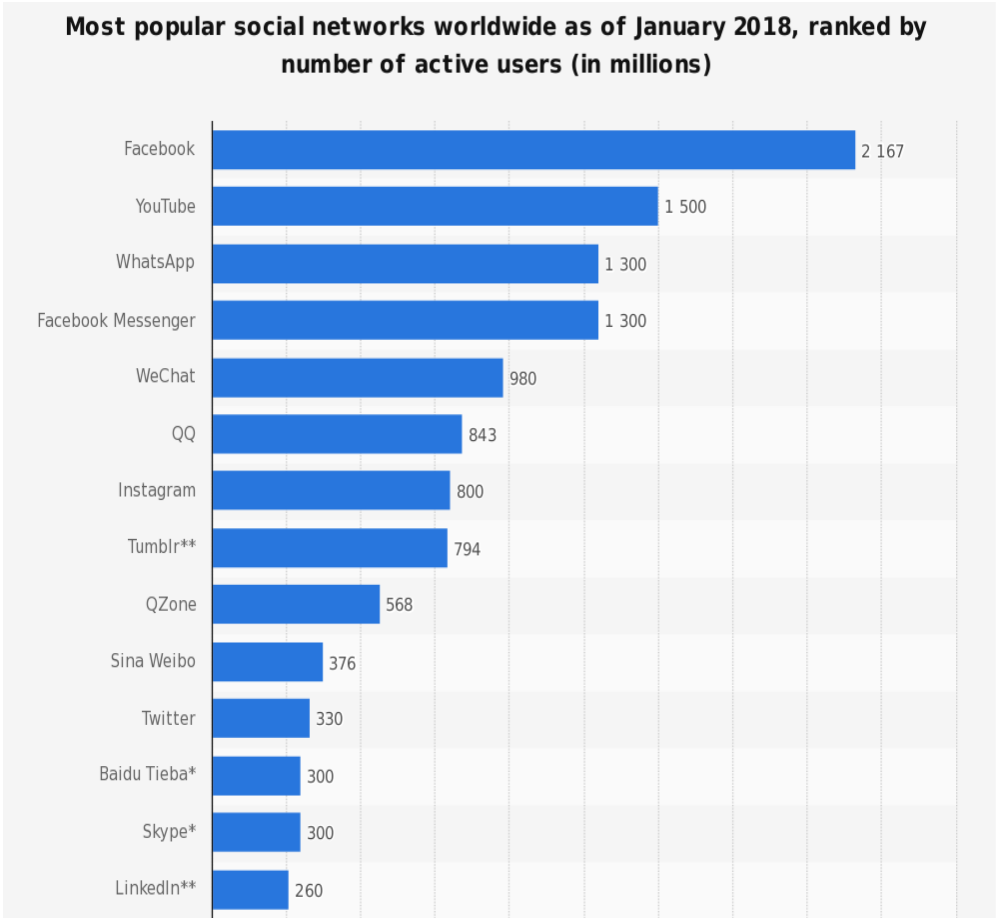
面向社会计算的网络表示学习

涂存超

导师：孙茂松 教授
刘知远 副教授

清华大学自然语言处理实验室

大规模社交网络


Most popular social networks worldwide as of January 2018, ranked by number of active users (in millions)
Facebook 2 167
YouTube 1 500
WhatsApp 1 300
Facebook Messenger 1 300
WeChat 980
QQ 843
Instagram 800
Tumblr** 794
QZone 568
Sina Weibo 376
Twitter 330
Baidu Tieba* 300
Skype* 300
LinkedIn** 260

大规模社交网络

- 社交网络 vs 传统网络
  - 规模更大、更稀疏
  - 异构信息网络
    - 文本信息、标签信息、属性信息、图像、视频等
  - 应用场景丰富
    - 用户画像
    - 个性化推荐
    - 异常检测
    - 广告推送

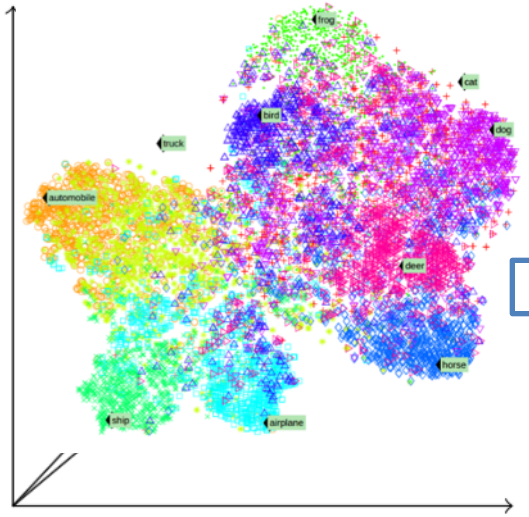
网络节点表示

网络

属性

文本

行为



统一语义空间

用户画像

个性化推荐

社交网络分析

研究框架

研究课题　　研究挑战　　解决方案　　具体工作

网络表示

可解释性

异构信息

计算效率

显式表示

隐式表示

基于词项的显式表示

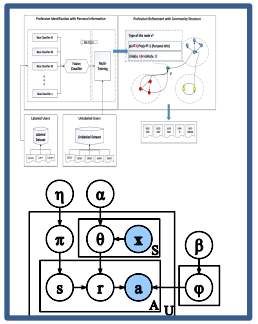
基于主题标签的显式表示

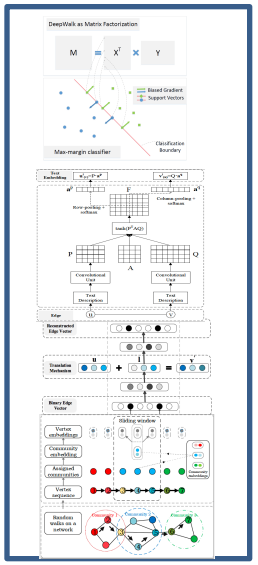
基于最大间隔的隐式表示

上下文相关的隐式表示

社会关系抽取的隐式表示

全局社区优化的隐式表示





融合多源异构信息的显式、隐式网络表示框架

具体工作

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

具体工作

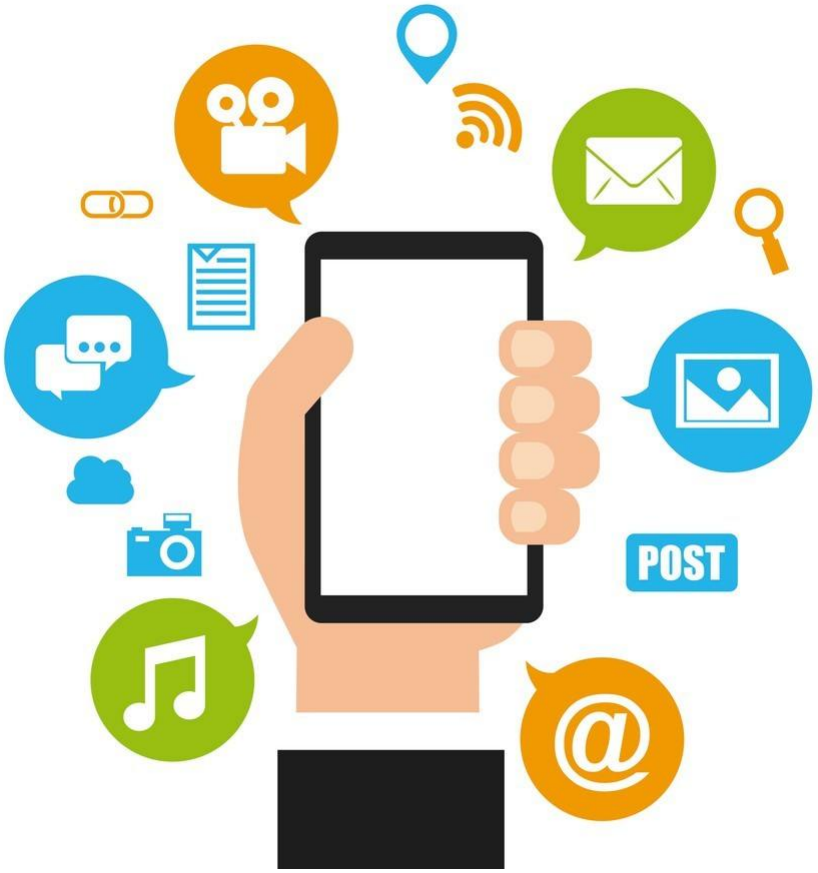
- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

基于词项表示的显式网络表示

- 社交网络用户多源异构信息

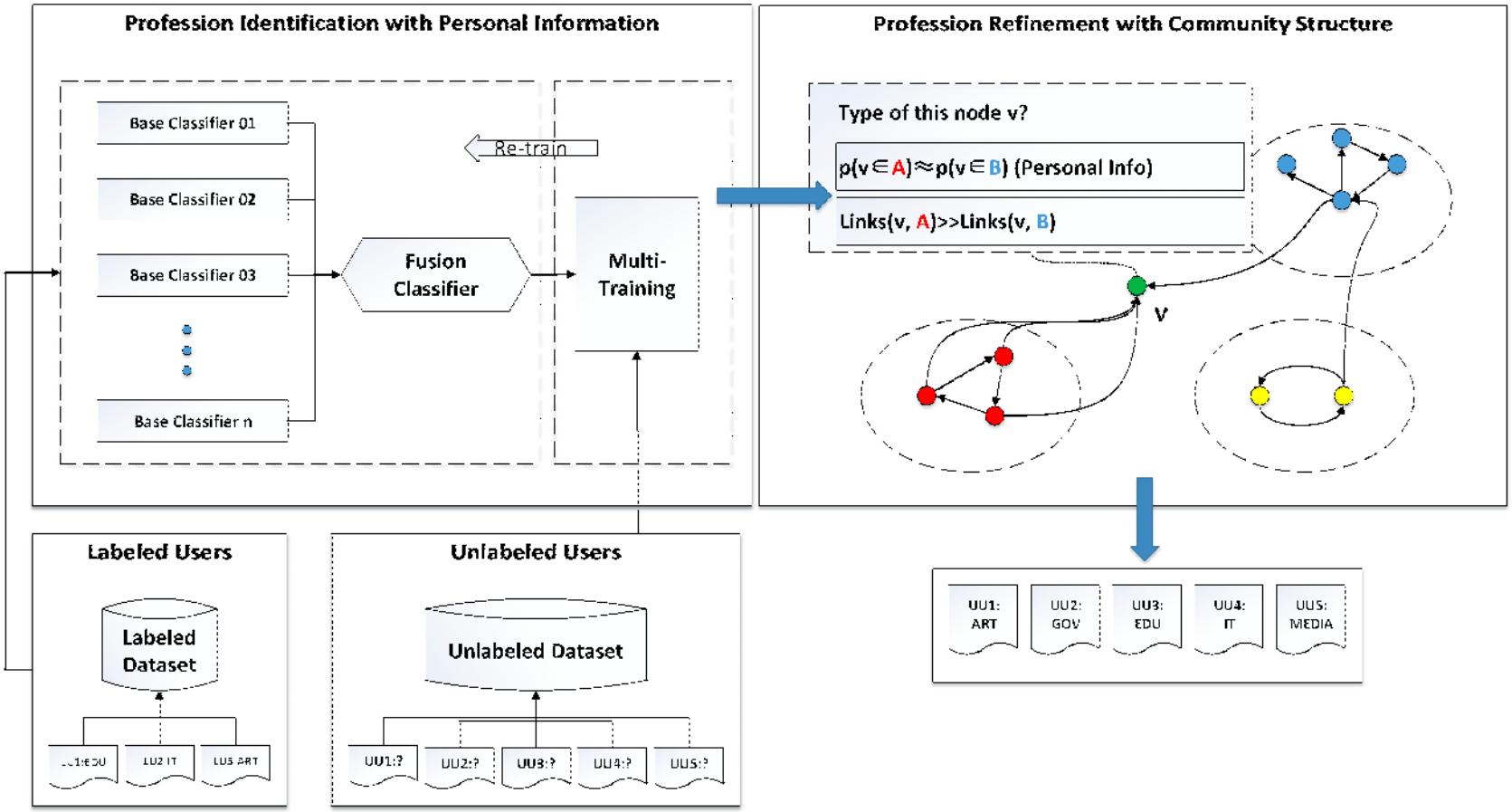
存在问题

- 多源异构信息融合
  - 个人信息
    - 如何选取特征源
    - 如何筛选有效特征
    - 如何融合多源异构特征
  - 标注数据与未标注数据
  - 网络结构信息



模型框架

- **PR**ofession **I**dentification in **S**ocial **M**edia (PRISM)

