# 5 实验结果与分析

在本章中，我们将对上一章实现的算法进行一次详细系统的实验分析。这其中包括对概念生成模型正确性的验证，涉及到实体集合的概念生成以及关系集合的概念生成。然后是验证初始参数对实验结果的影响。最后将从三个优化策略来比对算法性能的提升。

## 5.1 实验目的

我们的实验目的主要关注三大方面，模型正确性，模型初始参数对实验结果的影响，以及提出的优化策略对算法性能的提升是否显著。

**实体集合概念生成** 通过对实验原始数据集的修改验证概念是否能重新生成，而且此概念与原始数据集已有的一致来验证正确性；

**关系集合概念生成** 通过比对该实验结果与一般性的实体属性描述的一致性来验证正确性；

**实验参数对实验结果影响** 主要是评测两个参数，一个是MIN\_SUPPORT，另一个是MIN\_CONFIDENCE对实验结果的影响；

**优化策略对算法性能的影响** 通过比对简单算法以及优化过的剪枝算法评价优化策略对算法性能的影响，同时对分布式的作用做一个评估。

## 5.2 实验环境

除RDF标注部分用python 2.7.8实现，其它算法部分全部使用Java SE 1.7实现，整个程序通过shell脚本启动。

实验设备是以一台单机Linux设备（Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00GHz \* 32和250G内存）为单位的实验室集群（\*4）。在所有的实验中，要求实验次数为3次重复实验。

实验所用初始RDF知识库数据集为Dbpedia 在当天下载的副本，格式为TTL，每一行由RDF三元组表示。

实验设计的所有参数机器默认取值总结在表中。

|  |  |
| --- | --- |
| 实验参数 | 默认值 |
| MIN\_SUPPORT | 10000 |
| MIN\_CONFIDENCE | 0.75 |

## 5.3 实验结果及分析

### 5.3.1 模型正确性验证

**实验1.1实体集合概念生成**。为了验证实体层面上的概念生成模型的准确性，我们采取一种删证法，即删除一部分概念，看能否通过我们的模型成功将这部分概念补全。首先我们给出在默认参数下生成的概念集合，如图，其中1.0的可信度表示在原始数据集中对于主语。。。，该概念描述是已经存在的。然后我们将该描述从原始数据集中删掉或打上标记，重新用shell脚本启动整个流程，同样跟踪主语标注的结果，可以得出实验结果如下。由此，该模型的正确性得到充分验证。

**实验1.2关系集合概念生成**。为了验证关系集合的概念生成模型的准确性，我们采取的方法是对于一种关系模式概念，如果找出具体的主语描述符合该关系模式概念，并且此类主语出现次数在默认最小频度之上，则代表关系集合概念生成的模型正确性得到验证。

### 5.3.2 实验参数对实验结果影响

**实验2.1 MIN\_SUPPORT实验参数**。MIN\_SUPPORT是该理论模型中最重要的一个参数，因为它控制着相关度的表达，只有那些频繁出现的描述大于此最小频度阈值的时候才代表着它们之间的相关性，并且继续扩展发现。

本实验将分别设置阈值为10000，5000，1000，100几个量级来检测参数对实验结果的影响，包括提取出的概念集合密度，整体概念集合数量。

**实验2.2 MIN\_CONFIDENCE实验参数**。MIN\_CONFIDENCE

### 5.3.3 算法性能分析

**实验3.1 排序剪枝优化的影响**。

**实验3.2 前剪枝策略优化的影响**。

**实验3.3 分布式运行的影响**。

## 5.4 本章小结