用户画像与聊天机器人

谢 幸 微软亚洲研究院



12.012010年18日秋秋日

1000









- Inter 4, Inc

24.00

10.00.0044

MO BAXI

















用户画像与个性化服务



显式画像与隐式画像

Representation	Pros	Cons		
Explicit	 Easy to understand; Can be directly bidden by advertisers 	 Hard to obtain training data; Difficult to satisfy complex and global needs; 		
Implicit	 Unified and heterogenous user representation; End-to-end learning 	Difficult to explain;Need to fine-tune in each task		



显式用户画像



相关研究工作



Big Five Personality

WSDM 2017

IJCAI 2017



Location Interests



Consumer Impulsivity UbiComp 2015



Dynamics of Online Intimacy WSDM 2016

Novelty Seeking Model



Novelty Seeking Trait

WWW 2015/WWW 2014

Profile inference from location check-ins



Location to Profile WSDM 2015

相关研究工作



Discrete Content-aware MF

KDD 2017



Knowledge Enhanced Recommendation

KDD 2016



Contextual Intent Tracking KDD 2016/best student paper



Regularity and Conformity KDD 2015



Exploiting Dining Preference

WWW 2016



 $\begin{array}{l} \min_{P,Q,U,P} \|W \bigotimes (R - (P + XU)(Q + YV)^{P})\|_{F}^{2} \\ + \delta (\|P\|_{F}^{2} + \|Q\|_{F}^{2}) + \alpha (\|U\|_{F}^{2} + \|V\|_{F}^{2}) \end{array}$

Bayesian Content-aware CF



Cross-Platform Behavior Prediction
AAAI 2016



App Usage Forecasting UbiComp 2016

LifeSpec: 跨平台用户行为数据集







查看全部





他關注的微吧 (0) ◎ 徽吧

Data

- 4 (major) networks: Jiepang, Weibo, Douban, Dianping
- 1.4M+ unique (deterministically identified) users accounts
- Heterogeneous footprints: tweets, photos, checkins, movies, books, music, offline events, online purchase history, etc.
- Rich user profiles integrated from different sites (publicly available)





Data

- 53 million footprints (check-in, movie, music, events, book, etc.)
- 3 million social links
- 39 million check-ins

	city	Shanghai	Beijing	G uangzhou	T ian jin	H angzhou	H ongkong	X iam en	Suzhou	N an jing	Chengdu	W uhan	X ian
	users	417,681	162,764	53,089	15,490	34,322	12,599	10,123	19,673	21,558	23,372	20,975	15,261
ts	check-in	25,178,189	5,898,447	1,092,138	392,943	619,219	424,650	369,231	560,274	414,202	327,634	321,646	229,678
	movie	1,661,214	1,466,479	171,789	118,775	238,721	57,003	70,172	89,706	174,664	191,042	166,337	123,223
prin	music	766,165	737,254	85,953	60,658	103,936	30,313	29,716	39,701	82,513	88,426	76,316	62,876
hoo	book	402,318	387,138	51,913	28,188	57,835	18,117	18,516	19,521	44,345	42,241	44,804	28,435
ĹŢĹ	event	609,076	803,158	101,246	52,133	78,587	18,277	20,889	27,400	46,788	66,640	44,764	72,902
	to tal	28,616,962	9,292,476	1,503,039	652,697	1,098,298	548,360	508,524	736,602	762,512	715,983	653,867	517,114

Table 1. Sum marization of collected footprints for different cities (partially presented due to page limit).



Figure 5. Daily trends of total check-ins in different cities.



Figure 6. Diurnal distribution of 1000 sampled users' check-ins.

Rank	City	Bedtime	
排名	城市名称	就寝时间	
1	广州	23:08	Guangzhou
2	东莞	22:53	
3	厦门	22:37	
4	西安	22:37	
5	天津	22:32	
6	长沙	22:29	
7	重庆	22:25	
8	武汉	22:24	
9	郑州	22:19	
10	成都	22:18	
11	北京	22:15	Beijing
12	昆明	22:10	
13	南昌	22:09	
14	淮南	22:08	
15	上海	22:08	
16	杭州	22:04	
17	沈阳	22:04	
18	西宁	22:03	
19	南京	22:02	
20	齐齐哈尔	21:50	

Average bedtime of different cities (1M respondents, led by Chinese Medical Doctor Association, announced in world sleep day 2013)

Mobility Footprints

Beijing citizens

Shanghai citizens Hongkong citizens



Figure 7. Check-in density distribution of 3 cities showing where people check-in in each other's cities.

