异质信息网络的研究现状和未来发展

石 川[1] 孙怡舟[2] 菲利普·俞(Philip S. Yu)[3]
[1]北京邮电大学
[2]美国加州大学洛杉矶分校
[3]美国伊利诺伊大学芝加哥分校

关键词：异质信息网络 数据挖掘

引言

现实生活中的大多数系统，是由大量相互作用、不同类型的组件构成的。为了更好地进行分析，通常将其建模为同质信息网络 (Homogeneous information network)。这种建模方法，往往只抽取了实际交互系统的部分信息，或者没有区分交互系统中对象及关系的差异性，这样往往会造成信息的不完整或信息损失。最近，越来越多的研究人员开始将这些互连的多类型网络化数据建模为异质信息网络[1](Heterogeneous information network)，并且通过利用网络中丰富的对象和关系信息来设计结构分析方法。与广泛研究的同质信息网络相比，异质信息网络包含全面的结构信息和丰富的语义信息，这也为数据挖掘提供了新的机遇与挑战。

基本概念

异质信息网络 [1] 被定义为一个有向图。它包含多种类型的对象或者关系，每个对象属于一个特定的对象类型，每个关系属于一个特定的关系类型。**网络模式** [1] (Network schema) 是定义在对象类型和关系类型上的一个有向图，是信息网络的元描述。

图 1(a) 是一个由科技文献数据构成的典型的异质信息网络的实例 [1]。该网络包含三种类型的对象：论文、会议和作者，每篇论文有作者和会议的链路关系，每条链路属于一种关系类型。图 1(b) 是该网络的网络模式，描述了文献网络包含的对象类型（会议、论文、作者）和相应的关系（撰写 / 被撰写、出版 / 被出版、引用 / 被引用）。

异质信息网络分析中一个重要的概念是**元路径** [2](meta-path)。元路径是定义在网络模式上的链接两类对象的一条路径，形式化定义为 $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$，表示对象类型之间的一种复合关系 $R = R_1 \circ R_2 \circ \dots \circ R_l$，其中$\circ$代表关系之间的复合算子，$A_i$ 表示对象类型，R_i 表示关系类型。

元路径不仅刻画了对象之间的语义关系，而且

[1] 作者将Homogeneous/Heterogeneous information network翻译成同质/异质信息网络，也有学者将其翻译为同构/异构信息网络。由于同构/异构网络容易和通信网络中的同构/异构网络的概念混淆，所以作者认为同质/异质网络更能反映网络中节点和边的类型和性质不一样的特性。

能够抽取对象之间的特征信息。图 1(c) 显示了文献网络中两个元路径的例子，分别简记为“APA”和“APVPA”(A、P、V 分别表示作者、论文和会议类型)。可以看出，基于不同的元路径，对象之间的语义关系是不同的。元路径“作者 - 论文 - 作者”(APA)，表示两个作者合作撰写了同一篇论文；元路径“作者 - 论文 - 会议 - 论文 - 作者”(APVPA)，表示两个作者在同一会议上发表了论文。链接两类对象的不同元路径，表示了不同的语义关系和链接网络，这也造成了不同的分析结果和特征表示。

