|  |  |
| --- | --- |
| 学号 ： | 2016317200210 |
| 密级 ： |  |

华中农业大学本科毕业论文

基于信息扩散的异质信息网络链路预测模型

摘 要

作为一个新兴的研究课题，异质信息网络上的链路预测问题仍有待于深入挖掘，主流的研究方法有三种：基于分解的、基于随机游走的及基于深度网络的方法。这些方法为该领域内的研究人员指明了可行的方向，但也存在一些不足：如何生成和选择合适元路径是一个极其基本却复杂的问题，但目前为止缺乏权威的标准；网络的拆分和融合以及融合后是否能有效的还原原始网络仍缺乏足够的理论依据；深度网络如同一个黑盒，也许经过训练后能获得比较理想的效果，但其存在可解释性差的缺陷，不利于对问题机理的深入理解。本文针对以上方法的一些不足，参考随机游走模型的部分设计，结合蚁群算法中信息素传播的思想，设计出一种基于信息扩散思想的、适用于异质信息网络的链路预测模型。该模型为降低复杂网络的处理时间提供了一种新思路，并指出了另一种使用网络异质信息的可行方向，且在真实网络中获得了较好的效果。

在本文提出的基于信息扩散的异质信息网络链路预测模型中，粒子从网络中的某些节点出发，通过在网络中游走来传播概率。粒子游走的机制主要取决于两个策略：概率递减策略和概率传播策略。概率递减策略参考了真实世界中“亲疏有别”这一基本人际关系理念，通过机制的宏观调控，可使粒子在靠近源节点的其他节点上停留的概率大，在远离源节点的其他节点上停留的概率小，由此导致的结果是：节点总是趋于与其附近的节点建立联系，这和我们真实世界的情况是相符的。概率传播策略则参照了蚁群算法信息素传播的思想，粒子选择路径时受到连接类型和连接权重的影响。连接类型因子反映了某种类型的边在整个网络中的分布情况，这是一个宏观概念，它是历史的总和（类似于蚁群算法历史累积的信息素）；而连接权重因子反映了特定节点对边的偏向性，这是一个微观概念，它对应的是当下（类似蚁群算法蚂蚁经过某条路径是遗留下的信息素）。通过这两个策略，粒子即可在网络中游走并传播概率，进而通过统计所有目标节点的概率分布，得出节点对的相遇区以及总体相遇（建立连接）的概率。

模型的验证性实验主要采用Last.fm网站2011年提供的数据作为测试集，在该集合中，共有24164个节点，用户、歌手、歌曲、标签四类角色。在对网络中的用户-用户关系进行预测时，训练集和测试集按9：1进行划分。经过计算后，得出本模型预测的AUC为：92.21%，精确率为：89.42%，相比于CN模型的AUC：83.56%和精确率：82.71%，分别提高了和8.65%、6.71%。此外，本模型还与其他5个模型在该数据集上进行了用户-歌手、用户-标签、歌手-标签关系预测的对比实验，本模型的AUC指标平均比其他选中参加实验的模型高7.75%；为了验证模型的普适性，我们还将本模型与其他共6个模型在DBLP数据集上进一步进行对比实验，得出的结论是本模型效果普遍比其他6个模型要优，具体数据见表3。

关键字：异质信息网络、链路预测、随机游走模型、PageRank算法、蚁群算法

目录

[第 1 章 引言 1](#_Toc42014290)

[第 2 章 相关工作 3](#_Toc42014291)

[2.1 基于相似度的算法 3](#_Toc42014292)

[2.1.1 局部相似性指标 3](#_Toc42014293)

[2.1.2 全局相似性指标 5](#_Toc42014294)

[2.1.3 准局部相似性指标 6](#_Toc42014295)

[2.2 基于最大似然估计的方法 6](#_Toc42014296)

[2.2.1 层次结构模型 6](#_Toc42014297)

[2.2.2 随机块模型 7](#_Toc42014298)

[2.3 基于概率模型的方法 7](#_Toc42014299)

[2.3.1 概率关系模型 8](#_Toc42014300)

[2.3.2 概率实体关系模型 8](#_Toc42014301)

[2.3.3 随机关系模型 9](#_Toc42014302)

[第 3 章 问题定义 9](#_Toc42014303)

[第 4 章 基于信息扩散的异质信息网络链路预测模型 10](#_Toc42014304)

[4.1 算法框架 11](#_Toc42014305)

[4.2 初始化网络 12](#_Toc42014306)

[4.3 节点游走 13](#_Toc42014307)

[4.3.1 概率递减策略 13](#_Toc42014308)

[4.3.2 概率传播策略 14](#_Toc42014309)

[4.4 统计相遇情况 15](#_Toc42014310)

[4.5 计算得出排名 16](#_Toc42014311)

[4.6 与随机游走算法的区别 16](#_Toc42014312)

[第 5 章 实验与分析 16](#_Toc42014313)

[5.1 数据集 17](#_Toc42014314)

[5.1.1 Last.fm数据集 17](#_Toc42014315)

[5.1.2 DBLP数据集 17](#_Toc42014316)

[5.2 度量指标 17](#_Toc42014317)

[5.3 基准方法 18](#_Toc42014318)

[5.4 有效性验证 19](#_Toc42014319)

[5.5 灵敏度分析 21](#_Toc42014320)

[第 6 章 结果与展望 23](#_Toc42014321)

[参考文献 24](#_Toc42014322)

# 引言

网络数据分析是一个重要的研究范畴，网络在现实世界中无处不在，例如社交网络、论文引文网络、万维网。网络数据分析通常涉及节点或边上的节点分类、节点聚类、链接预测等。节点分类与聚类皆为比较传统的问题，目前该研究领域已经硕果颇丰；链接预测虽也诞生有一段时间了，但仍存在不少值得深挖的地方。链接预测的目的是发现节点对之间的关联关系，它具有许多现实世界的应用，例如社交网络中的朋友推荐，社交网络中检测虚假的节点或边，科学协作网络中的论文引用关系预测，复杂网络的建模等。

同质信息网络上已经提出了许多链路预测办法，大体上被分为局部/全局相似性指数、有监督学习、概率方法等。构造为局部/全局相似性指数的研究者的目标是为节点或链接提取一些局部（基于节点）或全局（基于路径）相似性特征。常见的局部指标有共同邻居（），Jaccard系数（），偏向连接Prefrential Attachment（），Adamic Adar（）和Resource Allocation（）等指标；而Katz指数，Leicht-Holme-Newman，平均通勤时间，随机游走，SimRank等被称为全局相似特征。尽管局部指标的计算简单，但全局指标可能提供更准确的预测。在用有监督学习来进行链路预测时，链接预测被当作为一个二分类问题处理，任何相似性指标都可以形成所需的特征向量，之后，可以将任何常规的监督学习算法应用于训练监督链接预测器；基于概率的方法的主要思想是优化目标函数，以便建立使目标参数达到最优的模型。通过在建立关于要学习的参数的条件概率模型，来获得其后验概率。

