Finding High-Quality Content in Social Media

Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

Eugene Agichtein et al.

Identifying Topical Authorities in Microblogs

Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011.

Aditya Pal and Scott Counts

Presented by Matt Chaney

CS 834 - Presentation 5

Web is Changing

1990s - traditional content producers / consumers

2000s - user-generated content ➔ “pro-sumers”

- Blogs
- Photo/Video sharing
- Social Networking
- **Question / Answer**
    - **Yahoo! Answers**

Yahoo! Answers

- Users ask questions / provide answers on *any* topic

- Actively regulate the system through democratic quality management
  - Mark an “interesting” question
  - Vote on answers - thumbs up / down
  - Report offensive / inappropriate behavior

- Users have threefold role
  1. Asker
  2. Answerer
  3. Regulator

- Results in a heterogeneous web of social interactions

Question Lifecycle

- Question is **Open** - about 90,000 questions *per day*
- Other users submit answers
- Question becomes **Closed**
  - Time limit reached - default 4 days
  - Asker closes question to more answers
- Question becomes **Resolved**
  - Other users vote on the best answer
  - Asker selects best answer

Study Goals

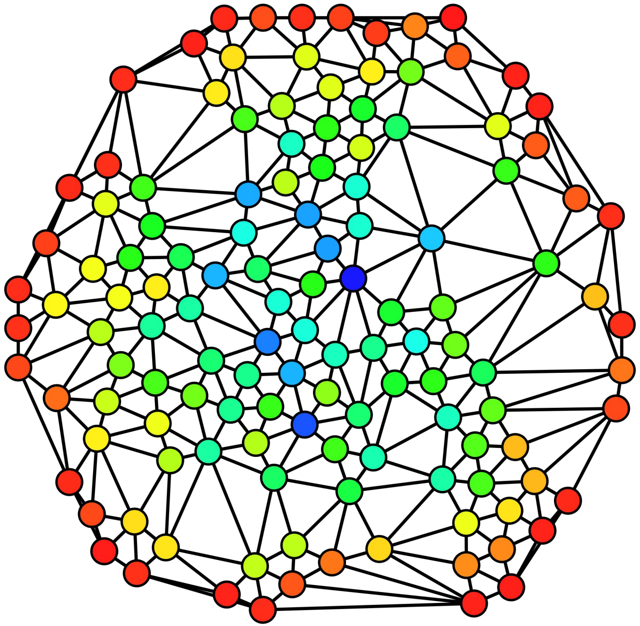
- Quality of content - traditional content quality range is much narrower than unmediated, user-generated content, which varies greatly

- How to find high quality question/answer content?

- Additional content sources
  - Document content
  - Link analysis
  - User-to-document relation types
  - User-to-user interactions

Related Work

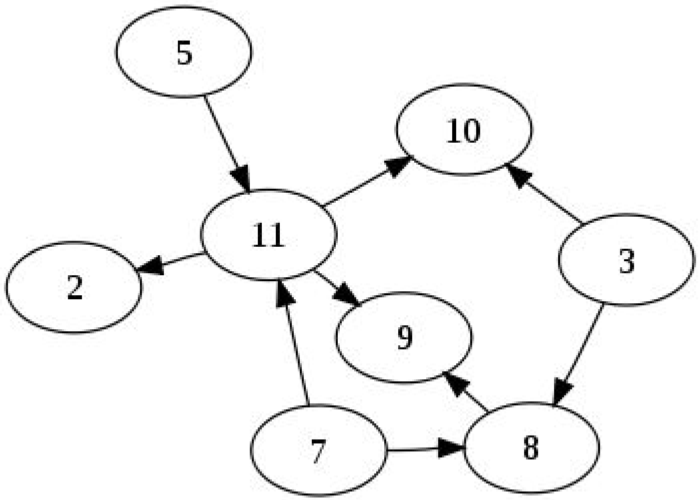
- Link Analysis in Social Media
- Propagating Reputation
- Question / Answering Portals
- Expert Finding
- Content Quality Text Analysis
- Implicit Feedback Ranking



Hue scale representing node betweenness on a graph.

Identifying Quality Content

- Intrinsic content quality, primarily text-related
  - Punctuation / typos
  - Syntactic and semantic complexity
  - Grammaticality
- User relationships
  - Answerer (u) answers user (v) question   *U* → *V*
- Usage statistics
  - Click count / dwell time
- **Classification → High Quality vs The Rest**



A directed graph.
Maat. Public Domain. 2010.

