网络表格间的快照关系发现

王 宁 任红伟

(北京交通大学计算机与信息技术学院 北京 100044)

摘 要 近年来,互联网上涌现出大量结构化的表格数据,网络表格的价值不仅在于数据本身,还在于数据之间的关系。只有探测出表格之间潜在的关系,方能更好地利用这些结构化数据。因此提出发现网络表格间的快照关系,并给出发现快照关系的框架以及检测与给定表之间满足某种匹配关系的快照表的算法,快照表可用于优化查询以及在大数据环境下实时地返回部分查询结果。提出了基于实体和属性重合度的评分方法,并引入实体新鲜度的概念,使得算法在快照关系的发现过程中更多地关注能提供新鲜实体的表;与此同时,基于Bayes模型的表格内容增强算法能更加准确地判断属性列上值的一致性,从而提高快照关系发现的准确率。大量实验表明,该评分模型能发现高质量的快照表,且在快照的查询精度和召回率上表现出色。

关键词 网络表格,关联关系,快照,数据集成,查询优化

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.002

Detecting Snapshots for Web Tables

WANG Ning REN Hong-wei

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract In recent years, a large number of structured tabular data have emerged on the Internet constantly. However, the value of Web tables depends not only on the data itself, but also on the relatedness between the data. Only when the potential relatedness between them is detected, can these structured data be fully utilized. We proposed a new type of relatedness between Web tables called snapshot relationship, and a framework for capturing snapshots that meet a certain matching condition with a given table. The snapshots are beneficial for query optimization, and also helpful for returning partial results rapidly when querying on big data. The relatedness between an original Web table and its snapshot can be computed based on entity consistency and schema consistency. In order to assign high weights on tables which provide more fresh entities, the concept of entity freshness was introduced into our scoring method. Meanwhile, the content consistency of Web tables can be enhanced by applying Bayesian analysis to our relatedness capturing framework. As a consequence, accuracy of finding snapshots is improved. Extensive experiments demonstrate that the algorithms can capture snapshots with high quality, and perform well in query precision and recall.

Keywords Web tables, Relatedness, Snapshot, Data integration, Query optimization

1 引言

随着信息技术的发展,互联网上的资源越来越丰富,除了非结构化数据外,还有大量的网络表存在,这些网络表覆盖面广且信息量大,因此受到人们的关注。Google公司启动了WebTables项目,以研究如何更好地抽取和利用网络上广泛存在的结构化数据[1];最近推出的Fusion Tables旨在帮助人们在云环境下进行数据集成和合作,用户可以上传表格状的数据并与其他用户分享,通过合作的方式利用众人的智慧来解决数据合并过程中可能引起的冲突[2]。为让机器容易地处理来自网络的表格数据,Microsoft公司利用知识库对网络表进行语义恢复,恢复其表头,并实现实体列的探测以及典型属性的提取[3-5]。

事实上,网络数据的价值不仅在于数据本身,还在于数据之间的关系。只有探测出数据间潜在的关系,方能更好地利用这些数据。Xin Luna Dong尝试发现网络数据间的复制关系,以便找到数据的真正来源,从而检测数据的真假[6]。Anish Das Sarma等人已经发现了表格之间的合并和连接关系,合并关系可以实现实体级的补充,而连接关系可以实现实体属性级的补充,这两种关联关系的发现不仅有助于表格内容的扩展,也给搜索带来了很大的帮助[7]。除了复制、合并以及连接关系外,网络表格之间还存在其它可以利用的关系。

图1和图2分别是来自Tennis和skysports官方网站的2014年ATP成绩列表,图1示出了2014年ATP网球成绩汇总,图2示出了2014年ATP网球成绩大于4000的运动员信息。显然,图2中的信息是图1中信息的一部分。此外,从

到稿日期:2014-07-29 返修日期:2014-11-05 本文受国家自然科学基金项目(61370060),江苏省自然科学基金项目(BK2011454)资助。

王 宁(1967-),女,博士,副教授,CCF会员,主要研究方向为Web数据集成、大数据管理、数据挖掘,E-mail:nwang@bjtu.edu.cn;任红伟(1988-),女,硕士生,主要研究方向为Web数据集成。

规模上来说，图 2 中的表比图 1 中的小很多。进一步观察可以发现，图 2 中的信息是图 1 中的表根据某个匹配关系 R 进行筛选的结果，R 为网球成绩大于 4000 的运动员的名次、姓名和成绩。网络表格之间的这种关系很像关系数据库中的快照关系，但不同于 Xin Luna Dong 发现的复制关系，复制关系主要用于判断数据的真假，追溯数据的起源，类似于整体和整体的关系，而这种快照关系更像整体和部分的关系。

Rank	Previous Rank	Name	Country	Ranking Points
1	1	Rafael Nadal	Spain	13,730
2	2	Novak Djokovic	Serbia	11,810
3	3	Stanislas Wawrinka	Switzerland	5,740
4	5	Roger Federer	Switzerland	5,225
5	7	Tomas Berdych	Czech Republic	4,720
6	4	David Ferrer	Spain	4,640
7	8	Juan Martin del Potro	Argentina	4,260
8	6	Andy Murray	Great Britain	3,975
9	10	John Isner	USA	2,715
10	12	Milos Raonic	Canada	2,710
11	9	Richard Gasquet	France	2,635
12	11	Jo-Wilfried Tsonga	France	2,615
13	14	Fabio Fognini	Italy	2,140

图 1　2014 年 ATP 成绩

Pos	Player	Pts
1	R Nadal	13,730
2	N Djokovic	11,680
3	S Warwinka	5,760
4	R Federer	5,355
5	T Berdych	4,720
6	D Ferrer	4,640
7	J M Del Potro	4,260
8	A Murray	4,040

