**基于RDF图的语义地点skyline查询计算**

我们现在手里有一张图，这张图不是地图，而是对欧洲很多国家地理位置以及文化、艺术的抽象描述图。例如图1中的图G中有9个顶点，9个顶点中的是带有地理位置的顶点，除了地理位置，这些顶点也还有其他信息，例如顶点的名字是Florence，它的地理位置是意大利的佛罗伦萨，除了这个地理位置，的其他属性包括：city（城市）, European（欧洲）, Italian（意大利）；顶点的名字是Florence Cathedral，它的地理位置是意大利的佛罗伦萨大教堂，除了这个地理位置，的其他属性还包括：art（艺术）, myth（神话）, ancient（古老）。类似地，顶点分别是Athens（雅典）、Ancient Greece（古希腊）、Acropolis（卫城），他们也都有各自的属性，见图2。

除了这些带有地理位置的顶点，还有一些顶点，这些顶点没有地理位置信息，只有一些描述信息，例如的名称是House of Medici（麦地奇家族），的属性包括：renaissance（文艺复兴）, commerce（商业）, dynasty（王朝），这个顶点描述的是佛罗伦萨15世纪至18世纪中期在欧洲拥有强大势力的名门望族——麦地奇家族；的名称是Michelangelo（米开朗基罗），的属性包括：Italian（意大利）, sculptor（雕塑家），painter（画家），architect（建筑），这个顶点描述的是意大利文艺复兴时期著名的艺术家米开朗基罗；的名称是David（大卫），的属性包括：Sculpture（雕塑），art（艺术），history（历史），这个顶点描述的是米开朗基罗著名的雕塑大卫。类似点，顶点描述的是希腊哲学家Socrates（苏格拉底）。

图1中的顶点如果相互有某种关系，则存在边相互连接，例如佛罗伦萨大教堂（）在佛罗伦萨（），所以和之间有边连接，麦地奇家族（）是佛罗伦萨（）的名门望族，所以和之间有边连接。麦地奇家族在建筑和艺术方面进行了赞助，米开朗基罗（）得到过他们的赞助，所以和之间有边连接，米开朗基罗创作了著名的雕塑大卫，则和之间有边连接，这样就得到了图1。我们称图1是RDF图，图2是图1中各顶点的属性。

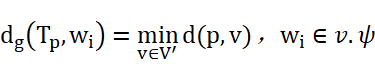
|  |  |
| --- | --- |
| 图 1 RDF图 | 图 2 顶点的文本属性 |

下面我们给出RDF图的数学定义。令有向图表示一个RDF图，其中是图中顶点的集合，是图中边的集合。在图中，带有空间坐标（地理信息）属性的顶点称作地点，例如图1中的等。图中，顶点的文本属性是一组词的集合，注意这个符号表示的是顶点的文字属性，例如，{renaissance, commerce, dynasty}。

定义1（**相关语义地点**）. 给定一组关键词和一个图，一个与相关的语义地点是一棵树，这颗树以地点为根节点，并且满足条件,，，这个符号的意思是所有顶点（顶点属于树）的文本属性的并集包含了关键词集合。

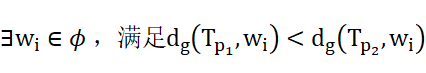
在我们给出图1的例子中，关键词{sculpture, art, history}，图1中有两棵树是与相关的语义地点，第一棵树以为根，树中其它所有顶点的文本属性的并集包含了这些关键词；第二棵树以为根，树中其它所有顶点的文本属性的并集包含了这些关键词。

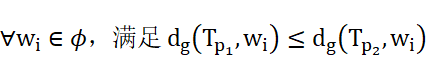
在一个相关语义地点中，根节点到一个关键词的距离用下面公式计算：



其中，是相关语义地点中到的最短路径的长度.