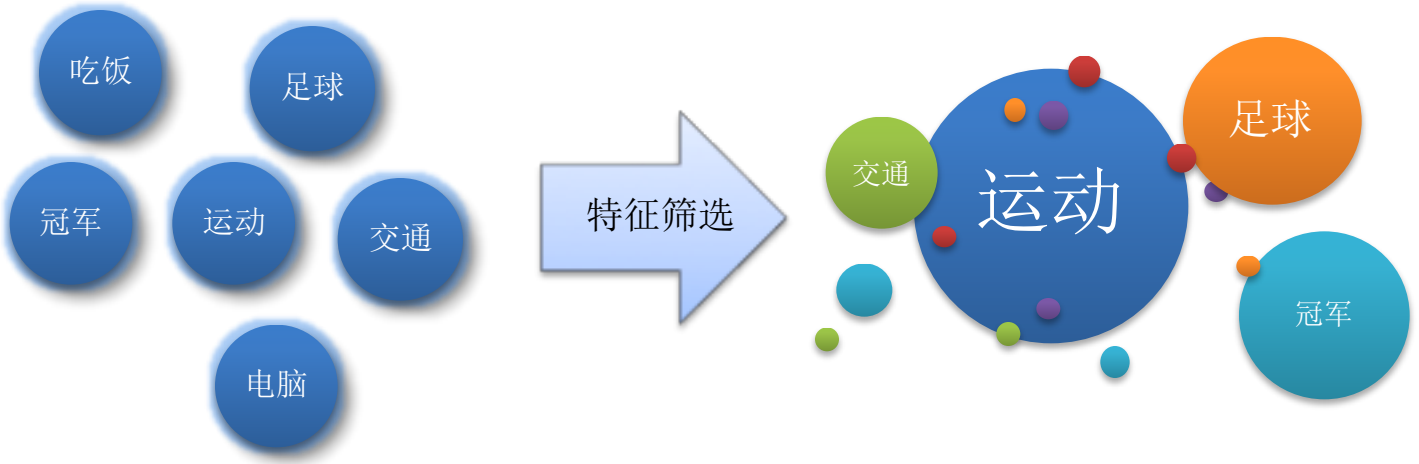


个人信息特征源选取

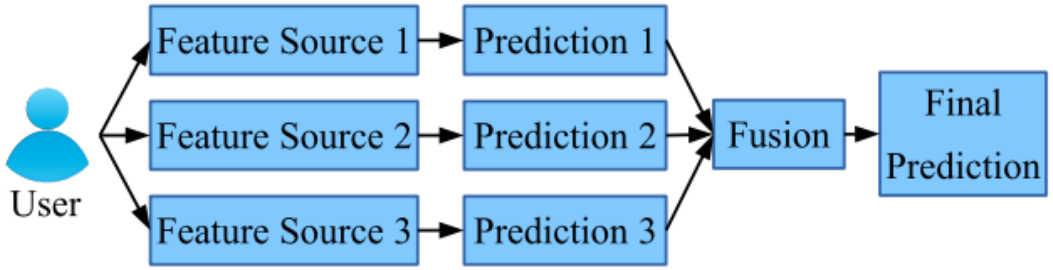
特征源	示例
个人简介	创新工场CEO，媒体联系：press@chuangxin.com
标签	风险投资、创新工场、教育、科技、创业、IT互联网
认证信息	创新工场董事长兼首席执行官
发布微博	陪同科技部部长参观创业大街，介绍我们的投资项目。
提及用户	好文分享～这是@汪华也是我们创新工场对资本寒冬的思考
提及链接	【怀念7年前的春节】7年前，谷歌中国和美国华人团队大都没有休假，趁着春节悄悄地推出了中国版Google.cn。
提及命名实体	陪同科技部部长参观创业大街，介绍我们的投资项目.....
提及哈希标签	#李开复阅享周末# 比尔盖茨推荐的9本书，大部分是有关人类社会未来的书.....

特征筛选及融合

- 特征筛选：卡方统计



- 特征融合：双层分类器



Multi-Training

- 采用协同训练（co-training）的方式，丰富训练数据
  - 利用双层分类器，对未标注数据进行属性预测
  - 投票：选取有一半以上基础分类器预测结果一致的样例
  - 将选取的样例加入训练集，重新训练双层分类器
  - 迭代上述步骤，直到分类器效果收敛

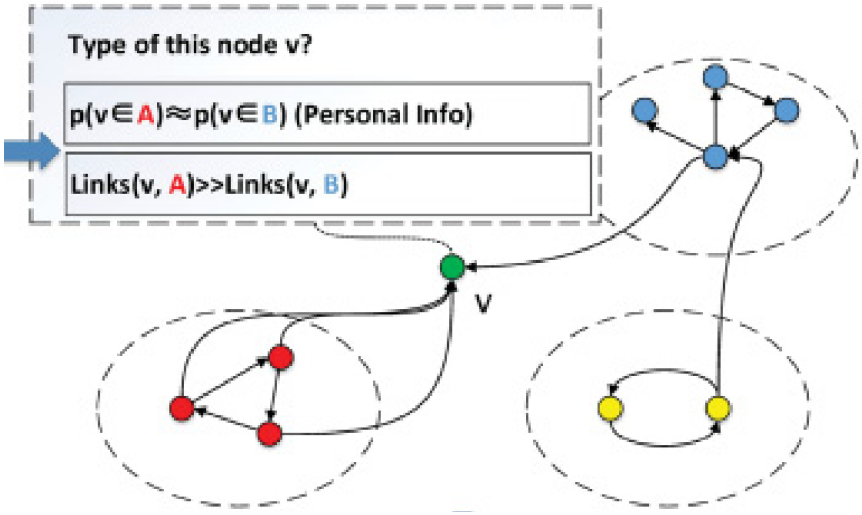
丰富训练数据，提升分类的效果以及模型的泛化能力

网络结构信息

- 拥有共同兴趣、属性的用户会行成社区

  [McPherson et al. 2001]

- 利用社区结构进行改进: normalized conductance



微博用户职业预测

- 实验结果

Method	Accuracy	Precision	Recall	F
DES	31.25	51.82	28.90	37.11
TAG	38.11	50.55	31.04	38.46
VER	78.63	75.73	74.89	75.31
MSG	47.47	49.58	42.79	45.93
MEN	38.22	42.85	30.59	35.70
URL	26.38	36.47	13.68	19.90
ENT	33.86	36.88	26.95	31.15
HAS	30.91	37.44	17.60	23.94
Single Vector	39.25	48.33	34.92	40.54
Fusion	81.25	79.60	76.27	77.90
Fusion+MT	**83.38**	**82.24**	**81.35**	**81.79**

Method	Accuracy	Precision	Recall	F
LPA	58.86	57.05	54.53	55.76
CD	64.20	65.11	60.78	62.87
PRISM				
$\lambda = 0.1$	84.17	83.15	81.62	82.37
$\lambda = 0.2$	**84.92**	**83.78**	**81.89**	**82.82**
$\lambda = 0.3$	81.12	79.10	77.42	78.25
$\lambda = 0.5$	77.56	76.53	75.08	75.79

微博用户职业预测

- 不同职业用户语言使用习惯

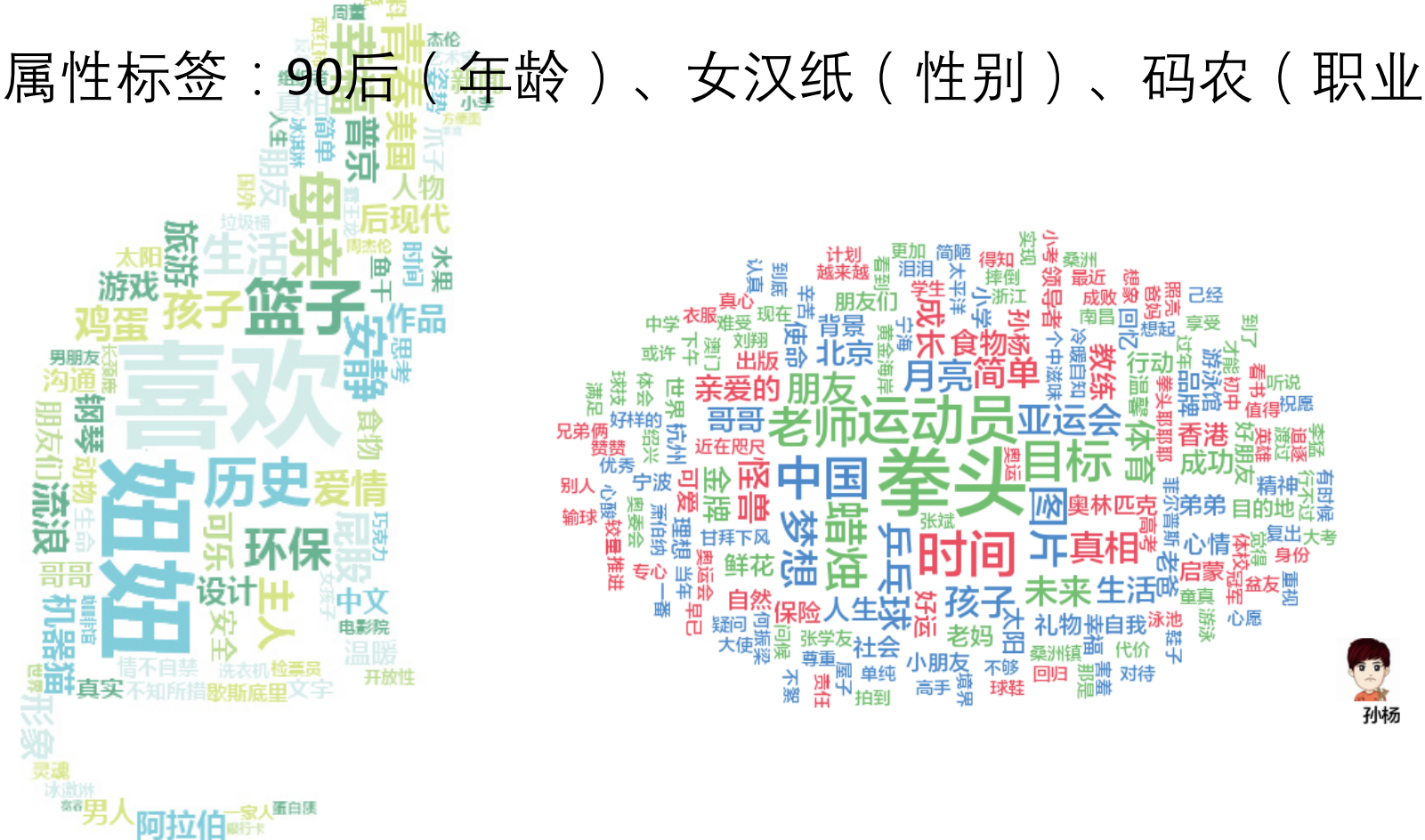
No.	Profession	Conj.	Interj.	M.P.
1	media	1.19▽	0.22△	2.16△
2	government	1.29	0.17	1.70
3	entertainment	1.08▽	0.26△	2.38△
4	estate	1.26	0.15	1.72
5	finance	1.39△	0.15▽	1.65▽
6	IT	1.35△	0.15▽	1.66
7	sports	1.04▽	0.25△	2.60△
8	education	1.42△	0.16▽	1.55▽
9	fashion	1.25	0.22	1.95
10	games	1.34	0.16	1.26▽
11	literature	1.31	0.27△	2.25
12	services	1.29	0.18	1.94
13	art	1.11▽	0.22△	2.06△
14	healthcare	1.76△	0.11▽	1.15▽

研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

基于主题标签的显式网络表示

- 社交网络用户标签
  - 兴趣标签：旅行、音乐、运动、篮球
  - 属性标签：90后（年龄）、女汉纸（性别）、码农（职业）



利用标签进行用户表示

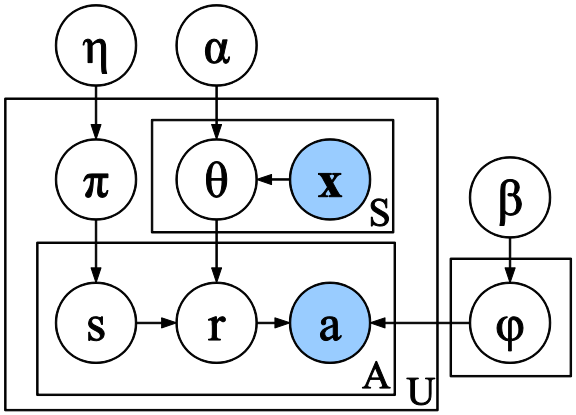
- 针对200万用户进行标签统计
  - 共计3,013,816标签
  - >=500次标签：4,127

如何建立多源异构特征与标签之间的语义关系？

如何对未标注用户构建标签表示，进行标签推荐？

Tag Correspondence Model (TCM)

- 标签关联模型



- 标签生成过程
  - 根据狄利克雷先验η选取来源分布π_u
  - 根据来源分布π_u选取一个来源s
  - 对于每个来源s，根据分布$\theta_{u,s}$选取对应r
  - 根据标签分布$\varphi_{s,r}$选取标签t

标签关联模型

- 联合概率

$$\Pr(\boldsymbol{a}, \boldsymbol{s}, \boldsymbol{r}|\boldsymbol{x}, \alpha, \eta, \beta) = \Pr(\boldsymbol{a}|\boldsymbol{r}, \beta) \Pr(\boldsymbol{r}, \boldsymbol{s}|\boldsymbol{x}, \alpha, \eta).$$

- 给定用户**u**，来源**s**，以及对应元素**r**：

$$\Pr(t|u, \phi) = \sum_{s \in S} \sum_{r \in V_s} \Pr(t|r, \phi) \Pr(r|u, s) \Pr(s|u),$$

标签对应关系

- 与不同来源最相关的标签

来源	Pr(s)	最相关的5个标签
UM	0.19	移动互联网, 方大同, 重庆, 深圳, 广州
UD	0.19	模特, 淘宝店主, 摄影师, cosplay, 电子商务
NT	0.42	网络购物, 小说, 媒体, 阅读, 广告
ND	0.20	豆瓣, 懒惰, 小说, 食物, 音乐