图 2　2014 年 ATP 排名前 8 的成绩

网络上日益增多的结构化数据可以和本地数据一起用于查询和分析，这种系统打破了传统数据库系统的封闭世界假设(Closed-world Assumption)，利用互联网上大众积累的开放数据实现“open-world”查询，帮助用户完成仅凭本地数据不能完成的分析任务[8]。由于网络表格数据有众多来源，具有不确定性，通过判别表格之间的快照关系，可以发现满足匹配条件的多张源表，top-k 个结果的呈现也有助于用户选择满意的答案。多数据源的选择是近年的研究热点[9]，与本文的研究是平行的。此外，快照关系可以帮助查询优化和在大数据环境下实时地返回部分查询结果。大数据环境下，实时性和查询的精确性是需要权衡的，有时并不需要一次返回所有结果，快速返回部分结果也是有益的[10,11]，而利用快照关系就可以较快地将一部分查询结果返回给用户。假设给定某种匹配条件，网络表 $T1$ 中匹配该条件的元组存在于网络表 $T2$ 中，$T2$ 就被称为 $T1$ 匹配该条件的快照，发现网络表之间的快照关系具有重要的意义。

值得注意的是，关系数据库中某张表的快照是通过查询生成的，快照中的数据与查询是准确匹配的。本文的研究不是针对网络表格生成其快照，而是发现来自不同数据源的网络表之间的快照关系，快照关系的发现在线下完成，其结果会产生一些索引，帮助在线查询找到来自不同源的数据，或根据用户要求快速返回结果。由于网络数据的异构性和不确定性，网络表之间的快照关系的发现会面临新的挑战，需要考虑新的问题：

首先，网络表格并不规范，往往没有完整的模式，表中数据存在噪音，即便是同一实体，其表现形式也会存在差异。

其次，对于来自不同数据源的网络表格而言，很难找到能准确匹配的快照关系。我们只能根据匹配程度评分，返回评分较高的快照关系。

本文的主要工作及贡献如下：

1) 首次提出发现网络表格间的快照关系，并设计评分的理论模型及算法，该算法权衡表中实体和属性重合度因素，能够发现较高质量的快照表；

2) 提出实体新鲜度的概念，并给出表中实体新鲜度的计算方法，在同样的匹配条件下，本文快照关系发现算法会更多地关注能提供新鲜实体的表；

3) 设计基于 Bayes 模型的表格内容增强算法，能更加准确地判断属性列上值的一致性，从而提高快照关系发现的准确率；

4) 大量实验表明，提出的评分模型可以高质量地发现网络表格间的快照关系，且在快照的查询精度和召回率上表现出色。

2　网络表格及其快照关系

定义 1　对于一张网络表格 T，其中的每条记录均代表一个实体，如果 T 中某个属性列能够标识相应的实体，则称该属性列为网络表格 T 的实体列，记作 $EC(T)$。

如表格 T 记录各个国家的名称及其 GDP 情况，则 T 的实体列就是国家名称。网络表格的实体列类似关系表的主键，不同之处在于，主键的值能唯一标识一个实体，对网络表而言，实体列仅存放实体的标识，它的值无须唯一。

定义 2　对于表 T 上的操作序列 $R=\pi(Poi)\sigma_{F_i(x)}$，其中 σ 为选择条件，$F_i(x)$ 是一个逻辑表达式，$\sigma_{F_i(x)}$ 是选择使逻辑表达式为真的实体；π 为投影操作，Poi 为一组属性，$\pi(Poi)$ 为投影表中各实体在 Poi 中对应的值。T 经过操作序列 R 可以得到另外一张表，定义 Poi 为兴趣列，R 为匹配关系，简记为 $R=\sigma'_{F_i(x)}(Poi)$。

定义 3　对于网络表 T_1 和 T_2，以及匹配关系 $R=\sigma_{F_i(x)}(Poi)$，如果满足以下 3 个条件：

1) $EC(T_2)=EC(T_1)$；

2) $E_{T_2}\cap E_{T_1|\sigma}\neq\emptyset$，其中 E_{T_k} 表示 T_k 表中实体的集合，$E_{T_1|\sigma}=\sigma_{F_i(x)}(T_1)$，$\sigma_{F_i(x)}(T_1)=\{x\mid x\in E_{T_1}\wedge F_i(x)=\text{true}\}$；

3) $A(T_2)\cap Poi\neq\emptyset$，其中 $A(T2)$ 为 T_2 表属性列的集合。

则称 T_2 是 T_1 上满足匹配关系 R 的快照表，T_2 与 T_1 的关系为快照关系。

本文拟解决的问题为：给定网络表格 T、网络表格的集合 Γ、常量 k 以及匹配关系 R，从 Γ 中找出 k 张与 T 满足匹配关系 R 的快照表，并根据匹配相关度对返回表进行排序。

3　快照的检测

因为网络表格可能存在结构不统一、模式不完整等问题，所以预先采用文献[4]提到的方法对数据集中的表进行语义恢复，规范表格的结构，检测表头及实体列。近年来语义恢复的成果颇多[3-5]，本文不一一赘述。网络表格的快照关系发现分为 3 步：

