Predicting Restaurant **Consumption Level** through Social Media Footprints

Yang Xiao, Yuan Wang, Hangyu Mao, Zhen Xiao
NC&IS LAB, Peking University, China

Outline

- **Motivation**
- Our Method
- Experiment
- Conclusion

The Social World

twitter

555 million Users
58 million Tweets
Per Day

新浪微博
weibo.com

560 million Users

facebook

1,310,000,000
Active Users
18 minutes Spent
Per Visit

WeChat微信

700 million Users
Wechat cover all
smartphones.

Demographic Prediction

- Demographic prediction is important for
  - ad recommendation
  - personalization
- Simple attributes
  - social media profile
  - age (Al Zamal et al., 2012; Nguyen et al., 2013)
  - gender (Ciot et al., 2013; Liu and Ruths, 2013; Rao et al., 2011)
- Complicated attributes
  - tweets, profile
  - tags (Feng and Wang, 2012)
  - political orientation (Pennacchiotti and Popescu, 2011)
- Economic status related attributes
  - useful for business
  - hard to collect ground truth

Outline

- Motivation
- **Our Method**
- Experiment
- Conclusion

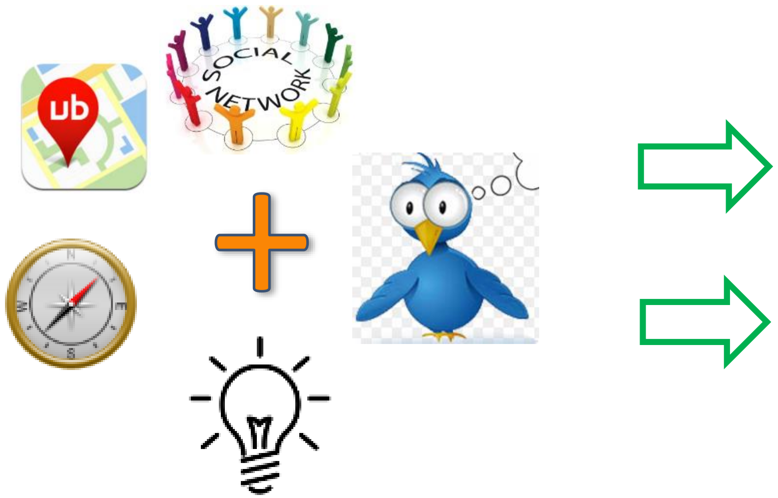
Dianping site

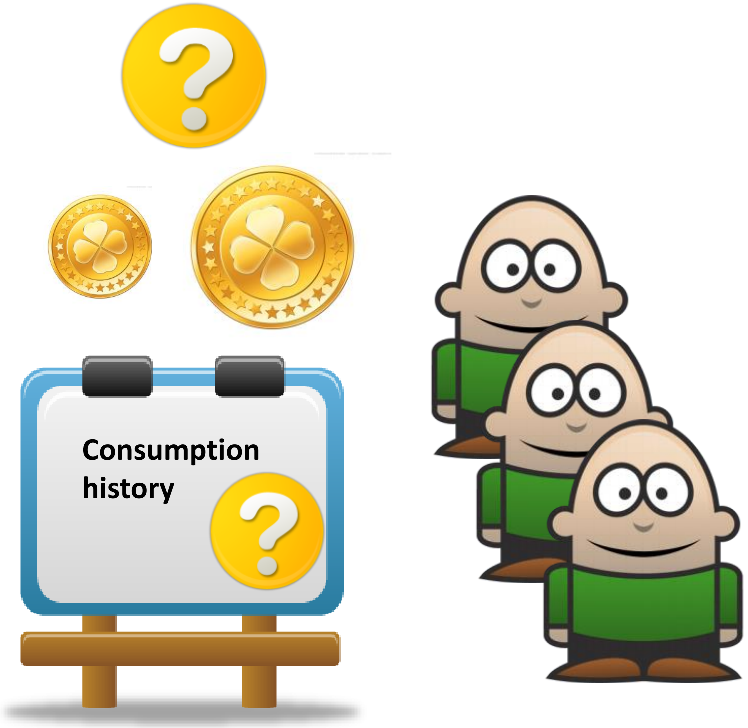
- Dianping is similar to Yelp
- Social network based review site
- Two components
  - users
  - local businesses
- Reviews
  - stars
  - comments
  - average spending

