基于Zipf分布与属性相关性的选择性估计

姜芳艽
(徐州师范大学智能信息处理研究所 徐州 221116) (中国人民大学信息学院 北京 100872)

摘 要 在Deep Web数据集成中,集成查询接口和很多Web数据库查询接口用合取谓词表达查询,但是也有相当一部分Web数据库的查询接口用互斥谓词表达查询,这意味着查询转换时每次只能选择一个谓词。因此,准确、高效地估计每个互斥查询的选择性是优化查询转换的关键。提出了基于Zipf分布与属性相关性的选择性估计方法。通过属性之间的相关性从Web数据库上获取该属性近似随机的属性级样本,在此基础上计算属性值的Zipf分布方程,进而推断该无限值属性的任意值的选择性。实验表明,该方法可以准确、高效地估计各互斥查询的选择性。

关键词 Zipf分布,属性相关性,选择性估计
中图法分类号 TP311 **文献标识码** A

Selectivity Estimation Based on Zipf Distribution and Attribute Correlation

JIANG Fang-jiao
(Institute of Intelligent Information Processing, Xuzhou Normal University, Xuzhou 221116, China)
(School of Information, Renmin University of China, Beijing 100872, China)

Abstract In Deep Web data integration, some Web database interfaces express exclusive predicates, which permit only one predicate to be selected. Accurately and efficiently estimating the selectivity of each exclusive query is of critical importance to optimal query translation. In this paper, we proposed a novel selectivity estimation method. Firstly, we computed the Attribute Correlation and access approximately random attribute-level sample through submitting the query on the least correlative attribute to the real Web database. Then we computed Zipf equation aided by the information of word rank from the sample and the actual selectivity of several words from the real Web database. Finally, the selectivity of any word on the infinite-value attribute was derived by the Zipf equation. An experimental evaluation of the proposed selectivity estimation method was provided and experimental results are highly accurate.

Keywords Zipf distribution, Attribute correlation, Selectivity estimation

查询转换在Deep Web[1]数据集成中扮演着一个重要的角色,但是由于Web数据库的巨大规模、高度异构和自治性,使得自动的查询转换面临着极大的挑战。其中一个很重要的原因是Web数据库查询接口可能用不同的谓词逻辑表达。集成查询接口和很多Web数据库查询接口一般用合取谓词表达,$Qc=P1 \wedge P2 \wedge \cdots \wedge Pm$,这里$Pi$是在单个属性上的谓词,这类查询我们称之为合取式查询。而另一些Web数据库查询接口用互斥逻辑表达,$Qe=Pi(Pi \in P1, P2, \cdots, Pm)$,这意味着任何给定的查询只能包含一个谓词,这类查询我们称之为互斥查询。Web数据库查询接口的互斥属性一般由一个选择列表或者一组选择按钮表达。无论采用的是何种形式,每次只能选择一个属性提交查询。无论选择哪一个属性,其实都是默认在其他属性上不再有任何限制条件,因此均放宽了原有的用户查询条件。为了降低由此引发的查询代价,在所有可能的互斥查询Qe中,需选择选择性最强的属性进行查询转换,这是优化互斥查询转换的关键,也是本文将要深入探讨的问题。传统的异质数据源集成中,已在合取式查询转换方面展开了广泛的研究,但是关于互斥查询转换,这一Web数据库接口的特有现象方面的研究工作仍非常缺乏。

1 实例分析与问题描述

1.1 互斥查询实例分析

如图1(a)所示,如果集成接口的查询Qc为{〈Title,"Java"〉∧〈Subject,"Computer"〉},互斥查询Qe则为{〈Title,"Java"〉}或者{〈Subject,"Computer"〉}。最佳查询转换是从所有互斥查询Qe中选择限制性最严格的谓词进行转换。根据常识可以判断,因为"Subject"是分类属性,所以查询Qe{〈Subject,"Computer"〉}将比查询Qe{〈Title,"Java"〉}返回更多的无用结果,因此查询Qe{〈Title,"Java"〉}是一个较优的选择。但是,在很多情况下,我们很难决定如何选择。例如,图1(b)所示,如果集成接口的查询为Qc{〈Title,"Java"〉∧〈Author,"Jim"〉},则互斥查询Qe为{〈Title,"Java"〉}或者{〈Author,"Jim"〉}。对于这样的两个互斥查询,我们很难凭借一般常识判断哪一个返回的结果更少一些。

到稿日期:2009-12-31 返修日期:2010-03-18 本文受国家自然科学基金(60773216)资助。
姜芳艽(1971—),女,博士,副教授,主要研究方向为Web数据集成,E-mail:jiangfj@gmail.com。

