ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.224−236
DOI: 10.3724/SP.J.1001.2008.00224
© 2008 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于属性相关度的Web数据库大小估算方法*

凌妍妍, 孟小峰+, 刘 伟

(中国人民大学 信息学院,北京 100872)

An Attributes Correlation Based Approach for Estimating Size of Web Databases

LING Yan-Yan, MENG Xiao-Feng+, LIU Wei

(School of Information, Renmin University of China, Beijing 100872, China)

+ Corresponding author: Phn: +86-10-62519453, E-mail: xfmeng@ruc.edu.cn, http://idke.ruc.edu.cn/xfmeng/

Ling YY, Meng XF, Liu W. An attributes correlation based approach for estimating size of Web databases. *Journal of Software*, 2008,19(2):224−236. http://www.jos.org.cn/1000-9825/19/224.htm

Abstract: An approach based on the word frequency is proposed in this paper to estimate the size of Web database. It obtains a random sample on a certain attribute by analyzing the attribute correlations among all the textual attributes in the query interface. The size of a Web database can be estimated by submitting probing queries which are generated by top-*k* frequent words to the query interface of a Web database. The experiments on several real-world databases have proved that this approach is effective and can achieve high accuracy in estimating the size of Web databases.

Key words: word frequency; Web database size estimation; attributes correlation

摘 要: 提出了一种基于词频统计的方法以估算Web数据库的规模.通过分析Web数据库查询接口中属性之间的相关度来获取某个属性上的一组随机样本;并对该属性分别提交由前 *k* 位高频词形成的试探查询以估算 Web 数据库中记录的总数.通过在几个真实的 Web 数据库上进行实验验证,说明该方法可以准确地估算出 Web 数据库的大小.

关键词: 词频;Web 数据库大小估计;属性相关度

中图法分类号: TP311 **文献标识码**: A

互联网的迅速发展使得 Web 中出现了越来越多的可以在线访问的数据库,通常把这些数据库称作 Web 数据库,所有的 Web 数据库就构成了 Deep Web(或 Hidden Web).据统计[1],目前整个 Deep Web 中 Web 数据库的数量超过了 45 万个,其中大约 3/4 的 Web 数据库存储的是结构化的信息.Deep Web 逐渐成为人们获取结构化信息的主要途径之一.以当当网为例,这是大家所熟悉的图书网站,人们通过向它提供的查询接口提交查询(如图 1(a)所示),得到满足查询的结果页面(如图 1(b)所示),并从中浏览和查找想要购买的图书.

* Supported by the National Natural Science Foundation of China under Grant No.60573091 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z155 (国家高技术研究发展计划(863)); the Program for New Century Excellent Talents in University of China (新世纪优秀人才支持计划); the Beijing Natural Science Foundation of China under Grant No.4073035 (北京市自然科学基金)

Received 2007-09-03; Accepted 2007-10-19

© 中国科学院软件研究所 http://www.jos.org.cn

(a) Query interface
(a) 查询接口

(b) Result Page
(b) 查询结果页面

Fig.1 Example of query interface and result page of Web database (Dangdang)

图 1 Web 数据库查询接口和查询结果页面示例(当当网)

为了帮助人们更加准确、高效地利用 Deep Web 上的海量信息,至今研究者们在 Deep Web 领域开展了大量的研究工作.文献[1]从宏观上对 Deep Web 的各方面特点进行了统计分析,包括 Deep Web 的规模、结构化程度以及覆盖的主题等,为研究者们从宏观上认识 Deep Web 提供了重要依据.另外,在 Deep Web 数据集成方面,对于 Web 数据库的发现[2,3]、查询接口的集成[4,5]、查询结果的抽取[6,7]等问题已经积累了相当多的研究成果.本文关注的则是 Deep Web 研究领域中的一个新的研究问题,即如何估算一个 Web 数据库的大小.该问题的研究意义主要有如下两个方面:

(1) 对 Deep Web 规模的宏观统计.随着 Web 的飞速发展,目前整个 Deep Web 中的 Web 数据库已经超过了 45 万个,但对于其所包含的信息总量却没有给出一个同步的估算数字,即目前整个 Deep Web 包含的信息量有多少 TB.这也使得 Deep Web 与 Surface Web 无法在信息量上进行对比.

(2) 每个 Deep Web 数据源只覆盖 Web 中局部的、有限的信息,于是 Deep Web 数据集成致力于对每个特定查询同时搜索多个与该查询相关的 Deep Web 数据源来扩大内容的覆盖性.出于对效率的考虑,我们需要选择出若干最相关的 Web 数据库进行查询.在此过程中,用户一般更倾向于选择信息更为丰富的 Web 数据库进行查询,这是因为信息越丰富的 Web 数据库可能会有越多的记录满足用户的查询.

基于上述两个方面的考虑,需要一种有效的方法来对 Web 数据库的大小进行估算,即较准确地估算出一个给定 Web 数据库中记录的总数.由于 Web 数据库具有高度的自治性,对它的访问只能通过其提供的查询接口,因此无法向 Web 数据库提交类似于 Select count(*) From Web database 的查询语句以直接得到记录总数.

为了进一步说明解决该问题所具有的挑战性,这里我们首先给出两种直观容易得到的方法,并同时指出它们的缺点或不可行性.一种方法是利用 Web 数据库在网页中提供的分类链接,比如当当网、易趣网等电子商务网站会给出所出售商品的导航式分类,引导访问者找到想要的商品.我们可以计算每个分类下的记录总数,然后累加得到总的记录数量.但并不是所有的 Web 数据库所在的网站都能提供分类链接,而且分类之间经常存在重复使得一个记录存在于多于 1 个的分类中,比如当当网中 GRE 相关的图书就存在于“考试”和“英语”两个分类中,这样就会使估计数字大于实际的记录数量.另一种方法是利用已经提出的爬取 Web 数据库的方法[8,9]获得 Web 数据库中的所有记录,这样可以得到最准确的记录总数.但是这种方法显然会使网络传输代价和本地存储代价过高,而且由于 Web 数据库经常处于频繁更新的状态,使得本地的副本难以维护.

