概率集相似性度量及应用

黄一夫
复旦大学计算机学院
学号 13210240015
ifhuang91@gmail.com

1. 主要内容

陈雷教授在新型数据管理技术课程上对不确定数据的研究进行了全面的介绍，主要包括不确定数据库上的查询（top-k, reverse top-k, nearest neighbor 等），以及不确定数据库上的频繁项挖掘。不确定数据的研究主要应用于数据清洗，数据整合等问题，例如文档的多标签分类，由于文档的标签一般是人为或者分类器给出，所以每个文档的每个标签拥有一定的可信度，该可信度即为不确定的表现，当然也可以从另外一个角度将其称为权重。不确定数据中最为基础并重要的一个问题是，如何度量两个概率集之间的相似度。基于两个概率集之间的相似性度量，我们可以进一步研究相似度查询处理，相似度连接等工作。

文献[1]对概率集从集合和元素的粒度分别给出了定义，但是我们采取文献[2]中的更加简单直接的定义方式。概率集与确定集类似，但是概率集中每个元素对应一个 0 到 1 之间的概率值，用来表示该元素在该概率集中出现的概率。形式化地，我们定义概率集 $A=\{a_i:p_{a_i}|a_i\in D,\forall i\in[1,n]\}$，其中 $\forall i\neq j,a_i\neq a_j,p_{a_i}>0$，例如{1:0.7, 2:1.0}。完成对概率集的定义之后，下一步是如何度量两个概率集之间的相似性。就概率集中每个元素的出现与不出现可以引入可能世界语义[3]，一个概率集 A 对应于许多可能世界，每一个可能世界 w(A)是概率集 A 的一个确定子集，其对应的概率可以表示为 $\Pr[w]=\prod_{t\in w}p_t\prod_{t\in w}(1-p_t)$，即把所有的情况列举出来，同时计算其出现的概率。因为对于两个确定集，现已有许多相似性度量的方法，所以对于两个概率集 A 和 B，首先将两者按其各自的可能世界语义展开成多个确定集{Ai}和{Bi}，然后定义方法计算多个确定集{Ai}与多个确定集{Bi}之间的相似度，最后按一定的策略汇总成为总的相似度。简单地，两个确定集之间的相似度可以由 Jaccard 系数给出，即两个集合 X 和 Y 的交集元素在 X 和 Y 的并集中所占的比例，即 $\mathrm{jac}(X,Y)=\frac{|X\cap Y|}{|X\cup Y|}$。对应地，与 Jaccard 系数相反的概念是 Jaccard 距离，表示为 $\mathrm{jac}_{dist}(X,Y)=1-\mathrm{jac}(X,Y)=\frac{|X\cup Y|-|X\cap Y|}{|X\cup Y|}$。参考文献[1][2]分别使用了 Jaccard 距离和 Jaccard 系数进行度量。上面的讨论给出了两个概率集间相似度的计算框架，可以发现其中的重点是，如何计算多个确定集{Ai}与多个确定集{Bi}之间的相似度。朴素地，我们可以对每一对确定集(Ai, Bi)求出相似度，然后对所有的相似度求取期望。但是这样不能表达所有(Ai, Bi)相似度的分布情况，因此还可以从累加概率分布的角度定义相似度。文献[2]正是从这两方面分别定义了两个概率集之间的期望相似度 ES 和基于自信度的相似度 CS。如果对计算性能有很高的要求，还可以考虑采用非可能世界的展开方法，或者考虑基于采样的近似计算。

在定义好两个概率集之间的相似度以后，接着要考虑如何将这些相似度高效地计算出。文献[2]指

出，因为采用了 Jaccard 系数度量公式，可以根据其自身特性，将所有的确定集对{(Ai, Bi)}划分成几个等价类，即拥有相同交集个数和相同并集个数的确定集对(Ai, Bi)为同一个等价类 H[i,j]，然后采用动态规划的方法高效地计算出。虽然这种算法效率高，但是其与相似度定义的耦合度也高，因此需要在确定相似度计算公式之后再进行启发式的设计。计算出概率集之间的相似度之后，便可以进一步应用于概率集的相似度查询处理和相似度连接。

概率集相似度查询处理。给定一个查询概率集 Q，给定一个有很多概率集的数据库{Oi}，同时还要给定一个相似度上限 t，要求查询后返回数据库{Oi}中与查询概率集 Q 相似度大于该相似度上限 t 的那些概率集。朴素地，可以用查询概率集和数据库中的所有的概率集进行一一的相似度计算，然后返回数据库中相似度大于该相似度上限 t 的所有概率集。但是其只适用于概率集数目较少的数据库，如果这个数据库是相当大的，这样的操作将花费很多的时间，因此需要启发式地对数据库中的概率集建立索引，并在查询处理时进行相应的剪枝。

概率集相似度连接。给定两个有很多概率集的数据库{Oi}和{Pi}，同时还要给定一个相似度上限 t，要求连接后返回相似度大于该相似度上限 t 的概率集对(Oi, Pi)。这个问题和上个问题的区别是，其对两个概率集数据库进行连接运算，找出那些相似度高于一个上限的概率集对。朴素地，可以采用上面概率集相似度查询处理的朴素思路，令数据库{Oi}中每一个概率集 Oi 为查询集，分别对数据库{Pi}进行查询处理，查询后返回相似度大于该相似度上限 t 的概率集对(Oi, Pi)。对于大的数据库，也需要启发式地建立索引，剪枝来加快处理。

