**异构学术网络中学者影响力的个性化预测**

**Abstract:**

然而，目前大多数关于学者影响力预测的文献忽略了几个重要事实，即学术网络的时间演变、不同学者影响力的不同动态以及不同学术实体之间的相互影响。

**1.Introduction：**

快速发展的研究技术导致学术论文数量的显著增长，同时也加速了科学的进步。然而，学术数据的多样性和海量性也使得追踪相关文章或有影响力的学者变得越来越耗时，因此，评估和预测科学影响性是很重要的。

借助科学影响评价，可以获得当前的影响，并有依据地解决许多问题：求职申请，研究基金决定，奖励。但是当前的很多问题不能通过当前的科学影响判断。比如现在很多期刊都是合著的，因此，寻找潜在的合作者对学术成果至关重要。也要可以预测一个研究人员对未来的影响，可以用来预测研究趋势，选择研究候选人。

**当前研究的问题**：

1.学术网络的发展使得以前的一些预测方法不再准确。

一般来说，**学术网络**包括几种类型的实体和链接，如文章、作者、地点、引用关系、合著和出版关系等，他们都不是一成不变的。这些现象体现了学术网络的复杂性和时变性。为了更好的预测学者的影响力，学术网络的变化是要考虑的。

2. 当前文献中存在的第二个问题是，他们倾向于使用统一的方法来预测不同学者的影响。然而，实际上不同的学者在他们的学术生涯中经常经历不同的轨迹。例如，一些研究人员的工作在发表后不久就被大多数人认可，同时也有一些工作需要时间来发现它们的意义。

本文在前人研究的基础上，针对存在的问题，提出了PePSI预测学者科学影响力的方法。我们的方法应用异构网络拓扑，这代表了学术网络的复杂性。我们首先根据不同的影响动态将学者分为不同的类型。然后在学术网络中应用不同的功能来表示它们的趋势。为了捕捉学术网络的动态，我们构建了异构的时间学术网络。结合上述趋势函数，通过在异构时间学术网络中使用随机游走算法考虑学术实体之间的相互影响，可以预测学者的未来影响

**本文的创新点：**

1.对科学影响预测的新见解

2.针对不用学者的个性化预测

3.用短期的数据得到高性能的预测

**2.相关工作**

目前主要有两种：第一种是基于动态引用来预测，另一种主要是基于利用学术网络中的重要性排序算法。

对科学影响的开创性研究是通过引用计数来评价期刊质量。遵循这一方向，研究人员提出了几个重要的基于**引用计数**的评价指标，这些指标考虑了很多的事实，至今仍广泛应用于学术实体的重要性排名：h-index是其中的代表。当然这种基于引用的有很多缺点：对新发布的文章不公平，还有一些人为的对引用的操纵。为了克服这些缺点，在美国，一些研究人员试图利用网络拓扑来评估科学的影响。

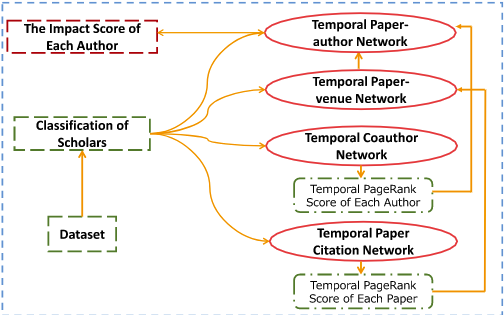
PageRank和HITS算法最初是为了在同构网络结构中对搜索引擎的结果进行排序而提出的，而现在它们被广泛用于影响评估。因此，学者们提出了一系列基于修改PageRank和HITS的方法来对学术实体在异构学术网络中的重要性进行排序。Wang等人[ 24 ]考虑时间信息，结合PageRank和HITS算法评估学者的影响。

对学者目前的影响可以通过上述工作来计算，但对他们未来的影响仍有待探索。为了预测未来的影响，目前的大部分工作可以分成两大类。第一类方法是通过分析学术数据(例如引用记录)，引用不仅在评价影响方面起着重要作用，而且对影响预测也有重要意义。考虑到引文的动态性，研究者们提出了许多研究来预测未来的影响。Wang等人[30]发现动态引用的普遍时间模式，并利用它来预测个别论文的影响。Cao et al .[ 32 ]使用历史引用数据来预测不同学科的文章的科学影响。

然而，一个重要的事实被现存的方法忽略了。也就是说，不同学者的不同的动态影响力。目前，大多数现有的著作都使用统一的方法来预测不同学者对的影响。因此，不同类型的学者在预测他们未来的影响时，应该考虑到他们的不同特点，而不是把他们看作是相同的。受这些现象的启发，我们提出了一种新的学者影响力预测方法，将学者分为不同类型，同时考虑异构学术网络的演化。