然而，多数真实网络通常包含各种不同类型的节点，这开启了一个名为异质信息网络的新研究主题。在这种网络中，针对同质信息网络中提出的大多数链路预测模型在异质信息网络中不再适宜。区别于同质网络的是：在复杂的异质信息网络中，两个对象可能通过不同类型的边相连，而这些边的语义并不一致。针对异构复杂网络中的节点间复杂的连接这一特性，有人提出了一种称为元路径[1]的元结构。元路径是一系列不同类型节点之间的连接，被用于充当节点与节点相似程度的评判指标。当不限制元路径长度时，可能的元路径的数量允许是无穷的。但是，在基于元路径的相似性预测中，生成元路径、选择最佳路径是一个比较重要且复杂的问题。因此，关于元路径的研究又大致分为两类。第一类研究的主体是发现并选择有效的元路径，尤其是在大型网络中，人们可以通过合理运用传统知识发现原则，在数十万种可能的选择中，找出最有意义的元路径。第二类问题研究时假设已经给出了元路径，目标是探索新的相似性度量或开发出有效的相似性度量标准，适用于元路径的常见相似度衡量标准主要有：路径计数（）、随机游走（）[2]、HeteSim[3]、SimRank[4]、PathSim[5]。

基于元路径的相似度指标也并非尽善尽美，它主要有两个缺点。首先，相似性指标对节点间的可达性的依赖非常有限，即它们不考虑元路径中的信息，因此，它们倾向于偏向高度可达或集中的对象。其次，大多数这些指标最初是为单个且通常是对称的元路径设计的。换句话说，即使可以使用一组有用的元路径，要使所有元路径都是有效的，从而提高预测质量也不是一件容易的事。

考虑到选择有效的元路径的困难，以及缺乏对元路径中信息的有效使用，本文试图从另一个方向对网络中的异质信息进行利用。在本模型中，粒子不再是基于元路径游走，而是在所有可达连接上运动，但边的类型会影响到连接的可达性，总的来说，粒子倾向于往出现概率高的类型的边上运动，倾向于往连接强度较大的边运动。这样考虑的好处是，不用费劲心机去选择合适的元路径，且由于粒子游走的方式简单，大大提高了模型的普适性。经过实际数据集验证，这种思路的确简单而有效。本模型主要有以下三点贡献：① 提出了一种利用网络中异质信息的新思路——参考蚁群算法的信息素这一概念，使用异质信息影响粒子游走过程中的路径选择；② 指出降低网络处理时间的一种方法——建立网络后，仅对要预测边两端节点所属类型的所有节点进行处理，对其他节点不处理；③ 将真实世界中的一种关系模式应用到网络处理——通过概率递减策略，使粒子游走后的概率分布在宏观上靠近源节点的稠密、远离源节点的稀疏。简洁起见，本模型，即“基于信息扩散的异质信息网络链路预测模型”，在下文中简称信息扩散模型（）。

# 相关工作

信息网络的研究已成为许多科学分支的共同焦点，为了了解网络的发展，拓扑结构与功能间的联系，科学家们已经做出了巨大的努力。非常基本的一个研究课题是进行链接预测，它试图估量出某条链接有多大的存在概率。链接预测是现代信息学研究工作者要花费大量时间翻越的一座高山，除了识别虚假链路、帮助分析缺少数据的网络外，链接预测算法还可预测动态变化的网络中有较大概率出现的连接。当前的链路预测办法大体上可分三种：基于相似度的、基于最大似然估计的以及基于概率模型的方法。

## 基于相似度的算法

基于相似度的算法是链接预测问题最朴素的解决方案。其中，对于每对节点和，都将分配到一个分数，定义为和之间的相似度。依据得分对那些可能出现的连接进行排序，并且连接相似度更大节点对被认为有更高的存在可能性。尽管这种想法很简单，基于相似性算法的研究依旧是一个主流方向。在一种方法中，节点的基本属性决定了它们的相似性：具有多个相同特性的节点对，被认为是相似的。但是，节点的属性一般情况下难以量化，因此我们更关注另一种相似性度量标准——结构相似性，其仅基于网络结构。结构相似性指标可以以多种方式进行分类，比如局部与全局、无参数与参数相关、节点相关与路径相关等。相似性指标也可以分为结构等价性和规则等价性。前者隐含了一个潜在假设，即链接本身就指明两个端点之间存在相似性；而后者则假定若一对节点的邻居相似则它们相似。我们将20个相似性指标分为三类：10个局部指标，7个全局指标，3个准局部指标（准局部指标不需要全局拓扑信息，但要使用比局部指数更多的信息）。

### 局部相似性指标

(1) 共同邻居[6]（）：对于节点，定义为节点所有邻居的集合。通常来说，若节点和节点有更多公共邻居，则它们有更大可能建立连接。衡量共同邻居的最简单度量是直接计数，即

表示集合Q的大小。显然，其中A是邻接矩阵：若节点x和y直接相连，则，否则，。注意也表示连结节点x和y，长度等于2的相异路径数。纽曼在科学家合作网络的研究中使用了这一指标，得出共同邻居的数量与两位科学家未来合作的可能性之间存在正相关的结论。Kossinets和Watts也分析了一个大规模的社交网络，结果表明两个具有大量相同朋友的学生未来也有较大概率成为朋友。

(2) 索尔顿指数[7]（））：

其中表示节点的度。索尔顿指数也常常被叫做余弦相似度。

(3) Jaccard指数[8]，该指标由一百多年前的Jaccard提出，其定义为：

(4) 索伦森指数[9]（）。该指标主要用于生态群落数据集

(5) 大度节点有利指标[10]（）：提出该指数用于量化代谢网络中底物对的拓扑相似程度。在这种测量下，由于分母仅依赖于较低的度数，与大度节点相邻的节点可能会被分配较高的分数。

(6) 大度节点不利指数[11]（）。与上述指标类似，我们还考虑对大度节点持相反态度的指标。

(7) Leicht-Holme-Newman指数[12]（）。该指标向具有许多公共邻居的节点对分配了高相似度，不是可能的最大邻居数目，而是此类邻居的预期数目。我们使用缩写将该指标与也由Leicht、Holme和Newman提出的另一个指标（LHN2指标）区分开。

(8) 优先连接指标（）。优先连接机制在生成动态的无标度网络中发挥重要作用，其中新链路连接到节点的概率与成正比。一种类似的机制也用于无增长的无标度网络，在这种机制中，每个时间步都删除了旧的链接，同时生成了新的链接。在于各种动态的网络中，优先连接指标已被广泛用于量化链接对某种功能的重要程度，例如渗滤、同步、运输。值得一提的是，该指标无需每个节点的邻域信息，因而导致它只有很小的计算复杂度。

指数： 为了完善仅对公共邻居简单计数这一算法，该指标为具有较少连接的邻居分配更大的权重。

资源分配指数（）：该指标灵感源于复杂网络上的资源分配动态。假设有一对不直接相连的节点和，节点可以向发送一些资源，二者的共同邻居负责资源的接收与转发。考虑极端情形，假定每个发送者将一单元资源平均地发送给所有邻居。从接收到的资源数，我们将其定义为和之间的相似性。显然，该度量是对称的，即。值得注意的是，尽管出于相异的动机，指数和指数具有非常相似的形式；事实上，那些高度数共同邻居的贡献在这两个指标都被压制了。

### 全局相似性指标

(1) 卡茨指数[16]（）：该指标基于所有路径的总和，它对路径集合直接进行求和。之后按路径的长度使指标呈指数衰减，以达到使较短的路径获得更大权的目的。

(2) 指数[17]（）：该指标是Katz指标的变体。根据若一对节点的直接邻居集相近，则它们相似这一理念，从而推出一个包含自洽矩阵的公式。

(3) 平均通勤时间[18]（）。定义为随机游走粒子从节点到所需的期望步数，则与之间的平均通勤时间被表示为：

(4) 基于的余弦相似性[19]：该指数是基于节点内部特性的度量，表示出节点特征向量后，再用节点向量来计算余弦相似度。

(5) 带重启的随机游走[20]（）。该指数为PageRank算法的直接应用。考虑一个从节点开始的随机游走者，该游走者每次有的概率返回初始节点，并迭代地以概率随机移动到当前节点的一个邻居，迭代足够多次数后将达到稳态。

