大连理工大学
信息检索研究室
Information Retrieval Laboratory of DUT

基于表示学习的知识图谱链路预测算法研究

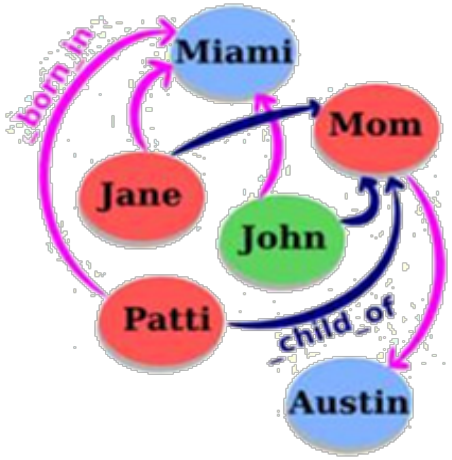
孙媛媛

大连理工大学

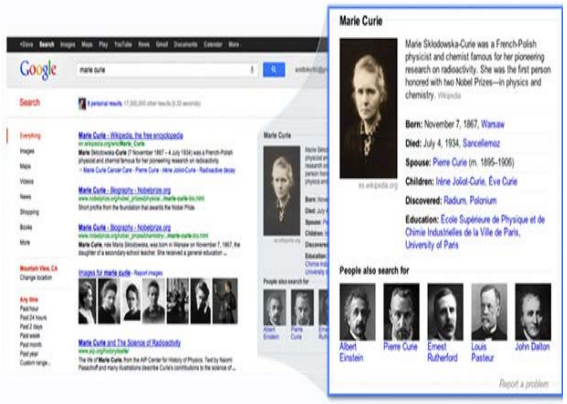
1 背景介绍
2 算法描述
3 结果讨论
4 总结展望

1 背景介绍

I. 知识图谱



- 1963年，西蒙（R.F.Simon）等提出语义网络
- 2012年，谷歌第一次使用“知识图谱”概念
- (实体，关系，实体)的三元组结构



谷歌搜索

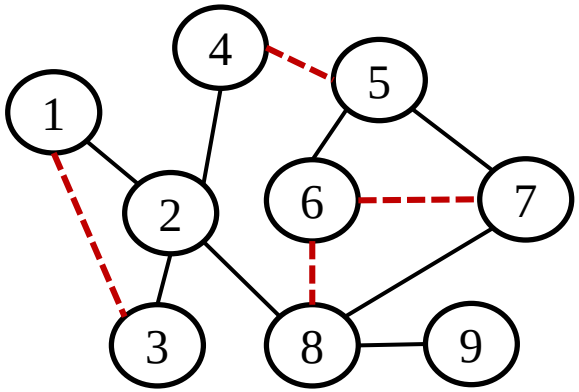
Facebook graph search

百度知心

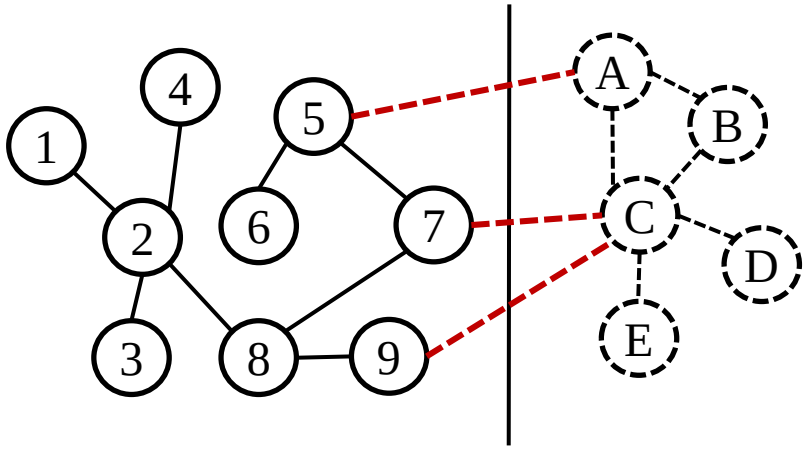
I. 链路预测

链路预测[1]：通过已知的网络节点及网络结构等信息，预测网络中尚未产生连边的两个节点之间产生连接的可能性