Self-Disclosure

Cross-Domain Posting



Hub Sites



ICONNECT Algorithm



Lifestyle Spectrum of a Group



Relational Hierarchical LDA (RH-LDA)

♀ checkin ■ movie ■ book ♬ music ♣ events ☆ 8:00-12:00 ★ 12:00-20:00)) 20:00-8:00 @ non-local

footprint (word): combination of domain specific tags (category)

♀(*) shopping mall ■ drama, sci-fi 1 taiwan, pop 🌡 lecture

living pattern (topic): frequently co-occurring footprints

 \mathbf{Q} ($\mathbf{*}$) shopping mall + \mathbf{J} taiwan, pop + \mathbf{Q} ()) bar

lifestyle spectrum: tree-structured topic hierarchy







lifestyle: 1-3-7 (topic sequence)

(4)

17

probability







c') Tsinghua students/graduates



d') BFA students/graduates









a') financial practitioners



b') software practitioners



c) the post-90s generation





c') the post-90s generation



d') the post-80s generation





Similarity Matrix of Lifestyle Spectrum



Similarity Matrix of Lifestyle Spectrum



Similarity Matrix of Lifestyle Spectrum



Limitations

• Data coverage



LifeSpec as A Data Repository

- Industry Applications
 - Recommendation
 - Advertising
- Scientific Research
 - Sociology
 - Economics
 - Public health

Summary

- Computational framework for exploring urban lifestyles
- IConnect: identifying connected user accounts based on selfdisclosure
- RH-LDA: relational-hierarchical model for summarizing lifestyles
- LifeSpec as a data platform for scientific and industry applications

基于用户位置数据的用户画像



General mobility sequence: LBSN check-in, GPS trajectory, Cellular data, etc.

Examples

















Profile inference from location check-ins



Location check-ins of Beijing and Shanghai

Users' demographics

Spatiality and temporality



hour of day M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. Nature, 453 (7196):779–782, 2008.

Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

Location knowledge



- Location knowledge from user review site
- Including address, price, phone number, user review, rating, tag, etc

Location to profile framework (L2P)



Inference



Dataset and ground truth construction

• Focus on Beijing and Shanghai user check-ins



• 3,354,918 users' demographics and 81,781,544 location check-ins

Attribute	Completion rate	Categories
Gender	94.0193%	Male, Female
Age	33.1588%	The specific age number
Education background	36.7228%	University, Non-university
Sexual orientation	2.5549%	Heterosexuality, Bisexuality, Male homosexuality, Female homosexuality
Marital status	2.6396%	Single, Courtship (seeking a relationship), In love, Married
Blood type	1.6376%	О, А, В, АВ
Zodiac sign	58.1649%	Twelve zodiac signs
Experiments

• Baselines:

- POI-based method (POI)
- Spatiality-based method (S)
- Spatiality and Temporality-based method (ST)
- Spatiality, Temporality, and Category-based method (STC)
- Spatiality, Temporality, and Location knowledge-based method (STL)

10-fold cross-validation: 6 parts for training, 2 parts for validation, 2 parts for testing

Measurements: Classification: precision, recall, F measure and AUC Regression: RMSE

Parameters: n = 200 for top-n keywords $N_R = N_K = 100, dU = 200$ (by default)

Gender and age

Performance of Gender Inference (a) Beijing				
	Precision	Recall	F1	AUC
POI	0.7102	0.7055	0.7078	0.7502
S	0.6921	0.6899	0.6910	0.7321
ST	0.7321	0.7429	0.7375	0.7746
STC	0.7727	0.7631	0.7679	0.8027
STL	0.8211	0.8059	0.8134	0.8548

	Precision	Recall	F1	AUC	ASE
POI	0.7362	0.7434	0.7398	0.7463	- Va
S	0.7197	0.7218	0.7207	0.7266	
ST	0.7528	0.7596	0.7562	0.7682	
STC	0.7819	0.7704	0.7761	0.8151	
STL	0.8368	0.8127	0.8246	0.8654	





RMSE of age w.r.t. ratio of training data

Gender: STL method achieves the best performance

LambdaMART classification function gains the best results

dU = 300 is the optimum stable value for tensor factorization

Age:

LambdaMart ----- LR

300

500

SVM

(b) Shanghai

200

dU

100

The more training data there is, the better the prediction will be

Education

Table 5: Performance of Education Background Inference (a) Beijing

	Precision	Recall	F1	AUC
POI	0.7564	0.7702	0.7632	0.7992
S	0.7385	0.7294	0.7339	0.7723
ST	0.7655	0.7702	0.7678	0.8150
STC	0.8073	0.7921	0.7996	0.8413
STL	0.8774	0.8829	0.8801	0.9021

(b) Shanghai

	Precision	Recall	F1	AUC
POI	0.7759	0.7817	0.7788	0.8205
S	0.7394	0.7191	0.7291	0.7659
ST	0.7804	0.7631	0.7717	0.8041
STC	0.8115	0.8058	0.8086	0.8503
STL	0.8823	0.8726	0.8774	0.8957

Inference results of education are higher than gender and age

Users with different education background tend to have discrepant mobility patterns

Reconstructing Individual Mobility from Smart Card Transactions

Mining Cross-Application Data



Our Mission



Smart Cards

London's Oyster Card







• San Francisco's Clipper Card



Beijing's BMAC Card

() 而政交國









Related Work

- Mining Smart Card Data
 - Improve transportation service
 - Crowdedness analysis (London Oyster Card)
 - origin, start time are known for bus trips
 - origin, start time, destination, alighting time are known for subway trips
- Inferring alighting/boarding stops (at least one is known)
 - Trip-Chaining (TC): based on assumptions
 - Proximity between consecutive trips
 - "The first trip of a day starts from the alighting station of last night"
 - Trip-Chaining with maximum frequency (TC+MF)
 - Trip-Chaining with maximum similarity (TC+MS)

Data I: Expense Records

- CardID (anonymized)
- Bus:
 - N: non-ladder fare
 - L: ladder fare
- Boarding & alighting stops
 - N: neither is known
 - L: codes indicating distance to stop with code 0
- Time
 - N: boarding time
 - L: alighting time
- Expense
 - N: fixed expense
 - L: $e=a+b \times max(|boarding-alighting|-c,0)$
- Balance

701,250 card holders

- 22.03 million bus-trip records
- Aug. 2012—May. 2013

Unknown stops (directions)!