因此,需要提出一种一般的方法,以便能够较为准确地估算一个任意 Web 数据库的大小.直觉上,给定 Web 数据库的某个文本属性 A,如果我们预先知道一个词 W 在 A 中所有属性值上的出现概率 p_W,并且知道在 A 的属性值中包含 W 的记录数量 n,就可以通过下面的公式估算出该 Web 数据库的大小 N:

$$N=n/p_W \tag{1}$$

通过观察发现,如果将 W 作为查询在属性 A 上进行发送,在 Web 数据库返回的结果页面中通常都会给出满足查

© 中国科学院软件研究所 http://www.jos.org.cn

询的记录总数(如图 1(b)中第 2 行文字所示),我们就可以确定公式(1)中的 n.基于这样的考虑,我们提出一种基于词频统计的解决方案.该方案的主要思想分为 3 步,简要描述如下:

第 1 步. 从 Web 数据库的查询接口模式中选取最合适的属性 A,并设法得到 A 上属性值的随机样本.通过随机样本进一步获取一组频繁的关键词以及它们各自在 Web 数据库中 A 属性上出现的概率.

第 2 步. 分别在该属性上提交这组关键词,并得到各自返回的记录数量.

第 3 步. 利用公式(1)为每个关键词计算出一个 Web 数据库大小的估算值,通过综合这组样本值最终得到唯一的估算值.

该方案的关键和挑战体现在第 1 步中的 3 个方面:(1) 如何从查询接口模式中选取最合适的属性 A 进行提交;(2) 如何获取 A 上属性值的随机样本;(3) 如何通过样本得到一组频繁出现的词及它们各自出现的概率.由于任何自然语言都是以词或字为基本组成单位的,因此我们所提出的方法与语言无关.

本文第 1 节对相关工作进行分析,指出以往工作的局限性和新的挑战.第 2 节首先提出一种 naïve 的方法,该方法利用自然语言上字频的一些统计成果来估算 Web 数据库的规模,并通过分析和验证指出该方法的缺陷;然后,针对存在的缺陷提出改进的策略,即本文的主体.改进的方法可以在很小的代价下比较准确地估算出一个 Web 数据库的大小.有关方法的详细内容分别在第 2 节和第 3 节加以介绍.第 4 节以中文为背景,通过实验对所提出的方法进行验证、比较,同时说明其有效性.第 5 节是结论和未来的工作.

1 相关工作

Web 数据库大小的估计.文献[10]是我们目前可以检索到的唯一一篇涉及 Web 数据库大小估算的文章.但是,这篇文章通过统计从宏观上将 Deep Web 与 Surface Web 作了对比,并非针对这个问题的研究论文.文中提出了若干并不可靠的估算方法,比如向 Web 数据库的管理者索取、Web 数据库在其网站上自动提供等.这些方法不是一般的方法,因而无法保证对任意一个 Web 数据库的大小进行估算.与之相比,我们的方法是一般性的,可应用于任何一个 Web 数据库,不受特定条件的约束.

搜索引擎大小的估计.与本文相关的工作还包括有关对搜索引擎或文档数据库大小的估算.首先我们将对这方面最新的工作进行简要的介绍.并在此基础上指出搜索引擎大小估计和 Web 数据库大小估计的本质区别以及现有方法的局限性和不适应性.

文献[11]于 2002 年最早提出估算文档集合大小的方法.该方法对待估算的文档集合进行采样,利用两个随机采样中重复的文档数目来估算总体的大小.用 N 代表文档集合的规模,如果我们分别独立随机采样了 a 个文档样本和 b 个文档样本,那么文档集合的大小可由下面的公式估算出来:

$$N=\frac{ab}{c} \tag{2}$$

其中,c 表示的是同时出现于两组采样中的文档个数.然而,文献[10]中并没有指明每次随机取样应该达到什么样的规模.该方法在很大程度上依赖于提交查询的数目以及样本的质量,导致一方面效率很低,另一方面该方法又是以不同采样样本相互独立为前提的,这很难得到保证.

文献[12]中介绍的是目前研究文献中在效率和准确性方面比较有名的工作.它假设为搜索引擎已经预先建立起了一个源描述(从搜索引擎中获得的一个样本文档集合),选取源描述中的一组词分别作为样本查询对该搜索引擎进行查询.因此搜索引擎的大小可以用下面的公式计算得到:

$$D=\frac{(D_T\times D_R)}{D_{RT}} \tag{3}$$

其中,D 表示搜索引擎的大小,D_T 表示包含词 T 的文档数目,D_R 是源描述中文档的数目,D_{RT} 是源描述中包含 T 的文档数目.最终搜索引擎的大小是每个词所得计算结果的平均值.该方法估算结果的准确性与预建立的源描述的质量存在非常强的依赖关系.而在搜索引擎的环境下,我们很难找到这样的一组词 T 使得 T 在源描述中的分布能够代表 T 在整个搜索引擎上的分布,因为搜索引擎是一个自治的黑盒,源描述对整体的代表性难以保证.

© 中国科学院软件研究所 http://www.jos.org.cn

文献[13]与前者的思想基本类似,同样是要求预先建立搜索引擎的源描述,最大的区别是它用文献[10]中的方法先对源描述的大小进行估计,并与源描述的实际大小比较得到估计的误差比例 CF,然后再将文献[10]中的方法应用于估计整个搜索引擎的大小,并用 CF 进行调整以降低估计误差.然而,文献[14]中通过实验指出,在对较大规模搜索引擎进行估计时[12],准确性下降得很快.文献[14]从本质上来说也是对文献[11]中的方法进行了扩展.不同于文献[11]中选取两个随机样本进行比较的方法,文献[14]依次从总体中选取 T 个大小均为 k 的随机样本,通过考察这 T 个样本两两之间重复的文档出现的情况,利用概率公式估算文档集合总体的大小.

总的来说,估计 Web 数据库的大小与估计搜索引擎的大小是不同的,可利用的信息也是不同的.我们的方法与上述搜索引擎大小的估计方法[11−14]有如下区别:

(1) 搜索引擎往往只提供基于关键词的文本搜索框,而 Web 数据库中蕴含的结构化数据意味着它必将提供给用户在多种不同属性上进行查询的服务.于是,我们在估计 Web 数据库大小时,会首先对属性进行分类.除了文本属性之外,分类属性和数值属性的存在使得我们可以更为快捷、方便地估计出一个 Web 数据库的大小.详见第 3.1 节.

(2) 上述所有估计搜索引擎大小的方法都是建立在获取独立的随机样本的基础上的.QBS[14]方法被用来实现样本的采集,即向搜索引擎的查询接口依次提交一批预定义的关键词,将返回的结果保存在本地,并从中随机选取关键词作为下一轮提交的查询.重复上述过程直至采集的样本数达到某一阈值.由于搜索引擎在返回结果时存在着固有的偏序性,长文本以及引用率高的文本往往更受青睐.很明显,这种不均匀性导致 QBS 产生的样本并不是真正随机的.相反,本文则充分利用了 Web 数据库查询接口模式丰富的特点,通过分析属性间的相关性关系,借助在相关性相对最小的属性上提交查询来获得本属性上一组真正随机的样本.实验显示,我们的方法具有很高的估算准确性.

