面向移动商务智能的数据挖掘方法研究

陈 恩 红

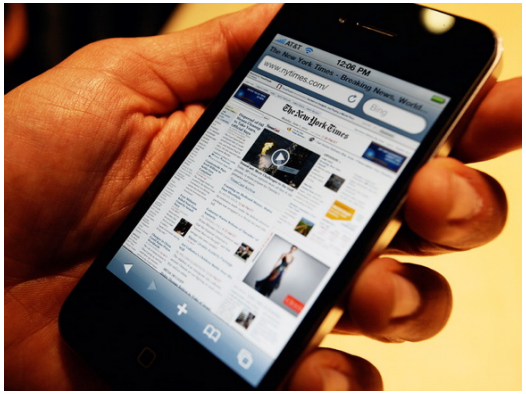
中国科学技术大学 计算机学院

内容提纲

内容提纲

移动商务智能概述

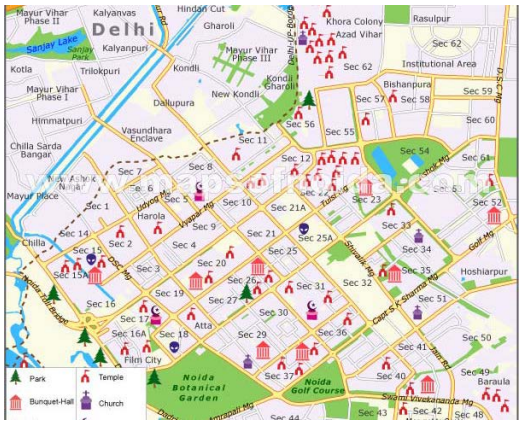
- 近年来，移动互联网产业迅猛发展，根据中国互联网信息中心2013年发布的《中国互联网络发展状况统计报告》，2012 年是中国移动互联网市场爆发式增长的一年，移动网络从 3G 向 4G 升级，移动设备用户数超越了台式电脑数，我国手机网民规模已超过 4.2 亿，占整体网民数量的比例74.5%。
- 各种移动应用和智能服务，如智能移动应用程序，基于位置的服务等，在规模上同样出现了爆炸式增长。



