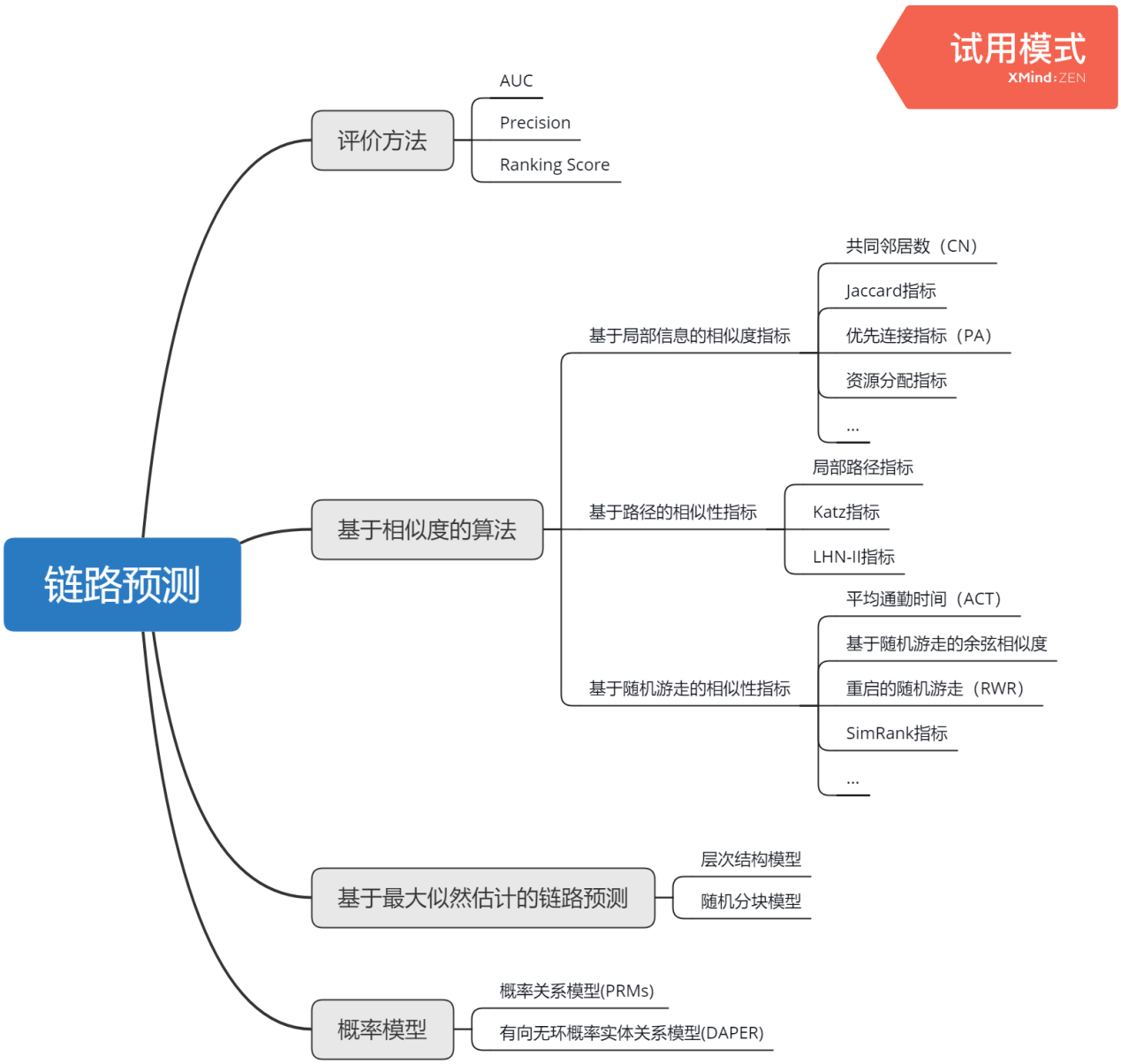
链路预测

- 秦晋琦 2018103664 2019年6月19日

问题背景

网络中的链路预测是**数据挖掘**的一种方法，它是通过已知的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。

