Identifying and Analyzing Researchers on Twitter

Robert Jäschke
Asmelash Teka Hadgu
Agenda

- Motivation
- Approach
- Results
Twitter is ...

- a communication platform,
- a social network,
- a system for resource sharing

... which researchers use ...

- to connect with other researchers,
- to announce calls for papers,
- to communicate and discuss,
- to stay up-to-date,
- etc.

source: twitter.com
Goals

- Understand how Twitter is used by researchers
  - Differences according to discipline, age, country, …?
  - Who's following/retweeting/mentioning whom?
  - Information flow between areas/disciplines
  - Impact and influence

source: mapequation.org
Goals

- Improve retrieval and discovery of **scientific content**
  - Researchers, topics, publications, conferences, ...
  - Trends, developments over time
  - Personalized recommendations

source: twitter.com
Goals

- Transfer peer review to social media
- What do researchers regard as important?
Challenges

- Data acquisition
  - Tweets and users from Twitter
  - Ground truth to train and evaluate algorithms
- Identifying researchers
  - One class problem: finding good counterexamples is difficult
- Brevity of tweets
  - How to extract meaning from 140 characters?
- Identifying and classifying scientific tweets
  - What is a scientific tweet?
- Ranking scientific content
  - How to evaluate a ranking?
Related Work

- Twitter directories (e.g., Wefollow, Twellow, JustTweetIt)
- User classification:
  - M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in Twitter (2011)
- Scholars on Twitter:
  - J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to widely spread scientific messages (2010)
  - K. Weller, E. Dröge, and C. Puschmann. Citation analysis in Twitter: Approaches for defining and measuring information flows within tweets during scientific conferences (2011)
- Typically: focus on tweets, not users
Agenda

- **Motivation**
- **Approach**
- **Results**
Approach

As a first step, we
- focused on computer science
- developed a pipeline to identify researchers
- analyzed their age, popularity, influence, and social network
Approach

Finding good seeds:

- requirements: small set, good coverage, followers likely scientists
- solution: Twitter accounts of computer science conferences
- started with a list from Wikipedia¹, searched for Twitter accounts
- 170 accounts for 98 conferences

1: http://en.wikipedia.org/wiki/List_of_computer_science_conferences
Approach

Generating candidates:
- follower, friends, retweeter of the seeds
- recursive approach possible but reduces precision
- 52678 accounts, mostly interested in one conference (83%)
Approach

Verifying candidates with ground truth:
- using computer science publications as evidence
- matching against 1304283 author names from DBLP
- matching: string matching of real name, ignoring duplicates
- manual validation of 150 verified accounts: 73% accuracy
Approach

Finding negative examples:

- challenging task: most users are not researchers
  - How to get a representative sample?
- randomly collected users from the Twitter stream
- removed candidates, their followers and friends
- added seeds and large companies
Approach

Extracting features:
- Which features can separate researchers from other users?
- profile (#tweets, #followers, website set, bio keywords, etc.) and
- content (#tweets with URLs, #scientific tweets, etc.) features,
- no network (#followed seeds, etc.) features
Approach

Classifying candidates:
- stratified 10-fold cross validation (2000 random cand. + neg. ex.)
- Random Forest: F1 of 0.94
- Baseline (SVM on Bag of Words): F1 of 0.93
- 38368 positive candidates, 5015 negative candidates
Agenda

- Motivation
- Approach
- Results
Which areas of computer science?
Are researchers on Twitter younger?
Are they more productive?
How are they connected with each other?

(a) Who follows whom?

- in general, the order of activity is follow, mention, retweet
How are they connected with each other?

(b) Who retweets whom?

- in general, the order of activity is follow, mention, retweet
How are they connected with each other?

(c) Who mentions whom?

- in general, the order of activity is follow, mention, retweet
How does closer scientific collaboration affect interaction on Twitter?

(a) following
(b) mentioning
(c) retweeting
Who are the most influential researchers?

<table>
<thead>
<tr>
<th>screen name</th>
<th>name</th>
<th>ranking by</th>
</tr>
</thead>
<tbody>
<tr>
<td>@timoreilly</td>
<td>Tim O’Reilly</td>
<td>1 2 1 16</td>
</tr>
<tr>
<td>@billgates</td>
<td>Bill Gates</td>
<td>2 1 2 1</td>
</tr>
<tr>
<td>@hmason</td>
<td>Hilary Mason</td>
<td>3 9 3 2</td>
</tr>
<tr>
<td>@zephoria</td>
<td>Danah Boyd</td>
<td>4 7 6 24</td>
</tr>
<tr>
<td>@csoghoian</td>
<td>Christopher Soghoian</td>
<td>5 51 12 5</td>
</tr>
<tr>
<td>@doctorow</td>
<td>Cory Doctorow</td>
<td>6 16 4 2</td>
</tr>
<tr>
<td>@ioerror</td>
<td>Jacob Appelbaum</td>
<td>7 30 7 5</td>
</tr>
<tr>
<td>@mattmight</td>
<td>Matthew Might</td>
<td>8 47 16 34</td>
</tr>
<tr>
<td>@kentbeck</td>
<td>Kent Beck</td>
<td>9 18 17 35</td>
</tr>
<tr>
<td>@mattcutts</td>
<td>Matt Cutts</td>
<td>10 15 9 2</td>
</tr>
<tr>
<td>@timberners_lee</td>
<td>Tim Berners-Lee</td>
<td>11 3 5 35</td>
</tr>
<tr>
<td>@codepo8</td>
<td>Christian Heilmann</td>
<td>12 87 14 1</td>
</tr>
<tr>
<td>@mattblaze</td>
<td>Matt Blaze</td>
<td>13 60 25 72</td>
</tr>
<tr>
<td>@digiphile</td>
<td>Alex Howard</td>
<td>14 42 13 1</td>
</tr>
</tbody>
</table>

verified candidates

retweet  follow  mention  #publications
Who are the most influential researchers?

top 200 influential researchers
other researchers

Matthew Might
@mattmight
CS prof

source: twitter.com

#science, Robert Jäschke
12.6.2014
Outlook

- improve matching accuracy
- analyze topics & interests of users
- social network analysis
- transfer to other disciplines
- build a web directory of researchers on Twitter

**dataset:** https://github.com/L3S/twitter-researcher

Thanks for your patience! Questions?

- Motivation

- Approach

- Results