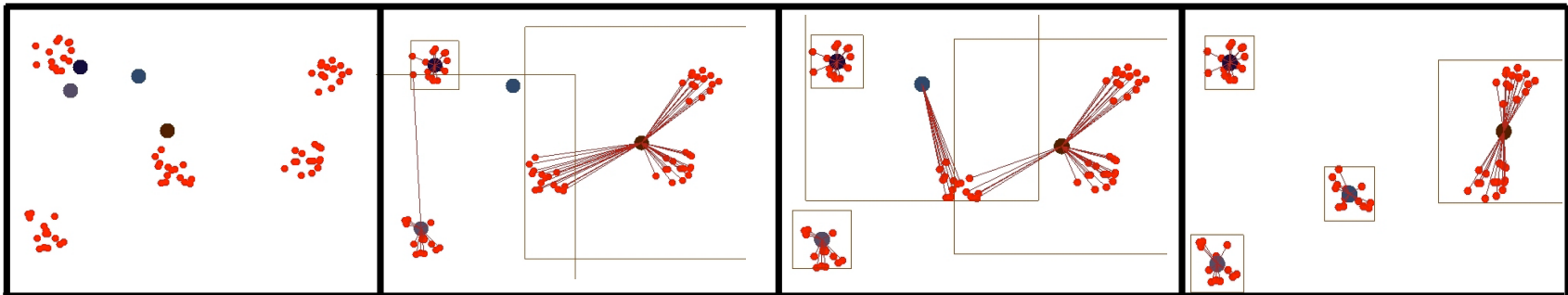


Biomedical Topic Detection and Tracking

jerry ye | jih-yin chen | johnson nguyen
Fall 2006

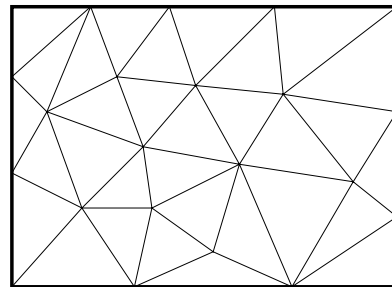


Motivation

- 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

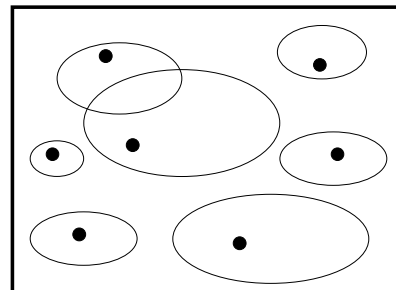
Topic Detection

- 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

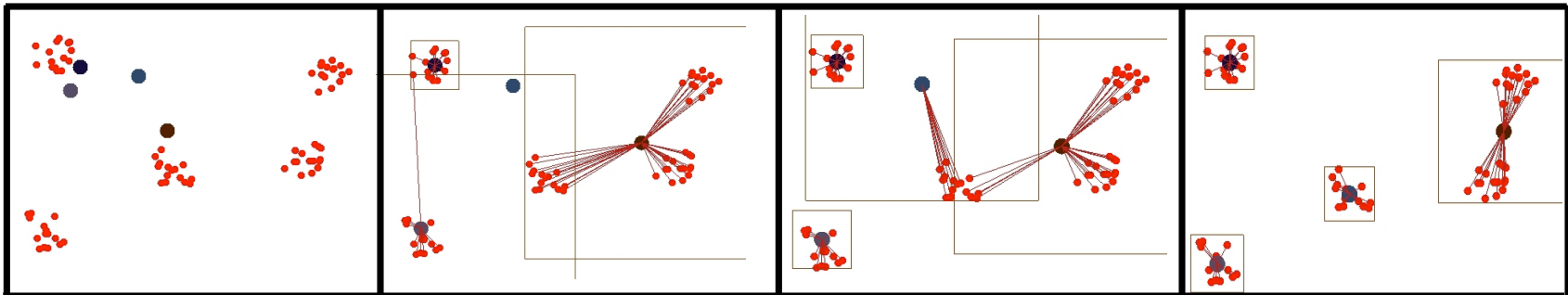
- Unsupervised clustering using K-means
- Represent training data in terms of k clusters with means u_k
- Minimize total intra-cluster variance