2 一种粗糙的基于词频的估算方法

在本节中,我们以中文为背景,首先利用自然语言处理方面的成果提出一种 näive 的基于词频的解决方案,然后对其局限性进行分析.由于中文以汉字为最小单位,因此文章后面部分的词频实质上等同于汉字的字频.在语言研究中,词频是一个很重要的参数,即代表了汉字 W 出现的概率.

定义 1(词频). 设语料含 n 个汉字,其中汉字 W 出现 r 次,则 W 在这个语料中出现的频率定义为

$$p_w=r/n \tag{4}$$

以清华大学统计资料中的汉字频度表为例,其中使用字数 6 763 字(国标字符集),范文合计总字数 86 405 823 个,获取前 n 个频繁汉字及其对应的出现概率.结果见表 1.出现频率最高的前 6 个汉字分别为“的”、“一”、国”、“在”、“人”、“了”.

Table 1 Frequent Chinese word frequency (clips)

表 1 频繁汉字字频(片断)

Chinese word	Frequency	Probability
“的”	2 948 833	0.034 128
“一”	974 062	0.011 273
“国”	921 530	0.010 665
“在”	708 916	0.008 204
“人”	697 930	0.008 077
“了”	684 656	0.007 924
“有”	670 720	0.007 762
…	…	…

为了在较小代价下估算一个 Web 数据库 WD_i 的规模,我们需要通过其查询接口提交尽可能少的查询.因此,如何有效地选取查询关键字是其中一个关键的问题.我们发现,如果将位于汉字字频表中前 n 位的频繁汉字用作查询关键字进行提交,将具有如下优势:(1) 用作提交的查询关键字越短(仅为单个汉字),则被该关键字覆盖的记录越广泛;(2) 用作提交的查询关键字字频越高(选取频繁汉字字频表的前 n 个),在 WD_i 中出现的可能性越

© 中国科学院软件研究所 http://www.jos.org.cn

大,对应返回的记录数越多.于是,为了估算 WD_i 的规模,我们从 WD_i 的查询接口中选取一个属性,将这些字作为属性值分别提交至 Web 数据库.如图 2 所示,我们分别将前 5 个频繁汉字提交至当当网(dangdang.com)的书名属性,观察返回的结果数,对应为(“的”,0),(“一”,15589),(“国”,328),(“在”,3294),(“人”,2054).需要注意的是,有些字提交之后返回的结果是 0,如上例中的“的”,这表示该字在 WD_i 中是被当作停止词来处理的,因此不必加以考虑.

Fig.2 Submit frequent Chinese-word to Dangdang

图 2 将频繁汉字提交至当当网

利用某个频繁汉字 W 的字频 p_w 以及将 W 作为查询关键字提交至 WD_i 返回的结果数目 n,我们可以利用公式(1)估算 WD_i 相对于汉字 W 的规模 N_w.为了综合由各个具体的汉字得到的估算数据库大小的样本值,进而最终得到唯一的估算值,我们采取如下方法:去掉最大的样本值以及最小的样本值,取其他样本值的平均数作为对 WD_i 规模的最终估计值.

局限性分析:在上述基于词频对 Web 数据库规模进行估算的方法中,词频的正确与否对估算的准确性有着至关重要的影响.一方面,上节中频繁汉字的词频是基于大规模语料统计分析得出的,可以被认为代表了汉字 W 在中文中最普遍的出现规律.另一方面,Web 数据库中的每个属性都具有某一特定的语义,因此,不同的属性其字频也不相同.以姓名属性为例,统计结果表明,在姓氏中频率最高的为“王”、“陈”、“李”、“张”、“刘”5 个姓,占了总样本数的 32%;在人名中频度最高的为“英”、“华”、“玉”、“秀”、“明”、“珍”6 个字,覆盖率达 10.35%.由此可以看出,WD_i 不同属性中的汉字词频各有不同,因此我们不能一概而论地用表 1 中词频的一般情况来估算 WD_i 的规模.

3 基于属性相关度的估算方法

根据上述分析,基于词频的估算方法对词频参数的准确性提出了很高的要求.因此我们需要一种改进的策略来对不同的属性估计其特有的词频信息.本节利用了 Deep Web 数据源特有的复杂查询接口模式,在对属性分类的基础上,根据不同 Web 数据库拥有的属性类型不同,制定自适应的估算数据库规模的方案.同时,在分析不同领域单属性值查询可行性的基础上,解决了占多数比例的文本型属性中属性值词频的估计问题.

3.1 属性分类

在 Web 数据库 WD_i 中,不同的属性不仅语义不同,而且在用于估算 WD_i 的规模上所起的作用也不同.由此,我们将属性归为如下 3 类:

(1) 分类属性.分类属性的属性值是一个有限的集合,在 Web 数据库查询接口上通常以下拉列表的形式出现,如图 1(a)中的属性“折扣”和“上架时间”.此外,还有一类隐藏的分类属性,它们在 Web 数据库的查询接口上并不以下拉列表的形式显式地罗列所有的分类值,但是它们的值域往往是容易获取的离散值集合.比如航班领域的属性“国家”和“城市”.如果 WD_i 允许在某个分类属性上单独提交查询,则该属性上所有分类返回的结果条数之和即为 WD_i 的实际规模.但是,绝大多数的 Web 数据库不支持单个分类属性上的查询.

(2) 数值属性.顾名思义,数值属性的属性值可以是不同类型的数值,如价格、时间或普通数字等.图 1(a)中

© 中国科学院软件研究所 http://www.jos.org.cn

的属性“当当价”和“出版时间”就属于该类属性.数值属性虽然不像分类属性那样有明确的几个取值,但数值属性的取值范围往往是比较容易估计的.因此,如果 WD_i 允许在某个数值属性上单独提交查询,那么该属性上所有可能的取值范围对应的结果数目之和就是 WD_i 的实际规模.比如,从图 1(a)中我们可以比较容易地估计出图书“当当价”的取值范围,并进行提交.但是,根据我们的观察,一般情况下,单个数值属性上的查询也是得不到支持的.

(3) 文本属性.文本属性是出现最广、处理最为复杂的一类属性,取值范围是无限的,往往以文本框的形式出现在 WD_i 的查询接口上.图 1(a)中的“书名”、“著译者”和“出版社”就属于文本属性.绝大多数的 Web 数据库在提交查询时都要求至少 1 个的文本属性上有值.因此下文将重点讨论如何通过 Web 数据库查询接口模式中的文本属性来估算 Web 数据库的规模.假设 WD_i 支持在某个文本属性 A 上单独进行查询,则我们必须设法获取 A 上属性值的随机样本,并进一步取得属性 A 上最频繁出现的若干个词及它们的出现概率(词频),从而根据公式(1)的原理估算 WD_i 的规模.第 3.3 节中将具体阐述利用文本属性间的相关性计算某单个文本属性上词频的方法.