Overview

Weibo Footprints

Dianping Consumption Level

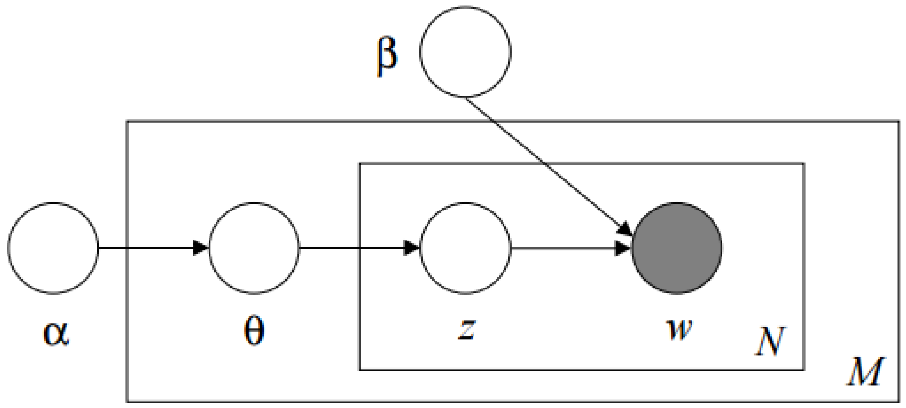




Weibo Features

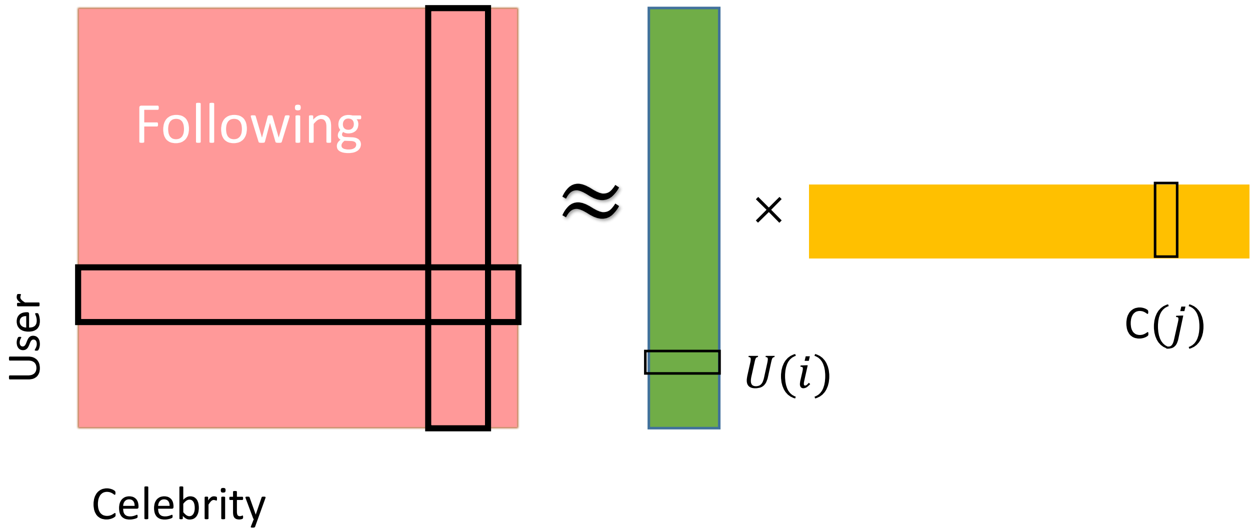
Content perspective

- Users of different **consumption level** have different **word** usage preference.
- Raw features
  - bag of words
- LIWC (Linguistic Inquiry and Word Count)
  - psychological meaningful categories usage preference
- Topics
  - Latent Dirichlet Allocation



Relationship Perspective

- Users of different **consumption level** follow different **celebrities**
- Raw feature
  - bag of celebrities
- Latent feature
  - logistic loss



$$U(i) = \arg min_w \sum_j \log\left(1 + \exp\left(-f_{ij} w^T C(j)\right)\right) + \lambda \|w\|^2$$

Ground Truth Estimation

- GMM over user average spending
  - compute average spending for each user
  - apply GMM with k=2 to cluster users into two groups
- Gaussian Mixture Model
  - natural structure
  - avoid manual threshold setting

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

$$p(x_i|\theta_z) = \frac{1}{\sqrt{2\pi\sigma_z}} e^{-\frac{(x_i-\mu_z)^2}{2\sigma_z^2}}$$

Outline

- Motivation
- Our Method
- **Experiment**
- Conclusion

Dataset

- Weibo dataset
  - Craw the search page to find linked users
  - Crawl the linked users' tweets, friends and followers
- Dianping dataset
  - Crawl all the linked users' reviews
  - Crawl all the related restaurants
- Dataset

Users	Tweets	Followings	Restaurants	Reviews
8844	13026078	3078497	35650	286069

Experiment Setup

- Task:
  - classification
- Classifier
  - Xgboost for latent features
  - Logistic regression for raw features
- Evaluation metric
  - precision
  - recall
  - F1
  - accuracy
- Profile feature is baseline feature

Experiment Result

Category	Name	Accuracy	Precision	Recall	F1
BASELINE	Age	0.5471	0.5547	0.5108	0.5318
	EDU	0.5507	0.5564	0.5324	0.5441
	TAG	0.5655	0.5629	0.6408	0.5993
	ALL	0.5889	0.5775	0.6715	0.621
RAW	RAWWORD	0.6574	0.6544	0.6715	0.6628
	RAWFOLLOW	0.6945	0.6783	0.7529	0.7137
	ALL	0.7118	0.6969	0.761	0.7276
LATENT	LIWCT	0.6066	0.5908	0.6982	0.64
	LDAT	0.7451	0.7303	0.7863	0.7573
	SVDF	0.7673	0.776	0.7635	0.7697
	ALL	0.8012	0.7821	0.8413	0.8106