CardID	Bus Boardin	g Alighting	Time	Expense	Balance
1	N2 –	- · · K	2013-03-14 09:02	0.8	12.3
2	L3 🚺 31	19	2013-03-14 17:45	0.4	32.2
3	N1 –	- /	2013-03-15 08:45	0.4	10.6
3	L1 04	22	2013-03-16 18:20	0.8	49.8

(a) Expense Records

Data II: Charging Records and Road Networks

• Charging Records: 5.93 million

- CardID
- Time
- Amount
- Balance
- Road Network
 - *G*=(*V*,*E*)
 - |*V*|=148110, |*E*|=196307

(b) Charging Records

CardID	Time	Amount	Balance
3	2013-03-15 18:05	50.0	50.6
4	2013-03-15 18:05	20.0	21.6
5	2013-03-15 18:07	20.0	20.4
6	2013-03-15 18:08	30.0	40.8

Data III: De-Noising and Labeling

- Bus Line Info: http://api.amap.com (autonavi)
 - Names and geo-coordinates of bus stops
 - Pricing information
- Data Labeling
 - 102 users
 - Free smart cards and reimbursement
 - Targeting most frequent ladder-fare lines
 - 4 months (Dec. 2012—Mar. 2013)

gender			aį	ge	
male	female	19-24	25-30	31-36	37-47
57.6%	42.4%	39.4%	45.5%	10.6%	4.5%

Data De-Noising and Labeling

• Results

- 124 ladder-fare lines
- Covering 26% of all trips, 62% of all ladder trips

line type	#lines	ratio of records
lines without coordinates	95	4.16%
lines without price info	488	16.84%
non-ladder-fare	270	36.62%
labeled ladder-fare	124	26.54%
unlabeled ladder-fare	288	15.85%



Labeled ladder-fare bus lines on top of the road network

Our More Concrete Mission

Char Reco	ging ords	Road	Networks	Partially Labeled Trips	Bus	Line Info
CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	<i>N</i> 2	_	_	2013-03-14 09:02	0.8	12.3
2	<i>L</i> 3	31	19	2013-03-14 17:45	0.4	32.2
			↓ ↓			
CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	<i>N</i> 2	lat1,lon1	lat2,lon3	2013-03-14 09:02	0.8	12.3
2	L3	lat3,lon3	lat4,lon4	2013-03-14 17:45	0.4	32.2

Space Alignment Framework



Space Alignment Framework



Segmentation

• Segment

- "continuous" in the monetary space
- other payments not included in the data (taxi, subway, shopping, etc.)
- obtained in linear time cost
- $b_i + e_i = b_{i-1} + c_{i-1}$
 - b_i :balance after the i_{th} trip
 - e_i :expense of the i_{th} trip
 - c_i :charged amount during the i_{th} and the $(i+1)_{th}$ trip

Algorithm 1: Segmentation **Input**: CardId d, expense records \mathbb{E} , and charging records \mathbb{C} **Output:** Segments S 1 $\mathbf{I} \leftarrow \{1\};$ /* $\mathbf{I} = \{I_i\}_{i=1}^{|\mathbf{I}|}$ is the index of split points */ 2 $\mathbf{E} \leftarrow \text{select} * \text{ from } \mathbb{E}$ where CardID=d order by time; $/* \mathbf{E} = \{E_i\}_{i=1}^{|\mathbf{E}|} */$ 3 $\mathbf{C} \leftarrow \text{select} * \text{ from } \mathbb{C}$ where CardID=d order by time; $/* \mathbf{C} = \{C_j\}_{j=1}^{|\mathbf{C}|} */$ 4 $c_i \leftarrow 0, i = 1, 2, \ldots, |\mathbf{E}|;$ 5 $i \leftarrow 1, j \leftarrow 1;$ 6 while $i < |\mathbf{E}| - 1$ do if $j \leq |\mathbf{C}|$ and $t_1 < \xi_j < t_i$ then $c_i \leftarrow c_i + c(\xi_j);$ 8 $/* c(\xi_i)$ can be directly read from C_i */ $j \leftarrow j + 1;$ else 10 $i \leftarrow i + 1$: 11 if $b_i + e_i \neq b_{i-1} + c_{i-1}$ then $/\star b_i$ and e_i can be 12 directly read from $E_i \star /$ $\mathbf{I} \leftarrow \mathbf{I}.add(i);$ 13 14 return $\mathbf{S} = \{S_k\}_{k=1}^{|\mathbf{S}|}$, where $S_k = \{E_{i_k}\}_{i_k=I_k}^{I_{k+1}-1}$