(6) 指数：与相似，在“两个节点连接到相似节点是相似的，则它们是相似的”这一前提下，SimRank被定义为一个自洽的指标。此外，也可用随机游走过程来解释SimRan，即SimRank指标度量了分别从节点和开始的两个随机游走者在某个节点上预期相遇的时间。

矩阵森林指标（）：矩阵和之间的相似度可以理解为：那些包含了节点和节点、且以为根的树，具有相同根的森林数量之比。该指标已用于量化协作推荐任务上衡量节点之间的相似性。结果表明，基于度量的相似性中，简单最近邻规则表现最佳。

### 准局部相似性指标

(1) 局部路径指标[22]（）： 该指标较好的平衡了准确性和计算复杂度，且考虑到了节点的局部路径，其范围比宽。指标的性能明显优于基于邻域的指标，例如，和。这是因为邻域信息难以区分，并且两个节点对被分配相同相似性分数的可能性很高，这就是LP指数显著提高预测准确度的原因。

(2) 局部随机游走[23]（）。出于测量节点和间的相似度的目的，需要在节点上放置一个随机游走者，节点的初始向量是一个分量为1的单位向量，之后节点游走若干步，每次更新为原来值的倍。

(3) 叠加的随机游走[24]（）。类似于指数，Liu和Lu提出了指数，其中随机游走者在起点处多次释放，从而导致目标节点与附近节点之间的相似度更高。Liu和Lu系统地比较了两个指标和，以及其他五个指标，包括三个局部（或准局部）指标，和，以及其他两个基于随机游走的全局指标，和，以及，Clauset和提出的分层结构方法（）。根据实验结果，和方法的性能优于其他指标，它们各自的最佳步行步长与网络的平均最短距离成正比。此外，和的计算复杂度低于和和的优势是能够保持较低的计算复杂度，尤其是在庞大的网络（即大）和稀疏的网络（即小）中。

## 基于最大似然估计的方法

本节将介绍最近提出两种基于最大似然估计的算法。这些算法以网络结构的一些组织原则为前提，并通过最大化观察到的结构的可能性来获得详细的规则和特定参数。之后，可以根据那些规则和参数来计算任何可能出现链接的概率。

### 层次结构模型[25]（Hierarchical Structure Model）

经验证据表明，许多真实世界中的网络是按层次结构组织的，它们的节点可以分为多个族，进一步细分为组，依此类推在应用到各个不同的尺度上（例如，代谢网络、大脑网络）。正如雷德纳所说，关注社会和生物网络固有的层次结构可能提供寻找丢失链接的明智方法。、Newman与提出了一种通用技术，能够从数据中推断出网络的层次结构，并将其进一步应用于预测缺失的链接。网络的分层结构可以用具有个叶子（对应于网络的节点）和个内部节点的树状图表示，该图中的每一内部节点都关联着一个概率，一对节点建立连接的可能性为，其中是这些节点中最近的公共祖先。

预测丢失链接的算法包含以下过程：

1. 对大量树状图进行采样，采样概率与它们的似然性成正比；
2. 对于每对未连接的节点和，通过平均所有采样树状图的对应概率来计算平均连接概率。
3. 以的降序对这些节点对进行排序，排名最高的将是要预测的节点对。

### 随机块模型[26]（Stochastic Block Model）

作为最通用的网络模型之一，随机块模型中的节点被划分到组，两个节点所在的组决定了它们建立连接的概率。随机块模型可以捕获社区结构、角色与角色的连接、可能建立连接的其他因素，尤其是当组成员决定了某个节点如何与其他节点交互时，它能起重要作用时，这种情况下通常不能通过简单的分类系数或度-度相关性来很好地描述。

随机块模型的可靠性描述了在给定观察到的结构的情况下，链接存在的可能性。它不仅可以用来预测丢失的链接，还可以识别可能的虚假链接。对五个不同网络的实证比较表明，基于随机块模型的最大似然方法的总体性能优于基于分层结构模型的方法和基于相似度的常见邻居算法。

## 基于概率模型的方法

概率模型诞生的初衷是为了从观察到的网络中提取出底层结构，然后通过使用训练后的模型来预测丢失的边。给定目标网络，概率模型将优化已建立的目标函数，以得出由一组参数组成的模型，该模型最能拟合目标网络的观测数据。之后，根据条件概率，估算没有被观测到的链接存在的概率。本节将介绍三种主流方法，分别为概率关系模型（ ），概率实体关系模型（）和随机关系模型（）。请注意，在某些文献中，术语仅通常指代关系贝叶斯网络的特定模型，而本文中我们采用的更一般用法。

### 概率关系模型[27]（Probabilistic Relational Models）

该模型表示关系数据集上属性的联合概率分布。它允许对象的属性概率性地依赖于自身的其他属性以及相关的其他对象的属性。传统图形模型通常运用单个图来建模同质实体属性之间的关系，而包含三个图：数据图，模型图和推理图。这些对应了Heckerman等人概述的骨架，模型和地面图。数据图代表输入网络，其中节点对应数据中的对象，而边对应了对象与对象的关系。模型图表示项目类型级别的属性之间的依存关系。推理图代表单个测试集中所有变量间的概率依存关系。

### 概率实体关系模型[28]（Probabilistic Entity Relationship Models）

概率实体关系模型的一种特定类型是有向非循环（简称），它通过有向弧来描述属性与属性的联系。使关系成为建模语言中的第一类对象，并鼓励显式表示条件概率分布。

模型由六个类组成：

1. 实体类：指定现实世界中的对象类；
2. 关系类别：代表实体类别之间的相互作用；
3. 属性类：描绘实体或者关系的属性。
4. 弧类：表示对应属性之间的概率依存相关性；
5. 局部分布类：为与该属性类别相对应的属性构建局部分布；
6. 约束类：指定如何在给定的实例化域上从相应的模型中导出推理图。

模型中关系分配与实体一样重要。当关系结构本身不确定时，模型的作用便体现出来了。它比或印版模型更具表现力。实际上，结合了印版模型和的功能。

### 随机关系模型[29]（Stochastic Relational Models）

随机关系模型的主要思想是通过多个高斯过程（）的张量交互来建模实体关系（即链接）的随机结构，每个高斯过程（）都在一种类型的实体上定义。实际上，该模型在无限维数张量矩阵上定义了一组非参数先验因子，其中每个元素代表一个实体元组之间的关系。通过最大化边缘似然函数，可以通过整个关系网络在参与的之间交换信息，从而将链接的依赖关系传递给实体的依赖关系，并由经过修改的内核反映出来。因为是在链接的条件模型上进行地训练，所以该模型为链接预测提供了一种判别方法，用以预测关系的存在与否、优势或基于部分观察到的链接网络以及实体的属性类型。Yu等人进一步升级了具有边缘协方差的SRM，可以降低整体计算复杂度。

# 问题定义

异质信息网络：通常用有向图来表示异质信息网络。是图中所有节点的集合，是所有边的集合，给出了节点类型的取值空间，代表了边类型的集合，任意两种节点类型、唯一确定了一种边类型，即存在映射：。设A大小为，则有大小为，其中代表类任选两不同类型节点的可构成边的可能数，代表所有相同类型节点的可构成的边数。

显然，当大小等于1或大小等于1时，网络退化为同质信息网络。在本文中，使用Last.fm于2011年公布的数据对模型进行验证，该数据集定义了4种节点类型：用户U、歌手A、歌曲M、标签T，由此导出10种元路径：UA、UM、UT、AM、AT、MT、UU、AA、MM、TT。

