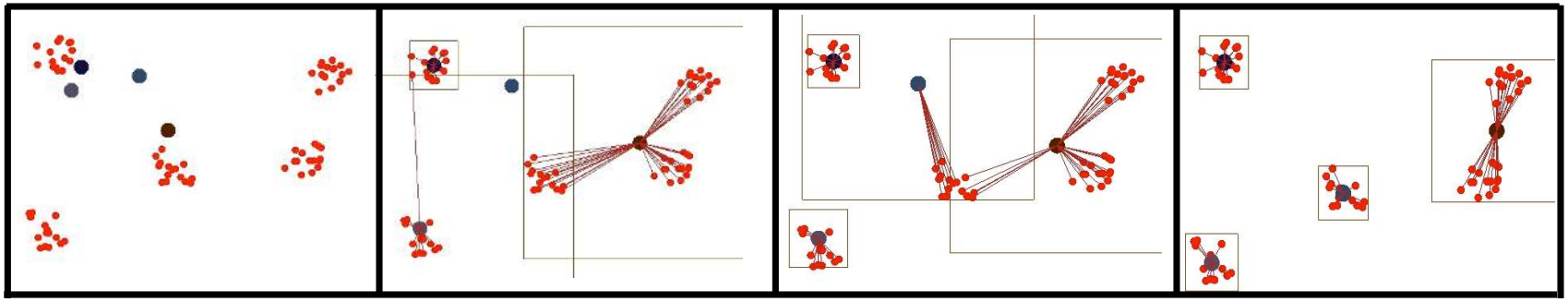
Biomedical Topic Detection and Tracking

jerry ye | jih-yin chen | johnson nguyen

Fall 2006

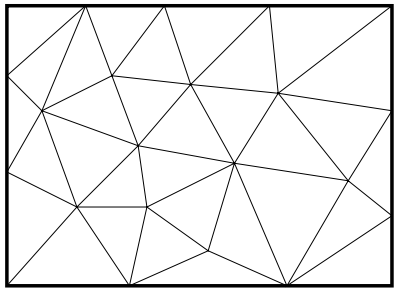


Motivation

- track RSS feeds of biomedical research papers from BioMed Central
- detect new research topics as they appear
- track and follow a topic
- email users when new paper belongs to a tracked topic

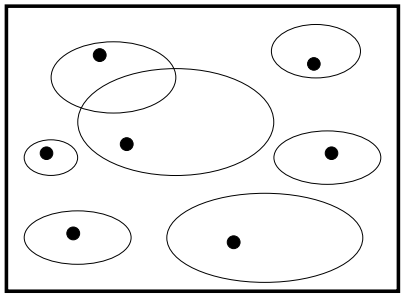
Topic Detection

- term vectors of title and abstract words
- papers in the same cluster considered similar topic
- compare new papers to existing topics (clusters) and classify as new topic if similarity below certain threshold



Topic Tracking

- classify new papers into currently identified topics
- given current clusters, compare new paper's term vector to centriods
- consider new paper to be on topic based on a similarity matrix



Algorithm

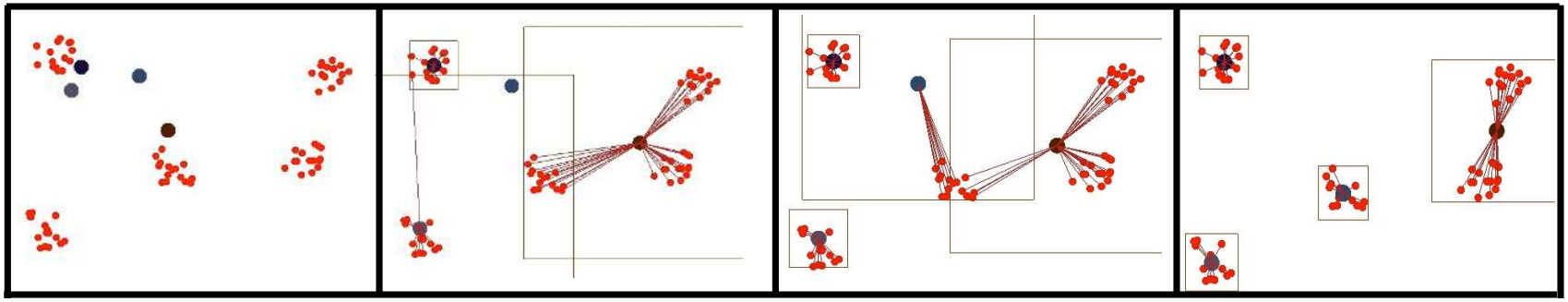
- Unsupervised clustering using K-means
- Represent training data in terms of k clusters with means u_k
- Minimize total intra-cluster variance