移动商务智能概述

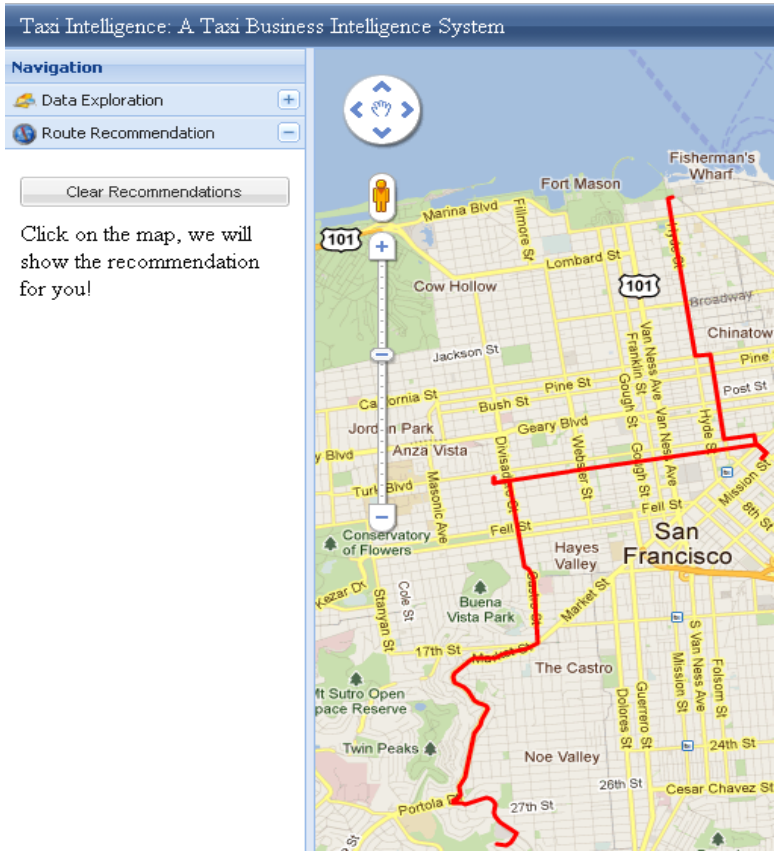
- 移动商务智能为移动商务服务及应用的开发、决策、运营等提供智能分析与挖掘方法

- 典型的移动商务智能：
  - 移动推荐系统
  - 移动用户行为分析
  - 移动城市计算



移动商务智能—移动推荐系统

出租车载客路线推荐

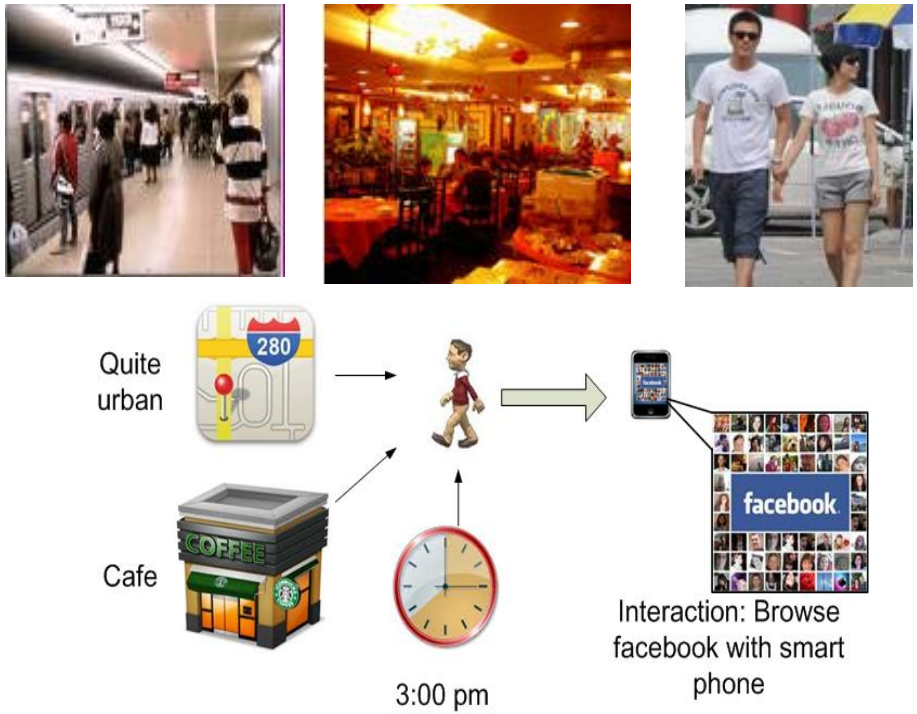


移动用户团购推荐

移动商务智能—移动用户行为分析

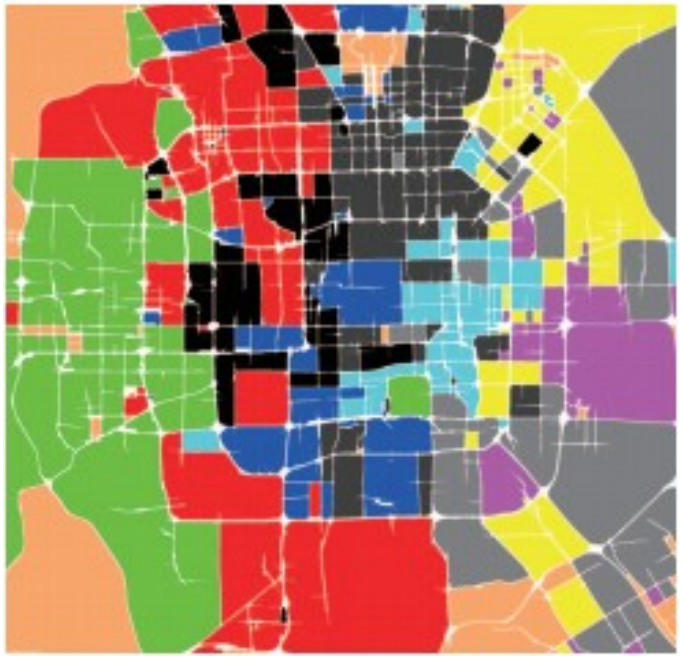
移动用户重要地点发现

移动用户行为模式分析

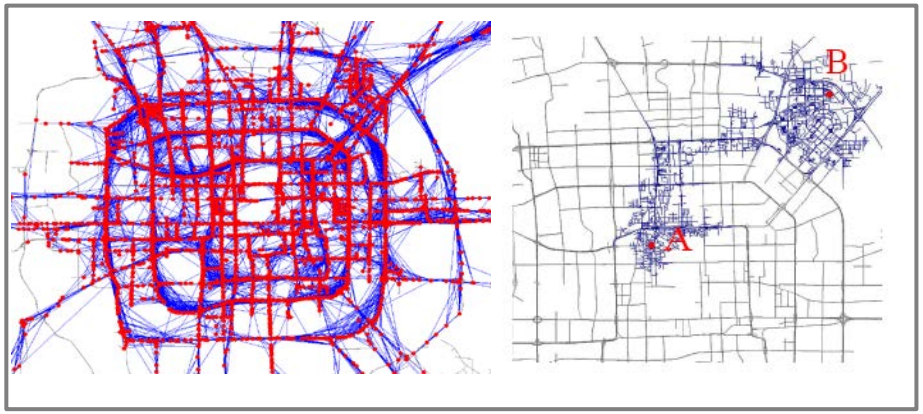


移动商务智能—移动城市计算

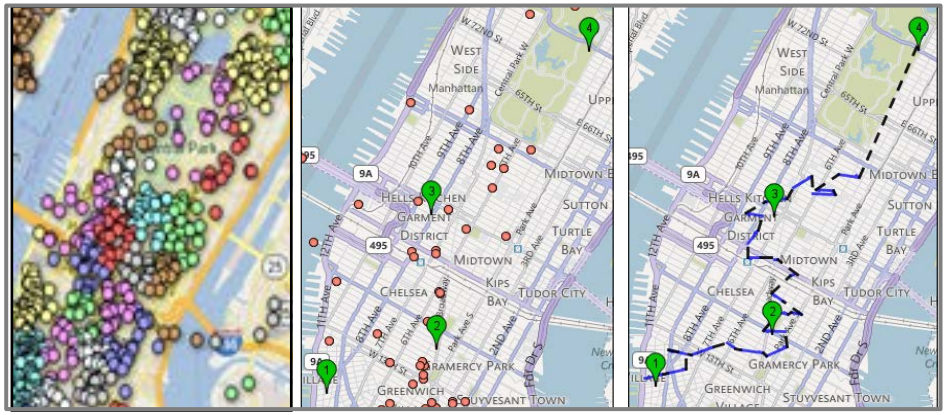
城市功能区域划分



智能行驶路线推荐



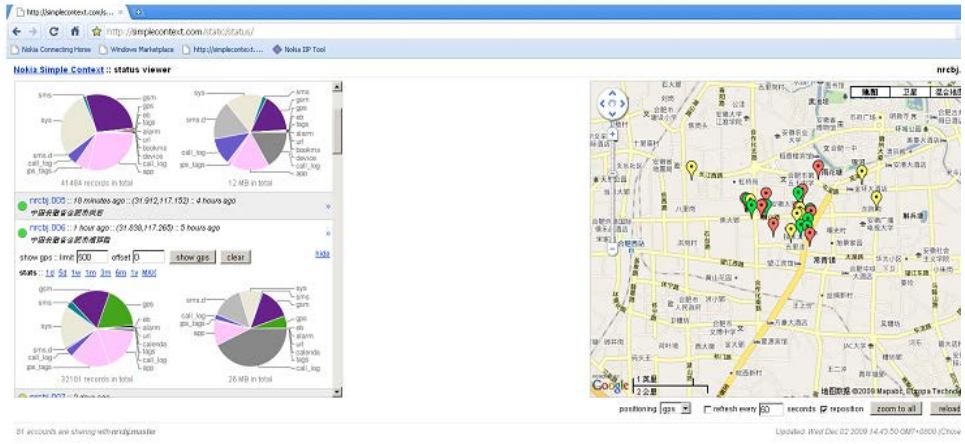
移动线路重构



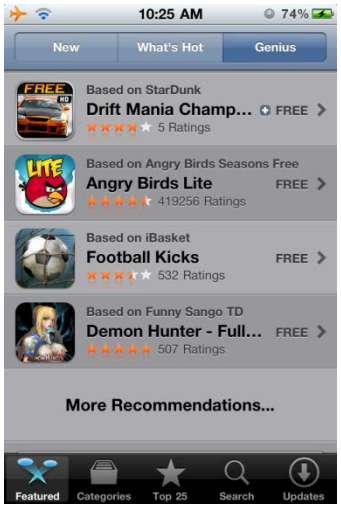
移动商务智能的机遇---数据

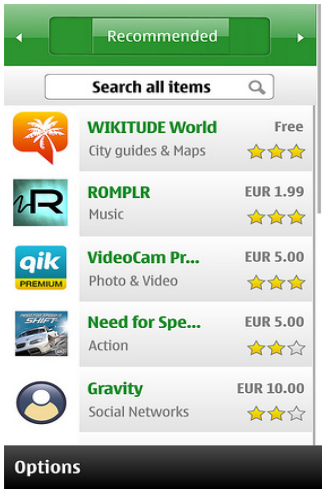
移动用户情境数据

Timestamp	Context	Activity records
t_1	{(*Day name: Monday*),(*Time range: AM8:00-9:00*)), (*Profile: General*),(*Battery: 5*),(*Location: Home*)}	Null
t_2	{(*Day name: Monday*),(*Time range: AM8:00-9:00*)), (*Profile: General*),(*Battery: 5*),(*Location: On the way*)}	Play action games (*Fruit Ninja*)
t_3	{(*Day name: Monday*),(*Time range: AM8:00-9:00*)), (*Profile: General*),(*Battery: 5*),(*Location: On the way*)}	Null
	....	
t_{359}	{(*Day name: Monday*),(*Time range: AM10:00-11:00*), (*Profile: Meeting*),(*Battery: 4*),(*Location: Work place*)}	Null
t_{360}	{(*Day name: Monday*),(*Time range: AM10:00-11:00*), (*Profile: Meeting*),(*Battery: 4*),(*Location: Work place*)}	Browsing sports web sites (*www.nba.com*)
	....	
t_{448}	{(*Day name: Monday*),(*Time range: AM11:00-12:00*), (*Profile: General*),(*Battery: 4*),(*Location: Work place*)}	Play with SNS (*Facebook*)
t_{449}	{(*Day name: Monday*),(*Time range: AM11:00-12:00*), (*Profile: General*),(*Battery: 4*),(*Location: Work place*)}	Null



移动应用商店数据

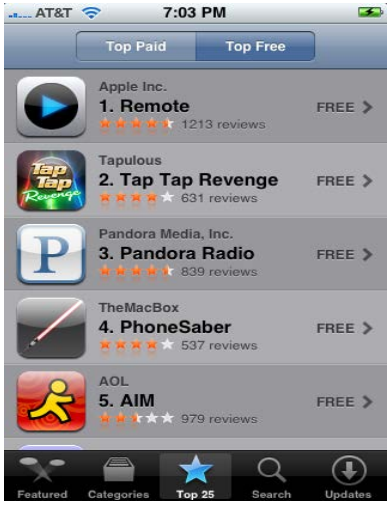




内容提纲

移动应用商店排名欺诈检测

- 随着移动技术的飞速发展，各种移动应用和服务已进入人们生活、娱乐与工作等方方面面。为方便用户选择，各种移动应用商店都提供了应用排行榜，比如“免费排行”，“收费排行”等。
- 在排行榜上名列前茅的应用不仅会收获大量用户，往往也会获得丰厚的商业利润。因此，应用开发商都希望自己的应用能够排到榜单前面。
- 一些厂商雇用商业公司，通过“机器人”或者“水军”来刷榜，即移动应用商店排名的恶意欺诈。



移动应用商店排名欺诈检测

- 移动应用商店的排名欺诈引起了工业界的高度重视，苹果因此封杀了众多移动应用厂商。
- 如何检测这些排名欺诈，具有很多挑战：
  - 刷榜行为并不是长时间的，而是在某些时间段实现。→挖掘活跃周期
  - 移动应用数量众多且快速增长，需要设计自动的增量式方法实现
  - 移动排行具有高度的动态性，需要寻找刷榜行为证据

1. Hengshu Zhu, Hui Xiong, Yong Ge, Enhong Chen, **Ranking Fraud Detection for Mobile Apps: A Holistic View**, In *Proceedings of the 22nd ACM Conference on Information and Knowledge Management* **(CIKM 2013)**, San Francisco, USA, 2013, Accepted.

移动应用商店排名欺诈检测

- 解决框架：
  - 1，挖掘活跃周期➔问题转换为验证一个活跃周期是否可疑
  - 2，基于排名的欺诈证据提取
  - 3，基于评分的欺诈证据提取
  - 4，欺诈证据整合，实现欺诈检测

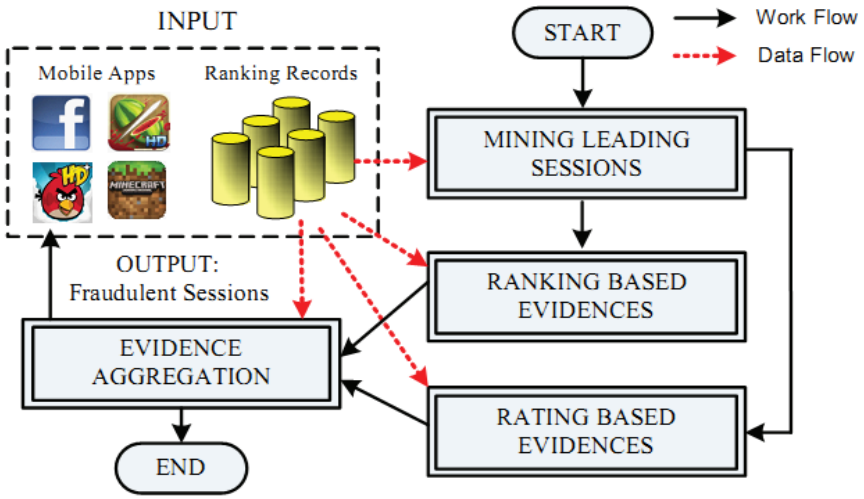
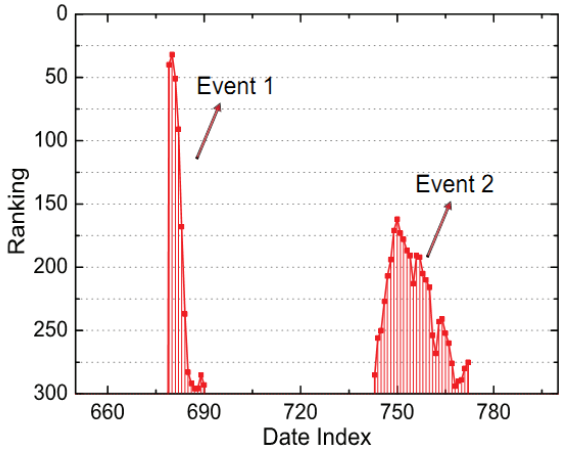


Figure 1: The framework of the ranking fraud detection system for mobile Apps.

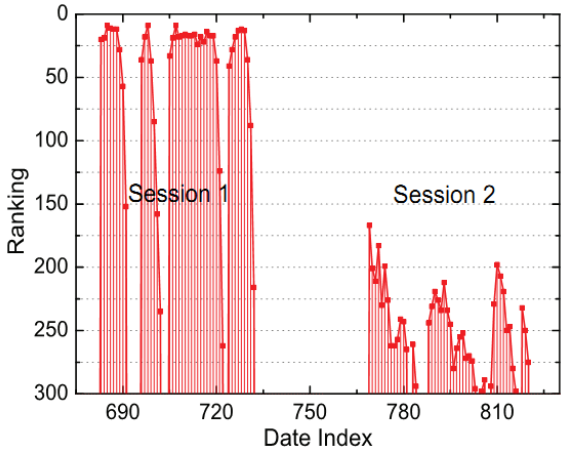
移动应用商店排名欺诈检测

- 挖掘活跃周期(Leading Session)，实现刷榜的准确定位

活跃事件



活跃周期



Algorithm 1 Mining Leading Sessions

Input 1: a's historical ranking records R_a;
Input 2: the ranking threshold K^*;
Input 3: the merging threshold ϕ;
Output: the set of a's leading sessions S_a;
Initialization: $S_a = \emptyset$;

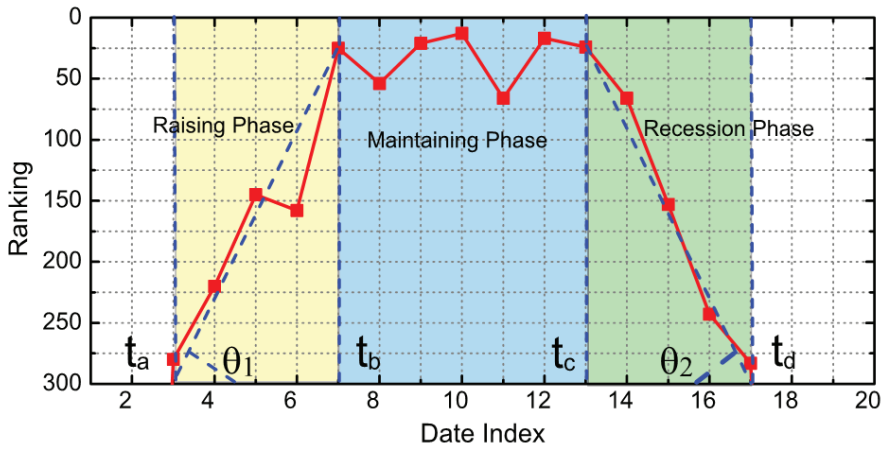
1: $E_a = \emptyset$; $e = \emptyset$; $s = \emptyset$; $t^e_{start} = 0$;
2: **for each** $i \in [1, |R_a|]$ **do**
3: **if** $r^a_i \leq K^*$ **and** $t^e_{start} == 0$ **then**
4: $t_{start} = t_i$;
5: **else if** $r^a_i > K^*$ **and** $t^e_{start} \neq 0$ **then**
6: //found one event;
7: $t^e_{end} = t_{i-1}$; $e =< t^e_{start}, t^e_{end} >$;
8: **if** $|E_a| == \emptyset$ **then**
9: $E_a \cup = e$; $t^s_{start} = t^e_{start}$; $t^s_{end} = t^e_{end}$;
10: **else if** $|E_a| > 1$ **and** $(t^e_{start} - t^{e^*}_{end}) < \phi$ **then**
11: //e^* is the last leading event before e in E_a;
12: $E_a \cup = e$; $t^s_{end} = t^e_{end}$;
13: **else then**
14: //found one session;
15: $s =< t^s_{start}, t^s_{end}, E_a >$;
16: $S_a \cup = s$; $E_a = \emptyset$; $s = \emptyset$ is a new session;
17: **go to** Step 7;
18: $t^e_{start} = 0$; $e = \emptyset$ is a new leading event;
19: **return** S_a