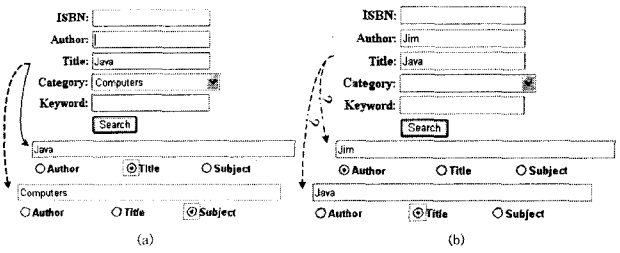

图 1　互斥查询转换的两个实例

查询选择性估计最杰出的解决方法是构建可以刻画属性值分布特征的直方图。其优点是直方图一旦构建好，在选择性估计时就可以多次使用。但是，构建直方图通常需要已知数据库的全部数据，而 Web 数据库的服务商一般不会提供所有的数据，这些不合作的 Web 数据库中的数据只能通过在其查询接口提交查询时获得。对于分类属性（例如 Subject 属性），由于可以枚举所有的属性值，因此可以将这些值通过 Web 数据库的查询接口提交并获得相应的返回结果的数量（绝大多数 Web 数据库提供返回结果数量信息）。由此可以获得属性值的分布特征，进而方便地构建直方图。但是直方图并不适用于所有的无限值属性（例如 Price 属性和 Title 属性）。对于数字型的无限值属性（例如 Price 属性），仍然可以构建直方图。许多已有的直方图的研究是基于数字型属性的，因此本文不再研究这个问题。

但是对于文本类型的无限值属性（例如 Title 属性），由于其属性值是无限的，枚举每一个属性值，然后逐个提交给 Web 数据库的代价是不可接受的，因此构建直方图的方法不再适用，需要寻找新的适用方法。这也是本文主要关注的问题，即文本无限值属性的选择性估计。

而且大量统计显示，Web 数据库查询接口上大部分属性都是文本属性。例如，在书的领域，无限值属性占 63%；在论文领域，占 71%；在电影领域，更是高达 76%。而且，在这 3 个领域中，Web 数据库查询接口用互斥逻辑表达的分别占 38%，27%和 9%，如图 2 所示。注意：本文以下部分所提无限值属性均为文本无限值属性。

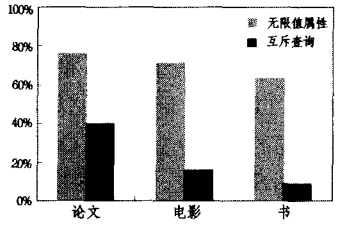

图 2　无限值属性及互斥查询所占的比例

1.2　问题描述

Web 数据库查询接口主要包括 3 种类型的属性：关键属性、无限值属性和分类属性。

定义 1（关键属性，Key Attribute，KA）　关键属性是指属性值能唯一标识表中元组的属性。

定义 2（无限值属性，Infinite-value Attribute，IA）　无限值属性是指有无限多个值的属性。在 Web 数据库查询接口上，这类属性一般用文本框表示。

定义 3（分类属性，Categorical Attribute，CA）　分类属性有一个较小的值的集合，这些值将所有元组分成有限的组。这类属性通常用一个选择列表或者一组选择按钮表示。

图 1 中，Subject 是分类属性，因为它通常有有限个值，例如文学、科学和技术等。ISBN 是关键属性，因为每本书的 ISBN 是唯一的。Title 和 Author 是无限值属性，因为两者都包含无限多个不同的属性值。

定义 4（互斥查询转换，Exclusive Query Translation）　给定集成接口的查询 $Qc=P1 \wedge P2 \wedge \cdots \wedge Pm$ 和某个用互斥谓词 $Qe=Pi(Pi \in P1, P2, \cdots, Pm)$ 表达的 Web 数据库查询接口，查询转换时将估计所有可能的本地查询 Qe 的选择性，并选择出查询结果最小的查询。

定义 5（查询的选择性，Query Selectivity，QS）　查询选择性即查询的限制性，表示为 Qe 的返回结果元组数占此 Web 数据库所有元组总数的比例。

对于一个给定的 Web 数据库，由于其全部元组的总数是一个常量，因此可以用返回结果的元组数衡量查询的选择性。

通常，3 种类型属性的限制性顺序，从严格到宽松分别是关键属性、无限值属性和分类属性，可以表达为

$$QS(KA) < QS(IA) < QS(CA)$$

显然，由于关键属性可以唯一标识元组，因此 QS（KA）的查询选择性等于 1。对于给定的查询以及给定的属性值，QS（CA）是查询的返回结果数（hit number）。对于给定的属性和属性值，查询的结果数可以通过发送探测查询给 Web 数据库获得。

