IDENTIFYING TOPICAL AUTHORITIES IN MICROBLOGS

Aditya Pal (University of Minnesota)
Scott Counts (Microsoft Research)

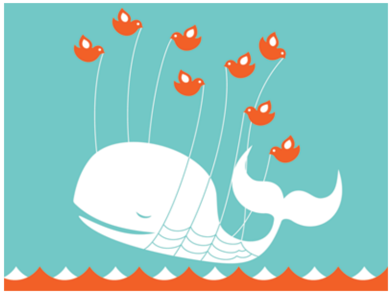
Problem Definition

- Given a topic, identify interesting and authoritative authors

- Benefits?
  - Stay updated on the topic
  - Recommendation to follow user
  - Topic summarization
  - Viral Marketing

Challenges with Microblogs

- Tens of thousands of authors posting on a topic per day
- Authors might not even exist prior to the topic (event)
  - e.g. HaitiReliefFund, WorldCupNews, iPhone4Reviews
- Avoid overly general authorities
  - e.g. CNN (oil spill)
- Avoid Celebrities.
  - e.g. Shakira (world cup)



Related Work

- Weng et al. [1] proposed TwitterRank
  - They compute topical distribution of a user (using LDA)
  - Estimate topical weights between graph neighbors
  - Use PageRank to find out the top influential

- Can we solve the problem in near real-time

[1] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM* 2010.

Presentation Outline

- Our approach
- Evaluation
- Results
- Conclusion and future work

Our Approach

- Near real-time method
- Can be implemented using Distributed framework

Feature Selection - Tweet Terminology

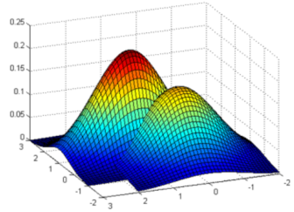
- **OT**: Original Tweet – tweet produced by the author
- **RT**: Repeated Tweet – tweet copied by the author
  - Typically contain keyword "RT"
- **CT**: Conversational Tweet – tweet directed towards another user
  - Typically start with @username

List of Features		
ORIGINAL TWEETS	OT1	Number of original tweets
	OT2	Number of links shared
	OT3	**Self-similarity** (how similar is author's recent tweet to her previous tweets)
	OT4	Number of keyword hashtags used
CONVERSATIONAL TWEETS	CT1	Number of conversational tweets
	CT2	Number of conversational tweets where conversation is initiated by the author
REPEATED TWEETS	RT1	Number of retweets of other's tweet
	RT2	Number of unique tweets (OT1) retweeted by other users
	RT3	Number of unique users who retweeted author's tweets
MENTION FEATURES	M1	Number of mentions of other users by the author
	M2	Number of unique users mentioned by the author
	M3	Number of mentions by others of the author
	M4	Number of unique users mentioning the author
GRAPH FEATURES	G1	Number of topically active followers
	G2	Number of topically active followers
	G3	Number of followers tweeting on topic after the author
	G4	Number of friends tweeting on topic before the author

List of Features

TEXTUAL FEATURES	Topical Signal (how much is user about the topic)	$(OT1 + CT1 + RT1) \ / \ \|\# \ tweets\|$
	Signal strength (how many topical tweets produced)	$OT1 \ / \ (OT1 + RT1)$
	Non-Chat signal	$OT1/(OT1 + CT1) + L * (CT1 - CT2)/(CT1+1)$
USER-TOPIC IMPACT	Retweet impact	$RT2 \cdot log(RT3)$
	Mention impact (how much the topic is about the user)	$M3 \cdot log(M4) - M1 \cdot log(M2)$
GRAPH FEATURES	Information diffusion	$\log(G3+1) - log(G4+1)$
	Network score	$\log(G1 + 1) - log(G2 + 1)$

Probabilistic Clustering



- Gaussian Mixture Model (GMM)
  - Soft clustering approach
  - Expectation Maximization (EM) is used to find local optimum

- Features
  - M – component
    - Akaike information criteria (AIC), Bayesian information criteria (BIC)
  - Initialize with Kmeans
  - Use regularization
  - Convergence based on likelihood
  - Run multiple iterations of GMM