3.2 单属性值查询可行性

在本节中,我们将对实际的统计数据进行分析,考察不同领域的 Web 数据库对单属性值查询的支持程度以及文本型属性所占的比例.

文献[9]给出的统计数据显示(见表 2),绝大多数领域的 Web 数据库都支持单属性值查询,即允许提交只涉及单个属性的查询.少数主题诸如汽车等查询接口的结构严密,语义限制较多,因此部分情况下只支持多属性值查询.但是,同时我们也发现,它们的文本型属性所占的比例相当小.因此,对于这样的网站,我们重点根据分类型属性和数值型属性上属性取值的有限种组合,即可估算出其底层 Web 数据库的规模(见第 3.1 节).不可否认,大多数主题下的 Web 数据库还是存在较难处理的文本型属性的,并且也支持在单个属性值上提交查询,比如图书(Book)领域支持在文本属性“书名”或“作者”上进行单属性值查询,工作领域支持在文本属性“职位描述”上进行单属性值查询.也就是说,在这样的情况下,一方面 Web 数据库允许我们选择合适的查询在某个单个的文本型属性上进行提交;另一方面,更重要的内容是,我们需要知道每一个提交的查询词在该文本属性上对应的特定词频,而不是表 1 中列举的统计的一般情况.

Table 2 Support degree of single attribute query

表 2 单属性值查询支持程度

Domain	Support degree (%)	Domain	Support degree (%)
Book	100	DVD	96
Job	96	Computer	96
Movie	100	Game	96
Automobile	58	Furniture	100
Music	100	Jewelry	100

3.3 基于相关性分析的词频获取

当仅通过查询接口中的分类属性或数值属性无法提交查询以估算 Web 数据库 WD_i 的规模时,选取某个合适的文本型属性并估算若干频繁出现的词在其上的出现概率就势在必行了.假设 A 是 WD_i 中的一个文本型属性,为了计算查询词 W 在 A 上的出现概率,一个直观的方法是向 WD_i 提交一批预定义的适合于属性 A 的查询,将所有返回结果构成的并集记为 R.用 R 作为样本来代表该 Web 数据库在 A 上属性值的全体,认为 W 在 R 中出现的概率就近似等同于 W 在 A 上出现的概率.用这样的方法估计出的词频与实际值之间存在着很大的偏差,因为预定义的查询始终不是随机的,将会不可避免地造成返回结果集 R 不是 A 上所有属性值的一个真正随机的样本,即 R 中属性值的分布与实际数据库内 A 属性值的分布具有很大的不同.因此,在本节中,我们将借助于对 WD_i 中不同文本型属性之间的相关性的分析来获取 A 上属性值的随机样本,从而估计查询词 W 在 A 上的词频.

首先,我们需要明确属性相关性的定义.假设用 A_1 和 A_2 分别代表 WD_i 中两个不同的文本型属性,向 A_1 中提交 t 个预定义的查询 $q_1,q_2,\ldots,q_t$(保证 t 个查询独立且各不相同),并设每一个查询 $q_i(1\le i\le t)$返回的结果在属性 A_2

© 中国科学院软件研究所 http://www.jos.org.cn

上的取值的集合为 R_i.如果在不同查询 q_i 得到的不同结果集 R_i 中,构成属性值的词的分布情况具有明显的差异性,则认为 A_1 和 A_2 这两个属性不是独立的,即存在着相关性联系.此外,不同属性之间的相关性程度有强有弱.不同 R_i 中词的分布情况差异性越大,则认为 A_1 和 A_2 越相关,即 A_1 的取值对 A_2 的取值决定性越强;反之亦然.

如果以图 1(a)中的文本型属性“书名”和“著译者”为例,在“著译者”上提交若干个作者名,每个作者名对应地返回一个书名的集合.往往不同的作者由于写作主题的不同,造成不同作者的著作在书名上的差异性很大.也就是说,“书名”和“著译者”之间存在着比较强的相关性,即“著译者”对“书名”有比较强的决定作用.作为对比,再以另两个文本型属性“书名”和“出版社”为例.在“出版社”上提交若干个出版社名称,对应地返回每个出版社出版的书名集合.往往每个出版社涉及的书都涵盖很多不同的主题,虽然具体的书名不完全相同,但是从词的分布情况来考察,各个出版社的差异性却不大.也就是说,“书名”和“出版社”之间的相关性相对来说较弱,即“出版社”的取值不同并不能直接造成“书名”上属性值分布的巨大差异.

于是,我们需要一种量化的手段来衡量不同文本型属性之间的相关性程度,简称属性的相关度.

如前面提到的,我们将在属性 A_1 上提交的第 i 个预定义的查询 $q_i(1\le i\le t)$返回的结果在属性 A_2 上的取值集合记为 R_i,并用向量 $\boldsymbol{V}_i$:$\langle p_{i1},p_{i2},\ldots,p_{in}\rangle$代表 R_i 中词频的分布情况.其中,每个元素 $p_{ij}(1\le i\le t,1\le j\le n)$代表第 j 个词 W_j 在 R_i 中的词频(这里假设 $\bigcup_{1\le i\le t} R_i$ 共存在 n 个不同的词).属性 A_1 对 A_2 的相关度定义如下:

定义 2(属性相关度). 相对于 A_1 上 t 个查询 $q_1,q_2,\ldots,q_t$,如果对应得到关于 A_2 的 t 个 n 元向量 $\boldsymbol{V}_1,\boldsymbol{V}_2,\ldots,\boldsymbol{V}_t$,则属性 A_1 对 A_2 的相关度计算如下:

$$Dependency(A_1,A_2)=1-\left(\frac{1}{n}\right)\times\frac{\sum_{i=1}^{t}\left(\boldsymbol{V}_i-\overline{\boldsymbol{V}}\right)^2}{t} \tag{5}$$

由公式(5)不难看出,向量之间的方差公式被用于评估 t 个 n 维空间向量之间的差异性,即由 t 个 A_1 上的查询所得到的 t 个 A_2 属性值上词频分布的差异.方差越小表示 t 个词频分布的差异越小,那么属性 A_1 对 A_2 的相关度就越大.其中,$1/n$ 只是一个用于对方差值进行规范化的因子,它使得我们将属性的相关度控制为 0,1 之间的实数.公式(5)中向量之间的距离定义如下:

$$\boldsymbol{V}_i-\overline{\boldsymbol{V}}=\sqrt{\sum_{k=1}^{n}\left(\left|p_{ik}-p_k\right|\right)^2} \tag{6}$$

其中,$\overline{\boldsymbol{V}}$:$\langle p_1,p_2,\ldots,p_n\rangle$代表 t 个 n 元向量 $\boldsymbol{V}_1,\boldsymbol{V}_2,\ldots,\boldsymbol{V}_t$ 的均值.