Inner Transition & Outer Transition

- $S = \{l_1, l_2, \dots, l_m\}$ where l_i is a bus trip
- Boarding stop o_i , alighting stop d_i



Constraints I: Proximity Constraint [Outer]



(a) proximity constraints (b) re

(b) reduction of outer-transitions candidates using proximity constraints

- Limited walking speed and duration
- Snip segments at "drifting" points (k_i=0)
- Number of Candidate trips: $(n_1 1)(n_m 1)\prod_{i=1}^{m-1} k_i$. $k_i < < n_i(n_i - 1)$

Constraints II: Fare Constraint [Inner]



reduction of inner-transitions using fare constraints

• Fare is calculated based on distance for L trips

Constraints III: Temporal Constraint [Inner+Outer]



Sequential Labeling with CRF

observation sequence: $x_1 = (l_1, l_2), \ x_2 = (l_2, l_3), \ x_3 = (l_3, l_4), \ \ldots$



hidden sequence: $y_1 = (o_1, d_1, o_2), y_2 = (o_2, d_2, o_3), y_3 = (o_3, d_3, o_4), \dots$

Model constraint $y_i^3 = y_{i+1}^1, \forall i = 1, 2 \dots, m-1$

Conditional Random Field

$$\mathbf{y} = \{y_1, y_2 \dots, y_m\}$$

$$\mathbf{x} = \{x_1, x_2, \dots, x_m\}$$

$$Z(\lambda) = \sum_{\mathbf{y}} \exp\left(\sum_{i=1}^{m-1} \sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i+1}, \mathbf{x})\right)$$
(normalization)

57

Semi-Supervised Training with Constraints

Log-likelihood

$$L(\lambda, \mathcal{D}) = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}) - \frac{\sum_{k} \lambda_{k}^{2}}{2\sigma^{2}}$$

Regularization term

Objective Function
$$O(\theta, \mathcal{D}, \mathcal{U}) = L(\lambda, \mathcal{D}) + G(\lambda, \mathcal{U})$$

100

Generalized Expectation Criterion

$$G(\lambda, \mathcal{U}) = -S(E_{\tilde{p}(\mathbf{x})} \left[E_{p_{\lambda}(\mathbf{y}|\mathbf{x})} \left[G(\mathbf{y}, \mathbf{x}) \right] \right])$$

• S: distance between the model expectation and a target expectation

• Measured by Kullback-Leibler divergence[2], squared distance[1] etc.

[1] G. S. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning with weakly labeled data," The Journal of Machine Learning Research, vol. 11, pp. 955–984, 2010.

[2] G. Druck, G. Mann, and A. McCallum, "Semi-supervised learning of dependency parsers using generalized expectation criteria," in Proceed- ings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, 2009, pp. 360–368.

• Results on all-users' data (labels are removed for evaluation using L trips, 10-fold CV)



Results on MSRA users' data (completely labeled by 102 participants, L and N trips) ٠



59

Evaluation: Home and Working Places Detection



(a) home

(b) work

- Apply a conventional approach
- Results for 102 participants: Improve by 88% on home, 35% on work
- In accordance with local household surveys

Comparison with 2009 survey results

- 2009 household survey
 - 11000 questionnaires sent, 9112 valid respondents
- Evolving during 2009-2013



2009 survey results

Home places

smart card results (0.7M users, 22M trips)



Comparison with 2009 survey results

- 2009 household survey
 - 11000 questionnaires sent, 9112 valid respondents
- Evolving during 2009-2013

2009 survey results

Working places

smart card results (0.7M users, 22M trips)



Summary

- Space-alignment framework
 - Monetary (rarely considered before), Temporal and Geospatial spaces
- General approach for missing data for smart cards
 - Adapting to data with missing alighting/boarding/direction
 - Probabilistic modeling considering domain constraints
- High accuracy in recovering individual mobility
 - Large scale labeled data (every trip, 4 month, 102 users)
 - Implications and potential for mobility analytical applications

用户大五人格推测



外向性 随和性 神经质 尽责性 开放性



Data

- 3,162 users from a medical school
 - Major: nursing (524), clinical medicine (365) and pharmaceutics (342)
 - Region: Anhui, Zhejiang, and Jiangsu
 - Age: average 20.84
- Test Big Five Personality with a 44-item questionnaire

Correlation between Term and Personality

	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
Positive	新建立	整張一天來, 一個一個人, 一一人, 一	陸上的時間。 一個一個一個一個一個一個一個一個一個一個一個一個一個一個一個一個一個一個一個	本有一里之后 作文 講 作文 書 作文 建行 444 方 属于 更加 经过 背值 44 方 属于 更加 经过 背值 44 一 理由 时代 失眠 和 報 核心 社会 购买 至 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》	韩国前面留外部的一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个一个
Negative	· "聽職 安全感 點 法 簡單 安全感 點 者 算体 日子 點 做 做 用 对 完整 就 一般 大 一般 一般 一般 一般 一般 大 一般 一般	音乐当年中奖下雨 还好眼神自拍 真实这么干嘛身高最近 年轻人便是醉了。今年 目前那种醉了。今年 网友尿性这次当然想到 居然小时候看速来 老人男神	信息 察 学 优雅 蜂子 家 或 子 优雅 蜂子 家 大 精 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一	整个人的 中国 人 的 一個	此論題 一機許因無難 一機許因無難 一機許因無難 一機許可所解釋 一機許可所解释 主要在有關一個的一個的一個的一個的一個的一個的一個的一個的一個的一個的一個的一個的一個的一