取节点类型函数：，将任意节点映射到对应的类型，即取出节点类型。

取边类型函数：，将任意连接映射到对应的类型，即取出边类型。

链路预测：对于任意两个节点和，确定边存在的概率，更一般的问题是，给定两种节点类型、，确定所有类型的连接存在的概率。

表1-符号声明表

|  |  |
| --- | --- |
| 变量 | 说明 |
|  | 用户节点类型 |
|  | 歌手节点类型 |
|  | 歌曲节点类型 |
|  | 标签节点类型 |
|  | 节点游走的步数 |
|  | 节点游走的最大步数 |
|  | 每走一步节点停留概率的衰减率 |
|  | 节点出现的第步的总概率 |
|  | 边的权重 |
|  | 节点游走到节点的概率 |
|  | 类型的边出现的总概率 |
|  | 节点与相遇的概率 |
|  | 节点所有可能出现的节点列表 |
|  | 节点的类型 |
|  | 边的类型 |

# 基于信息扩散的异质信息网络链路预测模型

基于信息扩散的异质信息网络链路预测模型的基本思想是：假设要预测节点、建立连接的概率，粒子从所有、类型的节点出发，依据某种策略在网络中随机游走，当游走收敛后，得到所有粒子在网络中的概率分布图。根据任意两个节点的分布概率图，计算节点相遇的概率，若两个节点同时出现在同一位置的概率之和或其排名大于给定的阈值时，认为它们会建立连接；反之，则认为它们不建立连接。

## 算法框架

算法主体描述了一个节点的游走过程，本模型由以下构件组成：

1. 概率递减策略

本策略基于某一节点与近距离节点的联系总是强于与远距离节点的联系这一基本原则，控制不同范围内节点的概率之和，使得节点到近范围节点的概率和总是大于到远范围节点的概率和。本策略的实现不一而足，可使用线性模型、指数模型、反比例函数模型以及其他模型，在本文中使用线性模型。

1. 概率传播策略

本策略实现概率在节点间的转移和传播。传播概率的基数为当前节点拥有的概率，当前节点的概率会按照一定比例传播给所有相邻节点（这个比例依赖于概率递减模型），每个相邻节点获得的概率可能不同。本策略参考了蚁群算法而设计，节点获得概率依赖于以下四个因素：

1. 概率基数：当前节点概率；
2. 衰减系数：反映概率递减策略，节点停留概率随步数增大在递减；
3. 连接类型：连接类型反映了同等情况下粒子选择下一经过边的偏向，类似于蚁群算法中历史累积的信息素。这是一个宏观因子，与整个网络各种边类型出现的概率相关；
4. 连接强度：当前节点与下一节点连接强度（表现为权重）越大，则其获得的概率越大，类似于蚁群算法中本轮行走留下的信息素。这是一个节点特定因子，与节点本身的性质相关。
5. 概率统计

通过概率递减策略和概率传播策略，可以得到一个节点游走后在图中的分布情况；通过对不同节点迭代以上过程，最终得出全部粒子游走后的分布情况。概率统计策略负责统计任意两个节点建立连接的概率，并对统计得出的节点对进行排序，选出最有可能出现的N个节点对。具体过程如下：

1. 对于节点和，统计它们可能出现的所有节点集合，记为、；
2. 选择、中的较小者，记为、剩余者记为；
3. 初始化节点和的相遇概率。对于中的每一个节点，如果同时出现在中，则；其中表示出现在中的概率，其中表示出现在中的概率。
4. 重复以上过程，直至所有节点相遇概率统计完成。

## 初始化网络

初始化图时应根据具体的节点类型建立不同的连接。

1. 对于所有用户类的节点，建立U🡪U，U🡪A，U🡪M，U🡪T类型的边，注意连接是单向的，这意味着存在U1🡪U2的同时，未必存在U2🡪U1。
2. 对于所有歌手类节点，建立A🡪M，A🡪T的边，注意并非所有类型的边都存在，比如A🡪U类型的边就不存在。
3. 对于所有歌曲类型的节点，建立M🡪A，M🡪T的边；对于所有标签类型的节点，建立T🡪M的边。

建立何种类型的边应当根据具体情况而定，在本次实验中，我们建立了U🡪A类型的边，却没有建立A🡪U类型的边，这意味着可通过用户达到歌手节点，却不能通过歌手到达用户节点。

依据该策略建立的一个典型图如下：

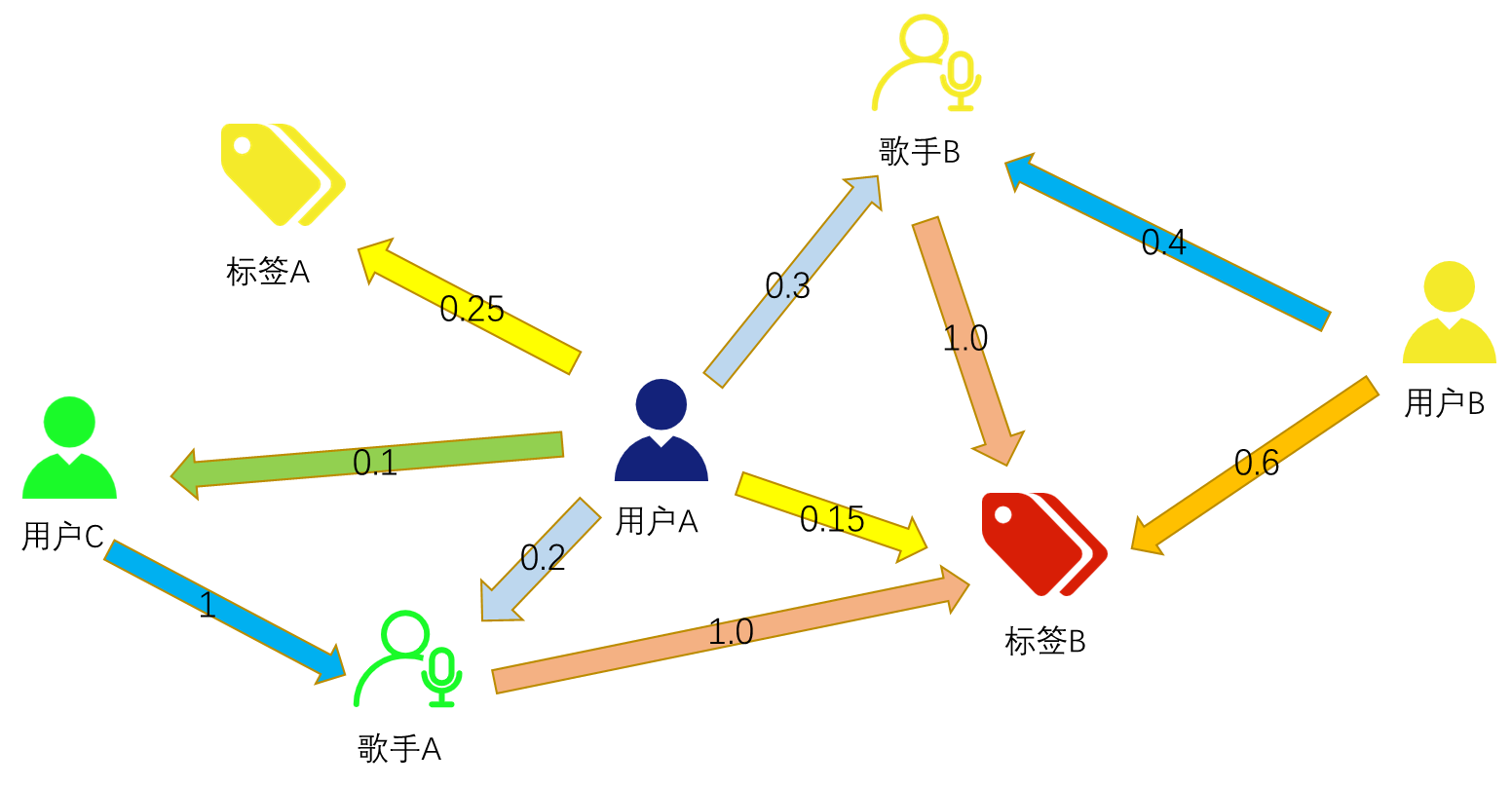


图1-示例图

## 节点游走

确定节点游走的策略是模型的关键之一，节点游走的策略越贴近真实情况，对节点可能出现位置的预测也越精确，由此统计得出可能存在的边也准确。节点游走策略包括概率递减策略和概率传播策略：

