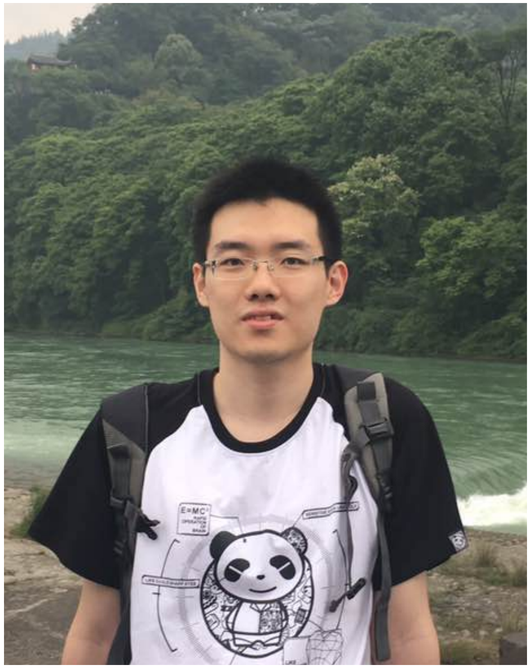
社会计算与表示学习

清华大学自然语言处理实验室

刘知远

合作者

涂存超



杨成

社会计算

社会计算的研究对象

社会网络

用户及其关系和行为

媒体信息

文本、视频、语音等信息

面临挑战

信息多源异构，难以建立语义关联

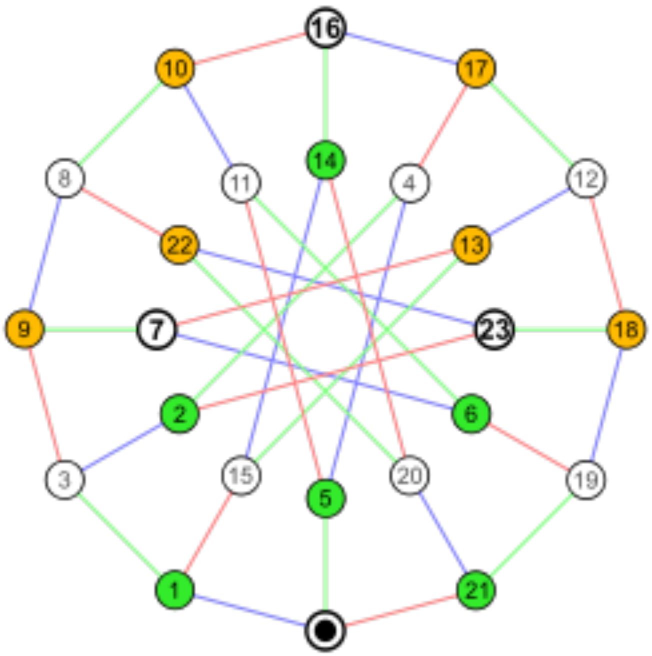
基于符号的表示方案

star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...]

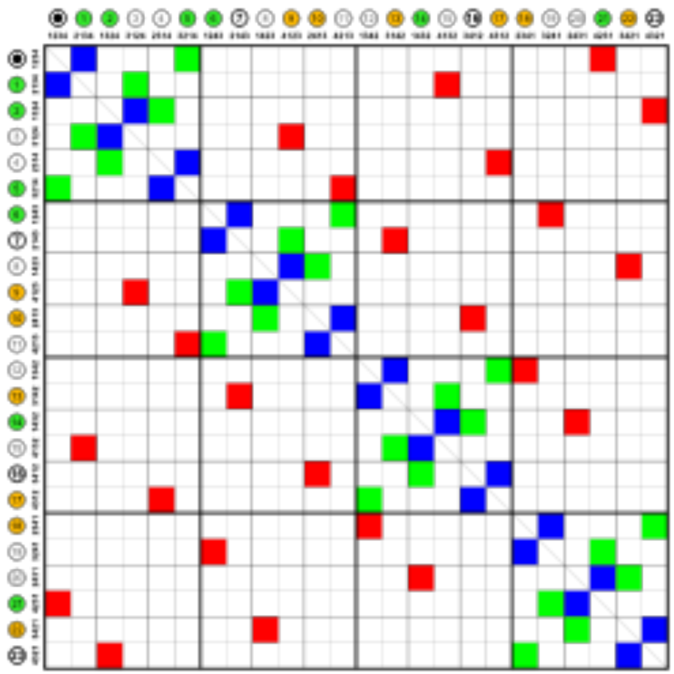
sun [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...]

