# 基于双曲空间图嵌入的科研热点预测

**戴筠**
上海大学，上海 200041

**摘要**
预测科研热点可以有效地开展科学研究和更好地分配科学资源。数据挖掘和机器学习算法已经被广泛应用到科研热点预测中，比如基于论文文本内容的主题模型建模和挖掘论文被引频次的算法等。提出一种新的将关键词信息嵌入双曲空间的双曲空间关键词图嵌入（PKGM）算法，利用关键词和它们之间的关系构建一个关键词网络，通过计算双曲空间中两个节点的距离来判别两个节点之间存在边的概率，从而对科研热点进行预测。该算法与7个基准算法的实验比较结果显示，PKGM算法与效果最好的欧氏空间嵌入算法相比有7.3%的AUROC和5.8%的AP提升；与双曲图神经网络算法相比，有10.8%的AUROC和7.2%的AP提升。这显示了PKGM算法的有效性。

**关键词**
科研热点；双曲空间；庞加莱模型；图嵌入；关键词网络；长尾效应

中图分类号：TP18　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2022041

# *Emerging scientific topic prediction based on Poincare graph embedding*

**DAI Jun**
Shanghai University, Shanghai 200041, China

***Abstract***
Scientific topic prediction is central to scientific research and can substantially advance the allocation of scientific resources. Machine learning and data mining approaches have been widely applied to scientific topic prediction, including paper content-based topic model and citation prediction models. A novel scientific topic prediction algorithm PKGM (Poincare keywords graph embedding) was proposed, which utilized keywords and their relations to build a keyword network, and calculated the distance between two nodes in this network to predict the probability that an edge existed. The result of comparing PKGM with seven baselines showed that PKGM obtained a 7.3% improvement by using AUROC and 5.8% improving by using AP in comparison to the best method in Euclidean space, and 10.8% improvement by using AUROC and 7.2% improving by using AP over the best approach in hyperbolic space. The results demonstrated the effectiveness of PKGM.

***Key words***
scientific topic, hyperbolic space, Poincare's model, graph embedding, keywords network, long tail effect

## 0 引言

科研热点是指在某个时间段里许多研究者在探讨的学问或专题，它承载着科学研究各个领域的最新研究成果，对科学研究的发展具有指导意义[1]。传统的科研热点预测，是相关领域高级专业人员通过科技论文查阅与市场调研来确定的。科技论文发表数量日益增长，使专业人员快速了解研究内容、跟进研究热点变得越来越困难。

随着大数据时代的到来和深度学习的崛起[1-3]，数据挖掘和机器学习算法已经被广泛应用到科研热点预测中。传统的机器学习主要从两方面对科研热点进行预测：一方面是基于论文文本内容的主题模型建模，包括考虑摘要和全文的主题模型，通过算法将论文的文字设计成一个词袋模型，通过对模型的统计分析，结合与时间相关的信息，得到出现频率显著增高的词[4-6]；另一方面是通过论文被引用数量的变化来预测科研热点[7-10]，即挖掘被引频次显著增加的论文，那些被引频次居高不下甚至不断增加的论文的研究内容，通常就是这些研究领域的热点。还有一种未被广泛使用的方法[11]，即通过分析论文关键词来预测未来可能会被广泛使用的词，这些词往往代表了科研热点。这种方法相比于前两种方法，优点在于能更好地避免全文和引文中的噪声，因为这些关键词由作者提供，能更好地反映论文的主题。

本文从论文关键词中寻找科研热点，研究思路是构建一个关键词网络，网络的节点为论文的关键词，当两个关键词出现在一个句子中时，这两个关键词之间就形成一个链路，即网络的边。连接某个关键词节点网络的边数量越多，这个关键词就越有可能是未来的一个科研热点。本文的研究目的是预测哪些关键词节点会有较多的网络边。现有网络边预测算法虽然被广泛应用于社交和物流等网络中，但它们并不能在关键词网络中得到好的效果[12]，主要原因是关键词网络中的关键词具有明显的长尾效应[13]，即有大量的关键词只有很少的边，但同时又有少量的关键词有大量的边。另外，现有的这些算法只能关注到出现频率高的关键词，而完全忽略那些目前出现频率低但在未来频率显著增高的关键词，也就是这些算法只能关注到近期的科研热点，而无法预测未来的科研热点。

