Sequencer

Distributed Temporal Tracking of Named Entities

Brett Walenz

University of Nebraska at Omaha
Concept

• A scalable and temporal tracking system for topic detection and named entity recognition
• Correlate topics across news articles and identify the key entities across the articles (people, places, organizations, etc)
• Research methods for monitoring evolving storylines
• Research visualization techniques for events that change through time
Research Status
October, 2009

• Evaluation sets created – CNN.com
• Apache Lucene and Nutch interfaces created
• Hierarchical clustering, first pass topic detection
Crawler

• Nutch – distributed crawler using Hadoop, creates an *inverted index* using Lucene.

• Lucene index structure:
  • docid
  • url
  • timestamp
  • parsedContent
  • originalContent
  • termVectors
Crawler

• Using CNN.com as the source of documents
• Three data sets created:
  – 1 week, depth 6 crawl (12k docs per index)
  – 1 month, depth 6 crawl
  – 4 days, depth 3 crawl (300 docs per index)
Clustering

• Find which documents have many words in common, and place them into an initial set of bins
• Hierarchical clustering used – both complete-link and single-link approaches
• Cutoff distance can be used to partition data
• First pass; second pass is named entity similarities in articles
Results

• Small Lucene index – depth 3 crawl, ~380 documents

• Considerations – high frequency words, partition cutoff, low frequency words

• Ignore documents under 100 characters – CNN has high frequency of ‘junk’ pages.
Results

Beaten teen had 'different attitude' than other boys, activist says - CNN.com
 Relatives, Chicago officials attend beaten teen's funeral - CNN.com
 Chicago police seek 3 more in teen's death CNN.com
 Family of beaten teen hope for healing after death - CNN.com

Obama's high-stakes Olympic gamble - CNN.com
Obama, Chicago come up short in Olympics bid - CNN.com
Rio de Janeiro to host 2016 Olympics - CNN.com
Olympic 'nopes' beat out hope in Chicago - CNN.com
Brazilians celebrate Olympics bid in Rio de Janeiro - CNN.com

U.S. envoy: Timeline for troops leaving Iraq on track - CNN.com
Odierno: May not be possible to declare victory in Iraq - CNN.com
4,000 U.S. troops expected to leave Iraq in October - CNN.com
Results

U.S. urges Karzai to allow vote fraud probe - CNN.com
Afghanistan commission orders partial vote recount - CNN.com
Hundreds of Afghan votes declared invalid - CNN.com
Karzai maintains lead in Afghan election - CNN.com
Karzai warns West over Afghan vote fraud claims - CNN.com

Father in Japan: 'I didn't do anything wrong' - CNN.com
Group calls for release of American dad jailed in Japan - CNN.com
Dad in Japan custody battle thought wife would take kids - CNN.com
American jailed in Japan for trying to reclaim his children - CNN.com
Father, kids in custody case Japanese citizens, officials say - CNN.com
Thoughts

• Slow process – $O(N^2 \log n)$
• Memory intensive – for the full crawl, gives an $N \times M$ matrix of 3600 documents and 53,000 terms.
• First step pass – named entities can be used to identify similar stories within the cluster.
• Next step – detect difference between clusters in sequential CNN crawls