    - Usage of time intervals (between accesses, …)
    - …

Citation-based Measures of Relatedness

- **Co-citation:**



- **Co-reference (Bibliographic Coupling)**

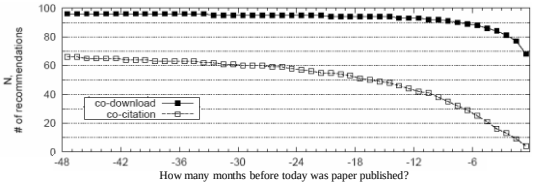


8

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.