sim(star, sun) = 0

基于符号的表示方案



N个节点的网络



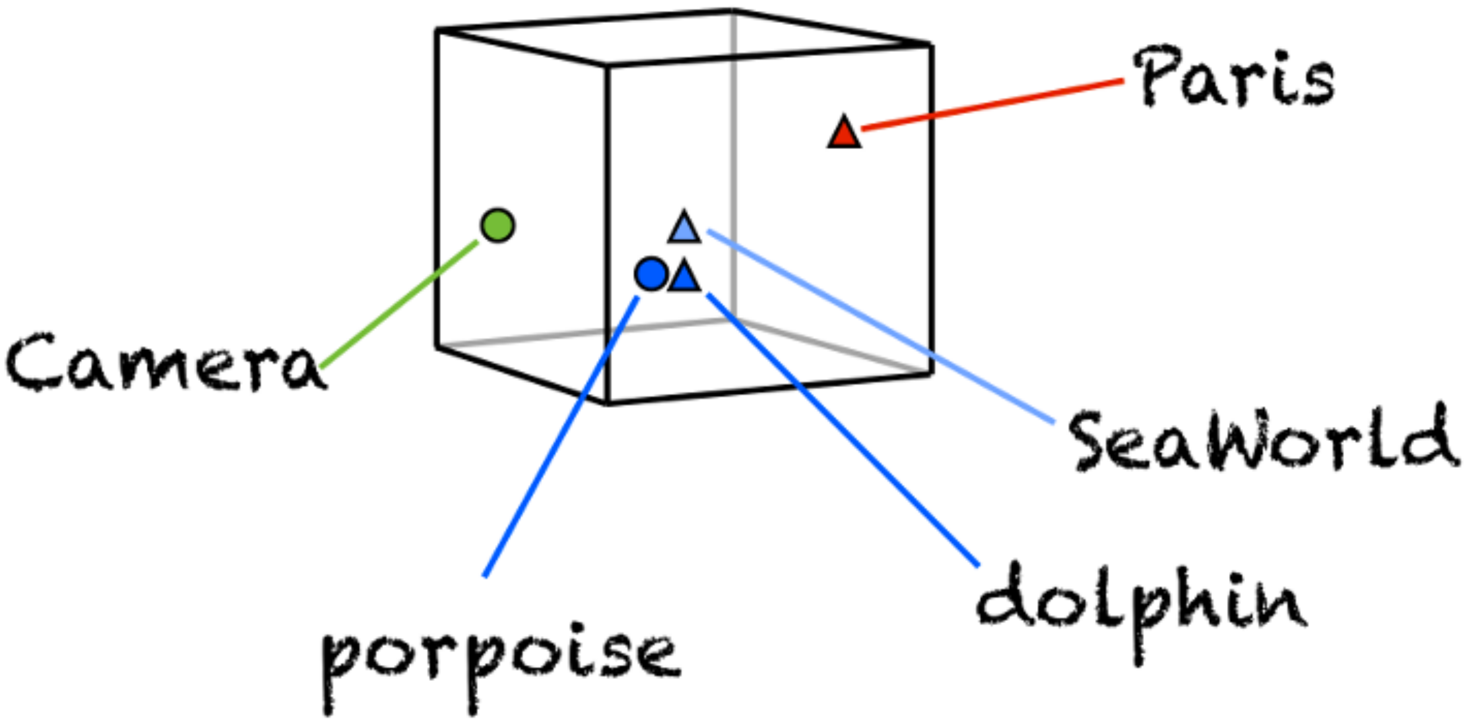
邻接矩阵

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



分布式表示的优势

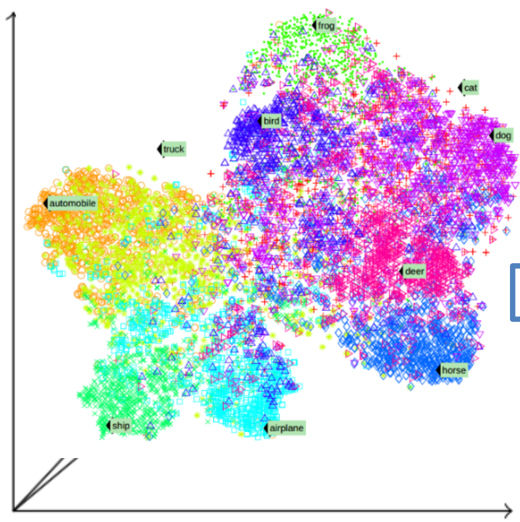
- 解决社会计算异质对象间的语义计算问题

知识

产品

文本

用户



统一语义空间

个性推荐

知识计算

社会网络分析

分布式表示的优势

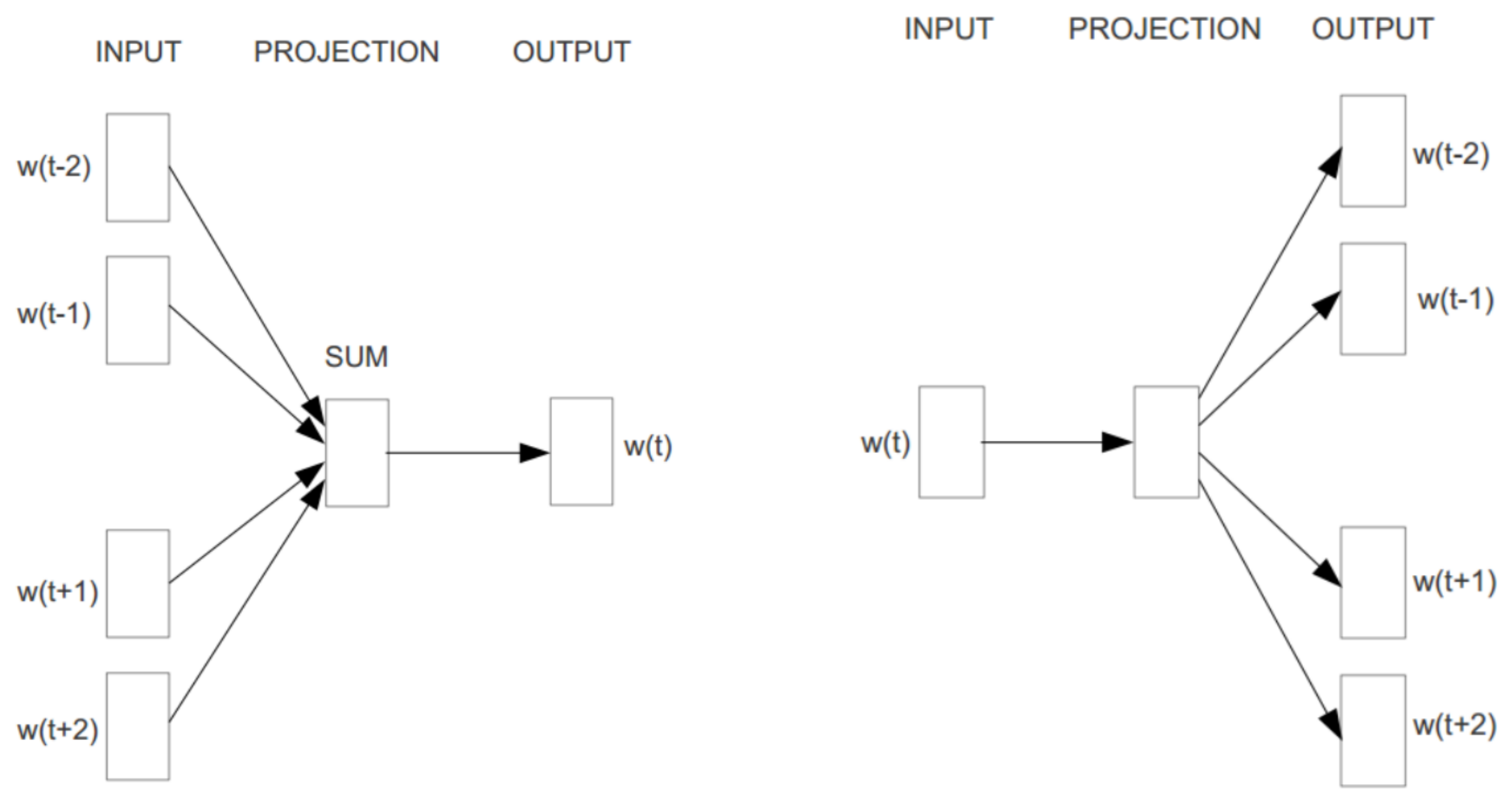
表层数据

深层语义

内容提要

- 网络表示学习方案
- 引入外部信息的网络表示学习
- 网络表示学习应用
- 展望

分布式词表示学习模型



word2vec

Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. NIPS 2013.

词汇表示用于词汇相似度计算

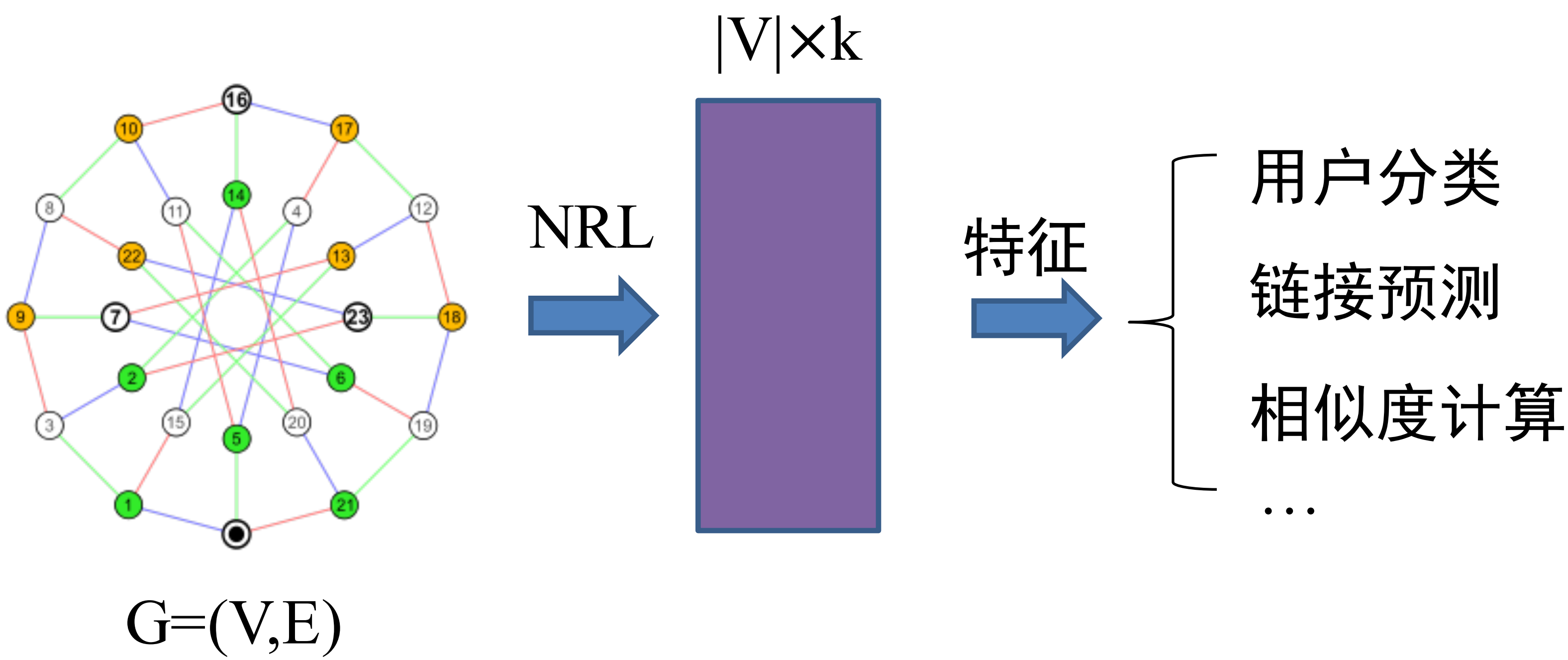
```
(EXIT to break): china

n vocabulary: 486

                          Word       Cosine distance
----------------------------------------------------
                        taiwan              0.768188
                         japan              0.652825
                         macau              0.614888
                         korea              0.614887
                           prc              0.613579
                       beijing              0.605946
                        taipei              0.592367
                      thailand              0.577905
                      cambodia              0.575681
                     singapore              0.569950
                      republic              0.567597
                      mongolia              0.554642
                       chinese              0.551576
```