User Relationships

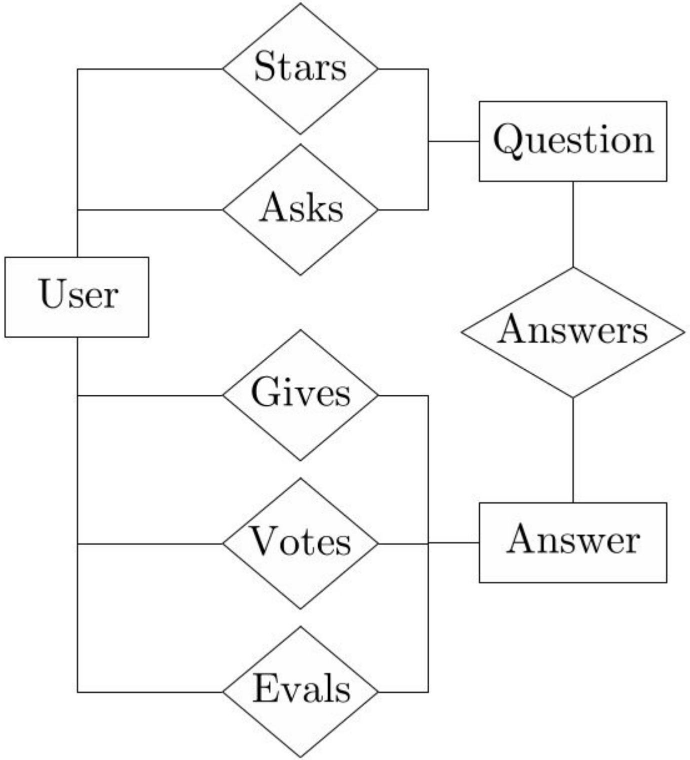


Figure 1. Partial entity-relationship diagram of answers.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

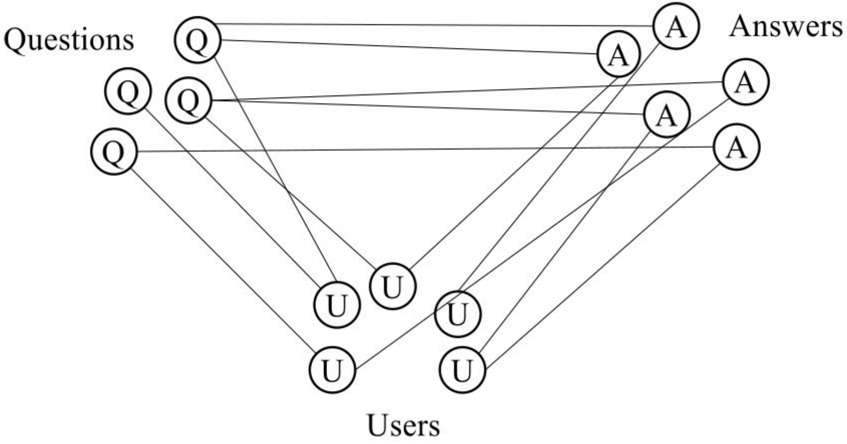


Figure 2: Interaction of users-questions-answers modeled as a tri-partite graph.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Question Feature Space