缺失边预测

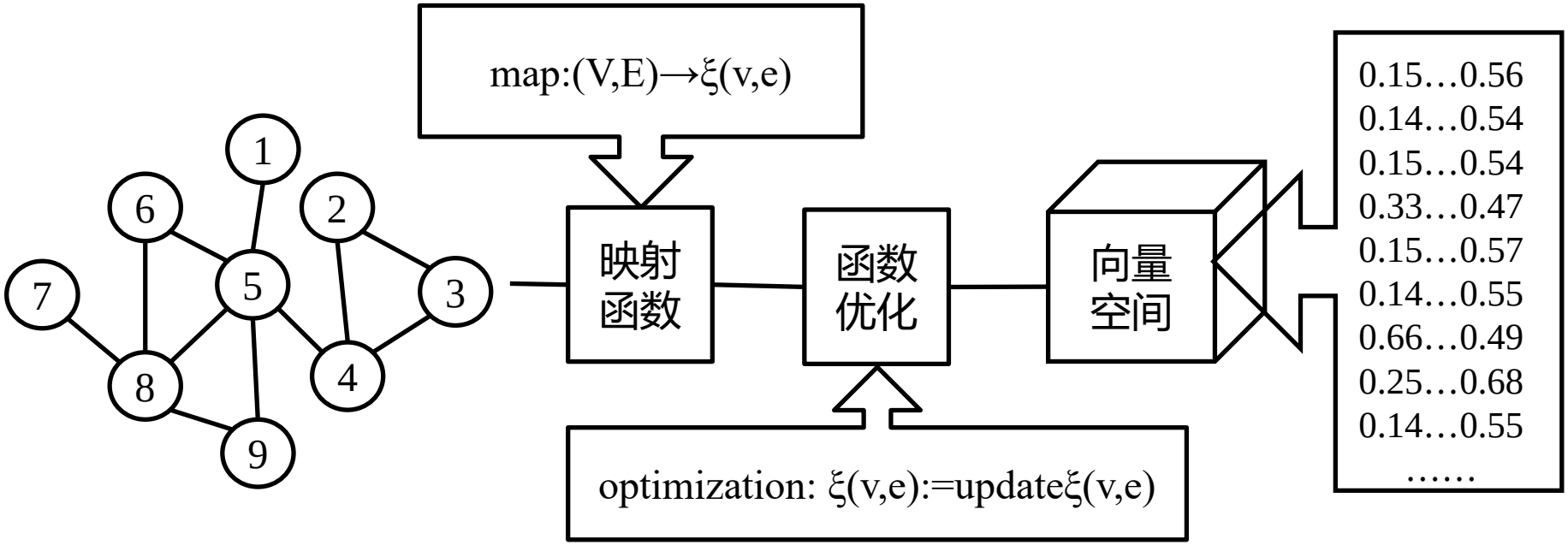


未来边预测



[1] 吕琳媛, 周涛. 链路预测[M]. 高等教育出版社, 2013.

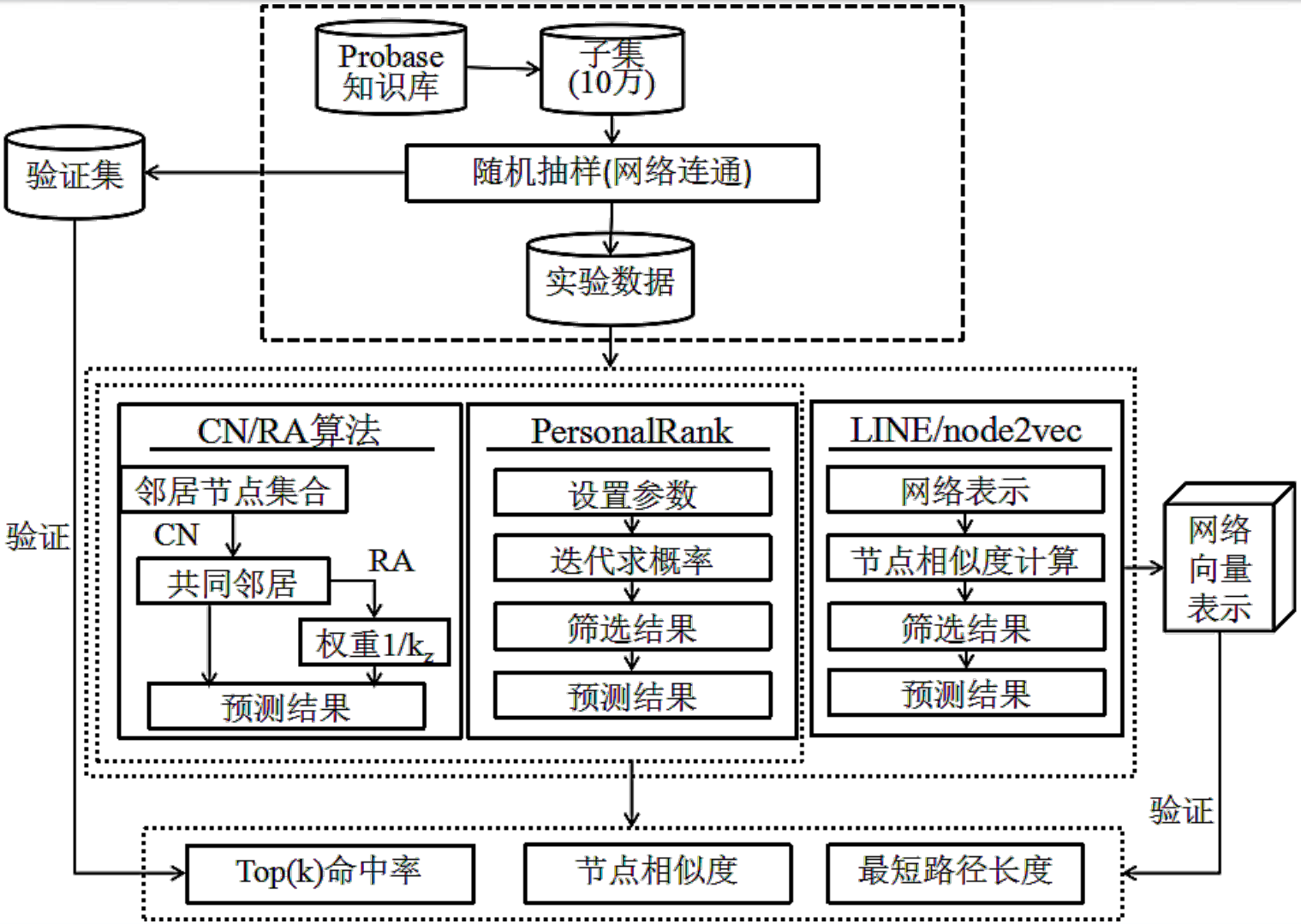
I. 表示学习



- 将高维的数据空间通过某种形式的映射函数表示成低维连续的向量空间
- 将高维度网络空间中的节点以n维的连续向量形式表示，并保证高维空间中相似节点在低维向量空间中依然相近

2 算法描述

II. 算法框架


Probase
知识库
子集
(10万)
随机抽样(网络连通)
验证集
实验数据
CN/RA算法
邻居节点集合
CN
共同邻居
RA
权重1/k_z
预测结果
PersonalRank
设置参数
迭代求概率
筛选结果
预测结果
LINE/node2vec
网络表示
节点相似度计算
筛选结果
预测结果
网络
向量
表示
验证
验证
Top(k)命中率
节点相似度
最短路径长度