但若比较两个包含无限值属性及其值的查询选择性则非常困难。主要原因如下：首先仅仅依靠属性粗略估计查询的选择性有时会出现较大的偏差，因为即使是在同一无限值属性上，只要属性取值不同，查询的选择性也有可能相差很大。例如，在学术搜索引擎 Libra 中，查询{〈Paper="network"〉}的选择性是 85889，查询 {〈Paper="autonomy"〉}的选择性是 1681，而查询 {〈Paper="metalearning"〉}的选择性则为 18。其次，若通过查询探测的方法来获得估计查询选择性所依赖的统计信息，则代价太高。如何以较小的代价获取较高的查询选择性估计，是本文的研究关键。

2　相关工作

数据库领域中应用最广泛的选择性估计方法是直方图[2-4]，其被成功地应用于几乎所有主要的商业数据库管理系统（例如 Oracle，Sybase，Microsoft，IBM 等）。各种直方图算法的主要不同之处是如何选择桶以减少近似数据分布的错误。主要的直方图算法有等宽直方图、等高直方图、最大差异直方图和 V-优化直方图。

通过随机样本进行选择性估计也受到了一些关注[5,6]。Goodman 的估计法[5]是唯一无偏的随机样本估计方法，但是由于其产生的方差很高，因此在实际中并不能很好地应用[6]。从数据库中选择随机样本还有其他的一些方法[7,10,11]。文献[7,10]中方法的前提是已知数据库的全部数据，这一前提条件在 Deep Web 数据集成环境中显然是不成立的，因此不再适用。文献[11]提出了在 Deep Web 环境中，用随机漫步的方法对 Web 数据库进行采样，取得了较好的随机样本，但是得到的样本是数据库级样本，相对于我们方法中的属性级的样本更为复杂。

另外，还有一些相关工作[8,9]。文献[8]关注 Web 数据库查询接口上绑定约束条件的文本属性的选择性估计。我们所关注的是同一属性不同值的选择性估计，而非同一属性不同约束的选择性估计。文献[9]为保证查询转换的有效性，构建了查询能力模型，但也只是针对合取式查询转换。因此我们所关注的是互斥查询转换的有效性和准确性。

3 选择性估计方法

由于获取 Web 数据库的所有数据很困难，我们考虑先获取样本，然后通过样本估计属性及其值的选择性。那么，如何提交查询才能获得随机样本？如何利用样本准确地估计其选择性？

据我们观察，不同的属性对之间的相关程度不同，而且无限值属性值的词频满足似 Zipf 分布。因此我们提出了一种新的选择性估计方法，主要由以下两部分组成。

属性相关性计算。对于一个给定的域（如书、论文），首先计算每对属性的相关性（Attribute Correlation calculation），并为每个指定的属性 $Attr_u$ 找到其最不相关的属性 $Attr_i$。由于同一个域中，每个属性对的相关性是与具体的 Web 数据库无关的，因此一对属性的相关性可以应用于本领域中所有的 Web 数据库。

选择性估计。给定无限值属性 $Attr_u$ 及某一特定的 Web 数据库，在 Web 数据库查询接口的 $Attr_i$ 属性上提交一系列的探测查询，获得属性 $Attr_u$ 上的近似随机样本（即基于属性相关性的采样，Correlation-based sampling）。从样本中可以计算出属性 $Attr_u$ 上的词序。由于样本是随机样本，因此样本的词序也即 Web 数据库所有元组在 $Attr_u$ 上的词序。然后，在属性 $Attr_u$ 上挑选任意若干个词，探测实际的 Web 数据库，获取这些词的实际词频（即词频探测，Word frequency probing）。在已获得词序和若干实际词频的基础上，计算出 Zipf 方程（即 Zipf 方程的计算，Zipf equation calculation）。最后，对于属性 $Attr_u$ 上任意的词，通过 Zipf 方程及相应的词序估计其选择性（即选择性估计，Selectivity estimation）。

4 基于属性相关性的采样

4.1 属性级样本

众所周知，商业 DBMS 通常采用的是一维直方图，不能提供属性之间相关性的信息。而且为全部属性的所有组合构建多维直方图通常是不可行的，因为随着属性的增加，属性组合的数量是按指数级迅速增长的。相似地，所有属性上的值的全部组合的数量也是巨大的。因此，对于只能通过查询接口获取数据的 Web 数据库，完成所有的探测查询也是不可行的。而实际上，我们只需要属性级的样本，并不需要数据库级的样本，因为属性级的样本已经可以计算出下一步选择性估计所需要的重要因素——无限值属性的词序。因此取而代之的方法是获取随机的属性级样本。

