HeteClass：基于元路径的转换框架

异构信息网络中对象的分类

Mukul Gupta [\*](https://translate.googleusercontent.com/translate_f#1) ，Pradeep Kumar，Bharat Bhasker

*管理勒克瑙* ， *印度* *的印度理工学院*

在异构信息网络中使用标记和未标记对象的转换分类

知识提取是一个有趣和具有挑战性的问题。 大多数现实世界的网络都是

其自然环境中的异质性和传统的均匀网络分类方法，

作品不适合异构网络。 在异构网络中，各种元路径

对要进行分类的目标类型的对象进行分类

任务更具挑战性。 每个元路径的语义将导致分类的不同精度，

阳离子。 因此，元路径的权重学习需要同时利用其语义

通过加权组合。 在这项工作中，我们提出了一个新的基于元路径的框架，HeteClass

目标类型物体的转换分类。 HeteClass探讨了给定网络的网络模式，

工作，也可以结合领域专家的知识来生成一组元路径。 该

HeteClass提出的基于正则化的权重学习方法有效计算权重

对称以及网络中的非对称元路径，生成的权重是一致的，

帐篷与现实世界的理解。 使用学习的权重，均匀的信息网络

通过加权组合形成目标类型对象，并且转换分类为

形成。 所提出的框架HeteClass是灵活的，以利用任何合适的分类算法

转移分类，可以应用于异构信息网络，

工作模式。 实验结果表明HeteClass对未标记分类的有效性

使用现实世界数据集的异构信息网络中的对象。

©2016 Elsevier Ltd.保留所有权利。

**介绍**

信息网络是一种最自然的代表性方式，

真实世界的实体/对象及其关系。 在这些网络 -

作品，现实世界的对象和他们的关系被表示

作为节点和链接/边缘。 传统网络是

本质上是同质的，因为这些网络由sin-

[对象GLE类型，一个类型的关系（](https://translate.googleusercontent.com/translate_f#17) [孙，韩，燕，](https://translate.googleusercontent.com/translate_f#17)

[羽，与吴，2011](https://translate.googleusercontent.com/translate_f#17) [）。](https://translate.googleusercontent.com/translate_f#17) [然而，大多数真实世界的信息网络 -](https://translate.googleusercontent.com/translate_f#17)

[作品在其天然环境异质性（](https://translate.googleusercontent.com/translate_f#16) [古普塔，库马尔](https://translate.googleusercontent.com/translate_f#16)

[＆Bhasker，2015;](https://translate.googleusercontent.com/translate_f#16) [石，岗，黄，宇，吴和2014年](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [Heteroge-](https://translate.googleusercontent.com/translate_f#16)

新的信息网络由多种类型的ob-

对象和/或关系。 [图。1](https://translate.googleusercontent.com/translate_f#2) （a）和（b）

分别是信息网络和信息网络。

[图。1](https://translate.googleusercontent.com/translate_f#2) （a）显示了均匀信息网的片段，

作者的作品，由作者关系相互关联，

\* 通讯作者：信息技术与系统，印度理工学院

管理勒克瑙，印度。

[*E-mail地址：*](mailto:pradeepkumar@iiml.ac.in) [fpm13008@iiml.ac.in](mailto:fpm13008@iiml.ac.in) [（M·古普塔），](mailto:pradeepkumar@iiml.ac.in) [pradeepkumar@iiml.ac.in](mailto:pradeepkumar@iiml.ac.in)

[（P·库马尔），](mailto:pradeepkumar@iiml.ac.in) [bhasker@iiml.ac.in](mailto:pradeepkumar@iiml.ac.in) [（B·巴斯克）。](mailto:pradeepkumar@iiml.ac.in)

船。 [图。1](https://translate.googleusercontent.com/translate_f#2) （b）显示了参考书目数据集的片段，

具有不同类型对象的异构信息网络

像论文，作者和不同关系的会议

它们之间。

链接信息的常规表示，

基因信息网络非常受欢迎，方便

各种采矿任务。 但是，异构实体和他们的

复杂的关系不能用均匀的方式表示

信息网络。 最近，研究人员和从业者都是

在各种采矿中利用异构信息网络

[任务获得的信息块（](https://translate.googleusercontent.com/translate_f#16) [邓黎，王，方和2012年;](https://translate.googleusercontent.com/translate_f#16)

[张，胡，何，王和2015年](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [信息丰富多样](https://translate.googleusercontent.com/translate_f#16)

与其相比，信息网络提供了更好的采矿结果

[齐次变换（](https://translate.googleusercontent.com/translate_f#16) [Gupta等人，2015; Shi等，2014;](https://translate.googleusercontent.com/translate_f#16)

[Sun等，2011](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [例如，](https://translate.googleusercontent.com/translate_f#16) [图1](https://translate.googleusercontent.com/translate_f#16) [（b）表示的异构IN-](https://translate.googleusercontent.com/translate_f#16)

形成网络由作者，论文和会议组成，

分配办法。 这个网络比ho-

mogeneous变换了在示出 [图1中](https://translate.googleusercontent.com/translate_f#2) （a）和显示

作者之间的共同作者关系。 对于不同的采矿

异构信息网络上的任务如聚类，

我们需要测量物体之间的相关性。

[http://dx.doi.org/10.1016/j.eswa.2016.10.01](http://dx.doi.org/10.1016/j.eswa.2016.10.013) 3

0957-4174 /©2016 Elsevier Ltd.保留所有权利。

|  |
| --- |
| **第2页** |

*Gupta等人* */ Expert Systems With Applications 68（2017）106-122*

107

**图1**均相和多相信息网络的例子。

[传统的基于链接的措施，如个性化的PageRank（](https://translate.googleusercontent.com/translate_f#16) [JEH](https://translate.googleusercontent.com/translate_f#16)

[＆Widom，2003;](https://translate.googleusercontent.com/translate_f#16) [立本-诺维尔和克林伯格，2007年](https://translate.googleusercontent.com/translate_f#16) [），SimRank（](https://translate.googleusercontent.com/translate_f#16) [JEH](https://translate.googleusercontent.com/translate_f#16)

[＆代表智慧，2002](https://translate.googleusercontent.com/translate_f#17) [）并不适用于异构信息](https://translate.googleusercontent.com/translate_f#17)

网络由于对象和关系的异质性

网络（ [Gupta等人，2015;](https://translate.googleusercontent.com/translate_f#16) [Shi et al。，2014;](https://translate.googleusercontent.com/translate_f#16) [Sun等，2011](https://translate.googleusercontent.com/translate_f#16) ）。

基于元路径的相关性测度，如PathSim，HeteSim，DPRel

最近已经提出了措施之间的相关性

物体（ [Gupta等人，2015;](https://translate.googleusercontent.com/translate_f#16) [Shi et al。，2014;](https://translate.googleusercontent.com/translate_f#16) [Sun等，2011](https://translate.googleusercontent.com/translate_f#16) ）。 运用

基于元路径的相关度测度，

路径语义，同时测量对象之间的相关性。

在这项工作中，

基因信息网络。 分类是一个重要的min-

因为它需要用于各种应用程序，如链接谓词 -

[重刑，社区检测和目标建议（](https://translate.googleusercontent.com/translate_f#17) [香港，余秋雨，](https://translate.googleusercontent.com/translate_f#17)

[丁，野，2012;](https://translate.googleusercontent.com/translate_f#17) [太阳和汉族，2013](https://translate.googleusercontent.com/translate_f#17) [）。](https://translate.googleusercontent.com/translate_f#17) [分类问题，](https://translate.googleusercontent.com/translate_f#17)

使用常规中存在的链接信息来对象

均匀信息网络得到了很好的研究，

[研究人员（plored](https://translate.googleusercontent.com/translate_f#17) [路与Geetor，2003; Macskassy及教务长，](https://translate.googleusercontent.com/translate_f#17)

[2007;](https://translate.googleusercontent.com/translate_f#17) [Zhou，Bousquet，Lal，Weston，＆Schölkopf，2004;](https://translate.googleusercontent.com/translate_f#17) [Zhu等人，](https://translate.googleusercontent.com/translate_f#17)

[2003](https://translate.googleusercontent.com/translate_f#17) [）。](https://translate.googleusercontent.com/translate_f#17) [各种关系和转换分类方法](https://translate.googleusercontent.com/translate_f#17)

对于同质信息网络，如加权投票Rela-

相邻（wvRN）分类器（ [Macskassy＆Provost，2003](https://translate.googleusercontent.com/translate_f#17) ），和

[与本地和全球的一致性（LLGC）分类（学习](https://translate.googleusercontent.com/translate_f#17) [周](https://translate.googleusercontent.com/translate_f#17)

[等人，2004](https://translate.googleusercontent.com/translate_f#17) [），已经提出并广泛使用。](https://translate.googleusercontent.com/translate_f#17) [然而，](https://translate.googleusercontent.com/translate_f#17)

异构信息网络中对象的分类是

比较新。 对于异类物体的分类

信息网络，我们利用转换分类

与传统分类不同。 在传统的超级 -

视觉分类，假定数据对象是独立的

并相同分布; 然而，

将一组链接对象之间的标签相关性缩小到

隐藏网络中其他对象的标签。

转换分类是非常具有挑战性和复杂性的

异构信息网络的情况。 由于不同

对象类型和/或对象之间的关系，标签信息 -

不应该使用一种类型的物体来确定

[另一种类型的对象的标签（](https://translate.googleusercontent.com/translate_f#16) [Angelova，Kasneci，＆Weikum，](https://translate.googleusercontent.com/translate_f#16)

[2012;](https://translate.googleusercontent.com/translate_f#16) [Kong等，2012](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [在异构信息网络中，](https://translate.googleusercontent.com/translate_f#16)

一个对象类型的标签集在概念上不同于

其他对象类型的标签集由于不同的特征 -

[不同类型的对象的抽搐（](https://translate.googleusercontent.com/translate_f#16) [。Angelova等人，2012; Kong等，](https://translate.googleusercontent.com/translate_f#16)

[2012](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [例如，在异构Flickr网络的情况下，](https://translate.googleusercontent.com/translate_f#16)

照片分类的标签集概念将不同

从标签集的用户概念（ [Angelova等，2012](https://translate.googleusercontent.com/translate_f#16) ）。 距离