第一步：预处理。此阶段的主要任务是将网络表格的集合用 Freebase 知识库按表的域进行分类，以降低快照关系发现的计算复杂度。

第二步：初筛选。该阶段的主要任务是根据匹配关系，从

表集中选择与给定表结构相似的表，而匹配关系由选择条件和兴趣列两部分组成。

第三步：内容增强。这一阶段采用 Bayes 模型探测表格在共同属性上内容的一致性，从而更加准确地发现表格间的快照关系。图 3 给出发现网络表格间快照关系的框架。

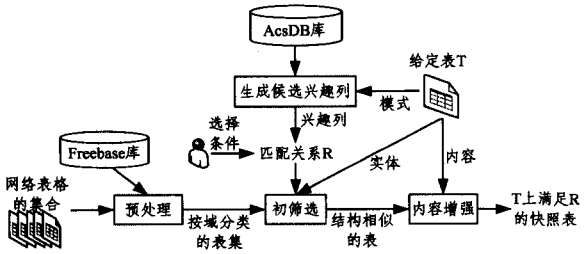

图 3　网络表格间的快照关系发现框架

目前，匹配关系 R 中的选择条件由用户指定，兴趣列则由系统推荐用户指定，兴趣列的推荐见 3.2.2 节。

3.1　预处理

来自不同数据源的网络表格涉及内容广泛，根据定义可知，具有快照关系的表格之间应有相同的实体。换句话说，快照关系只会存在于描述实体的领域一致的网络表格之间。为了提高快照关系的发现效率，减少后续计算复杂度，需要对网络表格的集合进行预处理，根据实体列的值判断表所在的领域，然后按领域对表进行分类。

Freebase 可以用来判断实体在现实世界中所属的类别和其拥有的特性[13]。目前的研究主要针对英文表格，而处理中文表格需要相应的中文知识库，例如 HowNet 等，对中文表格的处理是我们未来的研究工作。表格中的实体由实体列标识，其中实体组成表格的实体集 $E=\{e_1,e_2,\cdots\}$，对每一个实体 e_i，Freebase 返回其可能的域的集合 $D(e_i)$，$D(e_i)=\{d_i^1,d_i^2,\cdots\}$。表的域 d 满足两个条件：

1) $d\in D(E)$，其中 $D(E)=\bigcup D(e_i)$；

2) $\max\limits_{d\in D(E)} count(d)$，其中 $count(d)$ 为 d 在所有 $D(e_i)$ 中累计出现的次数。

3.2　基于实体和属性重合度的初筛选

初筛选的主要任务是从表集中筛选出与给定表 T_1 有较高的实体和属性重合度的网络表格。如果网络表格 T_2 是 T_1 上满足匹配关系 R 的快照表，两者首先应该在结构上一致；同时，T_2 中应尽可能地含有用户指定的兴趣列，即具有一定的实体和属性重合度；此外，它们的内容也应该具有一致性，即实体在共有属性上值的一致。本节介绍结构一致的评分方法，表格内容一致的评分方法将在 3.3 节中介绍。

3.2.1　实体因素

若 T_2 是给定表 T_1 上满足匹配关系 R 的快照，那么 T_2 中的实体应该尽可能多地包含 T_1 上满足 R 中 σ 条件的实体，两表在匹配关系 R 下的实体覆盖率见式(1)。

$$E_{base|\sigma}(T_1,T_2)=\frac{|E_{T_2}\cap E_{T_1|\sigma}|}{|E_{T_1|\sigma}|} \tag{1}$$

其中，$E_{T_1|\sigma}$ 表示 T_1 表中满足条件 σ 的实体集合，E_{T_2} 为 T_2 的实体集合。采用 Jaccard 相似度计算公式计算实体间的相似度。

然而，不能简单地只将实体的覆盖率作为实体部分的评价标准，对于给定表 T_1 上满足匹配关系 R 的快照 T_2 而言，在理想情况下，T_2 中的实体应该尽可能多地包含 T_1 中满足选择条件的实体，同时尽量少地包含不满足条件的实体。为此，我们引入一个调节因素 $E_{adjustment}(T_1,T_2)$，计算式见式(2)。

$$E_{adjustment}(T_1,T_2)=\frac{|E_{T_2}-E_{T_1|\sigma}|}{|E_{T_1|\sigma}|} \tag{2}$$

此外，实体新鲜度(Entity Freshness)也是一个不容忽视的因素。如果两张网络表格 A 和 B 都是 T_1 的快照，则它们对于 T_1 的实体覆盖率相等，但 A 中描述的实体在其它表中普遍存在，而 B 中描述的实体在表集中很少出现，那么就认为表 B 的实体新鲜度高于 A，并赋予这样的表 B 一个较高的实体因素的得分。为此引入另一个参数 $Fr(T_2)$ 表示表 T_2 中实体的新鲜度，计算方法见式(3)。

$$Fr(T_2)=\sum_{e_i\in E_{T_1|\sigma}}\frac{1}{count(e_i)}*exist(e_i\in E_{T_2}) \tag{3}$$

其中，$count(e_i)$ 为初筛选阶段处理的表集 Γ' 中 e_i 在每张表中出现次数的和，$exist(e_i\in E_{T_2})$ 根据 e_i 在表 T_2 上存在与否取值 1 或 0。将式(3)归一化为式(4)。

$$Fr^{\mu}(T_2)=\frac{Fr(T_2)}{\max\limits_{T_k\in\Gamma'}Fr(T_k)} \tag{4}$$

最后，结合实体覆盖率、调节因素和表中实体的新鲜度，将 T_2 表在实体因素方面的评分记为 S_{Entity}，如式(5)所示。

$$S_{Entity}(T_1,T_2)=Fr^{\mu}(T_2)*\exp\left(-\frac{(E_{base}-E_{adjustment}+\mu)^2}{2}\right) \tag{5}$$