预测已经存在但尚未被发现的连接实际上是一种数据挖掘，发现的过程，而对于未来可能出现的连边的预测则与网络演化相关。有关链路预测的研究方法总结如下：



链路预测的一些应用场景：

- 生物领域的蛋白质相互作用网络，新陈代新网络

  酵母菌蛋白质之间80%的相互作用不为人所知，相互作用关系（网络的连边）是需要大量的实验才能进行推断的；优化；
- 解决社会网络分析中数据不全的问题
- 分析演化网络：社交网络中推测用户潜在关系
- 通过部分标记的网络预测剩余节点的标签

数据集介绍（放后面）

我们采用的数据集是基于高能物理引文网构建的作者合作关系网络，具体说明如下：

引文网给出了一段时间内该领域发表的论文信息（包括名称及作者信息等），网络本身描述了论文间相互引用的关系，是一个有向的网络。我们要构建的是**描述这些作者之间合作关系的网络**，即如果两个作者合写过论文，则用一条边连接起来。该网络的构建通过引文网提供的节点信息：每篇论文的作者之间构建连边。

High Energy Physics - Theory collaboration network

Dataset information

Arxiv HEP-TH (High Energy Physics - Theory) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to High Energy Physics - Theory category. If an author *i* co-authored a paper with author *j*, the graph contains a undirected edge from *i* to *j*. If the paper is co-authored by *k* authors this generates a completely connected (sub)graph on *k* nodes.

The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP-TH section.

- 数据来源：High-energy physics theory citation network from SNAP
- 网络构建代码

网络常见参数描述：

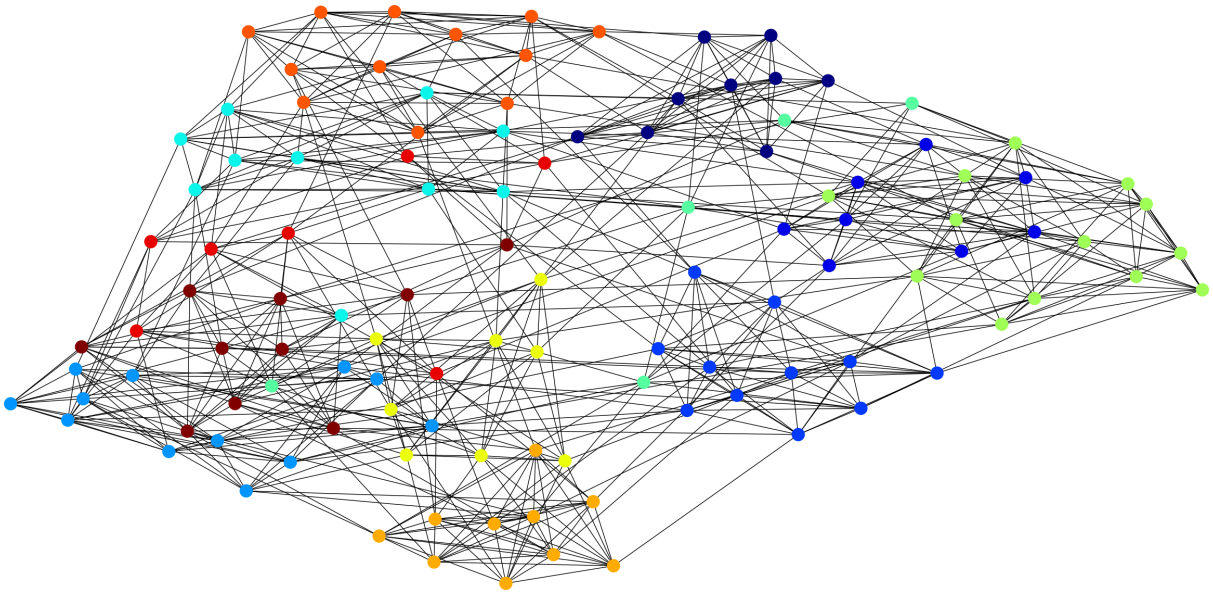
（表格）

链路预测方法简介

问题描述及评价方法

1、问题描述

定义$G(V,E)$为一个无向网络，其中V为节点集，E为边集合，网络总节点数为N，边数为M，节点对全集为U，给定一种链路预测方法，对每对没有连边的节点对(x,y)赋予一个分数$S_{x,y}$，然后将所有未连边的节点对根据分数进行排序，排在最前面的节点对出现连边的概率最大。