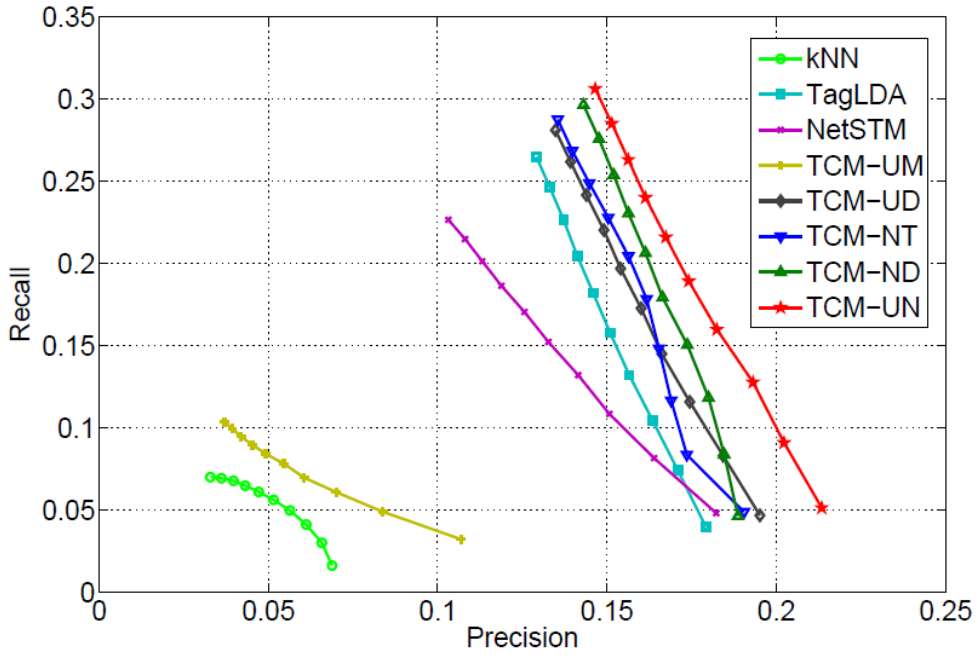
标签对应关系

- 与标签最相关的元素

标签	对应的元素
教育	互联网（NT）、教育（D）、教育（M）、政治（NT）、学习（NT）
电子商务	B2C（NT）、IT（NT）、电子商务（M）、电子商务（NT）、市场（NT）
移动互联网	SNS（NT）、移动（D）、互联网（M）、移动（D）、IT（NT）
创业	创业（NT）、风险投资（NT）、电子商务（NT）、创业者（NT）、互联网（D）

标签推荐结果

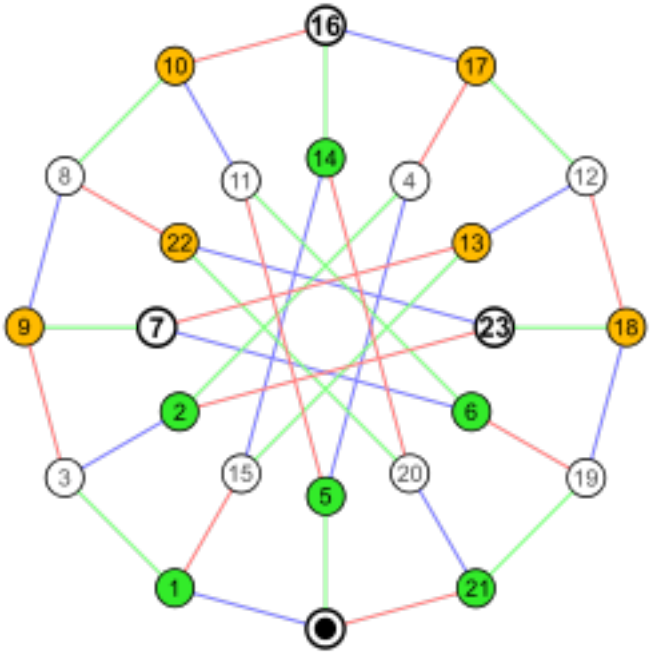
- 与不同方法对比
  - kNN, TagLDA, NetSTM



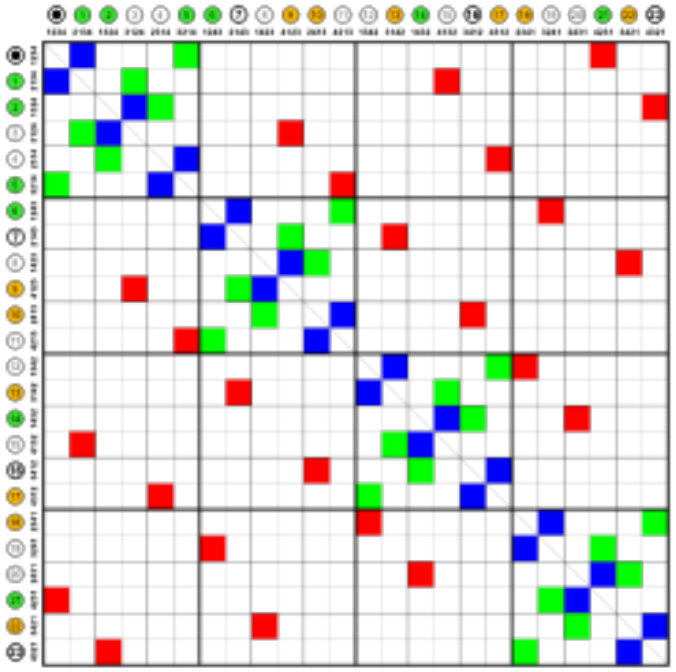
研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

基于符号的表示方案



N个节点的网络



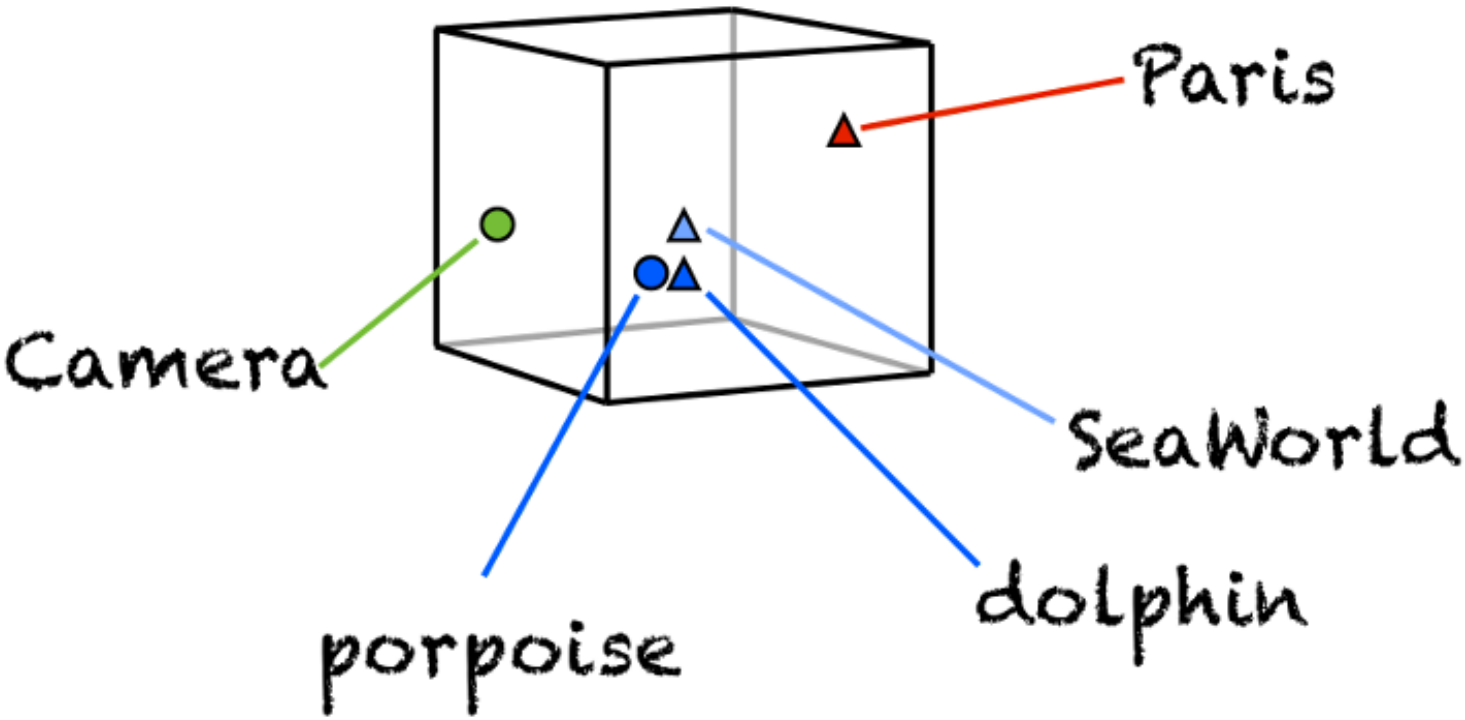
邻接矩阵

需要N×N个元素来表示

稀疏！不利于存储计算

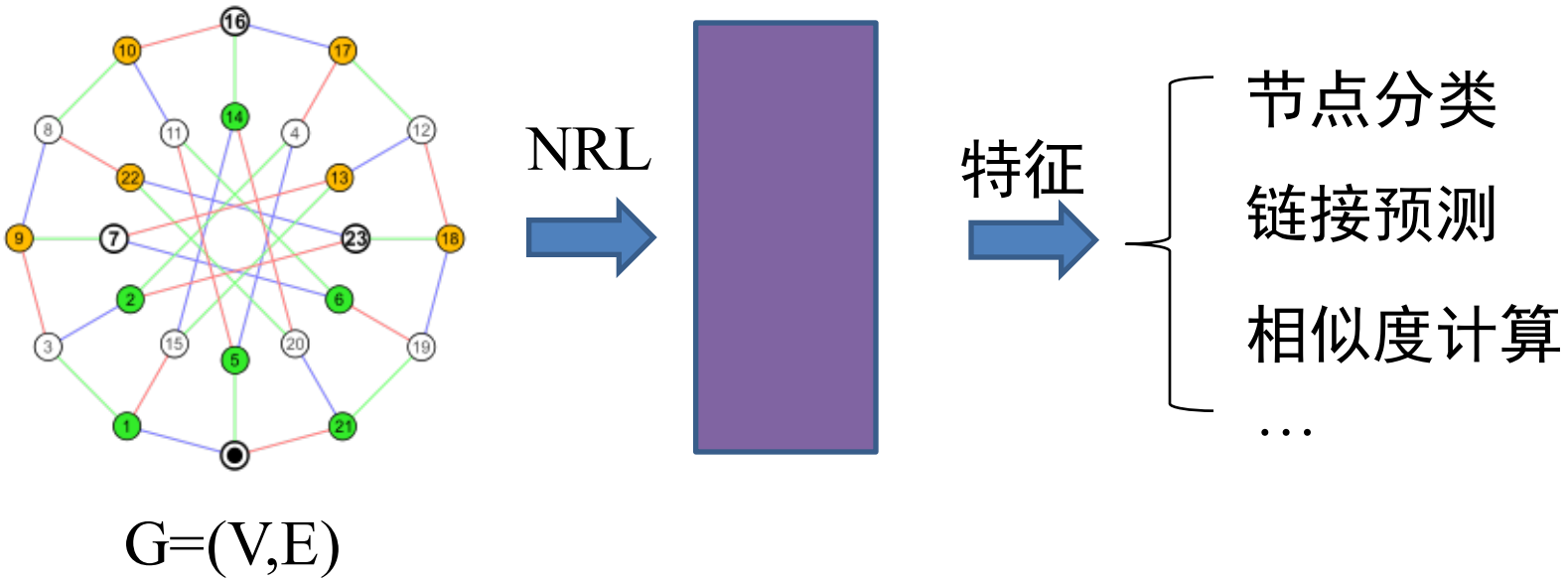
分布式表示方案

- Distributed Representation
- 对象均被表示成稠密、实值、低维向量



网络表示学习

- 将网络中节点的语义信息表示为低维向量



已有网络表示学习工作

- 基于矩阵特征向量的方法
  - 依赖于关系矩阵构建
  - 时间复杂度高
- 基于矩阵分解的方法
  - 已有模型可以近似成对关系矩阵的矩阵分解
- 基于简单神经网络的方法
  - 针对相关节点，设计损失函数
  - 采用随机梯度下降算法进行优化

存在问题

- 没有考虑真实网络中的异构信息
  - 节点的文本、标签类别等信息
  - 边上的文本、标签信息
  - 全局的社区信息

用户及其关系和行为

文本、视频、语音等信息

信息多源异构，难以建立语义关联

存在问题

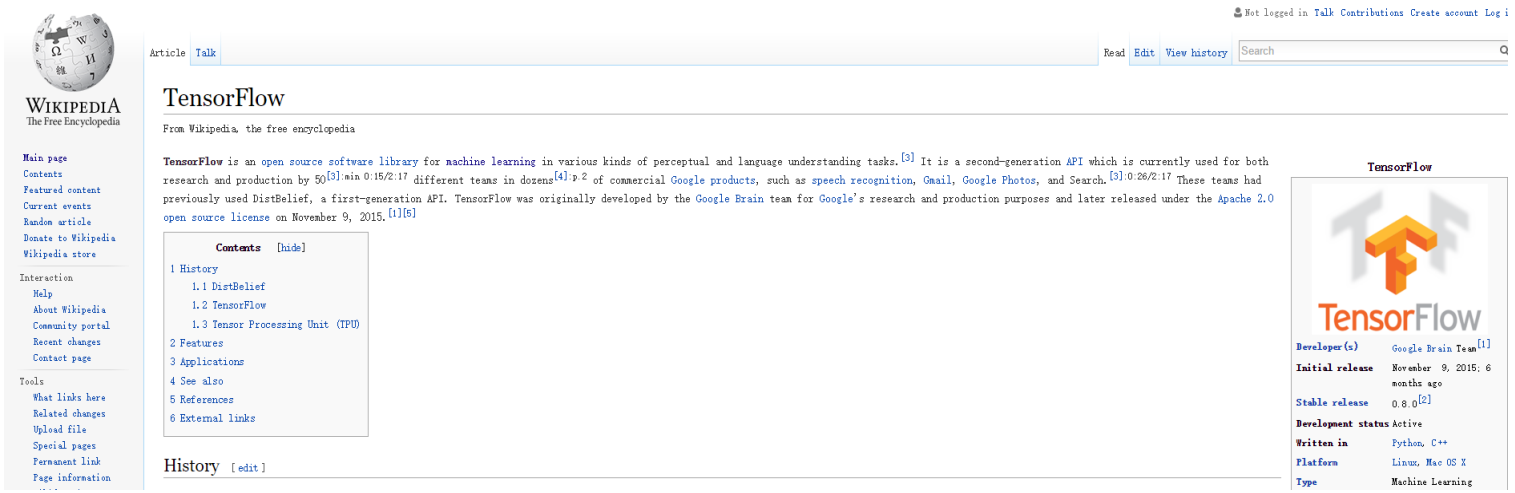
- 对网络结构进行建模和重构
- 忽略在网络分析任务中的效果
  - 节点分类
  - 社区发现
  - 链接预测
  - 社会关系抽取