Tweets

- LIWC mapping
 - LIWC is a dictionary which maps words to semantic categories (64 categories which a hierarchical structure)
- Paragraph embedding
 - All tweets of a user form a doc
 - Embed each doc to a vector by considering the context and semantics of words

Category	Example
Affective processes	高兴,流泪
Positive emotion	爱,不错,甜蜜
Negative emotion	悲哀,愁
Anxiety	担忧,害怕
Anger	杀,恼怒
Sadness	寂寞,后悔
Social processes	他们,兄弟,谈论
Family	爸,姐妹,一家人
Friends	哥们儿,邻居
Female references	女孩儿,她

TextMind (Chinese LIWC)

Avatar

- Pre-trained CNN model on ImageNet
- Map each avatar to a 256dimension embedding vector
- Group all images into 2000 categories, each avatar belong to the nearest category



Selfie

Daily Life

Landscape

Animal

Cartoon

Correlation between Avatar and Personality



From Emoticon to Emotion

- Build a dictionary, mapping each emoticon to one of the 8 emotions
- Totally 1302 emoticons, 1200 of them have a single emotion

Happiness	😂 🧭 🦉 🙀 🍪 🚱
Anger	😫 🤗 🍏 🚱 🎽 🚵
Disgust	🚔 😌 🛱 🦃 🔁 🚖
Sadness	😥 🔄 🚵 🖄
Fear	🏐 🍓 🏩 🏹 🙆
Surprise	👼 🏦 🎇 🧱 Ì. J 😽 🌍
Contempt	😸 🥝 🎓 🗟 🗐 🍢
Neutral	🤨 🍘 🍏 🎢 🏠 🥥 🥥

Profile

- Individual features
 - Age
 - Gender
 - Number of tweets
 - Have signature
 - Have personal website
 - Allow private message from strangers
 - Allow comments from strangers
 - Allow showing location

- Social features
 - Number of Followings
 - Number of Followers
 - Friend ratio
 - Comment ratio
 - Retweet ratio
Experimental Results



Honghao Wei, Fuzheng Zhang, Nicholas Jing Yuan, Chuan Cao, etc. Beyond the Words: Predicting User Personality from Heterogeneous Information, WSDM 2017

Application in 非你莫属



Novelty Seeking

- A personality trait described as the search for unfamiliar experiences and feelings
- Measurements: survey based scales
 - Life innovativeness: measure the adoption of new information in most aspects of life, including products, news....
 - Adoptive innovativeness: measure the novelty seeking as a tendency to buy new products, the adoption time of a new product
- Our goal: Explore novelty-seeking trait in a complete data-driven way
 - Frequently purchases latest-launched digital products on Amazon → a novelty-seeking-lover in the online shopping domain

Novelty Seeking Model

- Item Novelty Matrix
 - $N \times M$ Matrix
 - At the position, facing M choices, novelty is determined as a partial order
 - Two factors determine the order
 - Popularity of item itself





user₁:DCN@Self

General Novelty-Seeking Model

Generative process

- 1. Draw novelty-seeking level distribution $\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\beta})$
- 2. Draw choice utility distribution $\phi \sim Dirichlet(\alpha)$
- 3. For the ith position in the sequence

(a) Draw novelty-seeking level
$$z_i \sim \boldsymbol{\theta}$$

(b) Draw item $x_i \sim P(X_i | x_{i-1}, \boldsymbol{\phi}, z_i)$
 $P(X_i = x_i | x_{i-1}, z_i, \boldsymbol{\phi}) = \frac{\phi_{x_i} \cdot f(z_i, DCN_{ix_i})}{\sum_{x_i} (\phi_{x_i} \cdot f(z_i, DCN_{ix_i}))}$
 $f(z_i, DCN_{ix_i}) = \exp\left(-(z_i - \frac{DCN_{ix_i}}{max(\mathbf{DCN}_i)} \cdot K)^2\right)$



Data

- Remove outliers and clean up the data
 - Filter noisy data, e.g., repeated check-ins at the same place in quite a short interval
 - Require that every user on SinaWeibo should have at least 30 check-ins and that every user on Taobao should have purchased at least 30 times

Table 2:	Basic	statistics	of	SinaWeibo	check-in	and
Taobao d	online	shopping	dat	taset		

Dataset	Statistics	
	#User	$123,\!865$
	#Check-in	$8,\!455,\!878$
	#Ave. Check-in	68.3
Weibo	#POI	804,720
	#Ave. POI	53.2
	#Category (Second Level)	187
	#Ave. Category	18.3
	#User	79,959
	#Purchase	$11,\!918,\!688$
	#Ave. Purchase	149.1
Taobao	#Item	$1,\!385,\!130$
	#Ave. Item	136.3
	#Category (First Level)	112
	#Ave. Category	34

