Modeling Topics and Behaviors of Microbloggers: An Integrated Approach

Tuan-Anh Hoang[+] and Ee-Peng Lim[*]

[+]L3S Research Center, Leibniz University of Hanover, Germany

[*]Living Analytics Research Centre, Singapore Management University, Singapore

Microblogging: Rich Data Sources

- Social network + information network
  - Users interact with other users
  - Users generate and consume content
- Large number of users
- Heavily used in daily life

300M monthly active users
500M tweets every day

170M monthly active users

- Data is publicly shared

Microblogging: Multimodal Data

- User generated content

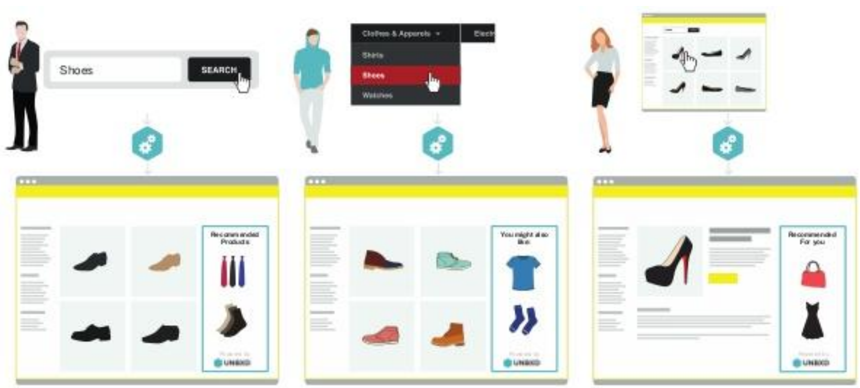
- User behavior of multiple types
  - Relationship: follow, unfollow other users, etc.
  - Communication: reply, mention other users, etc.
  - Propagation: retweet, share URL, etc.
  - Linguistic: use hashtag in tweets, etc.
  - etc.
- Etc.

Applications

User profiling

	Gender	Age	Religion	Party	...
					...
					...

Personalized recommendation



. . .

Personal Interest & Community Interest

- Interests may be shown in either content or behavior
- Users' personal interest is not always the same with interest of their topical communities (realms)

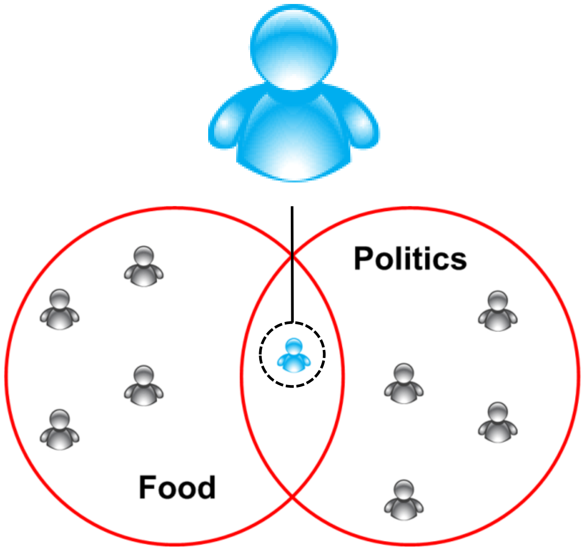
Been using @Microsoft #Windows8 on desktop & tablet. It's very promising.

New #HTML5 #Javascript book @Amazon HTML5 Game Development Insights 24 chapters 20 authors …

Avoid canned #foods, especially for your #kids

Good piece on @BarackObama, #OFA, and the midterm #elections: http://bit.ly/aZoeSb #p2.

.Net Dev, HTML5, JavaScript, Basketball



Follows: Microsoft, ForbesTech, NBA, MiamiHeat, BarrackObama, CNNPolitics

Retweets from: ForbesTech, MiamiHeat, BarrackObama, CNNPolitics

Mentions users: @Microsoft, @Amazon, @BarrackObama

Adopts hashtags: #windows8, #JavaScripts, #kids, #foods, #elections, #p2

Shortcomings of Existing Works

- Consider either user content or user behavior
  - E.g., Ramage et al. 2010; Zhao et al. 2011; Cheng et al. 2014; etc.
- Consider only a single type of behavior
  - E.g., Liu et al. 2010; Yan et al. 2012; Barbieri et al. 2014; etc.
- Do not differentiate between personal interest and realms
  - Determine a users' personal interest solely based on interests of their realms.
    - E.g., Yin et al. 2012; Sachan et al. 2014; etc.
  - Determine a realm's interest by aggregating interest of its members
    - E.g., Kim et al. 2012; Yang et al. 2014 ; etc.