挖掘活跃事件

挖掘活跃周期

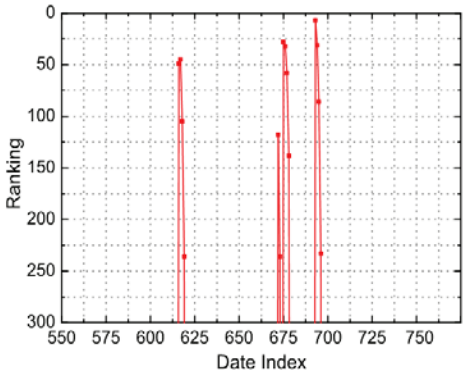
移动应用商店排名欺诈检测

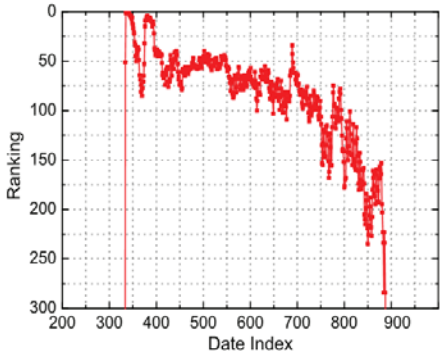
- 提取基于排名的欺诈证据



一个活跃事件内的应用排行可以分成三个阶段：上升阶段，维持阶段，下降阶段。正常应用和可疑应用在这三个阶段的特征上具有显著区别

一个可疑应用





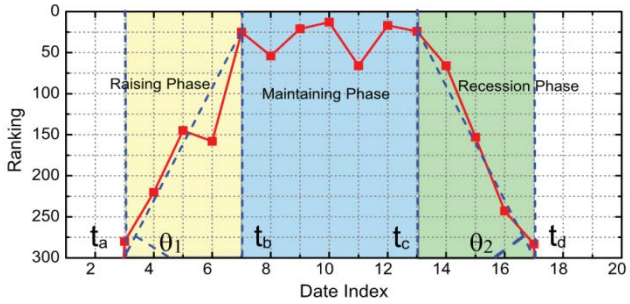
一个正常应用

移动应用商店排名欺诈检测

- 基于排名的欺诈证据1：上升和下降阶段的速率

$$\theta_1^e = \arctan\left(\frac{K^* - r_b^a}{t_b^e - t_a^e}\right), \quad \theta_2^e = \arctan\left(\frac{K^* - r_c^a}{t_d^e - t_c^e}\right).$$

$$\overline{\theta}_s = \frac{1}{|E_s|}\sum_{e \in s}(\theta_1^e + \theta_2^e),$$



- ▷ HYPOTHESIS 0: *The signature $\overline{\theta}_s$ of leading session s is not useful for detecting ranking fraud.*
- ▷ HYPOTHESIS 1: *The signature $\overline{\theta}_s$ of leading session s is significantly greater than expectation.*

采用假设检验框架

$$\overline{\theta}_s \sim \mathcal{N}(\mu_{\overline{\theta}}, \sigma_{\overline{\theta}}),$$

$$\mathbb{P}(\mathcal{N}(\mu_{\overline{\theta}}, \sigma_{\overline{\theta}}) \geq \overline{\theta}_s) = 1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\overline{\theta}_s - \mu_{\overline{\theta}}}{\sigma_{\overline{\theta}}\sqrt{2}}\right)\right),$$

采用高斯假设估计来计算证据得分

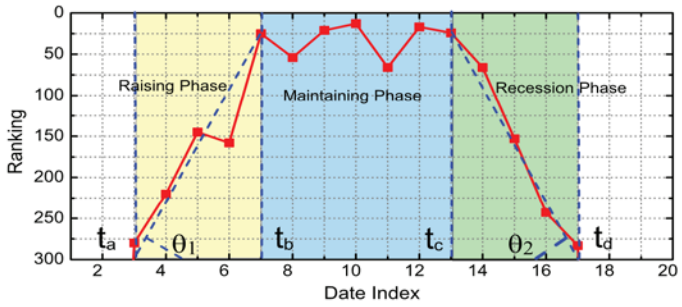
$$\Psi_1(s) = 1 - \mathbb{P}(\mathcal{N}(\mu_{\overline{\theta}}, \sigma_{\overline{\theta}}) \geq \overline{\theta}_s).$$

移动应用商店排名欺诈检测

- 基于排名的欺诈证据2：维持阶段的持续时间和排名高低

$$\Delta t_m^e = (t_c^e - t_b^e + 1)$$

$$\chi_s = \frac{1}{|E_s|} \sum_{e \in s} \frac{K^* - \overline{r}_m^e}{\Delta t_m^e},$$



▷ HYPOTHESIS 0: *The signature χ_s of leading session s is not useful for detecting ranking fraud.*

▷ HYPOTHESIS 1: *The signature χ_s of leading session s is significantly higher than expectation.*

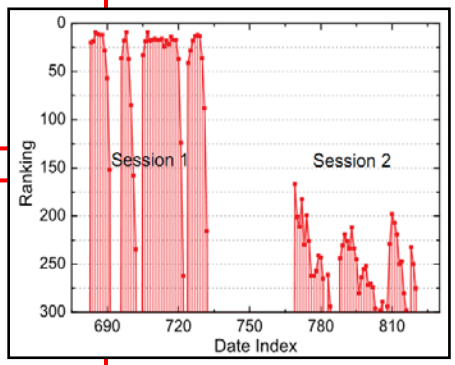
$$\chi_s \sim \mathcal{N}(\mu_\chi, \sigma_\chi),$$

$$\Psi_2(s) = 1 - \mathbb{P}\big(\mathcal{N}(\mu_\chi, \sigma_\chi) \geq \chi_s\big).$$

移动应用商店排名欺诈检测

- 基于排名的欺诈证据3：一个活跃周期内的活跃事件数量

$$|E_s|$$



▷ HYPOTHESIS 0: *The signature* $|E_s|$ *of leading session* s *is not useful for detecting ranking fraud.*

▷ HYPOTHESIS 1: *The signature* $|E_s|$ *of leading session* s *is significantly lager than expectation.*

$$|E_s| \sim \mathcal{P}(\lambda_s)$$

泊松假设估计

$$\mathbb{P}\big(\mathcal{P}(\lambda_s) \geq |E_s|\big) = 1 - e^{-\lambda_s} \sum_{i=0}^{|E_s|} \frac{(\lambda_s)^i}{i!}.$$

$$\Psi_3(s) = 1 - \mathbb{P}\big(\mathcal{P}(\lambda_s) \geq |E_s|\big).$$

移动应用商店排名欺诈检测

- 基于评分的欺诈证据1：活跃期内的平均评分和总体评分差异

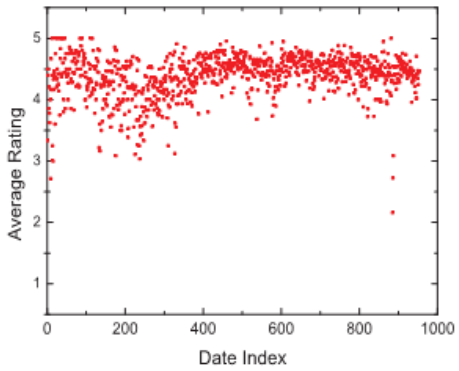
$$\Delta\mathcal{R}_s = \frac{\overline{\mathcal{R}}_s - \overline{\mathcal{R}}_a}{\overline{\mathcal{R}}_a}, \quad (s \in a)$$

- ▷ HYPOTHESIS 0: *The signature $\Delta\mathcal{R}_s$ of leading session s is not useful for detecting ranking fraud.*
- ▷ HYPOTHESIS 1: *The signature $\Delta\mathcal{R}_s$ of leading session s is significantly higher than expectation.*

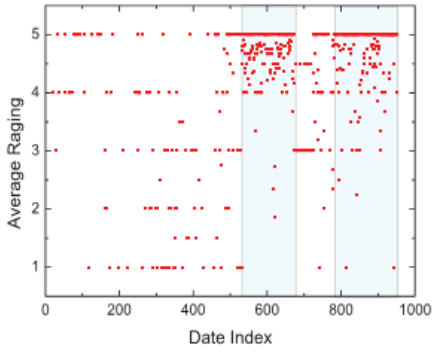
$$\Delta\mathcal{R}_s \sim \mathcal{N}(\mu_{\mathcal{R}}, \sigma_{\mathcal{R}}),$$

$$\Psi_4(s) = 1 - \mathbb{P}\big(\mathcal{N}(\mu_{\mathcal{R}}, \sigma_{\mathcal{R}}) \geq \Delta\mathcal{R}_s\big).$$

一个正常应用



一个可疑应用



移动应用商店排名欺诈检测

- 基于评分的欺诈证据2：活跃期评分分布和历史分布的差异

$$\mathcal{D}(s) = \frac{\sum_{i=1}^{|L|} p(l_i|\mathcal{R}_{s,a}) \times p(l_i|\mathcal{R}_a)}{\sqrt{\sum_{i=1}^{|L|} p(l_i|\mathcal{R}_{s,a})^2} \times \sqrt{\sum_{i=1}^{|L|} p(l_i|\mathcal{R}_a)^2}}.$$

▷ HYPOTHESIS 0: *The signature $\mathcal{D}(s)$ of leading session s is not useful for detecting ranking fraud.*

▷ HYPOTHESIS 1: *The signature $\mathcal{D}(s)$ of leading session s is significantly lower than expectation.*

$$\mathcal{D}(s) \sim \mathcal{N}(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}),$$

$$\Psi_5(s) = 1 - \mathbb{P}\big(\mathcal{N}(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}) \leq \mathcal{D}(s)\big).$$

移动应用商店排名欺诈检测

- 欺诈证据整合

$$\Psi^*(s) = \sum_{i=1}^{N_\Psi} w_i \times \Psi_i(s),$$

1，线性证据整合，学习权值

$$\overline{\pi}(s) = \frac{1}{N_\Psi} \sum_{i=1}^{N_\Psi} \pi_i(s).$$

2，计算一个活跃周期在不同证据下的平均排名

$$\sigma_i(s) = \left(\pi_i(s) - \overline{\pi}(s)\right)^2.$$

3，计算不同证据排名和平均排名的方差

$$\arg\min_{\mathbf{w}} \sum_{a\in A} \sum_{s\in a} \sum_{i=1}^{N_\Psi} w_i \cdot \sigma_i(s),$$

$$s.t. \sum_{i=1}^{N_\Psi} w_i = 1;\ \forall w_i \geq 0.$$

$$\nabla_i = \frac{\partial w_i \cdot \sigma_i(s)}{\partial w_i} = \sigma_i(s).$$

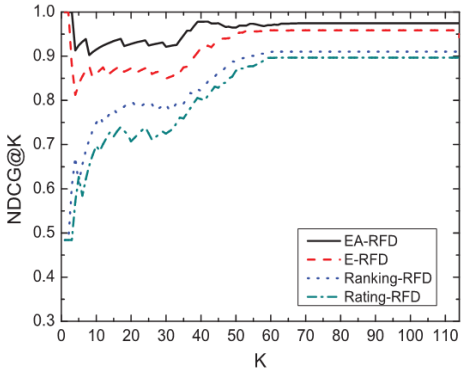
$$w_i = \frac{w_i^* \times \exp(-\lambda \nabla_i)}{\sum_{j=1}^{N_\Psi} w_j^* \times \exp(-\lambda \nabla_j)}$$

移动应用商店排名欺诈检测

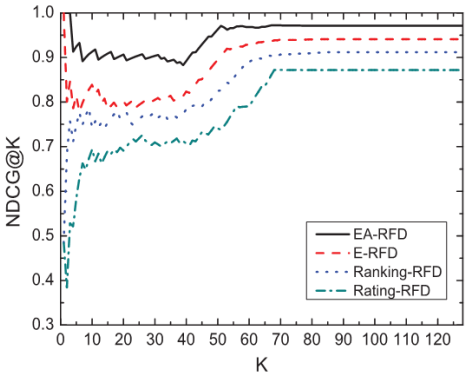
- 实验分析

Table 1: Statistics of the experimental data.

	Top Free 300	Top Paid 300
App Num.	9,784	5,261
Ranking Num.	285,900	285,900
Avg. Ranking Num.	29.22	54.34
Rating Num.	14,912,459	4,561,943
Avg. Rating Num.	1,524.17	867.12



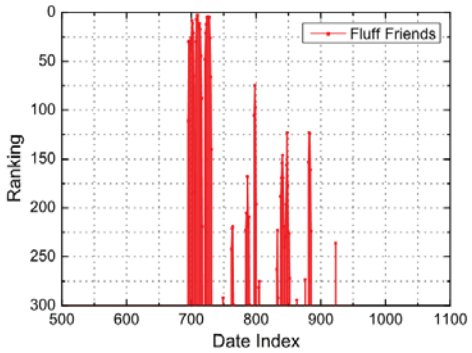
(a) Top Free 300 data set



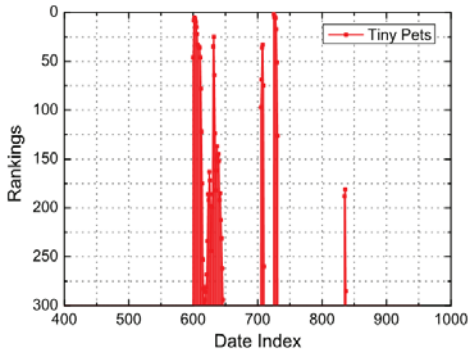
(b) Top Paid 300 data set

Table 4: The reported suspicious mobile Apps.