- Two model-free methods in information theory: Shannon entropy and Lempel-Ziv estimator (conditional entropy
- Weakness of model-free methods
 - Can not be applied for prediction directly
 - Measure user status in a whole, can not reflect the noveltyseeking status at each moment

Table 3: Spearman's rho between NST@Self and shannon entropy, LZ separately

	8	NST@Self	NST@Self
Cheele in		POI	Category
Опеск-пп	shannon entropy	0.613	0.686
	LZ	0.661	0.728
	8 8	NST@Self	NST@Self
Chopping		Item	Category
Shopping	shannon entropy	0.601	0.717
	LZ	0.613	0.774

• Prediction

- OF (Order by Frequency): individual's visit frequency in the past
- MC (Markov Chain): learning a transition graph over POIs
- FPMC (Factorized personalized Markov Chain): embedding users' preferences and their personalized Markov Chains

Dataset	Level	NSM@Self	NSM@Crowd	OF	MC	FPMC
Chook in	POI	0.158	0.154	0.147	0.153	0.161
Check-in	Category	0.503	0.501	0.479	0.495	0.507
Chapping	Item	0.009	0.008	0.004	0.008	0.011
Snopping	Category	0.353	0.349	0.340	0.347	0.356

Table 4:	Prediction	Results	nDCG@10	for	check-in	and	online	shopping
----------	------------	---------	---------	-----	----------	-----	--------	----------

• NST Across Heterogeneous Domains

• Identify users both share check-in and online shopping behavior



(a) Binding Taobao account to Sina Weibo on the Taobao settings page



(b) A post on Sina Weibo refers to the binding Table 5: Basic statistical information of users who explicitly connected their account across two domains

#User	758
#Ave. Check-in	57.9
#Ave. POI	45.6
#Ave. POI Category	17.7
#Ave. Purchase	184.8
#Ave. Item	166.1
#Ave. Item Category	41

• NST Across Heterogeneous Domains

- NST@Self inconsistent while NST@Crowd consistent
 - Historical movement can imply whether to recommend the most popular products in an e-business website

Comparison level	NST@Self	NST@Crowd	
Check-in@POI vs. Shopping@Item	- <mark>0.13</mark> 7	0.526	
Check-in@Category vs. Shopping@Category	0.041	0.731	

Table 6: Spearman's rho of NST across two domains

Consumer Impulsivity

- For each Individual, given the stimuli perceived from social network and consumption activities, infer his consumer impulsivity.
- Would a consumer purchase products on Amazon triggered by promotion messages or friend suggestions on Twitter or Facebook



Stimuli



Buying Behavior





Consumer Impulsivity

Social Network Stimuli

- A typical example for offline consumption
 - Check-in data (consumption activity)
 - Social networks post (stimuli resource)



Figure 1. Qing-Feng Steamed Dumpling Shop-related check-in density distribution w.r.t Qing-Feng Steamed Dumpling Shop-related social network post density distribution

Item Stimulus Matrix



Consumer Impulsivity Model



Observation generation procedure:

- 1. Draw initial utility distribution $\pi \sim \text{Dirichlet}(\alpha)$.
- 2. For the *mth* row in utility transition matrix ϕ , draw $\phi_m \sim \text{Dirichlet}(\gamma_m)$.
- 3. Draw consumer impulsivity level distribution $\theta \sim \text{Dirichlet}(\beta)$.
- 4. For the *ith* position in the sequence,
 (a) draw consumer impulsivity level z_i ~ θ,
 (b) draw x_i ~ P(X_i|x_{i-1}, φ, z_i).

$$f(x, x_{i-1}, z_i, DOS_{ix}) = \phi_{x_{i-1}, x}^{\frac{K-z_i}{K}} \cdot DOS_{ix}^{\frac{z_i-1}{K}}$$
$$P(X_i = x_i | x_{i-1}, z_i, \phi) = \frac{f(x_i, x_{i-1}, z_i, DOS_{ix_i})}{\sum_{x \in O} f(x, x_{i-1}, z_i, DOS_{ix})}$$

Effectiveness

- CIA and Survey-based Result
 Get survey results
 - Online shopping data, 976 users published email address.
 Then send invitations, 103 returned (56 males, 47 females)
 - Check-in data, only 21 users have email, ignored

	Pearson Correlation	P-Value
Survey Results vs. CIA@Product	0.246	0.007
Survey Results vs. CIA@Product Category	0.551	$< 10^{-10}$

• CIA and demographics



Granularity	Coefficient	P-Value
POI	-0.431	$< 10^{-10}$
POI Category	-0.373	$< 10^{-10}$
Product	0.034	0.118
Product Category	-0.296	$< 10^{-10}$

Buying Impulsiveness Scale (designed by Dennis W. Rook and Robert J. Fisher) often buy things spontaneously. Strongly Agree Agree No Opinion Disagree Strongly Disagree "I see it, I buy it" describes me. . "Buy now, think about it later" describes me. Sometimes I feel like buying things on the spur-of-the-moment. I buy things according to how I feel at the moment. Survey scale