其中，μ 为参数，用来控制函数在 E_{base} 和 $E_{adjustment}$ 两个指标下取得最大值的情况。若 $E_{adjustment}\neq 0$，则当 $\mu=\varepsilon$(ε 为参数)时，$S_{Entity}(T_1,T_2)$ 取最大值；若 $E_{adjustment}=0$，则当 $\mu=-1$ 时，$S_{Entity}(T_1,T_2)$ 取最大值。

3.2.2　兴趣列及属性重合度

快照表的发现基于匹配关系 R，而 R 中需要指定经常查询的列。本文用 Google 公司发布的 AcsDB 模式数据库为用户推荐表中经常查询的列，也称兴趣列 Poi。AcsDB 通过对网络表格的模式分析得到各模式出现的频率，利用它可以实现诸如自动模式匹配[1]、表格间连接关系的发现[7]等功能，本文尝试利用 AcsDB 发现表格的兴趣列。

对于给定的表格模式 S，AcsDB 可以返回库中该模式出现的频率，记作 $freq(S)$。在给定表 T_1 的实体列 a_e 的情况下，属性 a 出现的概率 $P(a|a_e)$ 按式(6)求得。

$$P(a|a_e)=\frac{freq(\{a,a_e\})}{freq(\{a_e\})} \tag{6}$$

若用户从推荐的兴趣列中指定了 n 个属性(包含实体列)作为最后的兴趣列 $Poi=\{a_1,\cdots,a_n\}$，其中 $a_1=a_e$，则 T_2 表在属性重合度方面的得分 $S_{attr}(T_1,T_2)$ 按式(7)计算。

$$S_{attr}(T_1,T_2)=\sum_{a_k\in Poi}\frac{P(a_k|a_e)}{\sum\limits_{a_i\in Poi}P(a_i|a_e)}*exist(a_k\in A(T_2)) \tag{7}$$

3.2.3　初筛选的评分

初筛选阶段主要考虑实体和属性重合度两方面的因素，这一部分的最终得分见式(8)。实体和属性这两个因素相当于表格的结构，经过这一步的判断，可以筛选出与给定的表 T_1 在结构上有较高相似度的表。

$$S_{frame}(T_1,T_2)=S_{Entity}(T_1,T_2)\times S_{attr}(T_1,T_2) \tag{8}$$

3.3 基于表格内容的增强算法

通过初筛选，已经选出与给定表 T_1 具有相似实体和较高属性重合度的表。但是，即使两张表有同样的实体和属性，同一实体对应属性上的值也可能不同，如果值不同的属性又恰是指定的兴趣列，则两表就不是快照关系。因此，快照的评分需要考虑实体的具体内容。

本文采用 Bayes 分析方法来计算 T_1 和 T_2 是快照关系的概率。若 T_2 和 T_1 只是结构相同，描述的实体相同，属性下的值都是独立给出的，并没有相互复制，那么就称 T_2 独立于 T_1。如果 T_2 和 T_1 是相互独立的，那么同一实体在某个属性上的值可能相同也可能不同。同样，如果 T_2 是 T_1 的快照，来自不同数据源的表格在同一实体上的值相同的概率应大于不同的概率。建立 Bayes 模型需要这样几个参数：$n(n>1)$，指一个实体在某个属性上可能给出值的个数；$r(0<r\leqslant 1)$，当 T_2 是 T_1 的快照时，T_2 表给出的值与 T_1 相同的概率。

我们对表格中的两个集合 $\overline{V}_s$ 和 $\overline{V}_d$ 感兴趣：$\overline{V}_s$ 表示 T_2 表在某个属性上给出与 T_1 相同值的实体集合，k_s 为该集合的大小；$\overline{V}_d$ 表示在该属性上给出不同值的实体集合，k_d 为相应集合的大小。

首先考虑 T_1 和 T_2 是独立的情况，记为 $T_2 \perp T_1$。由于一个实体在某个属性上的值有 n 种可能解，两张表各自给出某一个值的概率为 $1/n$。因此，对两张表来说，同一个实体在某个属性 a 上给出相同值和不同值的概率分别如式(9)、式(10)所示。

$$P_a(v\in\overline{V}_s \mid T_2\perp T_1)=\frac{1}{n} \tag{9}$$

$$P_a(v\in\overline{V}_d \mid T_2\perp T_1)=\frac{n-1}{n} \tag{10}$$

将式(9)和式(10)扩展到该属性的所有实体上，则在属性 a 上的条件概率见式(11)。

$$P_a(\overline{V}_s,\overline{V}_d \mid T_2\perp T_1)=(\frac{1}{n})^{k_s}\times(\frac{n-1}{n})^{k_d} \tag{11}$$

为方便表示，将 $(\overline{V}_s,\overline{V}_d)$ 简记为 Ψ，则式(11)简化为式(12)。

$$P_a(\Psi \mid T_2\perp T_1)=(\frac{1}{n})^{k_s}\times(\frac{n-1}{n})^{k_d} \tag{12}$$

用同样的方法考虑 T_1 和 T_2 是快照关系的情况。若 T_2 是 T_1 的快照，记为 $T_2\rightarrow T_1$。T_2 中属性的值应该与 T_1 中对应属性的值相同。但是对网络表格而言，也存在另一种情况，T_2 是 T_1 的快照表，但是 T_2 中的某些值被更改了，导致两张表的值不一致。为此，引入概率 r 来表示快照中的值为原值的概率。同理可得式(13)－式(15)。

$$P_a(v\in\overline{V}_s \mid T_2\rightarrow T_1)=r+(1-r)\times\frac{1}{n}=r+\frac{1-r}{n} \tag{13}$$

$$P_a(v\in\overline{V}_d \mid T_2\rightarrow T_1)=(1-r)\times\frac{n-1}{n} \tag{14}$$

$$P_a(\Psi \mid T_2\rightarrow T_1)=(r+\frac{1-r}{n})^{k_s}\times((1-r)\times\frac{n-1}{n})^{k_d} \tag{15}$$