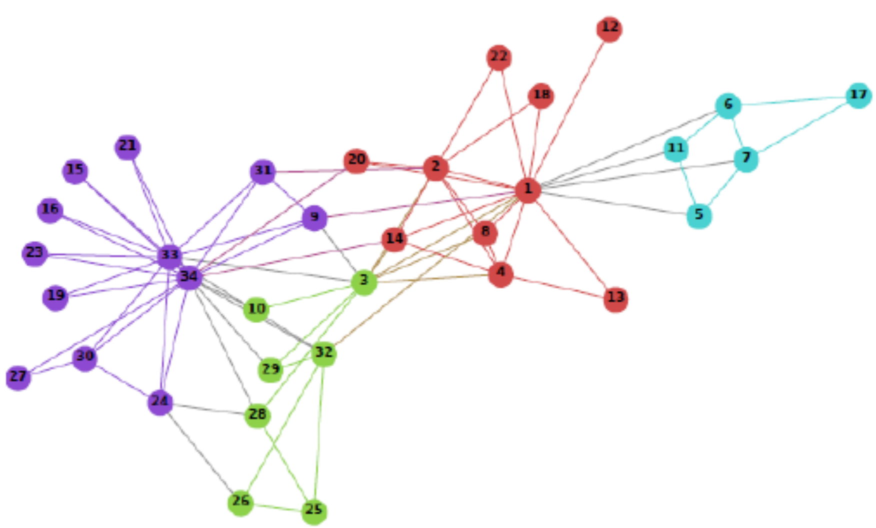
网络表示学习

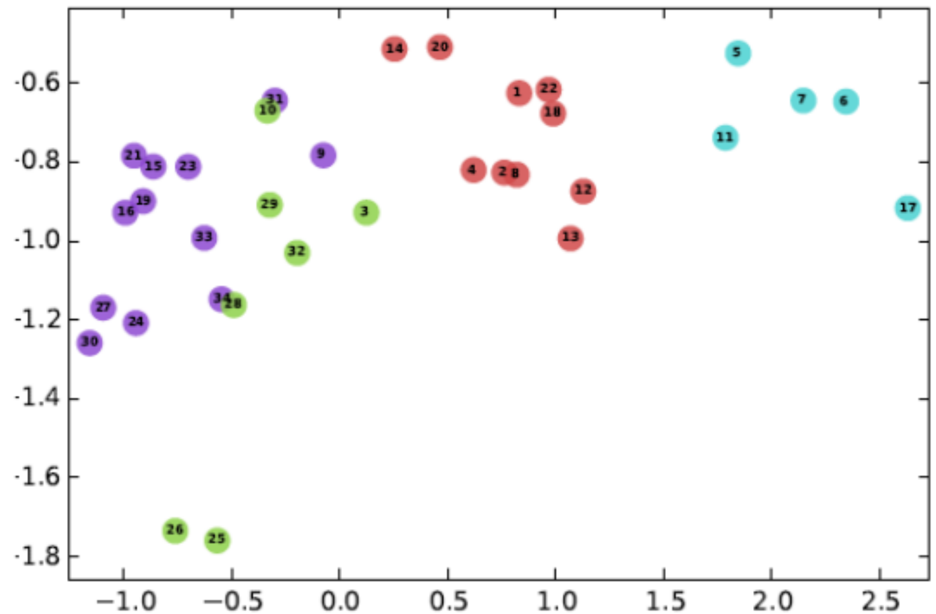
- 将网络中节点的语义信息表示为低维向量



网络表示学习

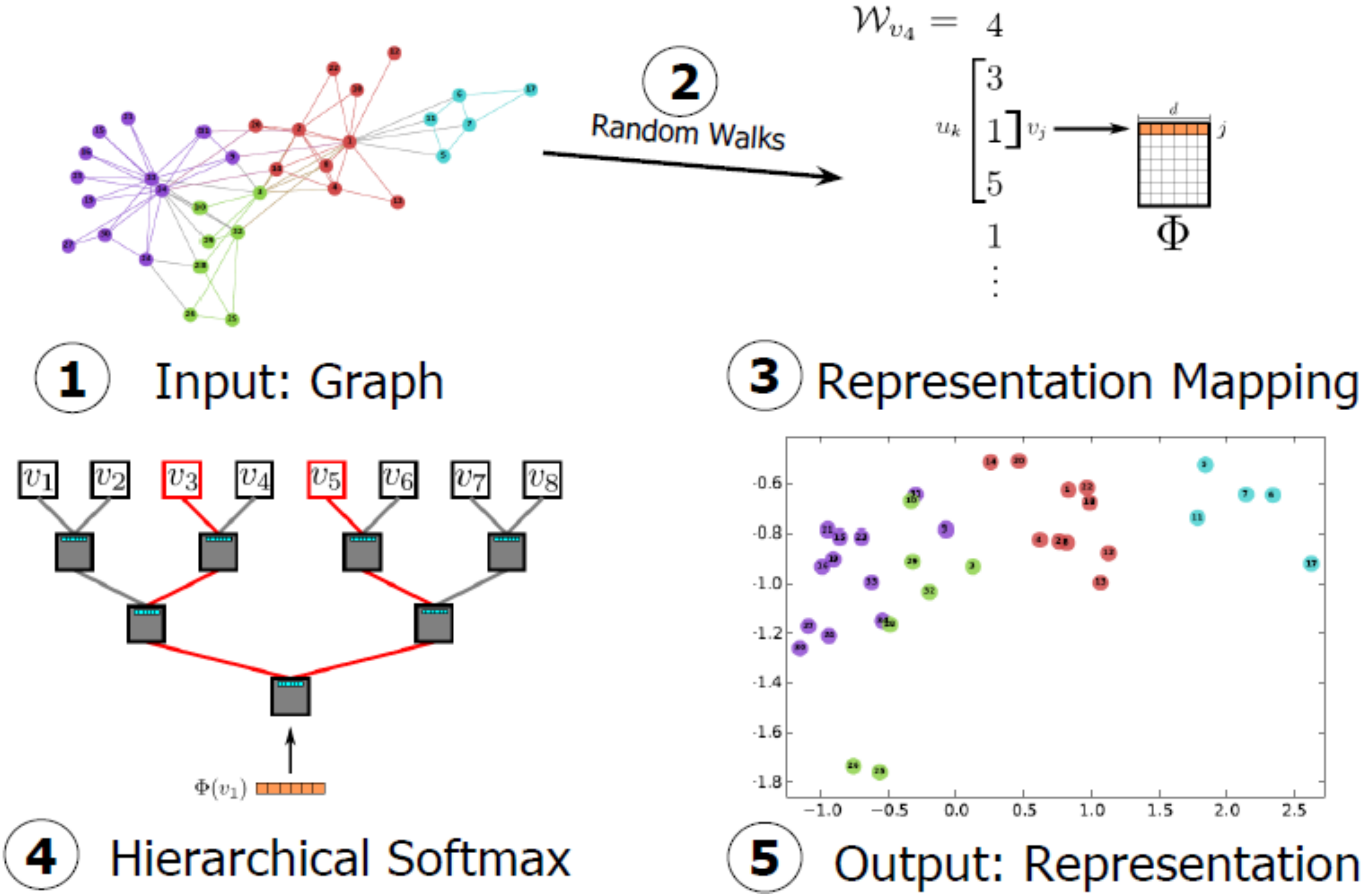
- 跆拳道俱乐部社会网络 (k=2)





Perozzi et al. DeepWalk: Online Learning of Social Representations. KDD 2014

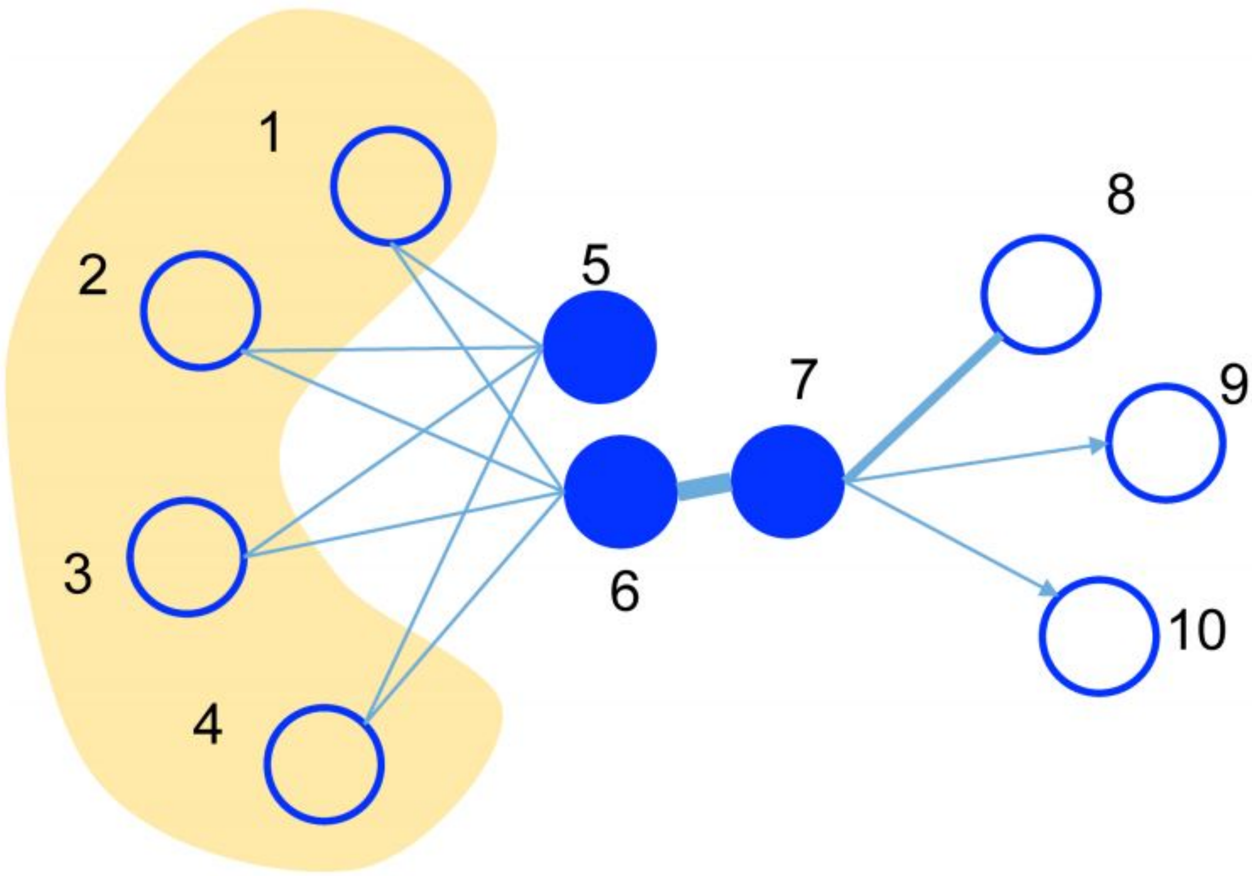
DeepWalk



1 Input: Graph
2 Random Walks
$\mathcal{W}_{v_4} = 4$
3 1 5 1
u_k v_j d j
Φ
3 Representation Mapping
v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8
$\Phi(v_1)$
4 Hierarchical Softmax
-0.6 -0.8 -1.0 -1.2 -1.4 -1.6 -1.8
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
5 Output: Representation

LINE

- 一阶和二阶邻近度

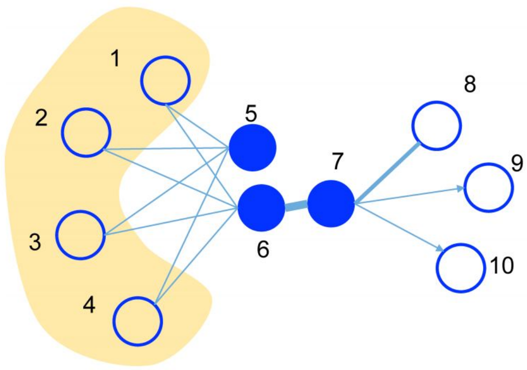


Tang et al. LINE: Large-scale Information Network Embedding. WWW 2015

LINE

- 一阶邻近度
  - 由强关系连接的6和7表示应该相近

$$p_1(v_i, v_j) = \frac{1}{1 + \exp(-\vec{u}_i^T \cdot \vec{u}_j)}$$

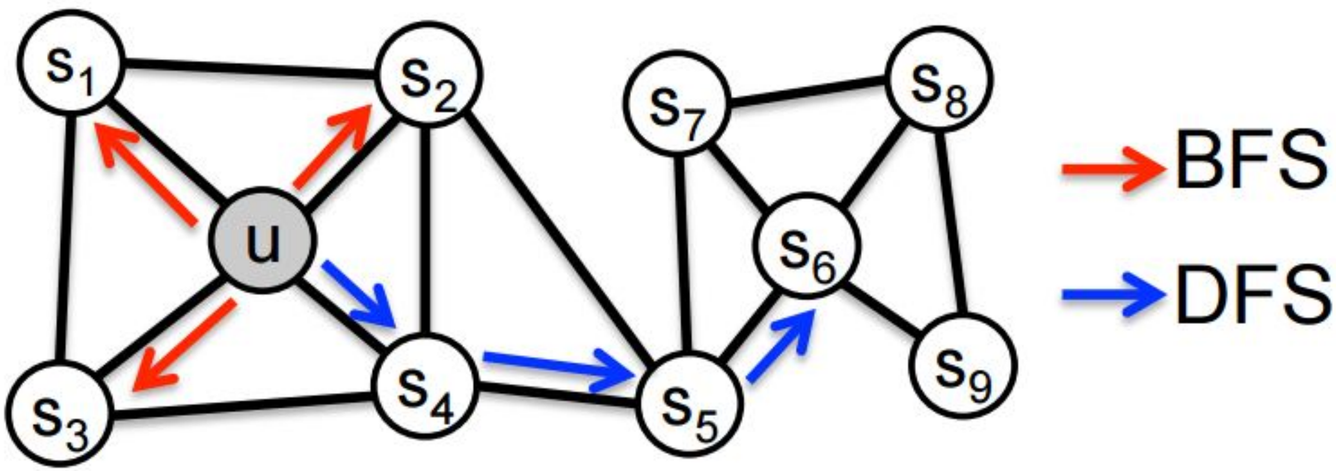


- 二阶邻近度
  - 共同邻居多的5和6的表示应该相近

$$p_2(v_j|v_i) = \frac{\exp(\vec{u}_j'^T \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_k'^T \cdot \vec{u}_i)}$$

node2vec

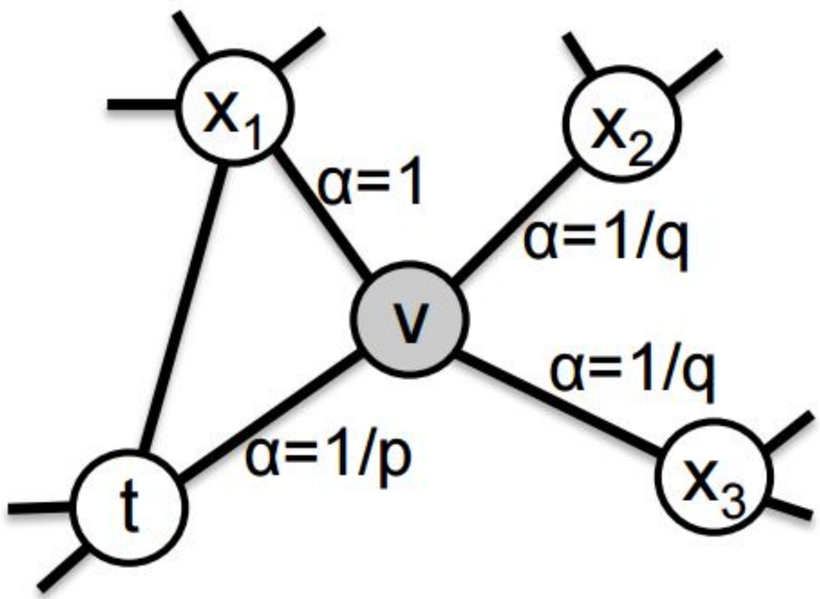
- 随机游走策略
  - 宽度优先搜索：微观局部信息
  - 深度优先搜索：宏观全局信息