依照定义 2 对图 3 中的 3 个文本属性“书名”、“著译者”和“出版社”进行属性相关程度的分析可知,“书名”对“出版社”之间的相关性是所有属性间相关性最小的.这与实际情况也是相符的.

	A_1	A_2	A_3	…	A_m
A_1	D_{11}	D_{12}	D_{13}	…	D_{1m}
A_2	D_{21}	D_{22}	D_{23}	…	D_{2m}
A_3	D_{31}	D_{32}	D_{33}	…	D_{3m}
…	…	…	…	…	…
A_m	D_{m1}	D_{m2}	D_{m3}		D_{mm}

Fig.3 Attributes correlation matrix

图 3 属性相关度矩阵

假设一个 Web 数据库 WD_i 的查询接口模式中含有 M 个支持单独提交查询的文本型属性 $A_1,A_2,\ldots,A_m$.任意属性 $A_i(1\le i\le m)$对 $A_j(1\le j\le m)$的相关度都可以藉由定义 2 得以量化.于是,对 WD_i 我们可以得到一个 $M\times M$ 的相关度矩阵,矩阵中的每一项 $D_{ij}(1\le i\le m,1\le j\le m)$代表属性 A_i 对 A_j 的相关度 $Dependency(A_i,A_j)$,如图 3 所示.

为每一个待估算的 Web 数据库建立一个相关度矩阵是后续一切工作的基础.藉由该矩阵,我们需要:(1) 选定最合适的文本属性 A_k 提交查询以估算 WD_i 的大小;(2) 通过在与 A_k 最不相关的属性 A_L 上提交一批查询来获取属性 A_k 上的随机样本,并统计 A_k 属性特有的频繁词及词频;(3) 利用相关度来对因样本的误差性造成的数据

© 中国科学院软件研究所 http://www.jos.org.cn

库规模估算偏差进行矫正,从而更准确地估计 WD_i 的大小.下面我们将具体阐述基于属性相关性的数据库大小估算策略.

从图 3 所示的相关度矩阵(记为 $\boldsymbol{M}_i$)中找出最小的单元值,设为 D_{KL},即属性 A_K 对属性 A_L 的相关度最小.前面提到过,通过属性 A_L 本身提交查询来获取样本从而估计 A_L 属性值上词频的方法是不可取的,因为该属性上预定义的查询很难保证样本的随机性.而一旦我们从 WD_i 的查询接口中找到了两个相关度最小的属性,则可以通过在其中一个属性上提交若干查询来得到另一个属性上属性值的随机样本.

沿用上述定义,属性 A_K 和属性 A_L 是两个最不相关的属性,那么假设在 A_K 上提交一批预定义的适合于属性 A_K 的查询,并将所有返回结果中由 A_L 属性值构成的并集记为 U.此时的 U 是最接近于 A_K 属性值词频真实分布的随机样本,因为 A_K 对 A_L 相对很弱的相关性决定了 A_K 上的查询对 A_L 上的取值造成的主观性影响很弱.因此,此时的 U 可以被认为相对最客观地代表了整个 WD_i 在 A_L 上属性值的全体.从属性值中词的角度来说,词 W 在 U 中的出现概率也就代表了 W 在属性 A_L 上的真实词频.

同样地,我们用向量 $\boldsymbol{V}$:$\langle p_1,p_2,\ldots,p_n\rangle$来代表随机样本 U 中词出现的概率(假设 U 中存在 n 个不同的词).其中,每个元素 $p_i(1\le i\le n)$代表 U 中第 i 个词 W_i 的词频.为了更准确地得到 WD_i 规模的估算值,我们取词频前 m 位的频繁词作为查询关键词在对应的属性 A_L 上进行提交,分别观察返回的结果数并利用公式(1)估算 WD_i 相对于每个频繁词的规模 $W_i(1\le i\le n)$.为了综合由各个具体的词得到的样本估算值,我们在去掉最大样本值及最小样本值的基础上取平均值作为对 WD_i 规模的最终估计值 N_{EST}.

$$N_{EST}=\frac{\sum_{2\le i\le m-1}\left(\frac{n_i}{p_i}\right)}{m-2} \tag{7}$$

其中,p_i 代表提交的第 i 个频繁词 W_i 在属性 A_L 上的词频,n_i 代表相应返回的结果数目.由于去掉了最大的样本估算值和最小的样本估算值,故公式(7)中只对 $m-2$ 个样本值计算平均数.本文实验部分说明了选取词频前 m 位的频繁词进行提交比随机选取 m 个词进行提交估算出来的 Web 数据库大小要更接近客观现实,详见第 4.2 节.

3.4 基于相关度的估算值矫正

上述方法作用于两个最不相关的文本型属性 A_K 和 A_L,通过 A_K 上的查询来获取 A_L 上属性值的随机样本进而客观地统计词频.但是需要指出的是,任意两个文本型属性 A_K 和 A_L 之间总是存在着一定的属性相关度,即并不完全独立.只是属性 A_K 对 A_L 的相关度越小,我们得到的 A_L 上属性值的样本就越随机,样本中词的分布情况也就越接近于 A_L 上实际属性值中词的分布情况.换言之,由于属性并不完全独立,我们得到的样本分布情况始终与实际属性值的分布情况存在着或多或少的偏差.因此,我们提出将属性之间的相关度作为影响因子来对由非完全客观样本估算出来的 WD_i 的规模 N_{EST} 进行矫正.

为了考察属性之间的相关度与估算 WD_i 规模时产生的误差这两者之间的联系,我们对训练集中的两个 Web 数据库 JobTong 和 Book 进行如下操作:统计其中所有支持单独提交的文本型属性(JobTong 中有 7 个,Book 中有 8 个),对每个 Web 数据库形成一个如图 3 所示的相关度矩阵.任意选取其中两个属性并利用其相关性来估算 WD_x 的规模.得到的估算值 N_{EST} 连同属性的相关度和 WD_x 的实际规模一起形成训练样本中的一行,见表 3,共形成 42+56=98 行.其中,估算误差ε为估算值与实际规模之间的差值.估算值小于实际值,ε为正,反之为负.