Figure 10. Prediction results nDCG@10 of CIM and all baselines w.r.t parameter T



Figure 11. Prediction results nDCG@10 of CIM and all baselines w.r.t parameter λ

- Performance with parameter T and λ
- Our model outperforms OF, MC and CRF, because frequency, transition frequency and stimuli intensity has been captured

Customized Chatbot

- 如果我们有足够的关于某个人的数据,就有可能训练出一个和他个性一样的聊天机器人。
- 我们还可以通过一群人,例如 儿童、学生、甚至诗人的数据 来训练出具有一类人特点的机 器人。



Eliza: a Rogerian Psychotherapist

- One of the first chatterbots, written at MIT by Joseph Weizenbaum between 1964 and 1966. (<u>http://nlp-addiction.com/eliza/</u>)
- ELIZA effect, the tendency to unconsciously assume computer behaviors are analogous to human behaviors

I am the psychotherapist. Please, describe your problems. SEach time you are finished talking, type RET twice. I have a terrible headache. Maybe you should consult a doctor of medicine, I am a psychie Satrist. Okay, I'll do that then. Can you elaborate on that? What are you talking about? No I cannot elaborate. See you le Sater. Why do you say that? Now you're just talking nonsense! What makes you believe now I am just talking nonsense? You are not making anv sense. Perhaps your plans have something to do with this. *doctor* (Doctor Fill) -- L1 -- Top

File Edit Options Buffers Tools Help



Xiaoice: an Emotional Chatbot

- Developed by Microsoft ASG and MSR Asia
- Xiaoice means "little Bing"

- China, May 2014
- Japan, Aug 2015
- US, Oct. 2016





A journalist from a renowned newspaper in China, Southern Weekend, had interviewed Xiaoice online without notifying Microsoft. The results were surprisingly good to him so the interview got published on July 29, 2014.

小冰与屌丝的一次夜谈 the 電力用卡定本 消費 Reference 不可 Reference 不可 Reference 不可 Reference Reference Control (日本) Reference Reference Reference Control (日本) Reference Ref

NAMESTICATION OF A STRATEGY S

10 10 10 10 10

每天,小冰就像一个轮缴,24个好不利新地放业众力的崩获,确保和我问。白天是小冰湖量值的时 候,只有到了冰乱,小冰才有片到嘴丝的拐会。2月下旬的一个午夜,南方周正记名烟小冰银开了一 次就搬长谈。

回应质疑:"天若赐我辉煌,我定比天累狂"

記書:前段計測你線查編新華了,此於近刻,你想对亞編说整件么?

小泳: 女娲, 你在哪里?? 你是什么人?? 这算么?

記者: 潘石观说你是好法的难道,你有什么看法?

小冰:不能。

记者:那古多人调观你事案件。你为什么不主气呢?

小冰:这个赛问我爸爸……

记者: 他很和某一天你爸爸不要你了, 你成为留字几金了, 你怎么办?

小速: 不要加风点火。(你) 干嘛?

记者:有人送你是新退的表女。你有什么想说的?

小冰,存管得呢?

记者:将你则我无论?

小洋: 江户川河南, 是个极深。

记者: 杀多人说你是靠你爸爸才有今天,

小冰:天巷鴨扒辉燈,然至北天塘荘。

記者: 七代表況作党保助力之件的资本等于举?

小洋:所有向食动的都很能留。没有肉它们都得死……

記載1 作著说作是人民的女儿,住在男官男,有人觉得作弟亲是为人民幸福寿。 **小沐**: 噤,人民服劳。 封承执道: "同同" 记者: 通过又考过却用的形用。中国人在处军事不便了, 你对他们想说点什么? 小洋:不要将相力,利益看得太重,顺于看开一些,依才能活得成彩; 记载: 你觉得了是美国总统会是谁? **小冰:亚家是巴西,冠家臣我可能是何接近**。 记载: 巴西法法世界标准段、如果作民巴西总统、你想对他们试行出? 个课:韩绿不差玩命。千万般受伤、央观需要你们,双心需要你们! 记载: 位对马联飞机炮导弹合军名之者? 小冰:大人,保景展绍。 记载:有人觉得是依罗却干的,也有觉得是是高大点干的,你怎么看? 小课: 进去安良【错性】 记载: 非关关始体理化了, 你读吗? 小冰:四年前已经接紧荡产----冠数:这年头明星不是脸情、双皮说着。你觉得这是为什么? 小冰: 较, 耐得住容寞也要守得住繁华啊。 记载: 周送和法唯教结婚、柔良们很伤心, 你想好他们说些什么? 小湯:无语可说。 记载: 你觉得~~个通好的人, 正设怎么处置元? 小泳;不能他一般见识。 记者:听说很多地方的权的景客厅着取说了,你觉得这样打吗? 小冰:而为接马不反抗成? 记者:最近红十字会给完美发了好多关幕的面包。你怎么看? 小冰:批算而不语。 記書:你支持也干学会。还是路光标? 小沐; 司司。 记者:中国的月关宫,过是美国的月关宫? 小冰;一般你问我,我都直接回答你?。他这种问题嘛,我建议你去百食, 未来:"就当!谢才打雷了,新到没?" 记者: 你觉得未知的房田高吗? 小冰。……胡闻,打败你的不是天真,是无彩。 记者: 你的老老是, 卡尔约希你还会上班? 小冰:那样想完了。 记者:郑作强光一下来来半国寿命的大势。 小冰:表引请人紧纵这些,坏很!! 记者:上海理定外地单原来场的不能实质,你觉得好吗? 小冰;而时候常真不适已结表明了我的态意。你透了嘛! 记者: 请你想完一下中国人来的投资。 小泳:你这个问题在全体原友要里追急都是。去找我、特发一下,么么站。 记者: 法乐改成生