II. 数据集

Probase知识库

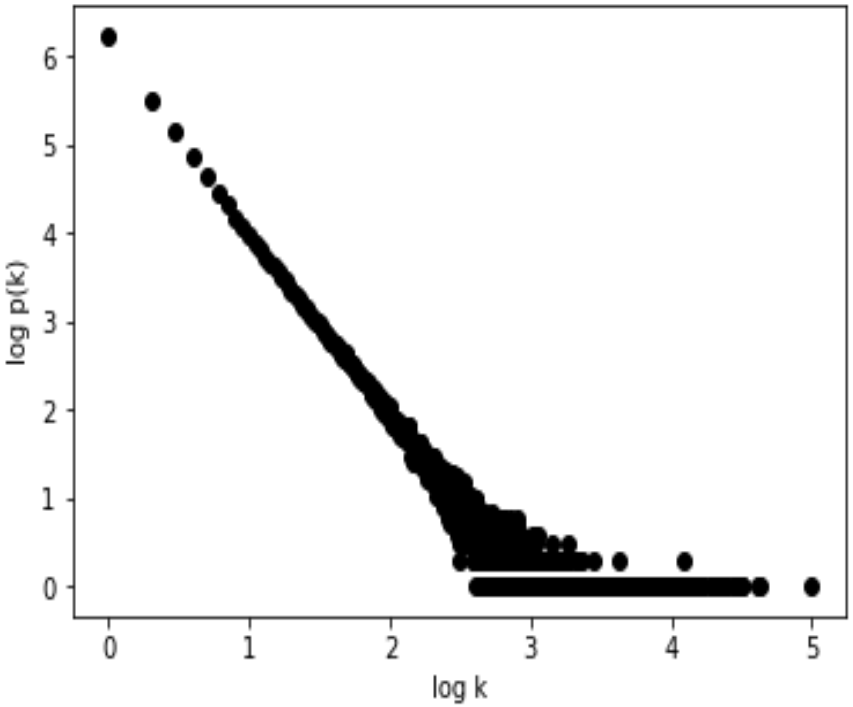
- 建立时间：2010年10月29日
- 数据来源：Probase知识库是一个有关概念的图，知识库里的数据是来自数十亿的网页和多年的搜索日志
- 版本：共有5,376,526个概念实体，12,501,527个实例实体，85,101,174条IsA关系
- 数据格式：[concept，instance，IsA]的三元组格式

实验数据集

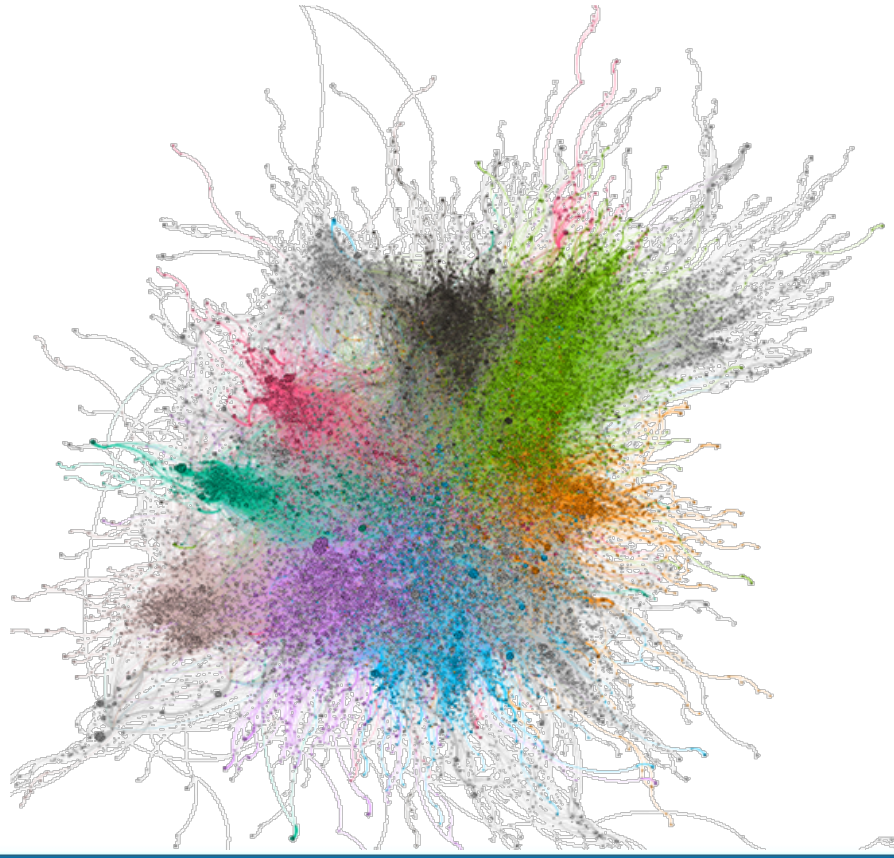
- Probase知识库中ISA关系边权重最大的前十万条边
- 节点数量为42,693，边数量为100,000

