关于论文《A Survey of Heterogeneous Information Network Analysis》的学习报告

过去人们研究的同构网络基本上只包含一种类型的网络节点，从而导致节点之间的连接也只有同一种类型。而现实的系统往往是由多种类型的部分构成的，而不同部分之间的连接类型因而也不仅限于同一种，这种由多种类型的节点、多种类型的连接构成的网络成为异构网络。相比于同构网络，异构网络包含了更丰富的网络架构信息与更丰富的语义信息，因此成为了近年来数据挖掘领域的一个较为热门的方向。研究者们可以从异构网络中提取出丰富的关于网络节点与网络链接的信息，并利用这些信息来协同挖掘出与网络相关的更重要的信息。

1. **关于异构信息网络的定义和基本概念**
   a) 异构信息网络：异构信息网络可以表示成一个有向图的形式，这个有向图包含点集 V 和边集 E，而这个点集 V 包含不止一种类型的节点或边集 E 包含不止一种类型的连接边。
   b) 网络模式：网络模式与网络实例相对应，网络模式是一个信息网络的模版。在网络模式中，每个组成部分都表示成类型的抽象形式，即某一类型的网络节点与其他（或同一）类型的网络节点之间存在了某种类型的连接关系。而网络实例就是网络模式的具体例子，网络实例中的每个组成部分都是具体的特定实体（如某一特定的 paper）。
   c) 元路径：针对一个特定的网络模式中的两种网络节点，这两种网络节点可能存在一定的关系，而这个关系是通过与其他节点的中间关系产生的。以这两种需要产生关系的网络节点为端点，依靠其他节点建立中间联系，这些所有的联系就构成了一条元路径。作为分析异构网络的方法中极为关键的一部分，元路径有着复杂而丰富的语义。即使对于相同的两个对象，以这两个对象为端点的不同元路径，一般都会有不同的含义，并且这些不同的元路径的侧重点也是不一样的。因此在异构网络的分析与运用中，需要针对不同的数据挖掘任务确定需要考察的元路径往往是算法需要考虑的基本点。

2. **为什么需要进行异构信息网络的分析?**
   a) 在数据挖掘领域中，挖掘出对象的特征与对象之间的连接关系已经成为一个主要的研究领域，而传统的同构网络中只存在一种类型的对象与连接已经不能真实地反映现实系统了，因此越来越多的研究者研究关于异构信息网络，以及异构信息网络中多种类型的连接关系，企图从中挖掘出更丰富的信息。
   b) 对异构信息网络的分析可以提取出更多的有用信息：由于异构信息网络的构成一般是糅合了来自多个源的数据，每个单一源可能只包含了整个网络的一部分网络成员信息，因此结合了多个源的数据的异构信息网络能比单一源的网络更加综合地考虑网络的信息，达到互补不足的目的。
   c) 异构信息网络包含更丰富的语义信息：正如前面提到的元路径的特点：即使对于相同的两个对象，以这两个对象为端点的不同元路径，一般都会有不同的含义，并且这些不同的元路径的侧重点也是不一样的。因此，从不同的元路径中可以挖掘到更多感兴趣的信息。

3. 研究发展领域
   a) 相似度测量：相似度测量的目的是测量网络中对象之间的相似性。在异构信

息网络中，相似性测量不仅考察了对象之间在网络架构上的相似性，由于元路径的存在，不同的元路径具有其不同的语义信息，导在不同元路径上的两个对象具有不同的现实关系，因此不同元路径上的相似性的含义也是不一样的。

b) 聚类：针对 HIN 中的聚类，聚类得到的同一个簇可以包含多个不同类型的对象，而这些对象是基于某些方面具有一定的相似性的，如属于同一个话题。由于 HIN 包含了众多丰富的信息，因此在 HIN 中进行聚类任务时可以同时考虑属性信息、文本信息和用户指引信息等。并且可以与其他任务，如排名任务等结合起来进行。