研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

最大间隔网络表示

- 真实世界网络节点往往被标注类别标签



TensorFlow

From Wikipedia, the free encyclopedia

TensorFlow is an open source software library for machine learning in various kinds of perceptual and language understanding tasks.[3] It is a second-generation API which is currently used for both research and production by 50[3]:min 0:15/2:17 different teams in dozens[4]:p.2 of commercial Google products, such as speech recognition, Gmail, Google Photos, and Search.[3]:0:26/2:17 These teams had previously used DistBelief, a first-generation API. TensorFlow was originally developed by the Google Brain team for Google's research and production purposes and later released under the Apache 2.0 open source license on November 9, 2015.[1][5]

Contents [hide]

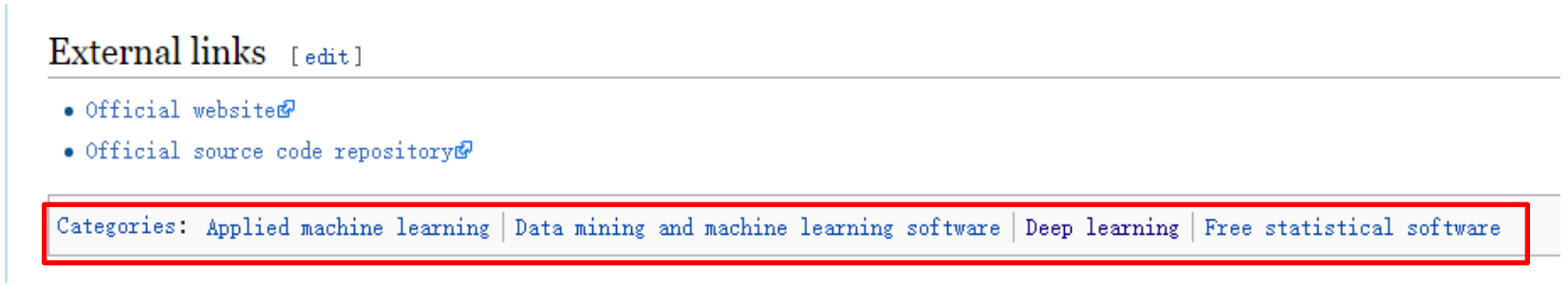
1 History
- 1.1 DistBelief
- 1.2 TensorFlow
- 1.3 Tensor Processing Unit (TPU)

2 Features
3 Applications
4 See also
5 References
6 External links

History [edit]

TensorFlow

Developer(s)	Google Brain Team[1]
Initial release	November 9, 2015; 6 months ago
Stable release	0.8.0[2]
Development status	Active
Written in	Python, C++
Platform	Linux, Mac OS X
Type	Machine Learning



External links [edit]

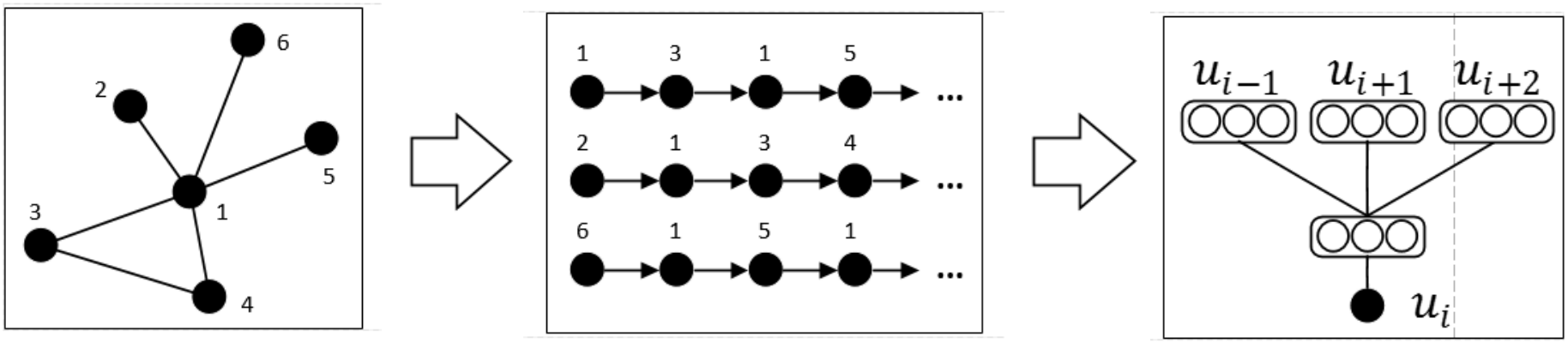
- Official website
- Official source code repository

Categories: Applied machine learning | Data mining and machine learning software | Deep learning | Free statistical software

传统NRL是无监督方法，无法考虑标签信息
在预测任务上效果差

DeepWalk

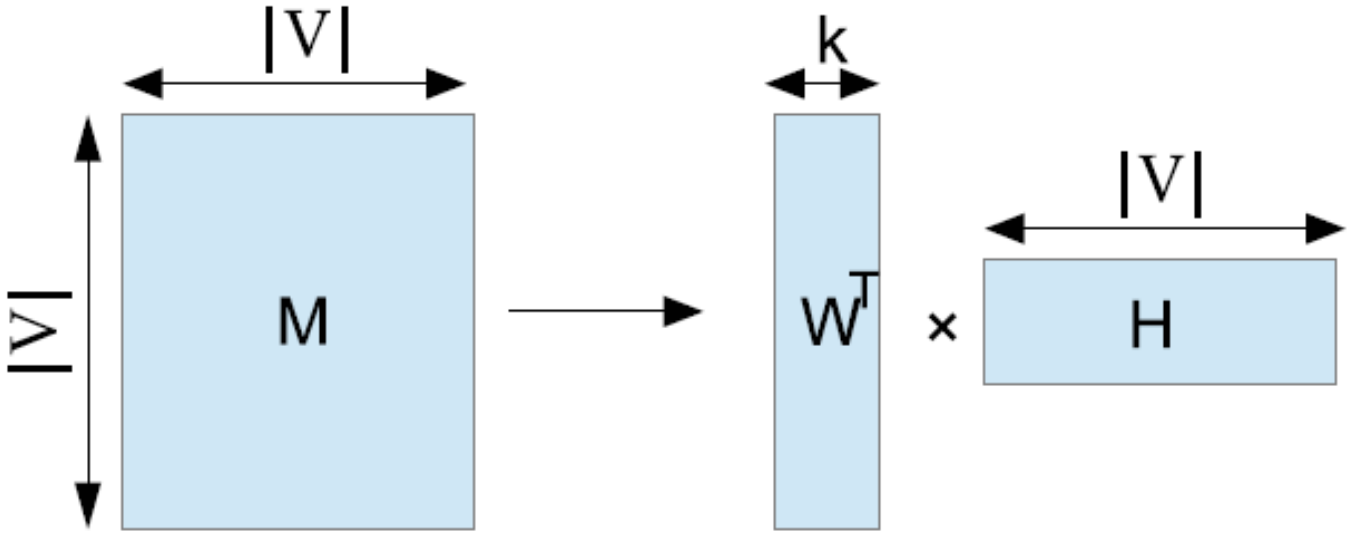
- 基于随机游走的网络表示学习模型



网络　　随机游走序列　　Skip-Gram

网络表示学习与矩阵分解的关系

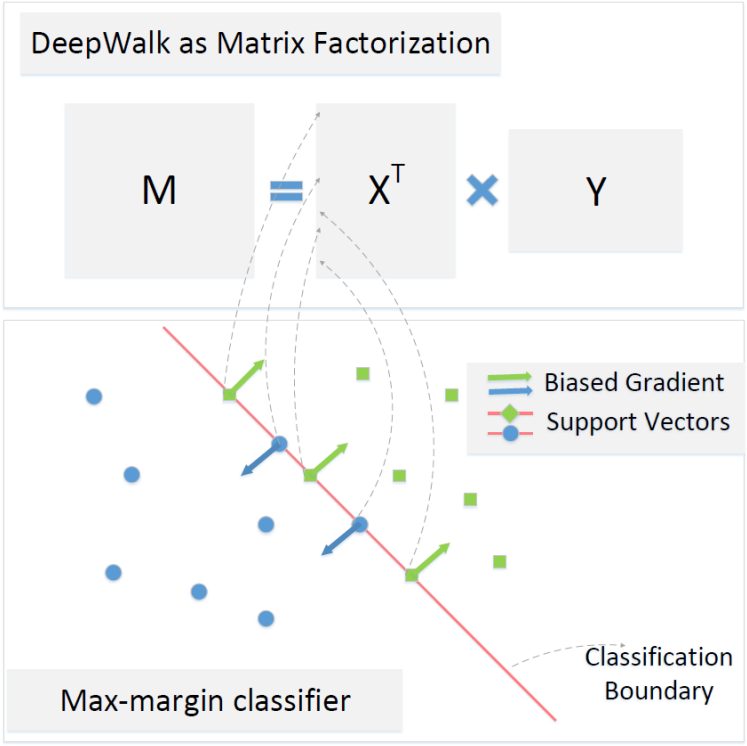
- DeepWalk等网络表示学习算法等价于矩阵分解



$$M_{ij} = \log \frac{[e_i(A + A^2 + \cdots + A^t)]_j}{t}$$

最大间隔网络表示

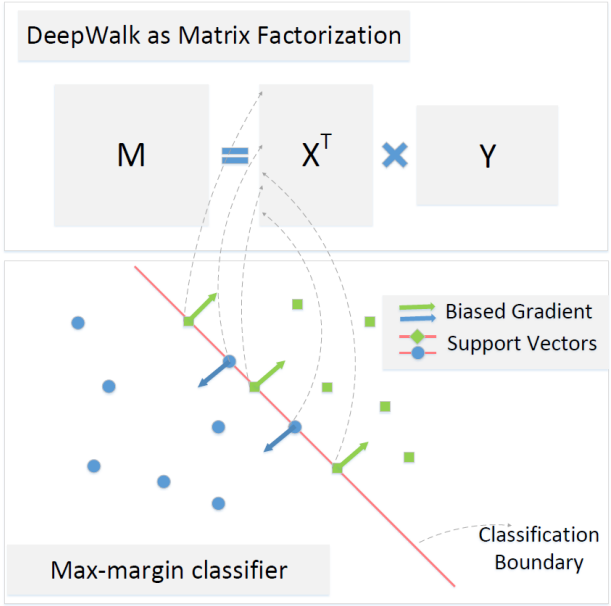
- 共同训练DeepWalk + 最大间隔分类器



$$\min_{X,Y,W,\xi} l_{MMDW} = \min_{X,Y,W,\xi} l_{DW} + \frac{1}{2}\|W\|_2^2 + C\sum_{i=1}^{T}\xi_i \quad \text{s.t. } \mathrm{w}_{l_i}^T \mathrm{x}_i - \mathrm{w}_j^T \mathrm{x}_i \geq e_i^j - \xi_i,\ \forall i, j$$

训练过程

- Max-Margin DeepWalk (MMDW)
  - 利用MFDW初始化节点表示
  - 利用标注节点训练SVM
  - 对于标注节点计算其偏置向量
  - 重新训练MFDW



使边界支持向量向各自类别移动，
让类别之间分类界限更加明显

实验结果

- 节点分类结果
  - >5%的提升

Table 2: Accuracy (%) of vertex classification on Citeseer.

%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
DW	49.09	55.96	60.65	63.97	65.42	67.29	66.80	66.82	63.91
MFDW	50.54	54.47	57.02	57.19	58.60	59.18	59.17	59.03	55.35
LINE	39.82	46.83	49.02	50.65	53.77	54.2	53.87	54.67	53.82
MMDW($\eta = 10^{-2}$)	**55.60**	60.97	63.18	65.08	**66.93**	**69.52**	**70.47**	**70.87**	**70.95**
MMDW($\eta = 10^{-3}$)	55.56	**61.54**	**63.36**	**65.18**	66.45	69.37	68.84	70.25	69.73
MMDW($\eta = 10^{-4}$)	54.52	58.49	59.25	60.70	61.62	61.78	63.24	61.84	60.25

Table 3: Accuracy (%) of vertex classification on Wiki.

%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
DW	52.03	54.62	59.80	60.29	61.26	65.41	65.84	66.53	68.16
MFDW	56.40	60.28	61.90	63.39	62.59	62.87	64.45	62.71	61.63
LINE	52.17	53.62	57.81	57.26	58.94	62.46	62.24	66.74	67.35
MMDW($\eta = 10^{-2}$)	**57.76**	**62.34**	**65.76**	**67.31**	**67.33**	**68.97**	**70.12**	**72.82**	**74.29**
MMDW($\eta = 10^{-3}$)	54.31	58.69	61.24	62.63	63.18	63.58	65.28	64.83	64.08
MMDW($\eta = 10^{-4}$)	53.98	57.48	60.10	61.94	62.18	62.36	63.21	62.29	63.67

实验结果

- 节点分类结果
  - >5%的提升
  - 仅用一半的训练数据即可达到baseline的效果

Table 2: Accuracy (%) of vertex classification on Citeseer.