本文提出双曲空间关键词图嵌入（Poincaré keywords graph embedding，PKGM）算法来预测科研热点。与传统的欧氏空间相比，双曲空间能更好地处理具有长尾效应的数据。双曲空间以指数形式进行建模，可以有足够的空间来表示罕见的数据点。双曲空间可以消除随机噪声对这些数据点的干扰，更好地处理长尾效应的数据。PKGM算法在双曲空间中进行图嵌入，而不在欧氏空间中进行图嵌入。首先构造一个关键词网络，然后将此网络嵌入双曲空间。即使两个关键词在原网络中没有边连接，如果在双曲空间中的距离非常近，就会认为这两个关键词之间未来会有一条边。对这些边的寻找可以预测未来出现频率高的关键词，从而找到科研热点。将PKGM算法在一个真实的数据集上进行验证，构建的关键词网络包括9 966个关键词节点和18 976条网络的边。实验发现，PKGM算法比7个基准算法有更好的表现，包括欧氏空间中的最佳算法。

## 1 相关工作

本节通过两部分来回顾相关工作，分

别是图嵌入算法和双曲空间嵌入算法。

图嵌入算法已在很多图结构中获得应用，并且取得较好的效果。一方面工作是通过图嵌入进行无监督学习，在低维空间还原高维空间的相似性[11,14-16]。例如，DeepWalk通过在网络中随机游走获得低维空间的图节点特征向量[17]。大规模信息网络嵌入（large-scale information network embedding，LINE）采用二阶相似性，利用神经网络和深度学习对图的离散结构进行分析[18]。另一方面，有监督学习也在图嵌入网络中被广泛应用。例如，图神经网络通过图卷积网络完成图结构的分类和回归任务[5,19-23]，还通过图注意力机制进行图嵌入，从而能动态地对图的边设置权重[24-25]。虽然这些算法都能够获得较好的结果，但它们无法在双曲空间中进行嵌入。

双曲空间嵌入算法是最近机器学习领域中的一个新热点算法[26-31]。它的思想是使用双曲空间代替欧氏空间，从而能更好地对长尾效应数据进行建模。双曲空间嵌入（Poincare embedding）是这方面的先驱工作，它通过将已有的数据映射到双曲空间，并且在双曲空间中找到一个潜在的层次结构来建模数据[32]。PoincareGlo Ve[26]用双曲空间嵌入算法对文本数据进行建模，从而获得文本的词向量表示。双曲图卷积神经网络（hyperbolic graph convolutional neural network，HGCN）算法是另一个后续工作[12]，它通过应用基于图网络的有监督学习模型来学习边的连接和点的分类。HGCN算法还指出，并不是所有图都适合在双曲空间中建模，只有双曲曲率较小的图才更适合在双曲空间建模。与这些工作不同的是，本文首次将双曲空间嵌入算法应用到论文关键词网络中，解决了欧氏空间算法中关键词存在长尾效应问题。

## 2 方法

本节首先给出提出的PKGM算法框架，然后描述关键词网络的构建，最后详细介绍PKGM算法。

### 2.1 算法框架

首先对文本数据进行预处理，构建一个关键词网络，其次通过对数映射将关键词连接嵌入双曲空间，然后利用庞加莱球（Poincare sphere）模型在双曲空间中计算两个关键词节点的距离，通过指数映射计算欧氏空间中新关键词概率，最后对新科研热点进行预测。科研热点预测算法框架如**图1**所示。