- Pick true cluster representatives (p>0.9)
  - Pick cluster with larger *Topic signal, Retweet impact, Mention impact*

User Ranking

- List based ranking
  - Rank users on individual features
  - Take average rank
- Gaussian based ranking
  - For every feature compute the Gaussian CDF
  - Compute their product:

$$R_G(x_i) = \prod_{f=1}^{d} [\int_{-\infty}^{x_i^f} N(x; \mu_f, \sigma_f)]^{w_f}$$

Dataset

- All public tweets (~90 Million) on Twitter between 6th-June-2010 to 10th-June-2010 (5 days)

- Three topics: *iphone, oil spill, world cup*
  - Keyword extraction to find topical tweets

	Users	Original Tweets	Conversational Tweets	Retweets
iphone	430,245	658,323	242,000	129,560
oil spill	64,892	111,000	8,140	29,224
world cup	44,387	308,624	28,612	47,837

Models

- **Our:** textual + graph properties

- **Baseline**
  - **b1**: graph properties (mention impact, retweet impact, …) + page rank
  - **b2**: textual properties
  - **b3**: random selection of outside cluster users
    - Validates cluster selection criteria

- Cluster and ranking algorithm is applied to build ranked lists.

Results

- Average number of followers for top 10 users per model:

	our	*b1*	*b2*	*b3*
iphone	282,665	1,364,015	117,250	1,252
oil spill	462,507	871,159	417,210	840
world cup	29,373	32,121	18,017	277

- Key observation: b1 > our > b2 > b3

Results – top 10 users (our)

iphone	oilspill	worldcup
macworld	NWF	TheWorldGame
Gizmodo	TIME	GrantWahl
macrumorslive	Huffingtonpost	Owen_g
macTweeter	NOLAnews	guardian_sport
engadget	Reuters	itvfootball
parislemon	CBSNews	channel4news
teedubya	LATenvironment	StatesideSoccer
mashable	Kate_sheppard	Flipbooks
TUAW	MotherNatureNet	nikegoal
Scobleizer	mparent77772	FIFAWorldCupTM

Human Evaluation

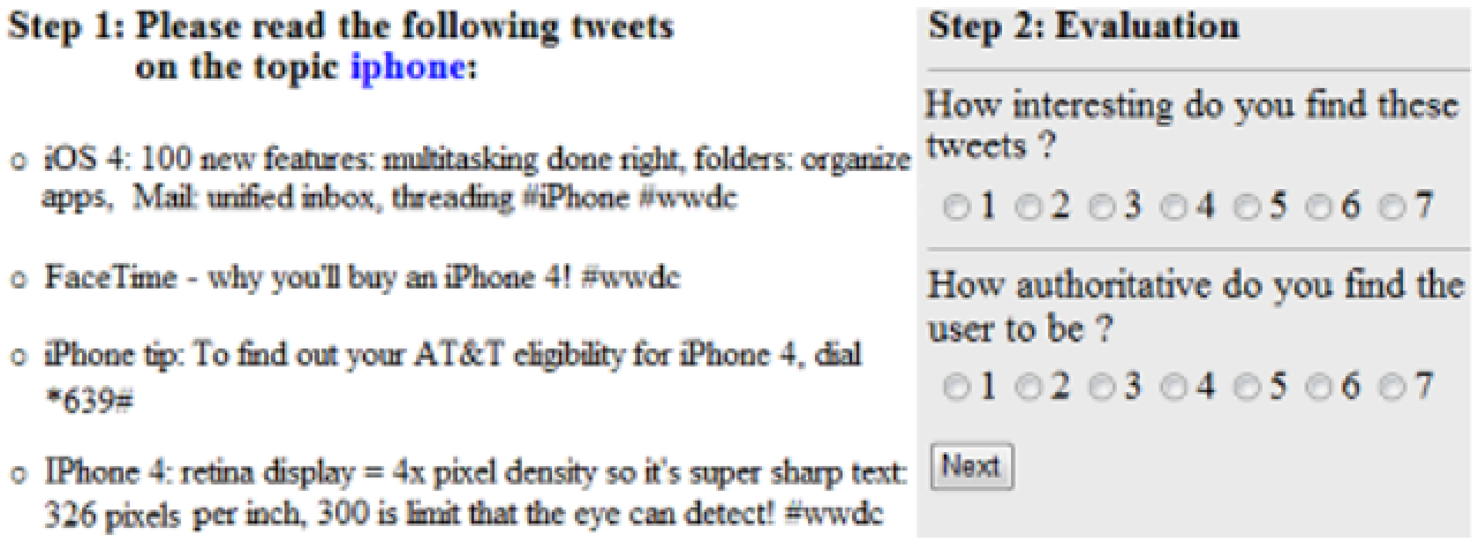
- Evaluate the results of different models