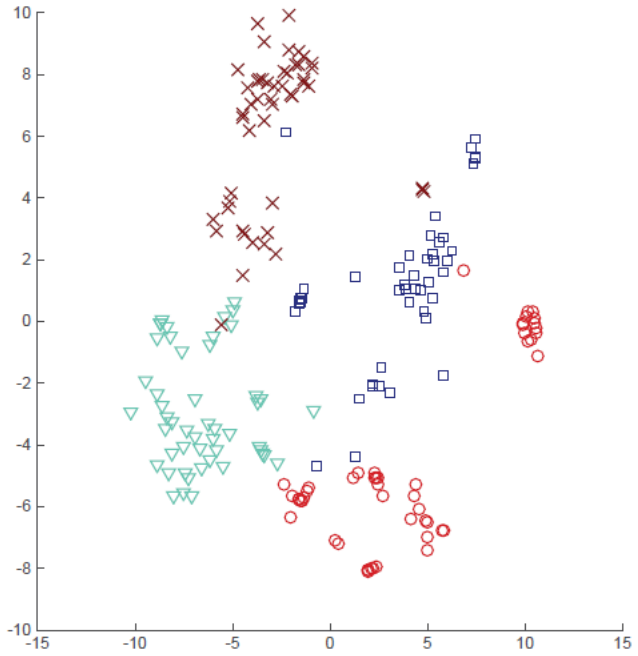
%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
DW	49.09	55.96	60.65	63.97	65.42	67.29	66.80	66.82	63.91
MFDW	50.54	54.47	57.02	57.19	58.60	59.18	59.17	59.03	55.35
LINE	39.82	46.83	49.02	50.65	53.77	54.2	53.87	54.67	53.82
MMDW($\eta = 10^{-2}$)	**55.60**	60.97	63.18	65.08	**66.93**	**69.52**	**70.47**	**70.87**	**70.95**
MMDW($\eta = 10^{-3}$)	55.56	**61.54**	**63.36**	**65.18**	66.45	69.37	68.84	70.25	69.73
MMDW($\eta = 10^{-4}$)	54.52	58.49	59.25	60.70	61.62	61.78	63.24	61.84	60.25

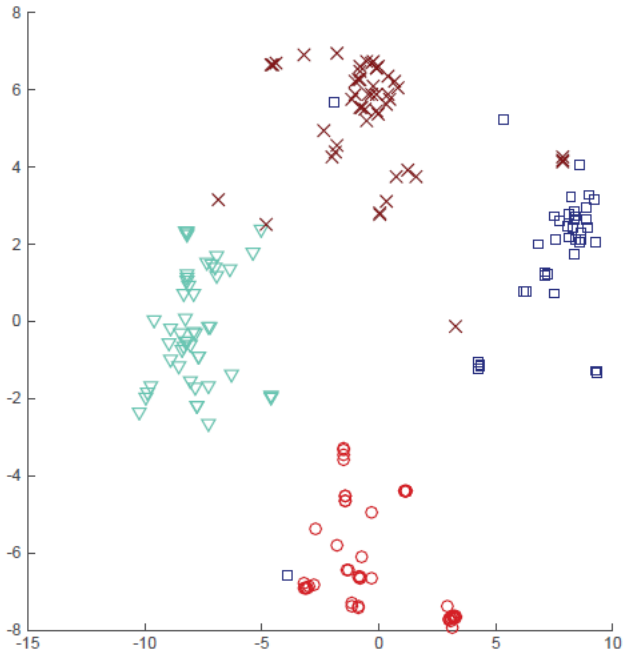
Table 3: Accuracy (%) of vertex classification on Wiki.

%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
DW	52.03	54.62	59.80	60.29	61.26	65.41	65.84	66.53	68.16
MFDW	56.40	60.28	61.90	63.39	62.59	62.87	64.45	62.71	61.63
LINE	52.17	53.62	57.81	57.26	58.94	62.46	62.24	66.74	67.35
MMDW($\eta = 10^{-2}$)	**57.76**	**62.34**	**65.76**	**67.31**	**67.33**	**68.97**	**70.12**	**72.82**	**74.29**
MMDW($\eta = 10^{-3}$)	54.31	58.69	61.24	62.63	63.18	63.58	65.28	64.83	64.08
MMDW($\eta = 10^{-4}$)	53.98	57.48	60.10	61.94	62.18	62.36	63.21	62.29	63.67

可视化结果

- 节点表示t-SNE可视化
  - DeepWalk与MMDW





研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

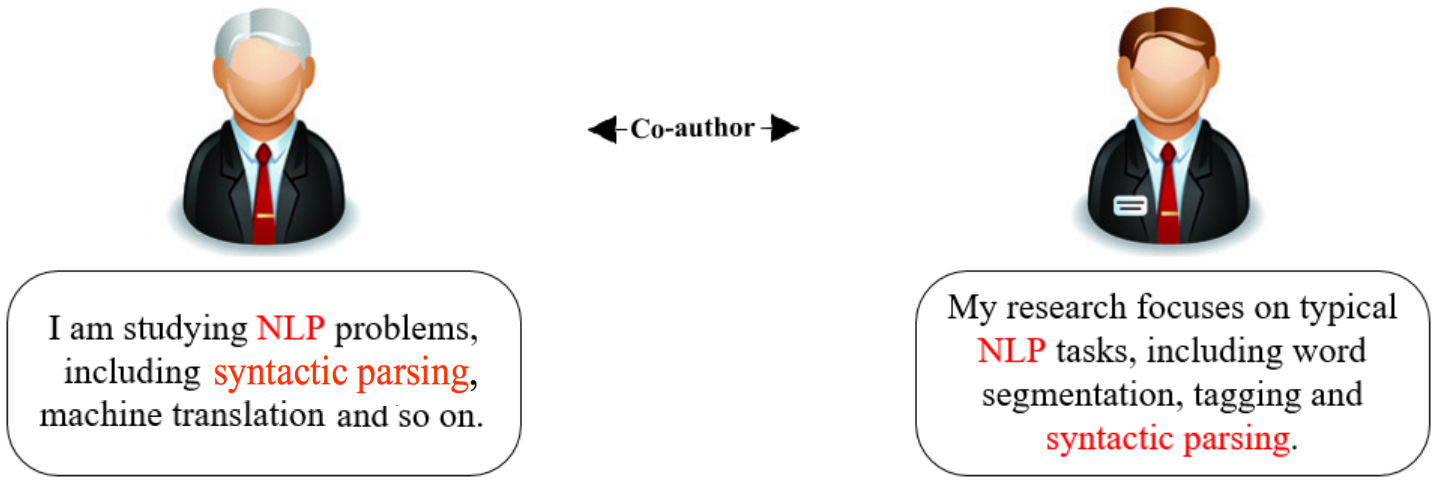
示例

- 一个NLP专家

I am studying NLP problems, including syntactic parsing, machine translation and so on.

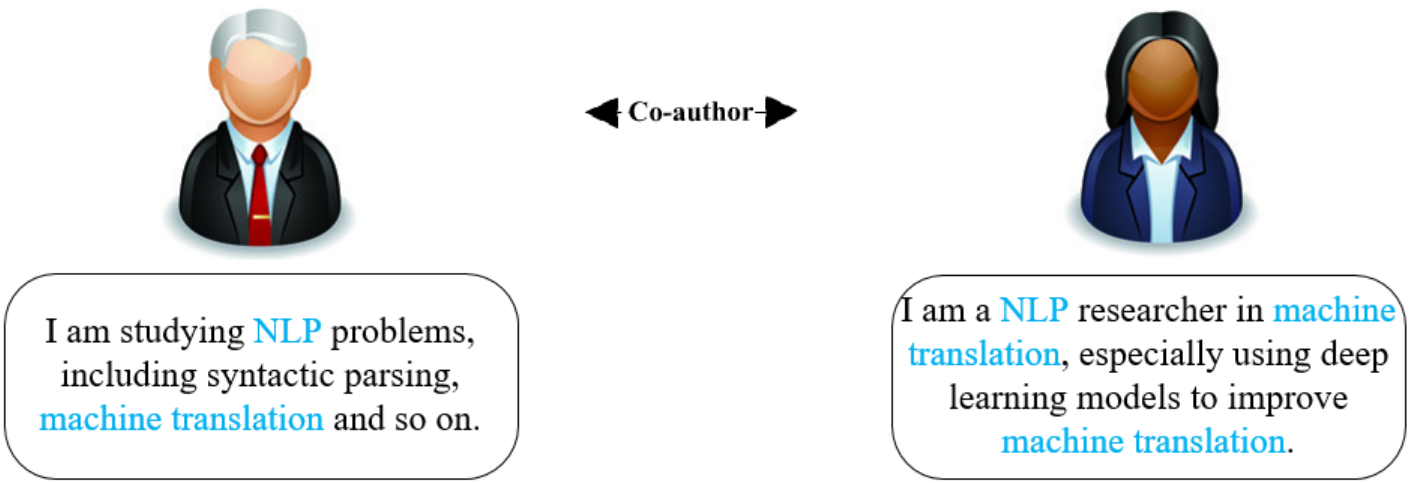
示例

- 与另外的研究者合作



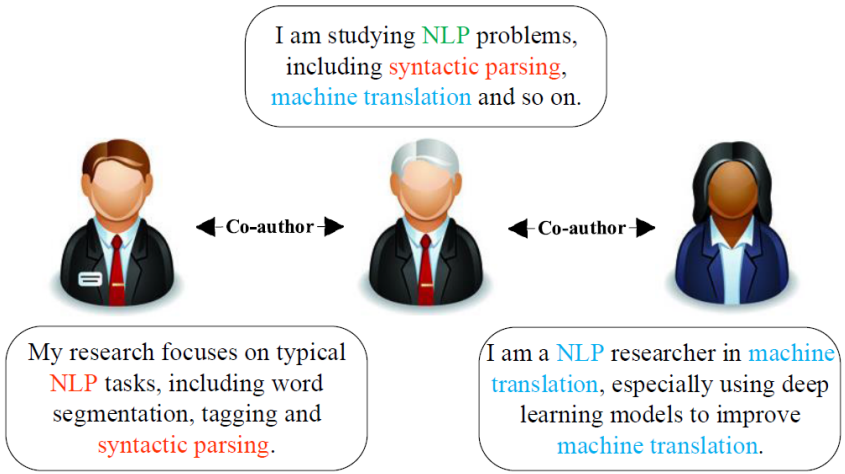
示例

- 与另外的研究者合作



上下文相关网络表示

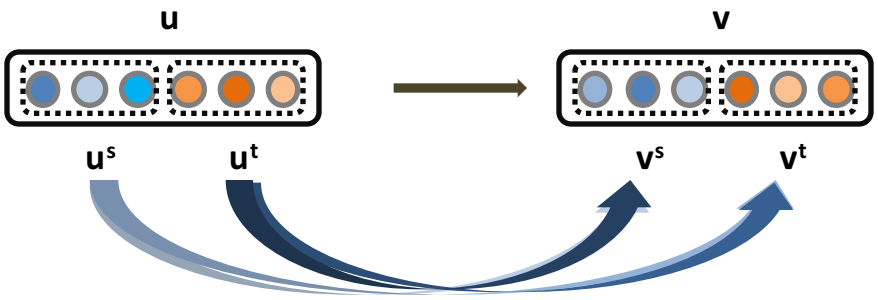
- 基于文本的信息网络
- 根据不同邻居为节点学习动态表示
- 对节点之间的关系进行更准确的建模



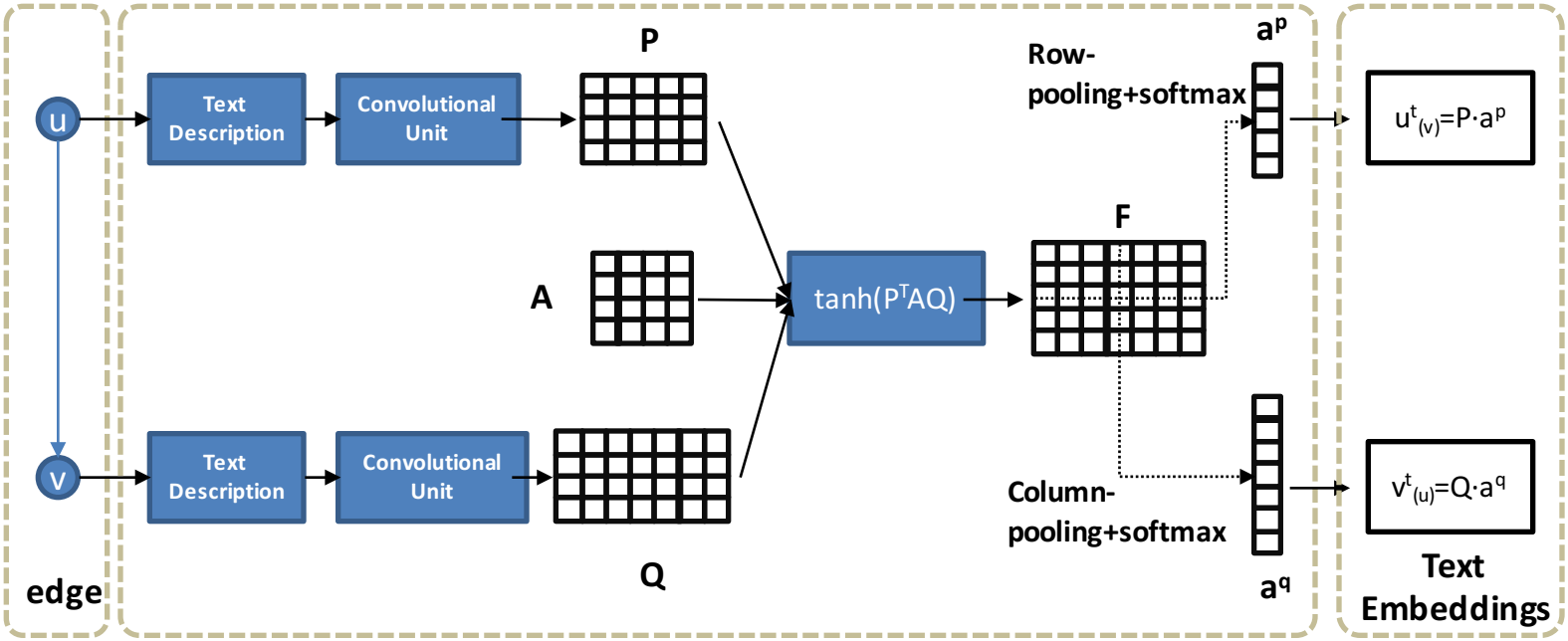
上下文相关网络表示

- 最大化每条边的对数似然

$$L_{ss}(e) + \alpha \cdot L_{st}(e) + \beta \cdot L_{ts}(e) + \gamma \cdot L_{tt}(e)$$