从此，异构数据的特点如

网络结构的复杂性，缺乏特征和稀缺性

的标签对象，也增加了对对象进行分类的难度

[异构信息网（](https://translate.googleusercontent.com/translate_f#17) [冀，太阳，Danilevsky，韩，＆](https://translate.googleusercontent.com/translate_f#17)

[高，2010](https://translate.googleusercontent.com/translate_f#17) [）。](https://translate.googleusercontent.com/translate_f#17) [开采这样的异构信息网络us-](https://translate.googleusercontent.com/translate_f#17)

传统的技术是不可行的，

如果没有，网络中存在的抽搐和微妙将会丢失

已搞定 （ [Shi et al。，2014;](https://translate.googleusercontent.com/translate_f#17) [Sun等，2011](https://translate.googleusercontent.com/translate_f#17) ）。

在传统分类中，进行监督学习

使用对象的本地特征或属性。 但是，在大多数情况下

真实世界的异构信息网络，没有

对象的自然本地特征或属性（ [Ji et al。，2010](https://translate.googleusercontent.com/translate_f#17) ）。

在异构信息网络中，如果链接信息是

被认为是物体的属性，那么很可能，

物体的尺寸将非常高，

[对象的折痕数，数据变得在自然界中稀疏（](https://translate.googleusercontent.com/translate_f#17) [籍](https://translate.googleusercontent.com/translate_f#17)

[et al。，2010;](https://translate.googleusercontent.com/translate_f#17) [洛，关，王，林和2014年](https://translate.googleusercontent.com/translate_f#17) [）。](https://translate.googleusercontent.com/translate_f#17) [但是当一些](https://translate.googleusercontent.com/translate_f#17)

对象类型具有与它们相关联的特征或属性，

传统的分类技术将不适用

因为他们的属性将无法比拟，

属性空间（ [Ji et al。，2010](https://translate.googleusercontent.com/translate_f#17) ）。 这就是为什么传统的classi-

支持向量机，逻辑回归 -

和朴素贝叶斯难以应用于异构信息 -

配合网络。 因此，对象的转换分类

可以使用标签和未标记的对象来执行

网络获取未标记对象的标签信息。

在这项工作中，我们提出了一个名为HeteClass的框架，

在异质场中称为目标类型的一种物体，

新的信息网络。 在这项工作中，我们的重点是分类 -

一种类型的对象的代替，而不是所有类型的对象collec-

tively。 此问题设置存在于各种现实世界中。

原因是由于不同的标签集概念不同

网络中的对象类型（ [Kong et al。，2012](https://translate.googleusercontent.com/translate_f#17) ）。 例如，

在用于参考书目的异构信息网络的情况下，

phy，会有几个对象类型，如作者，会议，

论文和关键字。 分类可以任意执行

一种类型的作品，如作者被称为

目标类型。 在提出的框架HeteClass中，

探索路径以结合与之相关联的语义

同时对目标类型对象执行分类。

对于分类，我们进行元路径的权重学习，

标记更高的权重到元路径，这将导致良好的分类，

使用先前标签信息的指示精度。 体重学习

元路径是HeteClass框架中的一个重要步骤

的元路径将导致良好的分类精度和

因此应该分配比其他更高的权重

|  |
| --- |
| **第3页** |

108

*Gupta等人* */ Expert Systems With Applications 68（2017）106-122*

元的路径。 如果相等权重分配给所有元路径，则

分类精度会因为影响而降低

进行分类时元路径的语义。

为了显示HeteClass的有效性，我们使用现实的书目 -

摄影数据集DBLP和Flick Fashion 10,000数据集。 亲爱的

提出的框架HeteClass与算法uti-

在框架中进行分类，即其性能可以

改进使用任何高级分类算法。 为了恶魔 -

重视HeteClass的有效性，

不同类别的分类算法，即LLGC

（ [Zhou et al。，2004](https://translate.googleusercontent.com/translate_f#17) ）和wvRN（ [Macskassy＆普罗沃斯特，2003](https://translate.googleusercontent.com/translate_f#17) ）clas-

sifiers。 我们还比较了HeteClass的性能，

用于异构信息网络的中等提出的算法

称为HetPathMine（ [Luo et al。，2014](https://translate.googleusercontent.com/translate_f#17) ），也可以利用各种各样的

用于目标类型对象分类的网络中的元路径。

实验结果表明HeteClass的有效性

与上述分类算法相比。

研究的预期贡献是基于元路径

新颖的框架HeteClass用于目标对象的分类

键入标签集的异构信息网络

对象类型在概念上是不同的。 HeteClass探索

网络的模式生成一组元路径，

同时执行目标类型对象的分类。

HeteClass允许将领域专家知识纳入se-

只能选择那些重要且足够好的元路径

以产生高分级精度。 HeteClass执行

元路径的权重学习更有效。 此外，

提出的框架HeteClass是灵活的，并允许利用任何

适用于物体分类的分类算法

生成目标单个均匀信息网络

键入对象。

在这项工作中，我们使用个性化PageRank算法

（ [Haveliwala，2002年](https://translate.googleusercontent.com/translate_f#16) ）用于对同质物体进行分类

网络上的目标类型。 但是，在拟议的框架内，

而不是Personalized PageRank算法，我们可以利用任何ad-

用于分类的高级算法将增加accu-

分类分类。 实验结果表明，精度

与LLGC，wvRN相比，拟议框架更好

和HetPathMine算法。

本文的其余部分安排如下。 在 [第2节](https://translate.googleusercontent.com/translate_f#3) ，我们

介绍相关工作。 本研究的初步和正式

问题定义在给定的 [第3节](https://translate.googleusercontent.com/translate_f#4) 。 拟议的框架

HeteClass在说明 [第4节](https://translate.googleusercontent.com/translate_f#6) 。 在[第5节](https://translate.googleusercontent.com/translate_f#9) ，实验 -

解释了设置。 在 [第6节](https://translate.googleusercontent.com/translate_f#12) ，结果和讨论是预先

sented。 最后，结论和未来的研究方向是

在呈现的 [第7节](https://translate.googleusercontent.com/translate_f#16) [。](https://translate.googleusercontent.com/translate_f#16)

**相关工作**

对象的转换分类问题

结构或网络受到重视。 这个想法

已经使用了少量的先前的标签信息

对象并对对象执行转换分类

网络利用当地和全球的结构

[网络。](https://translate.googleusercontent.com/translate_f#17) [通过早期的研究](https://translate.googleusercontent.com/translate_f#17) [Lu和Geetor（2003），Macskassy和](https://translate.googleusercontent.com/translate_f#17)

[Provost（2003），Zhou et al。](https://translate.googleusercontent.com/translate_f#17) [（2004年）](https://translate.googleusercontent.com/translate_f#17) [，和](https://translate.googleusercontent.com/translate_f#17) [Macskassy兼教务长](https://translate.googleusercontent.com/translate_f#17)

[（2007年）](https://translate.googleusercontent.com/translate_f#17) [已经执行对象的使用本地的分类](https://translate.googleusercontent.com/translate_f#17)

和/或网络的全局结构。 但是，这些方法

被设计用于均匀的信息网络，而不是

直接适用于异构信息网络。

在异构信息网络中，

不同类型的对象和/或关系使得网络不可用，

适合上述基于链路的分类。 当这些

基于链接分类的方法直接应用于

异构信息网络，分类准确性

会很低（ [Kong et al。，2012;](https://translate.googleusercontent.com/translate_f#17) [孙和汉，2013](https://translate.googleusercontent.com/translate_f#17) ）。 如果我们转变

异构信息网络成为相应的ho-

随后的一个元路径的信息网络

应用上述基于链接的分类方法，

如果选择的元路径se-

吝啬不传达有意义的关系

目标类型对象。 由于在异构信息网络中，

可能有几个具有不同语义的元路径

意思是，按照这些元路径，

基因信息网络成为相应的均匀

仅由目标类型对象组成的信息网络。 以来

我们不知道事先知道clas-

合理的准确性，我们可能会跟随一个无效的路径

这将导致分类精度低。 另外，

只有一个元路径，我们可能会错过其他语义上的

不可能的元路径，可能有助于良好的分类accu-

活泼。 因此，考虑开始和结束的所有元路径

目标类型节点对于良好的分类精度是必要的。

对于异构信息中的对象进行分类，

作品最近提出了不同的技术。 在里面

通过做工作， [罗西，安德拉德洛佩斯和雷森德（2016）](https://translate.googleusercontent.com/translate_f#17) ，au-

thor执行对象的转换分类，

异构网络 然而，他们的做法是为了二分网络，

作品。 [Angelova et al。](https://translate.googleusercontent.com/translate_f#16) [（2012）](https://translate.googleusercontent.com/translate_f#16) ）提出了一种称为涂鸦的技术

用于网络中集体对象的分类。 他们

在他们的方法中使随机步行模型化。 在他们的

他们考虑到不同的分类问题

具有概念上不同标签集合的节点类型即

甚至相同标签集的Cept / semantic也不会相似

[不同类型的标签的节点是特定类型（](https://translate.googleusercontent.com/translate_f#16) [Angelova等](https://translate.googleusercontent.com/translate_f#16)

[人，2012](https://translate.googleusercontent.com/translate_f#16) [）。](https://translate.googleusercontent.com/translate_f#16) [在所做另一部作品](https://translate.googleusercontent.com/translate_f#16) [姬等人。](https://translate.googleusercontent.com/translate_f#16) [（2010年）](https://translate.googleusercontent.com/translate_f#16) [，作者亲](https://translate.googleusercontent.com/translate_f#16)