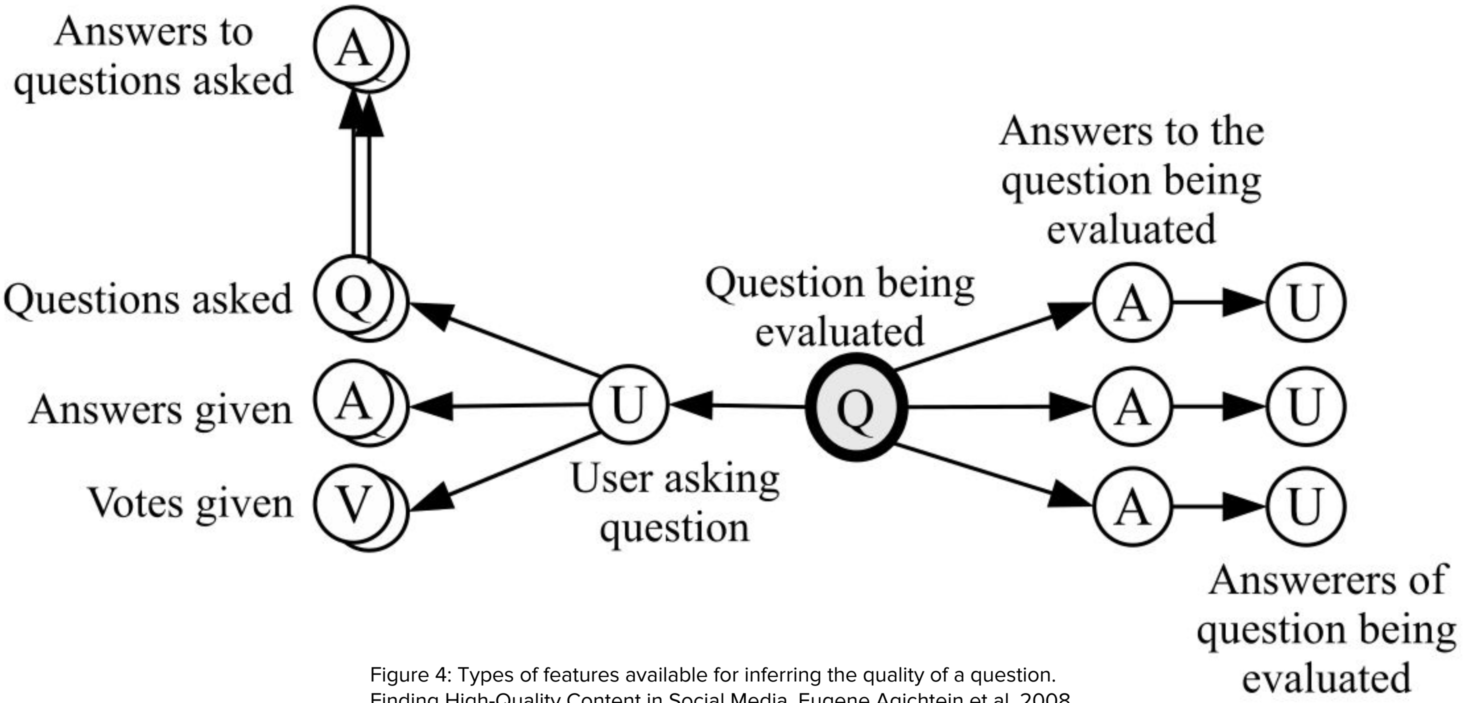


Figure 4: Types of features available for inferring the quality of a question.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Answer Feature Space