上下文相关文本表示


u
v
edge
Text Description
Convolutional Unit
Text Description
Convolutional Unit
P
A
Q
tanh(P^TAQ)
F
Row-pooling+softmax
Column-pooling+softmax
a^p
a^q
u^t(v)=P·a^p
v^t(u)=Q·a^q
Text Embeddings

链接预测

#Training edges	15%	25%	35%	45%	55%	65%	75%	85%	95%
MMB	54.6	57.9	57.3	61.6	66.2	68.4	73.6	76.0	80.3
DeepWalk	55.2	66.0	70.0	75.7	81.3	83.3	87.6	88.9	88.0
LINE	53.7	60.4	66.5	73.9	78.5	83.8	87.5	87.7	87.6
Node2vec	57.1	63.6	69.9	76.2	84.3	87.3	88.4	89.2	89.2
NC	78.7	82.1	84.7	88.7	88.7	91.8	92.1	92.0	92.7
TADW	**87.0**	89.5	91.8	90.8	91.1	92.6	93.5	91.9	91.7
CENE	86.2	84.6	89.9	91.2	92.3	91.8	93.2	92.9	**93.2**
CANE(Text)	83.8	85.2	87.3	88.9	91.1	91.2	91.8	93.1	93.5
CANE(w/o attention)	84.5	89.3	89.2	91.6	91.1	91.8	92.3	92.5	93.6
CANE	**90.0**	**91.2**	**92.0**	**93.0**	**94.2**	**94.6**	**95.4**	**95.7**	**96.3**

链接预测

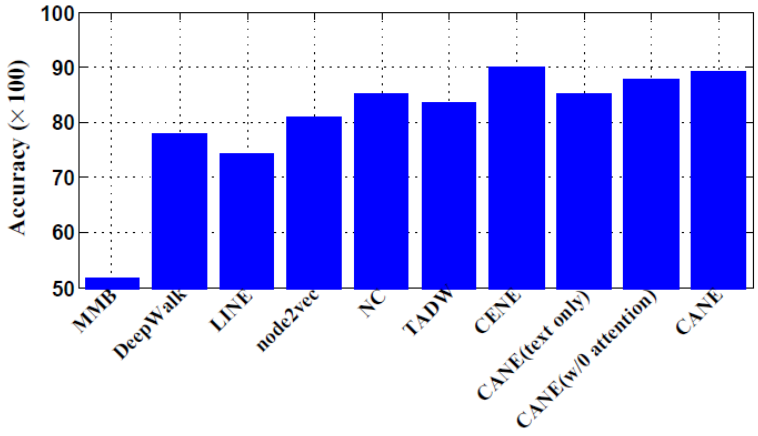
#Training edges	15%	25%	35%	45%	55%	65%	75%	85%	95%
MMB	54.6	57.9	57.3	61.6	66.2	68.4	73.6	76.0	80.3
DeepWalk	55.2	66.0	70.0	75.7	81.3	83.3	87.6	88.9	88.0
LINE	53.7	60.4	66.5	73.9	78.5	83.8	87.5	87.7	87.6
Node2vec	57.1	63.6	69.9	76.2	84.3	87.3	88.4	89.2	89.2
NC	78.7	82.1	84.7	88.7	88.7	91.8	92.1	92.0	92.7
TADW	87.0	89.5	91.8	90.8	91.1	92.6	93.5	91.9	91.7
CENE	86.2	84.6	89.9	91.2	92.3	91.8	93.2	92.9	93.2
CANE(Text)	83.8	85.2	87.3	88.9	91.1	91.2	91.8	93.1	93.5
CANE(w/o attention)	**84.5**	89.3	89.2	91.6	91.1	91.8	92.3	92.5	**93.6**
CANE	**90.0**	**91.2**	**92.0**	**93.0**	**94.2**	**94.6**	**95.4**	**95.7**	**96.3**

节点分类

- 将上下文相关表示转化为上下文无关表示

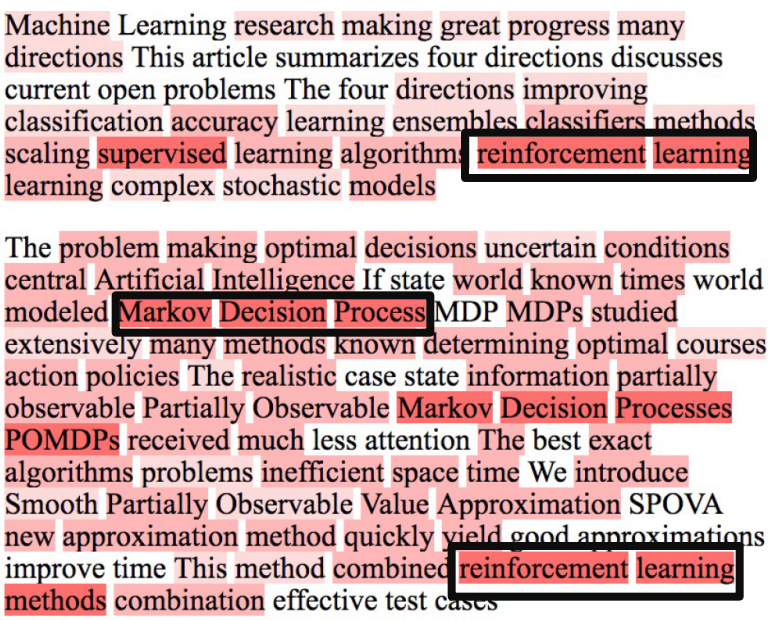
$$u = \frac{1}{N} \sum_{(u,v)|(v,u)\in E} u_{(v)}$$

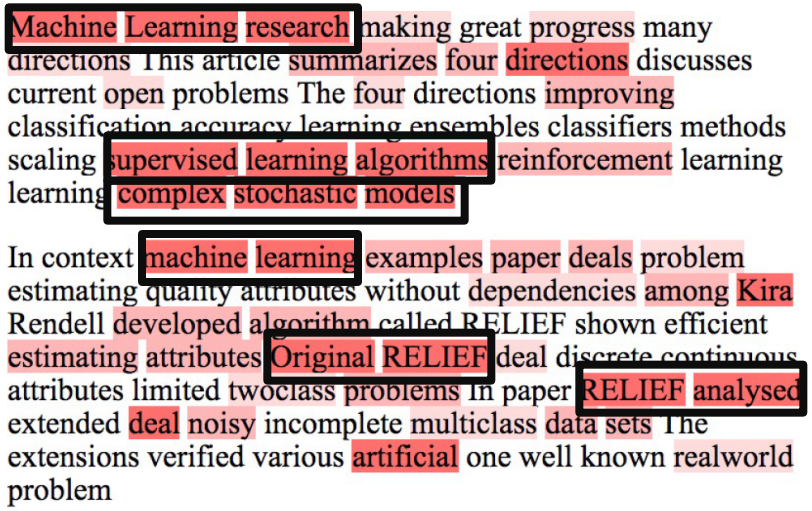
- 可比的结果



互相注意力机制

- Edge (A, B) and (A, C)