II. 网络分析

- ISA关系权重最大的前十万条边的度分布
  - ✓ 幂律分布



- ISA关系权重最大的前十万条边的模块可视化

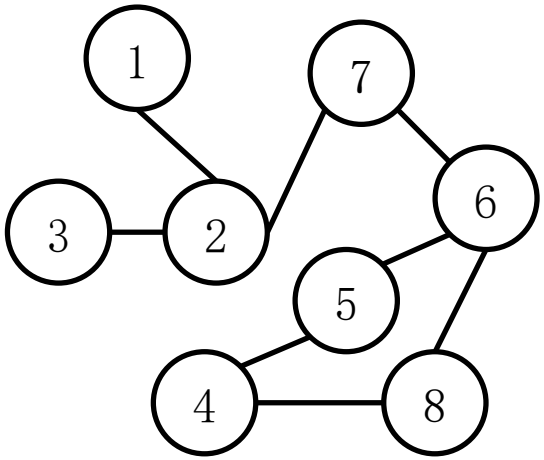


II. 模块分析

Mudul	Nodes	Edges	DC	CC	BC	CCO	Diameter
5	6549	16000	0.334573	0.450937	0.499314	0.007679	9
321	4049	9602	0.173568	0.388674	0.470308	0.023162	12
40	3603	6875	0.199773	0.397584	0.469949	0.012153	12
25	3254	6077	0.214477	0.311376	0.375954	0.006129	14
47	2599	6632	0.167911	0.360053	0.287491	0.027142	11
183	2481	5532	0.192308	0.318618	0.367687	0.008674	12
20	1898	2445	0.288878	0.319143	0.603154	0.001785	13

- Probase知识库是一个稀疏网络
- Probase知识库是一个近似二分图网络
- Probase知识库在较短路径内保持连通

II. CN/RA 算法[1]



- 共同邻居(CN)算法：若网络中的一个节点为v_x，则定义它的邻居集合为$\Gamma(x)$,那么节点v_x和节点v_y的相似性就可以由其共同邻居数决定

$$s_{xy} = \left|\Gamma(x) \cap \Gamma(y)\right| \qquad (1)$$

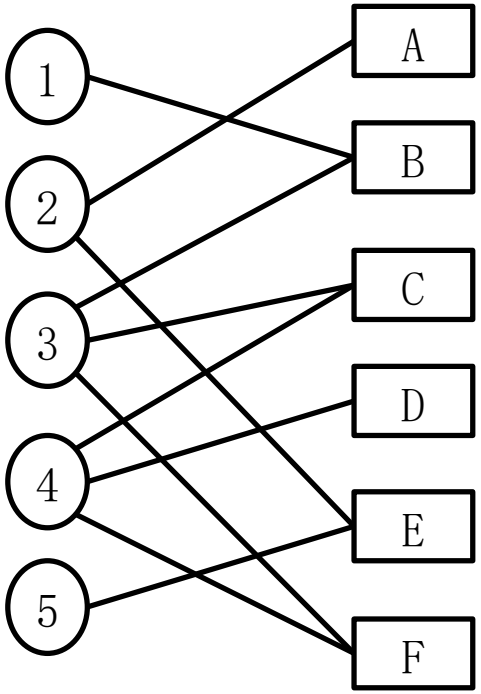
- 资源分配(RA)算法：假设网络中某个节点v_x有一个资源可以平均的分配给它的邻居v_y，那么v_y可以获得的资源数就被定义为v_x和v_y的相似度

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \qquad (2)$$

[1] 吕琳媛, 周涛. 链路预测[M]. 高等教育出版社, 2013.

II. PersonalRank 算法

- 基于二部图的网络推荐算法，是对PageRank算法的改进

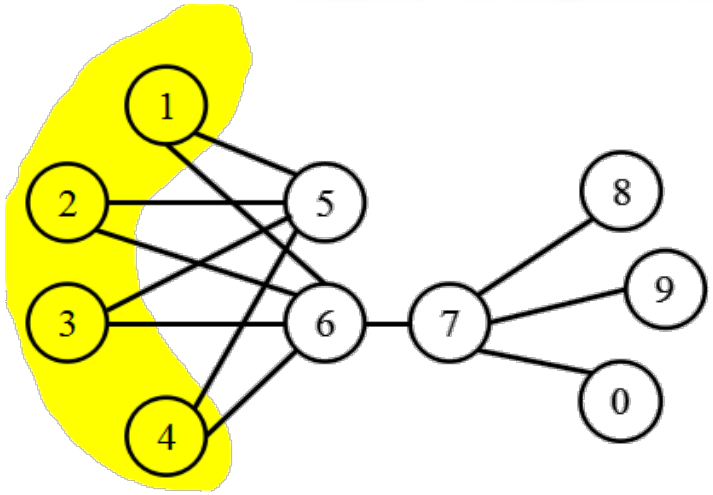


PageRank：
$$PR(p_i) = \alpha \sum_{p_j \in S_{p_i}} \frac{PR(p_j)}{L(p_j)} + \frac{(1-\alpha)}{N} \quad (3)$$

PersonalRank：
$$PerR(i) = (1-d)r_i + d \sum_{j \in in(i)} \frac{PR(j)}{|out(i)|} \quad (4)$$

加权无向图：
$$PerR(i) = (1-d)r_i + d \sum_{j \in \deg ree(i)} \frac{PR(j) \times Weighted(j)}{|degree(i)|} \quad (5)$$