### 概率递减策略

初始化一个权值，代表当前节点走步的概率。随着步数的增大，在逐步减小，且满足：

是一个衰减系数，代表节点继续行走的概率随着步数降低。此处参考了蚁群算法信息素衰减的思想，随着节点游走步数的增大，节点停留的概率在减小。

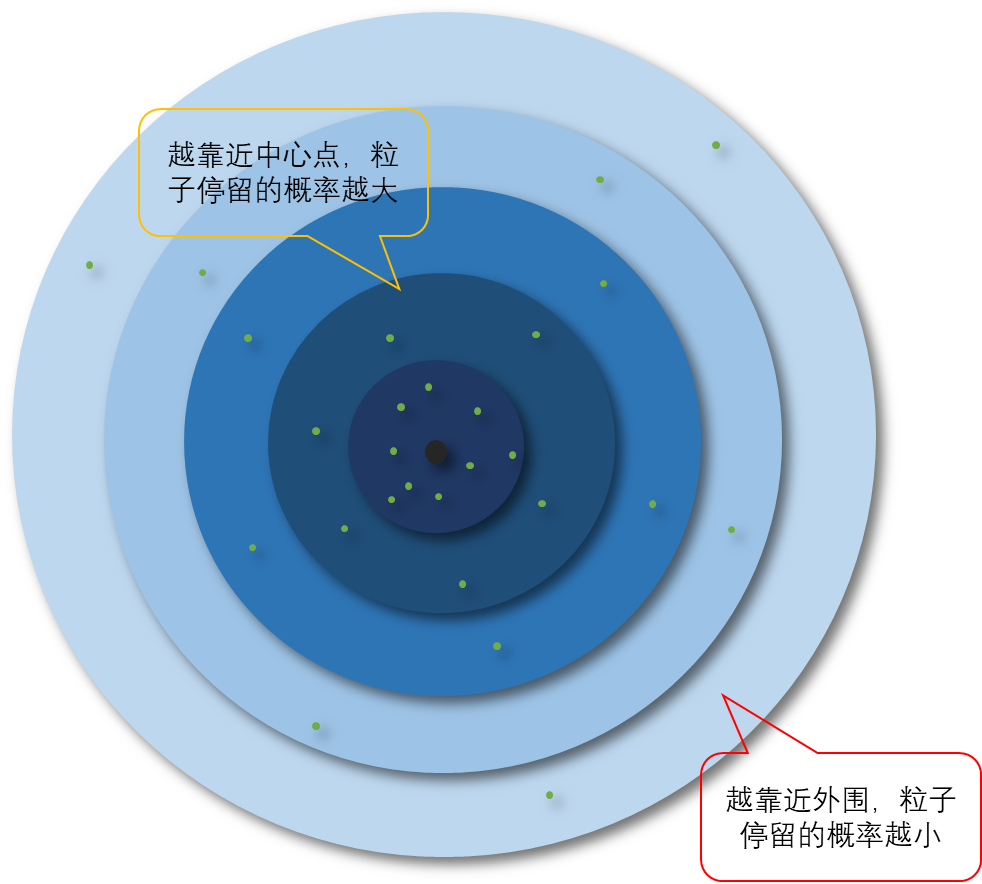


图2-节点停留概率分布示意图

### 概率传播策略

概率传播策略的灵感来源于蚁群算法和随机游走算法。决定下一节点概率大小的因素除了当前节点的概率基数及衰减系数外，还与边类型及连接强度有关。此处的边类型类似于蚁群算法中历史累积的信息素，它反映了同等情况下，粒子选择边的偏向。而连接强度则类似于当蚂蚁经过当前边时留下的信息素，它反映了无历史情况下，信息素将形成的分布，并且由所有当前（某条边的连接强度）构成了历史（边类型导致的偏向）。

假设源节点是，对于可达的某个节点，由则游走到的概率为：

其中代表边所属类型出现的概率，代表边的权。计算出节点游走概率的同时，将加入的可达节点列表。

整个过程示意图如下：

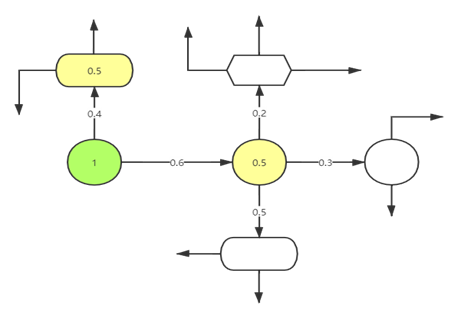
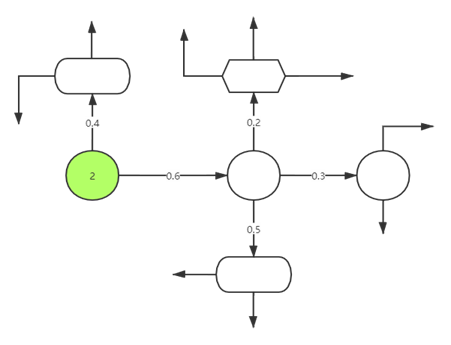