研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

面向社会关系抽取的网络表示

- 传统网络表示学习模型
  - 将边简化成二元或实数值
  - 忽略边上丰富的语义信息
- 传统网络分析任务
  - 不能衡量模型对于具体关系建模和预测的能力

社会关系抽取

- 关系的定义

用户

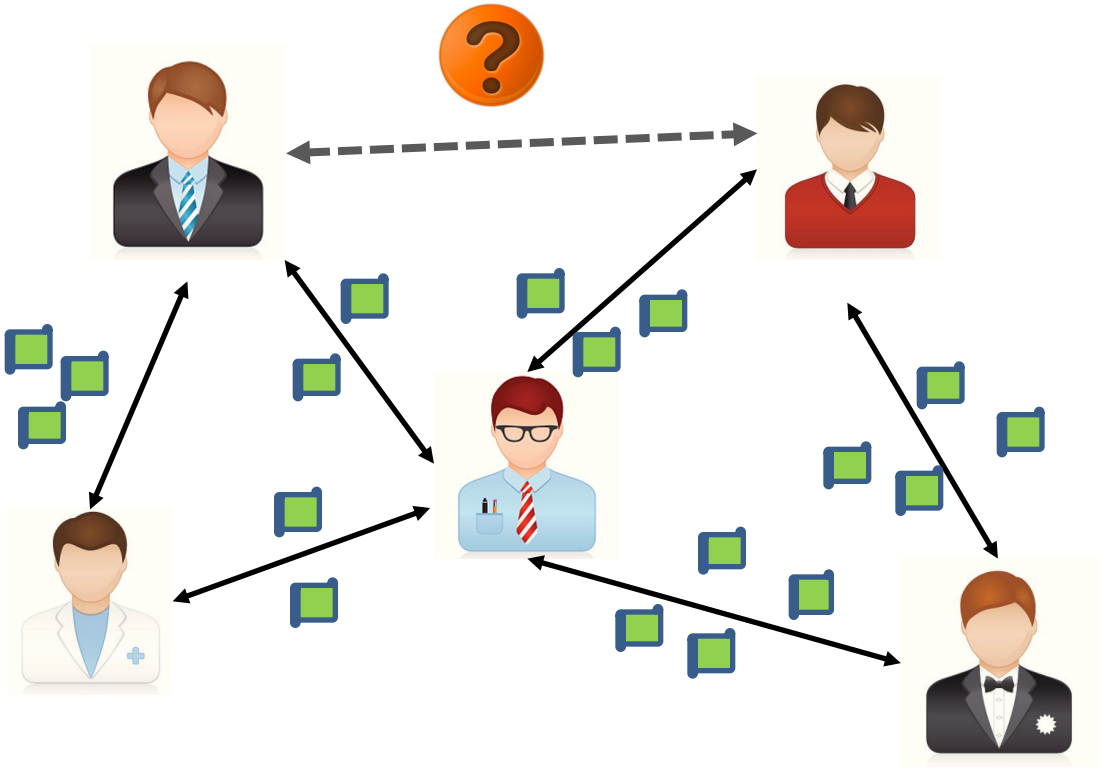
交互文本

命名实体识别/关键词抽取

标签集合

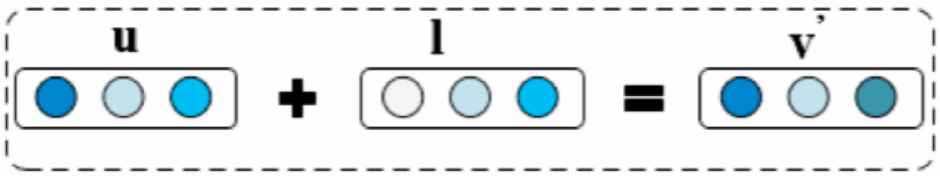
社会关系抽取

- 抽取的定义



面向社会关系抽取的网络表示

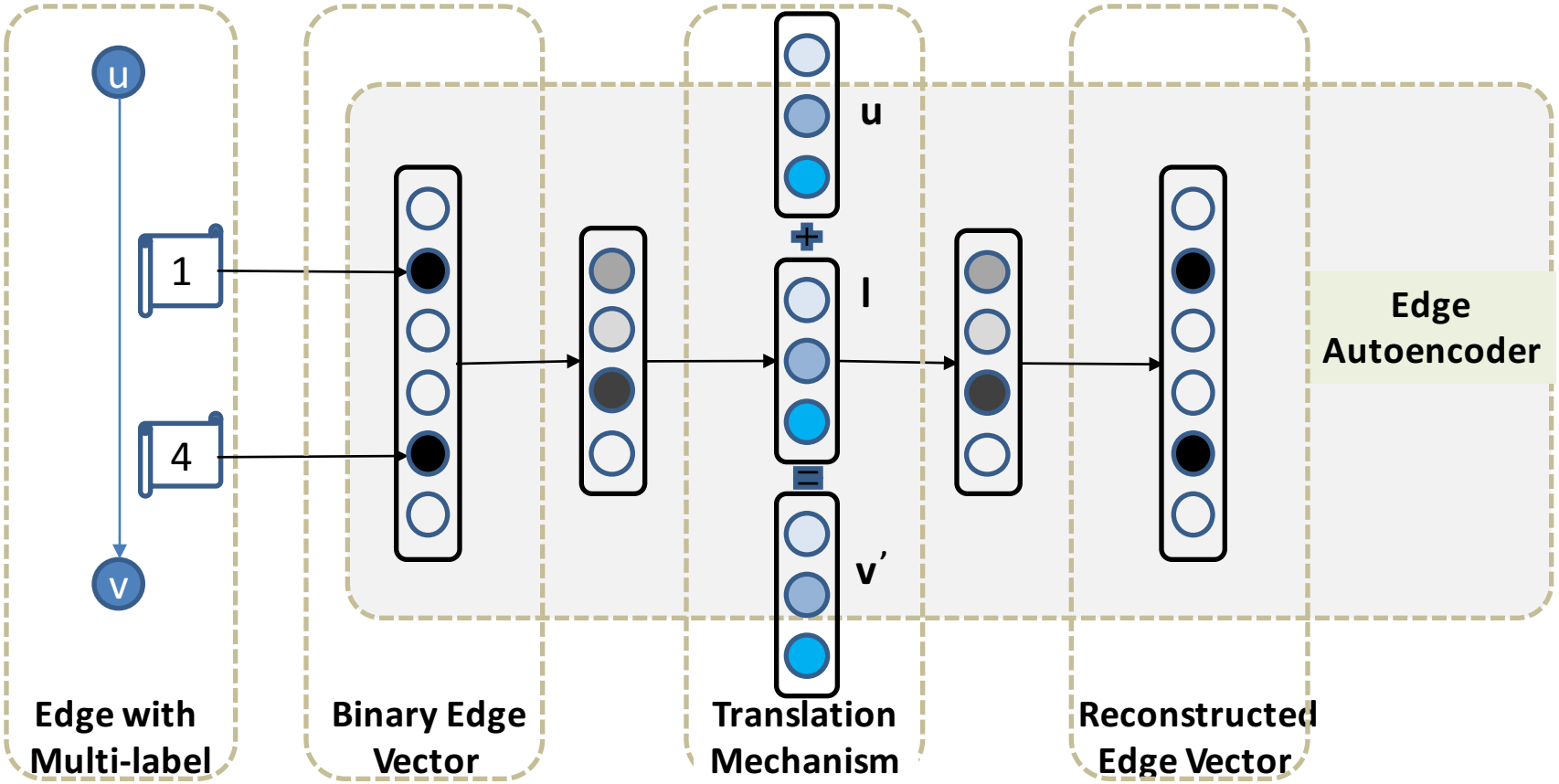
- 平移机制
  - 每个节点有两个表示向量，例如：**u, u'**
  - 一条边上节点表示及标签集合表示具有平移关系



$$\mathcal{L}_{trans} = \max(\gamma + d(\mathbf{u} + \mathbf{l}, \mathbf{v}') - d(\hat{\mathbf{u}} + \hat{\mathbf{l}}, \hat{\mathbf{v}}'), 0)$$

面向社会关系抽取的网络表示

- 标签集合表示构建



$$\mathcal{L}_{ae} = ||(\mathbf{s} - \hat{\mathbf{s}}) \odot \mathbf{x}||$$

面向社会关系抽取的网络表示

- 未标注边的关系预测
  - 近似计算边的表示

$$\mathbf{l} = \mathbf{v}' - \mathbf{u}$$

  - 对于近似的边表示进行解码

实验结果

- 社会关系抽取结果

Metric	*hits*@1	*hits*@5	*hits*@10	MeanRank	*hits*@1	*hits*@5	*hits*@10	MeanRank
DeepWalk	13.88	36.80	50.57	19.69	18.78	39.62	52.55	18.76
LINE	11.30	31.70	44.51	23.49	15.33	33.96	46.04	22.54
node2vec	13.63	36.60	50.27	19.87	18.38	39.41	52.22	18.92
TransE	39.16	78.48	88.54	5.39	57.48	84.06	90.60	4.44
TransNet	**47.67**	**86.54**	**92.27**	**5.04**	**77.22**	**90.46**	**93.41**	**4.09**

Metric	*hits*@1	*hits*@5	*hits*@10	MeanRank	*hits*@1	*hits*@5	*hits*@10	MeanRank
DeepWalk	7.27	21.05	29.49	81.33	11.27	23.27	31.21	78.96
LINE	5.67	17.10	24.72	94.80	8.75	18.98	26.14	92.43
node2vec	7.29	21.12	29.63	80.80	11.34	23.44	31.29	78.43
TransE	19.14	49.16	62.45	25.52	31.55	55.87	66.83	23.15
TransNet	**27.90**	**66.30**	**76.37**	**25.18**	**58.99**	**74.64**	**79.84**	**22.81**

Metric	*hits*@1	*hits*@5	*hits*@10	MeanRank	*hits*@1	*hits*@5	*hits*@10	MeanRank
DeepWalk	5.41	16.17	23.33	102.83	7.59	17.71	24.58	100.82
LINE	4.28	13.44	19.85	114.95	6.00	14.60	20.86	112.93
node2vec	5.39	16.23	23.47	102.01	7.53	17.76	24.71	100.00
TransE	15.38	41.87	55.54	32.65	23.24	47.07	59.33	30.64
TransNet	**28.85**	**66.15**	**75.55**	**29.60**	**56.82**	**73.42**	**78.60**	**27.40**

实验结果

- 不同频度标签对比

Table 5: Label comparisons on Arnet-S. ($\times 100$ for $hits@k$)

Tags	Top 5 labels				Bottom 5 labels			
Metric	$hits@1$	$hits@5$	$hits@10$	MeanRank	$hits@1$	$hits@5$	$hits@10$	MeanRank
TransE	58.82	85.68	91.61	**3.70**	52.21	82.03	87.75	5.65
TransNet	**77.26**	**90.35**	**93.53**	3.89	**78.27**	**90.44**	**93.30**	**4.18**

示例

- 对于“A. Swami ”不同邻居的关系标签推荐结果

Table 6: Recommended top-3 labels for each neighbor.

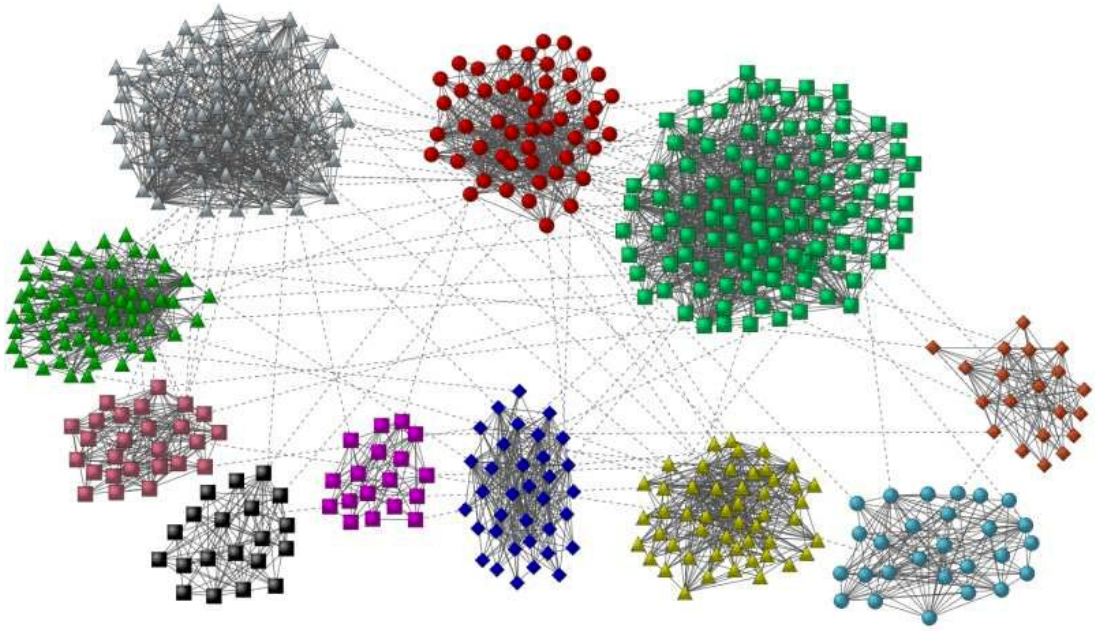
Neighbors	TransE	TransNet
Matthew Duggan	**ad hoc network**; wireless sensor network; wireless sensor networks	**management system**; **ad hoc network**; wireless sensor
K. Pelechrinis	**wireless network**; wireless networks; ad hoc network	**wireless network**; wireless sensor network; **routing protocol**
Oleg Korobkin	**wireless network**; wireless networks; **wireless communication**	**resource management**; **system design**; **wireless network**

研究框架

- 显式网络表示
  - 基于词项的网络表示：用户属性预测 (ACM TIST)
  - 基于主题标签的网络表示 ：用户标签推荐 (JCST)
- 隐式网络表示
  - 最大间隔网络表示：节点标签类别信息 (IJCAI 2016)
  - 上下文相关网络表示：节点文本信息 (ACL 2017)
  - 面向社会关系抽取的网络表示：边标签信息 (IJCAI 2017)
  - 全局社区优化网络表示：社区信息 (IEEE TKDE)

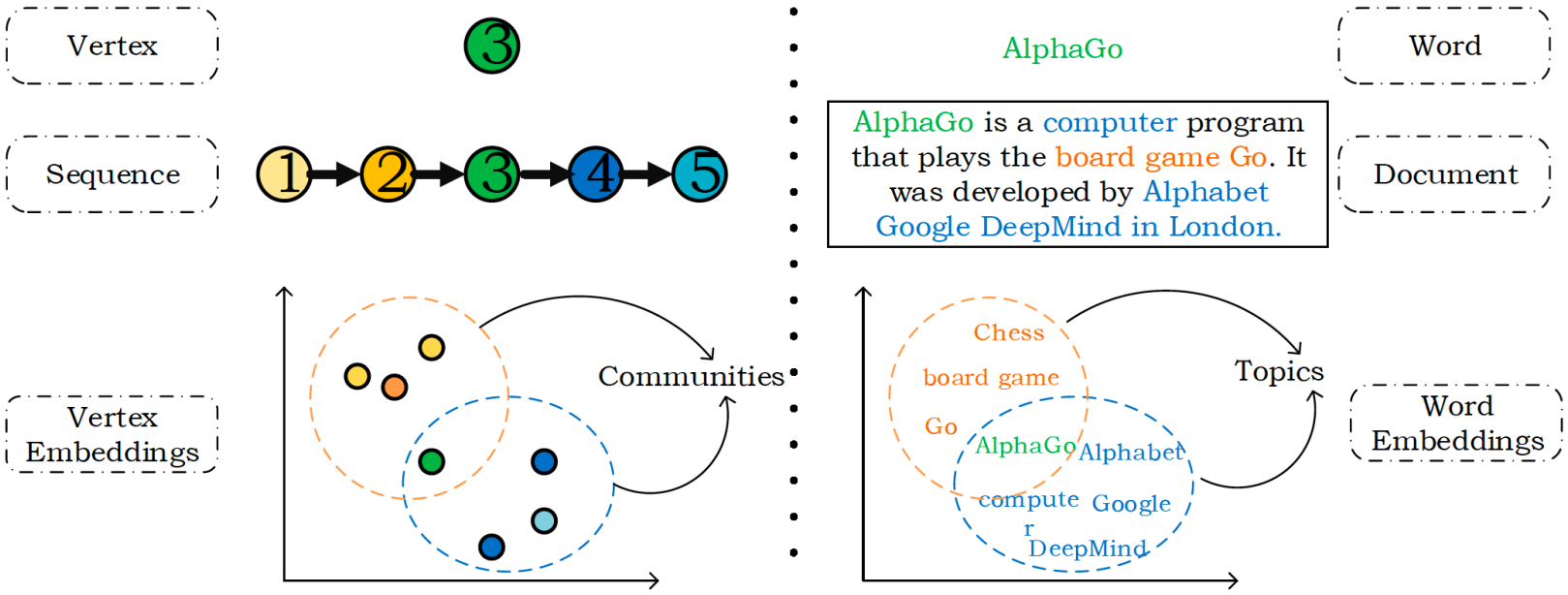
社区信息

- 真实世界中的网络往往拥有重要的社区信息
  - 社区内部的节点连接更加紧密
  - 社区内部的节点倾向于拥有同样的属性



社区优化的网络表示

- Communtiy-Enhanced NRL (CNRL)
  - DeepWalk: 将节点看做词，节点序列看做句子
  - CNRL: 将节点构成的社区看做词构成的主题



社区优化的网络表示

- 随机游走
- 主题模型训练
- 社区标签分配
- 节点表示学习

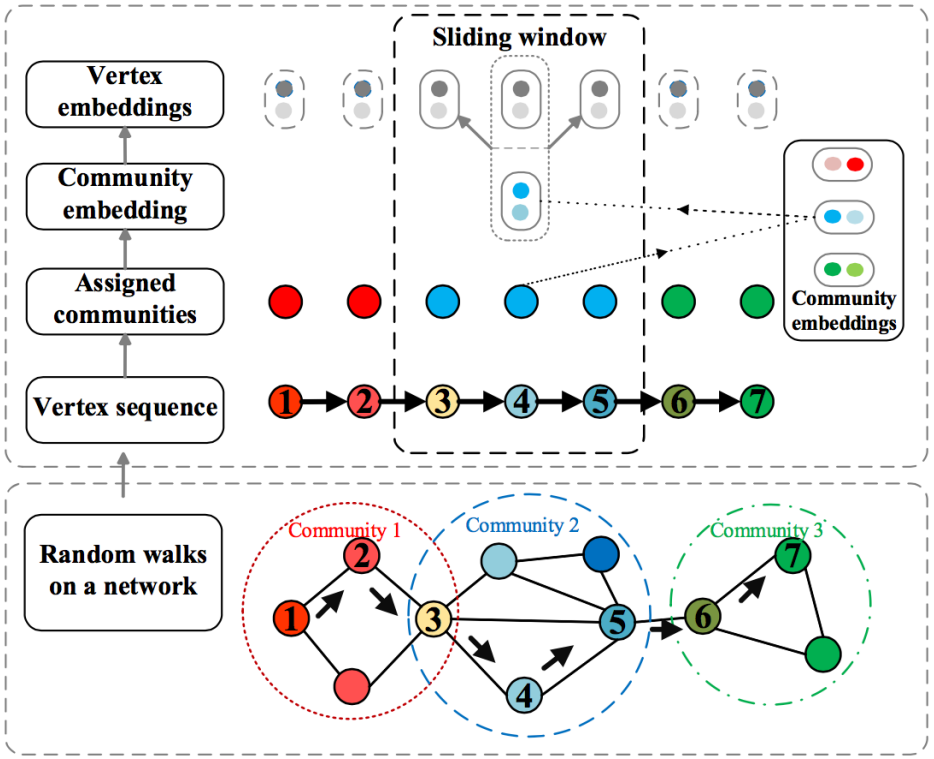


Figure 1: An illustration of CNRL.