获取某一属性随机样本的基本方法是通过在其他属性上提交查询来收集该属性上的数据。但是，由于属性间存在各种各样的相关性，因此很难保证得到的样本是随机的。例如，在音乐领域，属性 Title 与属性 Artist 存在相关性，不同的 Artist属性值，其相应的属性 Title 的值是不同的，因此在属性 Artist 上提交查询，在属性 Title 上并不能获得随机样本。而属性 Title 与属性 Format 存在弱的相关性，因为属性 Format 不同的值（例如 CD，Audio cassette，DVD Audio 等）将会在属性 Title 上返回相似的结果集。因此，为了尽可能获取属性 $Attr_u$ 上的随机样本，应发现与属性 $Attr_u$ 最不相关的属性 $Attr_i$。

4.2 属性相关性

通常，属性的相关性表示了不同属性的数据在数量和质量上的相互依赖性。本文中，我们用不同属性的词分布定义属性相关性概念。

定义 6（属性词分布，Attribute Word Distribution，AWD） 给定数据库 D，且其属性 A 上的所有值所包含的词为 w_1，w_2，…，w_m，则 A 上属性词分布为一个矢量 $\vec{r}$，其每个分量 w_i（$w_i \in (w_1, w_2, \cdots, w_m)$）是该词的词频。在每个属性值中任意词值出现一次的假设前提下，w_i 的词频即是由查询 $\sigma_A D(A = w_i)$ 返回的结果数量。

定义 7（属性相关性，Attribute Correlation） 属性相关性是任意属性对（$Attr_u$，$Attr_v$）之间的依赖性，可以在属性 $Attr_v$ 上提交探测查询，然后用属性 $Attr_u$ 上返回词频分布的差异性衡量。属性词分布的差异性越大，属性 $Attr_u$ 则更依赖于属性 $Attr_v$，属性 $Attr_u$ 和 $Attr_v$ 之间的属性相关性越强。

例 1 在学术论文领域，由于每位作者所关注的主题不同，因此通过在属性 Author 上提交查询获得的属性 Title 上的返回值的属性词分布差异性很大，即属性 Author 和属性 Title 是强相关的。但是，如果在属性 Year 上提交不同的年份值，因为每年会议的主题变化很小，所以返回的结果在属性 Title 上的属性词分布却很相似，即属性 Year 和属性 Title 是弱相关的。

注意：这里属性词分布的不同实际是指矢量的方向之间的差异。如果矢量 $\vec{x}$ 的每个分量 x_i 出现的概率等于矢量 $\vec{y}$ 的相应分量 y_i 出现的概率，则这两个属性词分布相同。例如，已知 3 个矢量分别为 $\vec{o} = \{o_1, o_2, o_3\} = \{100, 60, 20\}$，$\vec{p} = \{p_1, p_2, p_3\} = \{20, 60, 100\}$，$\vec{q} = \{q_1, q_2, q_3\} = \{300, 181, 60\}$，由于矢量 $\vec{o}$ 各分量 o_1, o_2, o_3 与矢量的相应分量 p_1, p_2, p_3 的概率相差较大，而与 $\vec{q}$ 的相应分量 q_1, q_2, q_3 的概率相差很小，因此 $\vec{o}$ 与 $\vec{p}$ 之间的差异远远大于 $\vec{o}$ 与 $\vec{q}$ 的差异。

我们可以用式(1)测量矢量之间的分布差异。如果在属性 $Attr_v$ 上提交不同的查询 $Q_1, Q_2, \cdots, Q_s$，则得到属性 $Attr_u$ 上不同的结果集 $S_1, S_2, \cdots, S_s$。假设 S 是结果集 $S_1, S_2, \cdots, S_s$ 的并集，而且 S 由词 $w_1, w_2, \cdots, w_k$ 组成，则属性 $Attr_u$ 上 S 与 S_j 之间的分布差异为

$$D(S||S_j) = \frac{1}{k}\sum_{l=1}^{k} |(prob(Attr_u = w_l | S) - prob(Attr_i = w_l | S_j)| \quad (1)$$

式中，$prob(Attr_u = w_l | S)$ 是 $Attr_u = w_l$ 在 S 中的概率，$prob(Attr_u = w_l | S_j)$ 是 $Attr_u = w_l$ 在 S_j 中的概率，S_j 未包含的词 w_l 则被忽略，词分布的差异性可以由 S_j 中包含的词 w_l 反映出来。