This work

- To learn users' personal interest and interest of their topical communities from both content and behavior
  - Topical community = *Realm*

- To differentiate between the two kinds of interests

- To learn users' dependency on their realms in generating content and adopting behavior

Integrated Approach

- To develop a unified model that considers:
  - Both content and behaviors of multiple types
  - Both personal interest and realms

- Modeling principles
  - Users may belong to multiple realms
  - Topic of content/ behavior may be chosen from either user' personal interest or one of her realms
  - The source of topic is determined by user's bias toward her realms

Data Representation

- Tweet = bag of words

  "He likes football and his brother likes basketball"

  = {and:1, basketball:1, brother:1, football:1, he:1, his:1, likes:2}

- Topic = multinomial distributions over words/ behaviors

  P{"match"| topic = "**sport**"} >> **P**{"programming"| topic = "**sport**"}

  P{following Barack Obama| topic = "**politics**"} >>

  P{following Justin Bieber| topic = "**politics**"}

- Interest = multinomial distribution over topics

  P{topic = "fashion"| "**interested in fashion**"} >>

  P{topic = "sport"| "**interested in fashion**"}

GBT Model

ϕ topic's word distribution

λ topic's behavior distribution

α user's topic distribution

σ realm's topic distribution

μ user's bias

π user's realm distribution

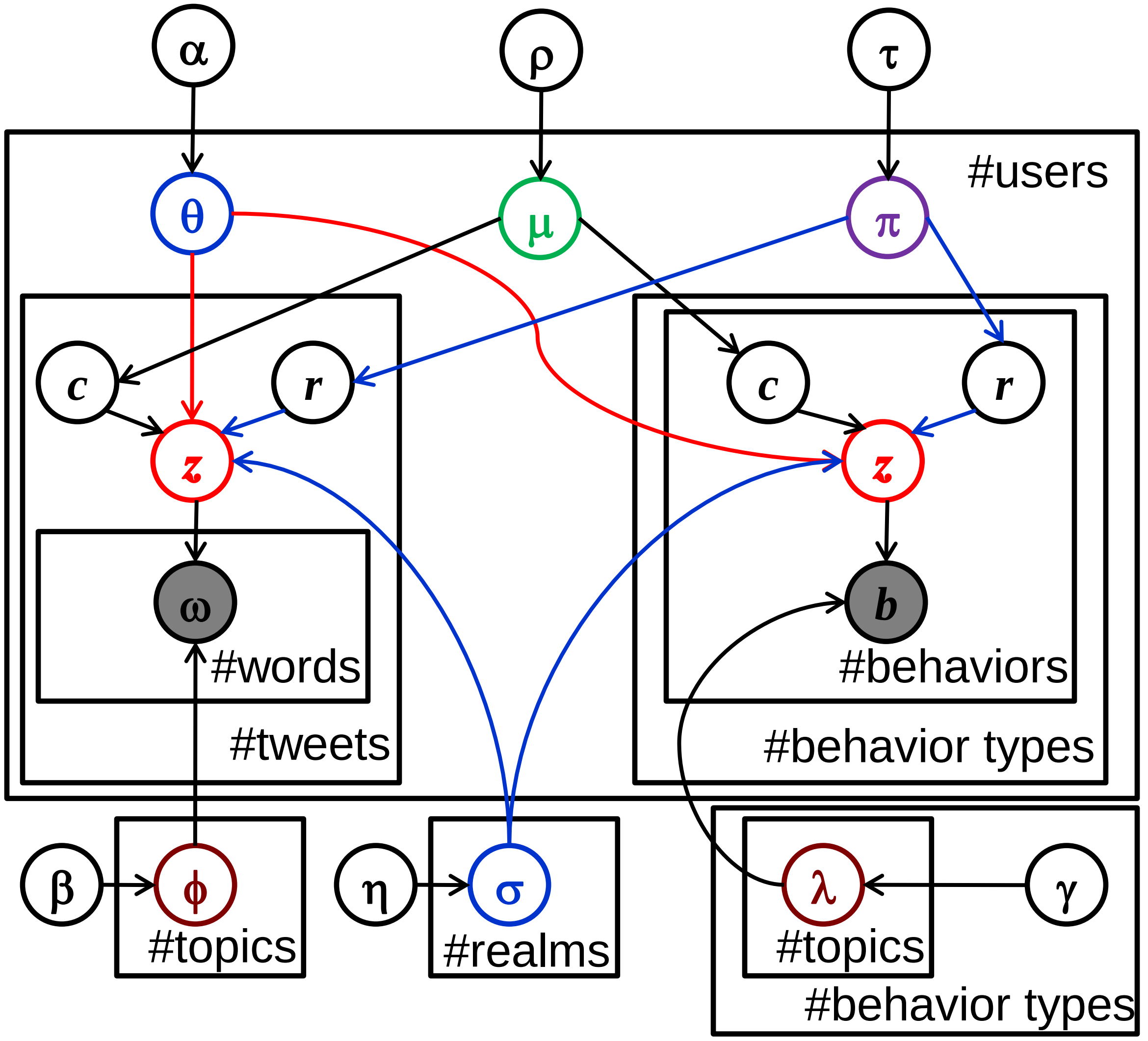
c source index

r realm index

z topic index

⟶ if $c = 0$

⟶ if $c = 1$



Sparsity Regularization

- To obtain semantically clearer realms
  - Realms and users focus on different topics
  - Different realms focus on different topics