提出了GNetMine用于对象的转换分类

网络。 在他们的方法中，他们利用了整个网络

转换分类。 他们假设了相同的标签集

保护网络中的所有对象类型。 他们丢弃了

标签集在对象类型之间传递的语义，

贝尔信息网络。 对于分类，他们使用

基于图形的正则化框架。 但是，他们没有

考虑不同标签集概念之间的区别

对象类型。 例如，标签的特征为

作者将不同于标签集的特征

会议（ [Angelova等，2012](https://translate.googleusercontent.com/translate_f#16) ）。

最近， [香港等。](https://translate.googleusercontent.com/translate_f#17) [（2012）](https://translate.googleusercontent.com/translate_f#17)提出的异构Collec-

tive分类（HCC）算法用于对象的分类

异构信息网络。 在他们的工作中，

形成一种称为目标的物体的集体分类

类型。 然而，他们将一个或多个对象类型作为特征ob-

目标类型的对象。 例如，在参考书目数据集中

DBLP，他们把论文的关键词作为作者的特征

被分类。 然后他们执行对象的特征增强

通过考虑各种元路径。 使用增强功能，

他们分两步进行分类：（1）Bootstrap by

考虑节点特征，（2）迭代推理，

关系特征。 他们在工作中，

一个对象类型的标签集的异端与其他对象不同

类型，并且它们执行被称为一种对象类型的分类

作为目标对象类型。 他们也考虑了各种元路径

功能集增加的网络。 但是，有一些

他们的方法的局限性。 首先，因为它们使用了功能

分类，需要大量的标签对象

模型一代。 在现实世界中，

对于大量的物体是昂贵的。 这使得HCC algo-

算法不适用于任务推断的现实情况

未标记的对象的标签信息使用小的num-

在网络中标记的对象。 其次，他们使用了一个或多个

对象类型作为目标类型对象的功能。 但是，它是

当我们难以确定对象的适当特征

|  |
| --- |
| **第4页** |

*Gupta等人* */ Expert Systems With Applications 68（2017）106-122*

109

**表格1**

符号说明

符号

描述

*G* *=（V，E），T g ^*

异构信息网络（HIN）及其模式

*A，R*

HIN中的对象类型和关系

*∅，ψ*

对象类型和链接类型映射功能

*R C，P*

复合关系和元路径

*w ^ AiAj，男*

邻接矩阵和加权路径矩阵

*SIM P（一个I，A j）的*

下列元路径*P* 之间*对象* *i*和*第j* 关联

*θK，W·K时*

元路径*P k*的元路径 *P k*和体重的重要性

只有链接信息。 接下来，他们考虑了他们的工作

元路径，但它们同等重要/重于所有元 -

路径。 但是，我们知道一些元路径可能是

比其他元路径要好，所以应该

给予较高的重量。

近日， [罗等人。](https://translate.googleusercontent.com/translate_f#17) [（2014）](https://translate.googleusercontent.com/translate_f#17)提出HetPathMine算法

目标类型对象在异构信息中的分类

网络。 作者在他们的工作中提出体重学习

元路径来整合各种路径语义。 他们用trans-

用于网络中物体标记的延性分类。 他们

解决了早期作品在分类中的局限性，

异构信息网络当一个ob-

对象类型在概念上与其他类型不同。 但是，有

HetPathMine有一些限制。 在HetPathMine中，没有

从网络模式生成元路径的方法。 另外，

HetPathMine的体重学习是无约束的监督学习，

使用可能导致否定的先前标签信息

路径重量并没有意义。 接下来，HetPath-

我们利用PathSim进行相关性测量，而不是

有效的异质信息相关性测量

只能使用对称的元路径

网络（ [Gupta等人，2015;](https://translate.googleusercontent.com/translate_f#16) [Shi et al。，2014](https://translate.googleusercontent.com/translate_f#16) ）。

为了解决HetPathMine和更早的甲基 -

ods在异构信息网络中进行分类

对于不同的对象类型，标签集在概念上是不同的，

我们提出了一个名为HeteClass的新框架进行分类

异构信息网络中的对象。 建议

框架HeteClass探索异构网络的网络模式，

基因信息网络生成一组元路径

可用于体重学习。 HeteClass也可以整合

领域专家的知识，如果可用，减少数量

的元路径用于体重学习。 减少的优点

路径数是计算时间会减少

体重学习。 因为可能会有一些路径不会

导致良好的分类准确性，从中删除这些路径

生成的路径集也会增加有效的重量

集合中的路径总数将减少

并且所有路径的权重之和将为一。 另外，

HeteClass允许使用选择的分类算法

由目标类型对象组成的最终同质网络。 这个

使HeteClass框架高效，这可能导致

准确度好。 对于这项研究，我们对性能进行了比较

的HeteClass与LLGC和wvRN。 我们还比较了

HeteClass与HetPathMine的性能显示效果

HeteClass。 对于实验，我们利用现实世界的参考书目

数据集DBLP和Flickr Fashion 10,000数据集。

**3.初步和问题定义**

在这部分，这个工作的背景和初步

被呈现。 正式的定义也给出了转换

目标类型节点在异构信息分类中的分类

网络。 一些重要和常用的符号是

在给定的 [表1中。](https://translate.googleusercontent.com/translate_f#4)

**图2**片段表示为信息网络书目的数据集。

**图** DBLP网络**3.**网络架构。

[这项工作利用信息网络中定义](https://translate.googleusercontent.com/translate_f#4) [的定义](https://translate.googleusercontent.com/translate_f#4)

[1](https://translate.googleusercontent.com/translate_f#4) [，这是类似于由定义](https://translate.googleusercontent.com/translate_f#4) [Sun等人。](https://translate.googleusercontent.com/translate_f#4) [（2011年）](https://translate.googleusercontent.com/translate_f#4) [。](https://translate.googleusercontent.com/translate_f#4)

**定义1（**信息网络**）。** 信息网络是

罚款作为与对象类型映射有向图 *G* *=（V，E）*

功能*∅：V→A*和链接类型映射函数*ψ：* *电子* *→R，*

其中每个物体 *V∈V*属于一个特定的对象类型*∅（v）的*

*∈A，*并且每个链路*e∈Ë*属于特定关系*ψ（E）∈R。*

当对象的类型| *A* | *>* 1或类型的关系| *[R* | *>*

1，网络是异构信息网络; 除此以外，

它是一个均匀的信息网络。

在参考书目数据集的网络片段中，如图所示

[图2](https://translate.googleusercontent.com/translate_f#4) 我们可以看到有三种不同类型的节点

和两个不同的关系。 在这个网络以上多

存在一种类型的节点和关系，

形成网络。 对于异构信息网络，我们

认为它是元级（即模式级）

理解（ [老挝，2010年](https://translate.googleusercontent.com/translate_f#17) ）中所定义的[定义2。](https://translate.googleusercontent.com/translate_f#4)

**定义2（**网络架构**）。** 网络模式表示

为 *T G* *=（A，R），*为异构的元级表示

信息网络 *G* *=（V，E）*与对象类型映射*∅：V→*

*A*和链路类型映射*ψ：* *电子* *→R，*这是一个有向图上方

对象类型 *A*和边缘作为与*R*的关系。

像DBLP这样的参考书目信息网络就是一个例子

的异构信息网络。 该网络由

四种类型的实体/对象：纸张*（P），*作家*（A），*会议

*（C）*和关键字*（K）。* [图3](https://translate.googleusercontent.com/translate_f#4)示出了所述元级表示

（网络架构）的DBLP网络。 由于关系是

在这个网络中已经使用了双向无向链路。

**定义3（**间路径**）。** 荟萃路径*P*被网络上定义

架构 *的Tg* *=（A，R）*和在*A* 1的形式表示

*，R* 1

*→A* 2

*R* 2

→...

*中 R 1*

→

*A L* 1，其定义了一个复合关系*R C* = *R* 1

◦

*R* 2

◦···◦ *中R 1*

的

源对象之间长度 *l* *A*型1个目标对象类型*A L* 1

使用组合运算符°关系。

|  |
| --- |
| **第5页** |

110

*Gupta等人* */ Expert Systems With Applications 68（2017）106-122*

**图**用于DBLP网络架构**4.**元路径。

双元路 *APA（* *作者-纸-作者* ）和*APC（* *作者*

*-纸-*在DBLP网络架构的*会议* ）显示在

[图4](https://translate.googleusercontent.com/translate_f#5) （a）和（b）。

在异构信息网络中，可以有两个对象

通过不同的路径连接，这些路径将有所不同

语义含义 例如，元路径 *APA*和*APCPA*在

DBLP网络模式是连接au-

对作者来说，对象和目标对象是相同的

这两个元路径。 但是，这两个元路径具有不同的se-

神秘的解释。 元路 *APA*意味着作家谁是合作

论文作者 元路 *APCPA*表明作者出版

论文在同一次会议上。 不同的语义含义

不同的元路径将导致不同的分类精度。

继元路 *APA* EM- 作者之间的相关性

phasizes上的文件，而在 *APCPA* 的情况下 ，会议

被强调。 因此，一个对象之间的相关性

异构信息网络依赖于元路径，

lowed。

如果对象之间没有多重关系，我们可以

仅使用对象类型如 *P* 代表所述元路径 =

*（A* 1 *A* 2 ... *A L* *1）。* *a* 1 和 之间 的 *路径实例P* *=（A* 1 *A* 2 ... *A L* *1）*

一个网络中的*G* *A L* 1是所述元路径*P，*即*P∈P*的一个实例，

如果对于每个 *A I，∅（A* *I）=* *A i*和对每个链路*为e i* = *<A I，A I* *1>，ψ（E I）=* *R I。*

元路径*P* 的 *反向路径* ，表示为*P* -1，定义了一个

对象类型之间的反比关系。 同样地， *反向路径IN-*

*姿态P* *-1∈P* -1为*p*的*G*中的反向路径。

对于使用HeteClass的对象进行分类，

确定对象之间的相关性。 为此，相关性度量

DPRel（ [Gupta等，2015](https://translate.googleusercontent.com/translate_f#16) ）被使用。 为了测量相关 -

使用DPRel的对象之间的关系，

基因信息网络组成二部网络

只有源和目标类型的对象。 为此，我们需要计算

加权路径矩阵中说明 [定义4。](https://translate.googleusercontent.com/translate_f#5)

**定义4（**加权路径矩阵**）。** 对于异构信息 -

mation网络及其模式级别表示，加权

用于元路径*P* *=（A* 1 *A* 2 ... *A L* *1）* 路径矩阵 *M*被定义为*M* =

*W A* 1 *A* *2×W A* 2 *A* *3×。* *。* *。* *×W A l 一个 升* 1

，其中 *W A I 第 j*

是邻接矩阵

类型的对象 *A I* *和第j* 之间 *。* *M* *[X I，Y j]的*表示数字

对象 *x* 与路径实例的 *我* ＆Element; *A* 1及*Y j∈A* *升* 1以下

元的路径 *P，*并且*M* *[X I，Y j]的* = *M* *[YĴ，X i]中* 。

在示出的异构信息网络 [图2](https://translate.googleusercontent.com/translate_f#4) ，的

加权路径矩阵的计算被示出在 [图5](https://translate.googleusercontent.com/translate_f#5)以下

元路径 *APC*其中源对象类型是*A（*作者）和目标

对象类型是 *C（*会议）。

DPRel定义在 [定义5](https://translate.googleusercontent.com/translate_f#5) 。 然后我们展示如何利用