Grover and Leskovec, node2vec: Scalable Feature Learning for Networks. KDD 2016

node2vec

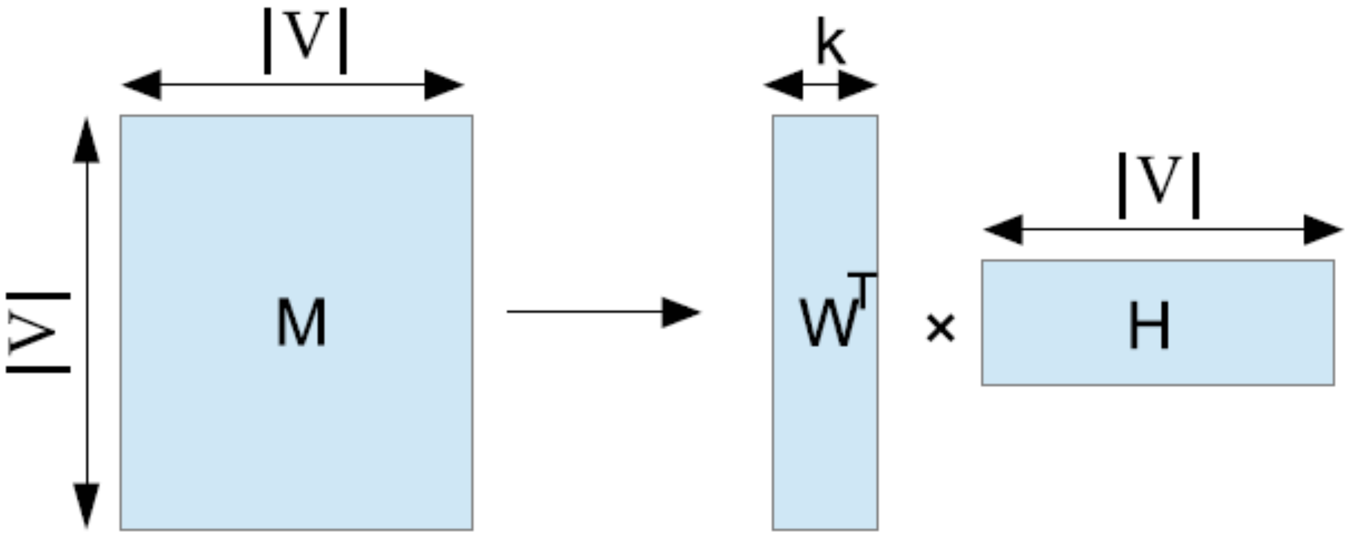
- 参数控制的随机游走
  - 返回概率参数p，对应BFS
  - 离开概率参数q，对应DFS



$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

网络表示学习与矩阵分解的关系

- 从数学上证明了Spectral Clustering, DeepWalk和GraRep等网络表示学习算法等价于矩阵分解



Yang et al. Network Representation Learning with Rich Text Information. IJCAI 2015.
Yang et al. Fast Network Embedding Enhancement via High Order Proximity Approximation

性能比较

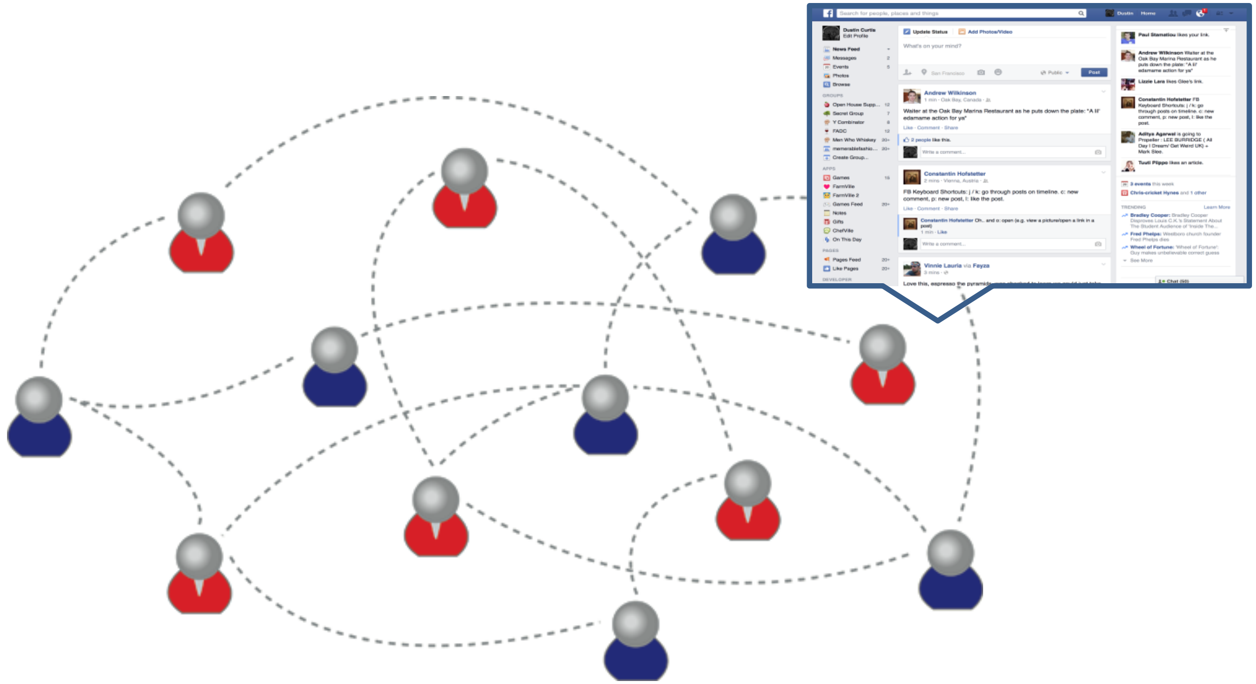
Algorithm	Dataset		
	BlogCatalog	PPI	Wikipedia
Spectral Clustering	0.0405	0.0681	0.0395
DeepWalk	0.2110	0.1768	0.1274
LINE	0.0784	0.1447	0.1164
node2vec	**0.2581**	**0.1791**	**0.1552**

内容提要

- 网络表示学习方案
- 引入外部信息的网络表示学习
- 网络表示学习应用
- 展望

引入文本信息

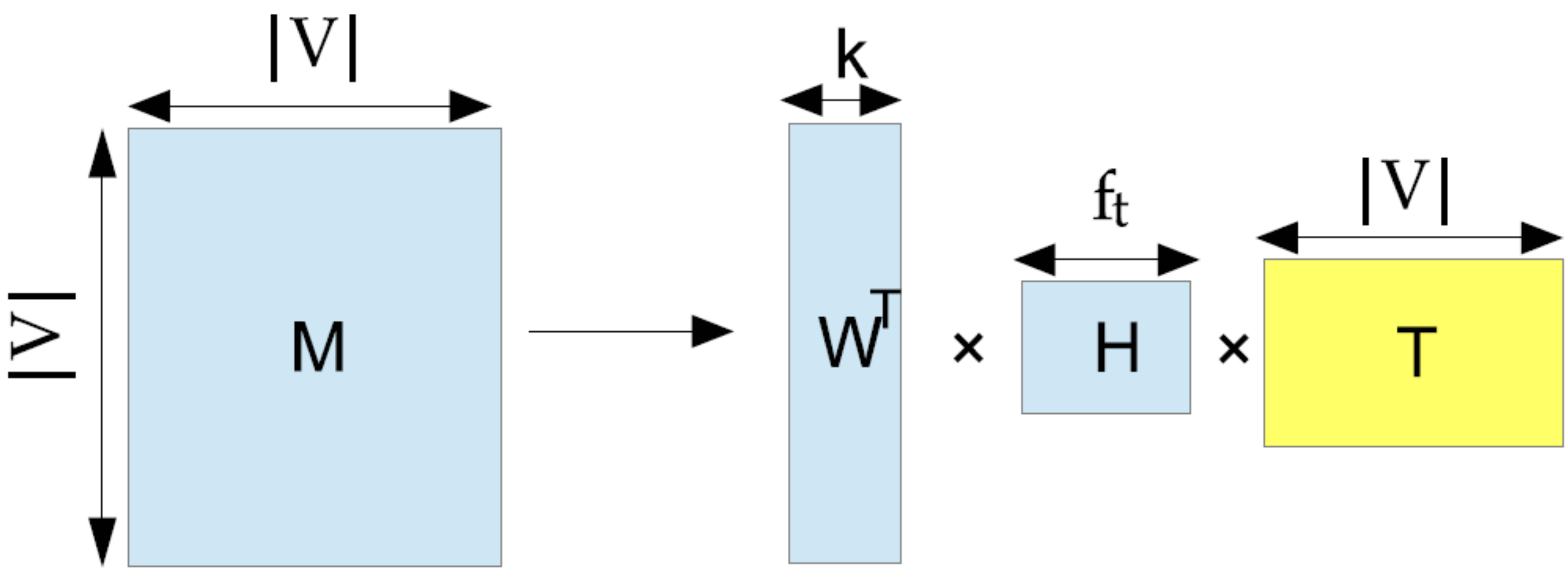
- 网络中存在丰富的文本信息
- 将节点文本信息嵌入网络表示学习



Yang et al. Network Representation Learning with Rich Text Information. IJCAI 2015.