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2$$

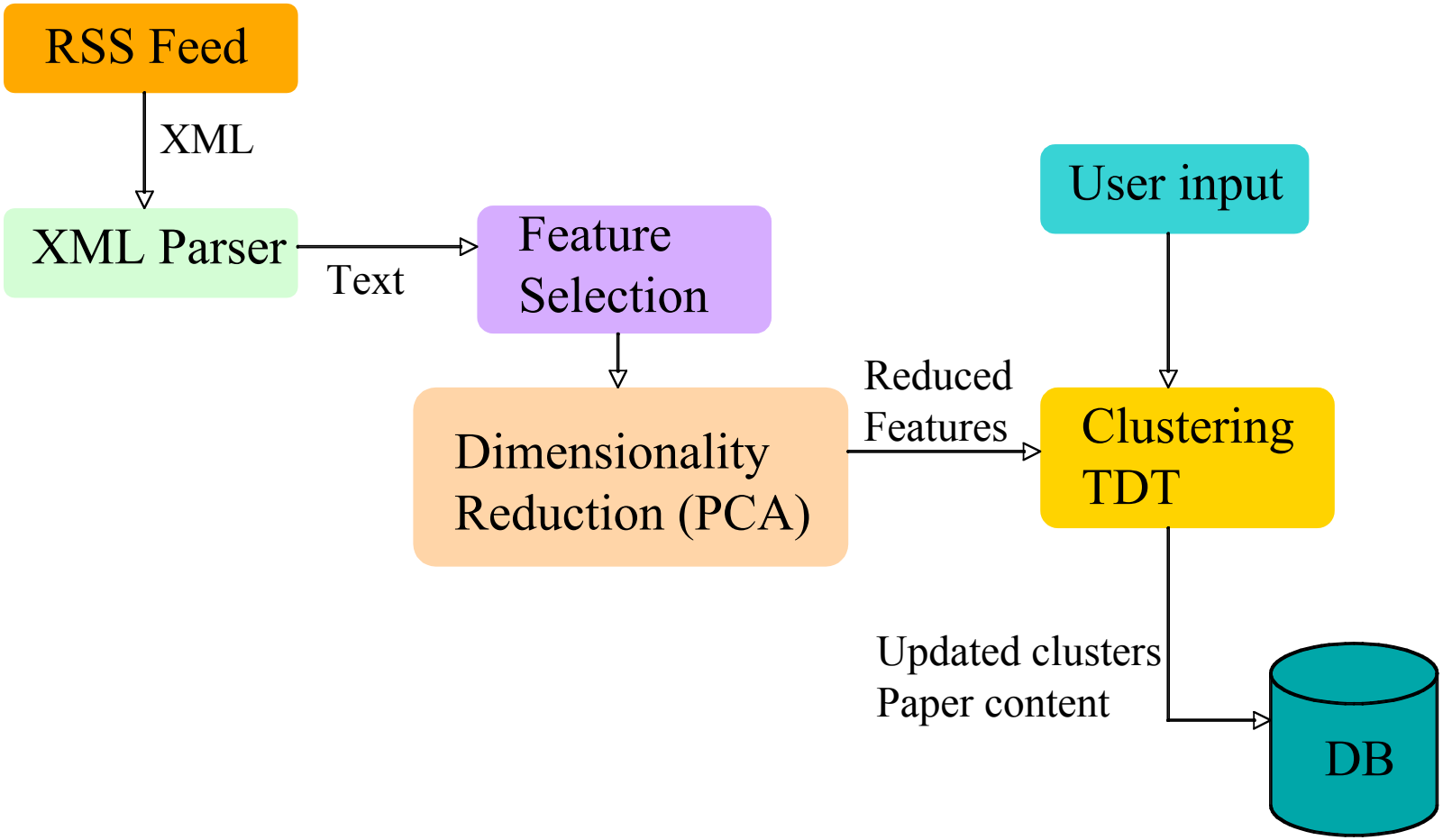
where there are k clusters $S_i, i = 1, 2, ..., k$ and u_i is the centroid or mean point of all the points $x_j \in S_i$