最终，根据 Bayes 模型可得 $T_2\rightarrow T_1$ 的条件概率如式(16)所示。

$$P_a(T_2\rightarrow T_1 \mid \Psi)=\frac{P(\Psi \mid T_2\rightarrow T_1)P(T_2\rightarrow T_1)}{P(\Psi \mid T_2\rightarrow T_1)P(T_2\rightarrow T_1)+P(\Psi \mid T_2\perp T_1)P(T_2\perp T_1)}$$
$$=(1+(\frac{1}{n\times r+1+r})^{k_s}(\frac{1}{1+r})^{k_d}(\frac{1+\theta}{\theta}))^{-1} \tag{16}$$

其中，$\theta=P(T_2\rightarrow T_1)$，而参数 n、r 满足 $0<\frac{1}{n\times r+1-r}<1<\frac{1}{1-r}$。对表集中的一张表而言，式(16)具有这种特性：

1) $k_s+k_d=|E_{T_2}\cap E_{T_1|_a}|$，即 k_s 与 k_d 的和不变；

2) $P_a(T_2\rightarrow T_1 \mid \Psi)$ 为关于参数 k_s 的递增函数。

将两张表在内容上的得分记为 $S_{content}(T_1,T_2)$，其计算方法有 3 种，见式(17)－式(19)。

$$S^{sum}_{content}(T_1,T_2)=\sum_{a\in A(T_1)\cap A(T_2)}P_a(T_2\rightarrow T_1 \mid \Psi) \tag{17}$$

$$S^{avg}_{content}(T_1,T_2)=\frac{\sum_{a\in A(T_1)\cap A(T_2)}P_a(T_2\rightarrow T_1 \mid \Psi)}{|A(T_1)\cap A(T_2)|} \tag{18}$$

$$S^{min}_{content}(T_1,T_2)=\min_{a\in A(T_1)\cap A(T_2)}P_a(T_2\rightarrow T_1 \mid \Psi) \tag{19}$$

这 3 种方法都对两张表的共有属性逐一进行判断，虽然计算精度高，但由于对某些属性进行了不必要的判断，导致耗时严重。为解决该问题，尝试利用 AcsDB 来获得最有判断意义的属性，称为代表属性。一般情况下，代表属性上的值最有可能出现不一致，而直接对这样的属性进行判断，会提高快照关系发现的效率。

给定某张表 T 的属性集合 $A(T)$，AcsDB 库可以返回该集合共同出现的频率 $freq(A(T))$。表格 T 中的代表属性可以通过式(20)得到。

$$rep(T)=\{a \mid \min_{a\in A(T)} freq(\{a_e,a\})\} \tag{20}$$

其中，a_e 为模式中实体列的列标签，利用代表属性的实体内容评分见式(21)。

$$S^{rep}_{content}(T_1,T_2)=P_{rep}(T_2\rightarrow T_1 \mid \Psi) \tag{21}$$

4.2 节中的实验结果证明，利用代表属性进行内容一致性判断不仅降低了计算复杂性，而且并不影响快照的提取效果。

3.4 结合实体内容的评分

T_2 是 T_1 的快照，需要保证 T_2 在描述的实体和其属性上保持一致，且内容上要尽可能一致。快照关系的最终评分式(22)结合了初筛选阶段的评分以及基于内容的评分。

$$S_{snapshot}(T_1,T_2)=S_{frame}(T_1,T_2)\times S_{content}(T_1,T_2) \tag{22}$$

4 快照关系的评估

4.1 实验设置

实验所使用的硬件环境为 Intel(R) CPU B960，4GB 内存，软件环境为 Windows7 操作系统，采用 Java 作为编程语言，开发环境为 eclipse。实验数据集 Γ[1)-4)] 来自互联网上抓取的 700 张表格，大部分为体育赛事的成绩。从中选取 10 张表并分别指定匹配关系，用本文提出的算法为每张表 T 发现

1) http://www.atpworldtour.com/Rankings/Rankings-Home.aspx

2) http://espn.go.com/tennis

3) http://www1.skysports.com/tennis

4) http://www.tennis.com/rankings/ATP

其满足指定匹配关系 R 的 top-5 张快照表。实验分两部分进行：其一，通过用户评分评估快照提取的质量；其二，通过查询的召回率和精度评估提取出快照的有效性。

4.2 快照提取算法的评估

初筛选阶段需要考虑 3 个因素，分别称为 base、freshness 和 adjustment，其中 base 代表只考虑实体的覆盖率，freshness 代表实体的新鲜度，adjustment 代表调节因素，all combined 表示考虑所有的 3 个因素。在基于 Bayes 模型判断内容一致性的阶段，$S_{content}(T_1,T_2)$ 的得分有 4 种计算方法，分别为取各列条件概率的 min、*sum*、*avg*，以及只计算代表属性的条件概率作为最终得分。为验证初筛选阶段各个因素的有效性以及判断内容一致性阶段各计算方法的效果，参照 Google 公司在文献[7]中提出的评价方法，以用户评分为准对使用上述各方法提取的 top-5 张快照表进行评估。为了尽可能减小用户评分带来的结果偏差，选取爱好运动的大学生作为评分用户，且事先使其了解实验所使用的表格内容。

