(i)DiversiNews:
a stream-based, on-line service for diversified news

Mitja Trampuš, Flavio Fuart, Jan Berčič, Delia Rusu,
Luka Stopar, Tadej Štajner
AI Lab, Jozef Stefan Institute, Slovenia

SiKDD 2013
Ljubljana, October 2013

Outline

- Motivation
- Demo
- Technical

Browsing News – mainstream sites

- How do you get the full picture of an event?
- http://news.google.com/

- 1000+ articles per story
- Sorted by popularity → echo chamber effect

Browsing News – iDiversiNews

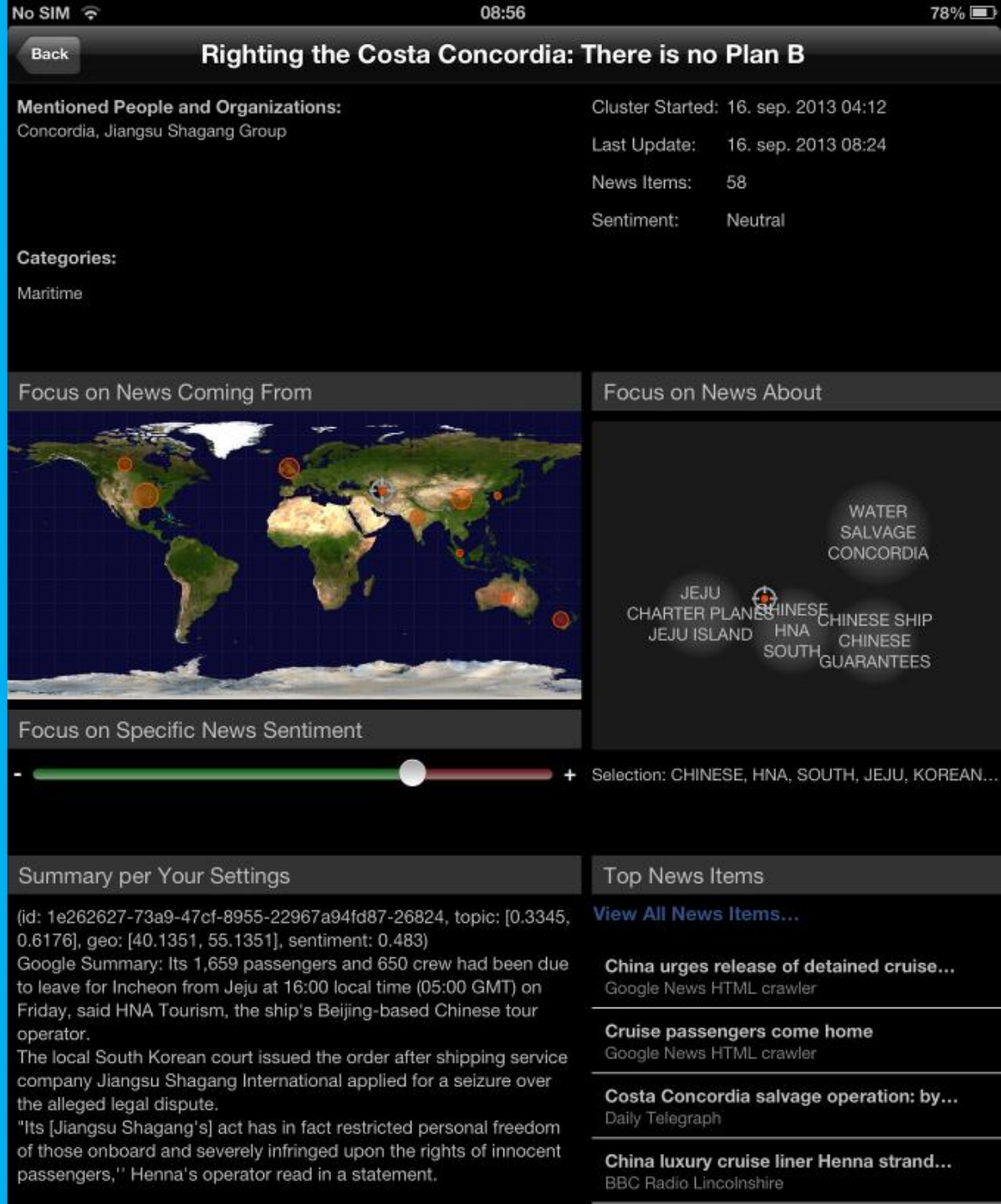
- Like Google News with **reordering, filtering**
- How do different people see a story? Re-rank on
  - Sentiment – get (un)favorable articles first!
  - Geography – where are the articles from?
  - Topic – what story aspects do articles emphasize?
- Adaptive summary
- Application for web, iOS

