精读报告——徐柴俊

论文标题：Summarizing semantic graphs: a survey(VLDB 2019)

研究背景和动机

RDF的种类非常繁杂，数据也非常多，但是RDF的Schema结构往往是缺失的，即使有，可能也有上百万的量级，所以如何简洁地表示RDF数据就是一个需要解决的问题。

解决什么问题

目前已有的RDF summary的方法很多，它们使用的算法、输入输出等都会有一些差异，本文对这些方法进行了详尽的分类总结。

主要思路

不同的分类方法如下：

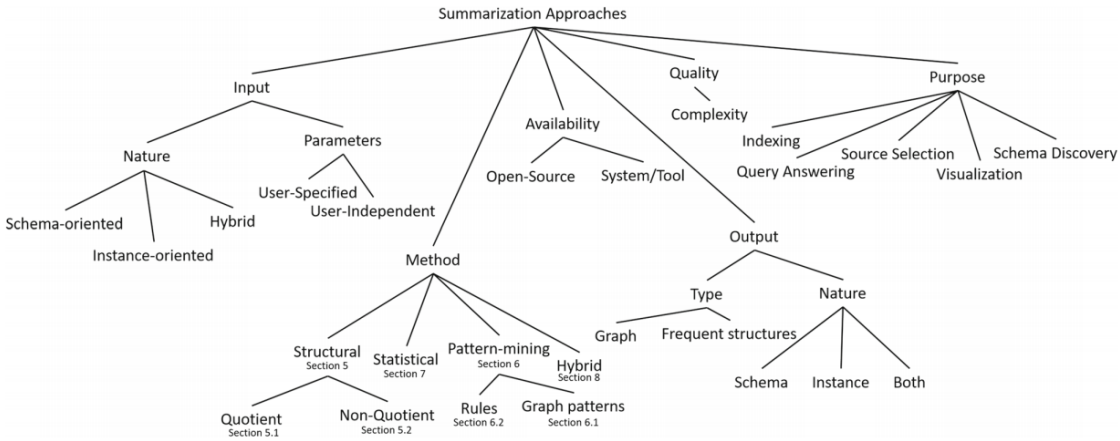

Fig. 4 A taxonomy of the works in the area

根据输入输出、目的、工具可用性、质量、算法思想等分类，本文的组织思路是按照算法思想来分类：结构化(图结构)、模式挖掘、统计信息等。

结构化(图结构)

商图

本文提到了商图的概念，即根据一定的算法将原图中的每个点划分成多个等价类，每个等价类在商图中是一个点，一种划分等价类的方法是根据它的拓扑结构，如果两个点的入边和出边的结构相同，则为同一等价类。具体的Backward bisimulation方法如下：

Definition 3 (*Backward bisimulation*) In an edge-labeled directed graph, a relation $\approx_b$ between the graph nodes is a **backward bisimulation** if and only if for any $u, v, u', v' \in V$:

1. If $v \approx_b v'$ and v has no incoming edge, then v' has no incoming edge;
2. If $v \approx_b v'$ and v' has no incoming edge, then v has no incoming edge;
3. If $v \approx_b v'$, then for any edge $u \xrightarrow{a} v$ there exists an edge $u' \xrightarrow{a} v'$ such that $u \approx_b u'$;
4. If $v \approx_b v'$, then for any edge $u' \xrightarrow{a} v'$ there exists an edge $u \xrightarrow{a} v$ such that such that $u \approx_b u'$.

Forward bisimulation即将入边改为出边，最终的Forward-and-Backward bisimulation则是两者的结合。

得到商图之后我们可以将商图作为索引，对于一个给定的查询，可以先在商图上做查询，匹配商图中的点，再根据索引匹配原图中的点。

非商图

商图将每个点划分成多个等价类，每个点只属于一个等价类。
但实际上可以让点属于多个等价类，其中RDF sentence graph是一个例子，它利用了RDF中空白节点周围的语义信息将RDF分成多个句子，不同句子中可以有相同的点。
还有一种想法是不用表示每个点，而是通过一些算法(比如PageRank)识别出最重要的点集合，构建summary。

模式挖掘

模式挖掘的基本思想是根据一些算法挖掘出图中出现的频繁的pattern，然后建立pattern和其对应的图数据之间的索引，查询中如果包含某个频繁pattern，则可以直接使用结果，比较类似于关系数据库中的视图。

统计信息

RDF图中的一些统计信息，比如某种类型的数量，属性的数量等等，这些信息可以帮助我们估计查询的代价，然后选择出最优的查询计划。

优缺点

- 优点：对于现有的RDF summary的方法做了非常详细的总结
- 缺点：

- 并没有对summary的质量做评估
- 没有考虑动态数据集上的summary方法

对工作的启发

文章中提到的在summary上做索引，然后查询时先在summary上查，再根据索引获得实际图数据上的结果的方法和我们目前在图数据库上的优化思路非常相近。
如果说之后我们在弱Schema约束的图数据库上做优化，也可以采用这种方法，先做summary，然后根据summary优化查询。