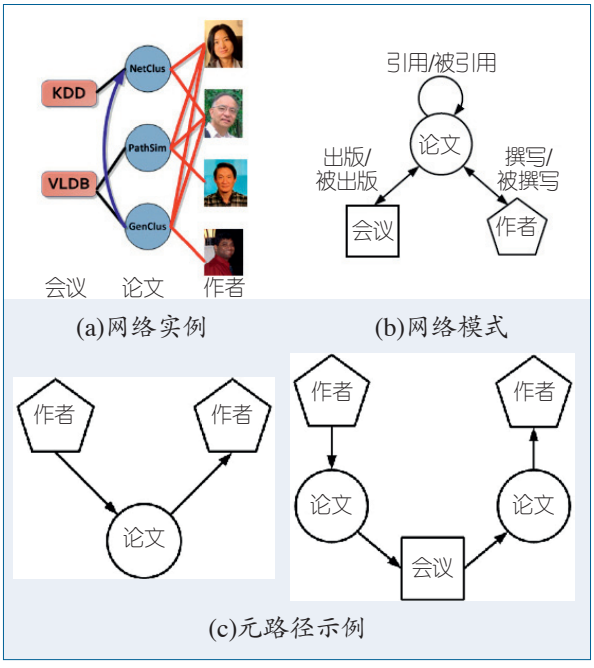


图1　由文献数据构建的异质信息网络

从实际情况看，大多数系统都存在多种类型对象的相互交互。例如，社交媒体网站（微信和微博）包含多种类型的对象（用户、帖子和标签）和这些对象之间的复杂交互。医疗系统包含医生、病人、疾病和设备等，以及他们之间的交互。一般来说，这些交互系统都可以被建模为异质信息网络。传统的同质网络建模方法只是抽取了这些交互系统的部分信息，而且这些信息往往也可以从异质信息网络中推导出来。例如，作者合作网络，就可以通过元路径“APA”从科技文献网络中得到。

异质信息网络分析

作为数据挖掘的重要研究方向，在过去的 20 年里，网络分析方法已经被深入研究，并且应用于很多数据挖掘任务，在这些工作中往往将网络化数据建模成同质信息网络。然而，异质信息网络的一些独特特征使得异质信息网络分析变得十分重要。**首先，异质信息网络分析是数据挖掘的新发展。**近年来，涌现出大量的社交媒体网站，以及不同类型对象之间的复杂交互。将这些相互作用的对象建模为同质网络是很困难的。然而，使用异质信息网络为其建模却是很自然的方式。大数据的一个显著特征是数据的多样性，作为半结构化的表示方法，异质信息网络可以对大数据中复杂多样的数据进行有效建模和处理。**其次，异质信息网络是融合更多信息的有效工具。**与同质网络相比，异质网络可以融合更多类型的对象及其之间复杂的交互关系，也可以融合多个社交网络平台的信息。**此外，异质信息网络包含丰富的语义。**在异质网络中，不同类型的对象和链接共存，它们具有不同的语义含义，在数据挖掘任务中考虑语义信息将导致更细微的知识发现。同质网络中的大多数方法并不能直接应用于异质网络中，因此，在异质信息网络中发现新的模式是十分必要的。