图3-初始时节点概率分布 图4-第1步节点概率分布

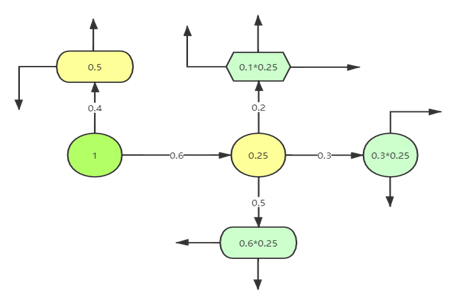


图5-第2步节点概率分布

## 统计相遇情况

表示节点所有可能到达节点的列表，同理。对于节点和，由以下算法计算它们相遇的概率：

(1) 对于节点和，统计它们可能出现的所有节点集合，记为、； (2) 选择、中的较小者，记为、剩余者记为；

(3) 初始化节点和的相遇概率。对于中的每一个节点，如果同时出现在中，则；其中表示出现在中的概率，其中表示出现在中的概率。

(4) 重复以上过程，直至所有节点相遇概率统计完成。

算法的伪代码表示如下：

|  |
| --- |
| for in :  if in : |

其中代表节点、相遇的概率，代表节点出现在的概率，代表节点出现在的概率。

一个相遇场景范例如下所示：

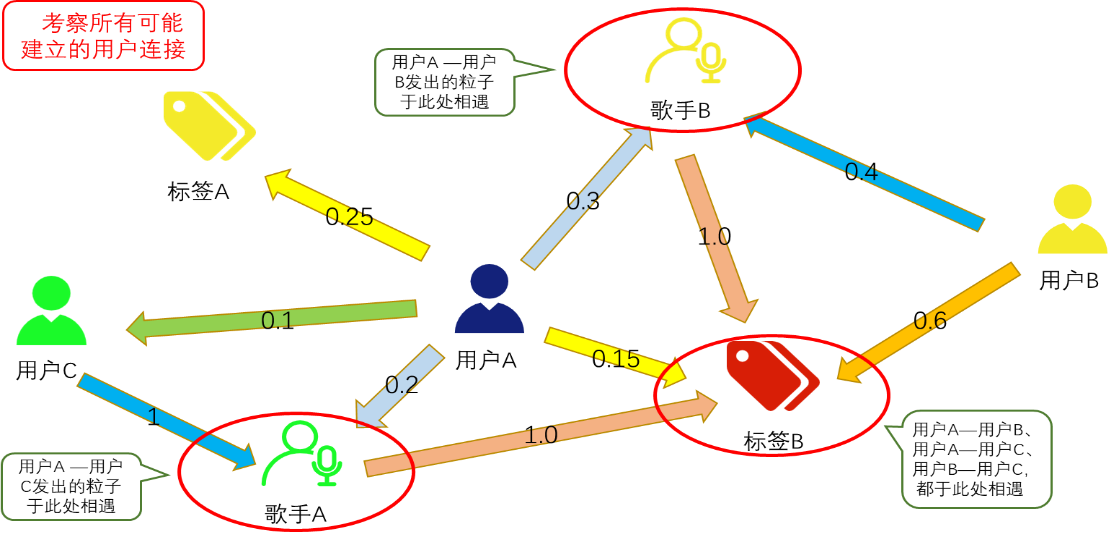


图6-节点相遇示意图

## 计算得出排名

对于所有的，将它们从小到大排列，假设测试集大小为，训练集大小为；取中排名前的节点对（即边），剔除集合中的节点，剩下的便是预测的结果。

## 与随机游走算法的区别

本算法参考了随机游走算法的基础游走步骤，它与随机游走算法的6大区别为：

1. 将随机游走算法的最终目标是刻画某种影响力在网络中的分布，而本算法的最终目标是为了得出一系列节点最终的分布情况。概括来说，随机游走算法寻找网络整体特征，本算法寻找网络中某些节点的特征；随机游走算法是整体的，本算法是局部的。
2. 本算法概率从源节点开始逐步往外扩散，源节点是特定的；随机游走算法概率可从任意节点开始迭代，源节点是任意的。
3. 随机游走算法中粒子选取下一节点的机制是：以的概率等概率选择周围节点，以的概率随机选择节点；本算法利用异质信息网络的特点，依据与周围节点连接的权和连接的类型非等概率选择下一节点。
4. 本算法概率扩散时，近范围节点概率和总是大于较远范围节点概率和；随机游走算法则无此限制
5. 随机游走算法不限制粒子行走的步数，本算法限制粒子游走步数。
6. 随机游走算法需不断迭代直至收敛，本算法只需对目标预测类型的所有节点执行一遍游走算法，即最坏时间是可预知的。

# 实验与分析

以下将本文提出的方法在2个真实数据集上，与其他6种方法（包括共同邻居、系数、优先连接指标、平均通勤时间、带重启的随机游走算法、基于元路径的异质网络链路预测模型），对和准确率2个指标进行比较，以验证方法的有效性。此外，还将考察本模型与共同邻居对测试集/训练集大小之比的的敏感程度，以研究本模型相对于其他模型，其处理效果受测试集/训练集大小多大程度的影响，为以后的研究者选择合适的训练集容量、以及确定模型预测结果的可信度提供参考。

## 数据集

### Last.fm数据集[30]

是国外的某个流行音乐网站，它提供了2011年的部分数据用于学术研究。该网络包含用户、歌手、标签；1892个用户节点、17632个歌手节点、11946个标签节点、25434个用户-用户关系、92834个用户-歌手关系、186479个用户-标签-歌手关系，总计31470个节点、491226条边。

实验的目标是预测最有可能建立用户-用户关系、用户-歌手关系、用户-标签关系、歌手-标签关系。对于25434组用户-用户关系，将训练集与测试集大小按9：1进行切分，得到训练集大小为：22825，测试集大小为：2609。对于其他关系，也是大致按9：1随机挑选边放入训练集与测试集中。

### DBLP数据集[31]

全称“数字书目索引与图书馆项目”,它为计算机领域内科学文献的检索提供服务。截至到目前为止，它收集了5,084,898个文章、2,512,935个作者、5,209个会议、1,698个期刊之间的关联关系，是网络信息挖掘研究中非常权威、应用也十分广泛的一个数据集。经过清洗后，我们选取了其1980年前的数据来完成此次实验，共37944条数据，其中包含27981个作者、38117篇文章、641个会议总计66739个节点、166560条边。

本次实验的目标是预测最有可能建立连接的文章-文章、文章-会议、文章-作者关系。在25464组文章-会议关系中，我们选取了22821个作为训练集、2643作为测试集。对于这三组关系的预测，我们都是大致按9：1随机选取边放入训练集与测试集，每种关系实验次数不小于3，以其结果的均值作为最终标准。

## 度量指标

主要使用AUC和准确率两个指标：

(1) AUC[32]（）是一个被广泛使用的、用于衡量链路预测算法精度的指标。被表示曲线下的面积，它的内涵是在测试集中那些边的评分，相比与随机选一非真实存在边的评分高的概率。每次随机在测试集中选取一条边和另一条非真实存在的边对比，若测试集中选取的边评分较高，则加1分，同等大小时加0.5分，小于时不加分。进行次比较后，为：

其中表示大于次数，表示相等次数。

(2) 准确率[33]（)表示被检测到的正样本占所有正样本的比例：

：被检测到的正样本，：未被检测到的正样本。

## 基准方法

本次实验共涉及6个基准方法，它们分别是：

(1) 共同邻居：本文简记为。、皆为一个集合，它们代表了节点、的所有直接邻居，则它们的共同邻居为：。对于任意两个节点，它们建立连接的概率正比于它们的共同邻居数目。

(2) 系数：本文简记为。该系数在考虑了两个对象之间拥有共同邻居的数目的同时，也考虑了对象总邻居的数目。假设我们要衡量节点和的相似度，记为节点的直接邻居数，为节点的直接邻居数，则节点与的Jaccard系数可表示为：

(3) 优先连接指标（）：本文中简记为。该理论认为，建立边的概率正比于节点的度。因此，对于节点与，若节点的度记作，节点的度记作，则它们的优先连接指标为：。

(4) 平均通勤时间（）：本文简记为。给定两个节点和，若从节点到平均需要走步，从节点到节点平均需要走步,则二者的平均通勤时间为：。

(5) 带重启的随机游走（）[18]：本文简记为。该理论的基本思想为：对于网络中的任意一个粒子，每走一步都以的概率转移到相邻节点、以的概率返回初始位置。时刻，节点处于网络中各节点的概率表示为：，式中代表了网络的马尔可夫概率转移矩阵（），代表初态。表示节点从出发有多大概率到达，由此得出指标为：。

(6) 基于元路径的异质网络链路预测模型（

）[34]：本文简记为。该模型研究的基元是元路径，通过梯度下降算法训练得出每种不同的元路径对目标类型边的权重，对于要预测的一对节点，组合这对节点之间的元路径和训练得出的权重，即可得出边存在的概率。该模型简单描述为：，其中代表第条相关的元路径，代表第条元路径训练所得的权。