- Bias toward skewness in Prob{source **c**| topic ***z***}
  - Topic ***z*** is mostly covered by either users' personal interest or realms

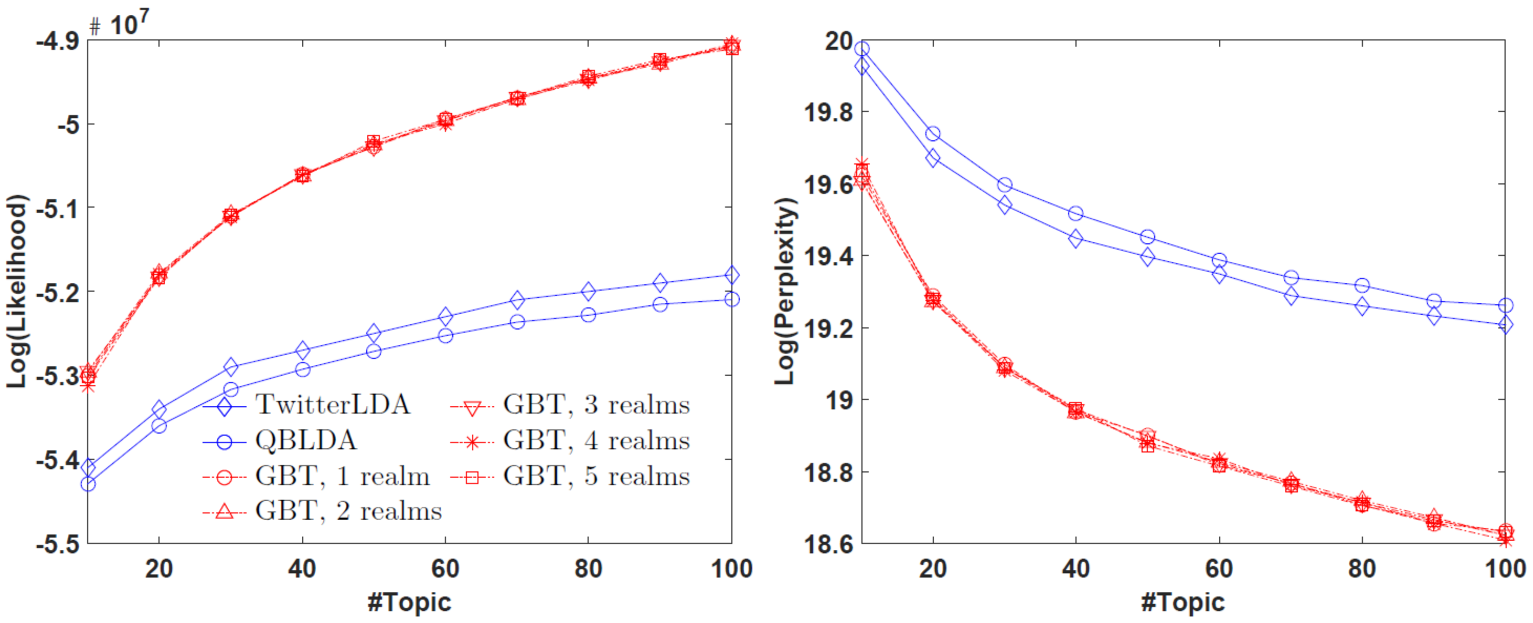
- Bias toward skewness in Prob{realm ***r***| topic ***z***}
  - Topic ***z*** is mostly covered by one or a few realms

SE Dataset

- Collected from a set of Twitter users following influential software developers in August – October, 2011
- 14K+ users
- Content: 3M+ tweets
- Behaviors
  - 350K+ user mentions
  - 890K+ hashtag adoptions
  - 900K+ retweeting

Likelihood & Perplexity in Content Modeling

- Baselines:
  - **TwitterLDA** (Zhao et al., 2011): consider content only
  - **QBLDA** (Qiu et al., 2013): consider content + behavior types
- Training set/ test set: 90%/ 10%



Realms' Top Topics

Top topics learnt by GBT model

Realm Id	Realm Label	Top topics		
		Topic Id	Topic Label	Probability
0	Software development	44	Scripting programming languages	0.760
		66	Email & social networking services	0.044
		26	Readings	0.043
1	Apple's products	38	iOS	0.369
		22	iPhone & iPad	0.231
		66	Email & social networking services	0.102
2	Daily life	76	Daily stuffs	0.536
		43	Foods & drinks	0.098
		26	Readings	0.089