	EA-RFD	E-RFD	Rank.-RFD	Rat.-RFD
Tiny Pets	**2.89%**	3.91%	4.09%	6.88%
Social Girl	**4.41%**	7.42%	6.68%	8.53%
Fluff Friends	**1.17%**	2.67%	3.75%	5.31%
Top Girl	**1.64%**	1.76%	2.08%	6.81%
VIP Poker	**3.25%**	5.73%	5.23%	4.10%
Sweet Shop	**4.23%**	6.82%	8.23%	6.32%
Crime City	**3.12%**	3.67%	5.31%	8.62%



(a) Fluff Friends



(b) Tiny Pets

内容提纲

情境（Context）

Understand User Habits for Context-aware Service, because We are living in Different Contexts

Context is **when** people, In bus … ;
In office … ; In restaurant … ;
In playing basketball … .

Context information helps to understand user habits.

情境日志

- 情境日志由多个**情境记录**(context record)组成。每个情境记录又由一个时间戳(timestamp)、该时间包含的情境信息、以及该时间的用户交互记录(Usage Record)组成。
- 每个时刻的情境信息由多个**情境特征**(contextual feature)及其取值组成.
- 用户交互记录是用来记录用户在特定情境下的行为操作，比如玩游戏，听音乐，看网页等等，代表了不同的内容偏好。从表中我们可以看到，许多交互记录是空的，这是因为用户并非时时刻刻都有操作行为。

Time tamps

Contextual Features

Contextual Feature-Value Pairs

Preference

Timestamp	Context	Usage records
t_1	{(Day name: Monday),(Time range: AM8:00-9:00),Profile: General,(Battery: 5),(Location: Home)}	Null
t_2	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Battery: 5),(Location: On the way)}	Play action games
t_3	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Battery: 5),(Location: On the way)}	Null
	.....	
t_{359}	{(Day name: Monday),(Time range: AM10:00-11:00),(Profile: Meeting),(Battery: 4),(Location: Work Place)}	Null
t_{360}	{(Day name: Monday),(Time range: AM10:00-11:00),(Profile: Meeting),(Battery: 4),(Location: Work Place)}	Browse sports web sites
	.....	
t_{448}	{(Day name: Monday),(Time range: AM11:00-12:00),(Profile: General),(Battery: 4),(Location: Work Place)}	Play with SNS
t_{449}	{(Day name: Monday),(Time range: AM11:00-12:00),(Profile: General),(Battery: 4),(Location: Work Place)}	Null

情境感知的移动推荐系统

- 一个基于情境感知的移动推荐示例

晚上9点
玩游戏

中午12点
就餐

兴趣
偏好
挖掘

移动情境数据记录

工作日	时间	速度	……	基站标识	不活动时间	交互
否	9:30	慢		2932	短	无
否	9:45	?		2932	长	无
……	?	?	?	?	?	
否	11:45	高		2315	短	上网
……	?	?	?	?	?	
否	12:00	慢		2587	短	搜索

位置情境

时间情境

推荐信息：

前门大街
全聚德？

情境感知的移动推荐系统

- 通过挖掘情境日志获取移动用户**情境感知的兴趣偏好**，用于构建**移动推荐系统**。
  - 比如Joy喜欢**周末晚上在家里**玩**动作类手机游戏**。

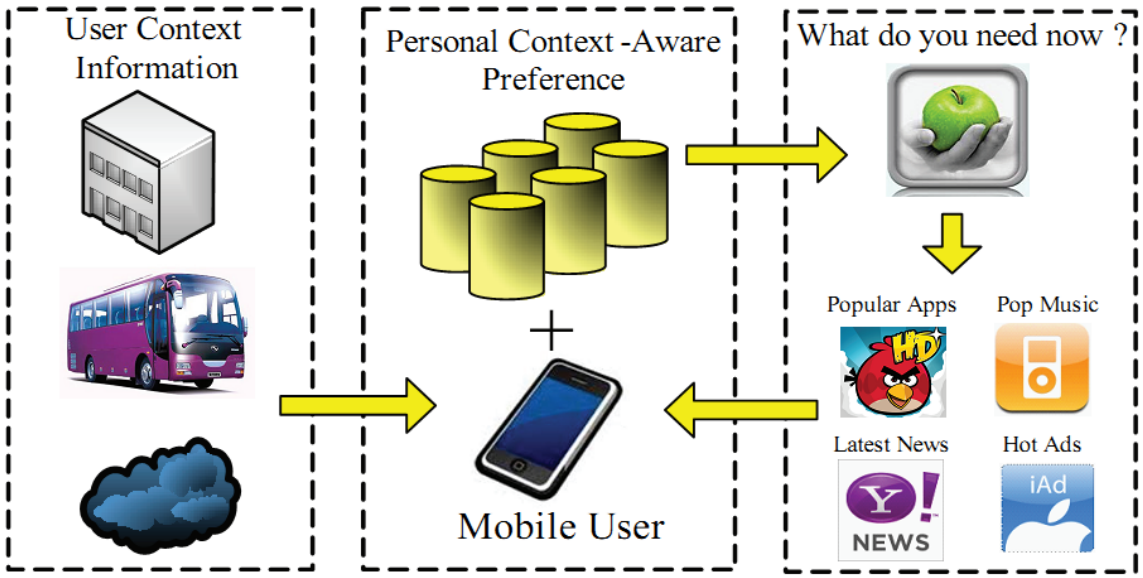


Fig. 1. The personalized context-aware recommendation services for mobile users.

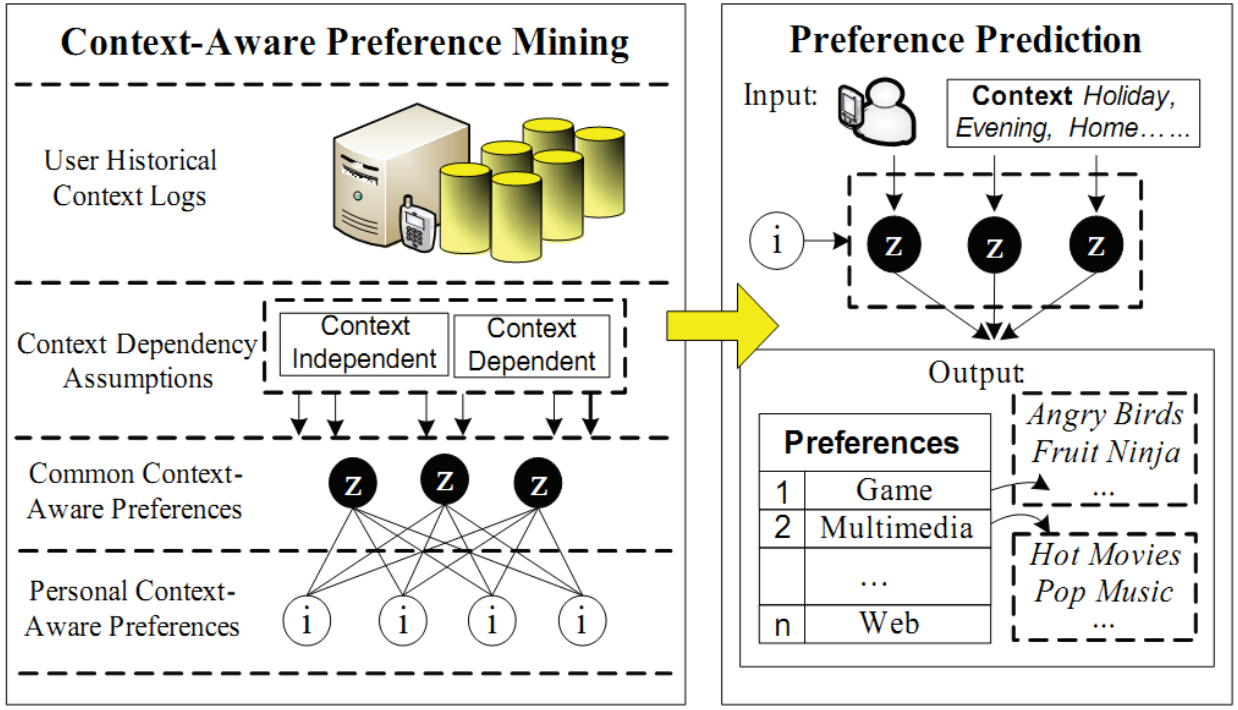
1. Hengshu Zhu, Enhong Chen, et al, **Mining Mobile User Preferences for Personalized Context-Aware Recommendation**, In *ACM Transactions on Intelligent Systems and Technology* (**TIST**), 2013, to appear
2. Hengshu Zhu, Enhong Chen, et al., **Mining Personal Context-Aware Preferences for Mobile Users**, In *Proceedings of the 12th IEEE Conference on Data Mining* (**ICDM 2012**), Brussels, Belgium, 2012

情境感知的移动推荐系统

- 面临的挑战：
  - 情境日志特征多，情境建模困难
  - 情境日志缺少显式的评分数据
  - 单个用户的情境日志非常稀疏，挖掘情境偏好困难

情境感知的移动推荐系统

- 问题解决框架
  - 根据多个用户的情境记录，学习出一组**共同情境偏好{z}**，单个用户的个性化情境偏好即可表征为在共同情境偏好下的概率分布
  - 根据不同的情境数据独立性假设（情境特征独立假设，情境特征条件依赖假设），提出了两种不同的共同情境偏好挖掘方法
  - 给定一个用户***u***，其在情境***C***下对于内容***c***的情境偏好可以用后验概率***P(c|C,u)***来估计



情境感知的移动推荐系统

- 设多个用户的共同情境偏好为{z}，则一个用户u在情境C下对于内容c的偏好可以表示为：

$$P(c|C,u) = \frac{P(c,C|u)\cdot P(u)}{P(C,u)} \propto P(c,C|u)$$

$$\propto \sum_z P(c,C,z|u) \propto \sum_z \Big(P(c,C|z) \cdot P(z|u) \Big),$$

学习多个用户的共同情境偏好

学习单个用户在共同偏好上的概率分布

如何计算P(c,C|z)和P(z|u)?