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

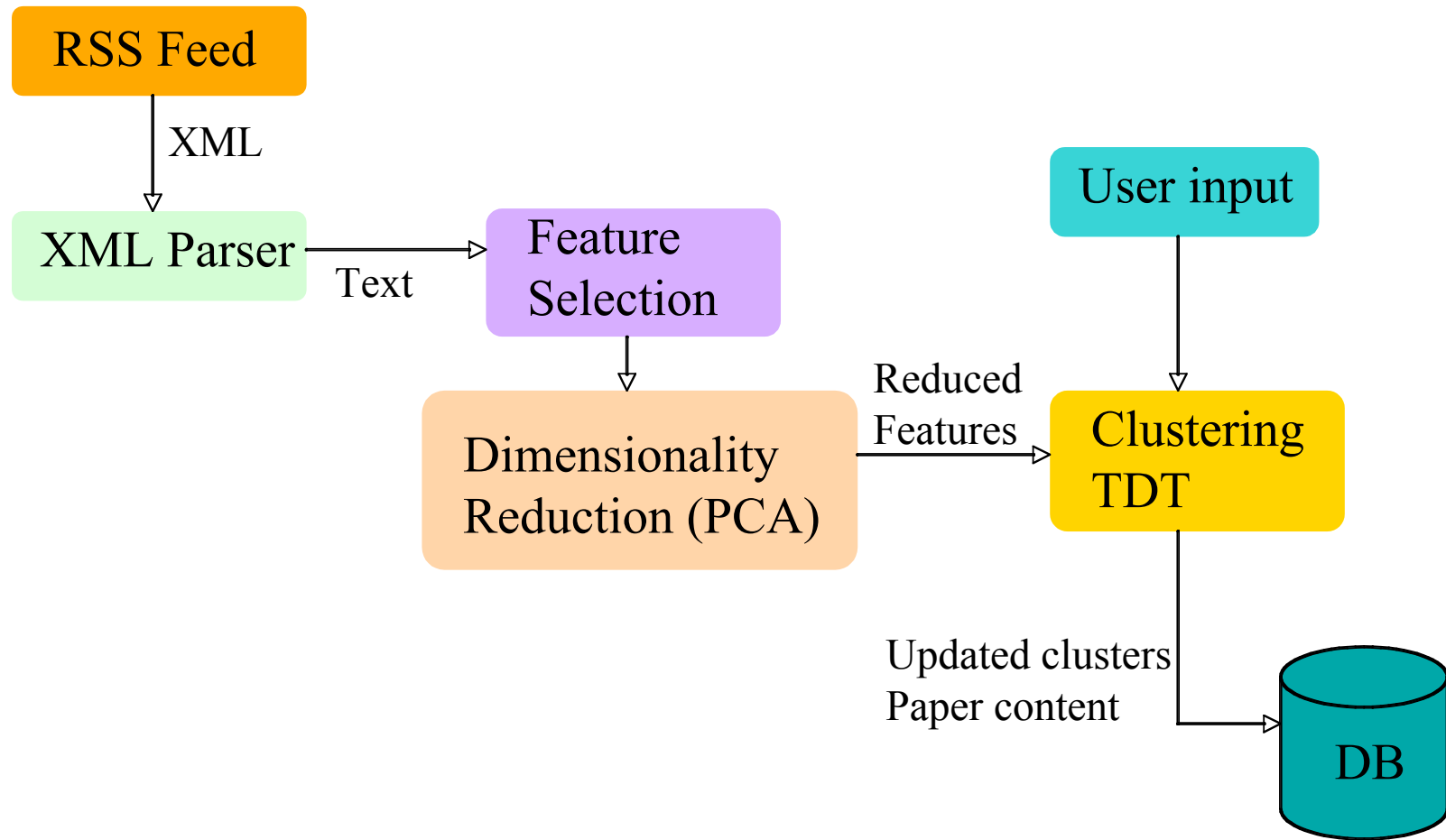
where there are k clusters $S_i, i = 1, 2, \dots, k$
and u_i is the centroid or mean point of all the
points $x_j \in S_i$