注：美国大学生足球俱乐部网络

2、评价方法

为了测试算法的准确性，将已知连边集E分成训练集E^T和测试集E^P两部分，测试集提供了正确的标签，因此可以采用常用的监督学习的评价指标进行测试。

- AUC

$$AUC = \frac{n' + 0.5n''}{n}$$

- Precision

$$Precision = \frac{m}{L}$$

基于相似性的链路预测

介绍

刻画节点相似性的方法有多种，最简单的就是利用节点的属性，比如人的年龄、性别、职业等。另一种思路是利用网络的结构信息，称为结构相似性。基于结构相似性的链路预测精度取决于该结构相似性的定义能否很好的抓住目标网络的结构特征。比如基于共同邻居的相似度在一些集聚系数较高的网络中表现非常好，甚至好于更复杂的算法。

方法

1、基于局部信息的相似性指标

- 共同邻居(CN)

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)|$$

共同邻居方法通过两个节点所拥有的共同邻居数来代表节点的相似性

- **Jaccard**指标

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

其中$\Gamma(x)$代表x的邻居集，y同理。

- **Adamic-Adar** 指标(AA)

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log k_z}$$

不是采用对公共邻居点简单计数的方法，而是对连接较少的近邻点赋予较大的权重。

- **Resource-Allocation** 指标(RA，资源分配指标)

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}$$

- **preferential attachment**指标（PA）

$$S_{xy} = k(x) * k(y)$$

只考虑两节点度的影响。

对比

表3　10种基于节点局部信息的相似性在6个网络链路预测中的精度比较

Index	PPI	NS	Grid	PB	INT	USAir
CN	0.889	0.933	0.590	0.925	0.559	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	0.933	0.590	0.882	0.559	0.901
Sorensen	0.888	0.933	0.290	0.881	0.559	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	0.933	0.590	0.877	0.559	0.895
LHN-I	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	0.590	0.922	0.559	0.925
RA	0.890	0.933	0.590	0.931	0.559	0.955

不同的相似度算法在不同结构属性的网络上效果各不相同。

基于football数据集，按照4：1的比例划分训练集和测试集，对比这几种方法的Precision：

Methods	Precision
CN	0.374
Jaccard	0.203
AA	0.407
RA	0.398

可以看出，不同的方法效果存在差异，这是因为特定的网络结构对不同方法的契合度不同。可以看出，AA和RA这两个指标对于描述这个网络的结构较为贴切。

2、基于路径的相似性指标

- Katz指标

  Katz指标考虑的是所有的路径数，且对于短路径赋予较大的权重，而长路径赋予较小的权重，它的定义为：

  $$S = \beta A + \beta^2 A^2 + \beta^3 A^3 + \ldots = (I - \beta A)^{-1} - I$$

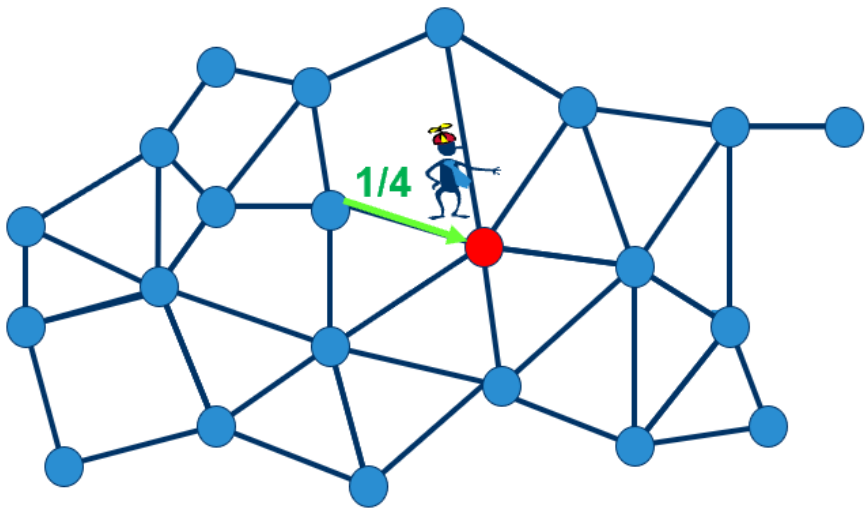
  其中β为权重衰减因子，为了保证数列的收敛性，β的取值必须小于邻接矩阵A最大特征值的倒数。

  这里我们取β为0.01.