对于选定的 10 张网络表格 $\{T_1,T_2,\cdots,T_{10}\}$ 及其相应的匹配关系 $\{R_1,R_2,\cdots,R_{10}\}$，考虑上述各种因素和计算方法分别为每张表 $T_i(i=1,\cdots,10)$ 提取 top-5 张满足匹配关系 R_i 的快照表，共计 10 组 50 张快照表。用户根据匹配程度为快照表评分，分数从 0 到 5，0 表示所评价的表不是 T_i 上关于匹配关系 R_i 的快照，5 表示所评价的表最有可能是 T_i 上关于匹配关系 R_i 的快照，分数越高表示匹配程度越高。实验结果来自 9 个用户对每张快照表的反馈，整理后分别得到 top-1、top-3 和 top-5 张快照表的平均得分，即所有用户对 10 组快照表中 top-k(k=1、3、5)张表评分的平均值。

用户的评分结果如表 1 和表 2 所列，表 1 列出了考虑不同实体因素的快照提取质量评估(使用 $S^{\min}_{content}(T_1,T_2)$ 对表格内容评分)，表 2 列出了考虑内容的评分方法在不同算法下的快照提取效果评价(使用 all combined 方法判断实体因素)。表 3 列出了采用不同方法计算 $S_{content}(T_1,T_2)$ 时的快照发现时间。

表 1 考虑不同实体因素的快照提取质量评估

$S_{frame}(T_1,T_2)$	Top-1	Top-3	Top-5
base	2.5	2.9	2.7
base+freshness	3.5	3.2	3.0
base+adjustment	3.1	2.9	2.7
all combined	3.7	3.6	3.1

表 2 考虑内容的评分方法在不同算法下的快照提取效果评价

$S_{content}(T_1,T_2)$	Top-1	Top-3	Top-5
min	3.7	3.8	3.1
sum	3.6	3.5	2.9
avg	3.3	3.2	3.2
rep	3.6	3.5	2.9

表 3 发现快照表的时间

$S_{content}(T_1,T_2)$	min	sum	avg	rep
Time(ms)	15562.5	15961	14852	9359.5

分析实验结果，可以得到如下结论：

1)在发现给定表的快照时，若只考虑两个表实体的覆盖率，则发现的快照表质量较差。覆盖率高只表示表中相同实体的数目多，却忽略了不同实体的数目，可能导致一个规模比给定表大很多的表成为其快照，因而失去了提取快照表的真正意义。在引入调节因素 adjustment 后(base+adjustment)，其效果较 base 方法好，因为 adjustment 有效地抑制了快照表中不同于给定表中实体的数目。加入实体新鲜度 freshness 后，得到了更好的评价结果，原因是 freshness 更关注快照表集中给出的实体是否新鲜，在实体覆盖率相同的情况下，实体越新鲜说明在其它快照表中出现得越少，相应的价值就越高。在所有 3 个因素都考虑的情况下(all combined)，快照表的得分最高，质量也最好。

2)在考虑内容的评分方法中，设计了 4 种计算方法，分别是 min、sum、avg 和 rep，从表 2 可以看出，使用 min 和 rep 的效果较好。但从表 3 发现快照表的时间来看，min 较费时，因为需要计算每一列的概率；rep 选择出最有可能给出不同值的代表属性计算概率值，虽然得分略低于 min 方法，但花费的时间较少。

3)分析用户评分的结果。单独使用 base 方法时，各项得分较低，随着考虑因素的增加，快照表的得分呈现上升趋势，当考虑所有 3 个因素时，top-1 到 top-5 得分都最高。总体而言，推荐使用 all combined+rep 方法，可以在较低时间开销的基础上获得高质量的快照表。

4.3 快照关系的有效性评估

快照表可以用来优化表格上的查询，如果存在表 T 在匹配关系 R 下的较小规模的快照表，那么当用户的查询列和条件与匹配关系 R 相当时，就可以将查询转移到快照表上，从而降低查询时间。另一方面，在大数据环境下，为实时地返回查询结果，有时并不需要一次返回所有结果，返回部分结果也是有意义的。

由于网络数据存在异构性和不确定性，网络表格之间的快照关系不同于关系数据库中的快照关系。关系数据库中表的快照是指表的一个静态视图，快照中的数据与查询是准确匹配的，对于来自不同数据源的网络表而言，很难找到能准确匹配的快照关系。只能根据匹配程度评分，返回评分较高的快照关系。本节将从精度和召回率两方面对快照关系的有效性进行评估。

在信息检索系统中，精度和召回率的计算方法如式(23)—式(25)所示。在评估实验中利用转移到快照表的 SQL 查询的结果来检测其相对于原查询结果的精度和召回率，将快照表中满足 where 条件的元组集合作为 *Ret*，而将原表中满足 where 条件且包含目标列的元组集合作为 *Rel*。因此，精度计算方法如式(24)所示。

$$Precision=\frac{|Rel\cap Ret|}{|Ret|},Recall=\frac{|Rel\cap Ret|}{|Rel|} \tag{23}$$

$$Precision=\frac{|S^{column}_{input}\cap S^{column}_{snapshot}|}{|S^{column}_{input}|} \tag{24}$$

其中，S^{column}_{input} 为查询语句中目标列的集合，$S^{column}_{snapshot}$ 为查询时转到的快照表中列的集合。*Recall* 的计算方法如式(25)所示。

$$\begin{aligned}Recall&=\frac{|Rel\cap Ret|}{|Rel|}=\frac{|Rel\cap Ret|}{|Ret|}\times\frac{|Ret|}{|Rel|}\\&=Precision\times\frac{|Ret|}{|Rel|}\end{aligned} \tag{25}$$