Text-Associated DeepWalk （TADW）

- 矩阵分解框架



$$\min_{W,H} ||M - W^T HT||_F^2 + \frac{\lambda}{2}(||W||_F^2 + ||H||_F^2).$$

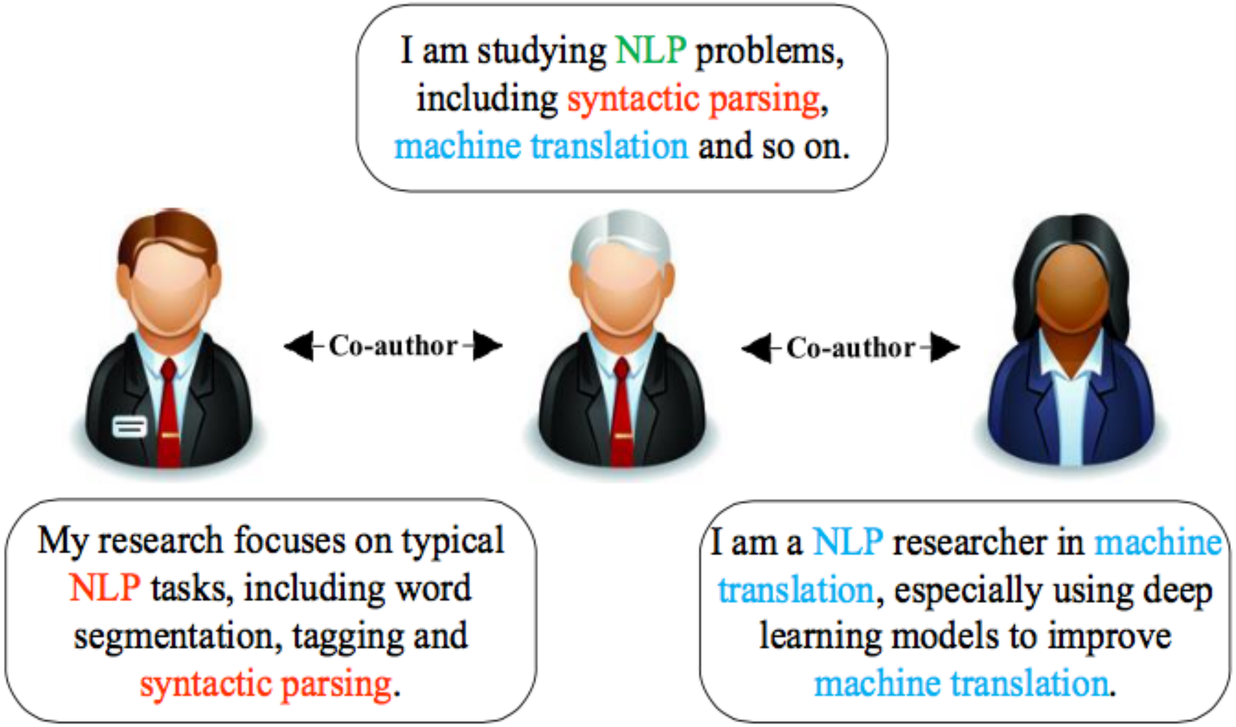
网络节点分类效果

Table 1: Evaluation results on Cora dataset.

Classifier	Transductive SVM				SVM				
% Labeled Nodes	1%	3%	7%	10%	10%	20%	30%	40%	50%
DeepWalk	62.9	68.3	72.2	72.8	76.4	78.0	79.5	80.5	81.0
PLSA	47.7	51.9	55.2	60.7	57.0	63.1	65.1	66.6	67.6
Text Features	33.0	43.0	57.1	62.8	58.3	67.4	71.1	73.3	74.0
Naive Combination	67.4	70.6	75.1	77.4	76.5	80.4	82.3	83.3	84.1
NetPLSA	65.7	67.9	74.5	77.3	80.2	83.0	84.0	84.9	85.4
TADW	**72.1**	**77.0**	**79.1**	**81.3**	**82.4**	**85.0**	**85.6**	**86.0**	**86.7**

语境感知问题

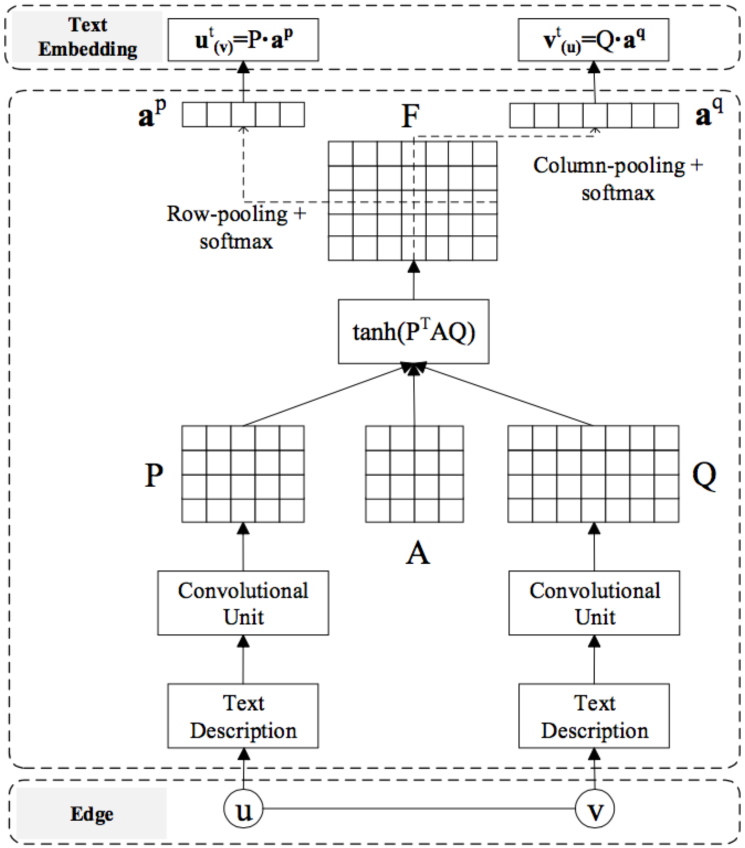
- 节点与邻居节点进行交互时，展现出不同方面



传统NRL简单使边上的两个节点表示相近
不能很好地对具体关系建模

Context-Aware Network Embedding

- 根据节点不同邻居，学习不同的向量表示
- 利用文本信息进行相互关注（mutual attention）



Tu, et al. CANE: Context-Aware Network Embedding for Relation Modeling. ACL 2017.

Context-Aware Network Embedding

- 链接预测结果

%Removed edges	15%	25%	35%	45%	55%	65%	75%	85%	95%
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
Naive Combination	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	89.8	91.2	92.3	91.8	93.2	92.9	93.2
CANE (text only)	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**

Table 3: AUC values on HepTh. ($\alpha = 0.7, \beta = 0.2, \gamma = 0.2$)

%Removed edges	15%	25%	35%	45%	55%	65%	75%	85%	95%
DeepWalk	56.6	58.1	60.1	60.0	61.8	61.9	63.3	63.7	67.8
LINE	52.3	55.9	59.9	60.9	64.3	66.0	67.7	69.3	71.1
node2vec	54.2	57.1	57.3	58.3	58.7	62.5	66.2	67.6	68.5
Naive Combination	55.1	56.7	58.9	62.6	64.4	68.7	68.9	69.0	71.5
TADW	52.3	54.2	55.6	57.3	60.8	62.4	65.2	63.8	69.0
CENE	56.2	57.4	60.3	63.0	66.3	66.0	70.2	69.8	73.8
CANE (text only)	55.6	56.9	57.3	61.6	63.6	67.0	68.5	70.4	73.5
CANE (w/o attention)	56.7	59.1	60.9	64.0	66.1	68.9	69.8	71.0	74.3
CANE	**56.8**	**59.3**	**62.9**	**64.5**	**68.9**	**70.4**	**71.4**	**73.6**	**75.4**

Table 4: AUC values on Zhihu. ($\alpha = 1.0, \beta = 0.3, \gamma = 0.3$)

Context-Aware Network Embedding

- Mutual Attention

Edge #1: (A, B)
Machine Learning research making great progress many directions This article summarizes four directions discusses current open problems The four directions improving classification accuracy learning ensembles classifiers methods scaling supervised learning algorithms reinforcement learning learning complex stochastic models

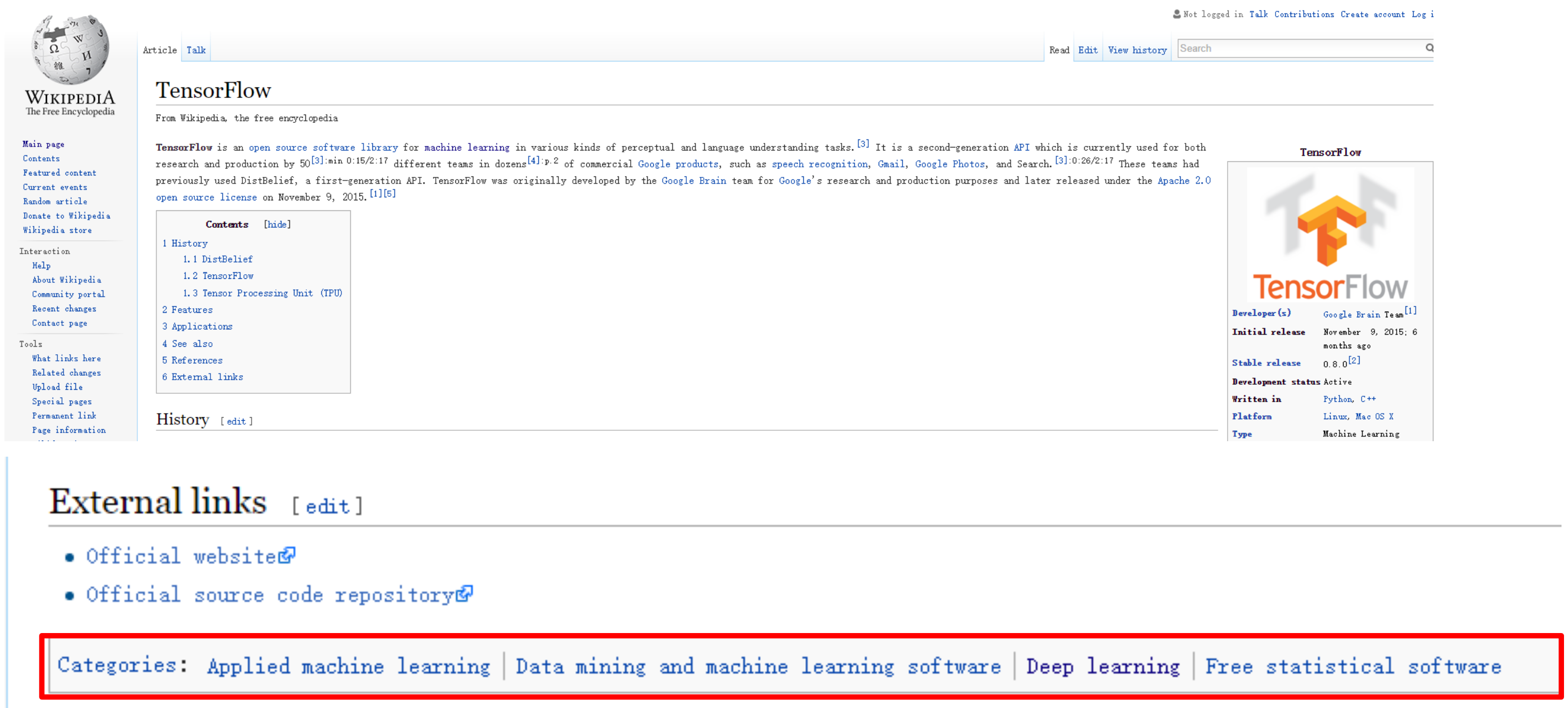
The problem making optimal decisions uncertain conditions central Artificial Intelligence If state world known times world modeled Markov Decision Process MDP MDPs studied extensively many methods known determining optimal courses action policies The realistic case state information partially observable Partially Observable Markov Decision Processes POMDPs received much less attention The best exact algorithms problems inefficient space time We introduce Smooth Partially Observable Value Approximation SPOVA new approximation method quickly yield good approximations improve time This method combined reinforcement learning methods combination effective test cases