研究现状

研究现状概述

异质信息网络为更好地分析网络化数据提供了

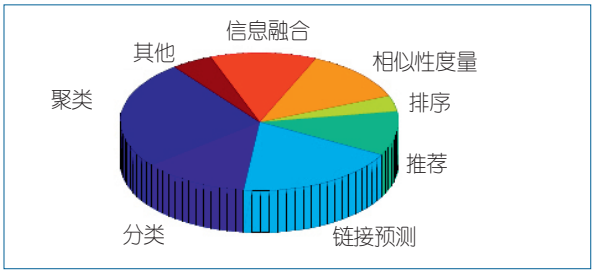


图2　异质信息网络分析相关论文的分布情况

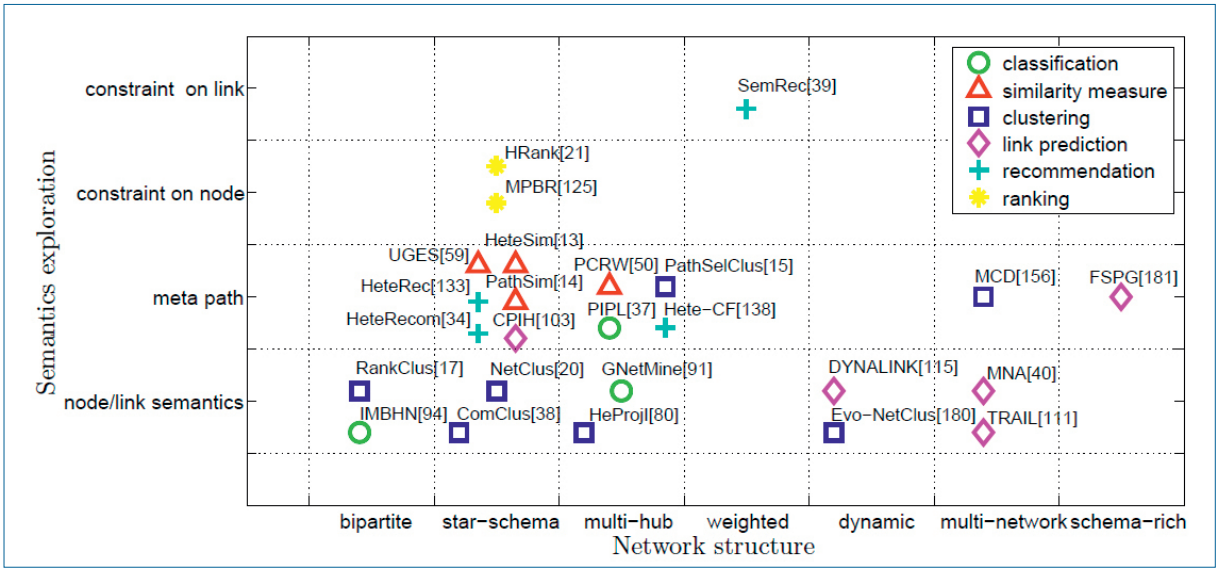


图3 从网络结构和语义探索两个维度对异质网络典型工作的总结

一种新的研究模式，同时也给许多数据挖掘任务带来了新的挑战。很多基于异质信息网络的数据挖掘问题已经被广泛研究，图 2 是对过去 6 年这一方向相关研究工作的近 200 篇论文按照研究问题分类的分布图 [1]。从图中可以看出，异质信息网络已经广泛应用于主要的数据挖掘问题，特别是相似性度量 [2]、聚类 [3]、分类 [4]、链接预测 [5]、推荐 [6] 等任务。

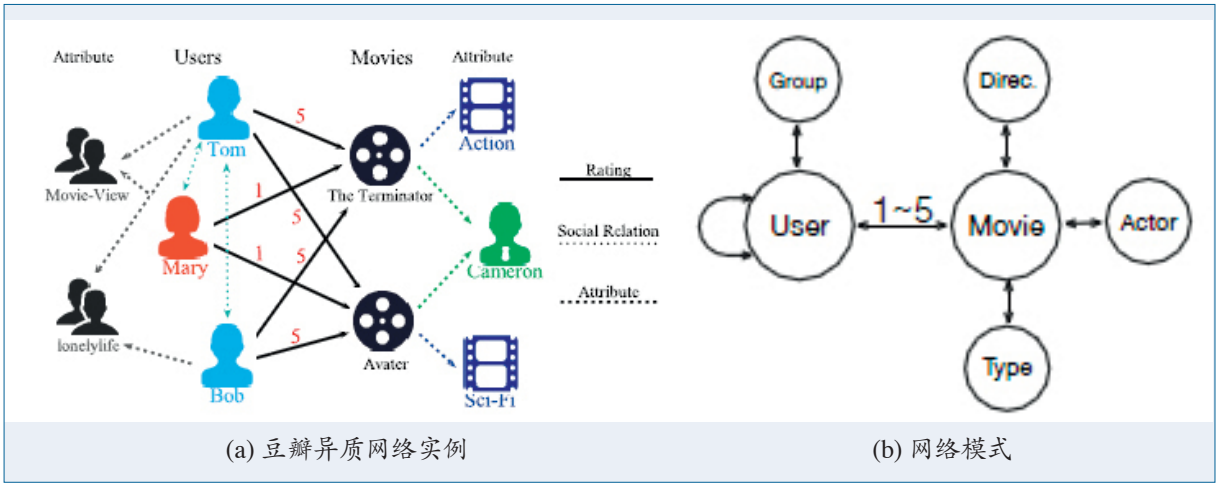
异质信息网络建模的优势在于可以整合更多信息和包含丰富语义，但是同时也会造成异质信息网络分析的难点，即如何有效地利用异质信息和探索丰富语义。作为有效利用异质信息和探索语义的工具，元路径将被广泛应用于异质网络分析。例如，PathSim[2] 利用对称元路径抽取两个节点之间的连通路径来度量二者的相似性，这样不仅利用了相关的异质信息，而且体现了节点和边的丰富语义。PathSelClus[3] 采用不同元路径抽取不同的网络结构进行节点聚类，并利用用户指导信息实现对聚类结果的整合。HCLP[5] 使用元路径构造边的特征向量，用于不同类型边的关系预测。很多机器学习技术都可以应用到异质网络分析中，例如随机游走模型 [7]、主题模型 [8]、矩阵模型 [6] 和概率模型 [3]。各类信息也都能够整合到异质网络分析中，例如属性信息 [2]、文本信息 [8] 和用户指导信息 [3]。