2. 主要难点

前面介绍了概率集的定义，概率集间相似度的定义，以及其如何应用于概率集的相似度查询处理和相似度连接。下面依次介绍其中的主要难点：

前面提到，文献[1]对概率集在集合和元素粒度分别给出了定义，考虑到问题的简化处理，我们采取[2]中的更加简单直接的定义方式，其为一种特殊的元素粒度的概率集。由此，概率集的定义是一个难点，因为其不仅需要考虑在数学上足够广义，又需要符合实际应用中的数据。

因为一般的做法都是将概率集转化为确定集进行相似度计算，这样的话，需要对两个概率集分别找出所有的可能世界，然后再一一计算相似度，而这是指数级别的。因此，当概率集的基数变得很大时，这样的相似度计算效率是不可接受的，难点就是如何更加高效地进行相似度计算。同时，这也是目前研究中处理的概率集基数偏小的原因。另外，我们在考虑集合的相似性的时候，为了不失一般性，还应当考虑集合中元素间的相关性。

另外，针对于概率集相似度查询处理和概率集相似度连接等应用，还需要解决如何高效地计算一个概率集和一堆概率集之间的相似度。由前面提到的，一般都会采用建立索引，剪枝。而如何索引，如何剪枝就成为了难点。对于索引，文献[1]使用了 M-tree 进行索引，文献[2]对概率集的期望基数进行了索引，一般地，索引与所要解决的问题耦合度较高，需要针对具体的问题启发式地设计出高效的索引结构。对于剪枝，文献[1]基于三角不等式以及概率上界进行剪枝，文献[2]提出基于 Chernoff 上界的批量剪枝和单独剪枝，一般地，剪枝是通过上界的设定来完成的，越小的上界越能提高剪枝率，因此如何求得更加小的上界是难点。

3. 难点看法

从头谈起，首先我对概率集数据的来源存在一定的疑惑，因为在现实生活中，我们观察到的值都是确定的，而不确定的值一般来说是由多次抽样建立起来的，世界上没有天生存在的概率数据，例如文献[1][2]中实验部分的真实概率数据就是按作者自定义的方法建立起来的，因此这样建模的可行性还有待考虑。至于概率集的形式化定义，理应考虑得更加一般性，因此我觉得应该参考[1]中的定义，在解决问题前从数学的角度进行严格的抽象。但对于解决实际问题，我觉得应该参考[2]中的定义，因为其足够简单，并且相同格式的数据很容易通过采样的方法获取。

对于概率集相似度的计算，这里考虑的是两个集合，一般来说广泛采用的是 Jaccard 系数度量公式，但是这是建立在集合间的元素相互独立的基础之上，形式化地我们应该考虑更加一般性的相似度度量方法。对于基数大的概率集，一一枚举其可能世界会是指数级别的复杂度，因此想要高效地处理，一定要避免枚举概率集的所有可能世界。因此，我觉得可以再定义一个概率上限 t，概率集中超过概率上限 t 的元素存在，反之则不存在，这样可以将一个概率集转化为一个确定集，当然这是建立在对概率集失真压缩的基础上，但是其可以极大地提高处理效率，并可以调节概率上限 t 以取得自定义的效果。另外，我觉得也可以采用采样的方式对相似度进行近似计算，可以将通过采样一定的样本数量将相似度控制到一定的误差范围内。

然后是概率集相似度查询处理和相似度连接，一般来说都是建立索引，采用剪枝的手段进行。对于索引，索引结构的目的就是在查询未知的情况下进行一些预计算，这些预计算会使得之后的查询更加便利，我觉得没有最优的索引结构，应该针对具体情况进行启发式的设计，比如经常更新的数据库，还应考虑索引的可维护性。对于剪枝，剪枝的目的是求得更加紧的上界和下界以达到最大程度的剪枝率。我觉得首先需要从数学的角度严格地推导出尽可能紧的上下界，同时考虑到数据库中概率集的规模，还应该考虑基于外存，分布式策略的查询处理手段。另外考虑到剪枝的误差，应当将上下界设置为一个函数而不是一个定值，这样可以进一步提高剪枝误差的可控性。

4. 课程感想

该门新型数据管理技术课程对前沿的数据管理研究进行了介绍。王晓阳教授主要介绍了数据隐私保护，李飞飞教授主要介绍了大规模数据总结，陈雷教授主要介绍了不确定数据。老师们的热情度很高，用心地制作了课件，对知识的解释也是深入浅出恰到好处。课程安排的时间也很到位，早上从邯郸赶来不是特别吃力。其中我对陈雷老师介绍的不确定数据最感兴趣，并且在课下就概率集相似性度量及应用问题研读了他的论文以及最近相关的研究成果，并将其中不清楚的地方向陈雷老师讨教，他也很是热情认真得解答了我的问题。唯一遗憾的是裴建教授因为生病而没有给我们上课，我对他的话题很感兴趣。另外还希望老师将上课的 PPT 发布，方便我们的课后的消化。其实在上课前也可以将相关的论文，课件，代码发给我们，这样我们能在上课前有一定的准备，将得到更好的吸收。最后感谢汪卫老师对新型数据管理技术课程的组织安排，我从中获益很大。

5. REFERENCES

[1] Xiang Lian and Lei Chen. 2010. Set similarity join on probabilistic data. Proc. VLDB Endow. 3, 1-2 (September 2010), 650-659.

[2] Ming Gao, Cheqing Jin, Wei Wang, Xuemin Lin, Aoying Zhou. Similarity query processing for probabilistic sets. ICDE 2013.

[3] N. N. Dalvi and D. Suciu. Management of probabilistic data: foundations and challenges. In PODS, pages 1–12, 2007.