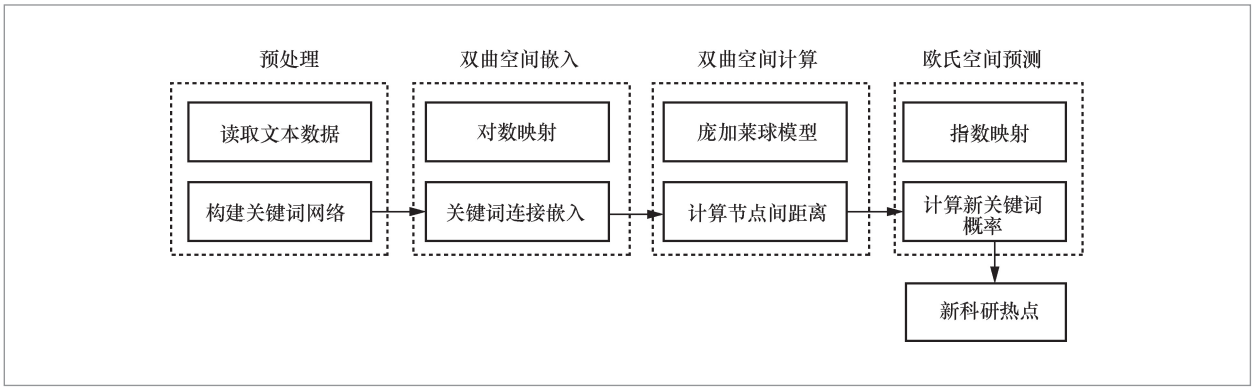


图1 科研热点预测算法框架

## 2.2 构建关键词网络

利用给定的关键词表，以关键词为节点，出现在同一个句子中的两个关键词之间形成一个链路，构成关键词网络。这个网络的权重为链路出现的频率。传统的方法会对这个网络直接建模，然而这个关键词网络具有长尾效应，即大量的以此关键词为节点的链路出现频率非常低，导致构建此网络时，模型仅关注出现频率高的关键词。与此相对的是，未来科研热点预测要捕获的关键词往往是更加新的词，与这些词相关的链路出现频率往往很低，且不能被传统模型捕捉到，导致传统模型的算法效果比较差。究其原因，是传统模型采用欧氏空间来建模，而欧氏空间不能对长尾效应数据进行有效的建模。因此，本文提出用双曲空间来解决这个问题。

## 2.3 双曲空间关键词图嵌入算法

本文用双曲空间对图中的节点进行建模，任意两个点的相似性和距离会用它们在双曲空间中的点嵌入进行计算，而不是传统的欧氏空间中的点嵌入。双曲空间有一些基本的空间模型，本文采用的是庞加莱球模型[32]。庞加莱球模型是一种更易于建模的多维空间模型，相比于欧氏空间模型，它常常仅需要少量的维度就能建模更复杂的数据。具体地说，所有在双曲空间的点被定义在一个$d$维度的单元球内，$B^d=\{x\in R^d:\|x\|<1\}$，其中$\|\cdot\|$是欧几里得范数，$d$是庞加莱球的维度，$x$是庞加莱球模型空间中的一点。

给定庞加莱球中的任意两点，PKGM算法计算它们在球内的距离，如式（1）所示：

$$d_B(x,y)=\operatorname{arcosh}\left(1+2\frac{\|x-y\|^2}{\left(1-\|x\|^2\right)\left(1-\|y\|^2\right)}\right) \quad (1)$$

随着$x$越来越接近庞加莱球的边缘，$x$和$y$之间的距离会趋向无限远，这样庞加莱球模型就能够建模复杂层次结构的数据，并能够对具有长尾效应的数据进行建模。相比于欧氏空间，庞加莱球空间具有更小的扰动性。

定义$g_x^B$是黎曼空间中测度张量，可以被用来计算庞加莱空间中的梯度，这个张量可以通过欧氏空间的张量简单表示为$g_x^B=\lambda_x^2 g_x^E$，其中$\lambda_x=\frac{2}{1-\|x\|^2}$是缩放参数，用来进行欧氏空间的梯度和庞加莱空间梯度的映射。$g_x^E=I_d$是对应的欧氏空间张量，也就是欧氏空间的点乘。

许多神经网络无法被应用到庞加莱空间中。为了解决这个问题，切空间被应用到庞加莱空间中，即通过对数和指数转换将庞加莱空间中的向量映射到欧氏空间。为了达到这个目的，PKGM算法通过对数映射和指数映射进行双曲空间和欧氏空间之间的转换。对数映射$\log_x\left(B^d\to T_xB^d\right)$被用来将$x$从庞加莱球映射到切空间的对应切向量，指数映射$\exp_x\left(T_xB^d\to B^d\right)$被用来将切空间的切向量映射到庞加莱球中的点$x$。

给定一个双曲空间$B^d$和切空间$T_xB^d$，对数映射如式（2）所示，指数映射如式（3）所示：

$$\log_x(y)=\frac{2}{\lambda_x}\operatorname{arctanh}(\|-x\oplus y\|)\frac{-x\oplus y}{\|-x\oplus y\|} \quad (2)$$

$$\exp_x(v)=x\oplus\left(\tanh\left(\frac{\lambda_x\|v\|}{2}\frac{v}{\|v\|}\right)\right) \quad (3)$$

其中，$x\in B^d$和$y\in B^d$都是庞加莱球中的点，$v\neq 0$是切空间中对应$x$的切向量，

$\log_x(y)$ 定义了从$y$映射到切空间中的$x$，$\oplus$代表莫比乌斯加法（欧氏空间中的向量加法）。

庞加莱球中两个点的欧氏空间PKGM算法如式（4）所示：

$$x \oplus y = \frac{\left(1+ < x,y > + \|y\|^2\right)x + \left(1 - \|x\|^2\right)y}{1 + 2 < x,y > + \|x\|^2 \|y\|^2} \quad (4)$$