Demo

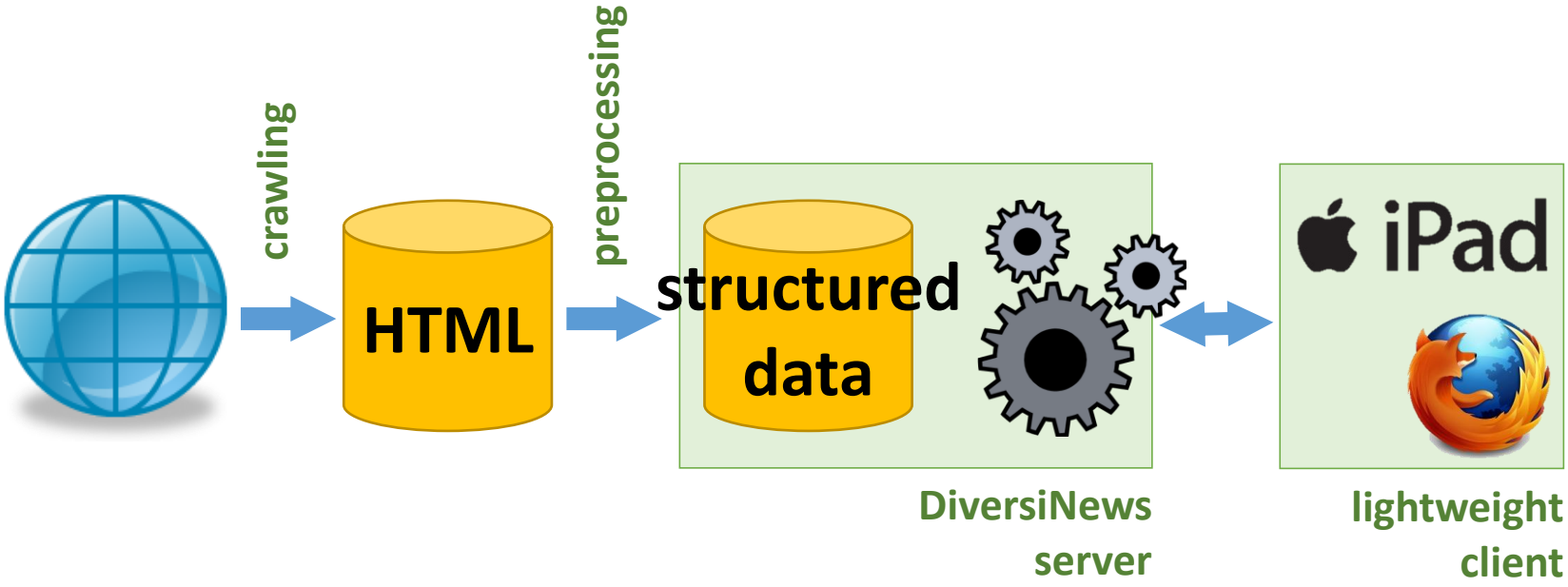
- Web: http://aidemo.ijs.si/diversinews/
- iOS: coming to App Store

Demo

- Web: http://aide
- iOS: coming to A



Tech


crawling
HTML
preprocessing
structured data
DiversiNews server
iPad
lightweight client

Tech: Raw Data

- http://newsfeed.ijs.si/
- RSS feeds of news, blogs
  - 50 000 – 100 000 English articles per day
- Discovery latency: median <3 hours

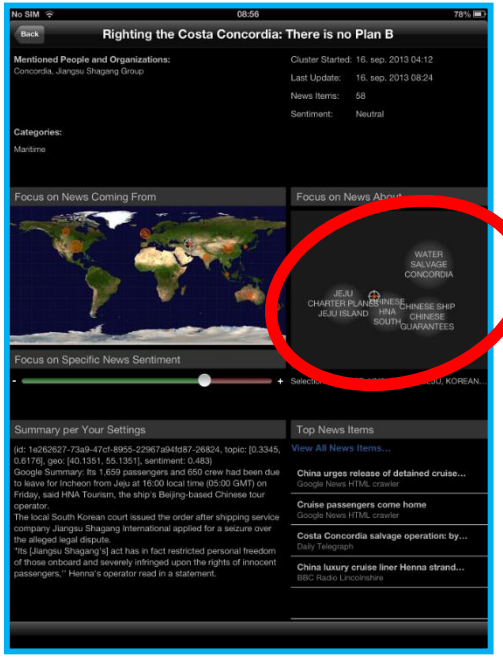
Trampus, Mitja and Novak, Blaz: *The Internals Of An Aggregated Web News Feed*. Proceedings of 15th Multiconference on Information Society (SiKDD 2012). [PDF]