属性相关性是上述分布差异的平均值

$$Correlation(Attr_u, Attr_v) = \frac{1}{s}\sum_{j=1}^{s} D(S||S_j) \quad (2)$$

差异性越小，属性 $Attr_u$ 越不依赖 $Attr_v$，属性 $Attr_u$ 与

$Attr_v$ 越不相关。反之，则属性 $Attr_u$ 与 $Attr_v$ 越相关。

4.3 基于属性相关性的采样

基于属性相关性采样方法的算法细节如下：给定一个 Web 数据库，其查询接口包含若干个属性。假设我们需发现属性 $Attr_u$（例如属性 Title）的最不相关的属性 $Attr_i$。首先，从查询接口选择属性 $Attr_u$ 之外的任意属性 $Attr_v$（例如，Subject），在属性 $Attr_v$ 上向 Web 数据库提交探测查询。然后收集和抽取在属性 $Attr_u$ 上的返回结果并存在本地表中（行 3—6）。第二步，分析每次返回结果中每个词出现的概率，计算属性 $Attr_u$ 与属性 $Attr_v$ 之间的相关性（行 7—11）。如此迭代，直到除了属性 $Attr_u$ 之外的所有属性都已选择过为止（行 2—12）。最后，计算得到与属性 $Attr_u$ 相关性的最小值，相应的属性 $Attr_i$ 即为 $Attr_u$ 的最不相关的属性（行 13—14）。

算法 基于属性相关性的随机采样算法

输入：m //the number of attributes of a Web database interface
Attru //an attribute in a Web database interface
$Attr_v$//other attributes except of $Attr_u$, v=1,2,…,m and v<>u

输出：$Attr_i$//attribute that is the least correlative with $Attr_u$

```
1. Begin
2.   For (v=1 to m) and (v<>u) Do
3.     For j=1 to k Do
4.       Submit Qj on Attrv
5.       Abstract and store the result set Sj
6.     End For
7.     Merge Sj(j=1,…,k) to S
8.     For j=1 to k Do
9.       DKL (S || Sj)
10.    End For
11.    Correlation(Attru, Attrv)
12.  End For
13.   Correlation(Attru, Attri)=Min(Correlation(Attru, Attrv))
14.   Return Attri
15. End
```

得到最不相关的属性 $Attr_i$ 后，在该属性上向 Web 数据库提交一些探测查询，将在属性 $Attr_u$ 上收集到的返回结果作为属性 $Attr_u$ 的属性级随机样本，然后按属性值中词出现的频率从高到低排序。

由于样本是近似的随机样本，因此属性 $Attr_u$ 样本的词序也真实地反映了实际 Web 数据库中该属性所包含的词的实际词序。

5 基于 Zipf 分布的选择性估计

从某一属性的属性级随机样本推导实际数据库的属性词分布（AWD）的一个简单方法可用如下公式计算得到。

$$AWD_{reality}=AWD_{sample}\frac{Size_{reality}}{Size_{sample}} \tag{3}$$

式中，AWD_{sample}，$Size_{sample}$，$Size_{reality}$ 分别是样本中的属性词分布、样本的大小以及实际 Web 数据库的大小。前两项可以获得，但是实际 Web 数据库的大小是未知的，因此这个简单的方法无法在这里应用，我们则提出了新的解决方法，即基于 Zipf 分布的选择性估计方法。

5.1 属性词分布的特征

众所周知，文档集中的词通常满足 Zipf 分布，但是不同领域中的文本属性的值所包含的词的分布有规律吗？如果有规律，那么满足什么分布？是否也满足 Zipf 分布？对此我们做了大量的统计分析。结果表明，不同领域的文本属性中的词确实满足似 Zipf 分布。图 3 是对 DBLP，Bookpool 和 IMDb 等 Web 数据库的无限值属性（如 DBLP 的 Title，Conference，Journal，Bookpools 的 Title，IMDb 的 Title，Director）的词分布统计，这些无限值属性值所包含的词均遵循似 Zipf 分布。因此我们考虑从属性词分布的特征着手，探寻选择性估计方法。

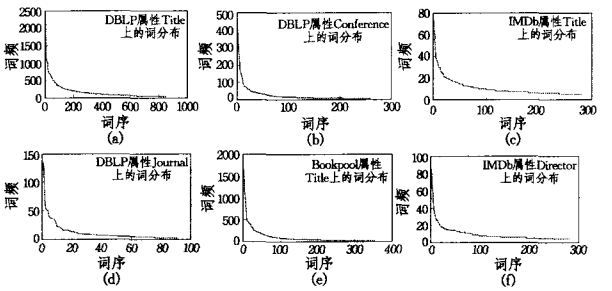

图 3 属性词分布的特征

5.2 基于 Zipf 分布的选择性估计方法

Zipf 分布可以表示为方程 $N=P(r+p)^{-E}$，其中 N 表示词出现的频率，r 表示词的排序，P，p 和 E 是正的参数。如果所有属性值中每个词的序都可以由属性级样本计算得到，则利用其中的几个词及其真实词频（通过提交这些词到真实的 Web 数据库上获取查询结果的数量），再对 Zipf 方程进行一系列公式转换与推导，则可以计算出 P，p 和 E 3 个参数，从而得到 Zipf 方程 $N=P(r+p)^{-E}$。

