**Meta Structure: Computing Relevance in  
Large Heterogeneous Information Networks**

# Abstract

异构信息网络（HIN）是一种图形模型，其中对象和边缘用类型注释。大型复杂数据库（如YAGO和DBLP）可以建模为HIN。HIN中的一个基本问题是计算两个HIN对象之间的接近度或相关性。相关性度量可用于各种应用程序，包括实体解析，推荐和信息检索。一些研究已经研究了HIN信息在相关性计算中的使用，然而，它们中的大多数仅利用简单的结构（例如路径）来测量对象之间的相似性。在本文中，我们建议使用元结构，它是对象类型的有向无环图，边缘类型在其间连接，以测量对象之间的接近度。元结构的优势在于它可以描述两个HIN对象之间的复杂关系（例如，DBLP中的两篇论文共享相同的作者和主题）。我们基于元结构开发了三个相关度量。由于这些测量的计算复杂性，我们进一步设计了一种具有数据结构的算法，以支持它们的评估。我们对YAGO和DBLP的广泛实验表明，基于元结构的相关性比现有技术方法更有效，并且可以有效地计算。

# INTRODUCTION

异构信息网络（HIN），如DBLP [8]，YAGO [15]，DBpedia [1]和Freebase [2]，最近受到了很多关注。这些图形数据源包含大量相互关联的事实，它们用于促进有趣知识的发现[5,7,12,13]。图1示出了HIN，其描述了不同类型的实体（例如，作者，论文，场所和主题）之间的关系。例如，Jiawei Han（a2）撰写了一篇VLDB论文（p2; 2），其中提到了“有效”这一主题（t3）。

给定两个HIN对象a和b，对其相关性的评估具有根本重要性。这量化了a和b之间的接近程度。在图1中，Jian Pei（a1）和Jiawei Han（a2）具有较高的相关性得分，因为他们在同一地点（KDD）发表了关键词“挖掘”的论文。相关性发现它在信息检索，推荐和聚类中的应用[18,22]：研究人员可以检索与DBLP中的主题和场所具有高度相关性的论文;在YAGO中，相关性有助于提取与特定导演关系密切的演员。作为另一示例，在实体解析应用中，可以识别具有高相关性得分的复制HIN对象对（例如，HIN中的两个不同对象指向相同的真实世界人），并从HIN中移除。

## Prior works

为了测量两个图形对象之间的相关性，提出了基于邻域的度量，如共同邻域和Jaccard系数[9]。其他基于对象之间随机游走的图论方法包括Personalized PageRank [3]和SimRank [6]。这些度量不考虑HIN中的对象和边缘类型信息。为了处理这些信息，最近提出了元路径的概念[7,18]。元路径是一系列对象类型，其间具有边类型。图2（b）示出了元路径P1，其表明两个作者（A1和A2）通过他们在相同场所（V）中的出版物相关联。另一个元路径P2表示两位作者撰写的论文包含相同的主题（T）。基于元路径，已经提出了几种相关性度量，例如PathCount，PathSim和Path Constrained Random Walk（PCRW）[7,18]。已经证明这些措施比不考虑对象和边缘类型信息的措施要好。

## Meta structures

我们提出了一个名为元结构的新概念来描述两个图形对象的关系。这本质上是对象和边缘类型的有向无环图。图2（b）说明了一个元结构S，**它描述了两个作者是否相关，如果他们在同一地点发表论文，并且也提到了相同的主题**。元路径（例如，P1或P2）是元结构的特殊情况。然而，元路径未能捕获可以由元结构（例如，S）方便地表达的这种复杂关系。我们的实验还表明元结构比元路径更有效。

我们为元结构提供了一个合理的定义。这不是直截了当的，因为元结构可能很复杂。然后，我们提出了三种基于元结构的相关性度量。这些度量在计算相关性的方式上有所不同。给定元结构S，StructCount评估与S匹配的子图的数量;结构约束子图扩展（SCSE）模拟受S限制的子图扩展过程;偏置结构约束子图扩展（BSCSE）是StructCount和SCSE的推广。

这些新措施的挑战在于其高计算成本。通常，评估这些度量需要在HIN上进行子图匹配操作。在包含数百万个物体和边缘的典型HIN（例如，YAGO）中，这可能非常昂贵。此外，应用程序（例如，聚类）可能需要计算许多对象对的相关性。因此，重要的是确保能够有效地评估这些相关性措施。为了应对这一挑战，我们设计了一个具有两个数据结构（称为Compressed-ETree和i-LTable）的递归遍历算法，以提高相关性计算的效率。

为了验证我们的方法，我们对YAGO和DBLP进行了大量实验。结果表明，我们的三个元结构度量在表达相关性方面比基于元路径的方法更有效。我们的算法还可以在大图上有效地计算元结构相关性，从而产生与元路径度量类似的运行时成本。

本文的其余部分如下。我们在第2节中描述了HIN模型并总结了现有的基于元路径的方法。我们在第3节介绍元结构。然后，我们在第4节中基于元结构定义相关性度量。我们开发了一个递归算法和两个数据结构，以便于在第5节中计算相关性度量。第6节介绍了我们的实验结果。我们在第7节结束了我们的研究。



