



博士论文答辩报告

# 基于知识迁移的社交媒体事件检测方法研究

博士生：黄威靖  
导师：王腾蛟教授

北京大学信息科学技术学院数据库实验室

2018-06-05

# 大纲

1. 引言

2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

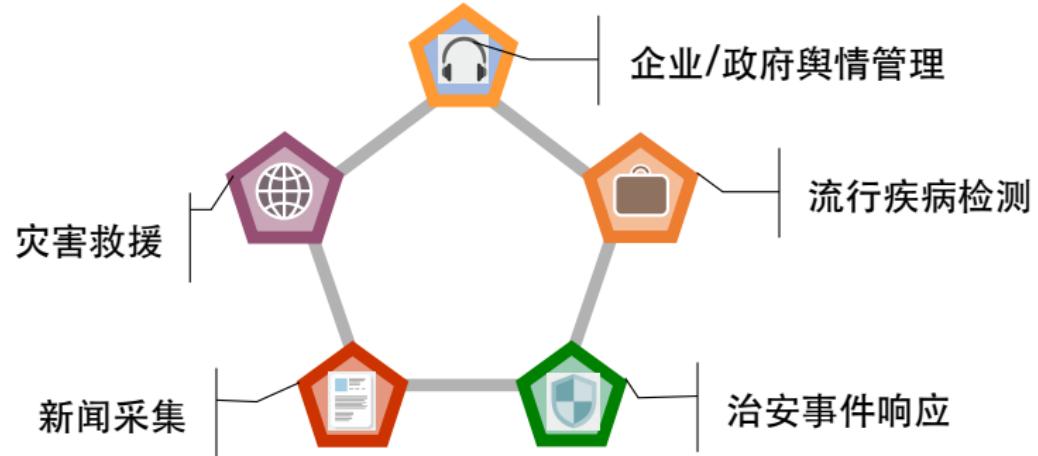
4. 用户兴趣建模与知识迁移方法

5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望

# 社交媒体事件检测问题的背景

社交媒体上进行事件检测可服务于众多应用：



# 社交媒体事件检测问题的背景

科研项目背景：

1. 非结构化数据管理系统北大部分——国家“核高基”重大专项  
(2010ZX01042-002-002-02)
2. 海量 Web 数据结构化内容提取与集成及大型示范应用——国家  
863 计划课题 (2012AA011002)
3. 微博用户社区及主题时序方法研究——中国信息安全测评中心合作  
项目

# 社交媒体的语境动态演变及带来的挑战

社交媒体用语不规范，主题动态变化，我们将其总结为社交媒体的语境动态演变

交流语境的变化：从正式交流语境到非正式交流语境的转变，导致用语不规范 [Gunraj 2016].

外部语境的变化：用户易于被外界影响 [Weng 2012]，热门话题随着时间快速变化。

前述两种变化的综合叠加，例子：  
Somehow made it through Irene.  
仅在 2011.8 的语境中知道这是关于飓风艾琳。

## 相关研究状况

传统媒体中事件的定义 [Allan, 2002]: 在特定的时间, 特定的地点发生的特定的事, 以及其附属的前提条件和后续结果, 称作事件。

社交媒体中事件的定义 [Becker, 2011]: 仅当上述定义中的事件被社交媒体报道了才能被定义为社交媒体上的事件。

### ↓ 形式化定义

社交媒体上的事件  $e$  由社交媒体数据流  $D_e = (d_{e,1}, d_{e,2}, \dots, d_{e,m})$  确定: 在时间窗口  $[t_{e,1}, t_{e,m}]$  上,  $D_e$  的文本集合  $w_{e,1}, \dots, w_{e,m}$  确定出的**特征项**的频率明显高于时刻  $t_{e,1}$  之前该特征项频率的期望值, 且  $D_e$  能够对应于现实中在特定时间已发生的特定的事。

# 相关研究状况

表: 社交媒体事件检测已有方法

方法类型	代表工作	事件粒度	特点
基于普通聚类的方法	[Chen, SIGIR 2013]	高频的相似微博	实现简单
基于敏感哈希的方法	[Petrovic, NAACL 2010]	高频的相似微博	响应速度快
基于词频统计的方法	[HE,SIGIR 2007]	高频词组	可区分周期性、非周期性事件
基于主题模型的方法	[Yan, AAAI 2015] [Jahnichen, AISTATS 2018]	主题	可利用上下文信息
基于分类器 + 后处理	[Yang, KDD 2014]	对微博分类后，再检测事件	进行领域事件检测

受限于社交媒体的语境动态演变，上述方法难于准确检测事件。

# 大纲

1. 引言

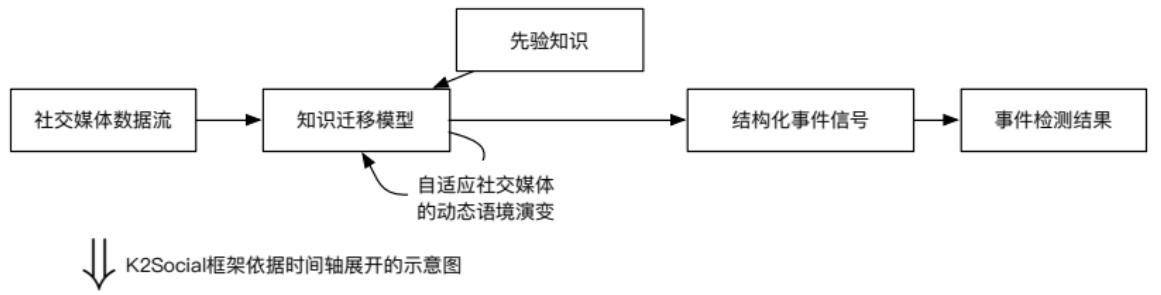
2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

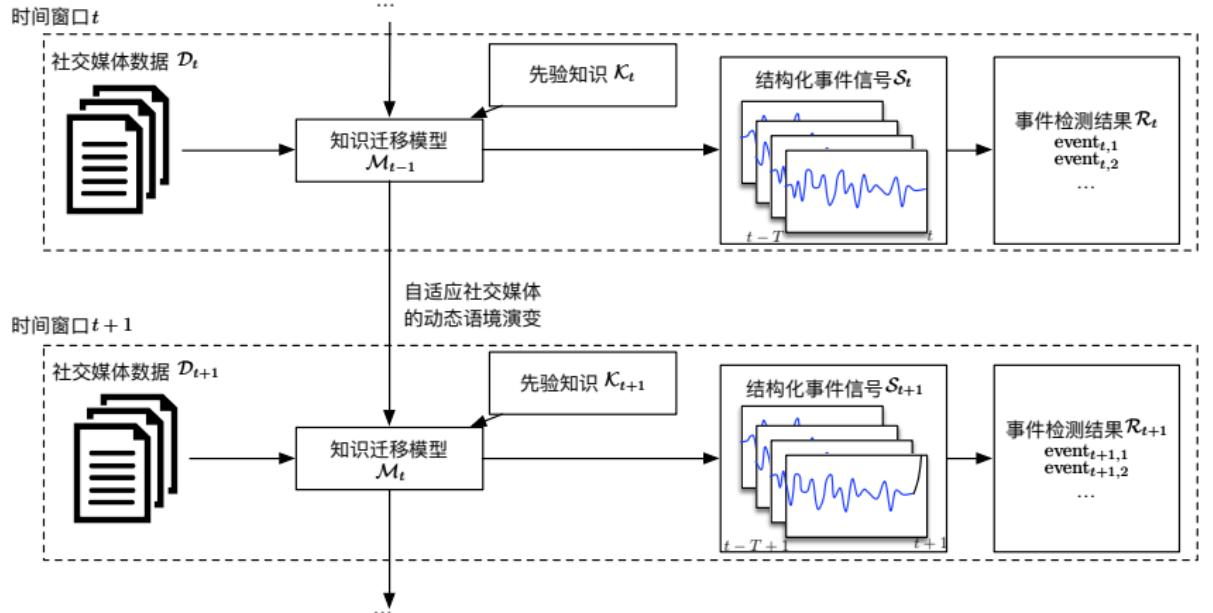
4. 用户兴趣建模与知识迁移方法

5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望



↓ K2Social框架依据时间轴展开的示意图



提升事件检测准确性

提升突发事件检测时效性

提升领域事件检测准确性

基于知识库结构的主题抽取与迁移方法  
TransDetector

用户兴趣建模与知识迁移方法 UMIETM

基于知识迁移的领域事件检测通用方法 Trans-Detector<sup>+</sup>

基于知识迁移的社交媒体事件检测通用框架 K2Social

# 大纲

1. 引言

2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

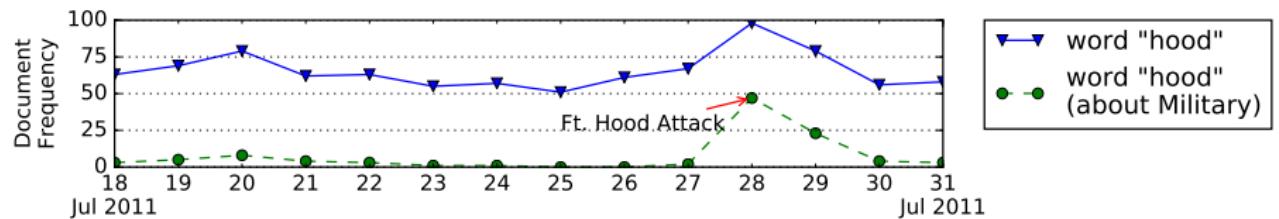
4. 用户兴趣建模与知识迁移方法

5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望

## 我们的观察：知识迁移与事件检测实例

图：词 *hood* 的原始时间序列，与经过迁移学习之后的军事相关的 *hood* 的时间序列对比图（在 *Edinburgh twitter corpus* 数据集上计算）。相关事件可参考[https://en.wikipedia.org/wiki/Fort\\_Hood#2011\\_attack\\_plot](https://en.wikipedia.org/wiki/Fort_Hood#2011_attack_plot).