c) 分类：与传统的分类任务不同的是，HIN 中的对象由于存在各种各样的连接，因此不属于独立同分布。除此以外，HIN 中涉及的多种不同类型的对象应该同时进行分类。并且，由于 HIN 中的对象存在基于元路径的多种复杂关系，一个对象的类别会直接影响其他与其相关的对象的类别，从而导致分类过程中对象的类别信息可以通过各种连接进行传输。因此，HIN 中的分类可以看成是一个通过整个网络的知识传递过程。

d) 链接预测：链接预测的任务是从当前的网络结构与信息中预测一对潜在出现关系的对象之间是否真的存在关系。由于 HIN 中存在多种类型的对象与关系，因此在异构信息网络中执行链接预测任务时，需要协同考虑多种多样的复杂对象之间的关系来判断两个对象之间是否应该存在关系，这一点是具有极大挑战性的。

e) 排名：网络结构中的排名任务是利用排名函数来评估网络中对象的重要性与流行度。然而，当应用到异构信息网络中时可能会带来许多问题，比如对不同类型的对象进行共同排名其实是无意义的，并且 HIN 中不同类型的对象和连接具有不同的语义信息，而且基于不同元路径的侧重点也不一样，这一个因素就导致了可能会出现不同的排名结果。除此以外，由于存在不同类型的对象，而某一种类型对象的排名会影响另一个类型的对象。因此在 HIN 进行排名任务的时候需要综合考虑并处理好以上的问题。

f) 推荐：基于异构信息网络中包含的全面与丰富的信息，利用 HIN 提取到更多的信息进行推荐能得到更好的推荐效果。

g) 信息提取：前面提到，组成异构信息网络的数据一般是来自不同数据源的，因此综合考虑多个数据源的数据来进行信息提取应该能得到更全面的整合信息。更加值得关注的一点是，目前的许多研究已经转向从多个异构信息网络中提取信息。在多个 HIN 中，公共的节点通过 anchor links 建立联系，不同 HIN 中的信息通过这些 anchor links 进行传递信息。还有一些研究方向是利用这些 anchor links，多个对齐网络可以相互传递信息，因此某个网络源中的网络节点不仅可以得到当前网络中其他节点的信息，还能得到其他网络源的节点信息，从而使得不同网络中的信息可以相互补充，进而完善各自的网络，或缓解新生成网络的极大稀疏性问题（冷启动问题）。

4. 潜在研究方向

a) 更复杂的网络构成：鉴于异构信息网络的复杂性，把现实系统表示成异构信息网络并不一定有效。这可能是因为网络中的对象并不能准确地与现实世界的实体相对应，或对象之间的关系不明确，或来自多个数据源的信息不可靠、存在矛盾等原因造成。这就产生了在 HIN 领域进行信息提取、自然语言处理和关系提取的研究课题。

b) 更有效的数据挖掘方法：1）网络架构：将异构网络的信息表示成不同的形式能方便不同的任务操作，如二分图，k 分图和星型模式图等。不同形式的表示方法能携带不同程度的网络信息。除此以外，对象之间的连接权重也能携带一定量的信息；2）语义挖掘：虽然 HIN 中的元路径能有效地捕获不同对象之间的关系信息，但是元路径本身也存在一些缺点，如当前对元路径的不带权重的分析忽略了对象之间连接关系的重要性程度，并且元路径所能表达的信息也只能针对特定的方面。因此，如何跳出元路径的约束从其他方面来进行更深层的语义挖掘值得研究人员去研究。

c) 更大型的网络数据：由于现实世界中存在的应用网络非常大型，并且非常复杂，跨域多个领域，具有多样性，而 HIN 的灵活和有效的集成不同对象的异构性为这些大型网络的分析提供了一个有用的工具。然而，这也存在一些挑战，比如大型的网络结构由于其复杂性难以存储在计算机的内存中，因此不难直接处理和分析，这就产生了对网络进行信息提取（例如 network embedding）得到潜在的网络表示以及并行计算处理等研究方向。