可知，精度与输入的查询列有关，而从快照表返回的元组数、原表中满足查询条件的元组数和精度共同制约着召回率。

在 4.2 节的实验已发现的快照关系的基础上，取其中一

张表 *rankTable*(*rank*, *name*, *country*, *points*, *week change*, *tourn player*)及匹配关系 $R=\sigma_{points>2000}$(*name*, *country*, *points*),用 all combined+rep 方法发现 *rankTable* 表上满足匹配关系 R 的 top-5 张快照表,并利用 SQL 查询的结果评估快照关系的有效性。

表 4 列出了实验所用的查询语句,其中 β 分别取 2000、2500 和 3000,β 的取值确保在 *rankTable* 上的查询能够获得足够数量的返回结果。显然,每个查询在原始表上的召回率和精度都为 1。将每个查询分别转移到 rankTable 的满足匹配关系 $R=\sigma_{points>2000}$(*name*, *country*, *points*)的 top-3 张快照表的每张表上,获得查询结果,并计算相对于原表上查询结果的精度和召回率。

表 4 查询示例

No.	Query
1	select name from rankTable where points > β
2	select points from rankTable where points > β
3	select name, points from rankTable where points > β
4	select country from rankTable where points > β
5	select name, country from rankTable where points > β
6	select country, points from rankTable where points > β
7	select name, country, points from rankTable where points > β

图 4 和图 5 分别示出了当查询条件和目标列与关系 R 匹配时,7 种查询示例在 top-1 和 top-3 张快照表上的查询召回率和精度的变化情况。对实验结果的分析如下。

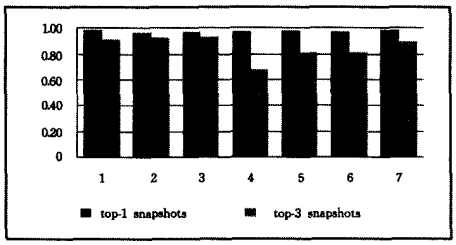

图 4 召回率

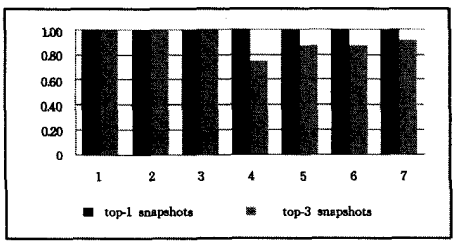

图 5 精度

1)就召回率而言,top-1 的快照表在所有的查询示例下均大于 0.96,top-3 的快照表在多数情况下(4/7 的查询示例)大于 0.90。召回率不够理想的查询示例(编号为 4、5、6)均含目标列"country",由于 rankTable 的 top-3 张快照表中存在不含该列的表,因此其召回率受到影响。其中,影响最大的是编号为 4 的查询,因其目标列仅"country"一项,相较而言,编号为 5 和 6 的查询的目标列虽包含"country"但非唯一项,对召回率的影响小一些。

2)在精度方面,根据度量公式,精度只与属性重合度相关,实验得到的 top-1 的快照表包含了查询的所有目标列,精度达到最高值 1。由于 top-3 的快照表中存在不包含"country"列的表,而编号为 4—7 的查询示例均包含"country"目标列,因此精度受到影响,其中,编号为 4 的查询将"country"作为其唯一的目标列,因此受到的影响最大。

3)综合来看,对于同样的查询示例,top-1 的快照表无论在精度还是召回率上都要高于 top-3。

一般情况下,当查询条件及目标列与匹配关系 R 匹配时,系统会默认推荐到 top-1 的快照表中查找,而从实验结果可以看出,top-1 的快照表无论在召回率还是精度方面都表现出了较高的性能。

5 相关工作

现有的研究已经发现了两种网络表格间的关联关系,即合并关系和连接关系。合并关系可以实现实体级的补充,而连接关系可以实现属性级的补充,这两种关联关系的发现为搜索带来了很大的帮助[7]。本文首次提出发现网络表格间的快照关系,用于优化查询以及在大数据环境下尽快地返回部分结果。

目前,数据管理系统需要融合来自不同源的数据,但不同源提供的数据往往会发生冲突,导致数据真假难辨。为了给用户提供更加可靠、可信的数据,Xin Luna Dong 等人尝试检测数据间存在的复制关系,旨在发现数据的本源,最终在冲突的数据中找出真值[15]。

本文提出发现快照关系,它与复制关系的区别在于:首先,关系发现的目的不同。发现快照关系的目的在于查询优化以及在大数据环境下实时地返回部分结果,而复制关系主要为了探测数据源间的依赖性从而发现数据的真假。其次,处理数据的单位不同。快照关系发现中以表为最小处理单位,复制关系中处理的是一条记录。

结束语 本文提出发现网络表格间的快照关系,并给出发现给定表上满足某种匹配关系的快照表的算法。实验结果表明,该快照关系发现算法能够高质量地发现满足匹配关系的快照表。为了让表格间的快照关系更好地服务于大数据环境下的检索,下一步的工作将尝试发现网络表格间的快照匹配关系。

参 考 文 献

[1] Cafarella M J, Halevy A, Wang D Z, et al. WebTables: Exploring the Power of Tables on the Web [J]. Proceedings of the VLDB Endowment, 2008, 1(1): 538-549

[2] Gonzalez H, Halevy A, Jensen C S, et al. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud [C]// Proc of the 1st ACM symposium on Cloud computing. New York: ACM, 2010: 175-180

[3] Wang J, Wang H, Wang Z, et al. Understanding Tables on the Web [M]. New York: Springer, 2012

[4] Venetis P, Halevy A, Madhavan J, et al. Recovering Semantics of Tables on the Web [J]. Proceedings of the VLDB Endowment, 2011, 4(9): 528-538