### Fort Hood

From Wikipedia, the free encyclopedia

**Fort Hood** is a U.S. military post located in Killeen, Texas. The post is named after Confederate General John Bell Hood. It is

(a) 常用词 hood

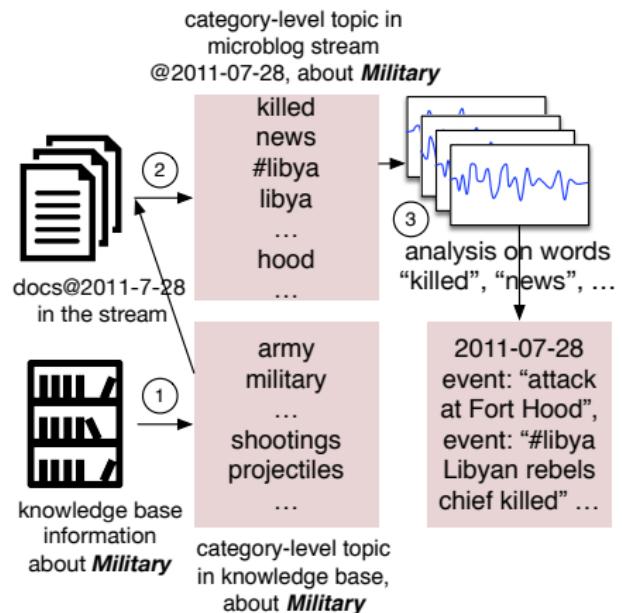
(b) 和军事相关的词 hood

# 我们的方法

面向概念类的迁移学习方法 TRANSDETECTOR，在 K2Social 框架下，分为三个阶段：

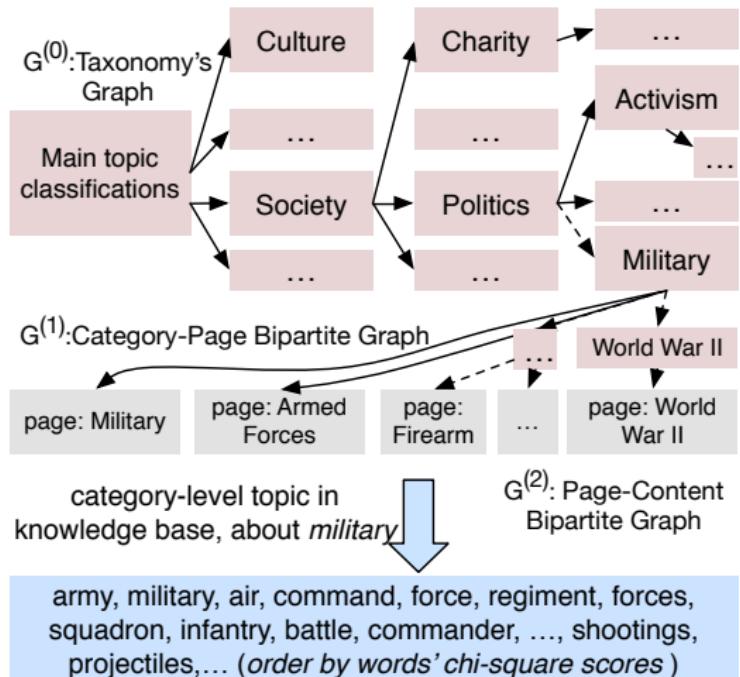
1. 主题抽取阶段
2. 迁移学习阶段
3. 事件检测阶段

图：TRANSDETECTOR 三阶段处理流程



# TRANSDETECTOR: 阶段 1 (基于知识库结构的主题抽取) (1/2)

图: 知识库层级结构与主题抽取 (以军事类别为例)



## TRANSDETECTOR: 阶段 1 (基于知识库结构的主题抽取) (2/2)

Military, Military\_terminology, Military\_terminology\_of\_Pakistan,  
Cold\_War\_terminology, Glossaries\_of\_the\_military, ...

分类类别

german 是一个常用词，在其他类别中也经常出现，所以卡方分数重计算之后在 Military 中的权重相应降低

## TRANSDETECTOR: 阶段 1 (基于知识库结构的主题抽取) (2/2)

Military, Military\_terminology, Military\_terminology\_of\_Pakistan,  
Cold\_War\_terminology, Glossaries\_of\_the\_military, ...

⇒  
Towarzysz\_pancerny, Derek\_Prince, No.\_21\_Squadron\_RAF,  
Benjamin\_Chew\_Tilghman, H.\_W.\_Gretton, ...

分类类别

实体对应的页  
面

german 是一个常用词，在其他类别中也经常出现，所以卡方分数重计算之后在 Military 中的权重相应降低

## TRANSDETECTOR: 阶段 1 (基于知识库结构的主题抽取) (2/2)

Military, Military\_terminology, Military\_terminology\_of\_Pakistan,  
Cold\_War\_terminology, Glossaries\_of\_the\_military, ...

⇒  
Towarzysz\_pancerny, Derek\_Prince, No.\_21\_Squadron\_RAF,  
Benjamin\_Chew\_Tilghman, H.\_W.\_Gretton, ...

⇒  
army (183815), air (141047), military (123486), force (116122), battle  
(95039), forces (94469), command (79866), ship (79478),  
german(76323), ... 括号中为原始词频

分类类别

实体对应的页  
面

原始主题

german 是一个常用词，在其他类别中也经常出现，所以卡方分数重计  
算之后在 Military 中的权重相应降低

## TRANSDETECTOR: 阶段 1 (基于知识库结构的主题抽取) (2/2)

Military, Military\_terminology, Military\_terminology\_of\_Pakistan,  
 Cold\_War\_terminology, Glossaries\_of\_the\_military, ...

⇒  
 Towarzysz\_pancerny, Derek\_Prince, No.\_21\_Squadron\_RAF,  
 Benjamin\_Chew\_Tilghman, H.\_W.\_Gretton, ...

⇒  
 army (183815), air (141047), military (123486), force (116122), battle  
 (95039), forces (94469), command (79866), ship (79478),  
 german(76323), ... 括号中为原始词频

⇒  
 army (742848), military (394538), air (390291), command (389632),  
 force (366014), regiment (364938), forces (319349), squadron (306569),  
 ..., german (107399), ... 括号中为卡方分数

分类类别

实体对应的页  
面

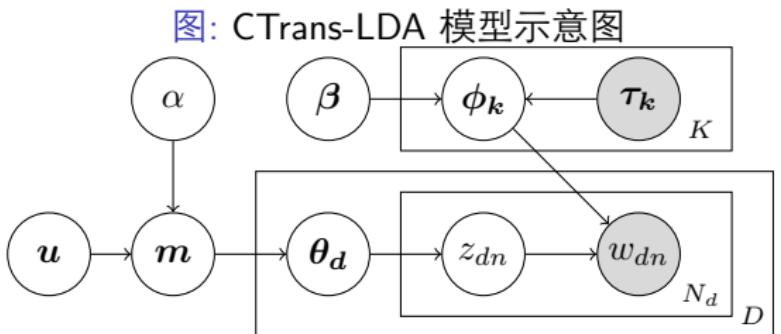
原始主题

经卡方分数重  
计算之后的主  
题

german 是一个常用词，在其他类别中也经常出现，所以卡方分数重计算之后在 Military 中的权重相应降低

## TRANSDETECTOR: 阶段 2 (从知识库到社交媒体的迁移学习) (1/3)

将知识库中分类类别  $c$  对应的主题  $\{\mathbf{h}_c\}_{c=1}^{K_{KB}}$  迁移到目标域，使用 CTrans-LDA.



在 CTrans-LDA 中卡方分数  $\{\mathbf{h}_c\}_{c=1}^{K_{KB}}$  用作先验知识， $S_k$  为和类别  $k$  相关的所有词的集合：

$$\tau_{kv} = \begin{cases} \lambda \frac{h_{kv}}{\sum_{v \in S_k} h_{kv}}, & v \in S_k \text{ and } k \leq K_{KB} \\ 0, & v \notin S_k \text{ or } k > K_{KB} \end{cases} \quad (1)$$

## TRANSDETECTOR: 阶段 2 (从知识库到社交媒体的迁移学习) (2/3)

下图是对称先验先验 VS 非对称先验的例子。 $\phi_k \sim Dir(\beta + \tau_k)$  中的  $\tau_k$  作用非常明显的，例如  $\tau_{Military, army}/\tau_{Military, basketball} = 203$ 。

图: (1.5, 1.5, 1.5)

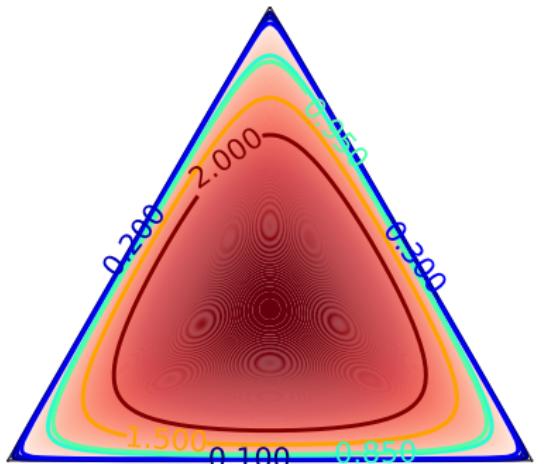
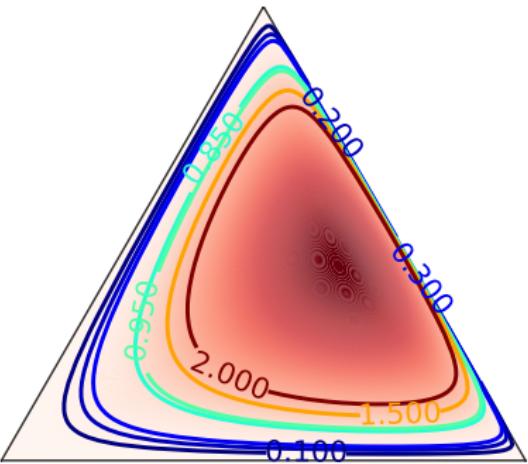


图: (1.5, 2.5, 2.5)



## TRANSDETECTOR: 阶段 2 (从知识库到社交媒体的迁移学习) (3/3)

使用吉布斯采样求解 CTrans-LDA:

- 初始概率  $\hat{q}_{k|v}$  使得社交媒体中的主题能够与知识库中的主题对齐.