**3.PePSI的设计**

为了预测学者的未来影响力，我们首先分析了以学者为中心的数据集，我们的研究基于微软学术网络。为了捕捉不同学者的各种动态，我们首先根据他们的引用动态将学者分为适当的类型，然后考虑到学术网络的动态性，提出了构建异构时态学术网络中针对不同学者的预测方法。PePSI的结构如下图所示：



**A.对不同学者进行分类**

学者们先前的信息是预测他们未来影响的基础，即在预测学者的影响时，历史引文和出版次数是非常重要的。我们认为，学者们实际上在整个学术生涯中表现出不同的趋势，例如，出版物的数量、总的引用次数。我们从MAG数据集中随机选取了计算机科学领域的100名学者，如图3所示，每个学者在其职业生涯的每一年的引文动态都有很大的差异。除了引用的动态，每位学者在研究生涯中每年的出版物数量也各不相同。学者之间的这些差异导致了他们未来不同的学术成就。所以我们首先将不同的研究者进行分类。

本篇文献中我们使用引用动态性来区分不同种类的研究者。我们从MAG数据集获得的子数据集由计算机科学领域不同学术年龄的学者组成。为了减轻学术年龄对学者的影响，我们选择了学术年龄在一个周期的学者，然后，我们从数据集中选择在不同学术年龄具有完整学术生涯的学者，以区分不同类型的学者。

首先我们定义了学者的分类。一般的，我们将学者分类任务转化成找到学者的最相似的动态引用，并且利用这个区预测影响力。比如说给一个学者A，如果她和另一组学者的动态引用很接近，那么他们就是一类的。学者分类的步骤可以形式化如下:

1.拿到A的前n年的引用记录xn和数据库中前n年的学者动态引用记录yn，随后计算匹配误差。

2.误差可以表示为：

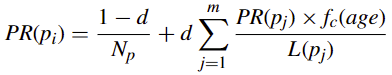
3.根据误差我们可以找到与学者A最相近的学者集合S（y）

**B时间学术网络中的影响预测**

对于每种学者，我们构建了四个学术网络，即时间论文引文网络、时间合著者网络、时间论文地点网络和时间论文作者网络。在这些学术网络下，我们应用随机游走算法来计算学者的影响力。

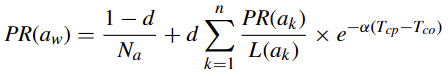
1）时间论文引用网络：

通过对MAG数据集的分析，我们发现最近发表的论文在未来会比以前的论文得到更多的引用。对于每一类学者，我们将每条边的权重设定为其在时间引文网络中动态引用的函数。然后通过论文之间的时间加权引用关系计算论文的影响力，计算公式可以表示为:



2）时间合著网络：

研究者认为，合作者的能力可以影响相应论文的质量，进而影响学者自身的影响。我们假设，如果一个学者最近与另一个学者合作，那么他或她将来更有可能与他合作。时间合著者网络包含学者和作者之间的边，如果他们在同一篇论文中合作的话。我们通过学者之间的加权合作关系来计算作者的影响力，计算公式可以表示为:



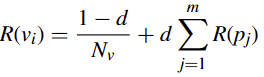
3）论文期刊网络：

论文的质量和期刊的质量是相互影响的，所以我们建立论文期刊网络来评估论文的影响。带时间权重的论文和期刊的出版关系为：

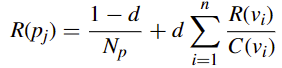


为了得到相互强化的过程，我们设置了两个更新角色来获得论文的影响。最初，每个期刊的值被设置为1/Nv，其中Nv是期刊的总数，根据等式( 1 )的计算作为论文的初始分数。

第一步：根据论文得分计算期刊价值;



第二部：通过期刊得分更新论文得分：



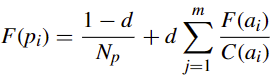
C（Vi）是期刊上刊登的论文总数。

持续迭代步骤1,2知道聚合。一般迭代两次就好了。

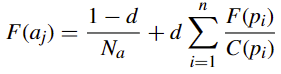
4）论文作者网络

论文和作者的影响力也是互相有影响的。因此，我们假设最近发表的论文能够更好地预示学者们未来的影响。为了描述论文和作者之间的相互影响，我们建立了更新规则来计算它们的影响。初始的时候，我们用等式（5）计算论文的初始值，用等式（2）计算作者的初始值。