Tech: Data Preprocessing

- HTML to cleartext:
  - DOM-based heuristics
- Clustering into stories:
  - Centroids of clusters of last 100 000 articles
  - Assign new article to nearest centroid
  - Split/merge clusters periodically (← information criterion)

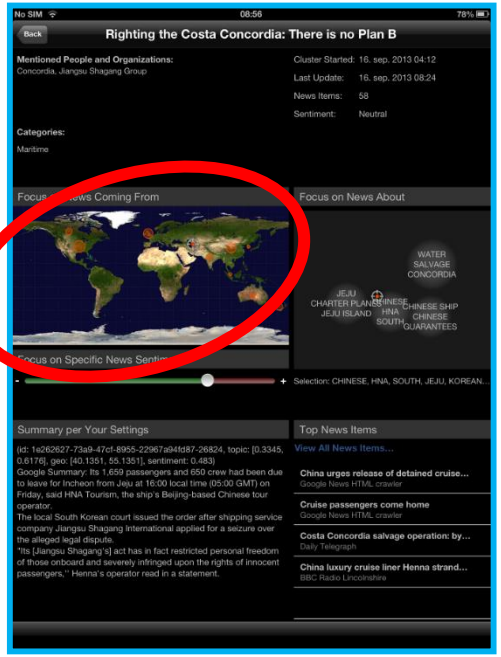
Tech: Topic Map

- BoW clustered articles
- Cluster keywords = top features
- 2D: MultiDimensional Scaling (MDS)



Tech: Article Source Geo

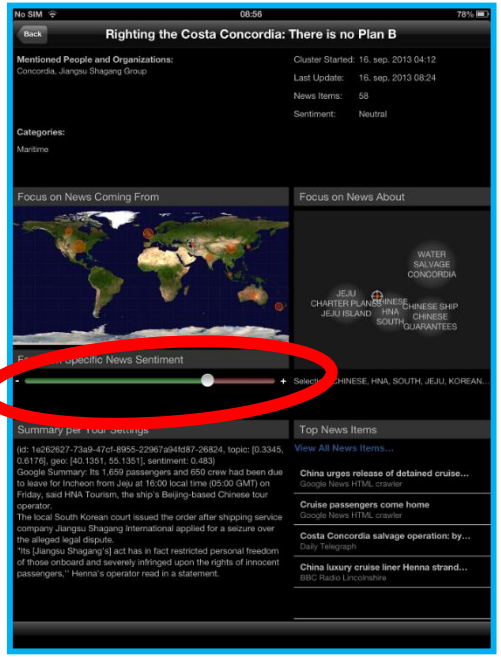
- Web listings of publishers with metadata
- RSS metadata
- Publisher homepage ccTLD
  - ccTLD – country-code top-level domain
- Publisher homepage WHOIS
  - Heuristic parsing



Tech: Sentiment

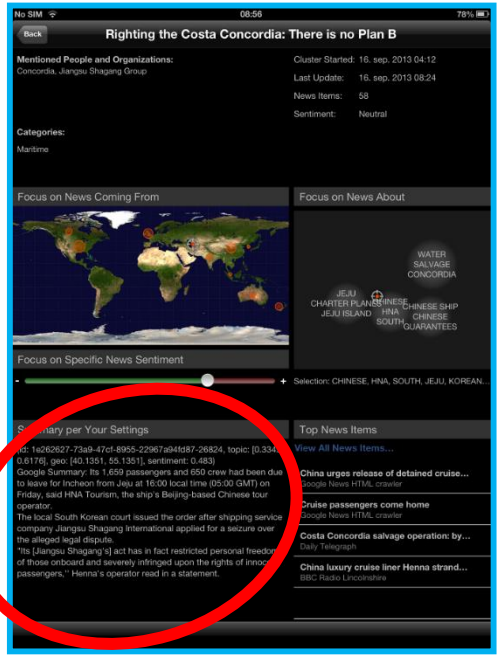
- Supervised methods
- BoW, surface and lexicon features

- Hard problem on newswire



Tech: Summarization

- Sentence-level extractive summaries
- LDA-like graphical model to obtain word distribution for the story
  - Weighted by article query-relevance
- Choose sentences that mimic this distribution
  - = minimize KL divergence between word distributions of input and summary



Delort, J. Y. and Alfonseca, E., DualSum: A Topic-Model Base Approach for Update Summarization, in proceedings of EACL2012

Questions?

mitja.trampus@ijs.si
jan.bercic@gmail.com

http://aidemo.ijs.si/diversinews/