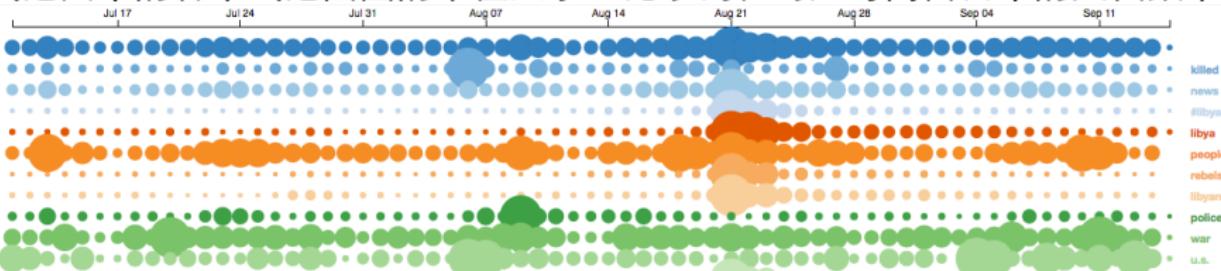
$$\hat{q}_{k|v} = \begin{cases} \frac{\tau_{kv}}{\sum_{k=1}^K \tau_{kv}}, \sum_k \tau_{kv} > 0 & (a) \\ 0, \sum_k \tau_{kv} = 0 \text{ and } k \leq K_{KB} & (b) \\ 1/(K - K_{KB}), \sum_k \tau_{kv} = 0 \text{ and } k > K_{KB} & (c) \end{cases} \quad (2)$$

- 吉布斯采样中的条件概率.

$$p(z_{dn} = k | .) \propto (n_{dk}^{(d)} + \alpha m_k)(n_{kv}^{(w)} + \tau_{kv} + \beta) / (n_{k,.}^{(w)} + \tau_{k,.} + V\beta).$$

# TRANSDETECTOR: 阶段 3 (在迁移学习后的时间序列上进行事件检测) (1/2)

图: 与 *Military* 概念类相关的词的时间序列的气泡图示意，在 *Edinburgh Twitter Corpus* 数据集上（2011 年 6 月 30 日至 2011 年 9 月 15 日）进行实验，气泡图中的各个气泡圆圈的半径大小正比于词在对应时间窗口中的文档频率。



## TRANSDETECTOR: 阶段 3 (在迁移学习后时间序列上进行事件检测) (2/2)

因自适应社交媒体语境动态演变，TRANSDETECTOR 能够更精确检测事件。

1. 检测事件候选词
  - 例如, *Ft., Hood, attack.*
2. 生成事件词组
  - 例如, *Ft. Hood attack.*
3. 召回事件相关的社交媒体文本
  - 例如, *Possible Ft. Hood Attack Thwarted* <http://t.co/BSJ33hk>.

# 实验设置 (1/3)

## 数据集

- 知识库：英文维基百科<sup>1 2</sup>，3,212,435 篇维基页面，973,125 个概念类别，TRANSDETECTOR 取  $K_{KB} = 50$ .
- 社交媒体数据流：Edinburgh twitter corpus<sup>3</sup>，36,627,434 条微博，时间跨度为 2011 年 6 月 30 日到 2011 年 9 月 15 日。

## 对比方法

- Tevent, BurstyBTM, LSH, EDCoW, TimeUserLDA

---

<sup>1</sup> <https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-categorylinks.sql.gz>

<sup>2</sup> <https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2>

<sup>3</sup> [http://demeter.inf.ed.ac.uk/cross/docs/fsd\\_corpus.tar.gz](http://demeter.inf.ed.ac.uk/cross/docs/fsd_corpus.tar.gz)

# 实验设置 (2/3)

## 事件评测基准

- 基准 1 (Benchmark1): 随 *Edinburgh twitter corpus* 数据集附带的手工标注的 27 个事件<sup>4</sup>.

表: 基准 1 中 27 个事件及对应的规模大小.

Event	Date	Event Size
S&P downgrade US credit rating	05/08/2011	656
Atlantis shuttle lands	21/07/2011	595
US increases debt ceiling	25/07/2011	485
Plane with Russian hocky team Lokomotiv crashes	07/09/2011	286
Amy Winehouse dies	23/07/2011	283
Gunman opens fire in youth camp in Norway	23/07/2011	260
Earthquake in Virginia	24/08/2011	246
First victim of London riots dies	09/08/2011	174
Explosion in French nuclear plant in Marcoule	12/09/2011	135
Google announces plans to bury Motorola Mobility	15/08/2011	127
NASA announces there might be water on Mars	04/08/2011	124
Car bomb explodes in Oslo, Norway	22/07/2011	114
...	...	...
Indian and Bangladesh sign a border pact	06/09/2011	25
Flight 4896 crash	13/07/2011	21
First artificial organ transplant	12/07/2011	18
three men die in riots in england	10/08/2011	16
rebels capture interntional tripoli airport	21/08/2011	13

<sup>4</sup> [http://demeter.inf.ed.ac.uk/cross/docs/Newswire\\_Events.tar.gz](http://demeter.inf.ed.ac.uk/cross/docs/Newswire_Events.tar.gz)

# 实验设置 (3/3)

## 事件评测基准

- 基准 2 (Benchmark2)：Tevent, BurstyBTM, LSH, EDCoW, TimeUserLDA, TRANSDETECTOR 提供的候选事件经人工和 Current\_events<sup>5</sup>对比，判断是否真实发生。总计395个事件。

The screenshot shows the Wikipedia Portal:Current events page for September 2011. At the top, there is a navigation bar with links for 'Portal', 'Talk', 'Read', 'Edit', 'View history', and a search bar. Below the navigation bar, the title 'Portal:Current events/September 2011' is displayed. A sub-navigation bar shows months from January to November. The main content area starts with a brief description of September 2011. Below this, a section titled 'Portal:Current events [edit]' is shown, stating it is an archived version from September 2011. It features a calendar for September 2011 with specific dates highlighted in blue. To the left of the calendar, there is a list of news items. One item is 'Armed conflict and attacks', and another is '2011 Libyan civil war: Libya's National Transitional Council'. The sidebar on the left contains links to various Wikipedia pages like Main page, Contents, Featured content, and Help.

<sup>5</sup> [https://en.wikipedia.org/wiki/Portal:Current\\_events](https://en.wikipedia.org/wiki/Portal:Current_events)

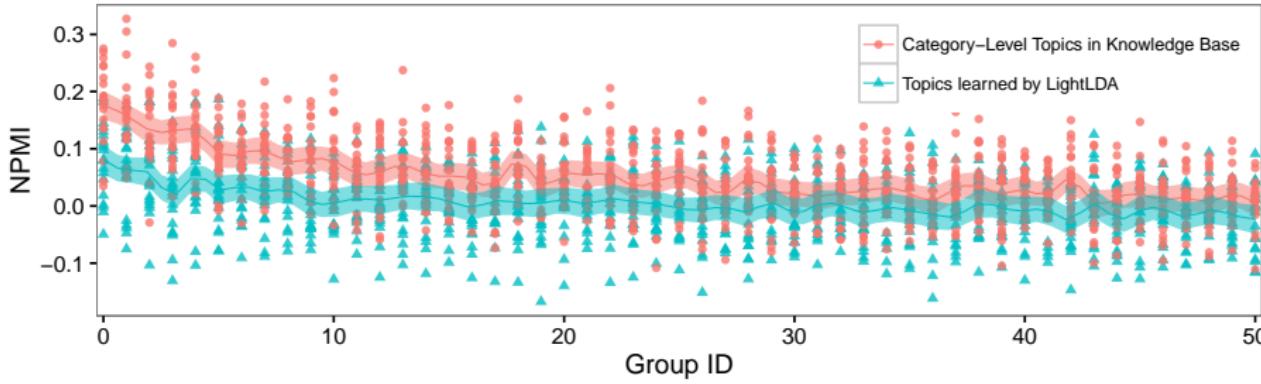
# 实验结果：基于知识库结构的主题抽取质量

表：以 *Aviation* 概念类别为例，TRANSDETECTOR 抽取的相关主题与 LightLDA(取主题个数为 100) 抽取的相关主题在主题连贯性 (NPMI) 指标上的对比。

Category-Level Topics extracted from Wikipedia by TRANSDETECTOR			Topics Learned from Wikipedia by LightLDA			
GID	#words*	words	GID	#words*	words	NPMI
-	1-5	aircraft air airport flight airline	-	1-5	engine aircraft car air power	-
0	1-5, 6-10	~, airlines aviation flying pilot squadron	0.113	0	1-5, 6-10	~, design flight model production speed
1	1-5, 11-15	~, flights pilots raf airways fighter	0.155	1	1-5, 11-15	~, system vehicle cars engines mm
2	1-5, 16-20	~, boeing runway force crashed flew	0.092	2	1-5, 16-20	~, fuel vehicles designed models type
3	1-5, 21-25	~, airfield landing passengers plane aerial	0.179	3	1-5, 21-25	~, version front produced rear electric
4	1-5, 26-30	~, bomber radar wing bombers crash	0.137	4	1-5, 26-30	~, space control motor standard development
5	1-5, 31-35	~, airbus airports operations jet helicopter	0.189	5	1-5, 31-35	~, film range light using available
6	1-5, 36-40	~, squadrons base flown havilland crew	0.088	6	1-5, 36-40	~, wing powered wheel weight launch
7	1-5, 41-45	~, combat luftwaffe aerodrome carrier fokker	0.159	7	1-5, 41-45	~, developed low test ford cylinder
8	1-5, 46-50	~, planes fly engine takeoff fleet	0.186	8	1-5, 46-50	~, equipment side pilot hp aviation
9	1-5, 51-55	~, fuselage helicopters aviator naval aero	0.157	9	1-5, 51-55	~, systems us sold body drive
10	1-5, 56-60	~, glider command training balloon faa	0.166	10	1-5, 56-60	~, gear introduced class safety seat
...	...	...	...	...	...	...
18	1-5, 96-100	~, scheduled carriers military curtiss biplane	0.131	18	1-5, 96-100	~, transmission special replaced limited different
19	1-5, 101-105	~, accident engines iaf albatross rcaf	0.068	19	1-5, 101-105	~, features machine nuclear even unit
						0.011