社区标签分配

- 根据p(c|v,s)向序列中的每个节点分配社区标签

$$p(c|v,s) \propto p(v|c)p(c|s)$$

- 基于统计的分配：

$$\Pr(v|c) = \frac{N(v,c)+\beta}{\sum_{\tilde{v}\in V} N(\tilde{v},c)+|V|\beta}, \quad \Pr(c|s) = \frac{N(c,s)+\alpha}{\sum_{\tilde{c}\in C} N(\tilde{c},s)+|K|\alpha}.$$

- 基于表示的分配：

$$\Pr(c|s) = \frac{\exp(\mathbf{c}\cdot\mathbf{s})}{\sum_{\tilde{c}\in C}\exp(\tilde{\mathbf{c}}\cdot\mathbf{s})},$$

节点表示学习

- 对于每个序列，优化目标为：

$$\mathcal{L}(s) = \frac{1}{|s|}\sum_{i=1}^{|s|}\sum_{i-t\le j\le i+t, j\neq i} \log \Pr(v_j|v_i) + \log \Pr(v_j|c_i),$$

- 节点最终的表示为：

$$\widehat{v} = v \oplus v_c$$

$$v_c = \sum_{\tilde{c}\in C} p(\tilde{c}|v)\tilde{c}$$

实验结果

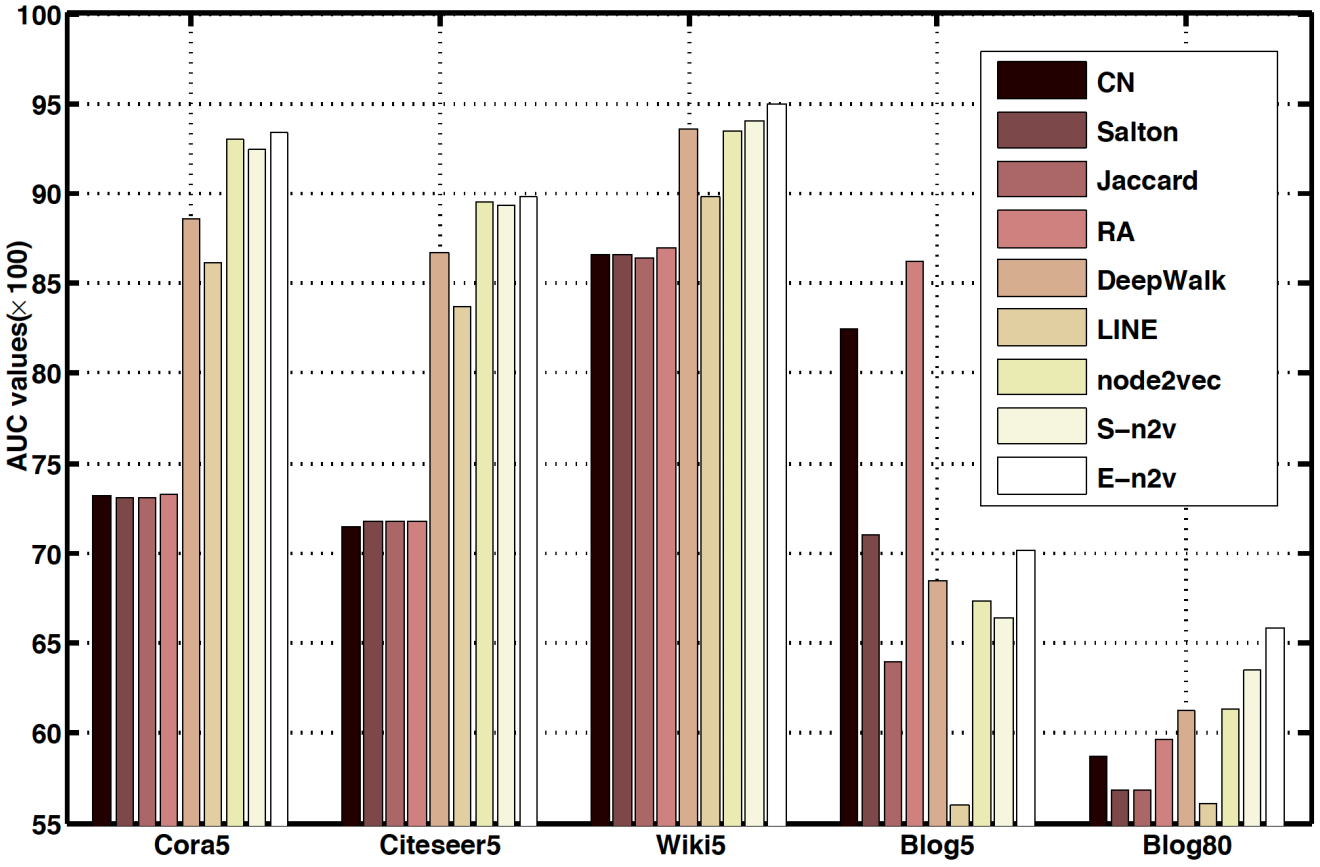
- 节点分类结果

Table 2: Vertex classification results. (%)

Dataset	Cora			Citeseer			Wiki			BlogCatalog		
%Training ratio	10%	50%	90%	10%	50%	90%	10%	50%	90%	1%	5%	9%
DeepWalk	70.77	75.62	77.27	47.92	54.21	55.33	58.54	65.90	67.56	23.66	30.58	32.31
LINE	70.61	78.66	79.67	44.27	51.93	54.01	57.53	66.55	68.31	19.31	25.89	30.04
node2vec	73.29	78.40	79.15	49.47	55.87	57.56	58.93	66.03	68.99	24.47	30.87	31.96
MNMF	75.08	79.82	79.41	51.62	56.81	57.22	54.76	62.74	64.77	19.26	25.24	27.16
S-DW	74.14	80.33	81.62	49.72	57.05	58.59	59.72	67.75	70.75	23.80	30.25	32.45
E-DW	74.27	78.88	79.41	49.93	55.76	57.19	59.23	67.00	67.89	24.93	31.19	32.76
S-n2v	75.86	**82.81**	**83.65**	**53.12**	**60.31**	**62.63**	**60.66**	**68.92**	**71.60**	24.95	30.95	32.34
E-n2v	**76.30**	81.46	82.80	51.84	57.19	58.51	60.07	67.64	70.71	**25.75**	**31.13**	**32.60**

实验结果

- 链接预测结果


AUC values(×100)
100
95
90
85
80
75
70
65
60
55
Cora5
Citeseer5
Wiki5
Blog5
Blog80
CN
Salton
Jaccard
RA
DeepWalk
LINE
node2vec
S−n2v
E−n2v

Community-enhanced NRL

- 社区发现结果

Datasets	SCP	LC	BigCLAM	S-DW	E-DW	S-n2v	E-n2v
Cora	0.076	0.334	0.464	0.464	**1.440**	0.447	1.108
Citeseer	0.055	0.315	0.403	0.486	**1.861**	0.485	1.515
Wiki	0.063	0.322	0.286	0.291	**0.564**	0.260	0.564

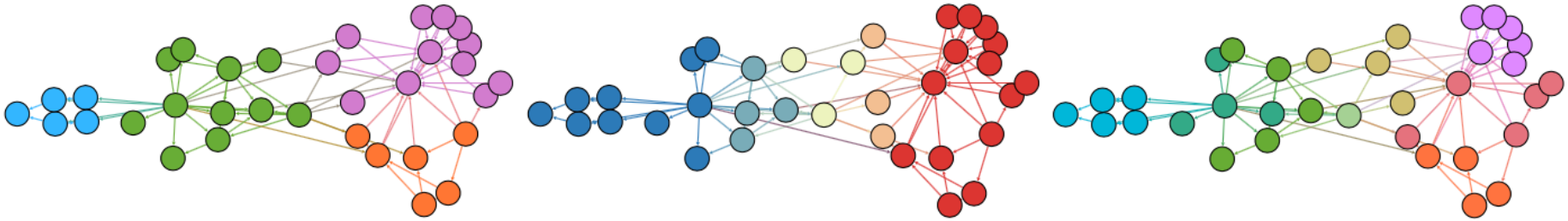


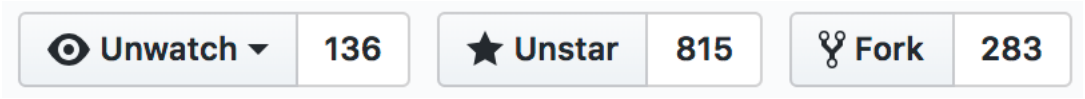
Figure 2: Detected communities on Karate (Fast Unfolding, 2 communities by CNRL, 4 communities by CNRL).

总结与展望

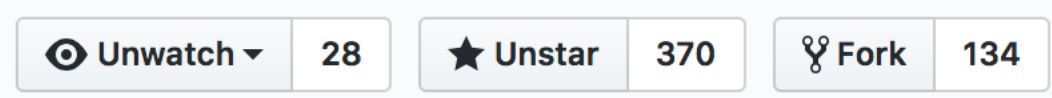
- 大规模网络表示学习
- 网络表示模型压缩
- 动态网络表示学习
- 知识驱动的网络表示学习
  - 为社会计算引入推理能力
  - 增强网络表示学习可解释性
- 面向具体应用的网络表示学习
  - 与具体应用场景相结合
  - 推荐系统
  - 用户画像

代码开源成果

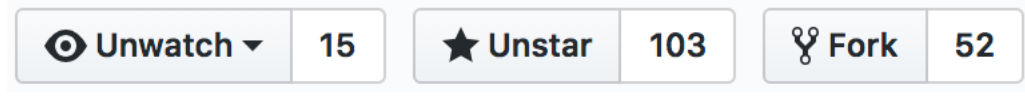
- http://github.com/thunlp/NRLpapers



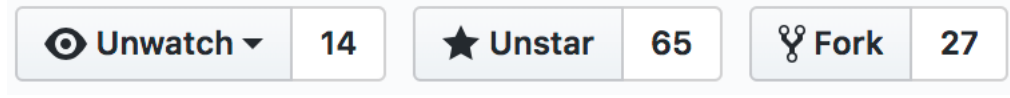
- http://github.com/thunlp/OpenNE



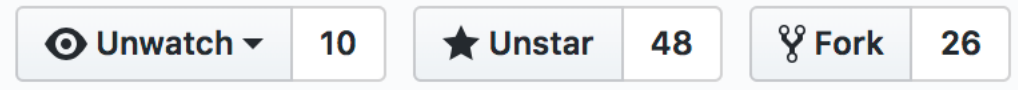
- http://github.com/thunlp/CANE



- http://github.com/thunlp/TransNet



- http://github.com/thunlp/MMDW



从2017年4月至今，
star累计1400+
fork累计520+

论文发表

- **Cunchao Tu,** Zhiyuan Liu, Huanbo Luan, Maosong Sun. 2017. PRISM: Profession Identification in Social Media. ACM TIST.
- **Cunchao Tu,** Zhiyuan Liu, Maosong Sun. 2015. Tag Correspondence Model for User Tag Suggestion. JCST.
- **Cunchao Tu,** Weicheng Zhang, Zhiyuan Liu, Maosong Sun. 2016. Max-Margin DeepWalk: Discriminative Learning of Network Representation. In Proceedings of IJCAI.
- **Cunchao Tu,** Han Liu, Zhiyuan Liu, Maosong Sun. 2017. CANE: Context-Aware Network Embedding for Relation Modeling. In Proceedings of ACL.
- **Cunchao Tu,** Zhengyan Zhang, Zhiyuan Liu, Maosong Sun. 2017. TransNet: Translation-Based Network Representation Learning for Social Relation Extraction. In Proceedings of IJCAI.
- **Cunchao Tu,** Xiangkai Zeng, Hao Wang, Zhiyuan Liu, Maosong Sun, Bo Zhang, Leyu Lin. 2018. Community-enhanced network representation learning for network analysis. IEEE TKDE (minor revision).
- **Cunchao Tu,** Zhiyuan Liu, Maosong Sun. 2014. Inferring Correspondences from Multiple Sources for Microblog User Tags. In Proceedings of SMP. **Best paper award**.
- **Cunchao Tu**, Zhiyuan Liu, Huanbo Luan, Maosong Sun. 2015. PRISM: Profession Identification in Social Media with Personal Information and Community Structure. In Proceedings of SMP.
- **Cunchao Tu**, Cheng Yang, Zhiyuan Liu, Maosong Sun. 2017. Network Representation Learning: An Overview. *Science China: Information Sciences. (In Chinese)*.

谢谢！