DPRel用于测量物体之间的相关性。

**定义5（DPRel）。** 给定的元路径*P* *=（A* 1 *A* 2 ... *A L* *1），*例如

是 *，A* 1和*A L* 1是不同的对象类型，则对于二分表象

发送异构信息网络只有ob-

类型 *A* 1 的jects和*A L* 1，源对象之间的相关性

**图5**计算加权路径矩阵的。

*一个* 1 *I＆Element;* *A* 1个对象物*B（L* *1）.J∈A 升* 1是：

*DPRel*

（

*1个* *I，B（L* *1）.J* | *P*

）

=

*w ^*

（

*1个* *I，B（L* *1）.J*

）（1

度 *（α1*个 *I）*

+

1

度 *（B（L* *1）j）的*

）

1

度 *（α1*个 *I）*

Σ

*Ĵw ^*

（

*1个* *I，B（L* *1）.J*

）

+

1

度 *（B（L* *1）j）的*

Σ

*I W*

（

*1个* *I，B（L* *1）.J*

）

（1）

其中*，w（1* *I，B（L* *1）j）*为值M *[1* *I，B（L* *1）j]的*加权从

路径矩阵即路径连接*对象* 1 *i* 中的数

＆Element; *A* 1

和 *b（L* *1）.J∈A 升* 1

遵循指定的元路径。

度 *（α1*个 *i）*和度*（B（L* *1）j）*是节点度*的*对象*的* 1 *i*和

*B（ 升* *1）*在二分表示*Ĵ*分别。

由于在这项工作中，我们必须对对象进行分类，

我们需要测量相同类型的对象之间的相关性

即，元路径的源和目标对象类型将相同。 在

在这种情况下，我们测量相同类型的对象之间的相关性

如在说明的 [Gupta等人。](https://translate.googleusercontent.com/translate_f#16) [（2015）](https://translate.googleusercontent.com/translate_f#16) 。 下面给出了一个例子。

在示出的异构信息网络 [图2](https://translate.googleusercontent.com/translate_f#4)和

加权路径矩阵中所示的该网络 [图5](https://translate.googleusercontent.com/translate_f#5) ，fol-

降脂元路 *APCPA，*我们计算非盟之间的关联性

雷神。 元路径的中途对象类型是 *C（*会议）。

使用这个对象类型我们把元路 *APCPA*为两个

长度相等的子路径，即*，P* *L* = *APC*和*P R* = *CPA。* 使用这些

元路径我们计算作者之间的相关性如图所示

下面：

*SIM P L（1，C* *1）=* *DPRel（1，C* 1 | *P L）=*

2

（1

2 + 1

3

）

（1

2×3

）

+

（1

3×5

）= *0。* 53

作者和会议之间的全面相关性矩阵，

下列元路径 *P L* = *APC*示于[图6中](https://translate.googleusercontent.com/translate_f#6) （一个）。 相关 -

作者和会议之后的元数据路径之间的矩阵

*P* -1

*[R*

= *APC*将是相同所示的矩阵[图6](https://translate.googleusercontent.com/translate_f#6) （a）这样

元路径是对称的。 使用这两个矩阵，我们计算

作者之间的相关性如下图所示：

*X* = *DPRel（α1，{C* *1，C* 2} | *P L* = *APC）=* *{0。* *53，0。* 33}

*Y* = *DPRel*

（

*3，{C* *1，C* 2} | *P* -1

*[R*

= *APC*

）

= *{0。* *5，0}*

*辛APCPA（A* *1，A* *3）=*

*XY*

*X* 2 + *Y* 2 - *XY*

= *0。* 71

|  |
| --- |
| **第6页** |

*Gupta等人* */ Expert Systems With Applications 68（2017）106-122*

111

**图**使用DPRel和下列元路径APCPA作者之间**6.**计算相关性。

**图7.**作者的分类。

的计算的其余部分的结果示于 [图6中](https://translate.googleusercontent.com/translate_f#6) （b）中

以矩阵形式。

在计算对象之间的相关性之后，

我们认为是按照不同的元路径进行的

创建与之相对应的均匀信息网络

元的路径。 之后，我们执行元数学的权重学习，

paths and then perform a weighted combination of those individ-

ual networks. Then transductive classification is performed to label

the unlabeled objects. The problem of transductive classification is

formally defined in [Definition 6](https://translate.googleusercontent.com/translate_f#6) [.](https://translate.googleusercontent.com/translate_f#6)

**Definition 6** (Transductive classification) **.** Given a heterogeneous

information network *G* = *( V, E ) ,V* = *U m*

*( i* =1 *)*

*V i*

where *V i* ∈ *A i , i* =

1 *, .* *。* *。* *, m* and a subset of target data objects *V T* ⊂ *V T* ∈ *A T* which

are labelled with values *C* = { *C* 1 *,C* 2 *, .* *。* *。* *,C n* } denoting class label to

which each object belongs to, predict the labels for the unlabelled

objects of target type *( V T* − *V T )* ∈ *A T* .

For example, in [Fig. 7](https://translate.googleusercontent.com/translate_f#6) , the target type objects are authors

on which we have to perform the classification. Authors *a* 1 and

*a* 4 are pre-labelled with class labels “Information Retrieval” and

“Data Mining” respectively. The task is to classify authors *a* 2 and

*a* 3 using the pre-labelled information. Since following various paths

we can connect authors to form homogeneous information net-

works, weight learning is required to determine the importance of

various paths. The proposed framework HeteClass explores various

meta-paths from the network schema and performs weight learn-

ing using the pre-labelled information to leverage the semantics of

various meta-paths for classification of objects.

**4. The framework of heteclass**

Transductive classification utilizes dependency among data ob-

jects for relational learning to classify unlabeled objects. 在这个

section, we present the proposed framework HeteClass for trans-

ductive classification on target type objects in a heterogeneous in-

formation network. The framework of HeteClass performs the clas-

sification task in two phases. In the first phase, a set of meta-

paths is generated and knowledge of domain expert(s) is utilized

(if available) to filter the meta-paths produced by exploring the

input network schema. This would reduce the number of meta-

paths to be explored for classification. In the second phase, weight

learning is performed for the final set of met-paths produced by

the first phase using prior label information in the network. After

the weight learning, a single homogeneous information network is

produced by a weighted combination of various homogeneous in-

formation networks corresponding to each meta-path. Then, using

the transductive classification algorithm, the unlabelled set of ob-

jects in the network is classified. [Fig. 8](https://translate.googleusercontent.com/translate_f#7) shows the detailed frame-

work of HeteClass. In this framework, there are two phases:

*•* Generation of meta-paths, and

*•* Weight learning and transductive classification.

*4.1。* *Phase 1: generation of meta-paths*

In this phase of the HeteClass framework, network schema of

the heterogeneous information network is explored to generate all

meta-paths that start from and end with the target object type.

For example, [Fig. 9](https://translate.googleusercontent.com/translate_f#7) (a) shows the network schema of a bibliography

dataset. In this network schema, if the target object type is Author

(A), the exploration of the network schema to generate meta-paths

that start from A and end with A would be as shown in [Fig. 9](https://translate.googleusercontent.com/translate_f#7) （b）中。

Meta-paths generated in this example would be in the format of

*author* − ∗ − *author* ie, the start and end type objects would be

same.

The pseudo code of the algorithm for meta-path generation is

given in [Algorithm 1](https://translate.googleusercontent.com/translate_f#8) 。 This algorithm takes the network schema of

the heterogeneous information network and target object type on

which classification is to be performed as input. It also takes as in-

put the maximum length of meta-paths to be generated. The start

node of the algorithm is the target object type. [Algorithm 1](https://translate.googleusercontent.com/translate_f#8) gen-

erates meta-paths by exploring the neighbors of the current node

and if any node is the target object type, it generates the meta-

path by storing the meta-path starting from the target object type

and ending with the same. [Algorithm 1](https://translate.googleusercontent.com/translate_f#8) has two functions. 首先

function is used to generate the neighbors of the current node and

the second function generates meta-paths by merging the two sub-

paths that have the same end and start object type. This process is

repeated for a number of iterations as shown in [Fig. 9](https://translate.googleusercontent.com/translate_f#7) (b) which

would be equal to the maximum length for meta-paths.

After generation of meta-paths, we can utilize domain expert

knowledge for reducing the number of meta-paths by discarding

the paths that would not lead to good classification accuracy. 对于

a network, it might be possible to have a large number of meta-

paths for weight learning. This would increase the computation

|  |
| --- |
| **第7页** |

112

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

**Fig. 8.** HeteClass framework.

**Fig. 9.** Meta-paths generation from the network schema.

时间。 However, if we remove the paths that would lead to low

classification accuracy, the computation time would be reduced.

Also, by removing ineffective paths, weight learning would be-

come more effective as more weight would be assigned to effective

paths.

*4.2。* *Phase 2: weight learning and transductive classification*

For the set of meta-paths produced from phase-1 using

[Algorithm 1](https://translate.googleusercontent.com/translate_f#8) of HeteClass, weight learning is performed using prior

label information. The weight learning for meta-paths is important

as each meta-path has a semantic meaning, which is different from

other meta-paths. For example, meta-path APA semantically signi-

fies the co-author relationship. However, meta-path APCPA seman-

tically means authors publishing papers in the same conference.

Each meta-path would yield a different accuracy of classification

and higher weights should be assigned to meta-paths that result in

good classification accuracy. We have posed the problem of weight

learning for meta-paths as an optimization problem as defined in

|  |
| --- |
| **第8页** |

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

113

**Algorithm 1** Generation of meta-paths.

[式。](https://translate.googleusercontent.com/translate_f#8) [(2)](https://translate.googleusercontent.com/translate_f#8) [.](https://translate.googleusercontent.com/translate_f#8)

∗ =

argmin

= { *θ* 1 *, θ* 2 *,..., θ K* }

*L ( )*

（2）

Here, *θ k , k* = 1 *, .* *。* *。* *, K* is the importance of the meta-path *P k , k* =

1 *, .* *。* *。* *, K* and *L* ( ) is the loss function, defined below in [Eq. (3)](https://translate.googleusercontent.com/translate_f#8) ，那个

is to be minimized.