第一步：通过相关的作者计算论文的值：



第二步：计算作者的值：



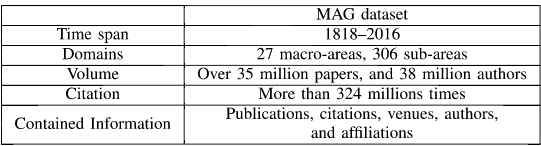
第三步：重复上述步骤，直到论文和作者的值收敛

**4.实验和结果：**

由于不存在预测影响的实际数值，所以采用常用引用计数作为实际。

A MAG数据集

包含文章标题，关键字，合著者，出版时间，期刊信息，作者附属关系和引用关系。



为了捕捉整个职业生涯中学者的显著特征，我们选择了计算机科学领域的学者，他们的学术生涯在数据集中是完整的

B 对比方法

1.PageRank

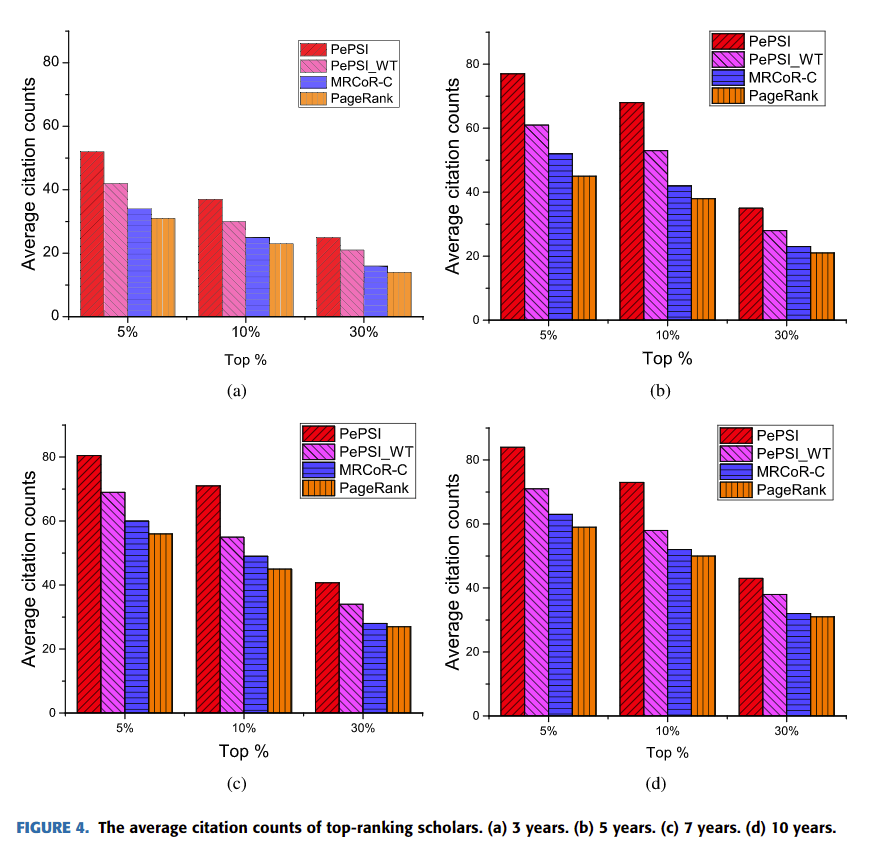
2.MRCoR

3. PePSI\_WT.：将PePSI中时间参数去掉。

C 顶尖学者的预测

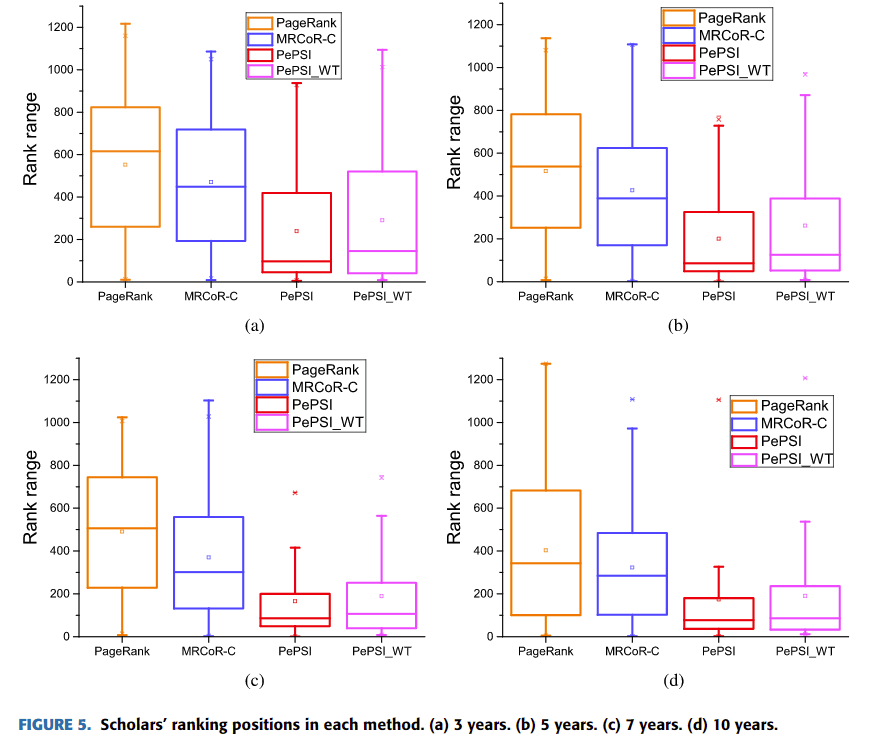
一般来说，学者们先前的信息对预测他们未来的影响是很有帮助的，使用的时间信息越长，预测结果就越准确。然而，较长的时间信息需要更多的计算，因此提高了计算成本。因此，理想状态是使用相对较短的时间周期和较高的预测精度。所以我们要找到到底需要多长的时间。