Edge #2: (A, C)
Machine Learning research making great progress many directions This article summarizes four directions discusses current open problems The four directions improving classification accuracy learning ensembles classifiers methods scaling supervised learning algorithms reinforcement learning learning complex stochastic models

In context machine learning examples paper deals problem estimating quality attributes without dependencies among Kira Rendell developed algorithm called RELIEF shown efficient estimating attributes Original RELIEF deal discrete continuous attributes limited twoclass problems In paper RELIEF analysed extended deal noisy incomplete multiclass data sets The extensions verified various artificial one well known realworld problem

引入标签信息

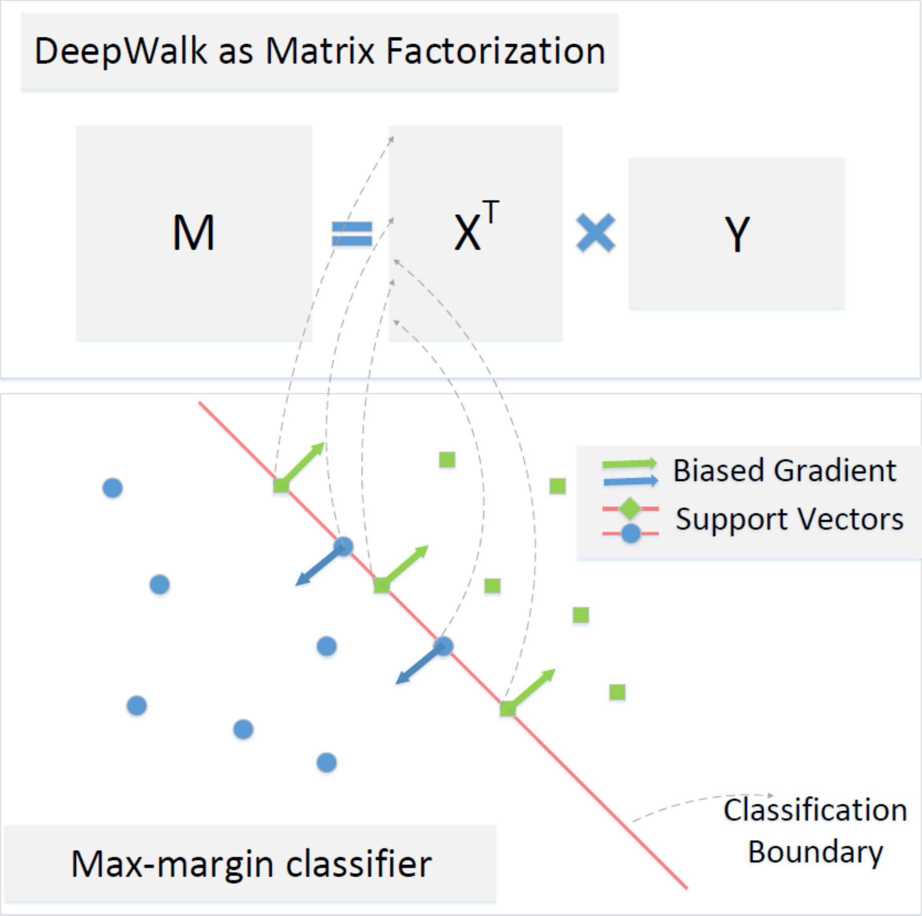
- 真实世界网络节点往往被标注类别标签



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

Max-Margin DeepWalk

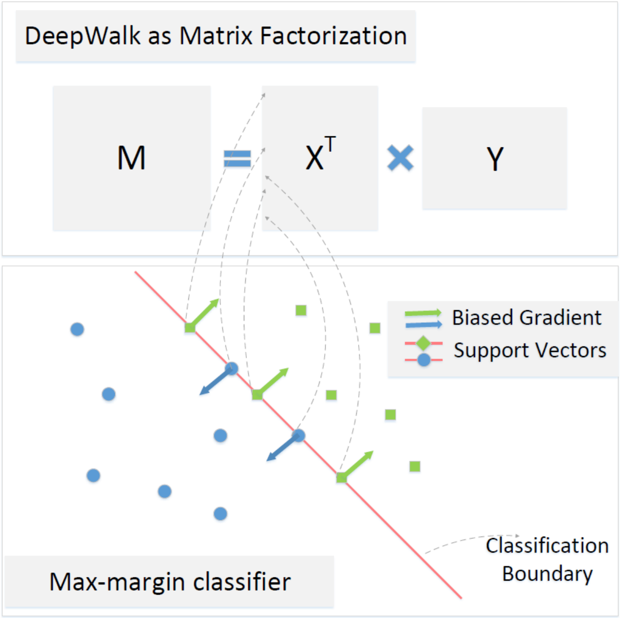
- 共同训练DW+最大间隔分类器



Tu, et al. Max-Margin DeepWalk: Discriminative Learning of Network Representation. IJCAI 2016.

Max-Margin DeepWalk

- 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

Max-Margin DeepWalk

- 节点分类结果
  - >5%的提升
  - 仅用一半训练数据即可达到baseline的分类效果

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

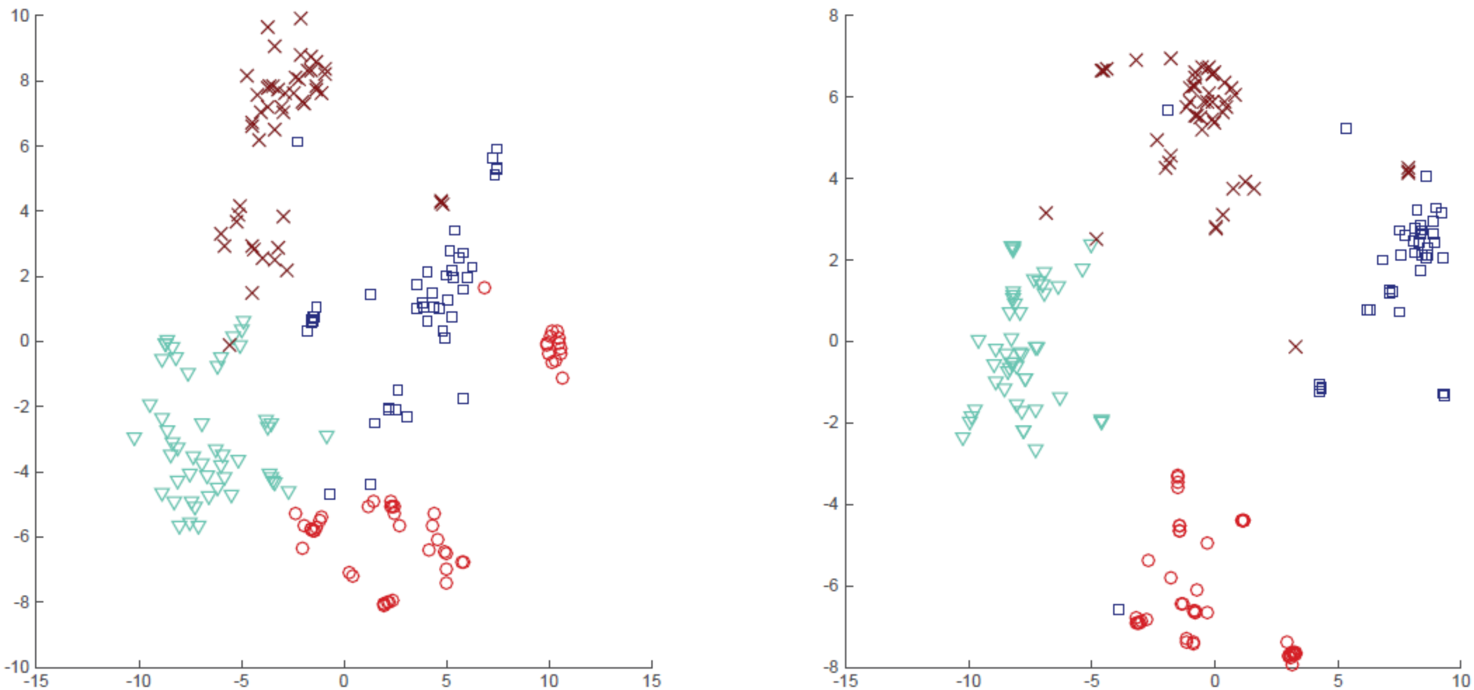
%Labeled Nodes	10%	20%	30%	40%	50%	60%	70%	80%	90%
DW	49.09	55.96	60.65	63.97	65.42	67.29	66.80	66.82	63.91
MFDW	50.54	54.47	57.02	57.19	58.60	59.18	59.17	59.03	55.35
LINE	39.82	46.83	49.02	50.65	53.77	54.2	53.87	54.67	53.82
MMDW($\eta = 10^{-2}$)	**55.60**	60.97	63.18	65.08	**66.93**	**69.52**	**70.47**	**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