图 4 是基于 Zipf 分布选择性估计方法的示意图。首先向 Web 数据库提交词 i 和词 j，分别获得相应的词频 F_{wi}（对应方程中的 N_i）和 F_{wj}（对应方程中的 N_j）。我们已基于属性相关性获得了属性级的随机样本，所以通过样本已知词 i 和词 j 的序（对应方程中 r_i 和 r_j），可以按如下方法计算出参数 P，p 和 E。

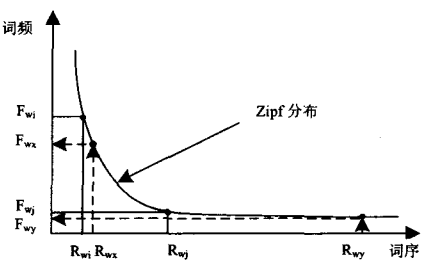

图 4 基于 Zipf 分布的选择性估计

方程转换：方程两边取对数，得到 $\ln(N)=\ln P-E\ln(r+p)$。因为参数 $p(0<p<1)$ 通常远远小于词序 r（一些应用甚至假设 $p=0$），所以参数 E 可以被近似看作直线 $\ln(N)=\ln P-E\ln(r+p)$ 的斜率，如图 5 所示。

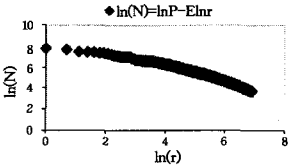

图 5 词分布的转换

参数 E:E 可以由公式 $E\approx\frac{\ln(N_i)-\ln(N_j)}{\ln(r_j)-\ln(r_i)}$ 计算。

参数 p:一旦计算出参数 E,则

$$\frac{N_i}{N_j}=\frac{P*(r_i+p)^{-E}}{P*(r_j+p)^{-E}}$$

所以参数 p 可以近似为

$$p\approx\frac{r_j-r_i*e^m}{e^m-1}\left(m=\frac{1}{E}*\ln\frac{N_i}{N_j}\right)$$

参数 P:最后,将已知的参数 p,E 代入 Zipf 方程 $N=P(r+p)^{-E}$,可以进一步推导出参数 P 近似为

$$P\approx N_j*(r_j+p)^E$$

通过上述公式转换和推导,可以计算出 Zipf 方程 $N=P(r+p)^{-E}$。例 2 是采用上述方法计算 Zipf 方程中各参数的实例。

例 2　考虑 DBLP 的一个子集(约 50,000 个记录),假设属性 Title 中的词遵循 Zipf 分布。首先,提交查询{〈Title,"learning"〉},返回结果数量为 1591;提交查询{〈Title,"modeling"〉},其相应的返回结果数量为 732。而且用基于属性相关性的采样方法,在属性"year"上提交查询,从此子集中选取了 8000 个属性值作为样本,并按词频从高到低的顺序对词进行了排序。在此例中,词"learning"和词"modeling"的序分别为 5 和 25。用词"learning"和词"modeling"的实际词频和它们的序,可以按上述方法估计出参数 P,p 和 E 分别为 3476,0.67 和 0.482。

接下来,可以用 Zipf 公式以及词序计算包含任意词查询的选择性,即包含该词实际的元组数。如图 4 所示,我们不必向 Web 数据库提交查询,而是可以用 Zipf 方程估计属性包含词 x 或词 y 的查询的选择性。例 3 是用 Zipf 公式以及词序推导任意词选择性的实例。

例 3　(例 2 续)如果向 Web 数据库的 Title 属性上提交包含任意词(如"automatic")的查询时将返回多少个查询结果?从子集的样本中,已知"automatic"的序是 60。由公式 $N=P(r+p)^{-E}(P=3476,p=0.67,E=0.482)$,则可以估计出"automatic"的词频是 484。而实际的 DBLP 子集中包含"automatic"的 Title 数量是 501,估计值非常接近于实际值。

值得注意的是,参数 P,p 和 E 并不唯一。当选择不同的词(词 i 和词 j),计算出的参数值并不相同,这会影响到下一步选择性估计的准确性。如何选择词 i 和词 j?我们探讨了选择性估计的准确性、词序和序差三者之间的关系。如图 6 所示,给定两个词(词 i 和词 j)之间的序差(如 10),当词序增加时,选择性估计的准确性将下降。为保持准确程度的稳定性,随着词序的增加,序差也应随之增加。实验表明,如果词 i 和词 j 的词序接近 10,当两者之间的序差大约 5 时,可以获得较高的准确性;而如果词 i 和词 j 的词序接近 100,两者之间的序差要达到大约为 50 时,才能获得较高的准确性。