Algorithm

- Online tracking of topics consist of comparing distance of new papers to centriods of existing clusters
- Label as new topic if distance above threshold



System Architecture



Feature Selection

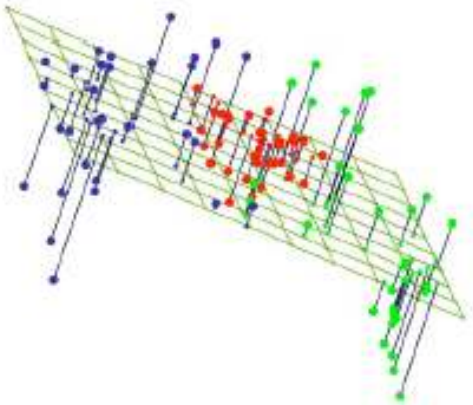
- stemming, frequency trimming, unigrams and bigrams, PCA
- title words are weighted 5:1 compared to abstract words
- term frequency x inverse document frequency

$$w(t,d) = (1 + log_2 TF_{(t,d)}) \times \frac{IDF_t}{||d||}$$

Dimensionality Reduction

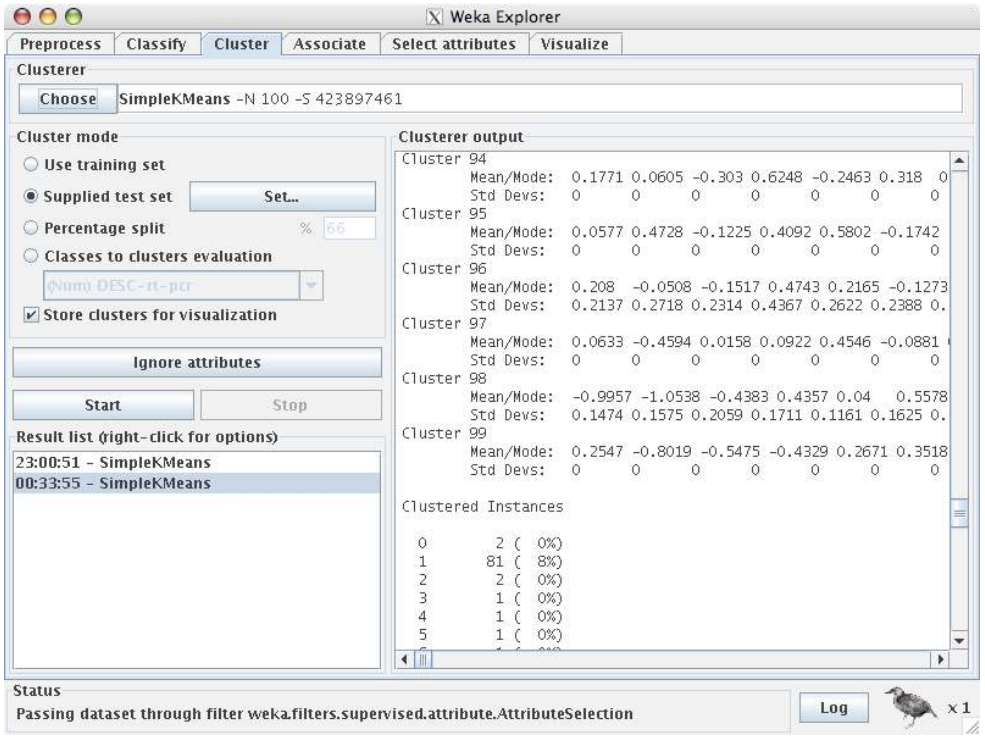
- Principle Component Analysis
  - Project higher dimension data to lower dimensional space
  - Maximize variance of projected data
  - Select eigenvectors with greatest eigenvalues