Max-Margin DeepWalk

- 节点表示可视化（t-SNE）
  - DeepWalk与MMDW

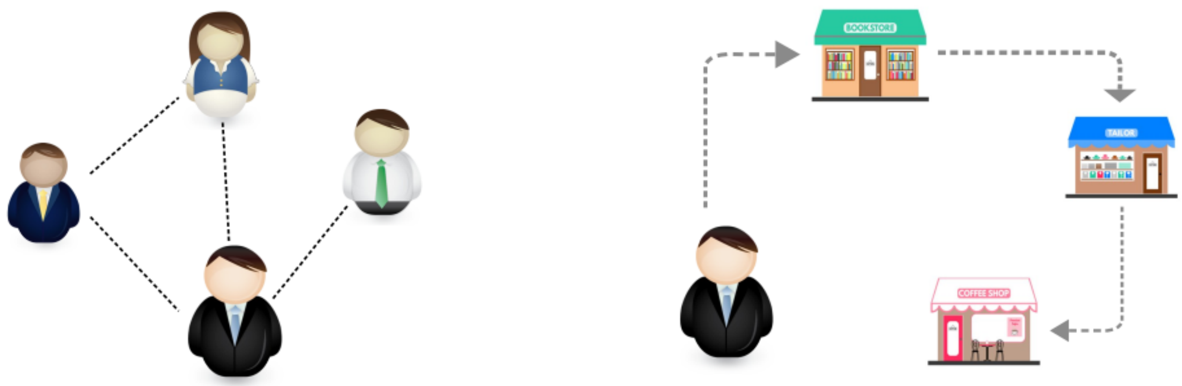


内容提要

- 网络表示学习方案
- 引入外部信息的网络表示学习
- 网络表示学习应用
- 展望

社交网络和用户轨迹的联合神经模型

- 将社交网络和用户的移动轨迹联合建模



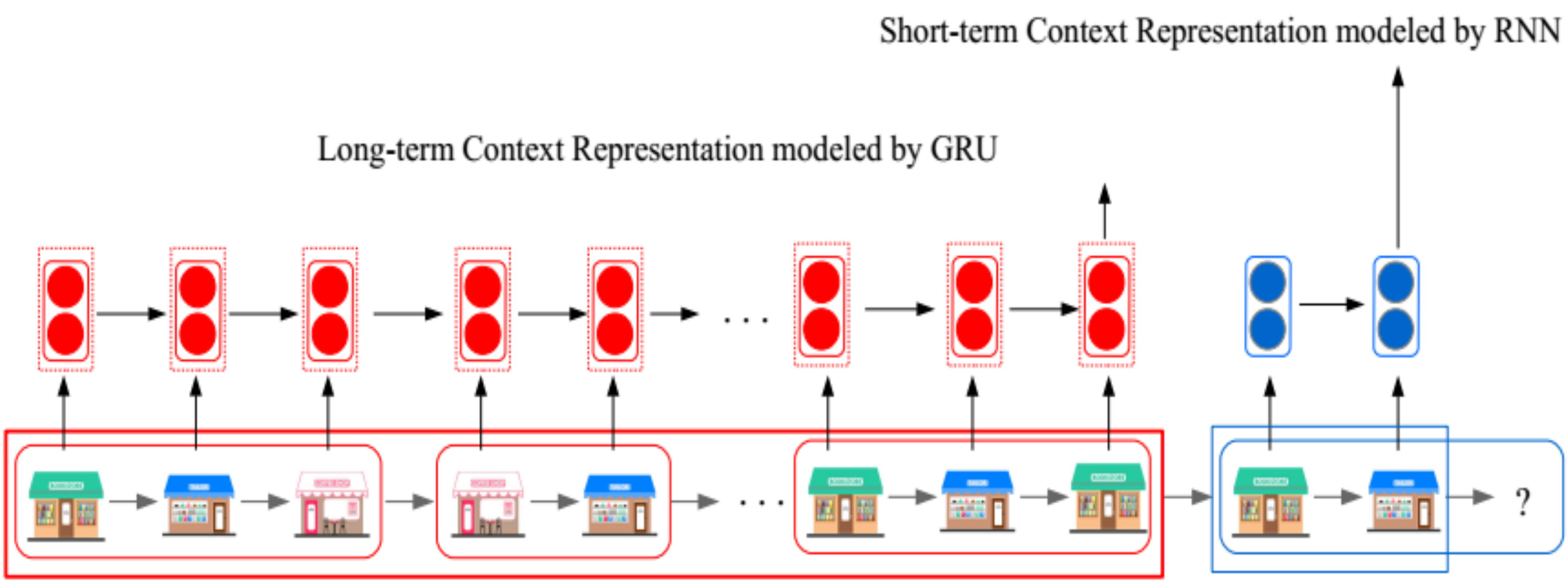
(a) Friendship Network (b) User Trajectory

Fig. 1. An illustrative example for the data in LBSNs: (a) Link connections represent the friendship between users. (b) A trajectory generated by a user is a sequence of chronologically ordered check-in records.

Yang et al. A Neural Network Approach to Jointly Modeling Social Networks and Mobile Trajectories. ACM TOIS.

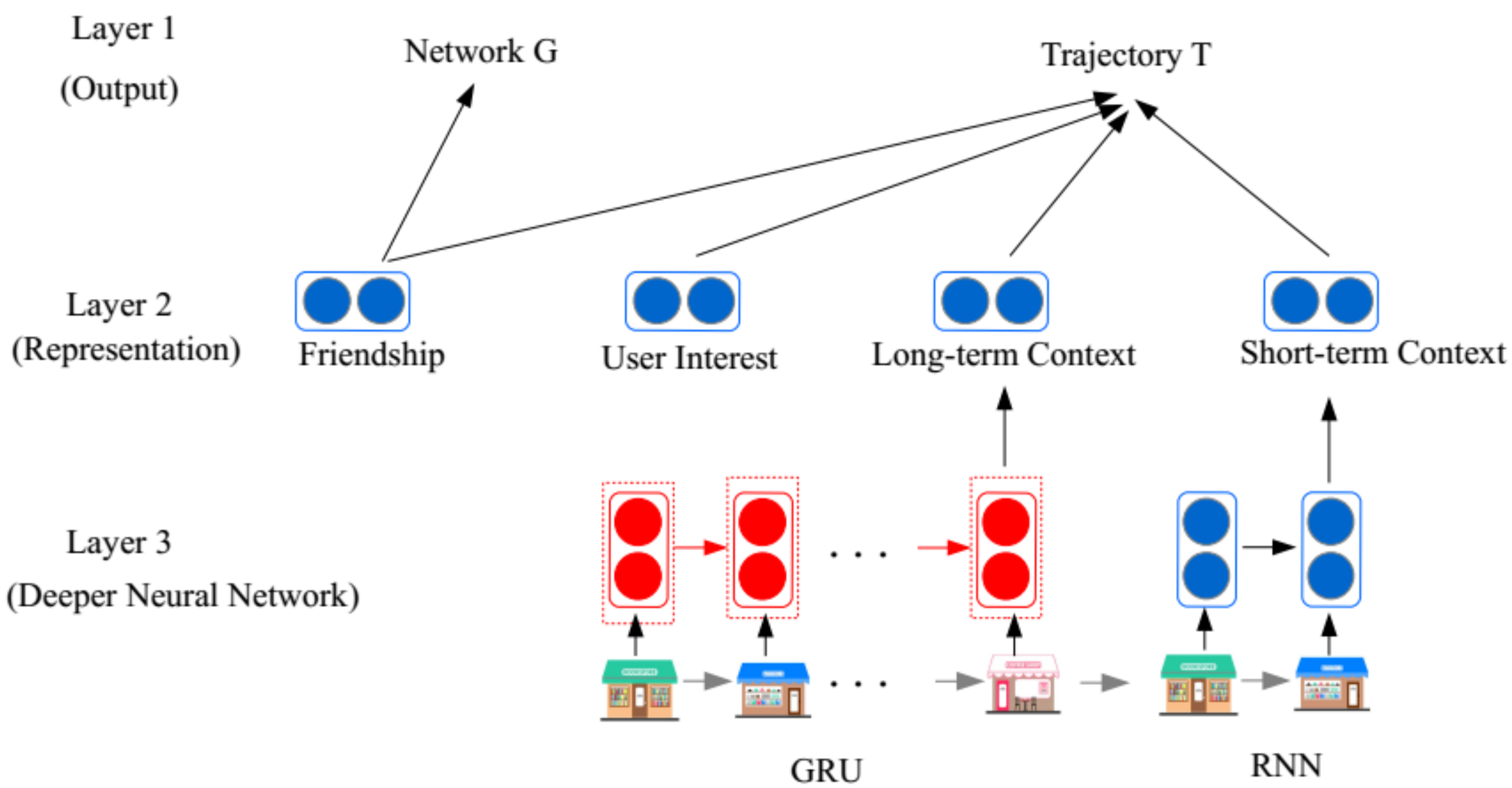
社交网络和用户轨迹的联合神经模型

- 使用循环神经网络对用户轨迹建模



社交网络和用户轨迹的联合神经模型

- 以用户表示为基础的整体神经网络模型



社交网络和用户轨迹的联合神经模型

- 下个地点预测任务实验结果

Dataset	Brightkite			Gowalla		
Metric (%)	R@1	R@5	R@10	R@1	R@5	R@10
PV	18.5	44.3	53.2	9.9	27.8	36.3
FBC	16.7	44.1	54.2	13.3	34.4	42.3
FPMC	20.6	45.6	53.8	10.1	24.9	31.6
PRME	15.4	44.6	53.0	12.2	31.9	38.2
HRM	17.4	46.2	56.4	7.4	26.2	37.0
JNTM	**22.1**	**51.1**	**60.3**	**15.4**	**38.8**	**48.1**

社交网络和用户轨迹的联合神经模型

- 下个新地点预测任务实验结果

Dataset	Brightkite			Gowalla		
Metric (%)	R@1	R@5	R@10	R@1	R@5	R@10
PV	0.5	1.5	2.3	1.0	3.3	5.3
FBC	0.5	1.9	3.0	1.0	3.1	5.1
FPMC	0.8	2.7	4.3	2.0	6.2	9.9
PRME	0.3	1.1	1.9	0.6	2.0	3.3
HRM	1.2	3.5	5.2	1.7	5.3	8.2
JNTM	**1.3**	**3.7**	**5.5**	**2.7**	**8.1**	**12.1**

社交网络和用户轨迹的联合神经模型

- 朋友推荐任务实验结果

Training Ratio	20%		30%		40%		50%	
Metric (%)	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10
DeepWalk	2.3	3.8	3.9	6.7	5.5	9.2	7.4	12.3
PMF	2.1	3.6	2.1	3.7	2.3	3.4	2.3	3.8
PTE	1.5	2.5	3.8	4.7	4.0	6.6	5.1	8.3
TADW	2.2	3.4	3.6	3.9	2.9	4.3	3.2	4.5
JNTM	**3.7**	**6.0**	**5.4**	**8.7**	**6.7**	**11.1**	**8.4**	**13.9**

Training Ratio	20%		30%		40%		50%	
Metric (%)	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10
DeepWalk	2.6	3.9	5.1	8.1	**7.9**	**12.1**	**10.5**	**15.8**
PMF	1.7	2.4	1.8	2.5	1.9	2.7	1.9	3.1
PTE	1.1	1.8	2.3	3.6	3.6	5.6	4.9	7.6
TADW	2.1	3.1	2.6	3.9	3.2	4.7	3.6	5.4
JNTM	**3.8**	**5.5**	**5.9**	**8.9**	**7.9**	11.9	10.0	15.1