我们首先比较了每种方法对不同时间跨度的顶级学者的预测能力。们比较了不同时期各种方法对一流学者的平均引文数，如下图所示：



显然，在所有基线方法中，使用PePsi得到的顶级学者的平均引用次数最高。显然还可以看到，以前信息的时间越长，顶级学者的平均引用次数越高。

此外，我们根据引用次数用每种方法比较了前100名学者的排名。学者的排名位置可以用每种方法来计算基于他们的影响分数获得，其中学者从影响最大到最小排列。由于本文选择引用数是根据实际的，我们假设每种方法中引用数排名前100位的学者在这个方法中的排名越高，那么这个方法识别有影响力的学者就越有效。



此外，从这些结果中，我们可以看出，预测时间越长，顶级学者在每种方法中的排名位置就越好。

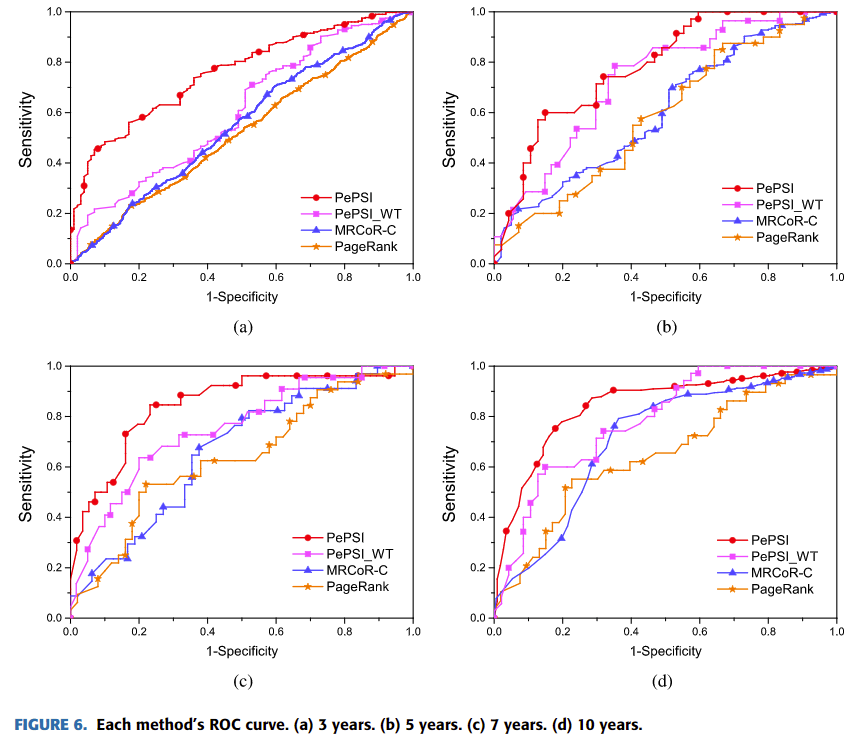
D 对所有学者的预测

除了探索识别顶级学者的能力之外，还研究了如何预测普通学者的影响的能力。

我们首先区分学者的不同影响。排名靠前的被视为积极实体，靠后的被视为消极实体。

一般来说存在四种分类结果：靠前的学者被分到了积极实体中，靠前的实体被分到消极实体，靠后的被分到积极实体中，靠后的被分到消极实体中。

利用这四个分类结果，可以获得每种方法的ROC曲线。



E 预测跨度

除了我们在上面小节中探索的预测性能，我们还通过每种方法研究预测周期。此外，在所有方法中，PePSI使用的过去信息的时间最短。通过所有用于预测的时间段，10年的性能是最好的。然而，使用10年的信息比使用5年的信息需要更多的计算，并且改进并不显著。因此，我们利用5年的信息对各种方法的预测周期进行了研究。

为了发现每种方法的预测时间段，我们分别在2006年和2014年进一步调查预测结果。如上图示，2010年优秀学者的预测结果要比2006年好得多，优秀学者的平均引用数和排名都较高。除了对优秀学者的预测外，还探讨了对整体学者影响预测的表现。