*L ( )* =

1

2

Σ

*v i , v j* ∈ *V T ,i* = *j*

∥

∥

∥

∥

∥

1− *Sign*

（

*v i , v j*

) *K*

Σ

*k* =1

*θ k Sim P k*

（

*v i , v j*

）

∥

∥

∥

∥

∥

2

2

+

*λ*

2

2

2

*st θ k* ≥ 0 *,* ∀ *k* = 1 *, .* *。* *。* *, K*

（3）

Here, *v i* and *v j* are the labelled objects and are of target type

objects on which classification is to be performed. *λ* is the regular-

ization parameter and

·

is the *l* 2 - norm. The function *Sign* () is

defined in [Eq. (4)](https://translate.googleusercontent.com/translate_f#8) 。

*Sign*

（

*v i , v j*

）

=

{

1 *, v i , v j* ∈ *C T and i* = *j*

−1 *, otherwise*

（4）

It returns +1 if both objects have the same label, otherwise

it returns −1. The function *Sim P k ()* computes the relatedness be-

tween target type objects following meta-path *P k* . For relatedness

[computation, we utilize the relevance measure DPRel (](https://translate.googleusercontent.com/translate_f#16) [Gupta et al.,](https://translate.googleusercontent.com/translate_f#16)

[2015](https://translate.googleusercontent.com/translate_f#16) [).](https://translate.googleusercontent.com/translate_f#16)

**Algorithm 2** Weight learning for meta-paths.

The value of loss function defined in [Eq. (2)](https://translate.googleusercontent.com/translate_f#8) , is high for a meta-

path if, following that meta-path, the relatedness between differ-

ently labeled objects are high. However, if a meta-path connects

objects with the same label with high relatedness, then the im-

portance assigned to that meta-path would be high. So, the idea is

to maximize the correlations between objects with the same la-

bel and to minimize the correlation between differently labeled

objects. To solve the optimization problem defined in [Eq. (2)](https://translate.googleusercontent.com/translate_f#8) ，我们

take the partial derivative of loss function with respect to *θ k , k* =

1 *, .* *。* *。* *, K* and equate it to zero to get the value of *θ k , k* = 1 *, .* *。* *。* *, K* as

defined in [Eq. (5)](https://translate.googleusercontent.com/translate_f#8) [.](https://translate.googleusercontent.com/translate_f#8)

*∂ L ( )*

*∂θ k*

= 0

（5）

[Now, we solve](https://translate.googleusercontent.com/translate_f#8) [Eq. (5)](https://translate.googleusercontent.com/translate_f#8) [and get the value of](https://translate.googleusercontent.com/translate_f#8) [*θ*](https://translate.googleusercontent.com/translate_f#8)[*k*](https://translate.googleusercontent.com/translate_f#8) [as given in](https://translate.googleusercontent.com/translate_f#8) [Eq.](https://translate.googleusercontent.com/translate_f#8)

[(6)](https://translate.googleusercontent.com/translate_f#8) [. Using this](https://translate.googleusercontent.com/translate_f#8) [*θ*](https://translate.googleusercontent.com/translate_f#8)[*k*](https://translate.googleusercontent.com/translate_f#8)[*, k*](https://translate.googleusercontent.com/translate_f#8) [= 1](https://translate.googleusercontent.com/translate_f#8) [*, . . . , K*](https://translate.googleusercontent.com/translate_f#8) [equation, we determine the optimal](https://translate.googleusercontent.com/translate_f#8)

value of the weights iteratively as given in [Algorithm 2](https://translate.googleusercontent.com/translate_f#8) [.](https://translate.googleusercontent.com/translate_f#8)

*θ k* =

Σ

*v i , v j* ∈ *V T ,i* = *j Sign*

（

*v i , v j*

）

*Sim P k*

（

*v i , v j*

）

*F*

（

*v i , v j*

）

*λ* +

Σ

*v i , v j* ∈ *V T , i* = *j Sign* 2

（

*v i , v j*

）

*Sim* 2

*P k*

（

*v i , v j*

）

（6）

where *f*

（

*v i , v j*

）

=

（

1 − *Sign*

（

*v i , v j*

) ∑

*r* = *k θ r Sim P k*

（

*v i , v j*

))

。

In [Algorithm 2](https://translate.googleusercontent.com/translate_f#8) , all importance values are first initialized with

positive values greater than zero. Then, using [Eq. (6)](https://translate.googleusercontent.com/translate_f#8) , we compute

the importance of each path using the importance value of other

meta-paths. We take the maximum between zero and the new

value of importance of meta-path to ensure that the importance

does not acquire negative value. This process is iterated till the

importance values are converged. To check the convergence con-

dition, we use convergence tolerance with the appropriate value.

After that, we compute the weight of meta-paths by normalizing

the importance values.

After getting the weights of meta-paths, a single homogeneous

information network is created on target type objects using meta-

paths. For each meta-path, the corresponding homogeneous infor-

mation network, represented using the relatedness matrix, is mul-

tiplied with the weight of that meta-path and summation is per-

formed to get a single homogenous network as given below in

|  |
| --- |
| **第9页** |

114

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

**Fig. 10.** Network Schema for DBLP database.

[式。](https://translate.googleusercontent.com/translate_f#9) [(7)](https://translate.googleusercontent.com/translate_f#9) [.](https://translate.googleusercontent.com/translate_f#9)

*Sim P* 1 *,...,P K* =

*ķ*

Σ

*k* =1

*w k* × *Sim P k*

（7）

Now, on this single homogeneous information network, which

is the weighted combination of different homogeneous informa-

tion networks, we perform transductive classification using prior

label information. For that, we utilize Personalized PageRank

（ [Haveliwala, 2002](https://translate.googleusercontent.com/translate_f#16) ）。 However, our framework can utilize any trans-

ductive classification algorithm; by taking more advanced and ac-

curate algorithm we can improve the classification accuracy.

**5. Experimental setup**

To validate the effectiveness of the proposed framework Het-

eClass for classification of objects in a heterogeneous information

network, we utilized DBLP “four-area” data set which is a bibliog-

raphy database, and a subset of Flickr Fashion 10,000 data set. 我们

compared the performance of HeteClass with the algorithms dis-

cussed in [Section 5.2](https://translate.googleusercontent.com/translate_f#10) 。 All experiments were performed on a system

with Intel Core i5 processor and 4 GB RAM using R version 3.0.3.

*5.1. Description of datasets*

In this section, we present the description of the DBLP “four-

area” dataset and Flickr Fashion 10,000 dataset used for classifica-

灰。

*5.1.1. DBLP “four-area” dataset*

For performance comparison, we utilized DBLP database, [1](https://translate.googleusercontent.com/translate_f)

which is a computer science bibliography database. DBLP database

[can be modeled as a heterogeneous information network (](https://translate.googleusercontent.com/translate_f#17) [Ji et al.,](https://translate.googleusercontent.com/translate_f#17)

[2010](https://translate.googleusercontent.com/translate_f#17) [). This network dataset consists of four types of objects and](https://translate.googleusercontent.com/translate_f#17)

three different relationships. The network schema of DBLP dataset

is shown in [Fig. 10](https://translate.googleusercontent.com/translate_f#9) 。 The network schema of DBLP database has

four different types of objects represented as nodes in the network

ie, *Author* (A), *Conference* (C), *Keyword* (K) and *Paper* (P). 三

different relationships, represented as links between objects, ex-

ist in the network. A bidirectional link between *Paper* and *Author*

nodes indicates that every paper in the database has been writ-

ten by some author(s) and vice-versa. Similarly, bidirectional links

exist between *Paper* and *Keyword* as well as *Paper* and *Conference*

nodes. The link from *Author* node to *Paper* node represents the re-

lationship “ *written by* ” which means the paper has been written

by the author and inverse relationship “ *writen by* −1 ” means that

the author has written the paper. Likewise, other relationships in

the network can be explained in forward and backward directions.

1 [http://dblp.uni-trier.de/](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://dblp.uni-trier.de/) 。

**表2**

Class label distribution of authors in

DBLP “four-area” dataset.

Class

# Authors

数据库

1,197

Data Mining

745

Artificial Intelligence

1,109

Information Retrieval

1,006

**Fig. 11.** Network Schema for Flickr Fashion 10,0 0 0 dataset.

When a relationship between two nodes in a network is bidirec-

tional, we can use an undirected link to represent that relationship.

For example, the relationship between author and paper nodes is

in forward and in the backward direction; therefore, we can use an

undirected link to show the relationship.

In our experiments, we utilized DBLP “four-area” dataset, [2](https://translate.googleusercontent.com/translate_f)

which is a subset of the DBLP database and has conferences in

four research area classes: Artificial Intelligence, Information Re-

trieval, Database and Data Mining. This dataset has been frequently

[utilized in various studies (](https://translate.googleusercontent.com/translate_f#16) [Gupta et al., 2015; Ji et al., 2010; Shi](https://translate.googleusercontent.com/translate_f#16)

[et al., 2014](https://translate.googleusercontent.com/translate_f#16) [). The dataset contains 20 conferences, 14,475 authors,](https://translate.googleusercontent.com/translate_f#16)

14,376 papers and 8,920 keywords with 170,794 links in total. 在

this dataset, 4,057 authors, 100 papers, and all 20 conferences are

labeled with one of the four research area classes. For accuracy

evaluation, we need to have the ground truth (label information)

for objects to be classified. The label sets are conceptually different

for different object types ( [Angelova et al., 2012; Kong et al., 2012](https://translate.googleusercontent.com/translate_f#16) ）。

In this work, we perform transductive classification on authors as

we have label information for authors and number of authors is

large enough to draw a meaningful conclusion from experiments.

[Table 2](https://translate.googleusercontent.com/translate_f#9) shows the four classes and a corresponding number of au-

thors in each class in the data set.

For experiments, we created two networks consisting of 2,500,

and 4,057 authors having label information. These two datasets

are named DBLP – 1, and DBLP – 2 respectively. The description

of these two datasets is given in [Table 3](https://translate.googleusercontent.com/translate_f#10) 。 The authors in these

datasets are selected randomly from the set of labeled authors.

DBLP – 2 has all the labeled authors.