内容提要

- 网络表示学习方案
- 引入外部信息的网络表示学习
- 网络表示学习应用
- 展望

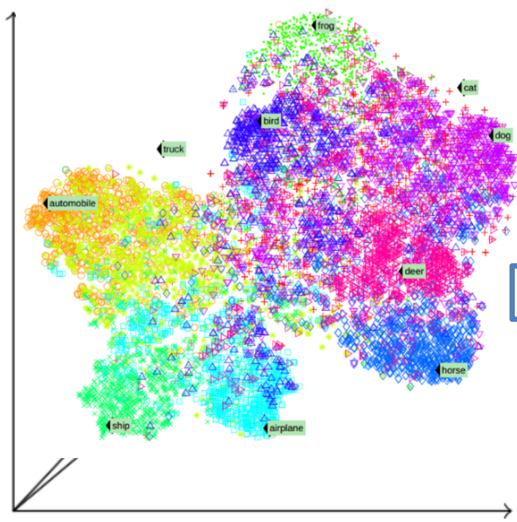
社会计算与表示学习

知识

产品

文本

用户



统一语义空间

个性推荐

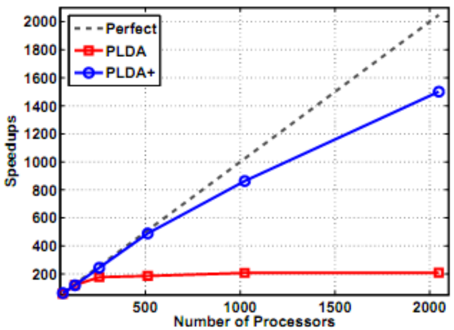
知识计算

社会网络分析

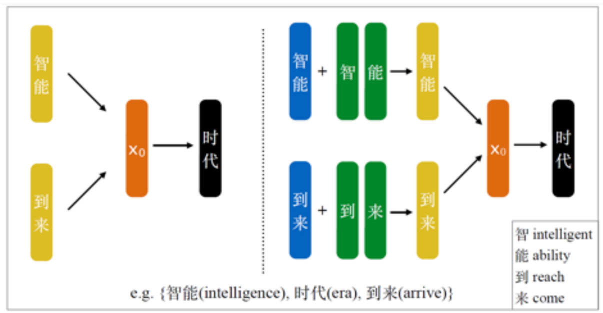
研究趋势

- 探索特殊社会网络表示学习
  - Bipartite Networks, Signed Networks, Hetereogeneous Networks, …
- 探索动态网络下的表示学习
- 改进社会计算典型任务
  - 链接预测，社区发现
  - 影响力分析，传播预测
  - 用户建模，个性推荐
- 知识驱动的社会计算
  - 为社会计算引入推理能力
  - 提高社会计算的可解释性

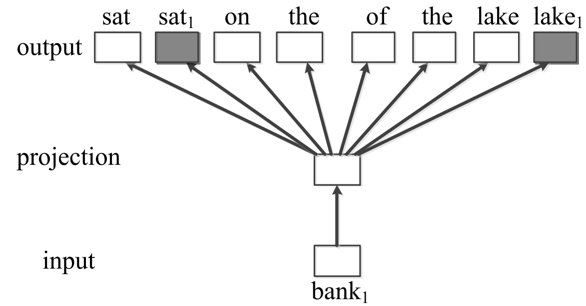
文本表示方法



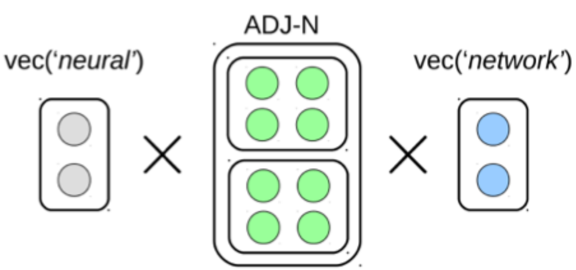
主题模型并行算法PLDA+
(ACM TIST 2011)



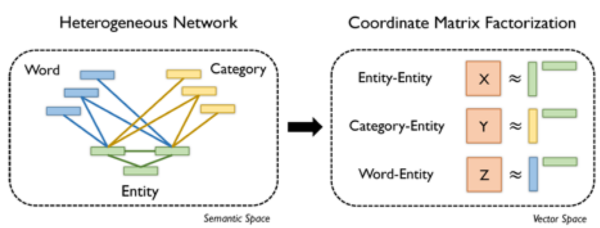
考虑汉字的词汇表示
(IJCAI 2015)



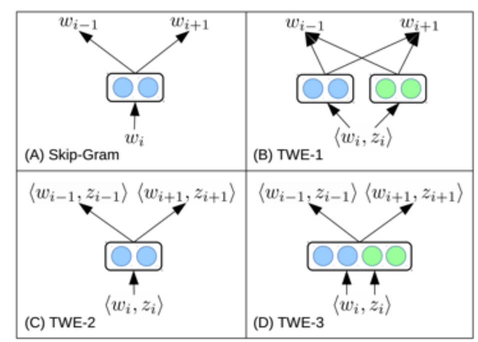
融合语言知识库的词义表示
(EMNLP 2014)



基于张量操作的短语表示
(AAAI 2015)

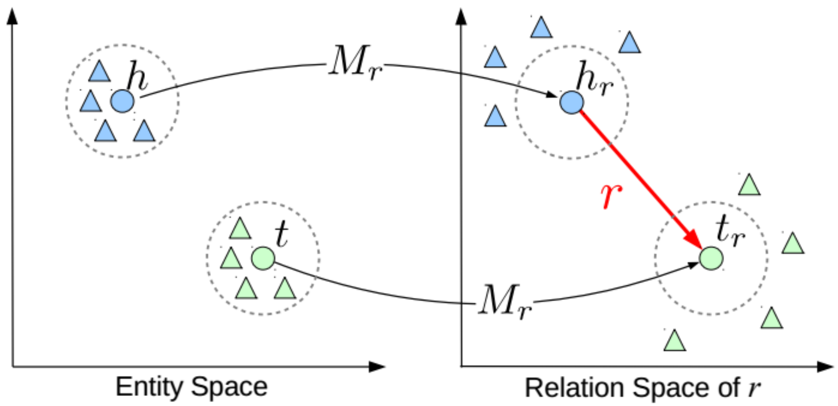


考虑丰富信息的实体表示
(IJCAI 2015)

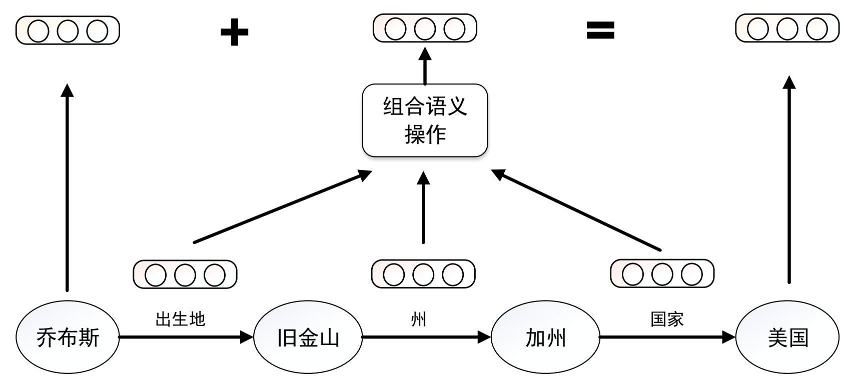


主题增强的文档表示
(IJCAI 2015)

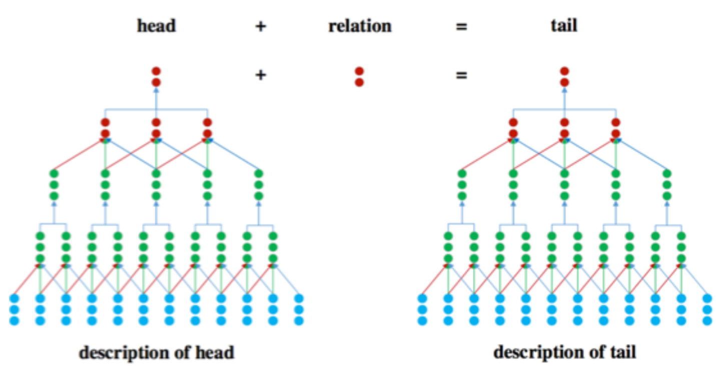
知识表示方法



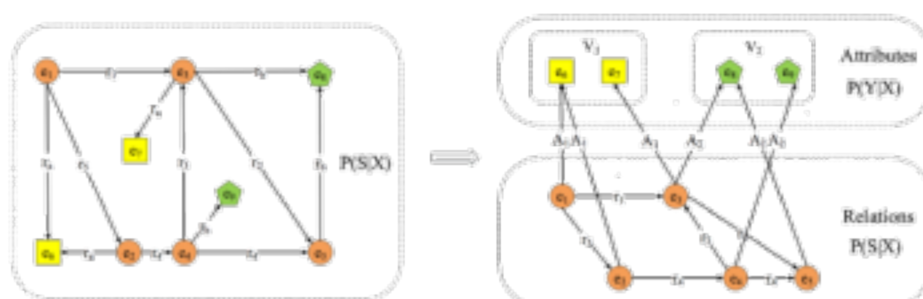
TransR (AAAI 2015)



Path-based TransE
PTransE (EMNLP 2015)



Description-Embodied KRL
DKRL (AAAI 2016)



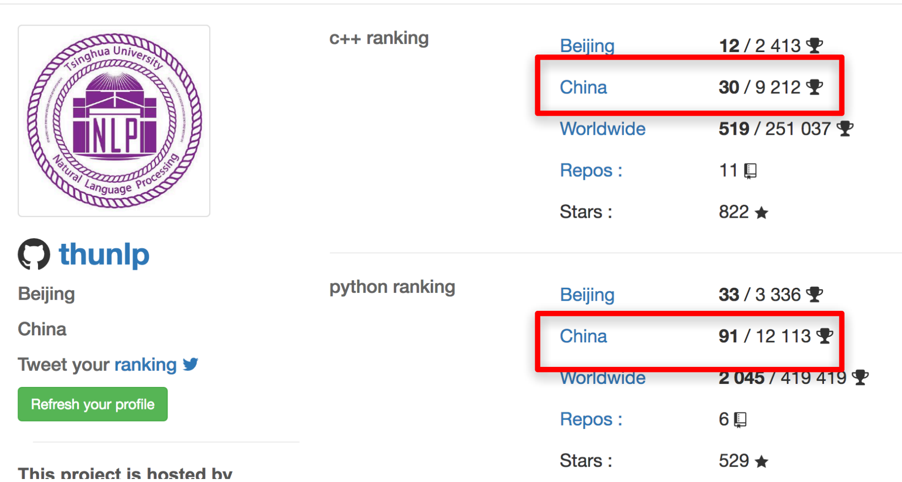
KRL with entities, attributes and relations
KR-EAR (IJCAI 2016)

开源工具

- 在中文分词、文本分类、关键词抽取、表示学习等方面开源数十项软件

https://github.com/thunlp

THULAC： 中文词法分析
THUCTC： 中文文本分类
THUTAG： 关键词抽取与社会标签推荐
KB2E： 知识表示学习
NRE： 神经网络关系抽取
NSC： 神经网络情感分类
MMDW： 最大间隔网络表示学习



Take Home Message

- 分布式表示将研究对象语义信息编码到低维向量空间
- 分布式表示可有效实现社会计算中异质对象语义计算问题
- 分布式表示已被广泛应用于汉字、词汇、词义、实体、短语、句子、文档、网络和知识的表示
- 分布式表示和深度学习技术已在社会计算和计算社会科学中崭露头角，并将发挥更大作用

感谢各位老师同学

http://nlp.csai.tsinghua.edu.cn/~lzy/

liuzy@tsinghua.edu.cn