Table 3 Attribute correlation affect the estimating accuracy

表 3 训练属性相关度对规模估算的影响

Attribute correlation D	Estimating size N_{EST}	Actual size N	Estimating error ε
1.33E−04	60 571	69 317	8 746
6.77E−05	70 939	69 321	−1 618
…	…	…	…
…	…	…	…
1.61E−04	79 802	69 322	−12 481

为了使基于属性相关度的估算值更接近于 Web 数据库的实际规模,我们通过回归分析来考察属性相关度

© 中国科学院软件研究所 http://www.jos.org.cn

D 和估计误差ε之间的关系,通过六次多项式函数逼近得到以ε为因变量、D 为自变量的回归方程:

$$\begin{cases}\varepsilon = N - N_{EST} = -(3\text{E}{+}24)D^6 + (8\text{E}{+}21)D^5 - (7\text{E}{+}18)D^4 + (3\text{E}{+}15)D^3 - (6\text{E}{+}11)D^2 + (2\text{E}{+}08)D + 12069 \\ R^2 = 0.825\end{cases} \tag{8}$$

其中,R^2 的值代表该回归方程的拟合度.

综上所述,在基于相关性分析估计 WD_i 规模的策略中,我们首先考虑在分类属性或数值属性上提交查询,将有限次查询结果之和作为对 WD_i 规模的估计.其次,如果 WD_i 要求文本型属性上必须有值,则通过对多个文本型属性进行相关性分析,找到最不相关的属性 A_K 和 A_L.借助于属性 A_K 上提交一批预定义的查询,获取属性 A_L 在 WD_i 中足够随机分布的一个样本,可以认为汉字 W 在该样本中的词频代表了 W 在属性 A_L 上的词频.当我们拥有了若干个频繁汉字在属性 A_L 上特定的词频信息之后,可以将这若干个汉字提交至查询接口的 A_L 属性以获得相应的结果数,此时公式(7)就可以很容易地被用来估算 WD_i 的规模了.最后,我们还需要借助公式(8)来估计因属性 A_K 对 A_L 的相关度 D_{KL} 引起的估算误差ε,最终得到合理的估算值.

为了能够得到在某个属性上比较客观的字频,理论上需要获取尽可能多的记录作为统计的样本,但获取大量的记录必然使代价提高,因此我们需要确定选取记录的数量,进而统计一个属性上的字频.我们将在实验部分证实记录超过一定数量时,字频就会趋于稳定.通过这个实验我们可以得到一个常数来确定获取记录的数量.

4 实 验

本节我们首先介绍用于实验的数据集,然后给出各项实验以及对实验结果的分析.

4.1 数据集

我们选取了内容丰富的数据集,分别用于训练和测试.每个数据集的来源和大小见表 4.训练集主要用于估计本文方法中涉及的一些参数,如我们将设法通过训练集来获取如公式(8)所示的矫正函数等.JobTong-1 和 JobTong-2 是来源于 WAMDM 实验室自行开发的本地数据库 JobTong 的两个独立子集.JobTong 数据库在本地存储了数十万条招聘信息,并可通过 http://www.jobtong.cn 进行访问.同样地,Book-1 和 Book-2 也是来源于本地数据库的两个独立子集,其中存储了数十万条图书信息.选用这 4 个数据集作为训练集的优势在于,这些数据来源于本地,我们可以精准地获取数据集的实际大小.测试集则主要由第三方的 Web 数据库组成,分别来源于图书、音乐、影视和求职领域.我们需要通过测试集来评估利用本文方法对 Web 数据库进行估算的准确度.需要指出的是,网络上现存的 Web 数据库由于商业原因一般都不对外公开自己的数据库规模.因此我们选取那些可以通过首页上的分类链接获得各个类别下记录总数的网站,通过把各个分类的记录数相加来近似地模拟测试集的大小.

Table 4 Training and testing dataset

表 4 用于训练和测试的数据集

Training set	Size	Testing set	Size
JobTong-1	69 317	Joyo	130 274
JobTong-2	69 321	ChinaHR	437 063
Book-1	57 368	Dangdang-Music	40 508
Book-2	58 426	Dangdang-Movie	48 900

4.2 实验及结果分析

我们设计了 4 个方面的实验.实验 1 主要考察属性相关度和估算准确性之间的关系,说明了选取两个相关度最小的属性来获取随机样本从而估算出来的 Web 数据库规模准确性最高.实验 2 主要考察词频准确性与随机样本记录数量之间的关系,说明了当随机样本的数量大于 4 000 条时,得到的词频就会趋于稳定并且与整个数据库上的实际词频基本一致.实验 3 主要考察了用于提交的词的词频大小与估算准确度的关系,说明了从随机样本中选取前 5 个频繁词进行提交,并将对应的 5 个估算值的平均数作为估算结果时,估算的准确性最高.实验 4 则是在 4 个不同领域的测试集上考察估算的准确度,说明了基于相关度的估算值矫正方法是行之有效的.

© 中国科学院软件研究所 http://www.jos.org.cn

实验 1:相关度和估算准确性的关系.

假设 Web 数据库 WD_i 中存在 M 个支持单独提交查询的文本型属性 $A_1,A_2,\ldots,A_m$,不同的属性之间相关度不同.我们需要从中选择一对属性来获取随机样本进而估算 WD_i 的大小.为了考察选择不同相关度的属性对是否会对估算准确性造成影响,我们从 JobTong-1(JobTong-2)中分别找出 7 个支持单独提交查询的文本属性构造如图 3 所示的相关度矩阵,并针对其中的每一个相关度值(共有 42 个)都进行了一次对 JobTong-1(JobTong-2)规模的估算,估算结果如图 4 所示.其中横轴代表了属性对的相关度,图中的每一个标记点代表使用相应相关度的两个属性进行估算所得到的 WD_i 大小的估算值(纵轴).从图中的趋势线不难看出,使用的属性对相关度越高,估算出来的规模与实际规模之间的差异越大;反之,如果我们使用的属性越不相关,估算出来的规模则越接近于实际值.我们对另两个训练集 Book-1 和 Book-2 也进行了相同的实验,结果如图 5 所示.可以看出,Book-1 和 Book-2 的趋势线也印证了同样的规律,即为了保证对 Web 数据库规模的估算值最接近于实际规模,我们需要选择两个最不相关的属性进行操作.

