

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

Clustered news for Tuesday, June 4, 2013

Read more...



View with Google Earth

Turkey riot zone [20] ar bg da de es fa fr it nl no pl pt ro ru sl sv sw tr

cnn 3:35:00 PM CEST

Dramatic rescue as river strands truck in China [19] da de es fr it nl pl pt sv sl

telegraph 1:51:00 PM CEST

Dozens of NGO workers, including Americans, sentenced in Egypt [11] ar da de pt sv
(CNN) -- An Egyptian court sentenced dozens of foreign workers, including several Americans, to prison after they were convicted on charges of illegal foreign funding, Egypt's state news agency reported Tuesday. The defendants received sentences of one to five years, MENA reported Tuesday.
cnn 12:19:00 PM CEST

Countries

	United States (466)
	United Kingdom (111)
	China (64)
	Syrian Arab Republic (56)
	France (55)
	India (51)
	Mali (41)
	Greece (39)
	Russian Federation (35)
	South Africa (32)
	Turkey (31)
	Germany (30)
	Korea, Democratic
	People's Republic Of (29)
	Palestinian Territory, Occupied (27)
	Japan (26)
	Afghanistan (25)
	Egypt (22)
	Mexico (21)
	Singapore (20)
	Nigeria (19)
	Ethiopia (18)

People

	Barack Obama
	Xi Jinping
	Recep Tayyip Erdogan
	John Kerry
	Bashar Assad
	Oscar Pistorius
	Kamla Persad-Bissessar
	Reeva Steenkamp
	Osama bin Laden
	David Cameron

This Week's New Stories

Chinese baby cut from 4in sewer will be returned to his mother
May 28, 2013 - May 31, 2013
Armenia assumes chairmanship in the CoE Committee of Ministers with pretentious agenda: Armenia's NA Chairman
May 31, 2013 - June 3, 2013
Judge in 'Blade Runner' Pistorius case warns of 'trial by media'
May 31, 2013 - June 4, 2013
Nigerian MPs ban same-sex marriage
May 29, 2013 - June 1, 2013
11 missing from Mexico City bar
May 28, 2013 - June 4, 2013
BREAKING NEWS: Letters containing poisonous RICIN sent to New York Mayor
Read more...

This Month's New Stories

Woolwich attack: Latest developments
May 15, 2013 - June 3, 2013
How to help
May 20, 2013 - June 3, 2013
Police hail rescued Ohio women
May 7, 2013 - May 16, 2013
'IRS must apply the law in a fair and impartial way': Obama blasts tax agency
Obama blasts tax agency

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)

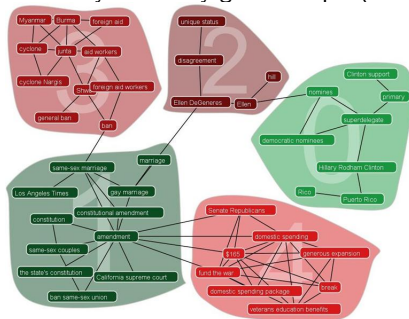
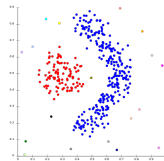
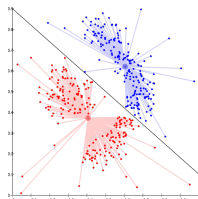


Figure: source: <http://www.cs.umd.edu/~sayyadi/keygraph.html>

Cluster

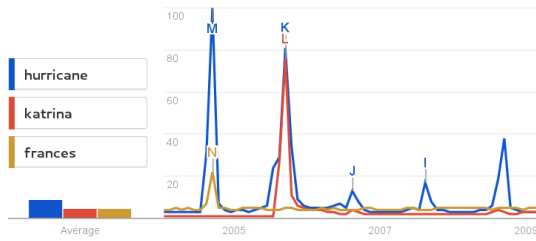
Stories are grouped through clustering using distance metrics based on the model (probabilistic, cosine similarity, etc.):

- **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)



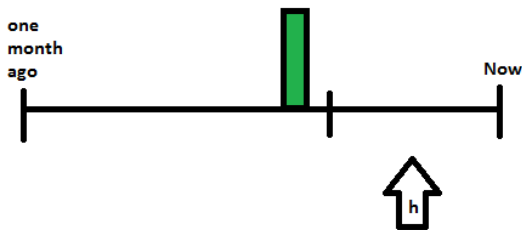
Time-Span (Swan and Allan, 1999)

- 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. Sjølvberg, 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. Sindhvani. Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization. In *Proceedings of the fifth*