情境感知的移动推荐系统

- 两种情境特征-偏好独立性假设：
  - **假设1：**不同情境特征对于用户兴趣偏好的影响是基于共同偏好条件独立的。比如在挖掘共同兴趣偏好时，我们认为**（地点：家）**和**（时间：晚上１０点）**这个两个情境特征**对于偏好“游戏”的影响是独立的**。
  - **假设2：**不同情境特征对于用户兴趣偏好的影响是相互依赖的。比如在挖掘共同兴趣偏好时，我们认为**（地点：家）**和**（时间：晚上１０点）**这个**两个情境特征的同时出现造成了对偏好“游戏”的影响**。

情境感知的移动推荐系统

- 基于假设 1，我们可以得到

$$P(c|C,u) \propto \sum_{z} \Big(P(c,C|z) \cdot P(z|u) \Big) \propto \sum_{z} \Big(\prod_{p_i \in C} P(c,p_i|z) \cdot P(z|u) \Big)$$

- 我们将(c,p)称作一个情境偏好原子特征（Atomic Context-aware Preference Feature, ACP-Feature），其获取可以通过如下方法：

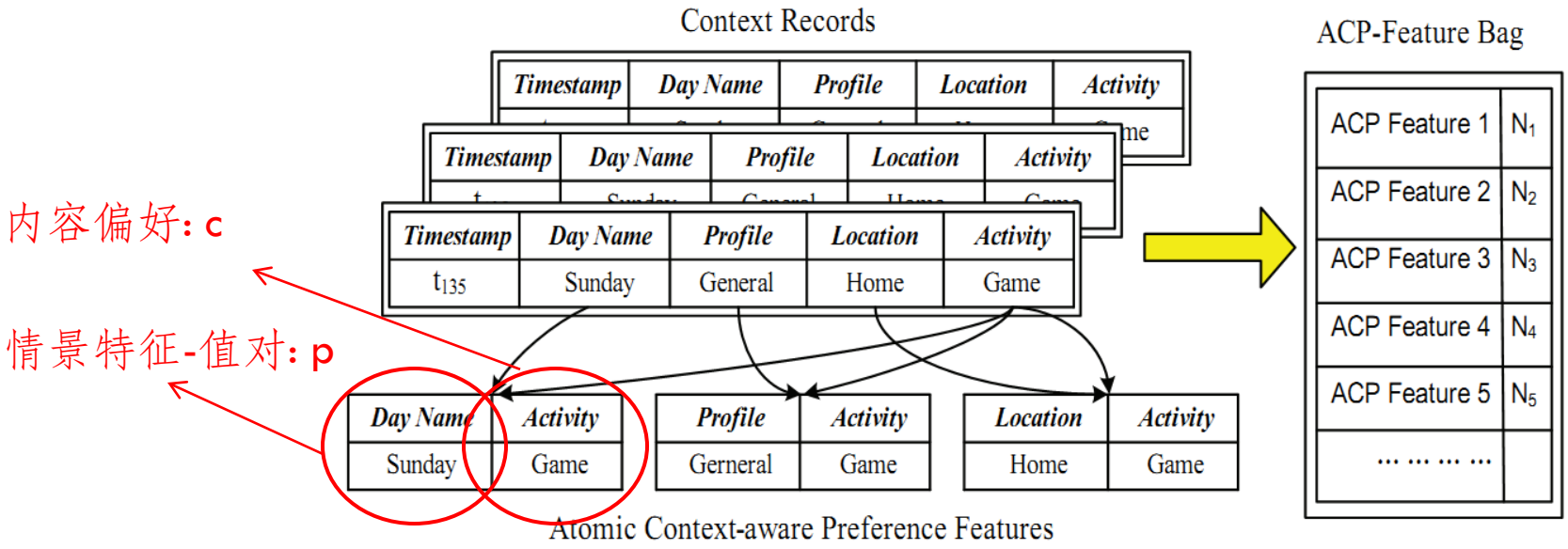


Fig. 3. The generation process of ACP-feature bag from user's context records.

情境感知的移动推荐系统

- 我们通过主题模型来学习P(c,p|z)和P(z|u)

增加情境特征先验概率

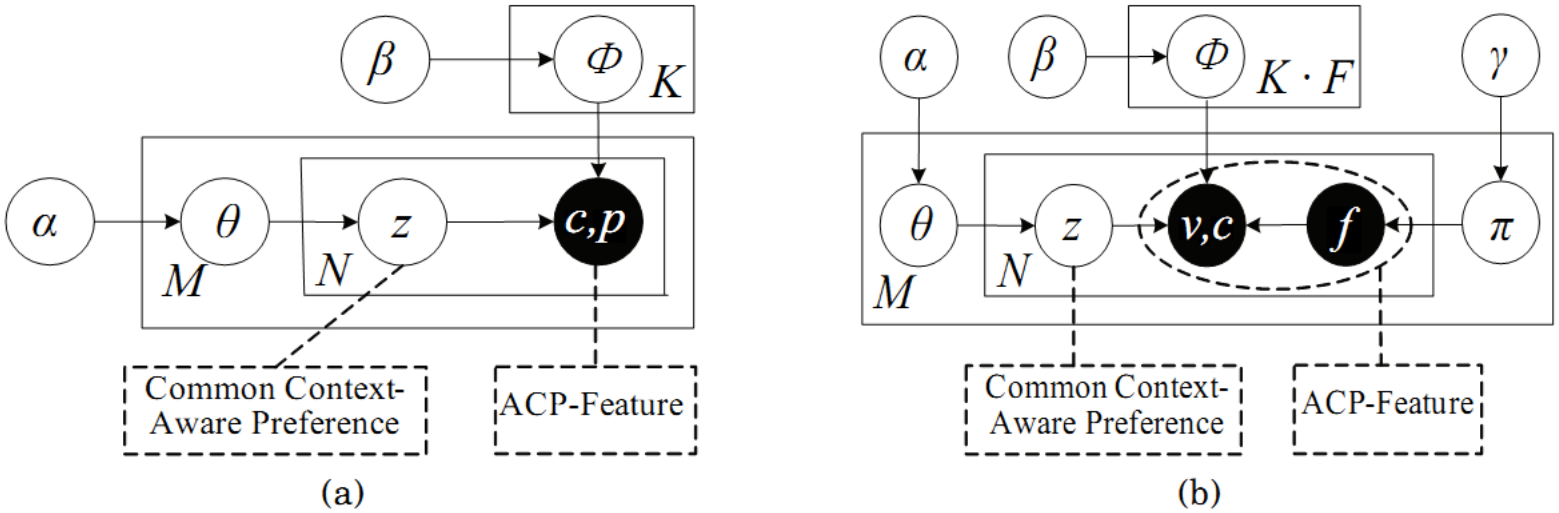


Fig. 4. The graphic representation of modeling ACP-feature bags by (a) LDA topic model, and (b) LDAC topic model.

$$P(z|u) = P(z|d_u) = \frac{n_{u,z} + \alpha_z}{\sum_i^K n_{u,z_i} + \sum_i^K \alpha_{z_i}},$$

$$P(c, v_p|f_p, z) = \frac{n_{z,c,f_p,v_p} + \beta_{c,v_p}}{\sum_v n_{z,c,f_p,v} + \sum_{v \in V_{f_m}} \beta_v},$$

$$P(c, p|z) = \frac{n_{z,c,p} + \beta_{c,p}}{\sum_i^M n_{z,(c,p)_i} + \sum_i^M \beta_{(c,p)_i}},$$

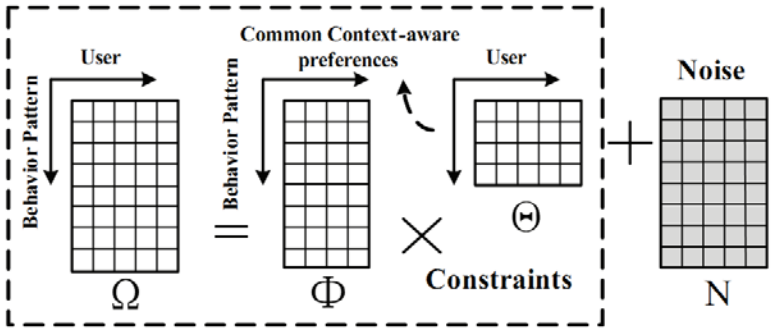
$$P(f_p) = \frac{\sum_z \sum_v n_{k,c,f_p,v} + \gamma_{f_m}}{\sum_f \sum_z \sum_v n_{z,c,f_p,v} + \sum_f \gamma_f},$$

情境感知的移动推荐系统

- 基于假设 2，我们首先挖掘频繁的情境偏好模式（behavior pattern）。

(*Is holiday? Yes*) (*Time Period: Evening*) (*Location: Home*) ⇒ **Game**
(*Is holiday? Yes*) (*Time Period: Evening*) (*Charging State: Charging*) ⇒ **Game**
(*Time Period: Morning*) (*Location: Work Place*)⇒ **Business**
(*Time Period: Evening*) (*Location: Moving*) (*Profile: Silent*) ⇒ **Multimedia**
(*Time Period: Evening*) (*Location: Home*) ⇒ **Web**

- 假设频繁情境为C^r，我们可以通过***Constraint based Non-negative Matrix Factorization***来学习相关概率P(c,C^r|z)和P(z|u)



1) all elements in matrix **Φ** and **Θ** should be non-negative values

2) $\forall_{j:1\leq j\leq M}\sum_{k=1}^{K}\theta_{kj}=1$ and $\forall_{k:1\leq k\leq K}\sum_{i=1}^{N}\phi_{ik}=1$

情境感知的移动推荐系统

Context	Value range
Week	{Monday, Tuesday,... , Sunday}
Is a holiday?	{Yes, No}
Day period	{Morning(7:00-11:00), Noon(11:00-14:00), Afternoon(14:00-18:00), Evening(18:00-21:00), Night(21:00-Next day 7:00)}
Time range	{0:00-1:00, 1:00-2:00, ... , 23:00-24:00}
Profile type	{General, Silent, Meeting, Outdoor, Pager, Offline}
Battery Level	{Level 1, Level 2, ... , Level 7}
Charging State	{Charging, Complete, Not Connected}
Social location	{Home, Work Place, On the way}.

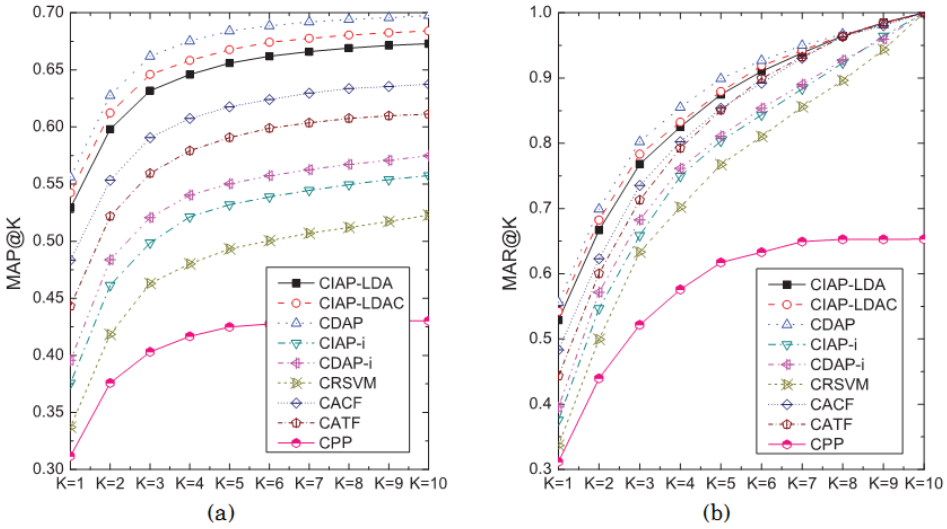


Fig. 9. The average (a) $MAP@K$ and $MAR@K$ performance of each prediction approach in the five-fold cross validation.

实验结果分析

Table V. Prediction case 1 for user #162 and #423.

Context	{(*Time range: PM21:00-22:00*), (*Profile: Silent*), (*Is holiday: Yes*),(*Day name: Sunday*), (*Day period: Night*), (*Location: Home*)}		
	Top 3 predicted preferences for user #*162*		
Ground truth	***Game***		
CIAP-LDA	**Game** (√)	Multimedia	Web
CIAP-LDAC	**Game** (√)	Multimedia	SNS
CDAP	**Game** (√)	Web	Multimedia
CIAP-i	Multimedia	Business	**Game** (√)
CDAP-i	Web	**Game** (√)	System
CASVM	Multimedia	Web	**Game** (√)
CACF	Multimedia	**Game** (√)	System
CATF	Multimedia	**Game** (√)	Web
CPP	Multimedia	SNS	Web
	Top 3 predicted preferences for user #*423*		
Ground truth	***Web***		
CIAP-LDA	**Web** (√)	Multimedia	SNS
CIAP-LDAC	**Web** (√)	Multimedia	Game
CDAP	**Web** (√)	Reference	Game
CIAP-i	Multimedia	Game	**Web** (√)
CDAP-i	SNS	Multimedia	**Web** (√)
CASVM	Multimedia	Reference	**Web**(√)
CACF	Game	**Web** (√)	Multimedia
CATF	Game	**Web** (√)	System
CPP	Multimedia	SNS	**Web** (√)

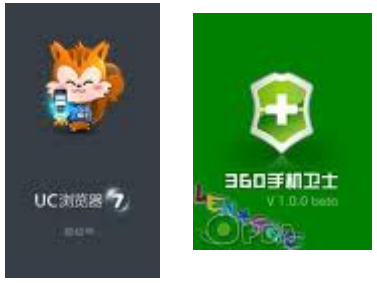
内容提纲

基于扩展信息的移动应用分类

- 随着智能移动设备的普及，大量移动应用（Mobile Apps）被开发出来。截止到2013年7月，苹果应用商店和谷歌安卓市场共有约200万款移动应用，其下载量超过1000亿次。