Background topics learnt by baseline models

Model	Top words of **background topic**
TwitterLDA	life,making,video,blog,change,reading,job,home,thought,line team,power,game,business,money,friends,talking,starting,month,company
QBLDA	video,life,blog,change,job,game,reading,business,power,making thought,line,home,#fb,giving,friends,team,money,talking,running

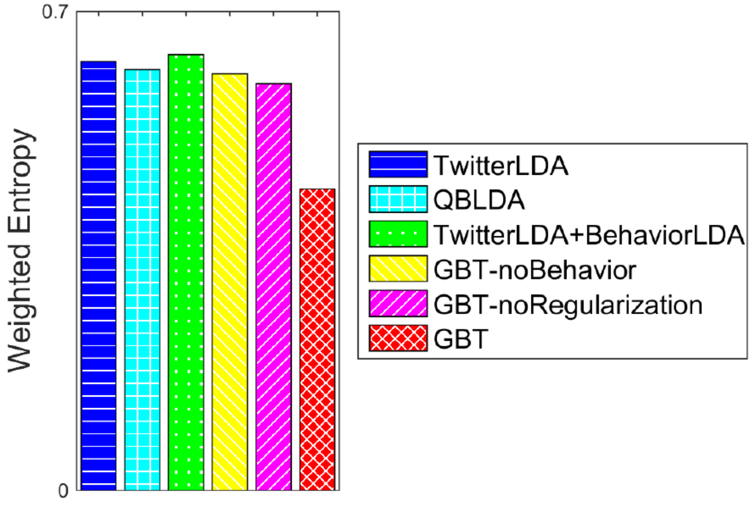
Developer Profiling

- To examine the ability of users' personal topics in determining their favorite programming language

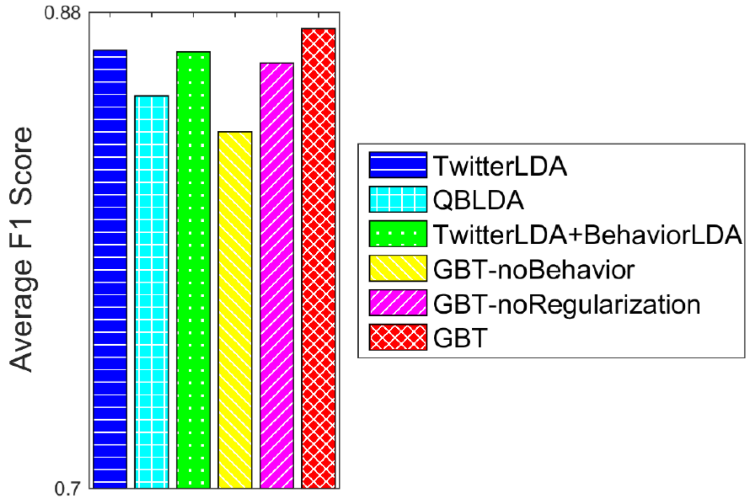
- Tasks
  - User clustering
    - Employ K-mean method
  - User classification
    - Employ SVM method

- Dataset
  - A subset of the SE dataset
  - 328 **.NET** developers
  - 363 developer of **non-.NET** languages

Performance



User clustering performance



User classification performance –10-folds cross validation

- Users' feature vector = topic distribution learnt by different models

Most Discriminative Topics

User label	TwitterLDA		QBLDA		TwitterLDA+behaviorLDA		GBT	
	Topic	Topic Label	Topic	Topic Label	Topic	Topic Label	Topic	Topic Label
.NET	66	Microsoft Visual Studio	5	Microsoft Visual Studio	tweet topic 66	Microsoft Visual Studio	69	Microsoft Visual Studio
	7	Windows Tablets & Phones	47	Windows Tablets Phones	tweet topic 7	Windows Tablets & Phones	35	Windows 8
	40	Lance Armstrong	58	Happenings in London	retweet topic 27	Windows developers	65	Windows Tablets & Phones
non-.NET	75	Data management	79	HTML & Web	tweet topic 75	Data management	44	Scripting programming languages
	47	iOS & iPhone	52	Internet & Media	tweet topic 47	iOS & iPhone	71	Java software development
	64	Entertainment	62	Web Browsers	tweet topic 9	Readings	48	Open-source data management systems

- Top topics learnt by the baselines models are not always representative
- Top topics learnt by GBT model are more reasonable

Future Works

- To incorporate social factors in modeling content generation and behavior adoption
  - E.g., a user may adopt some behavior due to either topical interests or social influence
- To combine more data sources
  - E.g., geo information and image embedded in tweets, mass media, etc.

Q&A

Thank you for your attention!