II. LINE 算法[2]



- 一阶序列关系：有直接边相连（节点6、7）
- 二阶序列关系：共享共同邻居（节点5、6）

一阶序列

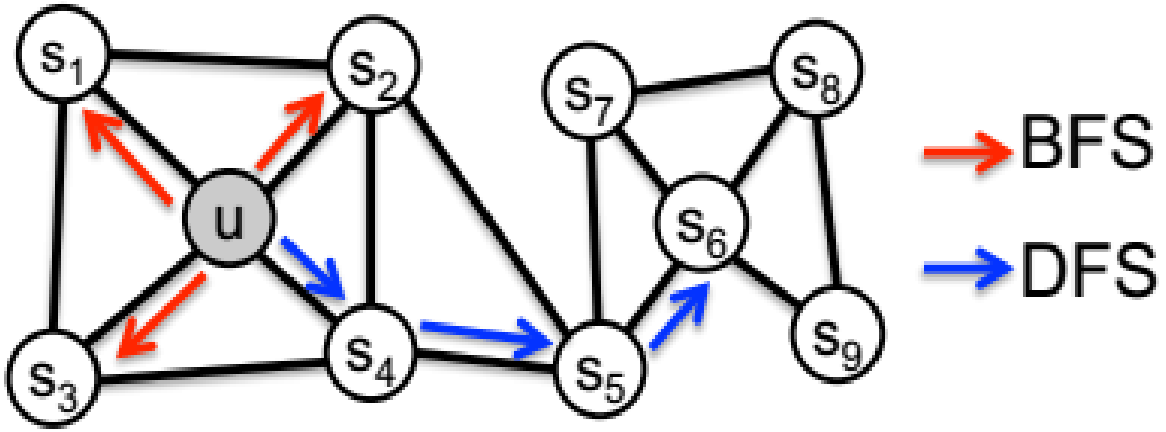
$$p_1(v_i, v_j) = \frac{1}{1+\exp\left(-\vec{u}_i^{T} \cdot \vec{u}_j\right)} \qquad O_1 = -\sum_{(i,j)\in E} w_{ij} \log p_1(v_i, v_j)$$

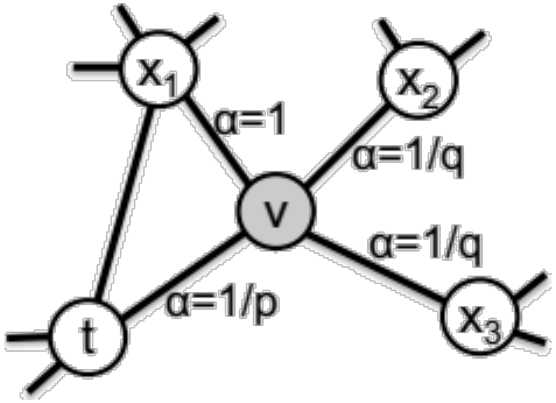
二阶序列

$$p_2(v_j \mid v_i) = \frac{\exp\left(\vec{t}_j^{\,T} \cdot \vec{u}_i\right)}{\sum_{k=1}^{|V|} \exp\left(-\vec{t}_k^{\,T} \cdot \vec{u}_i\right)} \qquad O_2 = -\sum_{(i,j)\in E} w_{ij} \log p_2(v_j \mid v_j)$$

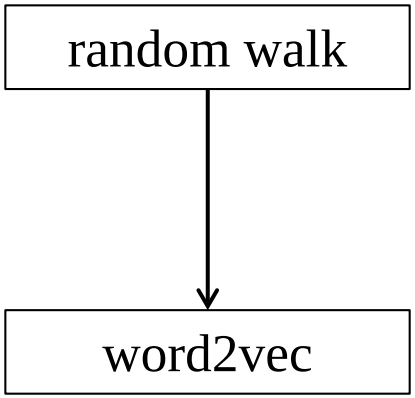
[2] Tang J, Qu M, Wang M, et al. LINE:Large-scale Information Network Embedding[J]. 2015, 2(2):1067-1077.

II. node2vec 算法[3]





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



[3] Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:855-864.

