

Genealogy-MBW

1. 简介

Genealogy-MBW 来源于华谱系统 (<https://www.zhonghuapu.com/>), 是一份真实的家谱数据集, 这份家谱以元代大德七年 (1303 年) 徽州府婺源县出生、明朝搬迁到安庆府桐城县 (现枞阳县) 的一位吴姓祖宗开始 (数据集里编号为 0), 记录了至 2020 年 12 月最近一次修谱竣工时他的所有后代数据, 包括 23646 位有具体名字且属于这位吴姓的直系后代。为便于同家谱文化和图数据研究者共享, 华谱系统提供此份仅包含单独立世人物的家谱数据, 隐私信息已匿名处理。

华谱系统的家谱图数据库中, 节点表示家谱人物, 边表示人物之间的关系, 节点的属性提供人物的描述信息。如图 1 所示, 0、1、7323 为三个节点, 1、7323 分别是 0 的儿子和隔 16 代的男性后代。至 2022 年 7 月, 华谱系统已存储超过 1855 万人物节点和 1130 份家谱。

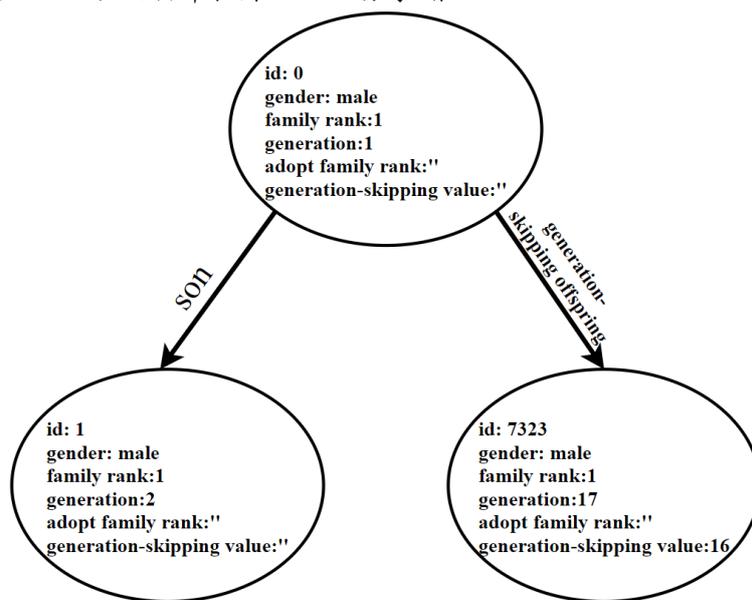


图 1 家谱中人物及关系示例

2. 详细信息

表 1 是 Genealogy-MBW 数据集的详细信息, 该数据集包含两个文件: Genealogy-MBW-nodes.txt 文件提供关于节点全部描述信息, Genealogy-MBW-edges.txt 提供关于边的信息, 详细描述见图 2、图 3。

表 1 数据集描述

图类型	节点总数	边总数	节点度数	节点属性数	边属性数	最大世辈
有向图	23646	24207	最小/最大出度: 0/13 最小/最大入度: 0/3 最小/最大度:1/14	5	10	28

表 2 是节点属性、属性值和特性的介绍,图 2 是 Genealogy-MBW-nodes.txt 文件中节点的介绍,其中属性值为空时表示该属性不适用。例如,“0,male,2,,1,,”表示 id 为 0 的男性人物,其家庭排行为 2,世辈为 1,过继家庭排行和隔代相连值不适用此人物;“8477,male,3,1,18,,”表示 id 为 8477 的男性人物,其家庭排行为 3,过继家庭排行为 1,世辈为 18,隔代相连值不适用此人物。

图 3 是 Genealogy-MBW-edges.txt 文件的介绍,其中的关系类型中英文对照表给在了表 3 中。表中隔代后代是一种特殊关系,它表示家谱中仅记录有某一人物的辈分和已知的最近直系先祖,而从这位直系先祖至该人物的中间人物均无法考证。为最大化保存家谱完整性,华谱系统中设计了这个新型关系来连接此人物和其直系先祖,在节点隔代相连值属性中保存其与直系先祖隔代值。

表 2 节点属性介绍

序号	属性名称(英文)	属性名称(中文)	属性值	特性
1	id	ID	0~23643	不可缺少,人物节点的唯一标识
2	gender	性别	male, female	不可缺少
3	family ranking	家庭排行	1~13	不可缺少
4	adoptive family ranking	过继家庭排行	1~4	可缺少
5	generation	世辈	1~28	不可缺少
6	generation-skipping value	隔代相连值	2, 5, 7, 8, 9, 14, 16, 17, 19, 20, 21	可缺少

0,male,2,,1,,
 1,male,1,,2,,
 2,male,2,,2,,
 3,male,1,,3,,

属性依次为：id，性别，家庭排行，
 过继家庭排行，世辈，隔代相连值，

图 2 节点文件 Genealogy-MBW-nodes.txt 介绍

0,1,son
 0,2,son
 0,7323,generation-skipping offspring
 0,9455,generation-skipping offspring

依次为：起始节点id，结束节点id，
 关系类型

图 3 边文件 Genealogy-MBW-edges.txt 介绍

表 3 关系类型中英文对照表

序号	关系类型 (英文)	关系类型 (中文)
1	son	儿子
2	daughter	女儿
3	step_son	继子
4	step_daughter	继女
5	heritor	嗣子
6	foster_son	养子
7	foster_daughter	养女
8	jiantiao_son	兼祧子
9	unconfirm_offspring	待考后代
10	generation-skipping offspring	隔代后代