其中，<>是点乘，$< x,x >= \|x\|^2$。

计算双曲空间中一个向量的线性变化，如式（5）所示：

$$W \otimes x = \exp_O\left(W \log_O(x) + b\right) \quad (5)$$

其中，O是庞加莱球的原点，$W$和$b$是神经网络中的权重。

根据这些定义，图中两个点$u$和$v$在双曲空间中存在边的概率如式（6）所示：

$$P\left((u,v)=1 \mid h_u^l, h_v^l\right) = \frac{1}{e^{\frac{\left(d_B(h_u^l, h_v^l)^2 - r\right)}{t}} + 1} \quad (6)$$

其中，$h_u^l$是点$u$在双曲空间中的点嵌入，$d_B(\cdot,\cdot)$是对应的庞加莱球，$r,t>0$是对应的超参数。

至此，可以用梯度下降法对图中每个点的点嵌入进行迭代优化，使图中存在边的两个点的概率最高，而图中不存在边的两个点的概率最低。在这个过程中，双曲空间可以使点嵌入的效果更好。算法优化结束后，可以根据图中未连接成边的点的距离判断可能出现的新科研热点，即距离越近的两个点之间的边越有可能是潜在的科研热点。

PKGM算法使用了如下的超参数：学习速率为$1\times10^{-5}$，降维维度$d$=16，$r$=2。这些超参数都是基于验证集合选取的。PKGM算法流程如算法1所示，第1行对庞加莱球模型中的点进行初始化，第2～6行求解庞加莱球中的点嵌入，其中第3行随机采样一条边进行优化，第4~5行对这条边的概率进行最大化。

**算法1** PKGM算法

**输入：**图$G(V, E)$

庞加莱球的维度：$d$

**输出：**图$G$中每一个点在庞加莱球模型中的点嵌入

1. 初始化：随机初始化图$G$中每个点在庞加莱球模型中的点嵌入
2. while 损失函数并不收敛
3. 随机采样图中一条边$(u, v)$
4. 根据式（1）计算两点在双曲空间中的距离
5. 根据式（6）优化$u$和$v$在双曲空间中的点嵌入，使它们的距离减小
6. end while

## 3 实验

### 3.1 数据集和评测指标

本文采用专用实体识别工具PubTator[33]从生物医学论文数据库PubMed中获得了1940年以来所有被PubTator处理过的科技论文摘要和关键词。不是所有论文都包含关键词信息，采用PubTator对论文进行处理，获得专用实体，以这些实体为关键词信息，共获得33 548 974篇论文。PubTator为每一篇论文标注出了多个关键词，平均每篇论文12个关键词，分别描述论文的类别、研究方法、研究方向、研究成果等。利用这些关键词组成一个关键词表，对所有论文的句子进行遍历。若一个句子中有两个关键词，这两个关键词之间就形成一个链路，即构成关键词网络的边。构建关键词网络的算法流程如**图2**所示。具体的是在33 548 974篇论文中，除去重复关键词后，

剩余392 522 996个关键词，随机抽取了10 000个关键词构建网络。在此网络中有34个节点由于与最大子图不联通被去除，最终获得了一个有9 966个关键词节点和18 976条链路的关键词网络。

如前所述，通过对关键词网络边的研究预测不同关键词之间是否存在边，就能够预测未来的科研热点。具体地说，本文研究就变为一个对网络边进行预测的研究，即预测未来可能出现的网络边[30]。实验将整个数据分成测试集、训练集和验证集，并且根据训练集大小，分为实验1（85%训练集）和实验2（60%训练集）。数据集的统计信息见**表1**。

实验选取7个基准算法来比较PKGM算法的效果，具体如下。

- Euclidean算法：欧氏空间嵌入算法是传统的数据降维算法，它将数据降维到欧氏空间进行后续的预测，本文用L2损失函数对欧氏空间进行降维。
- MLP算法：多层感知机（multi-layer perceptron，MLP）算法利用多层神经网络对目标函数进行非线性逼近。
- GCN算法：图卷积网络（graph convolutional network，GCN）算法额外考虑了数据中的图结构，同时通过对图和点向量进行降维来获得点的特征向量，从而进行连接预测。
- GAT算法：图注意力网络（graph attention network，GAT）算法通过注意力机制对图和点向量进行降维，从而进行连接预测。
- HNN算法：双曲神经网络（hyperbolic neural network，HNN）算法是在双曲空间中实现的神经网络算法，此算法比起传统的欧氏空间神经网络算法能更好地对长尾效应数据进行建模。
- HGCN算法：HGCN算法通过增加曲率参数被推广到双曲空间，这个算法在节点分类和边的连接预测上比GCN算法的效果更好。
- HGNN算法：双曲图神经网络（hyperbolic graph neural network，HGNN）算法以非卷积的形式对图结构进行建模。