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

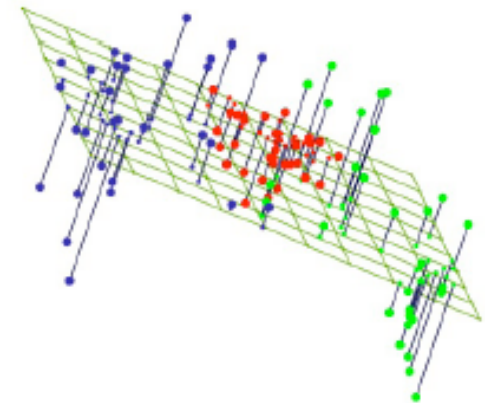
- stemming, frequency trimming, unigrams and bigrams, PCA
- title words are weighted 5:1 compared to abstract words
- term frequency x inverse document frequency

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

Dimensionality Reduction

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

$$\begin{array}{l} \mathbf{x} \in \mathbb{R}^{361} \\ \downarrow \mathbf{z} = \mathbf{U}^T \mathbf{x} \\ \mathbf{z} \in \mathbb{R}^{10} \end{array}$$



WEKA

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

The screenshot shows the Weka Explorer application window. The 'Clusterer' tab is active, displaying the 'SimpleKMeans' algorithm with parameters '-N 100 -S 423897461'. The 'Cluster mode' section has 'Supplied test set' selected. The 'Clusterer output' pane shows the following data:

Cluster	Mean/Mode	Std Devs
Cluster 94	0.1771 0.0605 -0.303 0.6248 -0.2463 0.318 0	0 0 0 0 0 0 0
Cluster 95	0.0577 0.4728 -0.1225 0.4092 0.5802 -0.1742	0 0 0 0 0 0 0
Cluster 96	0.208 -0.0508 -0.1517 0.4743 0.2165 -0.1273	0.2137 0.2718 0.2314 0.4367 0.2622 0.2388 0
Cluster 97	0.0633 -0.4594 0.0158 0.0922 0.4546 -0.0881	0 0 0 0 0 0 0
Cluster 98	-0.9957 -1.0538 -0.4383 0.4357 0.04 0.5578	0.1474 0.1575 0.2059 0.1711 0.1161 0.1625 0
Cluster 99	0.2547 -0.8019 -0.5475 -0.4329 0.2671 0.3518	0 0 0 0 0 0 0