Qualitative Analysis

- Topic preference difference between high spending users and low spending users
- Spearman correlation test
  - sort users by spending in descending order
  - group them into 100 buckets
  - correlation test over buckets level to capture trend

Qualitative Analysis

Topic ID	Label	Topic (most frequent words,translations)	rho	p value
13	Seafood	三文鱼,刺身,生蚝,日料,海胆,金枪鱼,鲍鱼,大闸蟹,鲜美,米其林 (*salmon, sashimi, oyster, Japanese cooking, urchins, tuna, abalone, steamed crab, tasty, Michelin*)	0.85	0.0001
32	Politics	反腐,受贿,公职,公安局长,批捕,缓刑,查清,名下,收受 (*anti corruption, accept bribes, public employment, public security bureau chief, ratify the arrest, probation, investigation, name, take*)	0.82	3.81E-05
71	Luxury brands	vogue, victoria, miranda, chanel, kerr, alexander, dior, collection,louis, mcqueen (*vogue, victoria, miranda, chanel, kerr, alexander,dior, collection, louis, mcqueen*)	0.75	0.0017
198	Driving	牌照,高架,成品油,中环,远光,私车,93号,车友会,立交,油门 (*vehicle license, elevated highway, product oil, median cycle, high beam, private car, No. 93 gasoline, car club, Interchange, gas*)	0.74	0.0014
120	Tennis	roger,莎拉波娃,罗杰,马卡洛娃,彭帅,阿扎伦卡,彭帅,郑洁,*oba* (*Roger, Sharapova, Roger, Makarova, Peng Shuai,Azarenka, Peng Shuai, Azarenka, Zheng Jie, oba*)	0.71	0.0001
45	Shanghai dialect	哪能,阿拉,今朝,老早,腔调,模子,白相,事体,闲话,辰光,喔唷 (*how, I, today, previously, cool, personal loyalty, play, thing, talk, time, ugh*)	0.69	0.026
192	Auto	车展,发动机,suv,保时捷,别克,沃尔沃,引擎,凯迪拉克,雷克萨斯,比亚迪 (*auto show, engine, suv, Porsche, Buick, Volvo, engine, Cadillac,Lexus, BYD*)	0.61	0.018

Qualitative Analysis

Topic ID	Label	Topic (most frequent words,translations)	rho	p value
135	Mass brands	美宝莲,宝洁,阿芙,origins,美优,olay,多芬,spa,玉兰油,梦妆(*Maybelline, P&G, AFU, origins, beaubeau.com, olay, dove, spa, olay, mamonde*)	-0.77	0.0054
19	Cooking	关火,八角,豆瓣酱,土豆丝,豆角,切末,桂皮,鸡丁,炸酱面,葱油(*take off heat, aniseed, thick broad-bean sauce, shredded potato, French bean, mince, cinnamon, chicken cubes, Noodles*)	-0.81	0.0008
112	Stars	吴亦凡,朴灿烈,张艺兴,吴世勋,exo-m,金钟仁,边伯贤,黄子韬,exok,泰妍(*exo Kris, Park Chan Yeol, exo Lay, Oh Se-hoon, exo-m, exo-k Kai, Baekyun, exo-m Tao, exo-k, Taeyeon*)	-0.81	9.19E-06
142	Character expression	2333, wwww, hhhh, OwO, hhhhh, 233333, QvQ, QuQ, wwwww, 0v0 (*2333, wwww, hhhh, OwO, hhhhh, 233333, QvQ, QuQ, wwwww, 0v0*)	-0.57	0.0322

Qualitative Analysis

- Interaction analysis between topic and gender or topic and age

Topic No.	Label	t(Age)	t(Gender)
13	Seafood	0.8837	2.2599
32	Politics	10.1372	-30.1144
71	Luxury brands	-1.8778	9.5550
198	Driving	7.8684	-7.2142
120	Tennis	-2.5192	-0.8891
45	Shanghai dialect	4.7150	5.9072
192	Auto	2.8303	-13.6531
135	Mass brands	-0.7032	7.0779
19	Cooking	-2.8099	7.3084
112	Stars	-2.7430	2.7556
142	Character expression	-7.4935	1.0283

Qualitative Analysis

- Celebrity Analysis

Celebrity		Celebrity	
Dianping Coupon Shanghai	-	Beijing subway	-
Beijing TV cusine programme	-	Reciting words app	-
Comic dialogue player	-	Beijing SKP	+
Tourism related company	+	International radio anchor	+
Waldorf astoria	+	UK shopping	+
Wine related magazine	+	Charity fund	+

Conclusion

- Weibo knowledge is effective to predict consumption level
- Users of different consumption levels uses different topics, words, celebrities
- Scalability
  - extend to other text based third party websites
  - many research work on user linking

Thank you