**Reference Type:**  Journal Article

**Record Number:** 24

**Author:** Jiaqing Liang, Yanghua Xiao

**Year:** 2017

**Title:** Graph-Based Wrong IsA Relation Detection in a Large-Scale Lexical Taxonomy

**Short Title:** Graph-Based Wrong IsA Relation Detection in a Large-Scale Lexical Taxonomy

**'File' Attachments:** internal-pdf://3324418910/document.pdf

===============================================================================

1. 针对有环图提出的两种方法：（基于Probase，一个能进行自动构建的分类系统（web））
2. 直接抽取无环的信息（无环子图）
3. 把有环的节点放在不同的level中

具体的针对对象是自动构建的KB，特别是词汇分类系统（lexical taxonomy）

常见方法：使用（查询）频率（词库很大，所以词库的幂律概率分布是一个长尾分布（follows a power-law distribution with a long tail, which implies that most relations with or without errors both have a low frequency），即在x轴上很长，但在y轴上很小（很短）的分布图，意味着不管使用频率大还是小，在分布图中的变化都不明显，即概率都非常小），使用外部知识库（缺点，一些KB中含有大量外部知识库没有的概念）

此方法被前人提出过，但存在两个问题：环是以指数级存在，多而难以枚举；不是所有的环都是错误的。

1. DAG decomposition based model（把识别错误的isA问题转化为识别有向无环子图）

定义：对于给定的有向图，找到一个有向无环图且为最小值，为剩余子图，即

MFAS model及其算法：

定义：给定一个边带权重的有向图，找到一个子图满足：1）是有向无环图，且为最小值。

算法流程：1）贪心算法——对于一个cycle，每次去掉weight最小的边，直到cycle变成acyclic

1. 对被移除的边按weight降序排列，并逐个逐个放回图中，如果不产生cycle则放回，否则，继续移除，时间复杂度为。具体步骤如下：

Basic metric：利用isA在XXX的frequency构建基础标准（边的置信度规则），表示边的frequency，显然，frequency越高，边代表的isA关系的置信度越高。文中分别在frequency为1、（2,10）、（11,100）、（100，+∞）的四个范围内任取isA50个样本，并人工判断其正确性，分别为78%、86%、94%、100%。basic metric的缺点是对低频的边（low frequency）的判别性较差，即如果一条边的frequency很低，则只有一小部分是错误的（即低频中还有很多错误的isA关系无法找到）。

Improved metric：对于数据驱动的分类系统，有个显著的特点——abstract concepts总是有较多的下位词，而specific concepts or entities的下位词较少，甚至没有，从而提出，使用“下位词的数量”作为metric。计算如下：，也就是X和Y的下位词的数量差距越大，边的isA的置信度越高。Improved metric是作为一个辅助，用于在frequency一样的情况下，判断isA关系。所以综上，总的置信度为：

1. Level assignment based model（转化成聚类问题）

思想：根据低level指向高level，得到图的有向无环子图，则剩下的子图可认为是可能的错误isA关系集合，所以将问题转化为求图中结点的level，即找到一个level assignment function使得最优，且最小，，则为最可能的错误isA关系集合。

Level assignment

定义：a function , mapping from V to N, which is a non-negative integer representing the level of v.

Subgraph implied by Level assignment

定义：对于定义在LA的有向图（即），子图

Baseline model:改良版拓扑排序，即每次移除入度最小的结点而非入度为0的结点

Penalty metric（惩罚规则）：1）错误的程度越高，惩罚度越高（1，即level差值越大，则错误程度越高，“+1”是表示在同一level时也需要惩罚）；2）在错误的前提下，边的可信度越高，惩罚度越高（可信度可以根据某一具体指标，如权重weight），综上，惩罚值为：

Agony model :（常用于有向图的分层）

详述：在baseline model的基础上，增加了weight和Penalty metric，即结点带level，边带权重和惩罚值。将问题升级为 ，即找到一个LA方案，使得的惩罚值之和最小（当惩罚值之和最小时，去掉是边最少，得到的最优。可转化为，因为时，isA关系是正确的。

1. 3
2. 3
3. 3
4. 3
5. 3
6. 3
7. 好词：

Machine-generated-----automatically constructed

Precise-----precision-----accuracy-----machine intelligence-----quality

Prohibitive（禁止，不提倡）

Trustworthiness（可信度）