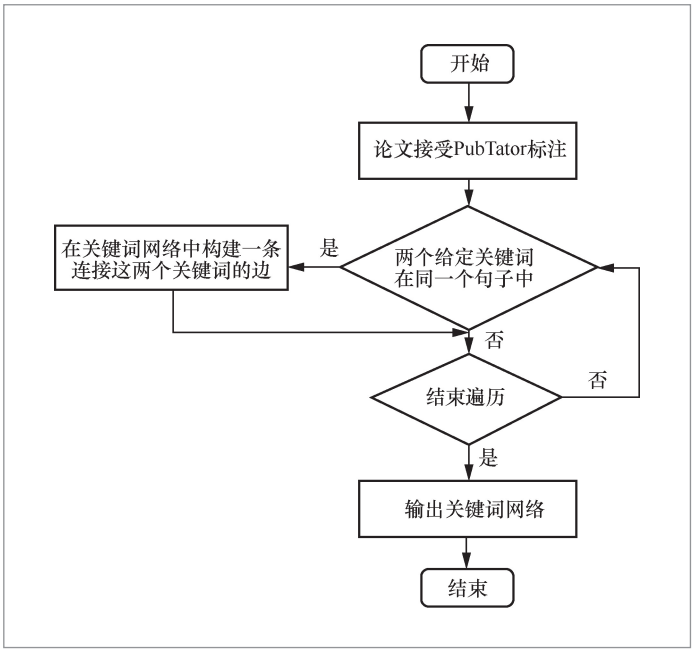


图2 构建关键词网络的算法流程

表1 数据集的统计信息

| 实验 | 测试集/条 | 训练集/条 | 验证集/条 |
|---|---|---|---|
| 实验1 | 1 897 | 16 130 | 949 |
| 实验2 | 5 692 | 11 386 | 1 898 |

上述算法中的Euclidean算法、MLP算法、GCN算法和GAT算法是欧氏空间嵌入算法，其余算法及PKGM算法是双曲空间嵌入算法。

实验将接受者操作特征曲线下面积（area under the receiver operating characteristic curve，AUROC）和平均精度（average precision，AP）作为算法的评价指标[11]。AUROC和AP在最佳情况下

趋近于1.0，而在随机的预测下趋近于0.5。AUROC和AP越高，说明算法对网络边的预测越准确。

### 3.2 关键词网络特性验证

在检验算法的有效性前，先对本文的假设关键词网络存在长尾效应进行验证。关键词数量与论文数量如**图3**所示，从**图3**中可以计算得到90%以上的关键词出现的论文篇数小于13，这样可以判断关键词网络存在明显的长尾效应。而这个长尾效应往往不能被传统模型所处理[13]，这也为本文提出的双曲空间建模提供了实验基础。

### 3.3 双曲空间嵌入算法与欧氏空间嵌入算法比较

本文用关键词网络中对连接的预测进行科研热点预测的验证。双曲空间嵌入算法与欧氏空间嵌入算法对比实验结果见**表2**，可以看到双曲空间嵌入要显著好于欧氏空间嵌入。比如实验1将85%数据作为训练集时，双曲空间嵌入算法获得了0.8822的AUROC和0.8906的AP，而基准算法中效果最好的欧氏空间嵌入算法也只有0.8180的AUROC和0.8389的AP。基于图神经网络的GCN算法和GAT算法的效果则更差，要比双曲空间嵌入算法至少降低13.5%。在实验2使用更少的数据集作为训练集时，双曲空间嵌入算法也同样获得了很大的提升，比如双曲空间嵌入算法获得了0.8751的AUROC和0.8857的AP，而基准算法中效果最好的欧氏空间嵌入算法也只有0.8062的AUROC和0.8276的AP。由此可见，双曲空间嵌入算法要比欧氏空间中所有的基准算法都要好，并且效果非常明显，AUROC至少上升7.3%，AP增加5.8%以上。同时，还可以看到基于嵌入的算法要优于基于图神经网络的算法，这也进一步验证了嵌入算法的有效性。