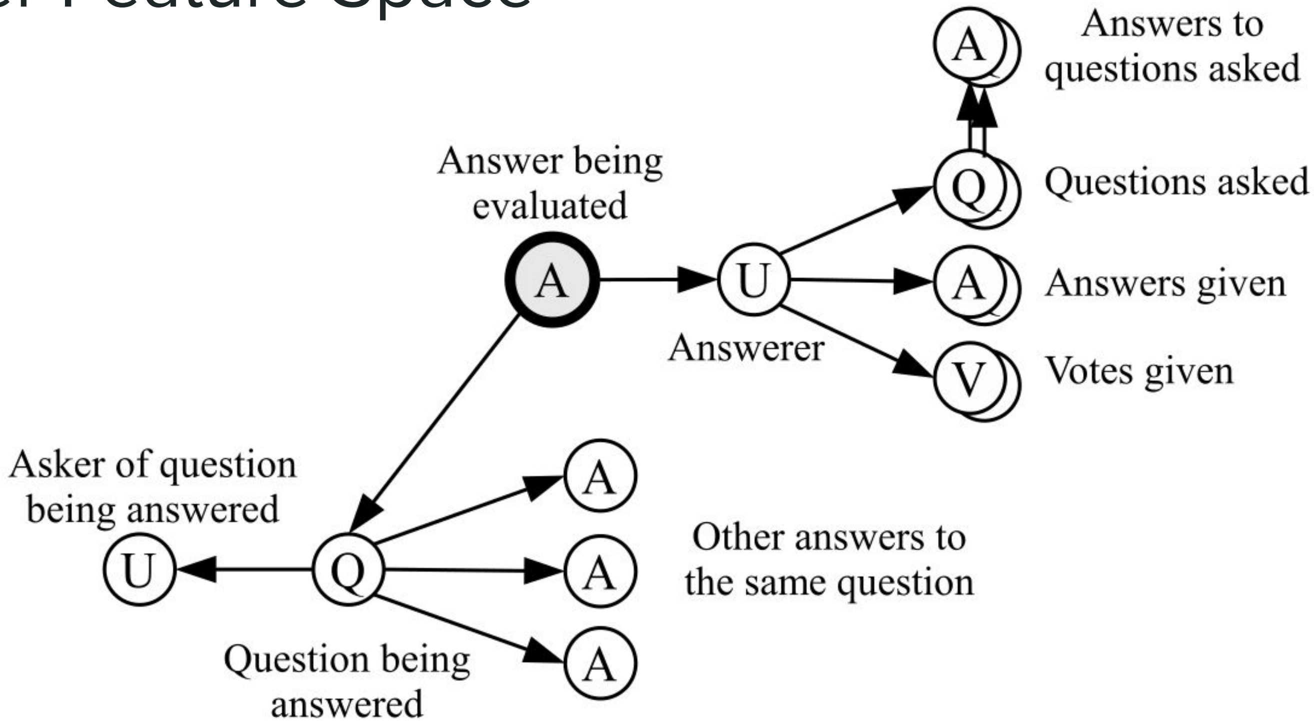


Figure 3: Types of features available for inferring the quality of an answer.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Implicit User-User Relationships

Graph G = (V, E)

- V = users
- E = edges connecting users ➜ $E = \{ E_a \cup E_b \cup E_v \cup E_s \cup E_+ \cup E_- \}$
- Apply HITS and PageRank on each graph

Graph $G_x = (V, E_x)$

- h_x ➜ HITS Hub scores
- a_x ➜ HITS Authority scores
- p_x ➜ PageRank scores

Experimental Setup

- 6,665 questions
- 8,366 Question-Answer pairs
- Labeled for quality by human editors
  - Well-formedness
  - Readability
  - Utility
  - Interestingness
  - Answers → Correctness
  - Type association → Informational, advice, opinion, poll

- 10-Fold cross validation

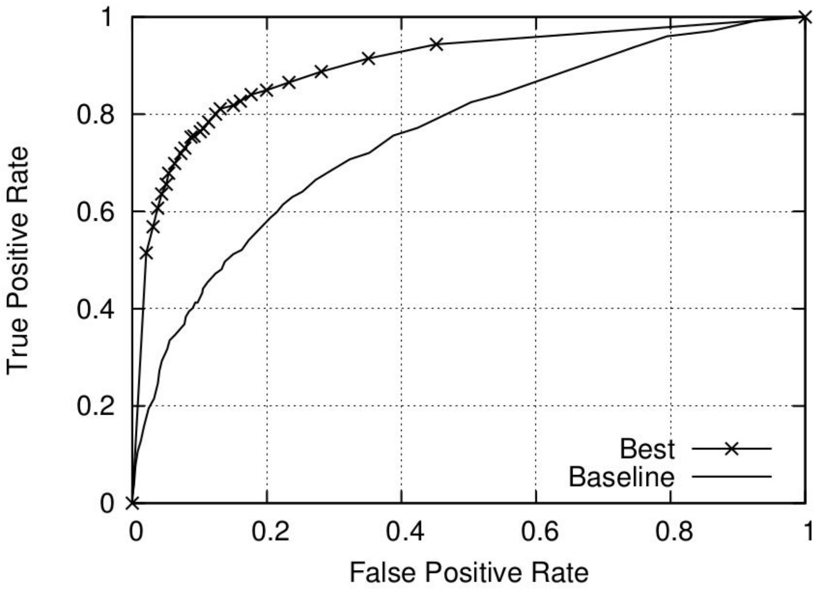
Question Quality Results

Method	High qual. P	High qual. R	Normal/low qual. P	Normal/low qual. R	AUC
Text (Baseline)	0.654	0.481	0.762	0.867	0.523
Usage	0.594	0.470	0.755	0.836	0.508
Relation	0.694	0.603	0.806	0.861	0.614
Intrinsic	0.746	0.650	0.829	0.885	0.645
T+Usage	0.683	0.571	0.798	0.865	0.575
T+Relation	0.739	0.647	0.828	0.881	0.659
T+Intrinsic	0.757	0.650	0.830	0.891	0.648
T+Intr.+Usage	0.717	0.690	0.845	0.861	0.686
T+Relation+Usage	0.722	0.690	0.845	0.865	0.679
T+Intr.+Relation	**0.798**	0.752	0.874	**0.901**	0.749
All	0.794	**0.771**	**0.885**	0.898	**0.761**