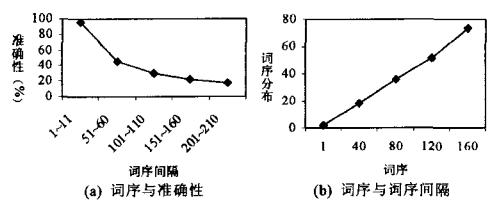

图 6　准确性、词序与词序间隔三者之间的关系

我们还做了准确性与样本大小之间关系的实验。当样本中属性值的数量在 5,000 时,选择性估计的准确性只有 60% 左右。当样本的数据量达到 15,000 个时,准确性较高,而且基本趋于稳定,如图 7 所示。

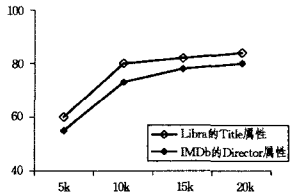

图 7　样本大小与选择性准确性的关系

6　实验

6.1　数据集

我们选择 DBLP,Libra,IMDb,Amazon 和 Bookpool 5 个网站作为我们的数据集,如表 1 所列。这些网站覆盖了 3 个域:学术论文、电影和书。

表 1　数据集

	Web database	URL	Domain
1	DBLP	http://www.informatik.uni-trier.de/~ley/db/	Papers
2	Libra	http://libra.msra.cn/	Papers
3	IMDb	http://www.imdb.com/	Movies
4	Amazon	http://www.amazon.com/	Books
5	Bookpool	http://www.bookpool.com/	Books

6.2　评价方法

我们定义了如下的准确性作为评价标准。

$$\text{准确性}=\frac{1}{n}\sum\left|\frac{N_r-E_s}{N_r}\right| \tag{4}$$

式中,N_r 是在某属性上提交查询时实际返回的结果数量,E_s 是用我们的方法得到的相应的估计值,n 是实验中测试的包含不同词的查询的数量。因此准确性是估计值偏离实际值占实际值的平均百分比。

6.3　实验结果

(1) 属性的词序

从属性级样本可以计算出每个无限值属性的词序。例如,Libra 的属性 Title 和属性 Conference 的前 20 个词分别如表 2 和表 3 所列。IMDb 的属性 Title 和属性 Director 的前 20 个词分别如表 4 和表 5 所列。

表 2　属性 Title 的前 20 个词 (Libra)

词	序	词	序	词	序	词	序
system	1	approach	6	algorithm	11	language	16
network	2	design	7	software	12	knowledge	17
model	3	information	8	multi	13	object	18
data	4	Time	9	agent	14	efficient	19
learning	5	analysis	10	distributed	15	mobile	20

表 3　属性 Conference 的前 20 个词 (Libra)

词	序	词	序	词	序	词	序
international	1	applications	6	distributed	11	logic	16
system	2	engineering	7	processing	12	security	17
computing	3	programming	8	artificial	13	networks	18
information	4	intelligence	9	management	14	knowledge	19
software	5	data	10	science	15	design	20

表 4　属性 Title 的前 20 个词 (IMDb)

词	序	词	序	词	序	词	序
black	1	day	6	life	11	dirty	16

die	2	world	7	city	12	karaoke	17
big	3	story	8	man	13	house	18
love	4	dead	9	time	14	Christmas	19
girls	5	hot	10	blue	15	night	20

表 5 属性 Director 的前 20 个词 (IMDb)

词	序	词	序	词	序	词	序
John	1	Thomas	6	Brian	11	Richard	16
David	2	Paul	7	Jim	12	Joe	17
Michael	3	James	8	Steve	13	Mike	18
Peter	4	Mark	9	Daniel	14	Christopher	19
Robert	5	George	10	Andrew	15	Tom	20

可以看出,相同词的词序在不同的属性上并不相同。比较表 2 和表 3,属性 Title 的前 20 个词和属性 Conference 的前 20 个词是有些重叠的(例如 Distributed,knowledge 等),因为属性 Title 和属性 Conference 两者在语义上是有相关性的。比较表 2 和表 4,即属性 Title(学术论文领域)的前 20 个词和属性 Director(电影领域)的前 20 个词;再比较表 4 和表 5,即属性 Title(电影领域)的前 20 个词和属性 Director(电影领域)的前 20 个词,可以看出,它们完全不同,这也证实了在无限值属性上没有通用的词序,每个属性都有自己与众不同的词序。