## 实验结果：基于知识库结构的主题抽取质量

图：对更多的 TRANSDETECTOR 从维基百科抽取的概念类相关的主题的连贯性与 LightLDA 在维基百科上学习的主题连贯性进行对比。在前 10 个分组上，TRANSDETECTOR 显著更优。



# 实验结果：迁移学习效果以及自适应社交媒体语境动态演变的效果

表: CTrans-LDA 迁移学习效果示例，展示从知识库中抽取的主题和经过迁移学习后微博中相关的主题。斜体标注的词为经过迁移学习后在社交媒体域中学习到的和概念类别相关的新词。

Aviation		Health		Middle East		Military		Mobile Phones	
Knowledge Base	Microblog Stream								
aircraft	air	health	weight	al	#syria	army	killed	android	iphone
air	plane	patients	loss	israel	#bahrain	military	news	mobile	apple
airport	flight	medical	diet	iran	people	air	#libya	nokia	android
flight	time	disease	health	arab	israel	command	libya	ios	app
airline	airlines	treatment	cancer	israeli	police	force	rebels	phone	ipad
airlines	news	hospital	lose	egypt	#libya	regiment	people	samsung	samsung
aviation	boat	patient	fat	egyptian	#egypt	forces	police	game	mobile
flying	airport	clinical	tips	ibn	news	squadron	war	app	blackberry
pilot	force	symptoms	treatment	jerusalem	#israel	infantry	libyan	iphone	tablet
squadron	fly	cancer	body	syria	world	battle	attack	htc	apps

# 实验结果：事件检测效果

TRANSDETECTOR 在保证召回率的同时，将准确率提升 9%

**表：**在基于 *Edinburgh twitter corpus* 数据集上构建的 Benchmark1 和 Benchmark2，各个事件检测方法的性能对比

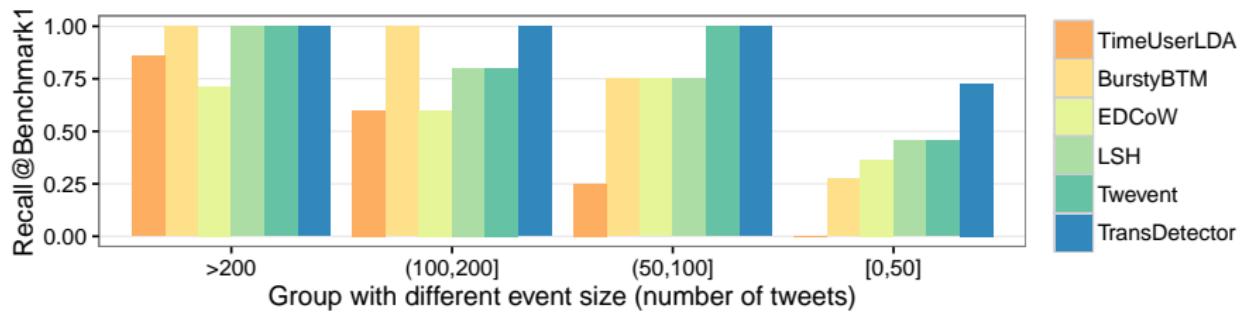
Method	Number of Events to be Evaluated	Recall@ Benchmark1	Precision@ Benchmark2	Recall@ Benchmark2	F@ Benchmark2	DERate <sup>a</sup> (Duplicate Event Rate)@ Benchmark2
LSH	500	0.704	0.788	0.651	0.713	0.348
TimeUserLDA	100	0.370	0.790	0.177	0.289	0.114
Twevent	375	0.741	0.808	0.658	0.725	0.142
EDCoW	349	0.556	0.748	0.511	0.607	0.226
BurstyBTM	200	0.667	0.825	0.384	0.497	<b>0.079</b>
TRANSDETECTOR	457	<b>0.889</b>	<b>0.912</b>	<b>0.876</b>	<b>0.894</b>	0.170

<sup>a</sup> DERate = (the number of duplicate events) / (the total number of detected realistic events)

## 实验结果：事件检测效果

TRANSDETECTOR 在较小规模的事件上召回率明显高于已有方法。

图：在基于 *Edinburgh twitter corpus* 数据集构建的 Benchmark2 上，召回率与事件规模之间的关系示意图



# 实验结果：事件检测效果

TRANSDETECTOR 在较小规模的事件上召回率明显高于已有方法。

表: 2011-07-22 至 2011-07-28, *Edinburgh Twitter Corpus* 数据集中军事相关事件列表以及各事件检测方法的具体表现

Date	Event key words	Representative event tweet	Number of event tweet	Methods <sup>a</sup>					
				L	TU	TW	E	B	TD
7/22/11	Norway, Oslo, attacks, bombing	Terror Attacks Devastate Norway: A bomb ripped through government offices in Oslo and a gunman... <a href="http://dlvr.it/cLbk8">http://dlvr.it/cLbk8</a>	557	✓	✓	✓	✓	✓	✓
7/23/11	Gunman, rink	Gunman Kills Self, 5 Others at Texas Roller Rink <a href="http://dlvr.it/cLcTH">http://dlvr.it/cLcTH</a>	43	-	-	✓	✓	-	✓
7/26/11	Kandahar, mayor, suicide, attack	TELEGRAPH]: Kandahar mayor killed by Afghan suicide bomber: The mayor of Kandahar, the biggest city in south _	47	✓	-	✓	✓	-	✓
7/28/11	Ft., Hood, attack	Possible Ft. Hood Attack Thwarted <a href="http://t.co/BSJ33hk">http://t.co/BSJ33hk</a>	52	-	-	-	-	-	✓
7/28/11	Libyan, rebel, gunned	Libyan rebel chief gunned down in Benghazi <a href="http://sns.mx/prfvyl">http://sns.mx/prfvyl</a>	44	-	-	-	-	-	✓

<sup>a</sup> L=LSH, TU=TimeUserLDA, TW=Twevent, E=EDCoW, B=BurstyBTM, TD=TRANSDETECTOR.

## TransDetector 方法小结

1. 提出了一种以知识库结构为导向的概念类相关主题抽取方法；
2. 提出了新的概率主题模型 CTrans-LDA 用于知识的迁移，并能自适应社交媒体的语境动态演变；
3. 在知识迁移后的时间序列上进行事件检测，准确地检测社交媒体中蕴含的事件；
4. Edinburgh Twitter Corpus 数据集上准确率相较于目前已有的最佳方法提升了 9%

# 大纲

1. 引言

2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

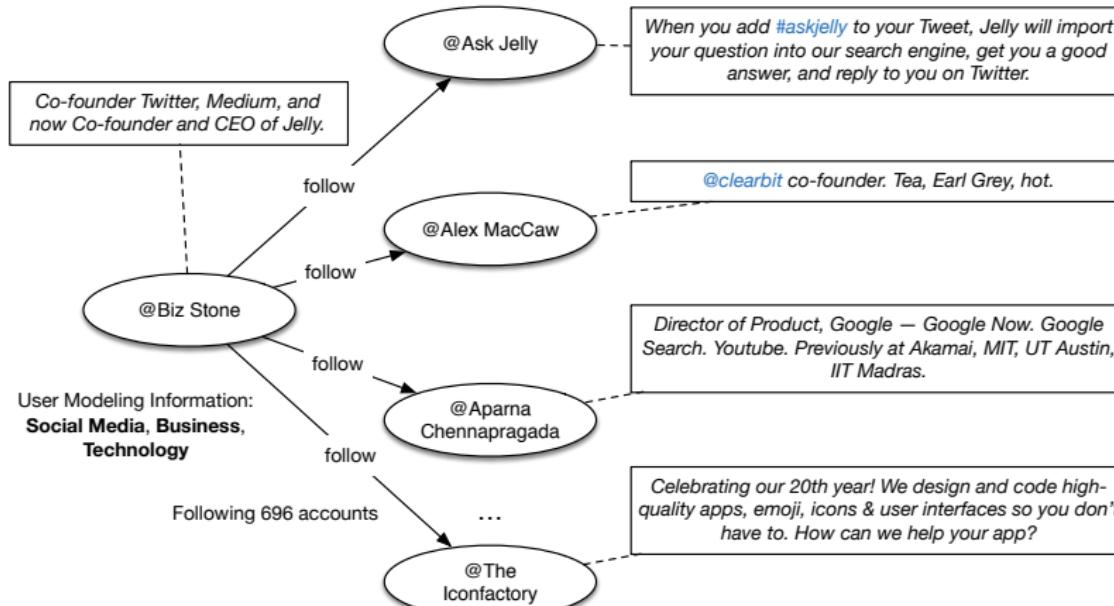
4. 用户兴趣建模与知识迁移方法

5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望

# 我们的观察

图：社交媒体关注网络用于扩充个人自我描述信息示意图

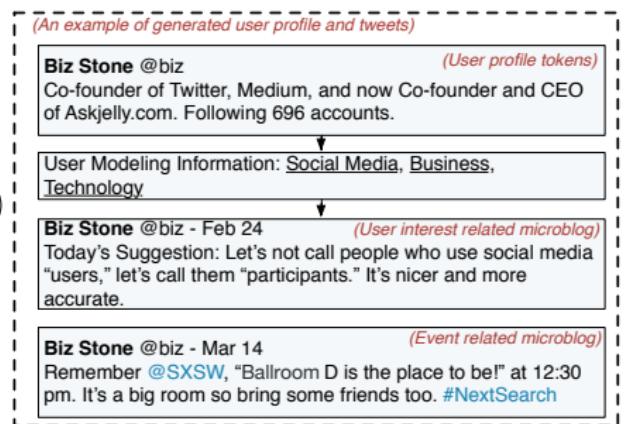
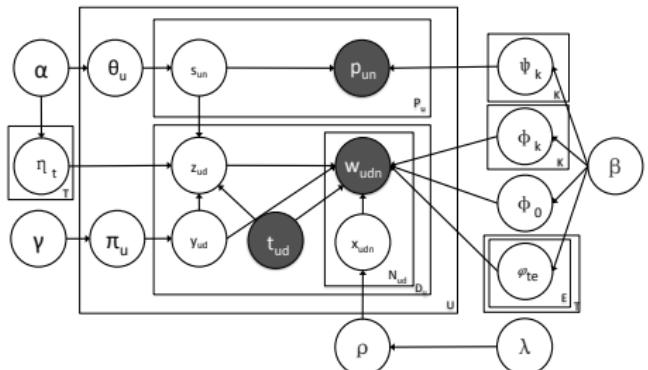