图 3 从网络结构和语义探索两个角度，总结了该领域的一些典型工作 [1]。沿着 X 轴，网络结构变得更加复杂；沿着 Y 轴，语义信息变得更加丰富。例如，PathSim[2] 可以处理星型模式网络，并使用元路径挖掘语义关系。SemRec[6] 在基本元路径上增加了链接的权值约束，以在带权异质网络中探索更微妙的语义信息。从图中我们可以发现，大多数研究都集中在简单网络结构（例如二分或星型模式网络）和基本语义探索（例如元路径）上，未来在利用更强大的语义探索工具分析更复杂的异质网络方面还需要做更多探索。

下面通过语义推荐这个应用，介绍异质信息网络分析的特点。更多应用可以参见相关论文 [1,9] 和专著 [7,10]。

基于异质信息网络的语义推荐

推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务上。面对推荐任务中常见的数据稀疏性问题，融合更多信息进行混合推荐是一种有效的解决方法。异质信息网络作为有效的信息融合方法，可以用于整合推荐系统中多种类型的对象和关系。图 4 展示了由电影推荐系统构建的异质信息网络，可以发现该异质网络整合了推荐系统



(a) 豆瓣异质网络实例　　(b) 网络模式

图4　豆瓣数据集构成的异质信息网络及其网络模式

中的打分、社交关系、属性等信息，而且网络中的节点和边包含了丰富的语义信息。因此，基于异质信息网络的推荐可能产生更加准确的推荐结果。

基于该框架，石川等人[6]提出了个性化语义推荐方法 SemRec，图 5 显示了该方法的基本思想。作为一种基本的推荐技术，协同过滤方法通过相似的用户对用户进行推荐。在异质网络中，可以利用元路径找到不同特性的相似用户。例如，通过元路径“UU”（U、M、T 分别表示用户、电影、电影类型），可以找到用户的朋友，这对应于社会化推荐；通过元路径“UMU”，可以找到具有相同观影记录的用户，这对应于传统的协同过滤。不同的相似用户有不同的推荐结果，有效整合这些推荐结果，可以产生综合的最终推荐。该方法还考虑了用户和电影之间打分关系上的分值（即关系权重），提出了带权异质信息网络和带权元路径等概念，以及相应的相似性计算方法。此外，该方法还采用组推荐技术对具有相同打分偏好的用户进行聚类。实验表明，由于融合了更多信息，该方法不仅具有更高的推荐准确性，而且能够有效缓解冷启动问题。此外，该方法能够根据用户的打分特性对用户进行聚类，较好地反映了用户群体特征。

推荐的可解释性一直是透明可信的推荐结果的必要条件，也是很多推荐模型所缺乏的功能。由于元路径的语义特性，SemRec 可以对推荐结果进行

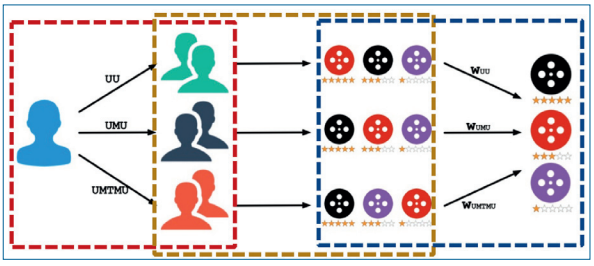


图5　基于元路径的语义推荐方法

解释，进而设计了可解释的语义推荐原型系统 RecExp[11]。图 6 的左上图显示了系统提供的由不同元路径产生的不同推荐功能；当用户选择混合推荐功能 (Hybrid recommendation) 时，图 6 的右上图将给出推荐原因，且根据权重最高的三条元路径的路径语义给出推荐原因。