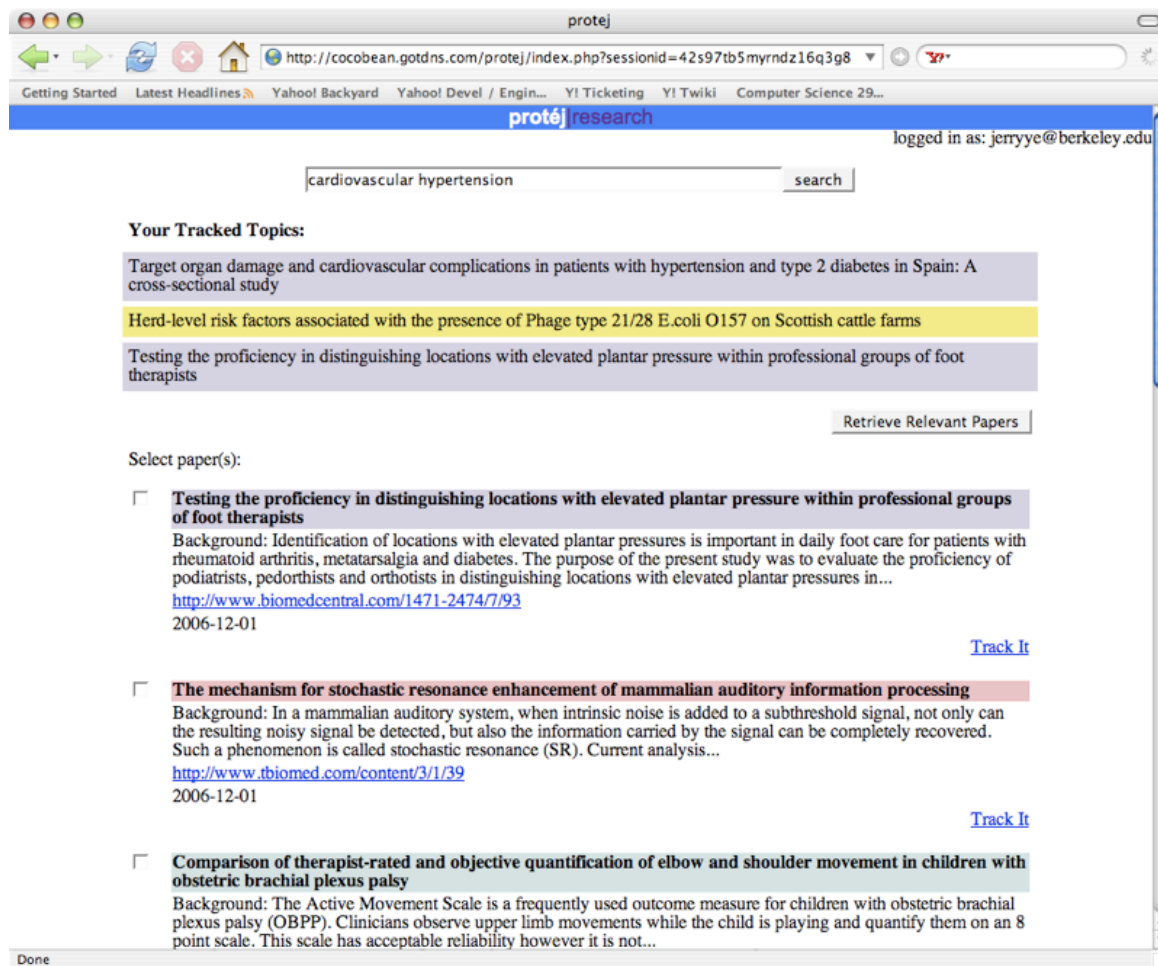
The 'Clustered Instances' section shows the following distribution:

Cluster	Count	Percentage
0	2	(0%)
1	81	(8%)
2	2	(0%)
3	1	(0%)
4	1	(0%)
5	1	(0%)

The status bar at the bottom indicates 'Passing dataset through filter weka.filters.supervised.attribute.AttributeSelection' and includes a 'Log' button and a window icon.

Demo

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



The screenshot shows a web browser window titled 'protej' with the URL 'http://cocobean.gotdns.com/protej/index.php?sessionId=42s97tb5myrndz16q3g8'. The browser's address bar and tabs are visible. The website header includes the 'protej|research' logo and a user login status 'logged in as: jerrye@berkeley.edu'. A search bar contains the text 'cardiovascular hypertension' and a 'search' button. Below the search bar, the section 'Your Tracked Topics:' lists three topics: 'Target organ damage and cardiovascular complications in patients with hypertension and type 2 diabetes in Spain: A cross-sectional study', 'Herd-level risk factors associated with the presence of Phage type 21/28 E.coli O157 on Scottish cattle farms', and 'Testing the proficiency in distinguishing locations with elevated plantar pressure within professional groups of foot therapists'. A 'Retrieve Relevant Papers' button is located below the topics. The 'Select paper(s):' section lists three papers with checkboxes, each followed by a title, background information, a URL, and a date. The first paper is 'Testing the proficiency in distinguishing locations with elevated plantar pressure within professional groups of foot therapists', the second is 'The mechanism for stochastic resonance enhancement of mammalian auditory information processing', and the third is 'Comparison of therapist-rated and objective quantification of elbow and shoulder movement in children with obstetric brachial plexus palsy'. Each paper has a 'Track It' link.