## 有效性验证

**(1) 准确率**：准确率作为一个重要参考指标，我们将共同邻居（）与信息扩散模型（）依据此指标进行对比实验。多次实验后取均值，最终得到的数据如下：

表2-CN与MSM模型准确率对比表

|  |  |  |  |
| --- | --- | --- | --- |
|  | TP+FN | TP | 准确率 |
| CN | 2609 | 2158 | 82.71% |
| MSM | 2609 | 2333 | 89.42% |

相比于模型，本模型的准确率提高了6.71%，大于5%，说明模型的优势是明显的，有进一步研究的必要。究其原因，是因为相比于共同邻居等传统同质网络链路预测模型，本模型对网络中的异质信息进行了利用，因此获得了更好的效果。毫无疑问，对信息利用的越是充分，模型的精度就越高，这也是异质信息网络诞生的初衷。

对和两个模型的初步实验表明，本文提出的模型是有效的。为了更全面地验证本模型的精度和效能，接下来将模型与其他6个模型的指标进行横向对比，如果模型对比的结果普遍较好，则说明相比于其他模型，本模型确实对网络中的异质信息进行了更进一步的运用。

(**2) AUC**：本实验将模型与其他6个模型（、、、、、）在指标上进行对比实验。我们限制节点允许走的最大步数为5，重复至少3次实验，取的均值作为最终的结果，最终得到的数据如下：

表3-7个模型效果对比表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 关系 | CN | JC | PA | ACT | RWR | MPLP | MSM |
| DBLP | PP | 0.7221 | 0.7112 | 0.731 | 0.7332 | 0.7484 | 0.7343 | 0.8534 |
| PV | 0.8334 | 0.8528 | 0.8425 | 0.8521 | 0.8456 | 0.8559 | 0.9136 |
| PA | 0.7274 | 0.7313 | 0.7287 | 0.7227 | 0.7049 | 0.7375 | 0.8235 |
| Last.fm | UU | 0.8356 | 0.8427 | 0.8312 | 0.8276 | 0.8512 | 0.8432 | 0.9221 |
| UA | 0.8438 | 0.8213 | 0.8518 | 0.8733 | 0.8696 | 0.8347 | 0.9142 |
| UT | 0.8221 | 0.8346 | 0.8536 | 0.8436 | 0.8469 | 0.8238 | 0.8937 |
| AT | 0.8186 | 0.8128 | 0.8238 | 0.8389 | 0.8457 | 0.821 | 0.8763 |

这7个模型在Last.fm数据集上的表现如下：

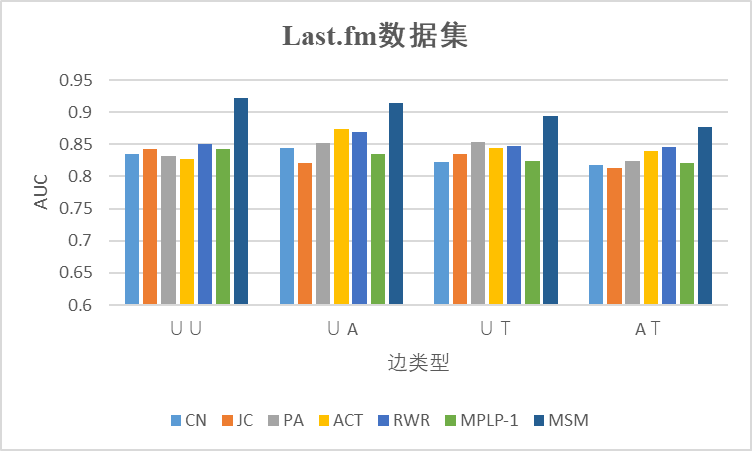


图7-Last.fm数据集上的实验结果

在DBLP数据集上的表现如下：

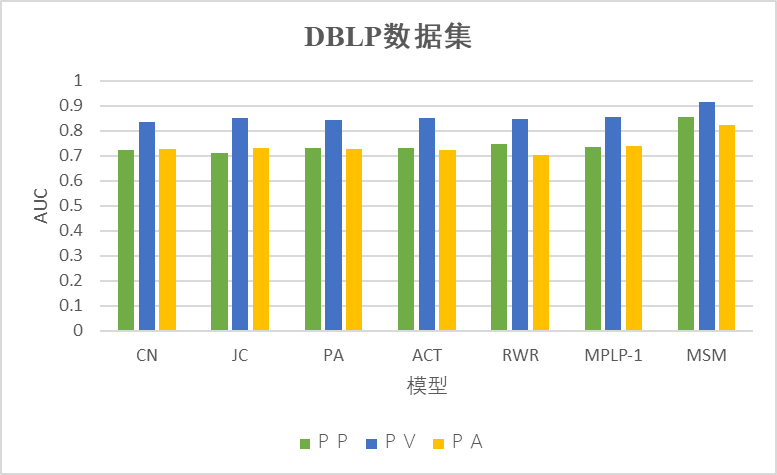
****

图8-DBLP数据集上的实验结果

由图可知，在2个真实的数据集上，相比于选取的6个基准模型， 模型在各组实验中都能取得理想的效果；最差比其他模型好3.06%，最好比其他模型好14.22%，平均比其他模型好7.75%。

对指标、系数来说，决定指标大小的节点邻域近限于与目标节点直接相连的其他节点，即其直接邻居；而本模型的则选取了更大的邻域，即为s步内可达目标节点的所有节点，更大的邻域虽然会增加处理时间，但也能获得更好的效果。对于优先连接指标（）、平均通勤时间（）、随机游走（）、基于元路径的异质网络链路预测模型（）等算法，它们的路径都是无权的，一对节点间建立连接的概率仅依赖于节点间的可达性；而在本模型中，节点之间建立直接联系的概率不仅和节点间的可达性相关，且于路径的权相关，而路径的权又蕴含了网络的异质信息。此外，本模型还对粒子游走后的概率分布进行了宏观控制，使粒子停留在靠近源节点范围内的节点概率大，停留在远离源节点范围内的节点概率小，这与真实情况是相符的，其他6个模型都缺乏这个机制。总言之，扩大节点邻域、路径赋权、实现概率递减策略是模型能获得更好效果的根本原因。

## 灵敏度分析

调整训练集/测试集比例，运用共同邻居和信息扩散模型分别对数据集上的U🡪U关系进行预测，从而比较两个模型受训练集大小的影响程度（即模型对训练集大小的灵敏度[35]）。其结果如下图所示：

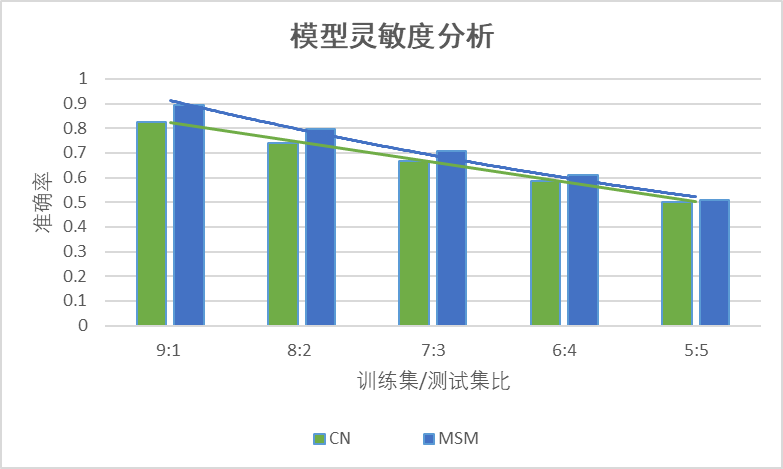


图9-训练集大小灵敏度

由上图可知，在训练集和测试集之比大于1:1的情况下，信息扩散模型的预测效果总是比共同邻居模型要好。但不可忽略的是，信息扩散模型受训练集大小的影响较大，这意味这训练集越大时，模型效果越好；与此同时，训练集越小，模型的优势越不明显。此外，由上述结果可知，当要求预测边数量为时，在预测得出的所有可能边集合中，前个结果的可信度大约为89.4%，前个结果的可信度大约为81.5%，前个结果的可信度大约为74.9%。