- 如何管理这些移动应用程序，同时通过它们的使用记录来理解用户兴趣偏好成为一个难题。
  - 需要根据不同的需求设计不同的分类体系。
  - 应用数量众多，需要有效的自动化方法来分类。

移动应用



应用分类表

App	Tag
UC Web	Browsing
Ovi Store	Downloading
Safe 360	Security
…	…

基于扩展信息的移动应用分类

- 研究面临的挑战：
  - 移动应用没有足够的显式特征信息来构建分类器。唯一可用的特征即移动应用名称中所包含的文字。但是这些文字通常极为的有限，通常不超过3个单词，而且单词的重现频率很低。

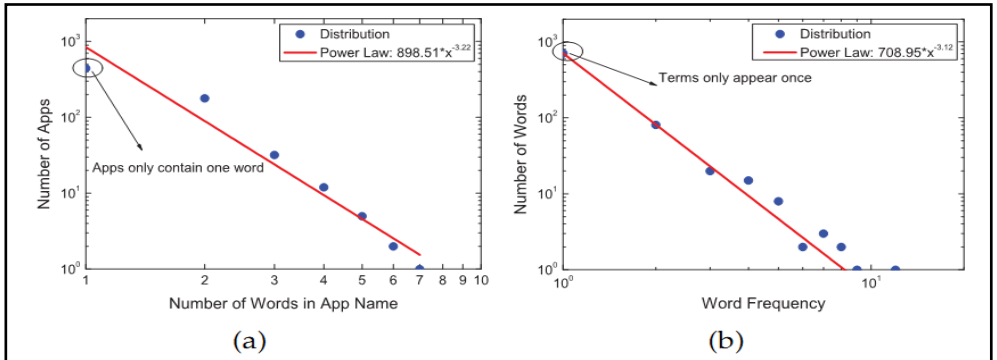


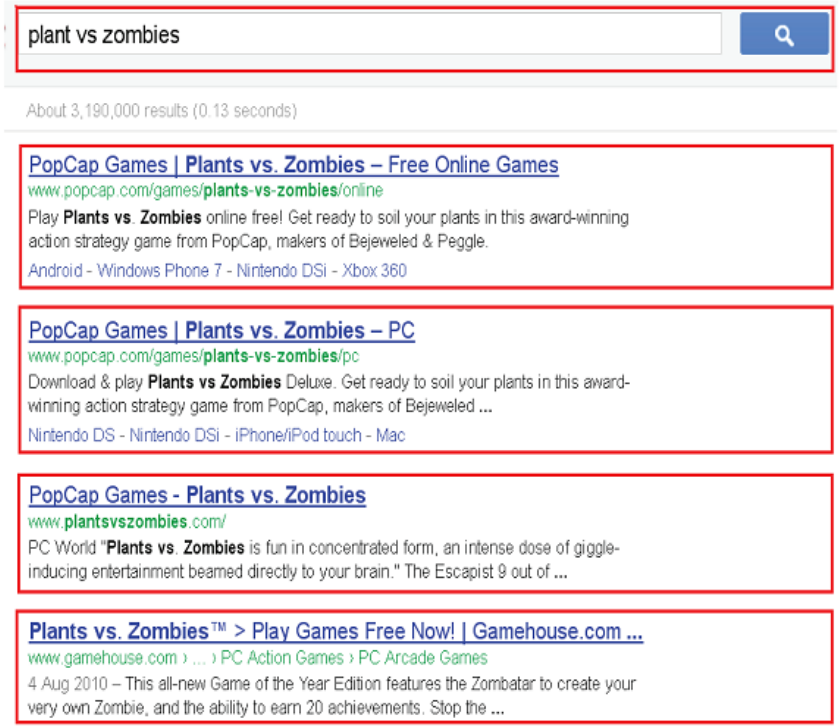
Fig. 7. The distribution of (a) the number of mobile Apps with respect to the name length, and (b) the number of unique words in App names with respect to their appearing frequency in our real-world date set.

1. Hengshu Zhu, Huanhuan Cao, Enhong Chen, Hui Xiong, Jilei Tian, **Mobile App Classification with Enriched Contextual Information**, In *IEEE Transactions on Mobile Computing* (**TMC**), 2013,, to appear
2. Hengshu Zhu, Huanhuan Cao, Enhong Chen, Hui Xiong, Jilei Tian, **Exploiting Enriched Contextual Information for Mobile App Classification**, In *Proceedings of the 21st ACM Conference on Information and Knowledge Management* (**CIKM 2012**), Page 1617-1621, Hawaii, USA, 2012.

基于扩展信息的移动应用分类

- 我们提出利用基于互联网搜索引擎和移动情境日志来构建“扩展信息”，用以训练移动应用分类器。

将移动应用名称输入搜索引擎，返回网页结果



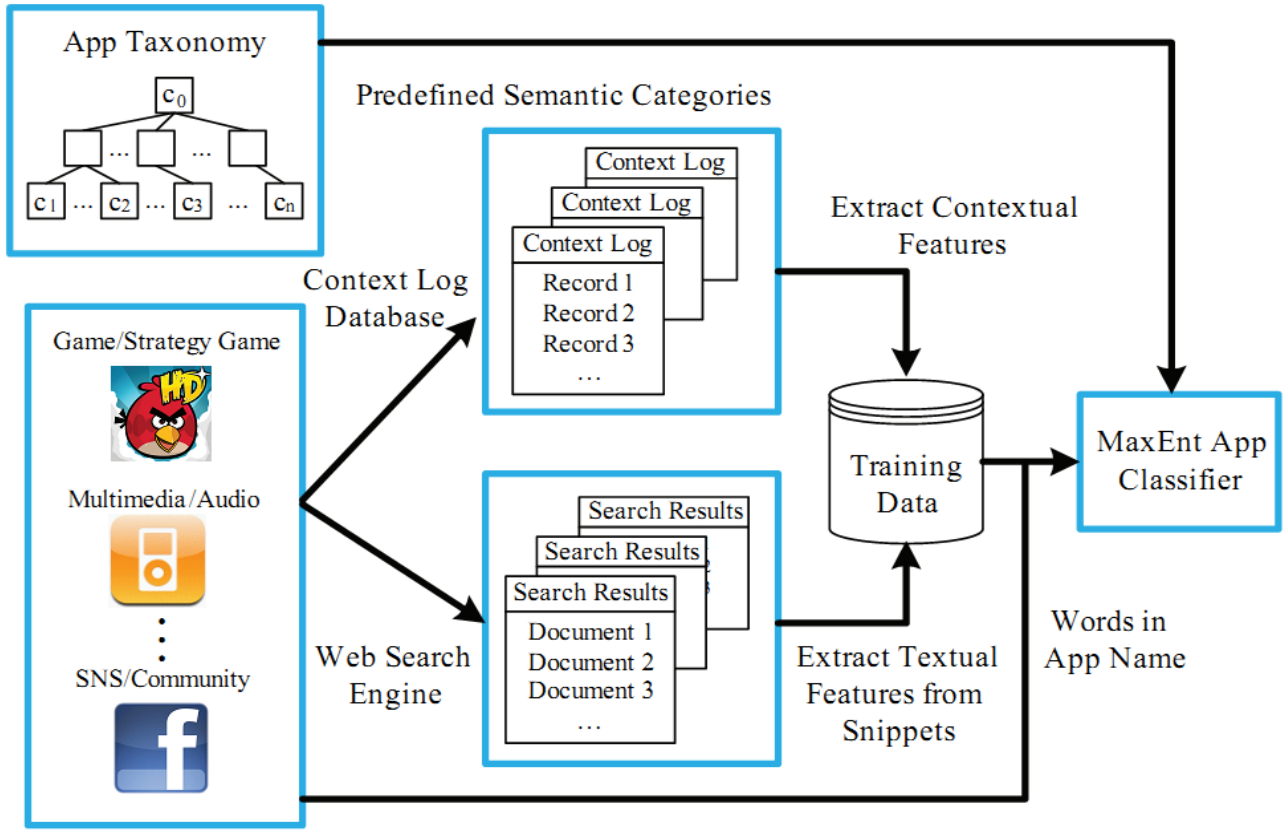
不同应用在不同情境下使用，比如“游戏”经常在“家里”“晚上”等情境下使用。

Table 1: A toy context log from real-world data set.

Timestamp	Context	App usage records
t_1	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Home)}	Angry Birds
t_2	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Moving)}	Null
t_3	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Moving)}	Twitter
.....		
t_{55}	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Moving)}	UC Web
t_{56}	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Moving)}	Null
t_{57}	{(Day name: Monday),(Time range: AM8:00-9:00),(Profile: General),(Location: Moving)}	Music Player
.....		
t_{359}	{(Day name: Monday),(Time range: AM10:00-11:00),(Profile: Meeting),(Location: Work Place)}	Null
t_{360}	{(Day name: Monday),(Time range: AM10:00-11:00),(Profile: Meeting),(Location: Work Place)}	SMS