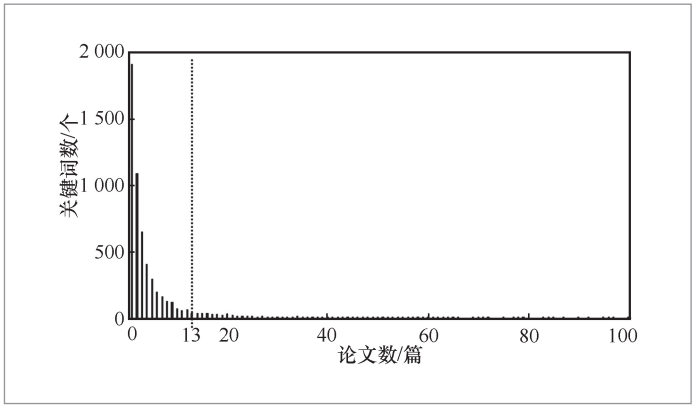


图 3 关键词数量与论文数量

### 3.4 双曲空间嵌入算法与双曲空间图神经网络算法比较

双曲空间嵌入算法与双曲空间图神经网络算法对比实验结果见**表3**，可以发现双曲空间嵌入算法依然好于所有的双曲空间图神经网络算法。如使用85%训练集时，双曲空间嵌入算法获得了0.8822的AUROC和0.8906的AP，而双曲空间图神经网络算法中最好的HGNN算法的AUROC和AP分别为0.7865和0.8264。也就是在实验评价指标AUROC和AP上，双曲空间嵌入算法比双曲空间图神经网络算法分别提升至少10.8%和7.2%。在较小的60%训练集上的实验也有相同的结论。同时还发现，虽然双曲图神经网络算法不如双曲空间嵌入算法效果好，但仍然好于欧氏空间嵌入算法，具体内容见**表2**。

上述实验结果表明，在科研热点预测上，PKGM算法整体上要好于欧氏空间算法，因为关键词网络的长尾效应更适合用双曲空间建模。PKGM算法解决了关键词网络的长尾效应问题，不但能够关注到近期的热点话题，还能预测到未来的科研热点。

表 2 双曲空间嵌入算法与欧氏空间嵌入算法对比实验结果

| 算法 | 实验1 | | | | 实验2 | | | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | 增加 | AP | 增加 | AUROC | 增加 | AP | 增加 |
| PKGM | 0.8822 | – | 0.8906 | – | 0.8751 | – | 0.8857 | – |
| Euclidean | 0.8180 | 7.3% | 0.8389 | 5.8% | 0.8062 | 7.9% | 0.8276 | 6.6% |
| MLP | 0.4989 | 43.4% | 0.4995 | 43.9% | 0.5003 | 42.8% | 0.5001 | 43.5% |
| GCN | 0.7222 | 18.1% | 0.7448 | 16.4% | 0.7406 | 15.4% | 0.7329 | 17.3% |
| GAT | 0.7590 | 14.0% | 0.7705 | 13.5% | 0.7575 | 13.4% | 0.7674 | 13.4% |

表 3 双曲空间嵌入算法与双曲空间图神经网络算法对比实验结果

| 算法 | 实验1 | | | | 实验2 | | | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | 增加 | AP | 增加 | AUROC | 增加 | AP | 增加 |
| PKGM | 0.8822 | – | 0.8906 | – | 0.8751 | – | 0.8857 | – |
| HNN | 0.6596 | 25.2% | 0.6466 | 27.4% | 0.6817 | 22.1% | 0.6645 | 25.0% |
| HGCN | 0.7764 | 12.0% | 0.7930 | 11.0% | 0.7750 | 11.4% | 0.7943 | 10.3% |
| HGNN | 0.7865 | 10.8% | 0.8264 | 7.2% | 0.7787 | 11.0% | 0.8162 | 7.8% |

## 3.5 算法性能分析

关键词网络具有长尾效应，因此它不能被应用到欧氏空间中，并且有更复杂且不直观的数学模型，但是其在双曲空间算法中所用的时间复杂度和空间复杂度并不比欧氏空间算法大。实验观察到双曲空间算法的运行速度与欧氏空间嵌入算法相近，因为它们的复杂度主要取决于网络中边的数量。更值得注意的是，双曲空间嵌入算法的空间复杂度要远小于欧氏空间嵌入算法，实验中算法的超参数和维度是通过实验选定的。

PKGM算法效果与学习速率超参数的关系如**图4**所示，可以看到PKGM算法对学习速率这个超参数非常稳定，学习速率超参数为0.005～0.350，AUROC和AP基本不变，这充分证明了PKGM算法的鲁棒性。