- We selected 20 author from our algorithm, 10 from each baseline
  - ~40 authors per topic, due to overlap

  - 4 original tweets per author
  - url's shortened
  - @username anonymize

- A participant rated 20 authors anonymously, 20 non-anonymously
  - Random author order
  - Equal anonymous and non-anonymous rating per author

Human Evaluation – Anonymous Screen



Step 1: Please read the following tweets on the topic iphone:

- iOS 4: 100 new features: multitasking done right, folders: organize apps, Mail: unified inbox, threading #iPhone #wwdc
- FaceTime - why you'll buy an iPhone 4! #wwdc
- iPhone tip: To find out your AT&T eligibility for iPhone 4, dial *639#
- IPhone 4: retina display = 4x pixel density so it's super sharp text: 326 pixels per inch, 300 is limit that the eye can detect! #wwdc

Step 2: Evaluation

How interesting do you find these tweets ?

1 2 3 4 5 6 7

How authoritative do you find the user to be ?

1 2 3 4 5 6 7

Next

Human Evaluation – Non Anonymous screen

Step 1: Please read the following tweets on the topic iphone:

username tweets:

- 5 Fun DIY iPhone Cases [PICS] - http:\\bit.ly\961WB7
- New iPhone Release Date Announced at WWDC 2010 - http:\\bit.ly\9pmOs6
- The iPhone 4 Is Here - http:\\bit.ly\bsGy7X
- iPhone Gets iMovie for HD Video Recording and Editing - http:\\bit.ly\bbwd3e

Step 2: Evaluation

How interesting do you find these tweets ?

1 2 3 4 5 6 7

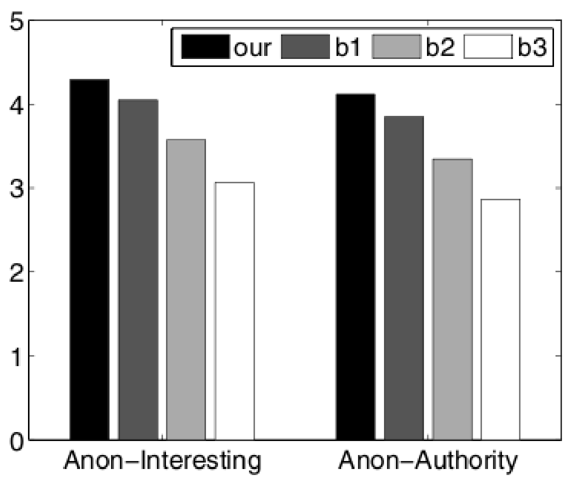
How authoritative do you find the user to be ?

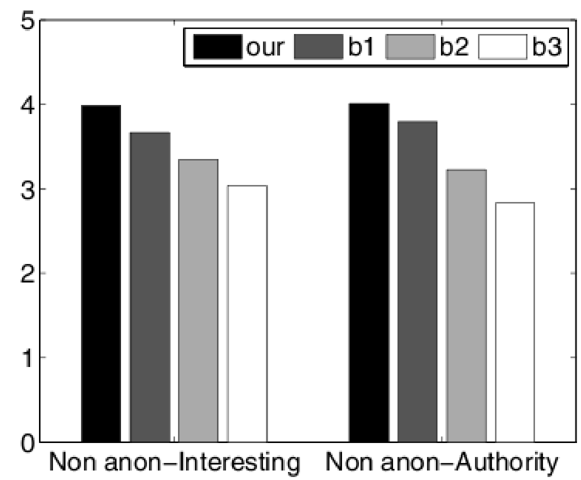
1 2 3 4 5 6 7

Next

Results – Author Rating Comparison

- Average Rating Comparison





- Our vs b1: two sample, one sided t-test
  - Our better for all topics ($p < 0.05$) in non-anon condition
  - Our better for all topics ($p < 0.05$) in anon condition, except world cup