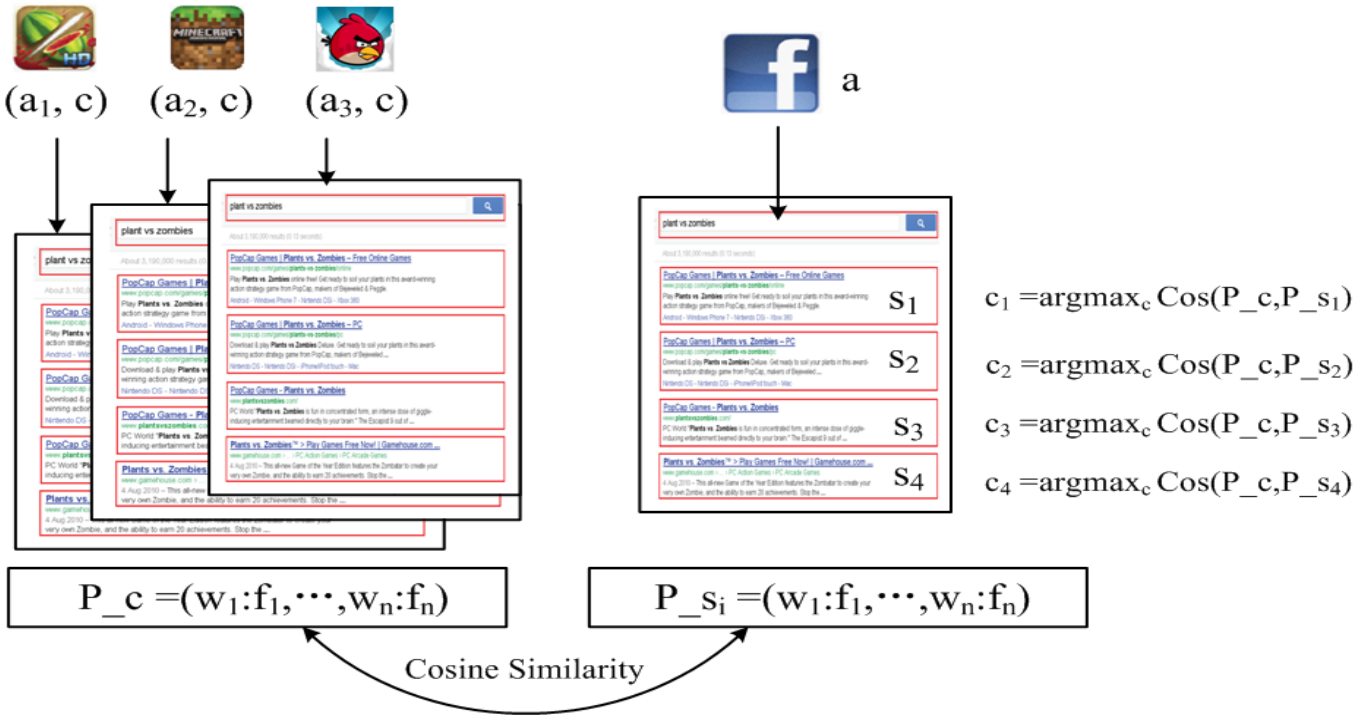
基于扩展信息的移动应用分类

- 我们提出的技术框架，核心问题在于如何从扩展信息中设计和提取分类特征。



基于扩展信息的移动应用分类

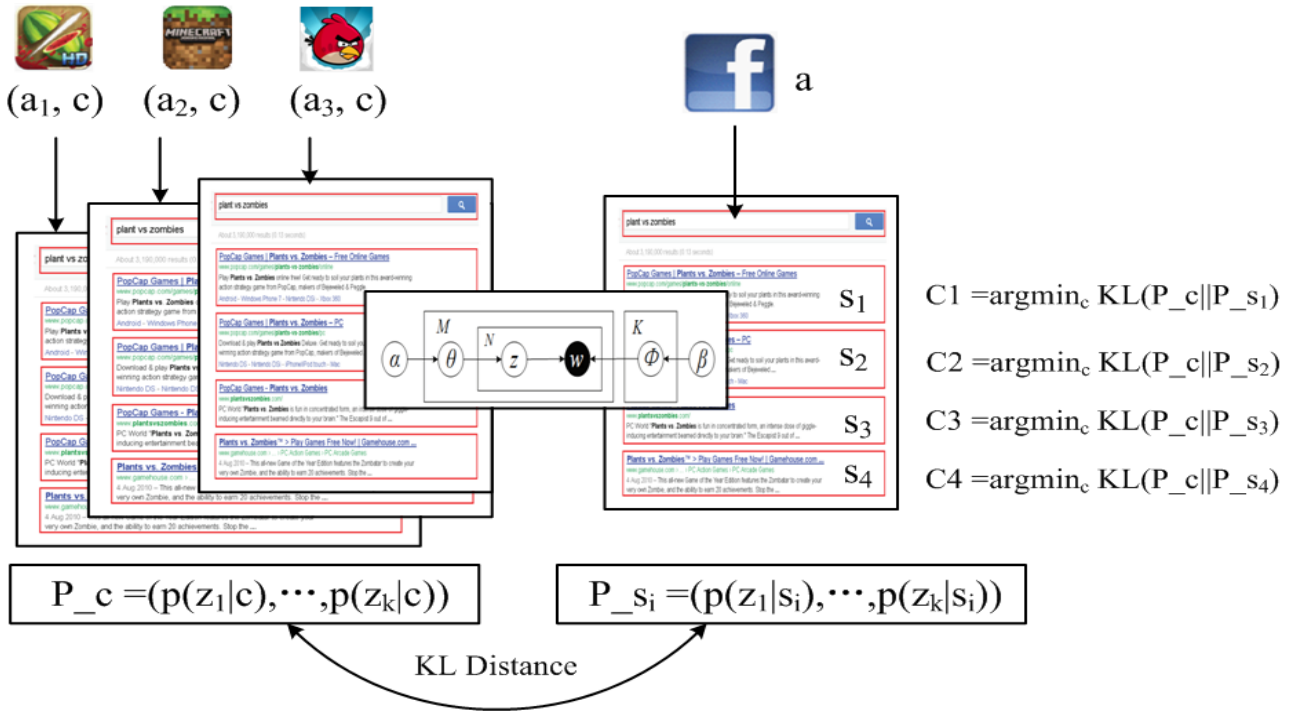
- 基于互联网的文本特征提取：
- **特征1：Explicit Feedback of Vector Space Models**
  - 将每个返回的互联网页通过VSM映射到特定的应用类别，然后计算



$$GConf(c,a) = \frac{M_{c,a}}{M}, \qquad GEnt(c,a) = -\sum_{c_i \neq c} P(c_i)\log P(c_i),$$

基于扩展信息的移动应用分类

- 基于互联网的文本特征提取：
- **特征2：Implicit Feedback of Semantic Topics**
  - 将每个返回的互联网页通过主题模型映射到特定的应用类别，然后计算

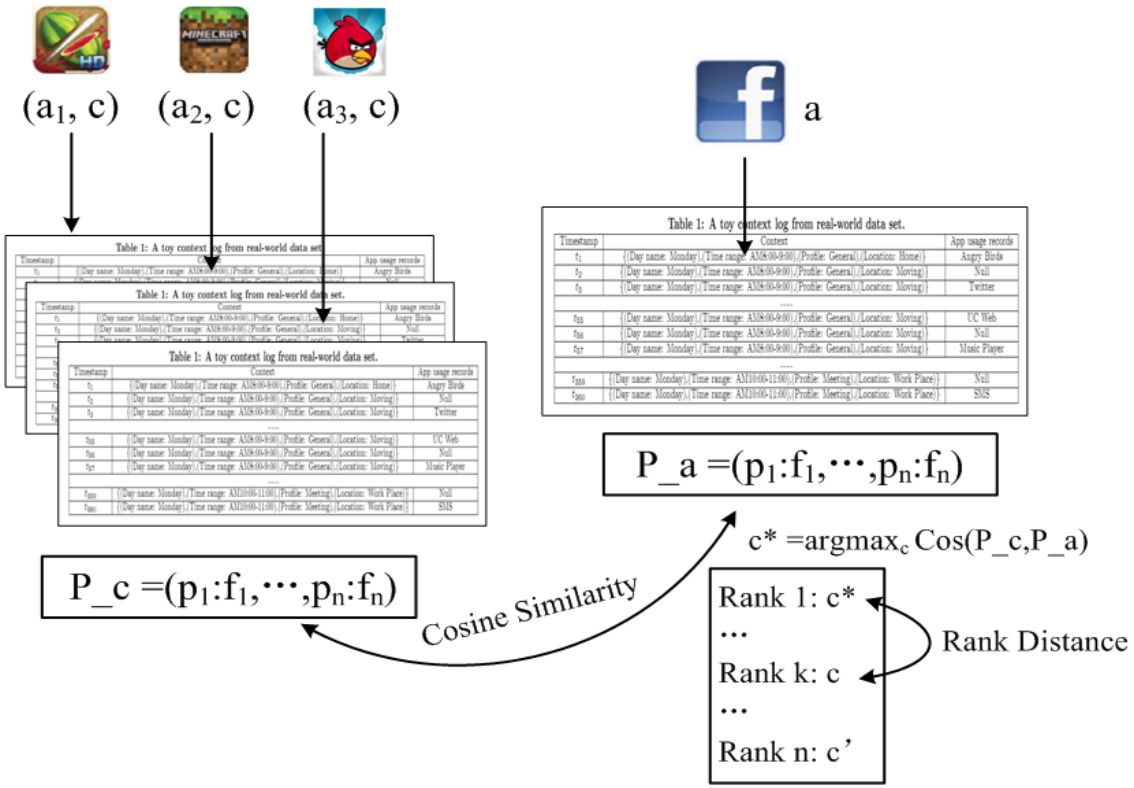


$$TConf(a,c) = \frac{T_{a,c}}{M} \qquad TEnt(c,a) = -\sum_{c_i \neq c} P(c_i)\log P(c_i),$$

基于扩展信息的移动应用分类

- 基于情境日志的情境特征提取：
- **特征1：Pseudo Feedback of Context Vectors**
  - 构建不同应用以及类别的情境特征向量，然后用VSM计算相近程度，然后计算

$$CRDistance(a,c) = Rk(c) - Rk(c^*) = Rk(c) - 1,$$

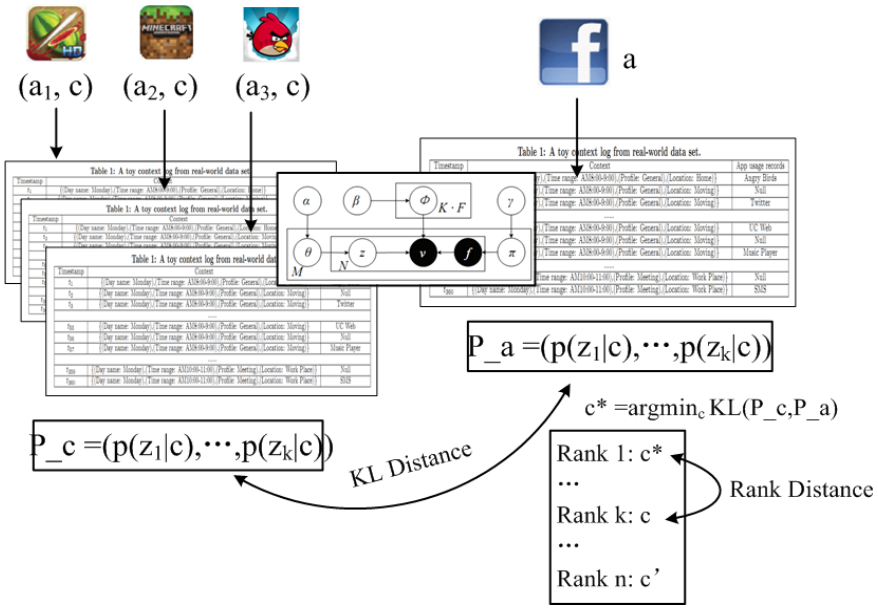


基于扩展信息的移动应用分类

- 基于情境日志的情境特征提取：
- **特征2：Implicit Feedback of Context Topics**
  - 通过主题模型计算不同应用及类别的KL相似度，然后计算

$$TRDistance(a,c) = Rk(c) - Rk(c^*) = Rk(c) - 1,$$

- **特征3：Frequent Context Patterns**
  - 挖掘和不同应用类别相关的频繁情境模式，作为布尔特征。



#1	(*Is a holiday? Yes*)(*Day period: Evening*)(*Location: Home*) ⇒ Angry Birds (**Game/Strategy Game**)
#2	(*Day period: Morning*)(*Location: Work Place*) ⇒ Yahoo Mail (**Communication/Mail&SMS**)
#3	(*Day period: Evening*)(*Location: On the Way*)(*Profile: Silent*) ⇒ Music Player (**Multimedia/Audio**)
#4	(*Day period: Afternoon*)(*Location: On the Way*) ⇒ Ovi Map (**Navigation/Maps**)

基于扩展信息的移动应用分类

- 采用最大熵模型分类器。

$$P(c|a) = \frac{1}{Z(a)} \exp(\sum_i \lambda_i f_i(a,c)),$$

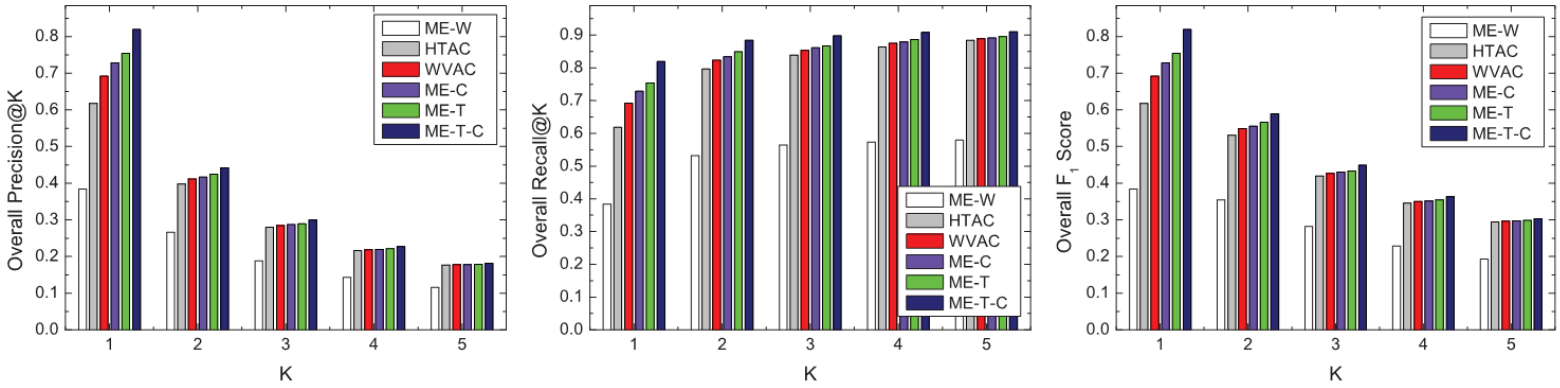
$$L(\Lambda|\mathcal{D}) = \log \prod_{d \in \mathcal{D}} P_\Lambda(c^{(i)}|a^{(i)}).$$

$$c_T^* = \arg\max_{c_T} P(c_T|a_T, \Lambda).$$

- 将应用名称中的单词作为基本特征。

基于扩展信息的移动应用分类

- 实验效果



(a) Overall Precision@K (b) Overall Recall@K (c) Overall F_1 Score

Fig. 9. The overall performance of each classification approach with different evaluation metrics in the cross validation.