*5.1.2. Flickr fashion 10,000 dataset*

To show the effectiveness of HeteClass, we utilized another

real-world dataset named Flickr Fashion 10,000. [3](https://translate.googleusercontent.com/translate_f) This dataset con-

sists of 32,398 photos (URLs of images on Flickr [4](https://translate.googleusercontent.com/translate_f) ) related to var-

ious fashion categories. These photos have been categorized into

262 distinct fashion classes containing at least 10 photos in a class

and at most 200 photos in a class. Each photo in the dataset has

been tagged by authors. The total number of distinct tags is 56,275.

Various meta-information related to photos like the number of fa-

vorites, the number of comments, and geo-location are available in

the dataset. However, we utilized only author and tag information

related to photos as these are the relevant information suitable for

classification task in our experiments. The resulting heterogeneous

information network has the network schema as shown in [Fig. 11.](https://translate.googleusercontent.com/translate_f#9)

In the network schema, three different types of nodes are there

ie *Author* (A), *Photo* (P), and *Tag* (T). Also, two bidirectional re-

2 [http://web.engr.illinois.edu/ ∼mingji1/](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://web.engr.illinois.edu/~mingji1/) 。

3 [http://www.st.ewi.tudelft.nl/ ∼bozzon/fashion10 0 0 0dataset/](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://www.st.ewi.tudelft.nl/~bozzon/fashion10000dataset/) 。

4 [https://www.flickr.com/](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=https://www.flickr.com/) 。

|  |
| --- |
| **第10页** |

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

115

**Table 3**

Description of the two DBLP datasets utilized for experiments.

作者

Papers

Conferences

Keywords

Total Objects

Total Links

DBLP – 1

2,500

9,899

20

7,316

19 ,735

100 ,963

DBLP – 2

4,057

14 ,328

20

8,898

27 ,303

148 ,246

**Fig. 12.** Accuracy results for DBLP – 1.

**Table 4**

Class label distribution of Photos in the subset of Flickr

Fashion 10,0 0 0 dataset utilized for experiments.

Class

# Photos

Class

# Photos

Haute Couture

206

Jeans

199

Band Collar

200

Mitre

199

Boat Neck

200

Robe

199

Cope

200

Style Line

199

Jodhpurs

200

Alb

198

Rubber Glove

200

Toile

197

Umbrella

200

Tuxedo

197

Wetsuit

200

Sari

196

Dirndl

199

Sash

196

Hijab

199

Toga

196

lationships are present in the network schema. For example, be-

tween Author and Photo nodes, the relationship “ *tagged by* ” indi-

cates that the photo has been tagged by author(s). For our exper-

iments, we utilized a subset of this Flickr Fashion 10,000 dataset.

The extracted subset contains total 3,980 photos categorized into

20 classes. The class label distribution of these photos has been

shown in [Table 4.](https://translate.googleusercontent.com/translate_f#10)

For experiments, we created two networks of the extracted sub-

set consisting of 2,500, and 3,980 photos having label informa-

tion and performed classification of photos. These two datasets are

**Table 5**

Description of the two Flickr Fashion 10,0 0 0 datasets utilized for experiments.

Photos

作者

Tags

Total Objects

Total Links

Fashion – 1

2,500

539

7,061

10 ,100

47 ,327

Fashion – 2

3,980

734

9,296

14 ,010

75 ,015

named Fashion – 1, and Fashion – 2 respectively. The photos in

these two datasets are selected randomly from the set of labeled

photos. The description of these two datasets is given in [Table 5](https://translate.googleusercontent.com/translate_f#10) 。

Fashion – 2 has all the labeled photos.

For performance evaluation of algorithms on each dataset, we

randomly select *x* %, (where *x* = 2, 4, 6, 8 and 10) of the labelled ob-

jects of the target type from that dataset, and use their label infor-

mation as prior knowledge. Using the prior knowledge, we perform

classification of the rest of the objects of target type and evaluate

the accuracy of the algorithms. This process is repeated 10 times

for each value of *x* and average value of accuracy is reported in

the results.

*5.2. Algorithms for comparison and evaluation*

To show the effectiveness of the proposed framework Hete-

Class, we compared its performance with two algorithms: Learning

with Local and Global Consistency (LLGC) ( [Zhou et al., 2004](https://translate.googleusercontent.com/translate_f#17) ）和

|  |
| --- |
| **第11页** |

116

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

**Fig. 13.** Accuracy results for DBLP – 2.

[weighted-vote Relational Neighbour classifier (wvRN) (](https://translate.googleusercontent.com/translate_f#17) [Macskassy](https://translate.googleusercontent.com/translate_f#17)

[& Provost, 2003](https://translate.googleusercontent.com/translate_f#17) [).](https://translate.googleusercontent.com/translate_f#17)

LLGC is a graph-based transductive algorithm for classification

of nodes. It utilizes the link structure of the network to propagate

label information to the rest of the nodes for classification. LLGC

utilizes the local as well as the global structure of the network

for classification. However, wvRN is a simple relational learning

algorithm, which utilizes only the local structure of the network

for classification of nodes. These two algorithms are very popu-

lar and widely utilized for classification task on networks. Neither

LLGC nor wvRN is directly applicable to a heterogeneous infor-

mation network. We can apply these algorithms only after trans-

forming the heterogeneous information network into a homoge-

neous information network by following a meta-path. In our ex-

periments, by following different meta-paths, we transform a het-

erogeneous information network into the corresponding homoge-

neous information network, which consists of only target type ob-

jects。 Then, we apply LLGC and wvRN algorithms on each ho-

mogeneous information network to perform classification on tar-

get type objects. We also compare the performance of HeteClass

with HetPathMine algorithm by [Luo et al. (2014)](https://translate.googleusercontent.com/translate_f#17) 。 HetPathMine

determines the weights of different meta-paths and combines

the corresponding network into a single homogeneous informa-

tion network. Then classification is performed on the target type

objects.

To evaluate the performance of a classification algorithm, we

utilize accuracy measure ( [Han, Kamber, & Pei, 2011](https://translate.googleusercontent.com/translate_f#16) ）。 准确性

measures compute the accuracy of the algorithm for classifica-

tion of unlabeled objects. It can be computed as the ratio of cor-

rectly classified objects to the total number of unlabelled objects

（ [Han et al., 2011](https://translate.googleusercontent.com/translate_f#16) ）。 Accuracy measure is defined as follows in

**Table 6**

Final set of meta-paths generated by Phase – 1 of HeteClass for DBLP – 1 and DBLP

– 2 datasets.

Meta-path

Length

Symmetric

*Author* – *Paper* – *Author* (APA)

2

是

*Author – Paper – Conference – Paper – Author* (APCPA)

4

是

*Author – Paper – Author – Paper – Author* (APAPA)

4

是

*Author – Paper – Keyword – Paper – Author* (APKPA)

4

是

[式。](https://translate.googleusercontent.com/translate_f#11) [（8）](https://translate.googleusercontent.com/translate_f#11) ：

*Accuracy* =

# *Correctly classi fied objects*

*ñ*

=

∑ *K*

*i* =1 *a i*

*ñ*

(8)

where *N* is the number of unlabelled objects need to be classified,

*K* is the number of classes in the dataset and *a i* is the number of

objects correctly classified to its actual class.

*5.3. Meta-path generation and selection*

To utilize the semantics of various meta-paths, HeteClass gen-

erates a set of meta-paths (in Phase – 1) that can be combined

to form the homogeneous information network consisting of ob-

jects that need to be classified. For our experiments, we consider

the network schema of DBLP as shown in [Fig. 10](https://translate.googleusercontent.com/translate_f#9) for HeteClass to

generate the set of meta-paths. In this work, we consider meta-

paths of length up to four since higher length meta-paths would be

[highly noisy for the classification process (](https://translate.googleusercontent.com/translate_f#17) [Kong et al., 2012; Shi et](https://translate.googleusercontent.com/translate_f#17)

[al., 2014](https://translate.googleusercontent.com/translate_f#17) [) and affect the accuracy of results. The final set of meta-](https://translate.googleusercontent.com/translate_f#17)

paths generated and utilized in our experiments for dataset DBLP

– 1 and DBLP – 2 is shown in [Table 6](https://translate.googleusercontent.com/translate_f#11) 。 The four meta-paths utilized

for experiments are symmetric ( [Gupta et al., 2015](https://translate.googleusercontent.com/translate_f#16) ）。

|  |
| --- |
| **第12页** |

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

117

**Fig. 14.** Accuracy results for Fashion – 1.

**Table 7**

Final set of meta-paths generated by Phase – 1 of HeteClass for Fashion – 1

and Fashion – 2 datasets.

Meta-path

Length

Symmetric

*Photo* – *Author* – *Photo* (PAP)

2

是

*Photo – Tag – Photo* (PTP)

2

是

*Photo – Author – Photo – Tag – Photo* (PAPTP)

4

没有

*Photo – Tag – Photo – Author – Photo* (PTPAP)

4

没有

*Photo – Author – Photo – Author – Photo* (PAPAP)

4

是

*Photo – Tag – Photo – Tag – Photo* (PTPTP)

4

是

For Fashion – 1 and Fashion – 2 datasets, the set of meta-paths

generated are listed in [Table 7](https://translate.googleusercontent.com/translate_f#12) 。 For these datasets, total six meta-

paths are generated in Phase – 1 of HeteClass considering network

schema shown in [Fig. 11](https://translate.googleusercontent.com/translate_f#9) 。 Out of six meta-paths, four meta-paths

are symmetric and the rest two meta-paths are asymmetric.

**6. Results and discussion**

In this section, we present the results of the comparison of

HeteClass with other algorithms. Since LLGC and wvRN algorithms

cannot combine the semantics of meta-paths, we apply these al-

gorithms to the homogeneous information networks created by

following each meta-path individually. However, HetPathMine can

perform weighted combination of the networks corresponding to

the symmetric meta-paths ( [Luo et al., 2014; Sun et al., 2011](https://translate.googleusercontent.com/translate_f#17) [)](https://translate.googleusercontent.com/translate_f#17) .

Therefore, HetPathMine utilizes only symmetric meta-paths simul-

taneously.