未来发展

虽然异质信息网络已经应用于很多数据挖掘任务，但它仍然是一个年轻的、正在快速发展的研究领域。未来可能有如下研究方向。

更加复杂的网络构建

当前的研究，大多是假定异质信息网络是明确定义的，网络中的对象和关系是清晰的。然而，在实际应用中，用真实数据构造异质信息网络会遇到

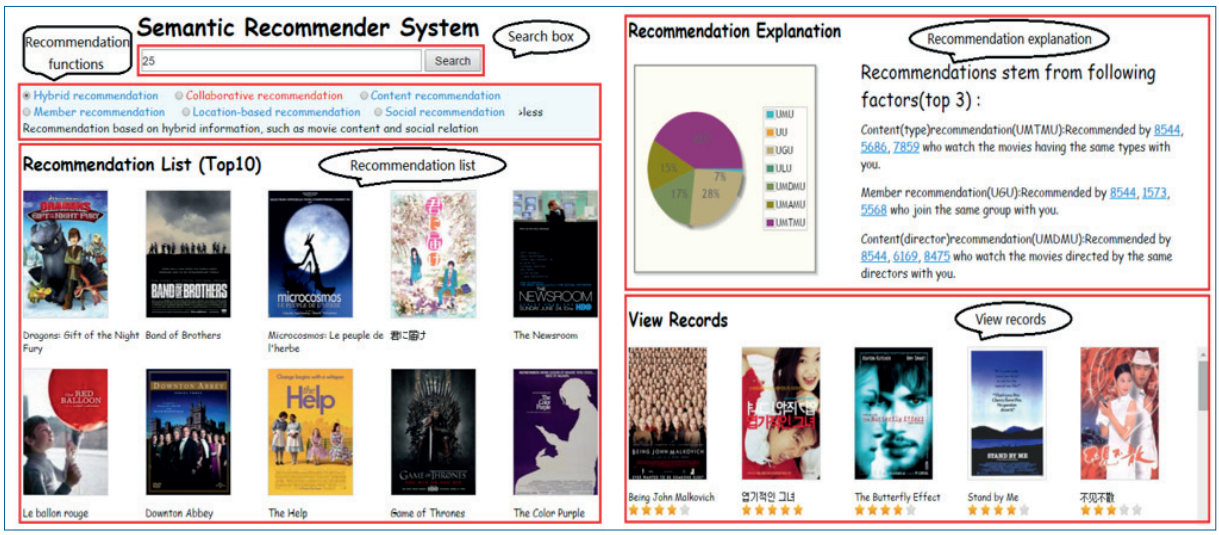


图6　可解释的语义推荐原型系统RecExp的主要功能界面

很多挑战。对于关系数据库之类的结构化数据，构造异质信息网络比较容易，然而即使是在这种网络中，对象和关系也可能具有噪声，会出现对象重名或关系不完整等问题；对于文本、图像等非结构化数据，如何准确抽取出相应的对象和关系，进而建立更加完善和准确的异质信息网络，也将面临更多挑战，在实践中会用到诸如信息抽取、自然语言处理、图像处理等技术。

更加强大的分析方法

在异质信息网络中，对象可通过不同的方式组织在一起。星型模式是广泛使用的异质信息网络类型，例如科技文献网络。之后，又出现了带环的星型模式和多中心网络等网络模式。在应用中，网络化数据通常更加复杂和没有规律性。某些网络中的链路会包含属性值，并可能包含重要的信息，这样就构成了带权异质信息网络，例如电影网络。在其他的一些应用中，用户可能存在于多个异质网络，这时需要对齐不同网络中的用户，有效融合不同网络的信息[12]。还有很多网络数据，例如知识图谱，就包含很多种类型的对象和关系，很难用简单的网络模式来描述[13]。这种丰富模式(schema-rich)的异质网络中也出现了很多新的研究问题，例如多种类型对象关系的管理以及元路径的自动产生等。这些复杂的网络化数据，给异质信息网络建模与分析提出了更多的机遇和挑战。

异质信息网络中的对象和关系包含着丰富的语义信息，而元路径可以捕捉这种语义信息。异质信息网络上的很多数据挖掘任务是基于元路径进行研究的，但是元路径在某些应用场景中并不能捕捉到精细的语义信息。例如，“作者 - 论文 - 作者”路径表示了作者之间的合作关系，但却不能描述特定条件下（例如KDD领域）的合作关系。为了克服这个不足，很多研究者提出了受限元路径[14]、带权元路径[6]、元结构[15]等概念，扩展了元路径的语义抽取能力。针对更加复杂的网络结构（如知识图谱），如何设计更加灵活精细的语义探索工具仍然需要进一步的研究。