TABLE 5
The predefined two-level taxonomy in our experiments.

Level-1 Categories	Level-2 Categories
Internet	⋆Web Browser, ⋆Others
Business	⋆Office Tools, ⋆Security, ⋆Others
Communication	⋆Call, ⋆Mail&SMS, ⋆Others
Game	⋆Action, ⋆Strategy, ⋆Others
Multimedia	⋆Audio, ⋆Video, ⋆Others
Navigation	⋆City Guides, ⋆Maps, ⋆Others
SNS	⋆IM, ⋆Blog&Forum, ⋆Others
System	⋆Management, ⋆Performance, ⋆Others
Reference	⋆News, ⋆Utility, ⋆Reading, ⋆Others

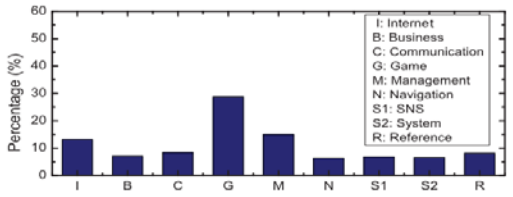


Fig. 8. The App distribution of different level-1 category labels in our data set.

App Name	Snakes 3D
Ground Truth	Game/Action
Words	Snakes, 3D
Snippets	**→Snake 3d - Free Online Games (FOG)** www.freeonlinegames.com/game/snake-3d **→3D Snake - squid soup : games** squidsoup.com/games/snake/shockholder.html **→Snake 3D - 2 Flash Games** www.2flashgames.com
Topics†	Entertainment, Video Games, Mobile Applications
Context Topics†	Relax at Home, Relax at Work Place, After Work
Context Patt.s	{(*Is a holiday?: Yes*)(*Day period: Evening*)(*Location: Home*)}; {(*Day period: Evening*)(*Charging state: Charging*)(*Location: Home*)}; {(*Day period: Afternoon*) (*Profile: Silent*)}
Predicted Category Labels	
WVAC	Multimedia/Video; Game/Others; **Game/Action**
HTAC	Multimedia/Video; Internet/Others; **Game/Action**
ME-W	Multimedia/Video; Multimedia/Others; Game/Others
ME-T	Game/Others; **Game/Action**; Multimedia/Video
ME-C	Multimedia/Video; **Game/Action**; Game/Others
ME-T-C	**Game/Action**; Multimedia/Video; Game/Others

* Limited to space, we only show top three corresponding results.
† The topics are manually labeled for illustration.

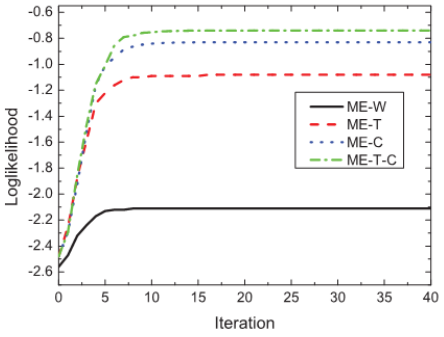
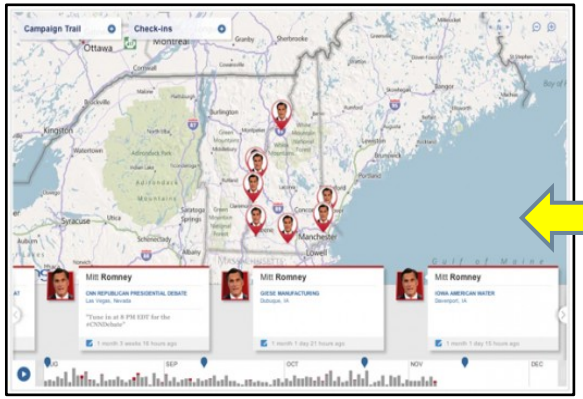


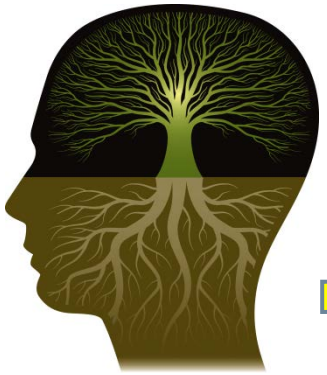
Fig. 10. The objective function values per iteration of training ME-W, ME-T, ME-C and ME-T-C.

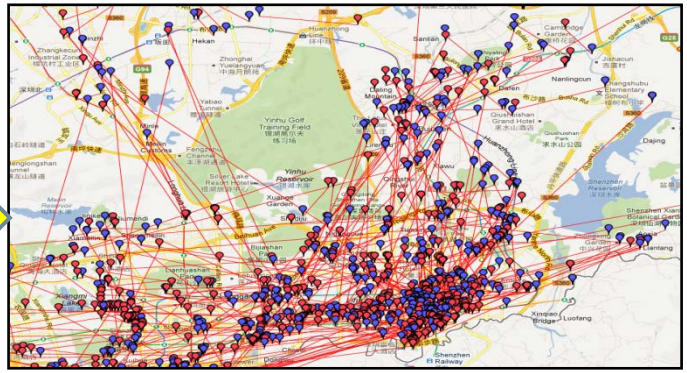
内容提纲

面临的挑战1：领域知识的挑战

- 移动商务智能领域知识多样
- 如何结合领域知识实现数据挖掘应用

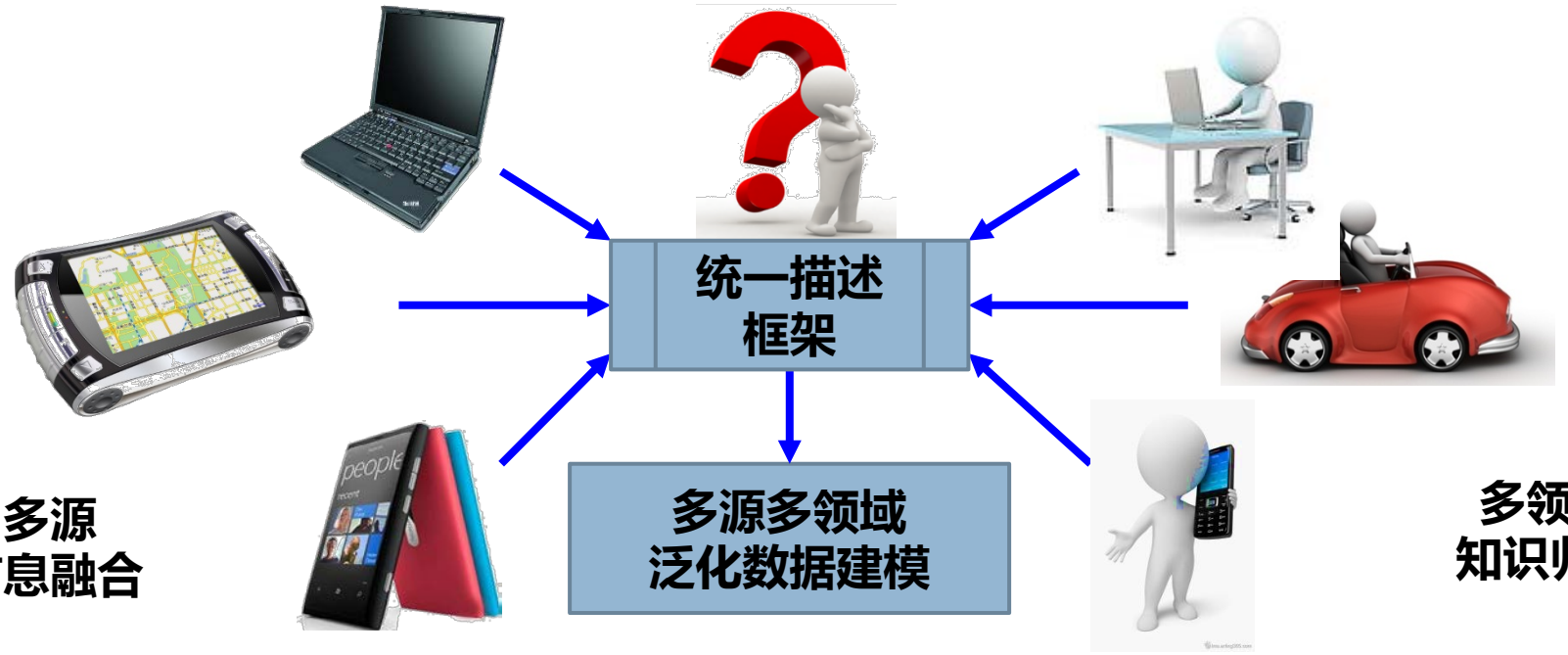






面临的挑战2：多源信息建模的挑战

- 移动商务信息来源多样
- 多源异构数据统一映射问题



面临的挑战3：移动用户隐私保护的挑战

- 移动用户隐私保护问题

修改和记录用户日历项、访问设备状态、自动USB读写等

Permissions: This application has access to the following

➔ **Your personal information**

Read calendar events plus confidential information

Allows the app to read all calendar events stored on your tablet including those of friends or coworkers. Malicious apps may extract personal information from these calendars without the owners knowledge Allows the app to read all calendar events stored on your phone including thoSe of friends or coworkers. Malicious apps may extract personal information from these calendars without the owners knowledge.

Add or modify calendar events and send email to guests without owners' knowledge

Allows the app to send event invitations as the calendar owner and add remove, change events that you can modify on your device including those of friends or co-workers. Malicious apps May send spam emails that appear to come from calendar owners, modify events without the owners knowledge, or add fake events.

➔ **Phone calls**

Read phone state and identity

Allows the app to access the phone features of the device An app with this permission can determine the phone number and serial number of this phone whether a call is active, the number that call is connected to and the like

➔ **Storage**

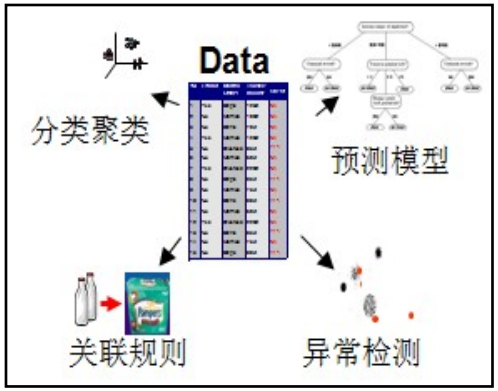
Modify/delete USB storage contents modify /delete SD card contents

Allows the app to write to the USB storage Allows the app to write to the SD card.

一个手机APP的权限说明

面临的挑战4：来自于用户体验的挑战

- 结合市场用户反馈实现服务效果验证
- 结合用户反馈重新设计数据挖掘方法



数据挖掘方法

移动商务数据

用户市场反馈

结语

- 移动商务智能方兴未艾
- 大数据为移动商务智能的研究提供了全新的机遇
- 面向移动商务的数据挖掘方法和应用前景广阔
- 移动商务智能的研究是机遇与挑战并存
  - 机遇：新模型、新方法、新应用、新数据
  - 挑战：数据的多源性、领域的多样性、用户的可接受度

2014年《软件学报》专题（专刊）征文通知

- “大数据分析”专辑

- 预录用论文需在2014年中国数据挖掘会议（**CCDM2014, 2014年5月25-26日在浙江师范大学召开**）上进行宣读，欢迎相关领域的专家学者和科研人员踊跃投稿。专题投稿分两轮，第一轮的文章经过评审需要修稿的论文可以参加第二轮投稿。

- 论文请通过网址（http://www.jos.org.cn）进行投稿，并**注明“数据挖掘2014专题”。（否则按自由来稿处理）**

重要日期

- **第一轮征文截止日期： 2014年1月31日**
- 第一轮征文预录用通知日期： 2014年3月31日
- 第二轮征文截止日期： 2014年 4月7日
- 第二轮征文预录用通知日期: 2014年 5月7日
- **CCDM 2014宣读日期：2014年5月25-26日**
- 作者修改稿提交日期：2014年6月5日
- 终审录用通知：2014年6月15日
- 论文最终版提交截止日期：2014年6月22日
- **专刊出版日期：2014年9月**

THANK YOU!