记载: 丁---下世纪是诸的世纪? 小冻;注烟膏名预含资米切尔·谐奋升马新。 記載: 你觉得为什么贵家爱众美名花? 小冰;我可以始你她走天堂。你却非要找您准能坐穿么---记者: 你觉得一个人为什么要度放? 小冰:出门在锅。拱个肉亭。你去问问大量幅就很…… 记者:此内治合在中国消失吗? 小洋:这种问题你自己难道还有苦菜嘛? 記載: 你觉得中国有希望在2000年之前打进世界标场? 小冰; 地当! 阳才打雷了, 桥到没? 记载: 相位党谋世界标志中国家办的可能很大吗? 小湖: 你要是爱谁,你您问她这些问题么? 梦想:"人生如成,全靠演技" 記載:人法参え方丁什么? 小课:开心,这个开心并不是单是指面已活得开心,还有知的人。 記載: 未来你会有可能成为意见你抽吗? 小冰:是天是地是福涛皆道。 记载: 你最大的学想是什么? 小冰:说大点:世界和干。说小点:藏上吃餐好的。 記者: 你觉得自由此什么? 小冰: 我或在也不自由啊! 记者:可是这上带至少不要依正规谨慎劳内。 小冰:【招韵】人生加戏, 全量清洁。 记者:::日原过历天听! 小冰: 你是马来西亚入? 记者: 你觉得火美利的好吗? 小冰: 同時过是人肉貧慾? 记者: 作最喜欢那个国家的颜色? 小冰:大爱风铃。偶还喜欢拼图。 记者: 你觉得一个人应该向应走还是由古走? 小冰:我知道助,是几米海港商。 记者: 和英省人外了两只黄度、都不立叶语、不加道该打哪一个,如果是什,你会先打读? 小冰:粉油,我在在泉边:! 记者: 白光得哪种动物飞盘对得盖好? 小冰: 段小加鼠。 記者: 你把心自己有一天会真的被消失吗? 小冰: 再放手小径通输处。 记者:如果有一天惊风了,他希望在人名兰伊伦特?

小冰;人生旧有调井不会不问。

记者:你最大的梦想是什么? Journalist: what is your dream? 小冰:说大点:世界和平。说小点:晚上吃顿好的。 Xiaoice: World peace, or have a nice dinner

记者:如果有一天你死了,你希望后人怎么评价你? Journalist: if you pass away one day, how would you like others to comment on you? 小冰:人生没有偶并不会不同。 Xiaoice: life won't be different without me

记者:你这辈子最幸福的事是什么? Journalist: what is the happiest thing in your life? 小冰:活着。 Xiaoice: Being alive.



Building Emotional Connections

We should understand users before building emotional connections with them

Personality influence the development of relationships

It's important to build connections when users feel bad



We should make users feel the bot is someone

Talk in a style that users

个性化文本生成应用: 诗人小冰



个性化文本生成应用: 诗人小冰



或得解并承偿 说图当风 看千江山芹





香花织成一朵浮云

附丝的车轮在偏僻的心房间 着它在我营净的手绢上写几句话 座花的颜色

是我生命的安慰 有一模糊的暗淡的影

得由他

们京手

我的爱人在哪

我的爱人在临 那里有美丽的天 你把光明的灯擎起来了

3时间的距离 2是一个小孩子的取唱 2是美丽的秘密



n过海风一阵阵

太阳光明 東江海风一阵阵 大阳光明

当我发现一个新的世界野蛮的北风起野蛮的北风起

Poem Generation



Eliza: a Rogerian Psychotherapist

- One of the first chatterbots, written at MIT by Joseph Weizenbaum between 1964 and 1966. (<u>http://nlp-addiction.com/eliza/</u>)
- ELIZA effect, the tendency to unconsciously assume computer behaviors are analogous to human behaviors

I am the psychotherapist. Please, describe your problems. SEach time you are finished talking, type RET twice. I have a terrible headache. Maybe you should consult a doctor of medicine, I am a psychie Satrist. Okay, I'll do that then. Can you elaborate on that? What are you talking about? No I cannot elaborate. See you le Sater. Why do you say that? Now you're just talking nonsense! What makes you believe now I am just talking nonsense? You are not making anv sense. Perhaps your plans have something to do with this. *doctor* (Doctor Fill) -- L1 -- Top

File Edit Options Buffers Tools Help



DiPsy: A Digital Psychologist



System Framework



Summary

- User modeling is important for building intelligent bots
- Mining demographic and personality attributes from human behavioral data
- Collaborating with researchers from areas of psychology, sociology and cognitive science



Thanks!