# 我们的方法

## UMIETM 模型 (User Modeling Based Interest and Event Topic Modeling)

- 使用用户个人描述信息与用户关注网络建模用户兴趣分布
- 将用户兴趣知识迁移到微博中
- 区分用户兴趣与突发事件

图: UMIETM 概率模型



(左侧) UMIETM 概率模型示意图。(右侧) 示例用户 Biz Stone 的两类文本（用户兴趣相关的文本与突发事件相关的文本）的示意图。其个人简档揭示出其个人兴趣主要集中在社交网络、商业、科技领域。依据 UMIETM 模型，可以检测出他 2016 年 2 月 24 日发布的有关社交网络中用户角色定位的文本可以被检测为与个人兴趣相关；2016 年 3 月 14 日发布的有关 SXSW 的文本则与个人兴趣都不相关，结合该时间窗口内更多的其他文本，可以将判断这条文本和突发事件相关，事实上 SXSW 是一个在 3 月 11 日到 20 日之间举行的盛大音乐节。

# UMIETM 知识迁移算法

1. 根据用户个人描述信息以及关注网络计算用户兴趣分布：

$$p(s_{un} = k | s_{-un}, \vec{p}, \alpha, \beta) \\ \propto \frac{c_{uk}^{(p)} + \alpha}{c_{u,.}^{(p)} + K\alpha} \frac{c_{kv}^{(p)} + \beta}{c_{k,.}^{(p)} + V\beta}$$

2. 用户兴趣知识迁移至社交媒体文本内容中：

- 和突发事件相关

$$p(y_{ud} = 0, z_{ud} = k | \vec{y}_{-ud}, \vec{z}_{-ud}, \vec{t}, \vec{w}, \vec{s}, \alpha, \beta, \gamma)$$

- 和用户兴趣相关

$$p(y_{ud} = 1, z_{ud} = e | \vec{y}_{-ud}, \vec{z}_{-ud}, \vec{t}, \vec{w}, \vec{s}, \alpha, \beta, \gamma)$$

**Algorithm 1:** UMIETM 知识迁移算法

```

1 initiate the topic label and the statistics
2 for  $i = 1 : l_1$  do
3   for  $u$  in user set  $\mathcal{U}$  do
4     for  $n = 1 : P_u$  do
5       sample profile's hidden topic  $s_{un}$ , update
           $s_{un}$ ,  $c_{u,k}^{(p)}$  and  $c_{k,v}^{(p)}$ 
```

```

6 for iteration  $i = 1 : l_2$  do
7   for  $t = 1 : T$  do
8     for  $u$  in user set  $U_t$  do
9       for  $d = 1 : D_u$  do
10      sample  $y_{ud}$  and  $z_{ud}$ 
11      if  $y_{ud} = 0$  then
12        update  $z_{ud}$ ,  $y_{ud}$ ,  $c_u^{(0)}$ ,  $c_{u,k}^{(0)}$ ,  $c_{k,v}^{(0)}$ 
13      else
14        update  $z_{ud}$ ,  $y_{ud}$ ,  $c_u^{(1)}$ ,  $c_{t,e}^{(1)}$ ,  $c_{t,e,v}^{(1)}$ 
15      for  $n$  in  $1, \dots, N_{ud}$  do
16        sample  $x_{udn}$ 
17        if  $x_{udn} = 0$  then
18          update  $x_{udn}$ ,  $M_0^p$ ,  $c_v^{(B)}$ 
19        else
20          update  $x_{udn}$ ,  $M_0^p$ ,  $c_u^{(0)}$ ,  $c_v^{(1)}$ 
```

# UMIETM 模型在线学习，提升突发事件检测时效性

图: UMIETM 批量学习



图: UMIETM 在线学习

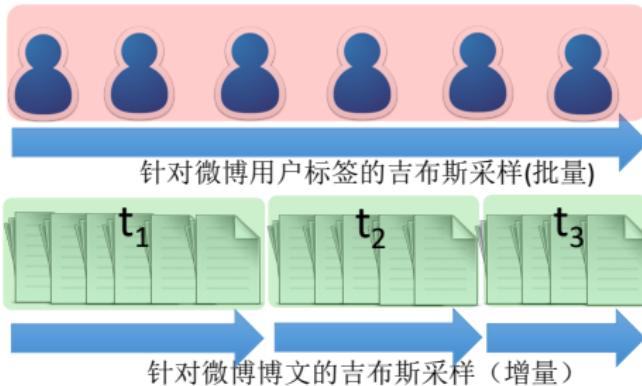
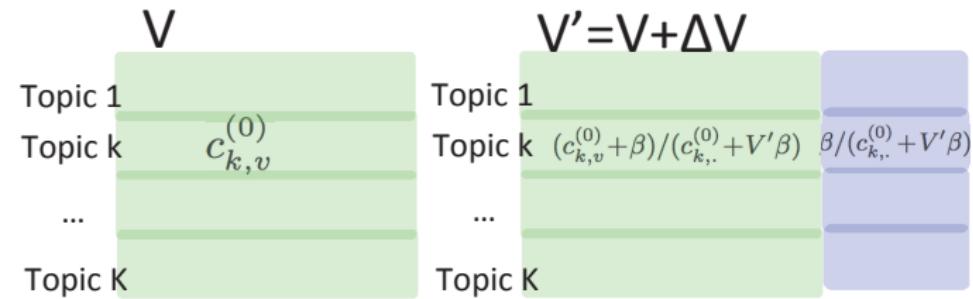


表: UMIETM 批量学习与在线学习时间复杂度与空间复杂度对比表，其中  
 $|W_t| \ll |W|$

	时间复杂度	空间复杂度
UMIETM 批量学习方法	$O(I_1 K P  + I_2(K + E) W )$	$O( P  +  W )$
UMIETM 在线学习方法	$O(I_1 K P  + I_2(K + E) W_t )$	$O( P  +  W_t )$

# UMIETM 对社交媒体动态语境演变的自适应

社交媒体的语境动态演变不断带来新词，新词也意味着新的事件，UMIETM 需要相应自适应机制。



图：社交媒体数据流上新增的词汇采用 *smoothing* 的方式初始化

## 实验设置 (1/2)

- 新浪微博数据集 2012.1.1-2012.12.31

- 数据预处理

1. 数据集按周进行分割，得到时间窗口文件
2. 中文分词（ICTCLAS 2013）
3. 移除停用词与低频词（时间窗口内文档频度 <3）
4. 移除过短的微博文本（词数 <3）

表：预处理后的新浪微博数据集上的统计信息

	用户数	微博数	微博中词的数量
全年	252,369	16,421,167	251,686,571
第 1 周	9,785	31,503	440,217
第 2 周	29,721	242,554	3,679,979
第 3 周	30,891	254,698	3,881,633
第 4 周	29,788	237,456	3,510,934
第 5 周	24,256	190,037	2,749,539
...	...	...	...

## 实验设置 (2/2)

这一页的实验说明检验模型的时效性  
是否要说是在 MALLET 上实现 UMIETM 算法，并在 8 核的 2.00HZ,  
64GB 内存的服务器上运行。

- 社交媒体数据流 Large twitter corpus (相比于 Edinburgh twitter corpus):
  - 61,2091,163 条微博
  - 7,695,256 个用户
  - 时间跨度: 2011 年 7 月 1 日到 2011 年 12 月 18 日。
  - 以 10 分钟为单位进行数据分片。
- 突发事件基准: Wikipedia 上用户整理的事件列表 2011 in the United States#Events<sup>6</sup>, 43 个突发事件。

<sup>6</sup>[https://en.wikipedia.org/wiki/2011\\_in\\_the\\_United\\_States#Events](https://en.wikipedia.org/wiki/2011_in_the_United_States#Events)

## 实验结果：UMIETM 模型有效性

- 对比方法：Author-LDA, twitterLDA, timeUserLDA，以及不使用用户个人描述信息进行用户建模的变体方法 IETM
- 度量指标：困惑度 (perplexity)

$$\text{perplexity}(D_{\text{test}}) = \exp \left\{ -\frac{\sum_{u=1}^U \sum_{d=1}^{D_u} \log p(w_{ud})}{\sum_{u=1}^U \sum_{d=1}^{D_u} N_{ud}} \right\} \quad (4)$$

