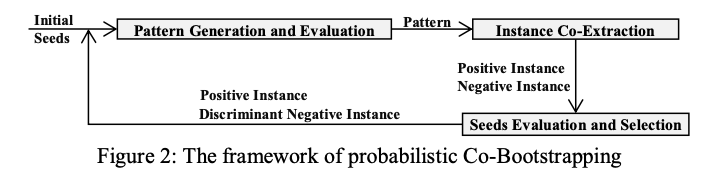
**论文阅读笔记一**

A Probabilistic Co-Bootstrapping Method for Entity Set Expansion

1. 贡献
2. 提出一种同时使用正负实例和判别负例来定义扩展边界的方法，boottrapping过程自动填充判别负实例。
3. 提出一种评分算法来估算新提取到的实例属于某一特定类别的概率，有效选择正例和判别负例
4. 通过上述联合扩展过程，维护和完善实体扩展的边界，解决语义漂移问题。
5. 方法



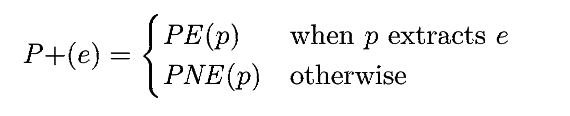
1. Pattern generation：统计正例和负例的比例来生成和评估模式；评估指标包括GE、PE、PNE

|  |  |  |
| --- | --- | --- |
| 指标 | 计算方法 |  |
| GE |  | P(p->e) |
| PE |  | P(p->e,e∈) |
| PNE |  | P(e∈p↛e) |

1. Instance Co-Extraction：根据统计信息和pattern质量来估计提取到的实例属于目标类别的概率

P+：实例为正的概率

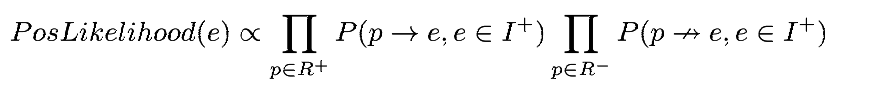
1. 只有一个pattern存在的情况下，P+的计算如下：



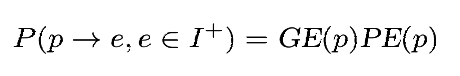
其中e表示实体，p表示提取到一个e的pattern

1. 多个pattern存在的情况，P+的计算

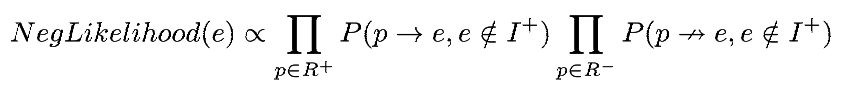
首先计算e为正的似然概率：



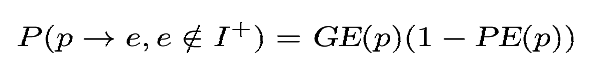
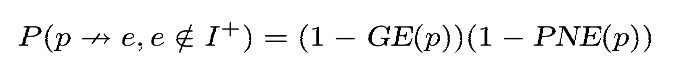
其中表示一个pattern集合生成一个e，表示所有的pattern集合都无法生成e,表示正例集合，乘积第一项是p生成e且e为正实例的概率，第二项为p无法生成e而e为正实例的概率，这两项军用贝叶斯公式进行计算。即



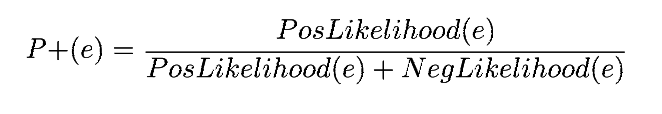
同样，一个e被预测为负例的情况：



两项依然用贝叶斯公式计算：

，

因此，P+的概率如下：



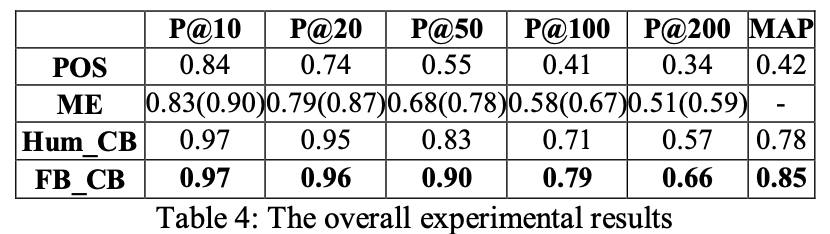
1. Seed selection：选择高分正例和判别负例

使用P+的阈值来确定实例的极性，高于特定阈值即为正例，否则为负例。

判别负例是那些与正例有比较高的重合度的负例，本文将那些在P+阈值附近的负例当作下一次迭代的判别负例，最终，相似pattern之间的边界即可用正例和判别负例来界定。