*6.1. Results for DBLP – 1 and DBLP – 2 datasets*

The accuracy results for DBLP – 1 and DBLP – 2 datasets are

shown in [Figs. 12](https://translate.googleusercontent.com/translate_f#10) and [13](https://translate.googleusercontent.com/translate_f#11) respectively. From the figures, it is clear

that HeteClass outperforms the considered algorithms for classifi-

cation of authors for these two datasets. HeteClass has consistently

outperformed other algorithms and its performance is stable. Hete-

Class gives an improvement in accuracy of more than 3% over Het-

PathMine. From the results, we can see that for LLGC and wvRN

algorithms, meta-path APCPA performs better as compared to the

rest of the meta-paths. The accuracy of LLGC and wvRN is not good

for meta-paths APA and APAPA. The reason might be that these

paths are not capturing the semantics required for good classifi-

cation accuracy. However, meta-path APKPA has given better per-

formance for LLGC and wvRN as compared to meta-paths APA and

APAPA.

This shows that the classification of objects in a heterogeneous

information network by leveraging semantics of various meta-

paths is more effective than transforming the heterogeneous in-

formation network into a homogeneous information network fol-

lowing a single meta-path. HeteClass combines the semantics of

various meta-paths and performs the classification of objects more

准确。

|  |
| --- |
| **第13页** |

118

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

**Fig. 15.** Accuracy results for Fashion – 2.

*6.2。* *Results for fashion – 1 and fashion – 2 datasets*

For Fashion – 1 and Fashion – 2 datasets, the six meta-paths

listed in [Table 7](https://translate.googleusercontent.com/translate_f#12) are utilized. However, HetPathMine cannot utilize

all six meta-paths as there are two meta-paths which are asym-

metric (ie PAPTP and PTPAP). Since HetPathMine works only for

symmetric meta-paths ( [Luo et al., 2014; Sun et al., 2011](https://translate.googleusercontent.com/translate_f#17) ), there-

fore, we separately compare the performance of HetPathMine with

HeteClass using only symmetric meta-paths.

*6.2.1. Comparison of heteclass with wvRN and LLGC using all six*

*meta-paths*

[图](https://translate.googleusercontent.com/translate_f#12) [14](https://translate.googleusercontent.com/translate_f#12) and [15](https://translate.googleusercontent.com/translate_f#13) shows the results of the accuracy of classifi-

cation for datasets Fashion – 1 and Fashion – 2 respectively. 从

the results, it is clear that the performance of HeteClass is better as

compared to wvRN and LLGC following all six meta-paths. 我们可以

also see that the performance of HeteClass is stable and it consis-

tently gives the almost same accuracy of classification for different

sizes of training data.

From the results, we can see that the meta-path PAPTP gives

the good accuracy for LLGC algorithm. However, for wvRN, meta-

paths PAP and PAPAP gives the good accuracy for classification.

From the results, we can see that leveraging simultaneously the

semantics of all meta-paths gives the better results than follow-

ing a single meta-path. We can also see that LLGC algorithm, for

meta-path PAPTP, has performed slightly better than HeteClass for

Fashion – 1 and Fashion – 2 datasets when labeled photos were

10% (ie for *x* = 10).

**Fig. 16.** Accuracy results for Fashion – 1.

*6.2.2. Comparison of heteclass with hetpathmine using only*

*symmetric meta-paths*

Since meta-paths PAPTP and PTPAP are asymmetric, therefore,

we cannot utilize these meta-paths for HetPathMine as it can use

only symmetric meta-paths ( [Luo et al., 2014](https://translate.googleusercontent.com/translate_f#17) ）。 Therefore, we have

separately compared the performance of HeteClass with HetPath-

Mine by considering only symmetric meta-paths. The results for

datasets Fashion – 1 and Fashion – 2 are shown in [Figs. 16](https://translate.googleusercontent.com/translate_f#13) and

[17](https://translate.googleusercontent.com/translate_f#14) respectively. For this comparison, we utilized meta-paths PAP,

PTP, PAPAP, and PTPTP for both HeteClass and HetPathMine. 从

|  |
| --- |
| **第14页** |

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

119

**Fig. 17.** Accuracy results for Fashion – 2.

the results, we can see that HeteClass outperforms HetPathMine

for both datasets. The classification accuracy of HeteClass is about

4% better than the accuracy of HetPathMine.

Also, from the results shown in [Figs. 14](https://translate.googleusercontent.com/translate_f#12) – [17](https://translate.googleusercontent.com/translate_f#14) , we can see that

the performance of HeteClass when utilizing all six meta-paths is

slightly better than the performance when only symmetric meta-

paths are utilized. This shows that the full utilization of semantics

could be better than partially utilizing it and HeteClass is able to

do that.

*6.3. Significance of classification algorithm in heteclass framework*

In the framework of HeteClass, after forming the homogeneous

network in phase-2, we can apply the classification algorithm for

classification of unlabeled target type objects in the network. 该

performance of HeteClass depends on and orthogonal to the clas-

sification algorithm utilized in the framework ie the performance

of HeteClass can be improved by taking an advanced classification

algorithm.

However, in our experimental results, the performance gain

of HeteClass as compared to other algorithms is not only be-

cause of the chosen classification algorithm but also because

of the ability of the framework to extract the various seman-

tics in the network following different meta-paths and leveraging

them simultaneously. To show this ability of the proposed frame-

work HeteClass, we performed experiments on DBLP and Fash-

ion datasets using LLGC as classification algorithm in the Hete-

Class framework instead of Personalized PageRank algorithm. 然后，

we compared the accuracy results of HeteClass utilizing LLGC

(named as *HeteClass* + *LLGC* ) with the accuracy results of LLGC.

HeteClass+LLGC would be able to leverage all the meta-paths si-

multaneously, however, LLGC can utilize only one meta-path at a

时间。

*6.3.1. Comparing Heteclass* + *LLGC with LLGC for DBLP – 1 and DBLP*

*– 2 datasets*

The results of the comparison of HeteClass+LLGC with LLGC for

datasets DBLP – 1 and DBLP – 2 are shown in [Figs. 18](https://translate.googleusercontent.com/translate_f#14) and [19](https://translate.googleusercontent.com/translate_f#14) re-

spectively. As we can see that the performance of HeteClass + LLGC

is better as compared to the performance of LLGC following any of

the meta-paths taken for experiments for both DBLP – 1 and DBLP

– 2 datasets. It shows that the performance gain of HeteClass is not

only because of the classification algorithm in the HeteClass frame-

work but also due to the power of the framework to leverage the

various semantics in the network.

**Fig. 18.** Accuracy results for DBLP – 1 using LLGC as classification algorithm in Het-

eClass framework.

**Fig. 19.** Accuracy results for DBLP – 2 using LLGC as classification algorithm in Het-

eClass framework.

**Fig. 20.** Accuracy results for Fashion – 1 using LLGC as classification algorithm in

HeteClass framework.

*6.3.2. Comparing Heteclass* + *LLGC with LLGC for fashion – 1 and*

*fashion – 2 datasets*

For Fashion – 1 and Fashion – 2 datasets, the results of the

comparison are shown in [Figs. 20](https://translate.googleusercontent.com/translate_f#14) and [21](https://translate.googleusercontent.com/translate_f#15) respectively. 对于这些

datasets also the performance of HeteClass+LLGC is better as com-

pared to the performance of LLGC.

|  |
| --- |
| **第15页** |

120

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

**Fig. 21.** Accuracy results for Fashion – 2 using LLGC as classification algorithm in

HeteClass framework.

Thus, these experiments prove that the performance gain of

HeteClass is not only because of the classification algorithm in the

framework but also because of the power of the framework to

leverage simultaneously the various semantics present in the net-

工作。 It also proves that it is important to consider and leverage

the various semantics present in the network to improve the clas-

sification accuracy.

*6.4. Weight learning for Meta-paths in heteclass*

For leveraging the semantics of various meta-paths, it is re-

quired to determine the weight of each meta-path as some paths

may be good for classification than others. Since higher weight

should be assigned to paths which lead to good classification ac-

curacy, weight learning is an important step. HeteClass performs

weight learning from the labeled data for various meta-paths and

assigns the highest weight to the meta-path which gives the high-

est accuracy among all the meta-paths.

[Fig. 22](https://translate.googleusercontent.com/translate_f#15) shows the accuracy of HeteClass for combined as well

as individual networks corresponding to different meta-paths for

DBLP – 1 and DBLP – 2 datasets. From these results, we can see

that the accuracy of HeteClass corresponding to meta-path *Au-*

*thor – Paper – Conference – Paper – Author* (APCPA) is highest for

these datasets. HeteClass is able to learn this subtlety and assigns

the highest weight to this meta-path. [Table 8](https://translate.googleusercontent.com/translate_f#16) shows the weights

assigned to different meta-paths by HeteClass and HetPathMine.

HeteClass has assigned the highest weight (0.87 ∼ 0.96) to the

meta-path APCPA. HetPathMine has also assigned highest weight

to the same meta-path APCPA; however, HetPathMine has assigned

only around 0.43 ∼ 0.58 wt to meta-path APCPA. Since meta-path

APCPA gives significantly better classification accuracy among all

the meta-paths considered in these experiments, the weight as-

signed to this meta-path should be close to 1. HeteClass has as-

signed the weight to meta-path APCPA close to 1.

For Fashion – 1 and Fashion – 2 datasets, the accuracy of Het-

eClass for the networks following individual meta-path and all

meta-paths simultaneously is shown in [Fig. 23](https://translate.googleusercontent.com/translate_f#16) (a) and (b) respec-

tively。 For comparison with weight learning process of HetPath-

Mine, we have taken only the symmetric meta-paths while per-

forming the classification of objects. From the figures, we can see

that the paths *Photo – Tag – Photo* (PTP) and *Photo – Tag – Photo*

*– Tag – Photo* (PTPTP) has given almost equal accuracy and better

than other two paths. The weights assigned to different meta-paths

by HeteClass and HetPathMine are listed in [Table 9.](https://translate.googleusercontent.com/translate_f#16)

From [Table 9](https://translate.googleusercontent.com/translate_f#16) , we can see that HetPathMine has assigned same

weights to the paths PTPTP and PTP as they have given almost

equal accuracy. Similarly, paths PAPAP and PAP have been assigned

almost equal weights but less than the weights of paths PTP and

PTPTP. From the results, we can understand that the reason for

weight assignment is that the accuracy given by paths PAP and PA-

PAP is low as compared to paths PTP and PTPTP. However, HetPath-

Mine has assigned the highest weight to the path PTPTP and path

PTP has been assigned very low weight, even though following that

path we can get high accuracy. This shows that the weight learn-

ing process of HeteClass is more effective than the weight learning

process of HetPathMine.