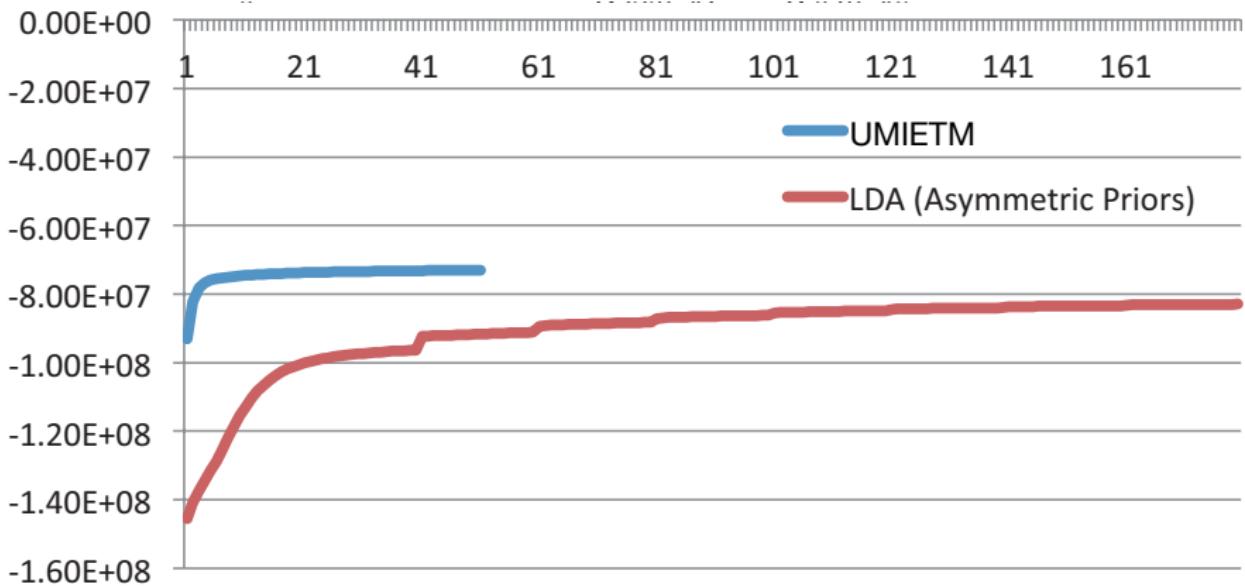
表：UMIETM 与对比方法的困惑度比较（数值较小者文本建模的有效性更强）

Author-LDA	TwitterLDA	TimeUserLDA	IETM	UMIETM
20422.25	6027.47	4810.92	3926.76	3107.83

- 从表中可以看出 IETM 和 UMIETM 都比其他方法更优，说明区分用户兴趣与突发事件有助于提升文本建模效果；
- 而 UMIETM 比 IETM 更优，说明通过用户个人简档和关注网络信息进行的建模比单纯建模社交媒体文本更优。

# 实验结果：UMIETM 算法收敛速度

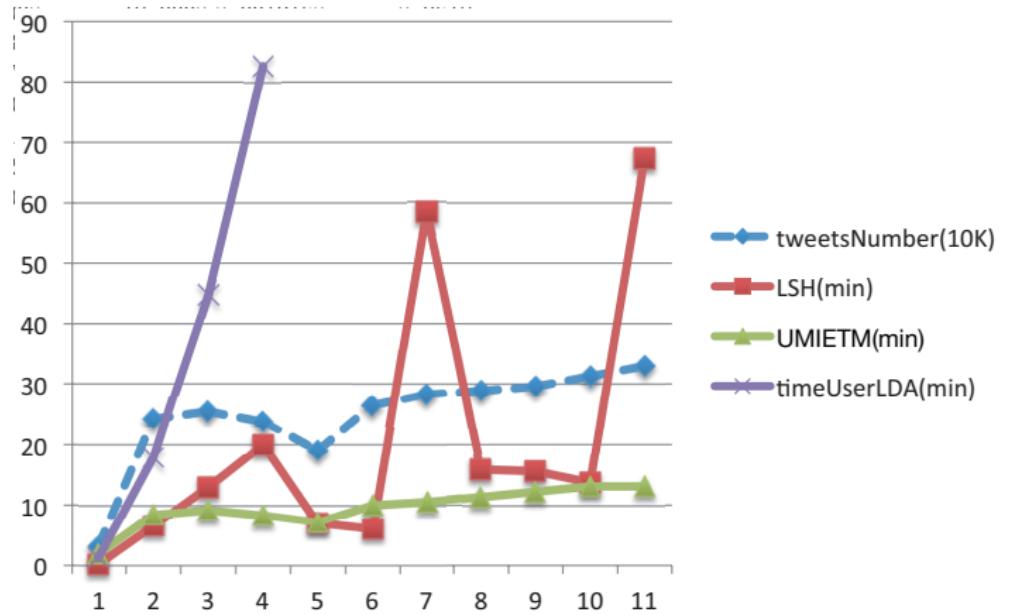
图：UMIETM 迭代 10 次之后，完全对数似然函数值的增长幅度小于 0.15%



## 实验结果：UMIETM 事件检测效率

timeUserLDA 使用批量学习方法，时间复杂度随着数据集大小线性增长  
基于 LSH 的方法也具备在线学习能力，但时间复杂度不稳定，随着桶内数据大小而变化

图：UMIETM 在线学习方法用于事件检测的运行效率



## 实验结果：突发事件检测准确率与召回率

注意是否要增加一个 Tevent 的实验，或者我们要不要增加一个新的结果（最好是 2016 或者 2017 的）

表：UMIETM 与对比方法在新浪微博数据集（2012.1-2012.12）上突发事件检测任务的准确率与召回率比较

	precision	recall	F
UMIETM	0.894	0.913	0.903
UMIETM(-)	0.847	0.697	0.765
IETM	0.824	0.536	0.650
LSH	0.394	0.913	0.550
EDCoW	0.731	0.435	0.545

# 实验结果：社交媒体数据流中检测突发事件时效性

表: Large twitter corpus 数据集中的突发事件，UMIETM 平均比 TW 提前 1.4 小时检测到

Date	Representative event tweet	Methods <sup>a</sup>		
		LSH	TW	UMIETM
9/17/11	Day of Rage occupies Wall Street Democracy in Action! Watching globalrevolution <a href="http://livestre.am/PINN">http://livestre.am/PINN</a> via @livestream	4:10PM	3:50PM	3:30PM
...	...	...	...	...
11/4/11	Hotmail - mikehuot@msn.com - Windows Live <a href="http://bit.ly/vRDsgg">http://bit.ly/vRDsgg</a> via @addthis Andy Rooney's Essay on Prayer! Gonna miss that old man!	11:20PM	9:10PM	7:30PM
11/7/11	Jerry Sandusky Accused Of Sexually Assaulting 8 Boys..	4:50PM	4:20PM	4:00PM
11/8/11	Here, I'm going to declare Phil Bryant a winner in the Mississippi gubernatorial race right now. Take that, CNN.	-	-	1:40PM
11/8/11	Miss. voters asked if life begins at conception: JACKSON, Miss. (AP) – Mississippi voters were asked Tuesday... <a href="http://apne.ws/sjgCMh">http://apne.ws/sjgCMh</a>	10:00PM	12:20PM	11:50AM
11/8/11	Looks like Steve Beshear is going to easily win reelection!	-	4:50PM	3:50PM
11/8/11	Will Republicans seize control of the Virginia state Senate? So far, our readers say yes. <a href="http://patch.com/A-n6G7">http://patch.com/A-n6G7</a> What do you think? #va31	-	4:40PM	3:20PM
11/8/11	Is it the end of Russell Pearce's future as state senator? Recall Vote today. - <a href="http://bit.ly/tZHNMM">http://bit.ly/tZHNMM</a>	-	2:40PM	1:30PM
...	...	...	...	...
12/18/11	The Iraq war is over, but the war on Iran is just beginning	11:20AM	9:50AM	9:30AM

<sup>a</sup> LSH, TW=Twevent

## UMIETM 方法小结

1. 提出了一种用户兴趣建模与知识迁移方法 UMIETM，检测社交媒体突发事件；
2. 使用用户个人描述信息与关注网络，建模用户兴趣并迁移到社交媒体文本中，区分用户个人兴趣相关的文本和由外部突发事件引起的文本；
3. 将 UMIETM 扩展为在线学习的方式，应用于实际的社交媒体数据流；
4. 相较于已有方法能够提前 1.4 小时检测突发事件

- 1, 本文提出了一种用户兴趣建模与知识迁移方法 UMIETM，检测社交媒体突发事件。
- 2, UMIETM 使用用户个人描述信息与关注网络，学习出用户个人兴趣，并将用户个人兴趣的知识迁移到社交媒体数据流中，区分用户个人兴趣相关的文本和由外部突发事件引起的文本。
- 3, 将 UMIETM 扩展为在线学习的方式，应用于实际的社交媒体数据流。
- 4, 相较于已有方法能够提前 1.4 小时检测突发事件

# 大纲

1. 引言

2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

4. 用户兴趣建模与知识迁移方法

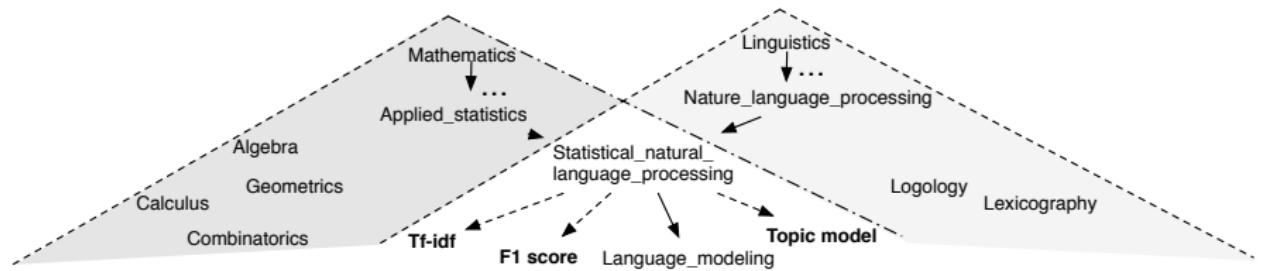
5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望

