Topic Detection and Tracking

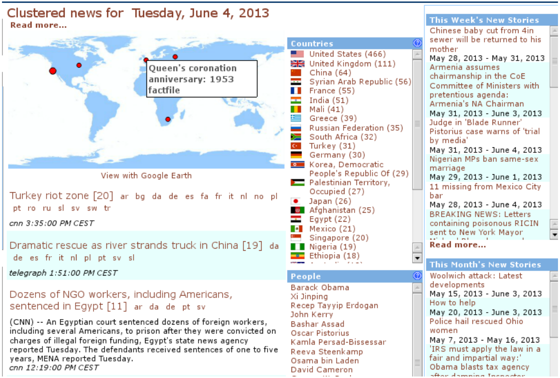
Marc-André Faucher
Jeff How
Jonathan Villemaire-Krajden

June 6, 2013

Topic Detection and Tracking

Wayne (1997) - *automatic techniques for finding topically related materials in streams of data*

- EMM NewsExplorer
- EMM NewsBrief



Terminology

Story an article or news broadcast with an underlying focus

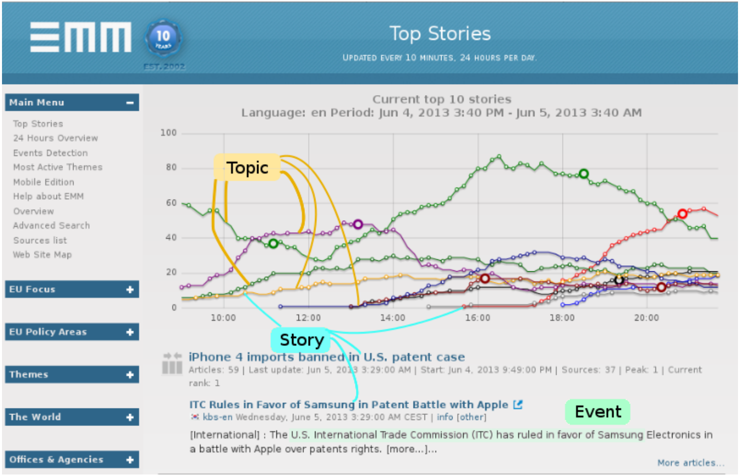
Event a unique thing that happens at some point in time

Topic a set of highly-related events

Input a continuous stream of real-time text (stories)

Output clusters organized by topic with a leading story

Terminology



Subtasks

- Topic Detection
  - 1st story
- Topic Tracking
  - Finding additional stories about a particular topic
  - Clustering

Feature Selection

Lexical	Words, Noun Phrases, Named Entities
Syntactic	POS tagging (nouns, verbs, proper names)
Semantic	Temporal language cues (verb tense and temporal NPs)(Makkonen and Ahonen-Myka, 2003)
Metadata	Timestamps

Similarity Measures and Model Selection

Similarity measures such as TF-IDF and language models such as smoothed Bayes are used to compare documents.

Vector space	TF-IDF (Carbonell et al., 1999)
Probabilistic	Chi-square(Swan and Allan, 1999)
Graph-based	Edge weights from word cooccurrences (Saha and Sindhwani, 2012)
Language Models	Probability of a story given a topic (Allan et al., 2000)

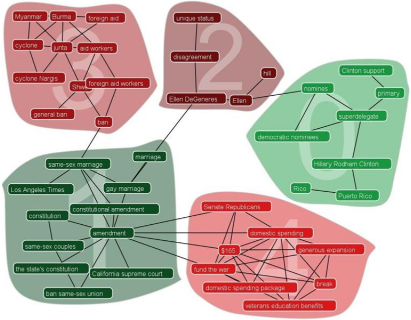
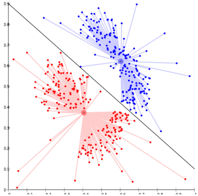


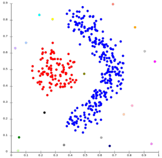
Figure: source: http://www.cs.umd.edu/~sayyadi/keygraph.html

Cluster

Stories are grouped through clustering using distance metrics based on the model (probabilistic, cosine similarity, etc.):

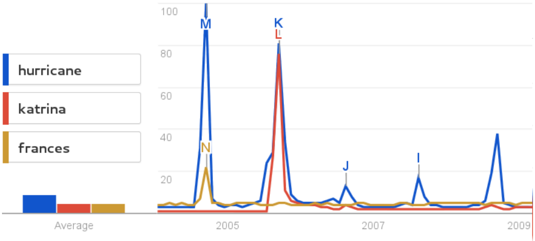


- **K-means:**
  - Selecting small k with small variance
  - Centroid represents dominant features
- **Hierarchical agglomerative clustering:**
  - Provides a hierarchy of clusters
- **KeyGraph link based clustering:**
  - Network of features and relations
  - Topics are identified using network theory
- **Advanced Techniques:**
  - Latent Dirichlet Allocation (LDA)
  - Non-negative Matrix Factorization (NMF) (Sayyadi et al., 2009)



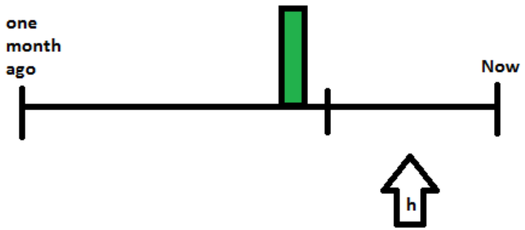
Time-Span (Swan and Allan, 1999)

- Entities & noun phrases with overlapping time-spans constitute a topic.
- Calculate probabilities of time-span overlap for features if independence is assumed.
- Merge features, which are statistically dependent (χ^2) in terms of doc occurrence.



Topic Tracking in Tweet Streams (Lin et al., 2011)

- High arrival rate (up to 4000+ tweets per second)
- Foreground model: tracks recent topic counts
  - History of *h* events
  - Smoothed with the background model
- Background model: long-term estimates of term distributions
  - Handles sparsity from limited history of foreground model
- Evaluation based off hashtags



Conclusion

- Topic Detection
- Topic Tracking
- Implementation: Feature $\rightarrow$ Model $\rightarrow$ Cluster
- Current Application:
  - EMM
  - Google News

References

J. Allan, V. Lavrenko, D. Malin, and R. Swan. Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of Topic Detection and Tracking Workshop (TDT-3)*, pages 167–174. Vienna, VA, 2000. URL http://maroo.cs.umass.edu/pdf/IR-201.pdf.

J. Carbonell, Y. Yang, J. Lafferty, R. D. Brown, T. Pierce, and X. Liu. Cmu report on tdt-2: Segmentation, detection and tracking. In *Proceedings of the DARPA Broadcast News Workshop*, pages 117–120, 1999. URL http://www.cs.cmu.edu/afs/.cs.cmu.edu/Web/People/jgc/publication/CMU_Approach_TDT_Segmentation_DARPA_1999.pdf.

J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 422–429, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020476. URL http://doi.acm.org/10.1145/2020408.2020476.

J. Makkonen and H. Ahonen-Myka. Utilizing temporal information in topic detection and tracking. In T. Koch and I. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, volume 2769 of *Lecture Notes in Computer Science*, pages 393–404. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-40726-3. doi: 10.1007/978-3-540-45175-4_36. URL http://dx.doi.org/10.1007/978-3-540-45175-4_36.

A. Saha and V. Sindhwani. Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization. In *Proceedings of the fifth*