*6.5.* *讨论*

From the results on both datasets ie DBLP and Flickr Fashion,

we can understand that meta-path based classification on target

type objects is effective. A meta-path contains the semantic which

could be highly significant for classification or other mining tasks.

Therefore, we should consider the semantics of meta-paths and

should leverage the semantics of various meta-paths simultane-

ously for effective results.

The proposed framework, HeteClass, demonstrates that lever-

aging the various semantics present in the network can improve

the classification accuracy on target type objects. The method pro-

posed in this work for assigning weights to various meta-paths

considered for a weighted combination of various semantics is

**Fig. 22.** Accuracy results of HeteClass for individual as well as for weighted combinations of networks corresponding to different meta-paths for DBLP – 1 and DBLP – 2

datasets.

|  |
| --- |
| **第16页** |

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

121

**Table 8**

Assignment of weights to different meta-paths for DBLP datasets.

Meta-path

Weights assigned by

Weights assigned by

HeteClass

HetPathMine

*Author – Paper – Author* (APA)

0 .001 ∼ 0.01

0 .24 ∼ 0.39

*Author – Paper – Conference – Paper – Author* (APCPA)

0 .87 ∼ 0.96

0 .43 ∼ 0.58

*Author – Paper – Author – Paper – Author* (APAPA)

0 .01 ∼ 0.08

0 .08 ∼ 0.13

*Author – Paper – Keyword – Paper – Author* (APKPA)

0 .001 ∼ 0.18

0 .1 ∼ 0.26

**Fig. 23.** Accuracy results of HeteClass for individual as well as for weighted combinations of networks corresponding to different meta-paths for Fashion dataset.

**Table 9**

Assignment of weights to different meta-paths for Fashion datasets.

Meta-path

Weights assigned by

Weights assigned by

HeteClass

HetPathMine

*Photo – Author – Photo* (PAP)

0.15 ∼ 0.2

0 ∼ 0.23

*Photo – Tag – Photo* (PTP)

0.2 ∼ 0.33

0 ∼ 0.1

*Photo – Author – Photo – Author – Photo* (PAPAP)

0.12 ∼ 0.19

0.1 ∼ 0.3

*Photo – Tag – Photo – Tag – Photo* (PTPTP)

0.23 ∼ 0.35

0.5 ∼ 0.9

effective and applicable for datasets from different domains. 该

weights assigned by HeteClass to different meta-paths are close

to the real-world understanding. Also, the flexibility of HeteClass

framework to utilize an algorithm of choice for classification of ob-

jects in Phase – 2 of the framework, makes HeteClass more useful

for different domains as we can choose the appropriate algorithm

for classification as per the context.

**7. Conclusion and future research directions**

In this paper, we studied the problem of transductive classifi-

cation on target type objects in heterogeneous information net-

works. For transductive classification on objects, we propose a

novel framework HeteClass, which is different from earlier ap-

proaches for classification in heterogeneous information networks.

The proposed framework explores the network schema to generate

a set of meta-paths for classification. By incorporating the knowl-

edge of domain expert using HeteClass, we can reduce the set of

meta-paths to select only those meta-paths that would be effec-

tive for classification. This would eventually reduce the computa-

tion time. HeteClass leverages the semantic subtleties of various

meta-paths by weight learning and performing the weighted com-

bination of various networks corresponding to each meta-path. Ex-

perimental studies performed in this paper demonstrate the effec-

tiveness of the HeteClass as compared to the baseline algorithms.

HeteClass also performs better as compared to the HetPathMine

algorithm in terms of classification accuracy and weight learning.

We can get meaningful insights and knowledge extraction by im-

proving the classification accuracy using the HeteClass framework.

Interesting future research directions include the extension of

the proposed framework for multi-label classification problem. 在

many real-world applications, a target type object in a heteroge-

neous network may acquire multiple labels. For example, in the

case of classification of movies according to genres is a multi-label

classification problem as a movie may have multiple genres. Also,

the proposed framework can be extended to, first, find the most

informative objects in the heterogeneous network and then acquir-

ing label information for those objects so that the overall classifi-

cation accuracy is improved for unlabeled objects. Also, it would

be interesting to see the behavior of the proposed framework for

those datasets which have objects having highly skewed class dis-

tribution.

**参考**

[Angelova, R.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0001) , Kasneci, G. , & Weikum, G. (2012). Graffiti: Graph-based classification

[in heterogeneous networks.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0001) [*World Wide Web, 15*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0001) [(2), 139–170 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0001)

[Deng, ZH](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0002) , Lai, BY , Wang, ZH , & Fang, GD (2012). PAV: A novel model for rank-

[ing heterogeneous objects in bibliographic information networks.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0002) [*Expert Systems*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0002)

[*with Applications, 39*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0002) [(10), 9788–9796 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0002)

[Gupta, M.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0003) , Kumar, P. , & Bhasker, B. (2015). A new relevance measure for hetero-

[geneous networks. In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0003) [*Big data analytics and knowledge discovery*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0003) [(pp. 165–177).](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0003)

[Springer International Publishing .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0003)

[Han, J.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0004) , Kamber, M. , & Pei, J. (2011). [*Data mining: concepts and techniques*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0004) [. Elsevier](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0004) .

[Haveliwala, TH](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0005) (2002). Topic-sensitive PageRank. In [*Proceedings of the 11th inter-*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0005)

[*national conference on World Wide Web*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0005) [(pp. 517–526). ACM .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0005)

|  |
| --- |
| **第17页** |

122

*M. Gupta et al. / Expert Systems With Applications 68 (2017) 106–122*

[Jeh, G.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0006) , & Widom, J. (2003). Scaling personalized web search. In [*Proceedings of the*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0006)

[*12th international conference on World Wide Web*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0006) [(pp. 271–279). ACM .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0006)

[Jeh, G.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0007) , & Widom, J. (2002). SimRank: A measure of structural-context similarity.

[In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0007) [*Proceedings of the eighth ACM SIGKDD international conference on Knowledge*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0007)

[*discovery and data mining*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0007) [(pp. 538–543) .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0007)

[Ji, M.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0008) , Sun, Y. , Danilevsky, M. , Han, J. , & Gao, J. (2010). Graph regularized transductive

[classification on heterogeneous information networks. In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0008) [*Machine learning and*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0008)

[*knowledge discovery in databases*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0008) [(pp. 570–586). BerlinHeidelberg: SpringerACM .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0008)

[Kong, X.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0009) , Yu, PS , Ding, Y. , & Wild, DJ (2012). Meta path-based collective

[classification in heterogeneous information networks. In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0009) [*Proceedings of the*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0009)

[*21st ACM international conference on information and knowledge management*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0009)

[(pp. 1567–1571). ACM .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0009)

[Lao, N.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0010) , & Cohen, WW (2010). Relational retrieval using a combination of path–

[constrained random walks.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0010) [*Machine learning, 81*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0010) [(1), 53–67 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0010)

[Liben-Nowell, D.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0011) , & Kleinberg, J. (2007). The link-prediction problem for social net-

[作品。](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0011) [*Journal of the American society for information science and technology,*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0011)

[*58*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0011) [(7), 1019–1031 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0011)

[Lu, Q.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0012) , & Getoor, L. (2003). Link-based classification. In [*ICML*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0012) [(pp. 496–503). AAAI](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0012) .

[Luo, C.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0013) , Guan, R. , Wang, Z. , & Lin, C. (2014). HetPathMine: A novel transductive clas-

[sification algorithm on heterogeneous information networks. In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0013) [*Advances in in-*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0013)

[*formation retrieval*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0013) [(pp. 210–221). Springer International Publishing .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0013)

[Macskassy, SA](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0014) , & Provost, F. (2003). A simple relational classifier. In [*Workshop on*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0014)

[*Multi-Relational Data Mining*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0014) [(pp. 64–76) .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0014)

[Macskassy, SA](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0015) , & Provost, F. (2007). Classification in networked data: A toolkit and

[a univariate case study.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0015) [*The Journal of Machine Learning Research, 8*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0015) [, 935–983 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0015)

[Rossi, RG](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0016) , de Andrade Lopes, A. , & Rezende, SO (2016). Optimization and label

[propagation in bipartite heterogeneous networks to improve transductive clas-](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0016)

[sification of texts.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0016) [*Information Processing & Management, 52*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0016) [(2), 217–257 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0016)

[Shi, C.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0017) , Kong, X. , Huang, Y. , Yu, PS , & Wu, B. (2014). HeteSim: A general frame-

[work for relevance measure in heterogeneous networks.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0017) [*Knowledge and Data*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0017)

[*Engineering, IEEE Transactions on, 26*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0017) [(10), 2479–2492 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0017)

[Sun, Y.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0018) , Han, J. , Yan, X. , Yu, PS , & Wu, T. (2011). PathSim: Meta path-based top-k

[similarity search in heterogeneous information networks.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0018) [*VLDB*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0018) [.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0018)

[Sun, Y.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0019) , & Han, J. (2013). Mining heterogeneous information networks: A structural

[analysis approach.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0019) [*ACM SIGKDD Explorations Newsletter, 14*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0019) [(2), 20–28 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0019)

[Zhang, M.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0020) , Hu, H. , He, Z. , & Wang, W. (2015). Top-k similarity search in heteroge-

[neous information networks with x-star network schema.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0020) [*Expert Systems with*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0020)

[*Applications, 42*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0020) [(2), 699–712 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0020)

[Zhou, D.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0021) , Bousquet, O. , Lal, TN , Weston, J. , & Schölkopf, B. (2004). Learning with

[local and global consistency.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0021) [*Advances in neural information processing systems,*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0021)

[*16*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0021) [(16), 321–328 .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0021)

[Zhu, X.](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0022) , Ghahramani, Z. , & Lafferty, J. (2003). Semi-supervised learning using gaus-

[sian fields and harmonic functions. In](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0022) [*ICML: Vol. 3*](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0022) [(pp. 912–919) .](https://translate.google.com/translate?hl=en&prev=_t&sl=en&tl=zh-CN&u=http://refhub.elsevier.com/S0957-4174(16)30546-2/sbref0022)