定义2（**语义地点支配**）. 给定一组关键词和一个图，令表示所有相关语义地点的集合，语义地点支配是上的一个偏序关系。这个偏序关系定义为：给定中的两个相关语义地点和，如果支配，则与之间具有偏序关系，记为。当且仅当下面两个条件满足，则称支配：



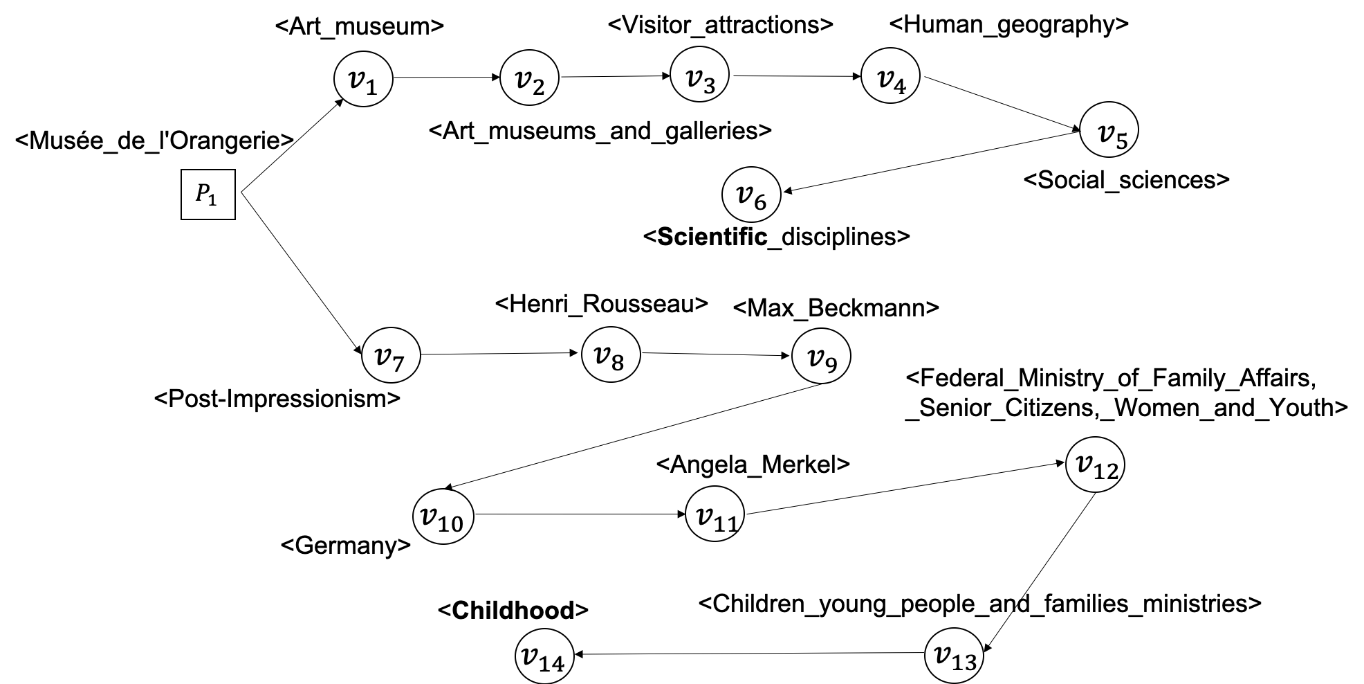


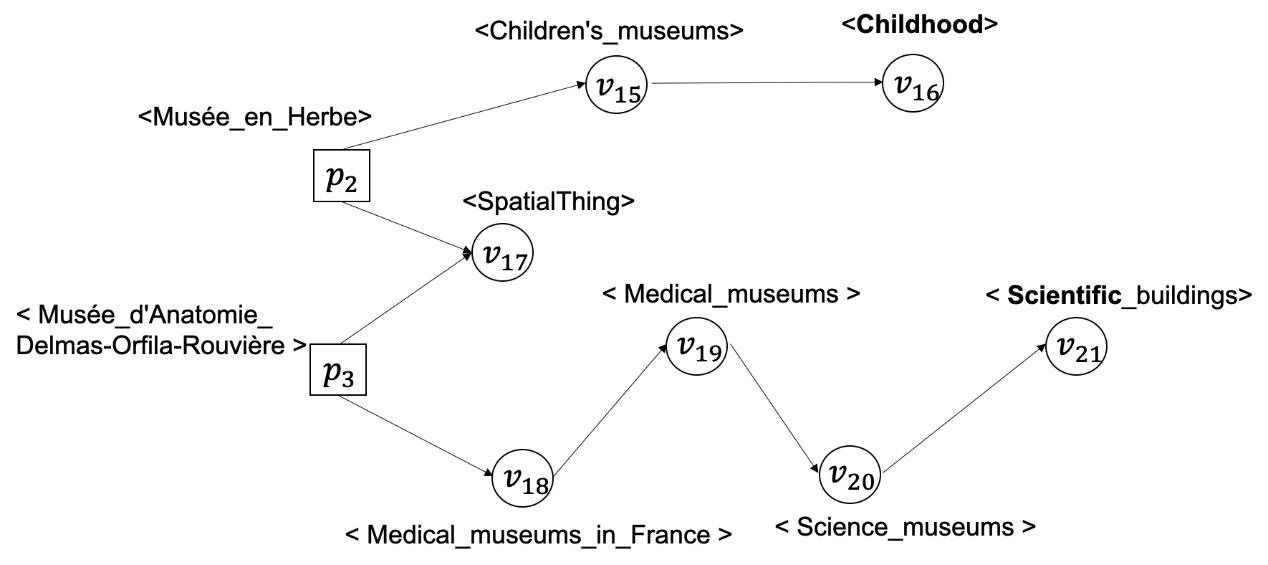
定义3（**语义地点skyline查询**）. 给定一组查询关键词和一个图,，令表示所有与相关的语义地点的集合，是一个语义地点skyline，当且仅当中不存在支配的语义地点。在图上,一个语义地点skyline查询的结果是中所有语义地点skyline的集合，记为。

举个例子，在图1中，给定一组查询关键词{sculpture, art, history}，图 1所示的RDF图，其中顶点的文本属性如图 2所示，其中和是两个相关语义地点。在这两个语义地点中，关键词到两个根节点的距离分别是，，，， ， 。根据语义地点支配的定义，可以看出支配，因为存在一些关键字，这些关键词在中的距离小于在中的距离，并且所有关键字在中的距离都小于等于在中的距离，这样就可以记为。那么语义地点skyline查询{sculpture, art, history}的结果是