(2) 选择性估计

分别向 Libra 的属性 Title 和属性 Conference 以及 IMDb 的属性 Title 和属性 Director 实际提交 90 个词。同时用我们的方法估计这些词在相应属性上的选择性。实验结果表明,属性中排序靠前的词,其选择性估计的准确性相对较高,如表 6 所列。由于准确性是估计值偏离实际值占实际值的平均百分比,对于排序靠后的词,其本身实际出现的频率较低,因此估计的偏差占实际值的百分比相对较明显。从总体来看,估计的准确性还是比较高的,如表 6 所列,选择性估计的平均准确性在 Libra 的属性 Title 上为 93.2%,在 IMDb 的属性 Director 上为 93.7%,在 IMDb 的属性 Title 上为 91.9%,在 DBLP 的属性 Conference 上为 86.5%。

表 6 选择性估计的准确性

Title 属性(Libra) 平均准确性:93.2%									
1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
97.8%	95.5%	94.7%	93.2%	92.9%	94.3%	93.7%	92.3%	88.9%	88.3%
Director 属性(IMDb) 平均准确性:93.7%									
1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
98.5%	95.5%	94.6%	95.3%	95.1%	93.8%	93.1%	90.6%	91%	89.8%
Title 属性(IMDb) 平均准确性:91.9%									
1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
96.7%	94.3%	93.5%	94.3%	93.1%	91.8%	92.9%	91.1%	86.6%	84.9%
Conference 属性(DBLP) 平均准确性:86.5%									
1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
93.1%	91.6%	87.8%	87.2%	85.6%	83.3%	85.3%	84.8%	84.1%	82.2%

估计值和真实值之间有一定的偏离,究其原因主要有两个方面:一是尽管我们已经选择出最不相关的属性,但是属性之间还是有某种程度的相关性。因此,我们提出的基于属性相关性得到的属性级样本是近似随机样本,并非真正的随机样本。另外,无限值属性值所包含的词并不完全符合 Zipf 分布,而符合近似的 Zipf 分布,这也会给估计带来一定的偏差。

由于我们提出的方法适用于所有无限属性而且是不依赖于领域的,因此是一种普遍适用的方法,可以优化 Deep Web 数据集成中的互斥查询转换。

结束语 选择性估计是 Deep Web 数据集成中互斥查询转换面临的关键问题。由于无法获取 Web 数据库的全部数据,因此传统的选择性估计方法不再适用。通过大量的统计分析,我们发现不同属性对之间的相关程度是不同的,而且无限值属性值所包含的词遵循似 Zipf 分布。我们的选择性估计方法利用这两个重要的特征,首先通过基于属性相关性的采样方法获得属性级的随机样本,然后利用随机样本和少量的探测查询结果计算 Zipf 方程,最后采用基于 Zipf 方程的方法推导包含任意词查询的选择性。由于不需要获取大量的 Web 数据库中的数据,因此大大简化了选择性估计的过程。实验表明,我们提出的基于 Zipf 分布的选择性估计方法具有很高的准确性。

参 考 文 献

[1] The Deep Web: Surfacing Hidden Value[EB/OL]. http://www.completeplanet.com/Tutorials/Deep Web/

[2] Piatetsky S G, Connell C. Accurate estimation of the number of tuples satisfying a condition[C]// the 3th International Conference on Management of Data. Boston: ACM Press, 1984: 256-276

[3] Poosala V V, Ioannidis Y, Haas P, et al. Improved histograms for selectivity estimation of range predicates[C]// the 15th International Conference on Management of Data. Montreal: ACM Press, 1996: 294-305

[4] Ioannidis Y E, Poosala Y. Histogram-based Approximation of Set-valued Query-answers[C]// the 25th International Conference on Very Large Data Bases. Morgan Kaufmann, Edinburgh, 1999: 174-185

[5] Goodman L. On the estimation of the number of classes in a population[J]. Annals of Math. Stat., 1949, 20: 572-579

[6] Haas P, Naughton J, Seshadri P, et al. Sampling-based estimation of the number of distinct values of an attribute[C]// the 21th International Conference on Very Large Data Bases. Morgan Kaufmann, Zurich, 1995: 311-322

[7] Oliken F. Random Sampling from databases[D]. Berkeley: University of California, 1993

[8] Zhang Z, He B, Chang K C-C. On-the-fly Constraint Mapping Across Web Query Interfaces[C]// the 23th International Conference on Management of Data Workshop on Information Integration on the Web. Morgan Kaufmann, Toronto, 2004

[9] Shu L, Meng W, He H, et al. Querying Capability Modeling and Construction of Deep Web[C]// the 8th International Conference on Web Information Systems Engineering Wise. Springer, Nancy, 2007: 13-25

[10] Vitter J. Random Sampling with a Reservoir[J]. ACM Transactions on Mathematical Software, 1985, 11(1)

[11] Dasgupta A, Das G, Mannila H. A random walk approach to sampling hidden databases[C]// the 26th International Conference on Management of Data. Beijing: ACM Press, 2007: 629-640