# 领域事件检测与知识抽取的优化

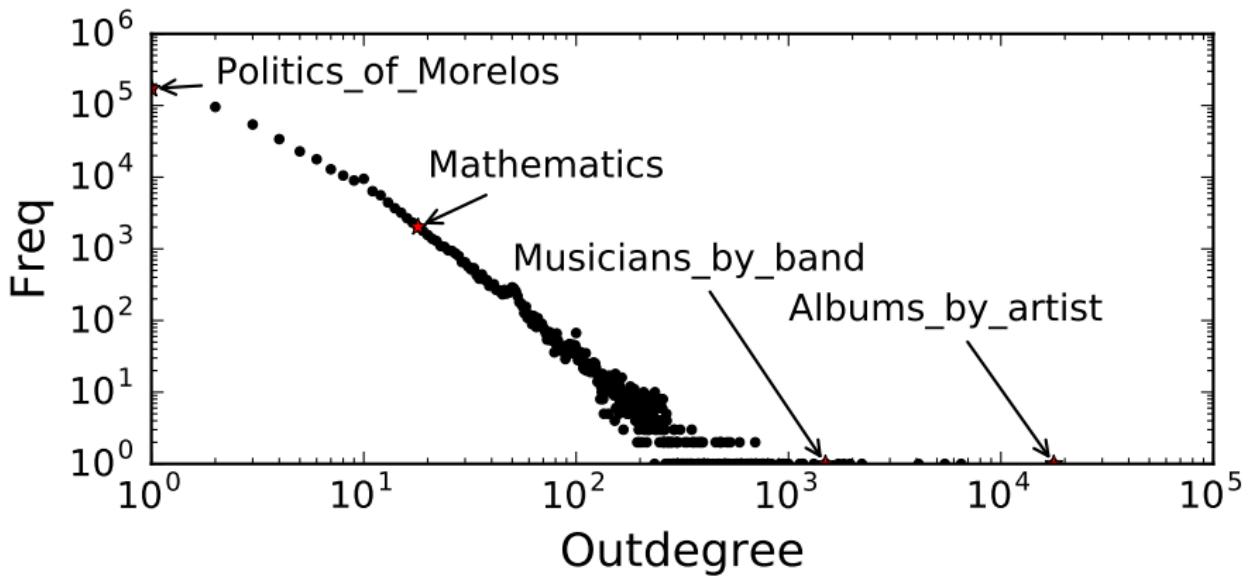
TaxoPhrase 用于从知识库中抽取领域知识的示意图。用户 1 关心数学领域中除统计自然语言处理（白色区域）之外的知识（深灰色区域），用户 2 关心语言学领域中除统计自然语言处理之外的知识（浅灰色区域），在此场景中 TaxoPhrase 可将上述三部分用主题建模的方法加以探索与抽取。

图：TaxoPhrase 用于从维基百科知识库中抽取领域知识的示意图



# 领域事件检测与知识抽取的优化

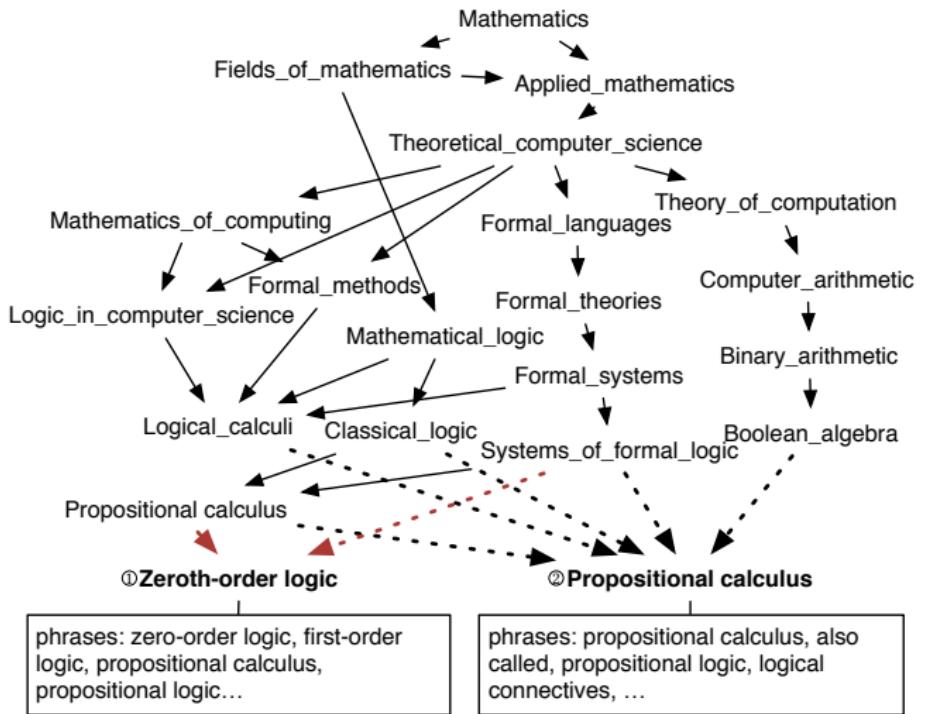
图：英文维基百科的分类体系中各类别节点的出度的分布



可以考虑给图的左侧加上分类类别、实体、短语的标注

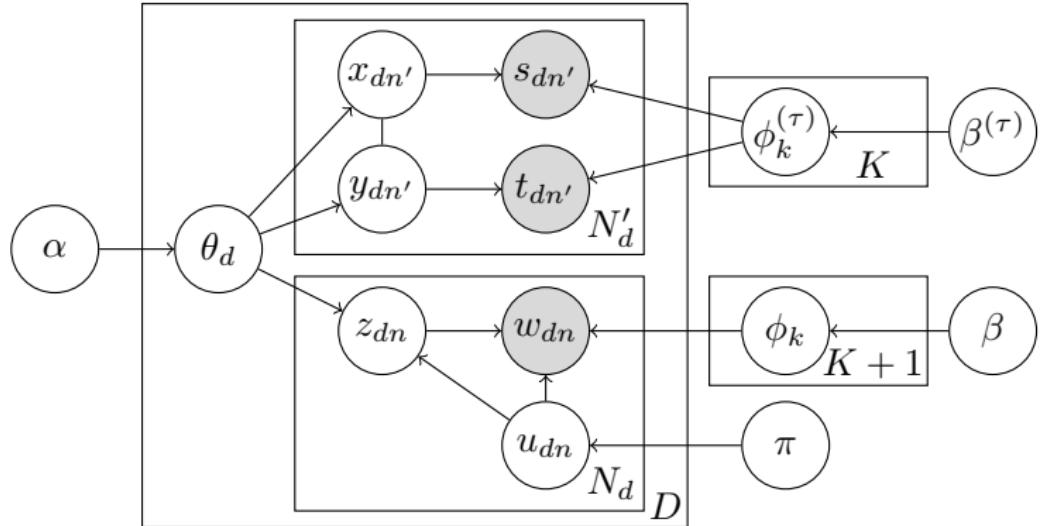
# 领域事件检测与知识抽取的优化

图：知识库中有互补关系的三部分的示例



# TaxoPhrase 模型

图：用于探索知识库中领域知识的概率模型 TaxoPhrase 的示意图



# 实验结果：TaxoPhrase 的建模质量

图：TaxoPhrase 方法在 Mathematics@Wiki 上学得的主题

<b>Topic 1</b>
(Categories) <i>Mathematics_awards</i> , <i>Mathematicians_by_award</i> , <i>Mathematicians_by_nationality</i> , <i>Mathematicians_by_field</i>
(Entities) John Cedric Griffiths Teaching Award, Santosh Vempala, Aisenstadt Prize, Subhash Suri, David P. Dobkin
(Phrases) university of california, american mathematical society, professor of mathematics, princeton university, computer science, harvard university, american mathematician, stanford university, massachusetts institute of technology, columbia university
<b>Topic 2</b>
(Categories) <i>Geometry_stubs</i> , <i>Differential_geometry_stubs</i> , <i>Elementary_geometry_stubs</i> , <i>Polyhedron_stubs</i>
(Entities) Enneadecahedron, Icosahedral pyramid, Expanded icosidodecahedron, Pentadecahedron, Cubic cupola
(Phrases) three dimensional, platonic solids, johnson solids, uniform polyhedron compound, symmetry group, regular dodecahedron, triangular faces, vertex figure, non-convex uniform polyhedron, four dimensional
<b>Topic 3</b>
(Categories) <i>Topology_stubs</i> , <i>Knot_theory_stubs</i> , <i>Theorems_in_topology</i> , <i>Theorems_in_algebraic_topology</i>
(Entities) Knot operation, Chromatic homotopy theory, Infinite loop space machine, Simple space, Base change map
(Phrases) topological space, algebraic topology, category theory, topological spaces, fundamental group, simply connected, homotopy theory, 3 manifold, 3 manifolds,

# 实验结果：TaxoPhrase 的建模质量

表：各数据集统计信息，以及各方法获得的短语和分类类别两种主题的质量对比（以 PMI 为评测指标）

		Maths	Chemistry	Argentina
#Entities		27,947	60,375	8,617
#Category Types		1,391	3,038	1,479
#Phrase Types		116,013	248,769	21,183
on phrases	LDA	4.55	4.30	3.52
	TaxoPhrase	4.67	4.55	3.81
on categories	SSN-LDA	4.01	3.97	3.06
	TaxoPhrase	4.51	4.48	3.73

# 实验结果：TaxoPhrase 抽取领域知识的实例

图：TaxoPhrase 方法在 Feminism@Wiki 上学得的主题

Topic 10	<i>Gender Equality</i>
(Entities)	Marriage bar, Timeline of women's rights (other than voting), Annestine Beyer, Equal Treatment Directive, Alliance for Female Equality
(Phrases)	gender equality, equal rights, equal pay, feminist movement, working women, higher education, equal opportunity, female education, female students, patriarchal society
Topic 11	<i>Feminist</i>
(Entities)	Clenora Hudson-Weems, We Real Cool: Black Men and Masculinity, Black feminism, Women of Labour, Redstockings
(Phrases)	black women, feminist movement, black feminist, black feminism, white women, radical feminist, radical feminists, white privilege, african americans, radical feminism
Topic 18	<i>Gender Identity</i>
(Entities)	Bem Sex-Role Inventory, Gender and emotional expression, Sex differences in intelligence, Sex differences in schizophrenia, Sex differences in leadership
(Phrases)	gender identity, sexual orientation, gender role, sex differences, gender differences, hegemonic masculinity, gender stereotypes, gender binary, biological sex, gender bias

# 实验结果：TransDetector<sup>+</sup> 领域事件检测

表：2011年6月30日至2011年9月15日，Edinburgh twitter corpus 数据集中 Feminist（女权主义）相关事件列表及各事件检测方法在各事件上的具体表现