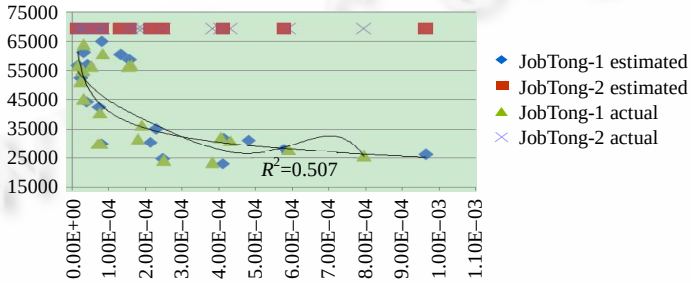

Fig.4 Relation between attribute correlation and estimating accuracy on JobTong-1 and JobTong-2

图 4 JobTong-1 和 JobTong-2 上属性相关度和估算准确性的关系

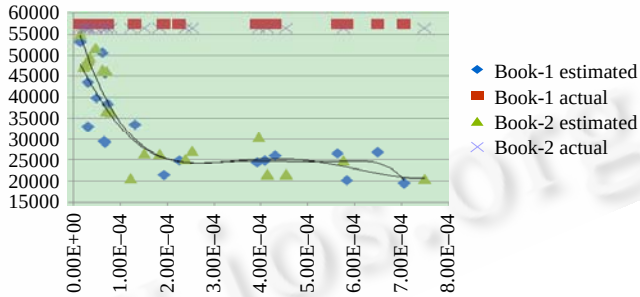

Fig.5 Relation between attribute correlation and estimating accuracy on Book-1 and Book-2

图 5 Book-1 和 Book-2 上属性相关度和估算准确性的关系

实验 2:词频准确性与样本记录数量的关系.

当在 Web 数据库 WD_i 中找到两个最不相关的属性 A_K 和 A_L 之后,我们需要借助于属性 A_K 上的查询来获取属性 A_L 上的随机样本以获取 A_L 上属性值的词频分布.于是,我们需要考察词频分布的准确性是否会受到随机样本数量的影响.

我们在 JobTong-1 和 JobTong-2 的并集 JobTong 上首先利用属性相关性分析获得两个相关度最低的属性:职位名称和公司名称.通过向职位名称提交若干查询(比如经理、教师等),得到公司名称上属性值的一个样本.我们取前 5 个频繁词(司,公,险,有,保),并分别在 500,1 000,2 000,3 000,4 000,5 000,10 000,20 000 和 25 000 条记录上观察公司名称属性中这些词出现的概率,如图 6 所示,横轴代表样本数量,纵轴代表某个频繁词在某个样本数量下统计出来的词频.其中,样本数量为 0 时的词频代表该词在整个集合 JobTong 中的实际词频.同样地,我们在 Book-1 和 Book-2 的并集 Book 上也获得了两个相关度最低的属性:出版社和书名.取书名属性的前 5 个频繁词(学,书,教,中,系)并分别在不同规模的样本记录集上观察词频的变化,如图 7 所示.

© 中国科学院软件研究所 http://www.jos.org.cn

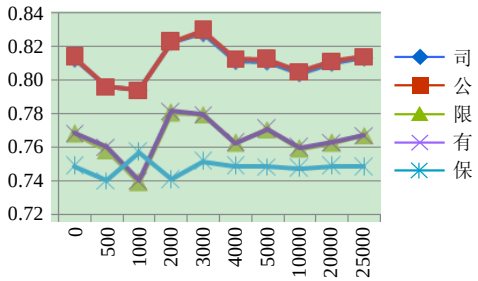

Fig.6 Word-Frequency and sample size (JobTong)

图 6 JobTong 上词频与样本数量的关系

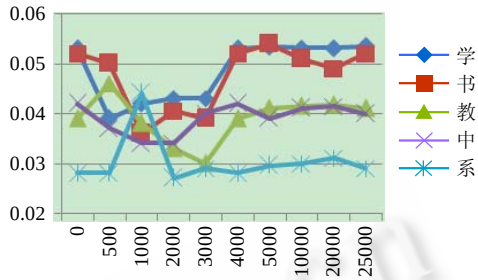

Fig.7 Word-Frequency and sample size (Book)

图 7 Book 上词频与样本数量的关系

不难发现,JobTong 和 Book 上词频的变化都遵循一个共同的特点,即在少于 4 000 条记录的样本上得到的词频并不稳定,而大于 4 000 条得到的词频就会趋于稳定.另外,通过与实际词频(横轴为 0)的比较发现,当样本的规模达到 4 000 条时,得到的词频就与整个数据库上的词频基本一致了.因此,我们把 4 000 作为在不相关的属性上获取随机样本记录数量的最小值.实验中我们对每一个测试集都至少获取了 4 000 个记录作为字频统计样本.

实验 3:词频大小与估算准确性的关系.

根据实验 2 的结果,我们将为 WD_i 获取一个属性 A_L 上的随机样本,并保证其至少包含 4 000 条记录.由此得出的 A_L 上属性值的词频分布将是最接近于实际分布的.为了从中选择词频最接近真实的若干词进行提交,我们需要考察词频的真实性是否与词频大小,即词的出现频繁程度有关.于是我们对训练集 JobTong-1 和 JobTong-2 都分别获取公司名称属性上属性值的一个随机样本,并根据词频的大小对样本中的所有词进行排序.取词的频繁度位于前 20 位的词分别进行提交得到对应于每个词的 WD_i 的估算值.图 8 和图 9 中的每个标记点表示取前 x 位(横轴)频繁词对应估算值的平均数作为最终的估算值(纵轴).很明显,在图 6 中,估算值在 x 取 5 时最接近于 JobTong 的实际大小,即当我们取前 5 个频繁词对应估算值的平均数作为估算结果时,估算最准确.我们对另两个数据集 Book-1 和 Book-2 进行了相同的实验,实验结果如图 7 所示.可以看出,虽然两组曲线走势起伏不同,但它们都有一个共同的特点,即估算值在 x 取 5 时达到峰值,之后有明显的下降趋势.也就是说,当我们通过属性 A_L 上的随机样本获得了一个词频的有序序列之后,如果取前 5 个频繁词对应估算值的平均值作为 WD_i 的最终估算值,则估计的准确性最高.