**要求：**

1. （1）写一段文字描述，说明上述语义地点skyline查询可以用在什么地方？（2）设计一个算法计算语义地点skyline查询的结果并用伪代码描述；（3）假定下图中每个顶点旁<>里面的内容是该顶点的文本属性，下划线是单词分割符，例如下图中有四个文本属性。查询关键词，计算下图的语义地点skyline。





1. 附件中提供了Yago\_small是一个小规模图数据，请编写代码，在该数据上完成查询。
2. 附件中提供的Yago数据集包含 8,091,179 顶点和 50,415,307 条边，是一个大规模图数据，请编写代码，在该数据上完成查询，要求高效率完成查询。

数据集格式说明：

* 数据中顶点和关键词都用整数表示。
* 文件edge.txt是有向图的邻接链表，每一行代表一个顶点的邻接顶点。
  + 例如，8: 6291031,5330605,6481451,6280292，表示顶点8有四个邻接顶点，它们是6291031,5330605,6481451,6280292。
* 文件node\_keywords.txt是顶点的文本属性，每一行代表一个顶点的文本属性。
  + 例如，0: 8973992,10029808,8435980，表示顶点0有三个关键词，它们是8973992,10029808,8435980。

评分标准：

1. 撰写报告，要求论述问题，讲述求解算法思想，并给出伪代码描述，进行数据测试，并进行分析。