Table 2: Precision P, Recall R, and Area Under the ROC Curve for the task of finding high-quality questions.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Receiver Operating Characteristic (ROC) Results

Questions



Answers

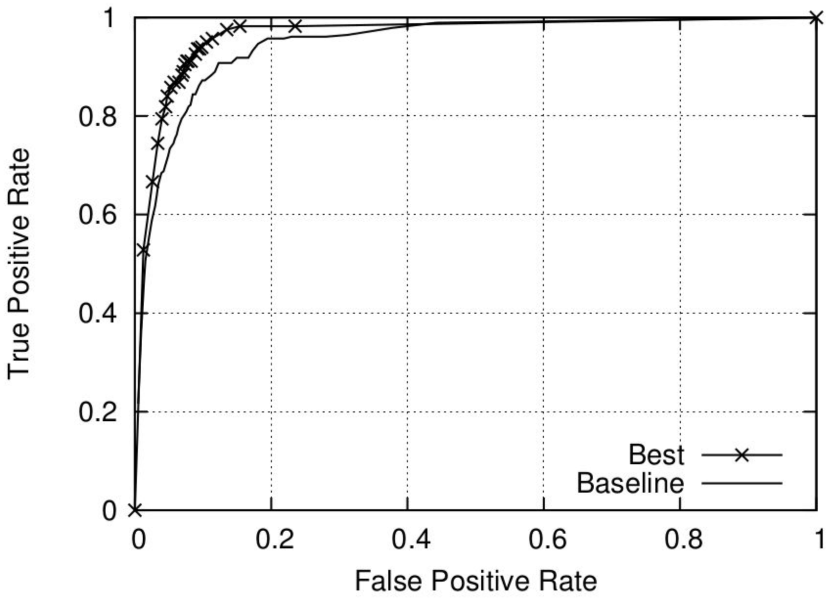


Figure 8. ROC curve for the best-performing classifier, for the task of finding high-quality questions (left) and high-quality answers (right).
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Finding High-Quality Content in Social Media

Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.

Eugene Agichtein et al.

Identifying Topical Authorities in Microblogs

Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011.

Aditya Pal and Scott Counts

Presented by Matt Chaney

CS 834 - Presentation 5

Advent of Microblogging

Short, simple messaging services, in the form of a blog, but on a much smaller scale

Motivation

- Huge number of content producers
  - \+ Great diversity
  - − Finding quality difficult

- Identify *topical* authorities
  - Avoid overgeneralized, well-known authorities, e.g. news outlets
  - Topical authority authors may not exist until an event occurs

- Graph analysis metrics insufficient
  - Sensitive to celebrity authors
  - Computationally infeasible for large-sized datasets in real time

- Algorithm that finds topical authorities automatically in near real-time using clustering

Related Work

- Most prior work dominated by computationally expensive **network analysis** approaches

- TwitterRank
  - Variant of PageRank
  - Latent Dirichlet Allocation
  - Weighted user graph ➔ weight = topical similarity
  - Method differs in using **clustering** over graph analysis and using additional **author features**

- PageRank; HITS

- Prior automatic authority identification
  - **Community Question and Answering**

Authority Features

- Focus on message impact metrics
  - Original Tweet (OT)
  - Conversational Tweet (CT)
  - Repeated Tweet (RT)
  - Mentions (M)
  - Graph Characteristics (G)

- Self Similarity $$S(s1, s2) = \frac{|s1 \cap s2|}{|s1|}$$

$$S(a) = \frac{2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{i-1} S(s_i, s_j)}{(n-1) \cdot n}$$

ID	**Feature**
$OT1$	Number of original tweets
$OT2$	Number of links shared
$OT3$	Self-similarity score that computes how similar is author's recent tweet w.r.t. to her previous tweets
$OT4$	Number of keyword hashtags used
$CT1$	Number of conversational tweets
$CT2$	Number of conversational tweets where conversation is initiated by the author
$RT1$	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
$M1$	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
$G1$	Number of topically active followers
$G2$	Number of topically active friends
$G3$	Number of followers tweeting on topic after the author
$G4$	Number of friends tweeting on topic before the author