1. 实验

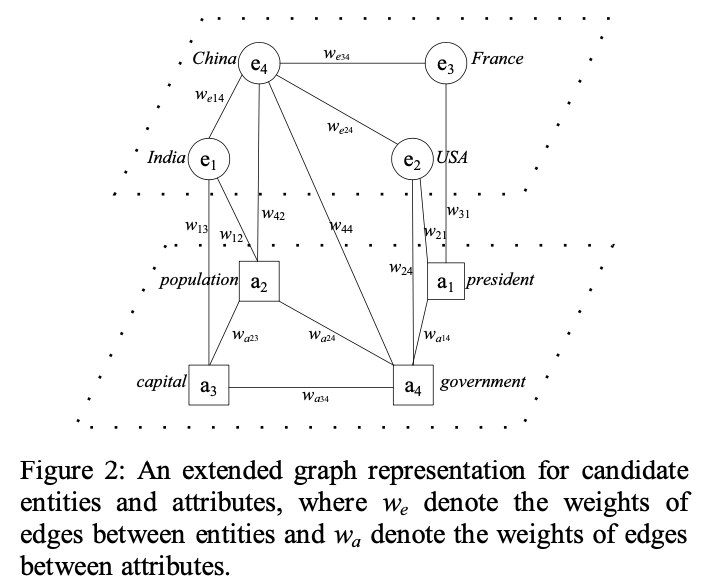
数据集：Google Web 1T



**论文阅读笔记二**

A Joint Model for Entity Set Expansion and Attribute Extraction from Web Search Queries

1. 贡献
2. 提出一个可以获取实体和属性之间关系的的图表示方法，
3. 一种基于无监督图的增强算法，挖掘新的实体和属性
4. 方法
5. 图的表示



其中上层的e表示候选实体，下层的a表示候选属性，表示实体与属性之间的相关性，表示实体之间的相关性得分、表示属性之间的相关性得分

1. 图的构建

Step 1: 产生候选实体和属性

通过种子实体去匹配模板，然后用模板去生成新的候选实体，然后用收集到的候选实体去提取候选属性；同时也在查询日志语料库中统计实体-属性对的频率,用于属性的过滤

Step 2：过滤候选实体和属性

利用候选实体与种子实体之间的相似性过滤候选实体；将每个候选实体表示为查询模式向量，向量每一维对应一个查询模式，将查询模式的频率作为权重；本文保留种子实体的前500个候选实体，然后保留≥5的前500个候选实体。

Step 3：添加边

共三种边：实体-实体，实体-属性，属性-属性

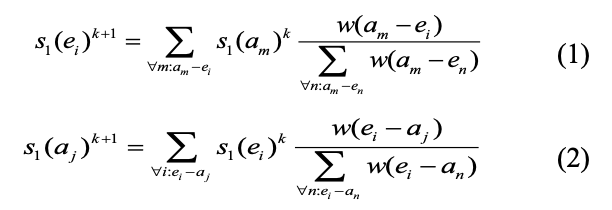
实体-属性：边由第二步确定，权重为1；

实体-实体：如果两个实体的查询模式向量的余弦相似度超过特定阈值δ(δ=0.6)，权重则为边的权重；

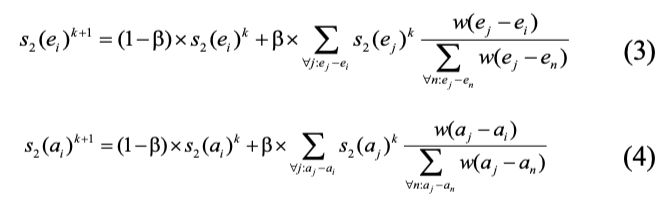
属性-属性：余弦相似度超过特定阈值θ(θ=0.2)，该边的权重即为阈值。

1. 节点重要性计算

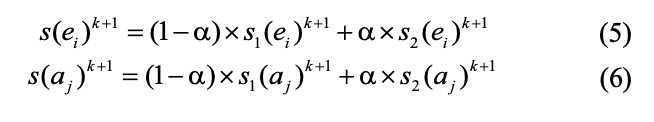
假设1：含有重要属性的实体趋于典型，由典型实体链接的属性趋于重要



假设2：由许多典型实体连接的实体是典型的，由许多重要属性连接的属性是重要的。



其中第一个部分是节点本身重要性，第二部分是邻居节点的重要性



通过公式5和6迭代更新实体和属性的重要性得分

1. 实验

数据集：AOL search data