[5] Yakout M, Ganjam K, Chakrabarti K, et al. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables [C]// Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 97-108

[6] Dong X L, Berti-Equille L, Srivastava D. Truth Discovery and

Copying Detection in a Dynamic World [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573

[7] Sarma A D, Fang L, Gupta N, et al. Finding Related Tables [C]// Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012: 817-828

[8] Eberius J, Thiele M, Braunschweig, et al. DrillBeyond: Enabling Business Analysts to Explore the Web of Open Data [J]. Proceedings of the VLDB Endowment, 2012, 5(12): 1978-1981

[9] Theodoros R, Xin L D, Divesh S. Characterizing and Selecting Fresh Data Sources[C]// Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 919-930

[10] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169
Meng Xiao-feng, Ci Xiang. Big Data Management: Concepts, Technology and Challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169

[11] Babcock B, Chaudhuri S, Das G. Dynamic Sample Selection for Approximate Query Processing[C]// Proc of the 2003 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2003: 539-550

[12] Cafarella M J, Halevy A, Wang D Z, et al. Uncovering the Relational Web[C]// Proc of the 11th Int Workshop on the Web and Databases(WebDB 2008). Vancouver, 2008

[13] Bollacker K, Evans C, Paritosh P, et al. Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge[C]// Proc of the 2008 ACM SIGMOD Intl Conf on Management of data. New York: ACM, 2008: 1247-1250

[14] Lee T, Wang Z, Wang H, et al. Attribute Extraction and Scoring: A Probabilistic Approach[C]// Proc of the 2013 IEEE Int Conf on Data Engineering (ICDE). Washington, DC: IEEE, 2013: 194-205

[15] Dong X L, Berti-Equille L, Srivastava D. Integrating Conflicting Data: The Role of Source Dependence [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 550-561

(上接第 4 页)

[18] Elson J, Girod L, Estrin D. Fine-grained network time synchronization using reference broadcasts[J]. ACM SIGOPS Operating Systems Review, 2002, 36(SI): 147-163

[19] Liu L, Xiao Y, Zhang J, et al. A bio-inspired time synchronization algorithm for wireless sensor networks[C]// 2010 2nd International Conference on Computer Engineering and Technology (ICCET). 2010, 114: 306-311

[20] Ganeriwal S, Kumar R, Srivastava M B, et al. Timing-sync protocol for sensor networks[C]// Proceedings of the 1st International Conference on Embedded Networked Sensor Systems. ACM, 2003

[21] Rahamatkar S, Agarwal A. An Approach towards Lightweight, Reference Based, Tree Structured Time Synchronization in WSN [D]. Advances in Computer Science and Information Technology, Springer, 2011: 189-98

[22] Kim B-K, Hong S-H, Hur K, et al. Energy-efficient and rapid time synchronization for wireless sensor networks [J]. IEEE Transactions on Consumer Electronics, 2010, 56(4): 2258-2266

[23] Sichitiu M L, Veerarittiphan C, et al. Accurate time synchronization for wireless sensor networks[C]// Wireless Communications and Networking, 2003(WCNC 2003). IEEE, 2003: 153-163

[24] Ping S. Delay measurement time synchronization for wireless sensor networks[J]. Intel Research Berkeley Lab, IRB-TR-03-013, 2003

[25] Maróti M, Kusy B, Simon G, et al. The flooding time synchronization protocol[C]// Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems. ACM, 2004

[26] Akhlaq M, Sheltami T R, et al. The recursive time synchronization protocol for wireless sensor networks[C]// 2012 IEEE Sensors Applications Symposium (SAS). 2012

[27] Shannon J, Melvin H, Ruzzelli A G, et al. Dynamic flooding time synchronisation protocol for WSNs[C]// 2012 IEEE Global Communications Conference (GLOBECOM). 2012

[28] Wang F, Zeng P, Yu H, et al. The design and realization of High-Precision Time Synchronization Protocol in Wireless Sensor Networks based on FTSP[C]// 2010 8th World Congress on Intelligent Control and Automation (WCICA). 2010

[29] Dutta P, Musaloiu-E R, Stoica I, et al. Wireless ACK collisions not considered harmful[C]// In HotNets-VII: Proceedings of the 7th Workshop on Hot Topics in Networks. 2008: 19-24

[30] Tseng Y-C, Ni S-Y, Chen Y-S, et al. The broadcast storm problem in a mobile ad hoc network[J]. Wireless networks, 2002, 8 (2/3): 153-167

[31] Sirkeci-Mergen B, Scaglione A, Mergen G. Asymptotic analysis of multistage cooperative broadcast in wireless networks[J]. IEEE Transactions on Information Theory, 2006, 52(6): 2531-2550

[32] Lenzen C, Sommer P, Wattenhofer R, et al. Optimal clock synchronization in networks [C] // Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. ACM, 2009

[33] Schmid T, Dutta P, Srivastava M B, et al. High-resolution, low-power time synchronization an oxymoron no more[C] // Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks. ACM, 2010

[34] Wang Y, He Y, Mao X, et al. Exploiting constructive interference for scalable flooding in wireless networks[C]// 2012 Proceedings IEEE INFOCOM. 2012

[35] Doddavenkatappa M, Chan M C, Leong B, et al. Splash: Fast data dissemination with constructive interference in wireless sensor networks[C]// NSDI: Proc of the USENIX Symposium on Networked Systems Design and Implementation. 2013

[36] 王义君. 面向物联网的无线传感器网络时间同步与寻址策略研究[D]. 吉林：吉林大学，2012
Wang Yi-jun. Time synchronization and addressing strategy of IoT-oriented wireless sensor networks[D]. Jilin: Jilin University, 2012