分析模型受训练集大小影响较大的原因，主要是模型对网络中蕴含的信息依赖较大。一方面，仅关心其直接邻居，缺少其他节点对其没有影响，而目标节点的直接邻居在网络中只占少数，因此减少训练集对其造成的影响较小；的领域比要大，与目标节点距离较远的节点也会影响到预测边的概率，减少训练集对其造成的影响较大。另一方面，模型中粒子选择路径时，边的类型会影响到其选择结果，当要预测类型的边在网络中急剧减少时，边类型影响因子势必也会受到影响，这造成了构建的网络中的异质信息与原始网络有较大偏差，粒子决策的信息不准确；而由于模型不关心网络中的异质信息，因此不存在此方面的偏差，这也是模型更稳定的原因。

# 结果与展望

链路预测是计算机网络科学中研究的一大基本问题，异质信息网络链路预测问题相比较于同质信息网络链路预测更加精确，同时也更加复杂。本文基于蚁群算法、算法、随机游走算法的部分思想，提出基于信息扩散的异质信息网络链路预测模型，该模型对网络中的异质信息进行了初步的利用，通过在真实的网络、网络数据集上实行的一系列实验，表明本模型确实有效的提高了预测的准确性。此外，本次实验也说明了，对网络节点及边的异质信息进行更加充分的使用，有利于提高模型链路预测结果的准确度。

本模型利用异质网络边类型不同的特点，为节点选择下一个到达的节点提供依据，即我们假设同等情况下，边类型会影响节点的游走方向。此外，我们利用节点类型不同这一信息，根据实际情况，在某些类型的节点间建立双向连接，在另外类型的节点间建立单向连接，这样节点的游走就有了方向。但在利用边类型这一信息上，为了不使模型过于复杂，我们利用的边类型整体的统计信息，而不是针对某一节点的具体选择偏向；这意味着模型可能存在比较大的偏差，如果能针对每一个节点，刻画它对各种类型边的选择偏向，也许我们能得到一个更加精确的模型，这就留给后人去研究了。

此外，为了降低节点遍历整个网络带来的时间开销，我们限制了节点所能游走的最大步长，这在提高模型效率的同时，势必会降低模型精度。如何在效率和精度中取得一个较好的平衡，仍需进一步的研究。

参考文献

[1] Niya Yang, Meta Path-Based Link Prediction Research for Heterogeneous Information Networks, 2019 TP393.02; TP311.13

[2] Pearson, K. The Problem of the Random Walk. Nature. 1905, 72. Bibcode:1905.

Nature.72.294P. doi:10.1038/072294b0.

[3] C. Shi, X. Kong, Y. Huang, P. S. Yu and B. Wu, "HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks,", vol. 26, no. 10, pp. 2479-2492, Oct. 2014, doi: 10.1109/TKDE.2013.2297920.

[4] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In KDD'02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 538-543. ACM Press, 2002

[5] Yizhou Sun†,Jiawei Han†,Xifeng Yan‡,Philip S. Yu§,Tianyi W, PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks, 2011

[6] Mitzenmacher A brief history of generative models for power law and lognormal distributions.Internet Mathematics, 2004,1(2):226-25

[7] Bartosz Chroł, Michał Bojanowski, Proximity-based Methods for Link Prediction, 2018-10-12

[8] Lipkus, Alan H, A proof of the triangle inequality for the Tanimoto distance, J Math Chem, 1999, 26 (1-3): 263–265

[9] Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons": 1–34.

[10] Tao Zhoua, Linyuan L¨u, and Yi-Cheng Zhang, Predicting Missing Links via Local Information, 2009

[11] Tao Zhou,a, Linyuan L¨u, and Yi-Cheng Zhang, Predicting Missing Links via Local Information, 2009

[12] Yu-Xiao Zhu1, Linyuan L¨u2, Qian-Ming Zhang, Tao Zhou, Uncovering missing links with cold ends, 2011

[13] Merton, Robert K. (1968). "The Matthew effect in science". Science. 159 (3810): 56–63. Bibcode:1968Sci...159...56M. doi:10.1126/science.159.3810.56. PMID 17737466.

[14] Lada A, Eytan, Adamic, Adar, (2003). "Friends and neighbors on the web" (PDF). Social networks. Elsevier. 25 (3): 211–230.

[15] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information. Eur. Phys. J. B, 2009, 71: 623

[16] L. Katz, A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39

[17] Chen Cheng, Junjie Chen, Xiaohua Cao, Hao Guo, Comparison of Local Information Indices Applied in Resting State Functional Brain Network Connectivity Prediction, 2016 Dec 27;10:585. doi: 10.3389.

[18] François Fouss， Marco Saerens， Masashi Shimbo. Algorithms and Models for Network Data and Link Analysis. Cambridge University Press

[19] Pickover, Clifford A. (2009). The Math Book: From Pythagoras to the 57th Dimension, 250 Milestones in the History of Mathematics. Sterling Publishing Company, Inc. p. 106. ISBN 9781402757969.

[20] Alberto Valdeolivas. Random walk with restart on multiplex and heterogeneous biological networks. 01 February 2019, Pages 497–505

[21] Pavel Chebotarev, Elena Shamis, The Matrix-Forest Theorem and Measuring Relations in Small Social Groups, Sat, 4 Feb 2006

[22] Linyuan Lü, Ci-Hang Jin, and Tao Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80, 046122 – Published 26 October 2009

[23] Aibo Song,Yangyang Liu,Zhiang Wub, Mingyu Zhai, A local random walk model for complex networks based on discriminative feature combinations, 15 March 2019, Pages 329-339

[24] Weiping Liu and Linyuan L¨u, Link Prediction Based on Local Random Walk, 2010

[25] Michael J. Kamfonas/Recursive Hierarchies: The Relational Taboo! Archived 2008-11-08 at the Wayback Machine--The Relation Journal, October/November 1992

[26] Krzakala, Cristopher, Decelle, Florent; Moore, Zdeborová, Aurelien, Lenka (September 2011). Physical Review E. 84 (6): 066106. arXiv:1109.3041. Bibcode:2011PhRvE..84f6106D. doi:10.1103/PhysRevE.84.066106. PMID 22304154.

[27] Friedman, Learning probabilistic relational models, 2001

[28] David Heckerman, Chris Meek, Daphne Koller, Probabilistic Entity-Relationship Models,

PRMs, and Plate Models, 2005

[29] Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, Zhao Xu, Stochastic Relational Models for Discriminative Link Prediction, 2006

[30] Last.fm. http://millionsongdataset.com/lastfm/ June 2011

[31] DBLP. https://dblp.uni-trier.de/ 2020-03

[32] Wikipedia. https://en.wikipedia.org/wiki/AUC 11 April 2019

[33] Wikipedia. https://en.wikipedia.org/wiki/Accuracy\_and\_precision 4 May 2020

[34] Cao, X., Zheng, Y., Shi, C. et al. Meta-path-based link prediction in schema-rich heterogeneous information network. Int J Data Sci Anal 3, 285–296 (2017).

[35] Wikipedia. https://en.wikipedia.org/wiki/Sensitivity\_and\_specificity 23 April 2020