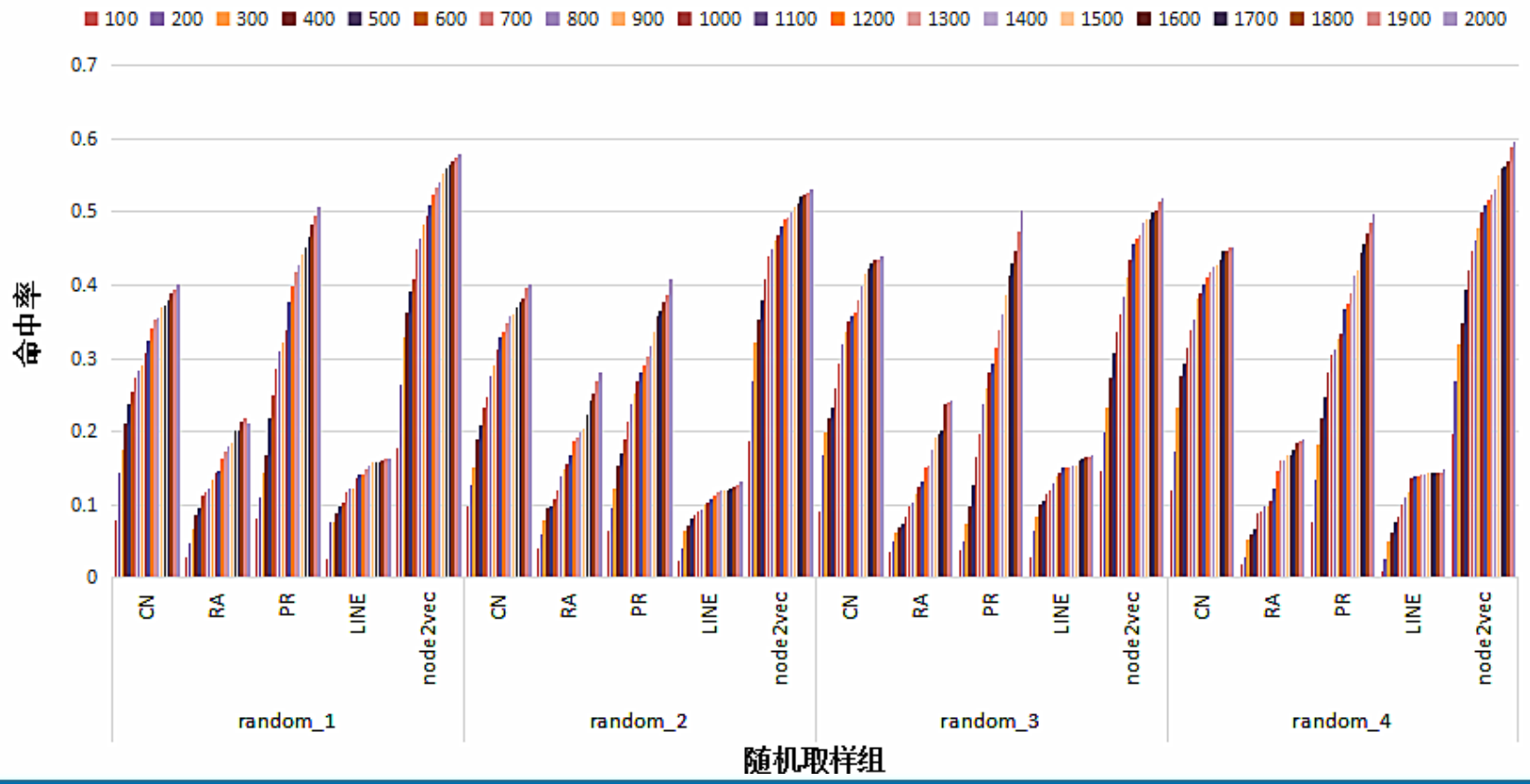
3 结果分析

III. 实验设置

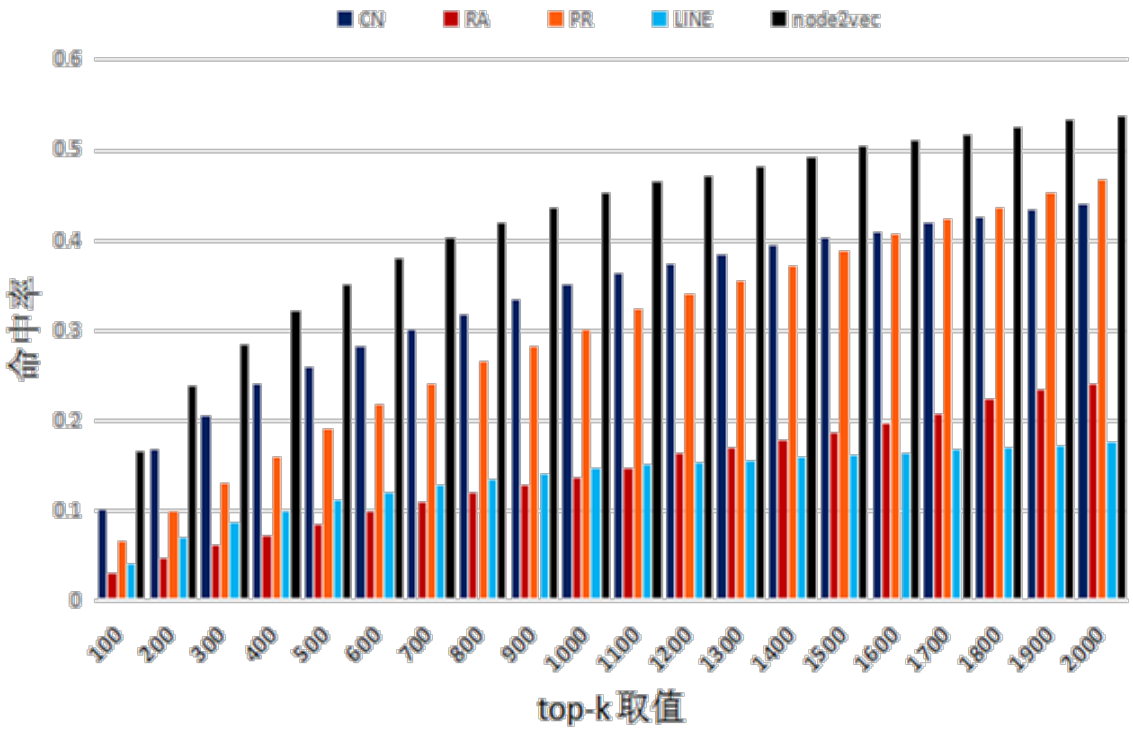
- 预处理：对Probase知识库中的实体赋以ID，形成[ID,ID,ISA]的数据形式
- 数据集划分：7:3 （70%的数据用于训练节点的低维向量，30%的数据用于结果验证），随机生成8组数据集
- 参数设置：所有用于实验的程序均选择其默认参数
- 初始节点选择：每组数据集中均随机选择10个节点作为初始节点
- 结果验证：Top-k命中率、节点相似度和最短路径长度

