Transient News Crowds in Social Media

Janette Lehmann[1], Carlos Castillo[2], Mounia Lalmas[3], Ethan Zuckerman[4]

1: Universitat Pompeu Fabra, Spain; 2: Qatar Computing Research Institute, Qatar;
3: Yahoo! Labs, Spain; 4: MIT Center for Civic Media, USA

Digital Journalism
News recommender
Crowds
Social media / Twitter

http://grupoweb.upf.es/

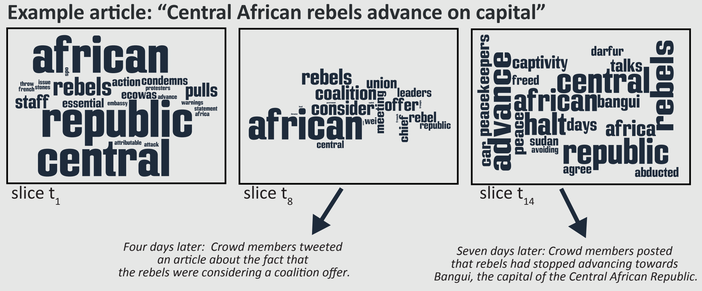
We introduce a recommender system to assists online news producers to rapidly detect follow-up stories to the articles they publish.

The beauty of this approach is that there is no need to pre-define topics and the related stories becomes available immediately, allowing journalists to cover news beats incorporating the shifts of interest of their audiences.

Transient news crowds

- A **news crowd** of an article is defined as the set of users that tweeted the article within the first 6h after the first tweet on that article
- During one week, the tweets of the users on each crowd were recorded
- The week is split in **time slices** of 12h

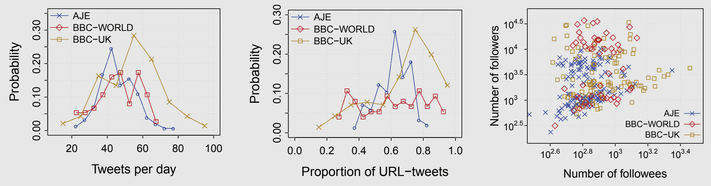
Example article: "Central African rebels advance on capital"



Data

Dataset	Articles	Users Total	Users Crowd	Tweets Total	Tweets Crowd
BBC World Service	75	13.3K	177	35.5K	201
BBC News UK	141	13.1K	92	47.8K	339
Al Jazeera English	155	8.3K	53	24.0K	154

Crowd-based news discovery



Given a news crowd and a time slice, find URLs in that time slice that are related to the article that created the crowd.

Candidate generation

- A **candidate URL** is defined as the set of equal URLs tweeted in a time slice
- Average number of candidates per time slice:
  BBC-WORLD: 41.2 (sd=23.8)
  BBC-UK: 54.8 (sd=23.8)
  AJE: 15.7 (sd=4.7)
- Crowdsourcing workers labeled the candidates as *strongly*, *weakly*, or *not related* to the original story

Features

Text

Similarity to original article tweet	Similarity to candidates in other crowds
*SimOriginalUnigrams**	*SimOtherHashTags*
*SimOriginalBigrams**	
SimOriginalHashTags	
*SimOriginalCapitalized**	

Frequency

TF and IDF
CandidateNormalizedFrequency
*CandidateInverseDocFrequency**

User

Topical focus	Activity
*WeightedSumInverseCrowds**	*UserFavorites*
WeightedSumInverseSections	*UserTweetsDaily**
Connectivity	*UserFracRetweets*
WeightedSumFollowerFollowees	*UserFracURL*
UserFollowers	*UserFracMention*
UserFollowees	*UserFracHashTag*

* - Most important features for the recommender.

Results

- Random forest classifier + asymmetric misclassification costs; 10-fold-cross validation
- R@2/3: A mandatory level of precision of two-thirds was required

Features	#	AJE AUC	AJE R@2/3	BBC-World AUC	BBC-World R@2/3	BBC-UK AUC	BBC-UK R@2/3
Frequency	2	0.65	-	0.64	-	0.54	-
Text-based	5	0.87	0.40	0.85	0.44	0.66	-
User-based	11	0.81	0.30	0.70	-	0.64	-
Frequ+Text	7	0.89	0.62	**0.88**	**0.52**	0.66	0.04
Frequ+User	13	0.79	0.32	0.72	-	0.64	0.11
Text+User	16	0.92	0.66	0.80	0.43	0.73	0.14
All	18	**0.92**	**0.72**	0.85	0.49	**0.71**	**0.14**

Conclusion and future work

- **Automatic-operated accounts**. Many Twitter accounts post tweets automatically, but are not spammers; e.g. corporate accounts (e.g. @AJELive) and third-party services (e.g. @AmmanHashtags)
- **Recurring crowd members**. Many crowd members (manually and automatic) tweet more than one article (74% from AJE, 61% from BBC-World, and 75% from BBC-UK)
- **Regathering**. Parts of some crowds come together again around new news-worthy events
- **Recommender system**. A method for uncovering related content to a news article was designed and validated

Future work

- Would an analysis of news crowds provide content producer with more information about their engaged audience?
- Can crowds influence content provider, for instance, in triggering them to create more content related to salient news?
- What are the roles of curators (e.g. opinion leaders) in news crowds?