  相比前面的局域方法，全域方法针对每个节点的运算都有较高的复杂度，因此计算代价较大。

3、基于随机游走的相似性指标

- RWR(random walk with restart)



重启的随机游走，简称RWR，该指标可以看成是网页排序算法(PageRank)的拓展。其假设随机游走粒子每走一步都以一定概率返回初始位置，设粒子返回概率为$1-c$，P为网络的马尔可夫转移矩阵（这里考虑等概的随机游走，如上图所示），初始时刻粒子在节点x处，则$t+1$时刻该粒子到达网络各个节点的概率向量为：

$$q_x(t+1) = cP^T q_x(t) + (1-c)e_x$$

其中e_x表示初始状态。解得粒子到达各个节点的概率的稳态解为：

$$q_x = (1-c)(I - cP^T)^{-1} e_x$$

其中q_{xy}表示从节点x出发的粒子最终以多少的概率走到节点y，由此定义RWR相似度为：

$$S_{xy}^{RWR} = q_{xy} + q_{yx}$$

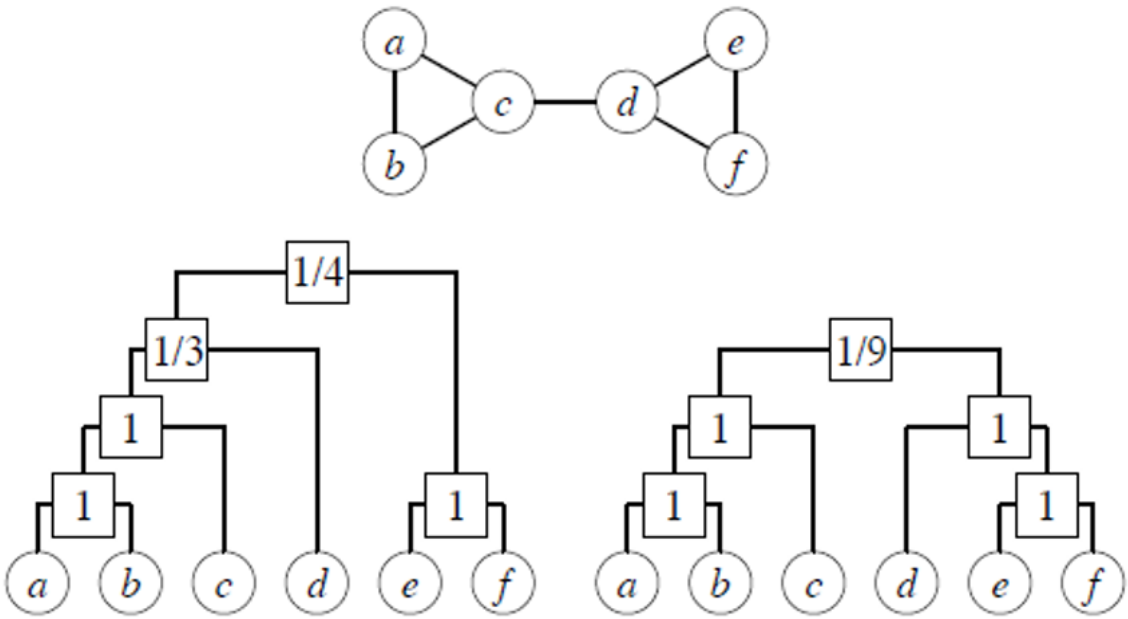
对比

Methods	Precision
Katz	0.423
RWR	0.455

这两个指标都是**全域指标**，考察的是网络整体的结构信息，而不是节点邻居等局部信息，虽然计算复杂度提高，但相比较而言，可以更好的利用网络的结构信息，因此也达到了更好的效果。

基于最大似然估计的链路指标

实际网络在很多情况下具有一定的分层结构，我们可以用树状图来表示。



给定一个网络G和对应的一个树状图D，D对G的似然估计值为：

$$L(D,\{p_r\}) = \prod_r p_r^{E_r}(1-p_r)^{L_r R_r - E_r}$$

使得似然估计最大的P_r为：$p_r^* = E_r/(L_r R_r)$.按此求得每一个树状图的似然值，然后按概率对树状图抽样，并以要预测的链路出现的平均概率作为最终的预测概率。

这种方法对分层结构明显的网络预测效果较好，但复杂度极高，不适合大规模网络的链路预测。

（可参照ppt）

概率模型（选）

前沿方向：

比如CNN和图模型的结合等等。。

总结

参考文献

[1] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651-661.