Date	Representative event tweet	Number of event tweet	Methods <sup>a</sup>					
			L	TU	TW	E	B	TD+
7/1/11	DSK Has Bail Lifted Over <b>Sex Assault Case</b> : Dominique Strauss-Kahn has had his bail lifted after prosecutors said... <a href="http://bit.ly/lIcWhN">http://bit.ly/lIcWhN</a>	46	-	-	-	-	-	✓
7/26/11	bad! David Wu resigns because of an <b>unwanted sexual encounter</b> with an 18-year-old	23	-	-	-	-	-	✓
8/17/11	Nevin Shapiro said he provided players with <b>sexual bribery</b> and cars over years, and NCAA is investigating <a href="http://dlvr.it/cAaMF">http://dlvr.it/cAaMF</a>	52	✓	-	✓	✓	-	✓
8/19/11	Obama relieves illegal immigrants who are students, veterans, the elderly, crime victims and those with family, including <b>same-sex partners</b>	21	-	-	✓	-	-	✓

<sup>a</sup> L=LSH, TU=TimeUserLDA, TW, E=EDCoW, B=BurstyBTM, TD+=TaxoPhrase+TRANSDETECTOR

# 实验结果：TransDetector<sup>+</sup> 领域事件检测

表: *Edinburgh twitter corpus* 数据集中 Natural Hazards (自然灾害) 相关事件

Date	Representative event tweet	Number of event tweet	Methods <sup>a</sup>					
			L	TU	TW	E	B	TD+
7/2/11	Exxon oil spill in Mont. river prompts evacuations [AP] - An ExxonMobil pipeline that runs under the Yellowstone River <a href="http://tiny.ly/IGuN">http://tiny.ly/IGuN</a>	22	-	-	-	-	-	✓
7/5/11	@438PM Watching storms form around #Phoenix, potential for a dust storm 6-9PM. #Tucson about to get hit. #azwx <a href="http://www.weather.gov/phoenix">http://www.weather.gov/phoenix</a>	40	-	-	✓	-	-	✓
7/11/11	Possible earthquake east coast of Honshu, JAPON ! 48hrs. close attention.	14	-	-	-	-	-	✓
7/21/11	DTN France: Deadly heat-wave spreads in US: A punishing heat-wave settles over the central and eastern US, with ... <a href="http://bit.ly/q5mRkl">http://bit.ly/q5mRkl</a> .	249	✓	-	✓	✓	-	✓
8/23/11	More Virginia Earthquake 2011: Philadelphia Eagles Feel Quake In Locker Room (VIDEO) <a href="http://post.ly/2yOQ8">http://post.ly/2yOQ8</a>	310	✓	✓	✓	✓	✓	✓
8/28/11	@CBSBigBrother Anything but Hurricane Irene	1458	✓	✓	✓	✓	✓	✓
9/1/11	Tropical storm Lee in the Gulf of Mexico showed up randomly like at mama's house looking to borrow a few dollars.	159	✓	-	✓	-	-	✓
9/5/11	Satellite loop of the wildfires in Texas <a href="http://fb.me/Zbve5o4E">http://fb.me/Zbve5o4E</a>	21	-	-	-	✓	-	✓

<sup>a</sup> L=LSH, TU=TimeUserLDA, TW=EDCoW, B=BurstyBTM, TD+=TaxoPhrase+TRANSDETECTOR

## TransDetector<sup>+</sup> 方法小结

1. 提出了一种适用于各领域的领域事件检测通用方法 TransDetector<sup>+</sup>；
2. 提出概率模型 TaxoPhrase 优化从知识库中抽取领域知识的过程；
3. TransDetector<sup>+</sup> 将抽取出的领域知识迁移至社交媒体数据流中，检测领域事件；
4. 在 4 个示例领域中进行实验，相较于已有方法，平均将 F 值提升了 21%

本文提出了一种适用于各领域的领域事件检测通用方法 TransDetector<sup>+</sup>。TransDetector<sup>+</sup> 方法利用本文提出的概率模型 TaxoPhrase 优化从知识库中抽取领域知识的过程，并将抽取出的领域知识迁移至社交媒体数据流中，提取和领域相关的词和短语的时间序列，进而检测领域事件。TransDetector<sup>+</sup> 在 4 个示例领域中进行了实验，相较于已有方法，平均将 F 值提升了 21%

# 大纲

1. 引言

2. 基于知识迁移的社交媒体事件检测通用框架 K2Social

3. 基于知识库结构的主题抽取与迁移方法

4. 用户兴趣建模与知识迁移方法

5. 基于知识迁移的领域事件检测通用方法

6. 总结与展望

# 总结

## 基于知识迁移的社交媒体事件检测通用框架 K2Social

基于知识库结构的主题抽取  
与迁移方法  
**TransDetector**

用户兴趣建模  
与知识迁移方  
法 UMIETM

基于知识迁  
移的领域事  
件检测通用  
方法 **Trans-  
Detector<sup>+</sup>**

提升事件检测准确性，  
实验中提升 9%

提升突发事件检测时  
效性，实验中提前  
1.4 小时

提升领域事件检测准  
确性，实验中 F 值提  
升 21%

# 未来工作展望

扩展基于知识迁移的社交媒体事件检测通用框架 K2Social

扩展面向领域的知识迁移方法

更多与领域相关的社会计算任务：如心理危机干预，网络亚文化社区检测等

1. **Weijing Huang**, Tengjiao Wang, Wei Chen, Siyuan Jiang, Kam-Fai Wong,  
PhraseCTM: Correlated Topic Modeling on Phrases within Markov Random Fields,  
ACL 2018, accepted, to appear.
2. **Weijing Huang**, Tengjiao Wang, Wei Chen, Yazhou Wang, Category-Level  
Transfer Learning from Knowledge Base to Microblog Stream for Accurate Event  
Detection, DASFAA 2017. (EI index number: 20174404323179)
3. **Weijing Huang**, Wei Chen, Tengjiao Wang, Shibo Tao. TaxoPhrase: Exploring  
Knowledge Base via Joint Learning of Taxonomy and Topical Phrases, the 2nd  
Open Knowledge Base and Question Answering Workshop at SIGIR 2017.
4. **Weijing Huang**, Wei Chen, Tengjiao Wang, Shibo Tao. Efficient Topic Modeling  
on Phrases via Sparsity, the 29th IEEE International Conference on Tools with  
Artificial Intelligence (ICTAI) 2017.
5. **Weijing Huang**, Wei Chen, Lamei Zhang, Tengjiao Wang, An Efficient Online  
Event Detection Method for Microblogs via User Modeling, the 18th Asia Pacific  
Web Conference (APWeb) 2016. (EI index number: 20164102880250)

6. Ruhui Wang, **Weijing Huang**, Wei Chen, Tengjiao Wang, Kai Lei. ASEM: Mining Aspects and Sentiment of Events from Microblog, the 24th ACM International Conference on Information and Knowledge Management (CIKM) 2015.
7. Yue Wang, **Weijing Huang**, Lang Zong, Tengjiao Wang, Dongqing Yang: Influence maximization with limit cost in social network. SCIENCE CHINA Information Sciences 56(7): 1-14 (2013)
8. Yue Wang, **Weijing Huang**, Wei Chen, Tengjiao Wang, Dongqing Yang. Informed Prediction with Incremental Core-based Friend Cycle Discovering, the 12th International Conference on Web-Age Information Management (WAIM) 2011.
9. 张腊梅, 黄威靖, 陈薇, 王腾蛟, 雷凯. EMTM: 微博中与主题相关的专家挖掘方法. 计算机研究与发展. 2016 年 (第 53 卷) (萨师煊优秀学生论文奖).
10. 吴良, 黄威靖, 陈薇, 王腾蛟, 雷凯, 刘月琴. ACT-LDA: 集成话题、社区和影响力分析的概率模型. 计算机科学与探索. 2013 年第 7 卷第 8 期 (718-728).

11. Shibo Tao, Xiaorong Wang, **Weijing Huang**, Wei Chen, Tengjiao Wang, Kai Lei, From Citation Network to Study Map: A Novel Model to Reorganize Academic Literatures, Big Scholar workshop at WWW 2017.
12. Wei Chen, Lang Zong, **Weijing Huang**, Gaoyan Ou, Yue Wang, Dongqing Yang, An Empirical Study of Massively Parallel Bayesian Networks Learning for Sentiment Extraction from Unstructured Text, the 13th Asia-Pacific Web Conference (APWeb) 2011.
13. Xilian Li, Wei Chen, Tengjiao Wang, **Weijing Huang**, Target-specific Convolutional Bi-directional LSTM Neural Network for Political Ideology Analysis, the Asia Pacific Web and Web-Age Information Management Joint Conference on Web and Big Data (APWeb-WAIM) 2017.
14. Xiao Zhang, Xiaorong Wang, Wei Chen, Jie Tao, **Weijing Huang**, Tengjiao Wang, A Taxi Gap Prediction Method via Double Ensemble Gradient Boosting Decision Tree, the 2nd IEEE International Conference on Intelligent Data and Security 2017.

## 硕博连读期间参与的项目：

1. 非结构化数据管理系统北大部分——国家“核高基”重大专项  
(2010ZX01042-002-002-02)
2. 海量 Web 数据结构化内容提取与集成及大型示范应用——国家 863 计划课题  
(2012AA011002)
3. 微博用户社区及主题时序方法研究——中国信息安全测评中心合作项目
4. 大数据驱动的航空航天装备创新研发与应用示范——十三五国家重点研发计划课题

## 硕博连读期间参与的专利：

1. 基于交互式文档聚类的信息检索方法及系统, 黄威靖, 于倩, 陈薇, 王腾蛟, 杨冬青, CN103514183A.
2. Web 社会网络核心用户信息交互演化分析方法, 王悦, 黄威靖, 陈薇, 王腾蛟, 杨冬青, CN102637182A.

感谢导师王腾蛟教授对我的悉心指导！

感谢关心、支持我的人们！

谢谢各位老师！