近些年出现的深度学习在图像、自然语言等高维复杂数据处理上展现了优异的特征抽取能力，因此可以利用深度学习方法处理异质网络数据。当前深度学习和表示学习已经开始用于网络的结构特征表示[16]。异质网络中包含不同类型的节点和边，而且元路径体现了丰富的语义信息，这些特征使得异质网络的特征表示学习表现出很大的不同。异质网络的表示学习对异质网络分析提出了新思路，也为结构信息与其他模态信息融合提供了新途径。

更大数据的处理

为了展现异质网络建模的优势，我们需要在更广泛的领域中对大型网络化数据设计实用的数据挖掘算法。多样性是大数据的重要特征，异质网络是处理大数据多样性的有效方法。然而，构建一个真正的基于异质网络的大数据分析系统也是具有挑战性的工作。实际上，异质网络是巨大的，甚至是动态的，所以通常不能在内存中直接进行处理。由于用户往往只对一小部分节点、链接或子网络感兴趣，我们可以根据用户需求，从现有网络中动态地提取子网络进行分析。另外，设计基于异质网络的快速算法和并行算法也是亟需研究的内容。

其他一些研究方向也值得关注。相比于学习大数据深层特征的深度学习方法，最近兴起的广度学习 (broad learning)，整合不同类型的多个数据源进行融合学习，并在一些应用中取得了显著效果 [17]。由于异质信息网络是大数据时代整合不同类型数据的天然工具，因此结合异质信息网络研究广度学习方法，不仅会推动新的机器学习方法的发展，而且将为解决大数据的多样性提供新的思路。针对具体问题的异质网络分析系统也是重要发展方向。2017年 KDD 的最佳应用论文，就是利用异质网络和元路径构建和描述 Android 手机的 APP 应用和 API 调用的丰富交互，并将其用于恶意软件检测 [18]。这也为采用异质信息网络解决实际问题带来了启示。 ■

石　川

CCF专业会员。北京邮电大学教授。主要研究方向为数据挖掘、机器学习、演化计算。

shichuan@bupt.edu.cn

孙怡舟(Yizhou Sun)

美国加州大学洛杉矶分校助理教授。主要研究方向为数据挖掘、机器学习、网络科学。

yzsun@cs.ucla.edu

菲利普・俞(Philip S. Yu)

美国伊利诺伊大学芝加哥分校讲席教授，清华大学大数据科学研究院院长。ACM/IEEE Fellow。主要研究方向为数据挖掘、数据库、数据隐私。

psyu@uic.edu

参考文献

[1] Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(1): 17-37.

[2] Sun Y, Han J, Yan X, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks[J]. *Proceedings of the VLDB Endowment*, 2011, 4(11):992-1003.

[3] Sun Y, Norick B, Han J, et al. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks[C]//*Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2012:1348-1356.

[4] Kong X, Yu P S, Ding Y, et al. Meta path-based collective classification in heterogeneous information networks[C]//*Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM Press, 2012:1567-1571.

[5] Cao B, Kong X, Yu P S. Collective Prediction of Multiple Types of Links in Heterogeneous Information Networks[C]//*Proceedings of the IEEE International Conference on Data Mining*. IEEE Press, 2015:50-59.

[6] Shi C, Zhang Z, Luo P, et al. Semantic Path based Personalized Recommendation on Weighted Heterogeneous Information Networks[C]//*Proceedings of ACM International on Conference on Information and Knowledge Management*. ACM Press, 2015:453-462.

[7] Shi C, Yu P S. *Heterogeneous Information Network Analysis and Applications*[M]. Springer International Publishing, 2017.

[8] Deng H, Han J, Zhao B, et al. Probabilistic topic models with biased propagation on heterogeneous information networks[C]//*Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 2011:1271-1279.

[9] Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach[J]. *ACM SIGKDD Explorations Newsletter*, 2012, 14(2):20-28.

更多参考文献：www.ccf.org.cn/dl/publications/cccf