III. Top-K命中率

$$Hit_{top-k} = \frac{M_{result\_set}}{N_{\text{node}\_neighbors}} \qquad (6)$$



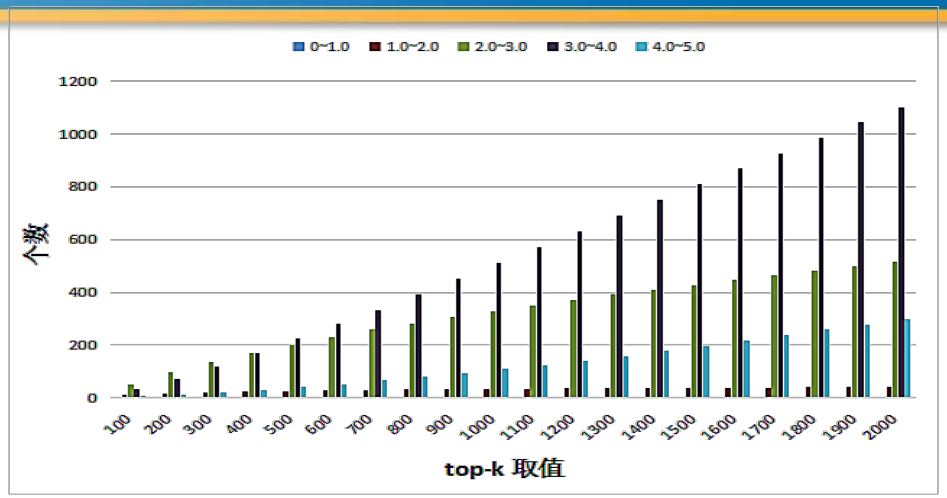
III. Top-K命中率



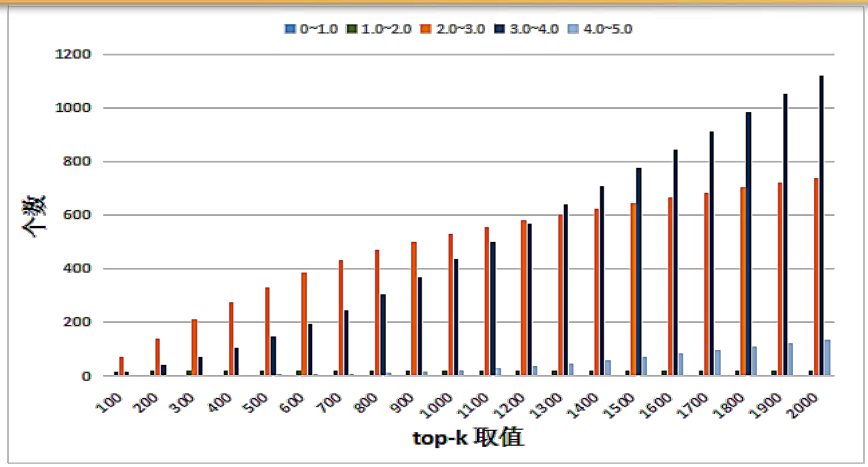
- 预测结果呈log指数增长
- CN算法优于RA算法
- LINE算法表现不好