Table 1: List of metrics of potential authorities.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Feature List

$$\text{Topical signal } (TS) = \frac{OT1 + CT1 + RT1}{|\#\ \text{tweets}|}$$

$$\text{Signal strength } (SS) = \frac{OT1}{OT1 + RT1}$$

$$\text{Non-Chat signal } (\bar{C}S) = \frac{OT1}{OT1 + CT1} + \lambda \frac{CT1 - CT2}{CT1 + 1}$$

$$\lambda < \frac{OT1}{OT1 + CT2} \cdot \frac{CT1 + 1}{OT1 + CT1}$$

ID	Feature
$OT1$	Number of original tweets
$OT2$	Number of links shared
$OT3$	Self-similarity score that computes how similar is author's recent tweet w.r.t. to her previous tweets
$OT4$	Number of keyword hashtags used
$CT1$	Number of conversational tweets
$CT2$	Number of conversational tweets where conversation is initiated by the author
$RT1$	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
$M1$	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
$G1$	Number of topically active followers
$G2$	Number of topically active friends
$G3$	Number of followers tweeting on topic after the author
$G4$	Number of friends tweeting on topic before the author

Table 1: List of metrics of potential authorities.
Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Feature List

$$\text{Retweet impact } (RI) = RT2 \cdot \log(RT3)$$

$$\text{Mention impact } (MI) = M3 \cdot \log(M4) - M1 \cdot \log(M2)$$

$$\text{Information diffusion } (ID) = \log(G3+1) - \log(G4+1)$$

$$\text{Network score } (NS) = \log(G1+1) - \log(G2+1)$$

ID	Feature
$OT1$	Number of original tweets
$OT2$	Number of links shared
$OT3$	Self-similarity score that computes how similar is author's recent tweet w.r.t. to her previous tweets
$OT4$	Number of keyword hashtags used
$CT1$	Number of conversational tweets
$CT2$	Number of conversational tweets where conversation is initiated by the author
$RT1$	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
$M1$	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
$G1$	Number of topically active followers
$G2$	Number of topically active friends
$G3$	Number of followers tweeting on topic after the author
$G4$	Number of friends tweeting on topic before the author

Table 1: List of metrics of potential authorities.

Finding High-Quality Content in Social Media. Eugene Agichtein et al. 2008.

Gaussian Mixture Model

- Probabilistic model for representing sub-populations within overall populations, without requiring prior identification of the target sub-population
- Group users into two clusters over entire feature space
- Less sensitive to outliers (celebrities)
- Aims to maximize the likelihood of the data

Maximizing Likelihood of the data

Consider n data points $x = \{ x_1, x_2, \ldots, x_n \}$ in *d*-dimensional space, density of a point x:

$$p(x|\pi, \Theta) = \sum_{z=1}^{k} p(z|\pi) \cdot p(x|\theta_z)$$

$\boldsymbol{\pi}$ is the prior over the k components (using k-means) and
$\Theta = \{ \theta_z : 1 \le z \le k \}$ as model parameters of k Gaussian distributions
with $\theta_z = \{\mu_z , \Sigma_z \}$

$$p(x|\theta_z) = \frac{1}{((2\pi)^d|\Sigma_z|)^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu_z)^T \Sigma_z^{-1}(x-\mu_z)\}$$

Expectation Maximization

Assuming points are independent and identically distributed, likelihood is:

$$p(\mathbf{x}|\pi, \Theta) = \prod_{i=1}^{n} P(x_i|\pi, \Theta)$$
$$= \prod_{i=1}^{n} \sum_{z=1}^{k} p(z|\pi) \cdot p(x_i|\theta_z)$$

Expectation Maximization (EM) - Iterative algorithm with 2 steps

1. E-Step ➔ Compute probability of *k* Gaussian components given data points using Bayes' theorem

2. M-Step ➔ Compute model parameters to maximize the likelihood of the data

Ranking within clusters

Gaussian Ranking Algorithm

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

- $N(x; \mu_f, \sigma_f)$ is the univariate Gaussian distribution with model parameters as μ_f and σ_f

- Computes Gaussian Cumulative Distribution (GCD)
    - Monotonically increasing function, useful for ranking
    - Scoring can be high or low with modification to the integral

Dataset

- All tweets posted from June 6th 2010 to June 10th 2010
- 89,622,039 tweets in total
- Three topics

	$\|U\|$	$\|OT\|$	$\|CT\|$	$\|RT\|$
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

Table 2: Dataset statistics. |U|, |OT|, |CT|, |RT| are overall count of users, original tweets, conversational tweets and retweets, respectively.
Identifying Topical Authorities in Microblogs. Aditya Pal and Scott Counts. 2011.