Rating Comparison

- Best rating: Pick the author with the highest rating per participant per model.
  - One sided ttest, p-value table:

		iphone	oil spill	world cup	overall
anon	interestingness	0.001	0.011	*0.084*	0.001
	authority	0.001	0.006	*0.5*	0.001
non-anon	interestingness	0.024	0.001	0.004	0.001
	authority	0.018	0.044	0.006	0.001

- Top 1-10 vs 11-20
  - $P < 0.057$

Model Precision

- Sort author on ratings and pick top 10
- Absolute performance:

		iphone	oil spill	world cup	overall
anon	interestingness	0.8	0. 8	0.6	0.73
	authority	0.8	0.7	0.5	0.63
non-anon	interestingness	0.7	0.7	0.6	0.6
	authority	0.6	0.7	0.6	0.6

- Compared to other models

		Our vs b1	Our vs b2
anon	interestingness	0.8	1
	authority	0.6	0.93
non-anon	interestingness	0.73	0.793
	authority	0.63	0.7

Algorithm Effectiveness

- Comparing top 10 users: our ranking *vs* anon-interestingness ratings:

- Pearson correlation of ranked lists:

	iphone	oil spill	world cup	overall
our (Gmm)	0.54	0. 41	0.22	0.39
our (Kmeans)	0.40	0.29	0.14	0.28
our (no clustering)	-0.07	-0.05	0.06	-0.02

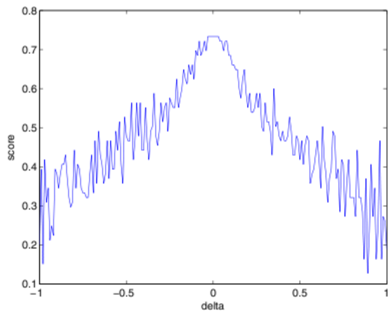
- Ranking algorithms
  - Gaussian ranking (0.39)
  - List based ranking (0.17)

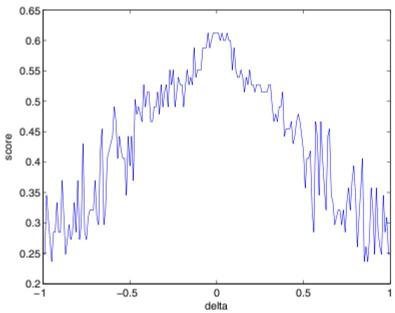
Estimating Optimal Weights

- Weighted Ranking:

$$R_G(x_i) = \prod_{f=1}^{d} [\int_{-\infty}^{x_i^f} N(x; \mu_f, \sigma_f)]^{w_f}$$

- Maximize Pearson Score





- Best correlation: 0.56 (iphone) and 0.61 (oil spill)
- Our correlation: 0.54 (iphone) and 0.41 (oil spill)

- Mention Impact and Topical signal should have higher weights than rest

Conclusion

- Near real-time algorithm

- Our method yields authors of greater interest and authoritativeness than the baseline models
  - Some combination of popular and less popular authors is a likely "sweet spot"
  - We isolated the role that name value of authors plays when evaluating their content
    - anonymous ratings higher than non-anonymous ratings
    - popular authors get a boost when their names are revealed (popularity matters)

- Probabilistic Clustering is useful
  - Removes outliers, robust ranking
  - Better than rule based filtering: e.g. OT > 5 (doesn't work)

- Microblogging is a more dynamic environment
  - Short lifetime of topics
  - Graph based approach can wrongly assign a celebrity as authority (like *shakira* for *worldcup*)
  - Graph based (b1) is better than pure text based (b2)

Future Work

- Explore precise balance of popularity vs topic
  - Depends on timing e.g. for *iphone* popular users like *mashable* might matter if topic is pressing

- Ways to filter human:{male, female}, organizations, topical names

Thanks