# 6 总结与展望

## 6.1 总结

用一句话总结我的工作就是，在知识图谱中，具体一点在RDF知识库中采用数据挖掘中的常用方法生成关于实体集合和关系集合的概念。所以大体上可以分为三个方面来叙述：RDF知识库隐藏的信息、数据挖掘中的频繁模式挖掘算法、概念生成定义与实践。

首先，知识表示的方式多种多样，RDF三元组的形式是其中最简单最直观的表达方式，并且有很多项目组采用的就是此类数据库来构建知识图谱和知识库。当然，数据存在的同时必然也带来一些额外的信息，那就是数据之间的联系，除非在数据层面上直接表示该层联系，否则这类关于数据的数据将无法得知，因此我们的工作就是将隐藏的数据信息挖掘出来。其实这也就是数据挖掘广义上的定义。也因此我们遇到了项目进展中的第一个问题，RDF的频繁模式挖掘。现有的研究大多上注重于RDF的语义表达能力，对数据表达能力的研究比较少，因此少有直接面向RDF知识库的频繁模式挖掘算法。通过仔细分析，我们发现RDF知识库与图的表现形式有着近似的关系，RDF知识库可以看作是一个有多条边的巨大的有向图，因此采用图理论中经典的频繁子图挖掘算法是一种可用解法。但是实际上也存在一些问题，由于RDF知识库大多表示一张大图，在一张大图上的频繁模式挖掘算法有个不可克服的性质，因此必须重新考虑一种新解法。由此，我们将RDF知识库用一种标注方式拆分，原始数据集便转换为有向图集，利用模式增长的思想，此问题也就迎刃而解了。

其次，在图集上的频繁子图模式挖掘算法上，我们也做了一定的工作。在模式增长思想和DFS编码树的基础上，我们构建了一套在RDF上频繁模式挖掘的算法rdf\_fpm，并且提出了一些优化策略，比如排序剪枝策略还有前剪枝优化策略，从而将算法复杂度降为线性时间，而且减小搜索复杂度。我们也使用了常用的分布式计算框架，在最小频度阈值设置较小的情况下，开启分布式运算单元，可以有效地加快实验进程。其次，我们也做了大量重复实验，确保实验在算法性能上不至于影响应用层面。

最后，在概念生成的应用层面上，我们提出了两类概念生成。一类是关于实体集合的概念生成，另一类是关系集合的概念生成。在两类集合中，我们都采取的是频繁模式挖掘算法，所不同的是在实体集合中提出了概念树，将概念生成的问题转换为分析概念树的问题；而在关系集合中，频繁出现的模式就是一个定义下的概念，简单来说是一个模式发现的过程。同样，我们也做了大量重复实验保证概念生成模型的正确性。

## 6.2 工作展望

### 6.2.1 弱同构方式的改进

我们的整个频繁模式挖掘算法在扩展匹配过程中，要求的是强同构方式，即要求主语、谓语、宾语都一致才被认定为匹配成功，然后继续搜索扩展下一条边。实际上，这样会导致一个问题，两个明明相关度很强的实体因为描述上的宾语不一致而被判断为不相关，从而将概念生成的总数量大大减小，概念生成集合的密度也大大降低。因此，在判断匹配的过程中，忽略宾语一致性的弱同构方式尤其显得有意义。

### 6.2.2 关系集合中的不确定因素

我们的方法中关系集合概念生成必须是在实体集合概念生成之后完成，这是因为关系集合需要的概念-概念标注方式需要实体集合生成的概念提供。然而，实体集合生成的概念是带有一定可信度的信息的，在实验中为了将问题简化，我们将其概念可信度全部认作是1.0，这样就忽略了概念的不确定因素，因此挖掘出来的频繁模式也不一定十分准确。因此，后续的工作可以将这一因素也考虑进去，使得实验结果更加精确。