PKGM算法效果和双曲空间维度的关系如**图5**所示，可以看到随着双曲空间

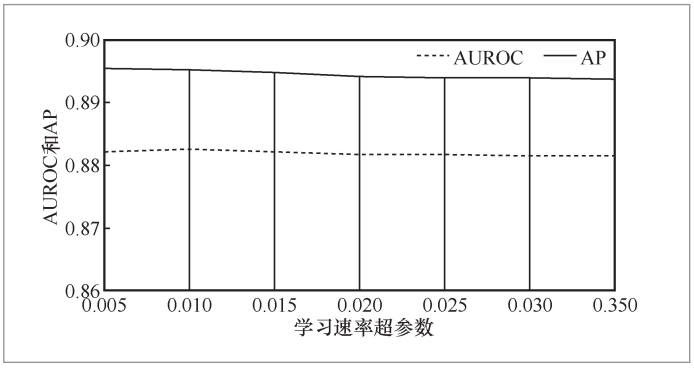


图 4 PKGM 算法效果与学习速率超参数的关系

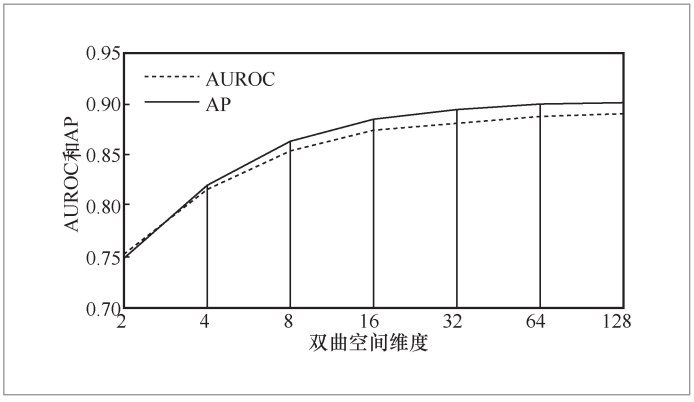


图 5 PKGM 算法效果与双曲空间维度的关系

维度的增加，AUROC和AP也在增加。从图5可以发现曲线斜率起初比较大，也就是AUROC和AP值变化较快，当维度为16时，曲线斜率变小，AUROC和AP数值变化变缓。综合考虑算法空间复杂度等方面因素，PKGM算法将16作为双曲空间的维度，而欧氏空间维度为200。双曲空间能够通过更好地利用空间的位置节省维度，在较小的维度中嵌入更多的信息，继而有效地模拟长尾效应中的罕见词。空间复杂度的减少，有助于在有限的空间资源内计算和存储更多的论文数据，更能适应双曲空间对科研热点预测。

## 4 结束语

本文提出了一种新的PKGM算法来预测科研热点。首先，利用论文关键词来构建一个关键词网络，然后将这个网络图嵌入双曲空间，通过计算双曲空间中两个节点的距离来判别两个节点之间存在边的概率，从而预测出未来科研热点。实验发现，PKGM算法比7种基准方法效果有显著提高，与效果最好的欧氏空间嵌入算法相比，有7.3%的AUROC和5.8%的AP提升；与双曲空间图神经网络算法相比，有10.8%的AUROC和7.2%的AP提升。其主要原因是双曲空间以指数形式进行建模，可以把数据点更均匀地分布于低维空间，有足够的空间来表示罕见的数据点。对于出现次数很多的数据点，指数运算的逆运算即对数运算对次数的降低就较大；而对于出现次数很少的数据点，指数运算的逆运算即对数运算对次数的降低就较小。这样就可以大大缩小数据点出现次数的差距，利用均匀的空间来表示出现次数多和出现次数少的数据点，这些空间可以抵消随机噪声对这些数据点的干扰，能更好地处理长尾效应的数据。

未来有3个研究方向：在关键词网络中加入文本信息，通过共同训练获得更高质量的节点表示；在关键词网络中加入作者、期刊名等数据，构建异质网络以获得更丰富的图表示；在关键词网络中加入时序信息，通过不同时间点关键词的差异获得更精准的关键词网络。

## 参考文献：

[1] YANN L, YOSHUA B, GEOFFREY H. Deep learning[J]. Nature, 2015, 521: 436-444.

[2] ATTARDI G. DeepNL: a deep learning NLP pipeline[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. [S.l.:s.n.], 2015: 109-115.

[3] VOULODIMOS A, DOULAMIS N, DOULAMIS A, et al. Deep learning for computer vision: a brief review[J]. Computational Intelligence and Neuroscience, 2018: 1-13.

[4] WANG C, BLEI D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.:s.n.], 2011: 448-456.

[5] ZHANG S, TONG H H, XU J J, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(11): 1-23.

[6] WALLACH H M. Topic modeling: beyond bag-of-words[C]//Proceedings of the 23rd International Conference on Machine Learning. [S.l.:s.n.], 2006: 977-984.

