论文研究大规模符号网络的高效关联聚类算法，提出了两种符号网络的约简方法，改进了基于ILP模型的局部搜索算法与基于随机游走的符号网络关联聚类算法。采用人工生成符号网络与真实网络对相关算法与改进方法进行了评估。论文选题具有一定理论意义。

论文试图解决的科学问题描述不具体，研究动机不明确：解决关联算法应用到大规模符号网络中的哪些问题？论文中将“需要大规模集群服务器”、“需要很多先验知识”列为现有工作的不足，缺乏提炼。本文提出的相关约简算法与改进方法如何克服这些问题？论文研究结论没有回答解决这些问题的效果。

论文结构基本合理，但论述存在严重的问题。一方面，文字表达不通畅、不规范、不严谨，论述逻辑差，存在多处病句与标点使用错误、多处术语使用不一致（例如“节点”、“顶点”、“点”混用）；另一方面，文中多幅图片来自已有文献（20余幅图直接拷贝相关文献），不符合学术规范。在介绍论文主体工作时，依然大量混杂现有工作的介绍，无法区分本文的工作。

论文研究整体思路不够清晰，理论创新不突出，部分工作缺乏科学依据。例如，针对符号网络的特定结构特征提出的相关关联算法的普适性如何（第三章的工作）？论文没有给出支持上述假设的观测；“我们认为这样一条路径上的顶点有很高的概率在一个类中”（第10页）论断缺乏解释。

此外，论文还存在如下不足之处：

1. 多个图的坐标轴缺乏标注（指标与单位均缺失），图的结果分析不深入。
2. 多处算法缺乏详细的描述与解释；数学符号表示随意、不合规范。
3. 摘要写法不符合要求，无法表达出本文的实质性贡献。例如，每个创新点的描述没有给出具体解决的问题、解决思路与解决的效果。建议重新撰写。
4. 参考文献过于陈旧，近三年的文献极少；多条参考文献不合规范。

**修改后重审。**

**北京交通大学博士毕业论文评审意见**

该论文以提出高效的适合大规模符号网络的关联聚类算法为目的，提出了一个“网络约简+局部搜索”的关联聚类框架、两种符号网络的约简方法、一种基于ILP模型的局部搜索算法及一种基于随机游走的符号网络关联聚类算法，并在人工生成及真实的网络中验证并比较了论文所提算法的有效性和效率。

论文总结了目前关联聚类算法主要通过ILP和Multicut得到聚类结果，并发现ILP和Multicut算法难以适应大规模符号网络这一问题。为了解决这一问题，论文提出了“网络约简+局部搜索”的关联聚类框架：首先约简大规模的复杂网络，然后利用局部搜索的方法得到最终的聚类结果。在网络约简方面，论文基于类特征和Drop of Water原理提出了两种网络约简方法：前者基于符号网络中存在少数节点具有很高的正度，并且其中的每一节点都聚拢了大量的节点这一现象，找出符号网络中正度较高的若干节点，并将每一个这样的节点及它的“正邻居”当成一个类，然后在符号网络中通过将每一个类作为一个节点达到约减网络的目的；后者基于从度较大的顶点出发，沿着一个正权重变化最小的路径，则路径上的点以极大的概率属于同一类这一直觉，首先找到一个正度较大的点，然后将它与具有最大权重的“正邻居”合并，然后从出发寻找下一个节点，重复以上步骤直到满足预设的条件。在局部搜索方面，传统ILP模型的约束条件数量随着网络中节点数的增长而快速增长，导致在大规模符号网络中开销巨大。论文针对这一问题，利用实际的符号图都是非常稀疏的这一观察，提出了约束条件的数量由实际的边数来确定的改进的ILP模型，并将该模型运用到约减的符号网络中。论文进一步在人工生成及真实的网络中验证并比较了提出的两种网络约简方法及改进的ILP模型的有效性和效率。



针对现有一些符号网络的聚类任务中使用随机游走时通常仅仅在正边中游走，而负边则在游走的过程中被忽略这一问题，提出了让负边参与随机游走的方法。该方法从原符号网络中衍生出两个无符号网络，分别在这两个无符号网络中进行随机游走，得到节点之间的转移概率，通过分析转移概率的变化情况推断边的类型，最后使用一个贪心算法得到最终的聚类结果。论文进一步在人工生成及真实的网络中验证并比较了提出的方法的有效性和效率。

论文的结构安排清晰合理。论文的第一章和第二章从研究背景出发，阐述了关联聚类的研究现状，并基于现有研究工作分析存在的问题，提出了针对这些问题的研究思路。论文的第三章至第五章分别针对三个不同问题，提出了解决问题的核心想法及方案，并通过经验研究的方式，验证并比较了解决方案的有效性及效率。论文的第六章总结全文，阐述了研究的不足及改进方向。论文的语言逻辑性较强、文笔有待加强。

不足之处：

1. 文中缺乏对图和算法的描述，书写的算法和引用的数学符号不规范，并且需要对部分图进行进一步地加工。
2. 第三章提出的基于类特征的网络约简方案的实用性有待验证，该方法将度较高的顶点以及它的一步“正邻居”归为一类，即合并为一个顶点。根据Power-law度分布：网络中顶点的度分布极不均匀，少数顶点拥有极高的度，大部分顶点的度很小。理论上，在大规模符号网络中利用该方法得到的类数目是十分有限的，即约简后的网络和原始网络差别不大。建议在真实的大规模符号网络中应用该方法，对比约简后的网络和原始网络的差异。
3. 第三章提出了基于ILP模型的局部搜索算法，该算法将原始ILP模型的约束条件替换成Chopra和Rao的环不等式约束。建议进一步阐述Chopra和Rao的环不等式替换原有约束条件的动机。

该论文表明作者掌握了关联聚类的基本理论和关键技术，达到了博士学位论文水平及毕业要求。