Baselines for Comparison

- ***b1*** - Graph Properties { RI, MI, ID, NS }
- ***b2*** - Textual Properties { TS, SS, $\overline{C}$S }
- ***b3*** - Random selection of authors that fall outside target cluster
- ***our*** - The authors' model

Evaluation - User Study

- Comparison of ***our*** vs three baselines ***b1***, ***b2***, and ***b3***
- 48 participants
- Users presented with 40 screens
  - 20 anonymous author
  - 20 author shown
  - 4 tweets
  - Topic
- Rate on scale 1 - 7 on *interestingness* and *authoritativeness*
- Validated statistical significance of results using **one-sided paired t-tests**

Results

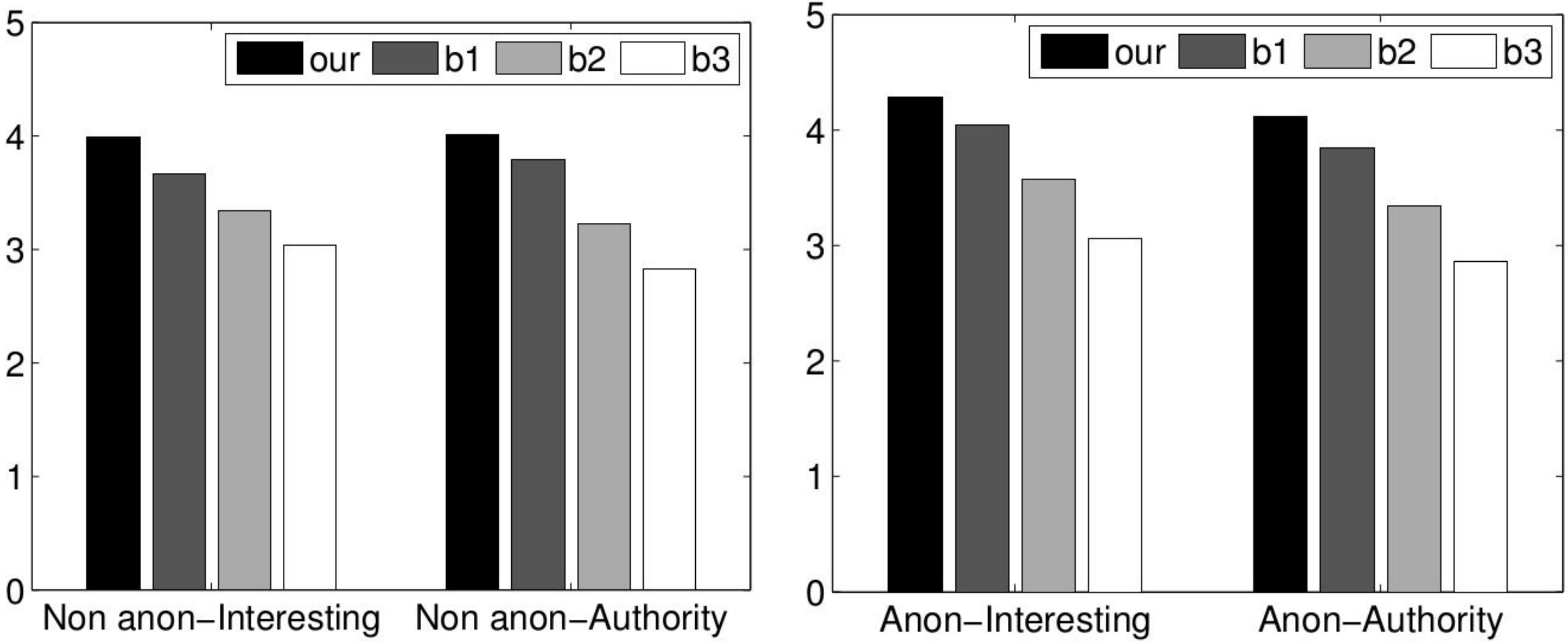


Figure 3: Average ratings per model per participating user.
Identifying Topical Authorities in Microblogs. Aditya Pal and Scott Counts. 2011.