protej

http://cocobean.gotdns.com/protej/index.php?sessionId=42s97tb5myrndz16q3g8

Getting Started Latest Headlines Yahoo! Backyard Yahoo! Devel / Engin... Y! Ticketing Y! Twiki Computer Science 29...

protej|research logged in as: jerrye@berkeley.edu

cardiovascular hypertension search

Your Tracked Topics:

Target organ damage and cardiovascular complications in patients with hypertension and type 2 diabetes in Spain: A cross-sectional study

Herd-level risk factors associated with the presence of Phage type 21/28 E.coli O157 on Scottish cattle farms

Testing the proficiency in distinguishing locations with elevated plantar pressure within professional groups of foot therapists

Retrieve Relevant Papers

Select paper(s):

Testing the proficiency in distinguishing locations with elevated plantar pressure within professional groups of foot therapists
Background: Identification of locations with elevated plantar pressures is important in daily foot care for patients with rheumatoid arthritis, metatarsalgia and diabetes. The purpose of the present study was to evaluate the proficiency of podiatrists, pedorthists and orthotists in distinguishing locations with elevated plantar pressures in...
<http://www.biomedcentral.com/1471-2474/7/93>
2006-12-01 [Track It](#)

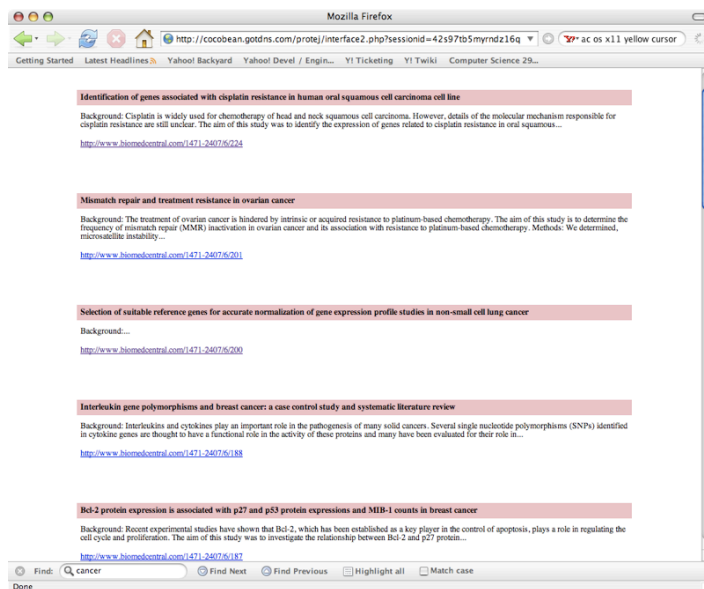
The mechanism for stochastic resonance enhancement of mammalian auditory information processing
Background: In a mammalian auditory system, when intrinsic noise is added to a subthreshold signal, not only can the resulting noisy signal be detected, but also the information carried by the signal can be completely recovered. Such a phenomenon is called stochastic resonance (SR). Current analysis...
<http://www.tbiomed.com/content/3/1/39>
2006-12-01 [Track It](#)

Comparison of therapist-rated and objective quantification of elbow and shoulder movement in children with obstetric brachial plexus palsy
Background: The Active Movement Scale is a frequently used outcome measure for children with obstetric brachial plexus palsy (OBPP). Clinicians observe upper limb movements while the child is playing and quantify them on an 8 point scale. This scale has acceptable reliability however it is not...

Done

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.