[7] BHAT H S, HUANG L H, RODRIGUEZ S, et al. Citation prediction using diverse features[C]//Proceedings of 2015 IEEE International Conference on Data Mining Workshop. Piscataway: IEEE Press, 2015: 589-596.

[8] MORADI B, PARENT M C, WEIS A S, et al. Mapping the travels of intersectionality scholarship: a citation network analysis[J]. Psychology of Women Quarterly, 2020, 44(2): 151–169.

[9] LIU L S, YU D J, WANG D J, et al. Citation count prediction based on neural hawkes model[J]. IEICE Transactions on Information and Systems, 2020, 103(11): 2379–2388.

[10] JEONG C, JANG S, PARK E, et al. A context–aware citation recommendation model with BERT and graph convolutional networks[J]. Scientometrics, 2020, 124(3): 1907–1922.

[11] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]// Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining. [S.l.:s.n.], 2016: 855–864.

[12] CHAMI I, YING R, RE C, et al. Hyperbolic graph convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 4869–4880.

[13] LEE J, LEE J N, SHIN H. The long tail or the short tail: the category–specific impact of eWOM on sales distribution[J]. Decision Support Systems, 2011, 51(3): 466–479.

[14] YANG C, LIU Z, ZHAO D, et al. Network representation learning with rich text information[C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence. [S.l.:s.n.], 2015: 2111–2117.

[15] YUE X, WANG Z, HUANG J G, et al. Graph embedding on biomedical networks: methods, applications and evaluations[J]. Bioinformatics, 2020, 36(4): 1241–1251.

[16] ASHOOR H, CHEN X W, ROSIKIEWICZ W, et al. Graph embedding and unsupervised learning predict genomic sub–compartments from HiC chromatin interaction data[J]. Nature Communications, 2020, 11: 1173.

[17] PEROZZI B, AL–RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 701–710.

[18] TANG J, QU M, WANG M Z, et al. LINE: large–scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. [S.l.:s.n.], 2015: 1067–1077.

[19] KIPF T N, WELLING M. Semi–supervised classification with graph convolutional networks[C]//Proceedings of the International Conference on Learning Representation. [S.l.:s.n.], 2017: 1–8.

[20] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning. [S.l.:s.n.], 2017: 1263–1272.

[21] BOSCAINI D, MASCI J, RODOLA E, et al. Learning shape correspondence with anisotropic convolutional neural networks[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. [S.l.:s.n.], 2016: 3189–3197.

[22] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5115–5124.

[23] KAWAMOTO T, TSUBAKI M, OBUCHI T. Mean–field theory of graph neural networks in graph partitioning[C]// Proceedings of the 32nd International Coference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 4366–4376.

[24] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]//Proceedings of the 28th International Conference on World Wide Web. [S.l.:s.n.], 2019: 2022–2032.

[25] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//Proceedings of the

International Conference on Learning Representations 2018. [S.l.:s.n.], 2018: 1-8.
[26] TIFREA A, BECIGNEUL G, GANEA O E. PoincaréGloVe: hyperbolic word embeddings[C]//Proceedings of the International Conference on Learning Representations 2019. [S.l.:s.n.], 2019: 1-8.
[27] DHINGRA B, SHALLUE C J, NOROUZI M, et al. Embedding text in hyperbolic space[C]//Proceedings of the Association for Computational Linguistics. [S.l.:s.n.], 2018: 59-69.
[28] GANEA O, BECIGNEUL G, HOFMANN T. Hyperbolic neural networks[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. [S.l.:s.n.], 2018: 5345-5355.
[29] LIU Q, NICKEL M, KIELA D. Hyperbolic graph neural networks[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems. [S.l.:s.n.], 2019: 8230-8241.
[30] ASLAN S, KAYA M. Topic recommendation for authors as a link prediction problem[J]. Future Generation Computer Systems, 2018, 89: 249-264.
[31] SONG W Z, CHEN H X, LIU X Y, et al. Hyperbolic node embedding for signed networks[J]. Neurocomputing, 2021, 421: 329-339.
[32] NICKEL M, KIELA D. Poincaré embeddings for learning hierarchical representations[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. [S.l.:s.n.], 2017: 6338-6347.
[33] WEI C H, KAO H Y, LU Z Y. PubTator: a web-based text mining tool for assisting biocuration[J]. Nucleic Acids Research, 2013, 41(W1): 518-522.

作者简介

**戴筠**(1966- ),女,上海大学副教授,主要研究方向为数据挖掘和机器学习。

**收稿日期:** 2022-01-18
**通信作者:** 戴筠, daijun@staff.shu.edu.cn