$$\mathbf{x} \in \mathbb{R}^{361}$$
$$\mathbf{z} = \mathbf{U}^T \mathbf{x}$$
$$\mathbf{z} \in \mathbb{R}^{10}$$



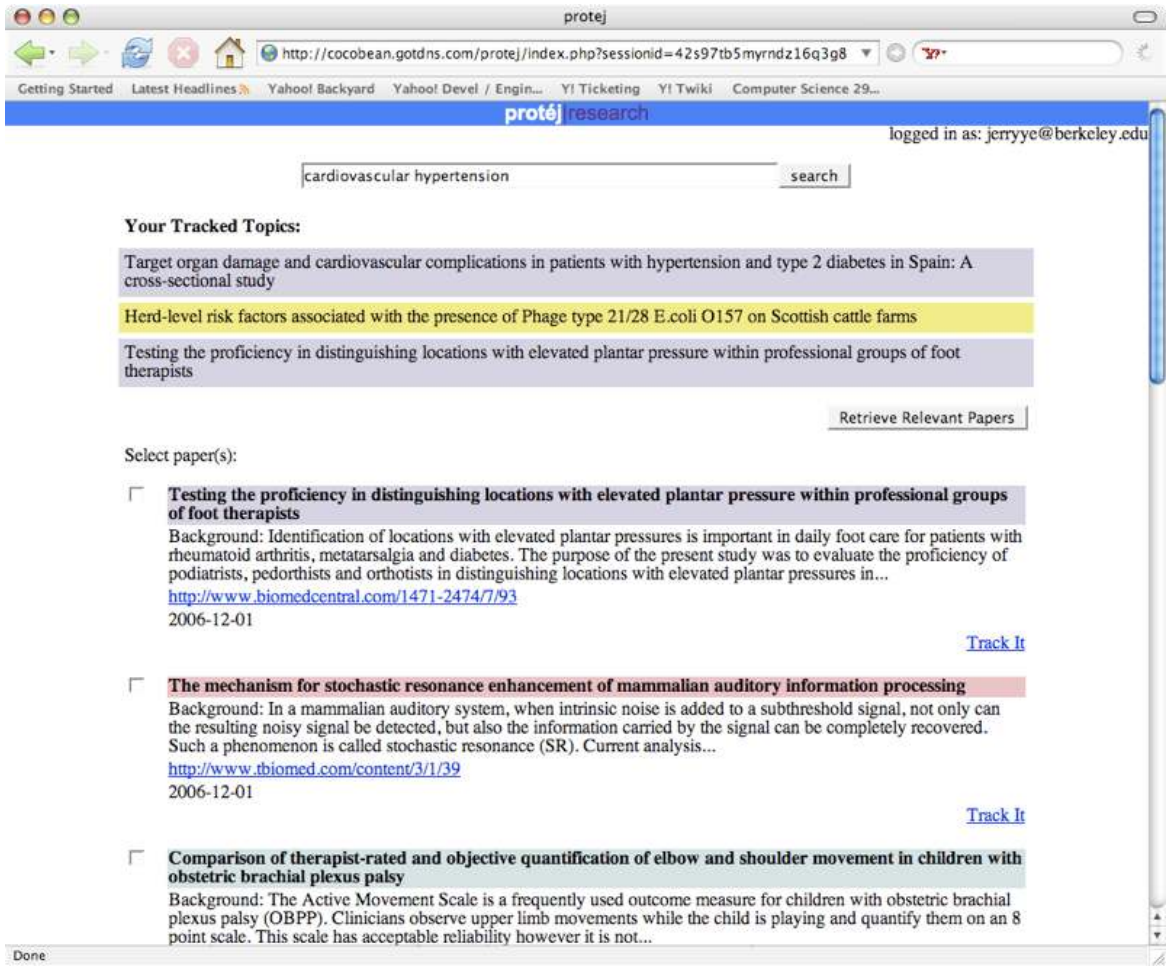
WEKA

- weka for clustering of initial research papers
- save centriods from clustering for topic tracking



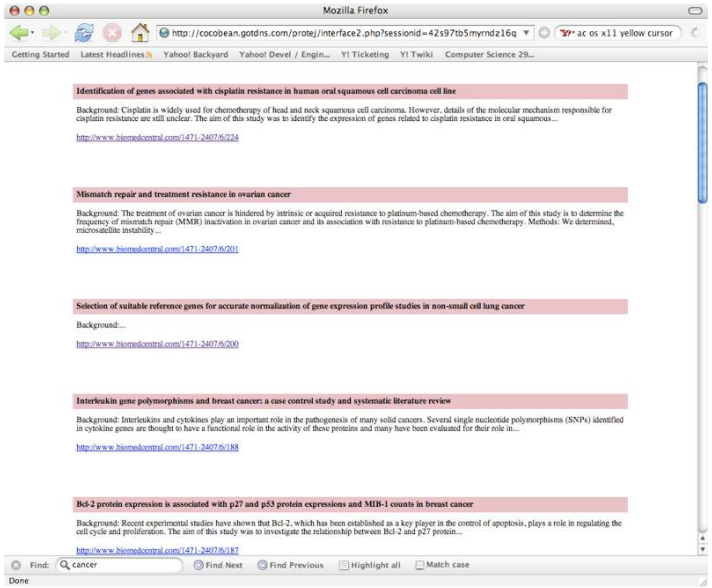
Demo

- http://cocobean.gotdns.com/protej



Results

- Results for clusters usually share a common topic
- In the example, papers in the cluster are about cancers afflicting females



Bcl-2 protein expression is associated with p27 and p53 protein expressions and MIB-1 counts in breast cancer
Background: Recent experimental studies have shown that Bcl-2, which has been established as a key player in the control of apoptosis, plays a role in regulating the cell cycle and proliferation. The aim of this study was to investigate the relationship between Bcl-2 and p27 protein...
http://www.biomedcentral.com/1471-2407/6/187

Mismatch repair and treatment resistance in ovarian cancer
Background: The treatment of ovarian cancer is hindered by intrinsic or acquired resistance to platinum-based chemotherapy. The aim of this study is to determine the frequency of mismatch repair (MMR) inactivation in ovarian cancer and its association with resistance to platinum-based chemotherapy. Methods: We determined, microsatellite instability...
http://www.biomedcentral.com/1471-2407/6/201

Conclusions

- TDT relies heavily on clustering
- Choice of number of initial clusters arbitrary
- Since data is so sparse, more features should improve results
- K-means is efficient and worked well, but other learners might do better

References

- J. Allan, R. Papka, and V. Lavrenko. On-line New Event Detection and Tracking. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

- Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. A., X. Liu. Learning Approaches for Detecting and Tracking News Events. IEEE Intelligent Systems, Volume 14 Issue 4, 1999.

- M. Franz, T. Ward, J. McCarley, W. Zhu. Unsupervised and supervised clustering for topic tracking. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.

- Y. Yang, T. Ault, T. Pierce, C. W. Lattimer. Improving text categorization methods for event tracking. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000.

- J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang. Topic Detection and Tracking Pilot Study: Final Report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.