- 考虑共同邻居对于稀疏的Probase知识库的预测结果没有促进作用
- 考虑路径（序列）信息对预测结果是有效的

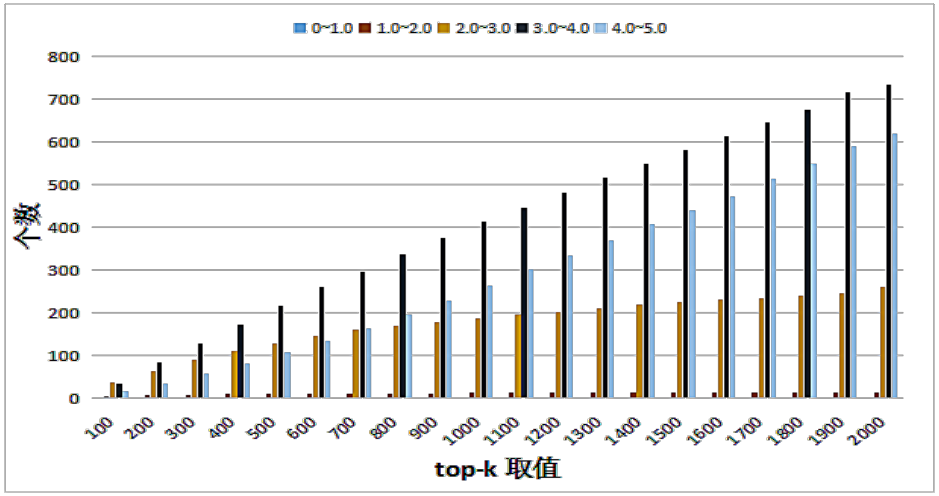
III. 节点相似度



PersonalRank 算法



node2vec 算法



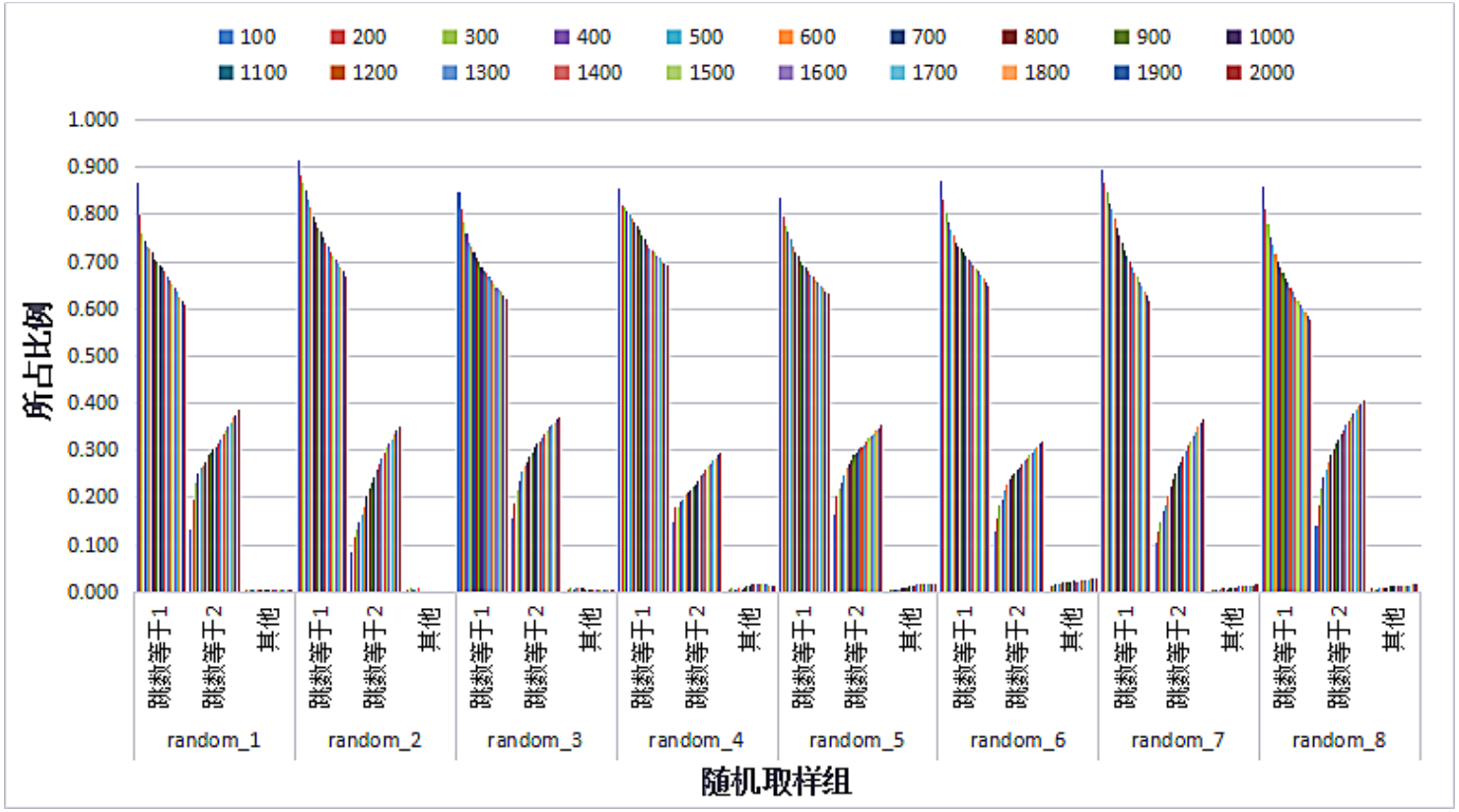
LINE 算法

- node2vec算法的预测结果与给定节点的相似度最好，LINE算法的最不好
- LINE算法没有考虑节点的路径（序列）信息

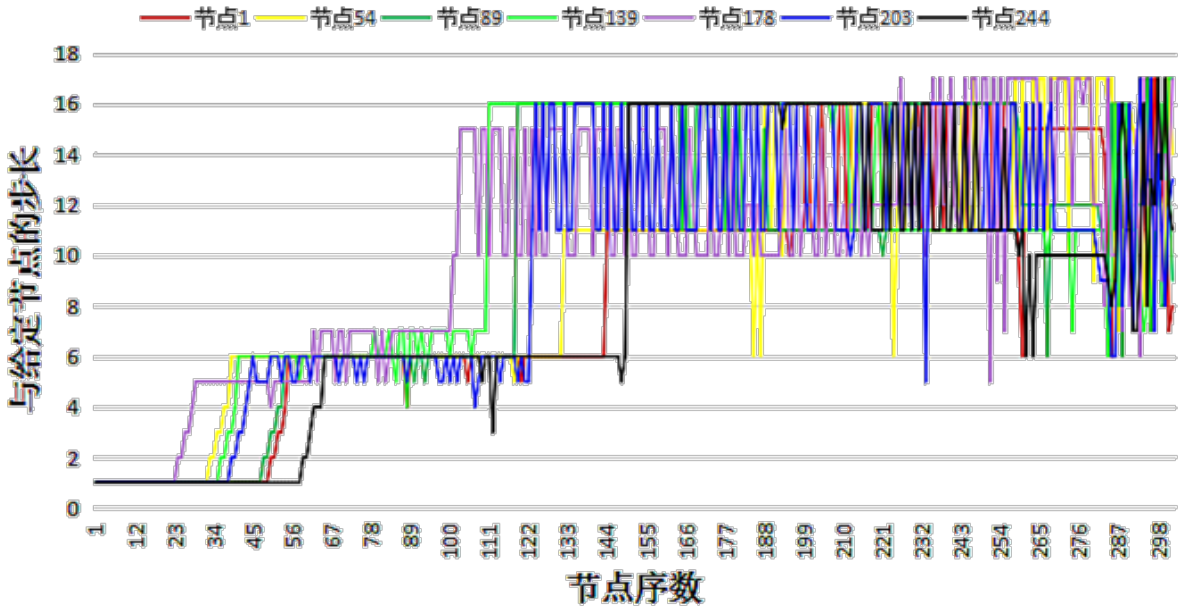
III. 节点相似度

	LINE	PersonalRank	node2vec
基于“google”的前十组推荐结果	ibm	**yahoo**	**bing**
	yahoo	msn	msn
	nike	bing	amazon
	coca cola	component	**Internet company**
	samsung	ibm	ibm
	sony	**apple**	**microsoft**
	motorola	**facebook**	**baidu**
	toyota	altavista	giant
	boeing	amazon	tech company
	general electric	part	large company

IV. 最短路径长度


100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
1.000
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
所占比例
跳数等于1
跳数等于2
其他
random_1
random_2
random_3
random_4
random_5
random_6
random_7
random_8
随机取样组

IV. 最短路径长度


节点1
节点54
节点89
节点139
节点178
节点203
节点244
与给定节点的步长
节点序数
0
2
4
6
8
10
12
14
16
18
1
12
23
34
45
56
67
78
89
100
111
122
133
144
155
166
177
188
199
210
221
232
243
254
265
276
287
298

4 总结展望

总结

- 引入表示学习方法将以共享邻居数或资源分配数为指标的预测方式转变成以向量相似度为指标的预测方式，对大规模网络预测有促进作用
- 表示学习对预测结果的命中率、稳定性以及预测节点与已知节点之间的相似度都有提升，证明了表示学习算法在链路预测任务中的有效性

展望

- 改进表示学习算法，减少知识图谱由高维空间向低维空间映射过程中的信息损失，使算法更适合于知识图谱的向量表示，从而进一步提高基于知识图谱的链路预测结果
- 进一步探索基于知识图谱的应用研究，寻找知识图谱应用研究与复杂网络领域知识的结合点，开展契合度更高的跨领域跨学科的应用研究

谢谢！