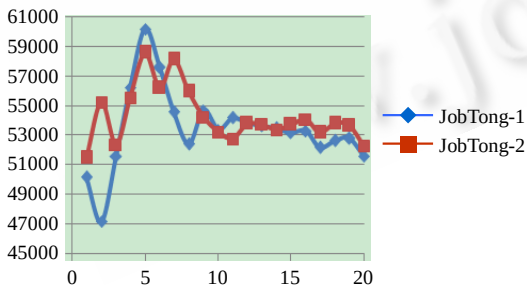

Fig.8 Word-Frequency and estimating accuracy (JobTong)

图 8 JobTong 上词的频繁程度与估计准确性的关系

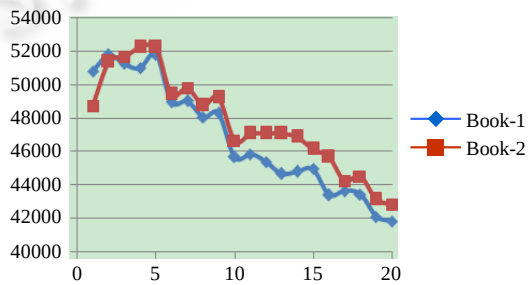

Fig.9 Word-Frequency and estimating accuracy (Book)

图 9 Book 上词的频繁程度与估计准确性的关系

实验 4:对测试集的估算结果.

我们将基于属性相关度的估算策略用于对 4 个不同领域的测试集估算其 Web 数据库的大小.以中华英才网(ChinaHR)为例,我们通过其首页提供的行业分类链接访问到各个行业的职位数目,相加后得到整个 Web 数

© 中国科学院软件研究所 http://www.jos.org.cn

据库的实际大小是 437 063 条记录.然后,对查询接口中所有的文本属性计算属性之间的相关度并进行排序,得到两个最不相关的属性“城市名称”和“职位名称”.我们将词频前 5 位的“程”、“主”、“师”、“务”、“工”分别作为关键词查询提交到相应的职位名称属性上,根据词频和返回记录数就可以估算出 ChnaHR 的大小,结果见表 5.从表中估算的准确性可以看出,我们用 5 个频繁词对应的估算值的均值作为矫正前的估算值,达到的估算准确度为 96.58%.通过公式(8)得到最小属性相关度(0.0024%)下的估计误差并对估算值加以调整,作为矫正后的估算值,此时,估算准确度提高为 99.87%.

Table 5 Estimating result for ChinaHR

表 5 对 ChinaHR 的估算结果

Key Word	Frequency	Returned result	Estimating size
“程”	0.164 506	87 711.38	533 178.8
“务”	0.191 924	94 332.16	491 507.3
“工”	0.232 081	98 512.23	424 473.8
“师”	0.235 498	72 933.84	309 699.9
“主”	0.312 261	109 811.1	351 664.1

同样地,在表 6 中我们给出了 4 个测试集上估算结果的汇总.不难看出,基于属性相关度分析来估算 Web 数据库大小的方法在不同的实际数据源上均能达到 90%以上的估算准确度.同时,实验数据也表明,以相关度为因子去估计可能的估算误差ε的方法是行之有效的,因为通过ε进行矫正后的估算精度与之前的估算精度相比都提高了至少 3 个百分点.

Table 6 Estimating results on testing dataset

表 6 测试集上的估算结果

	ChinaHR	Joyo	Music	Movie
Actual size	437 063	130 274	40 508	48 900
Estimating size (before tuning)	422 104	117 301	37 735	46 008
Estimating accuracy (before tuning) (%)	96.58	90.04	93.15	94.09
Estimating size (after tuning)	436 522	124 226	42 039	50 312
Estimating accuracy (after tuning) (%)	99.87	95.36	96.22	97.11

5 总结和未来工作

为了在较小代价下准确估算 Web 数据库的大小,我们提出了一种基于属性值词频的解决方案,同时给出了一种 naïve 的方法和一种基于查询接口上属性相关性分析的改进方法.通过在 4 个真实的 Web 数据库上进行实验验证,说明了改进的方法可以准确地估算出 Web 数据库的大小.

虽然我们的方法是与特定自然语言无关的,但本文的实验只在中文 Web 数据库上进行了验证.因此,在未来的工作中,我们将选取一些英文 Web 数据库进行实验.另外,我们的方法还无法处理查询接口上只有 1 个文本属性的情况,今后我们将针对这种情况加以解决.

References:

[1] Chang KCC, Cho J. Accessing the Web: From search to integration. In: Proc. of 2006 ACM SIGMOD Int’l Conf. on Management of Data (SIGMOD 2006). Chicago: ACM Press, 2006. 804−805.

[2] Cope J, Craswell N, Hawking D. Automated discovery of search interfaces on the Web. In: Proc. of the 14th Australasian Database Conf. (ADC 2003). Adelaide: Australian Computer Society Press, 2003. 181−189.

[3] Kabra G, Li C, Chang KCC. Query routing: Finding ways in the maze of the deep Web. In: Proc. of the Int’l Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2005). Tokyo: IEEE Computer Society Press, 2005. 64−73.

[4] He H, Meng W, Yu CT, Wu Z. WISE-Integrator: An automatic integrator of Web search interfaces for e-commerce. In: Proc. of the 29th Int’l Conf. on Very Large Data Bases (VLDB 2003). Berlin: ACM Press, 2003. 357−368.

© 中国科学院软件研究所 http://www.jos.org.cn

[5] Wu W, Doan A, Yu CT. WebIQ: Learning from the Web to match deep-Web query interfaces. In: Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE 2006). Atlanta: IEEE Computer Society Press, 2006. 44.

[6] Zhai Y, Liu B. Web data extraction based on partial tree alignment. In: Proc. of the 14th Int'l World Wide Web Conf. (WWW 2005). Chiba: ACM Press, 2005. 76−85.

[7] Zhao H, Meng W, Wu Z, Raghavan V, Yu CT. Fully automatic wrapper generation for search engines. In: Proc. of the 14th Int'l World Wide Web Conf. (WWW 2005). Chiba: ACM Press, 2005. 66−75.

[8] Raghavan S, Garcia-Molina H. Crawling the hidden Web. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB 2001). Rome: ACM Press, 2001. 129−138.

[9] Wu P, Wen JR, Liu H, Ma WY. Query selection techniques for efficient crawling of structured Web sources. In: Proc. of the 22nd Int'l Conf. on Data Engineering (ICDE 2006). Atlanta: IEEE Computer Society Press, 2006. 47−58.

[10] BrightPlanet.com. The deep Web: Surfacing hidden value. 2000. http://brightplanet.com

[11] Liu KL, Yu CT, Meng W. Discovering the representative of a search engine. In: Proc. of the 11th Int'l Conf. on Information and Knowledge Management (CIKM 2002). McLean: ACM Press, 2002. 652−654.

[12] Si L, Callan JP. Relevant document distribution estimation method for resource selection. In: Proc. of the 26th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR2003). Toronto: ACM Press, 2003. 298−305.

[13] Karnatapu S, Ramachandran K, Wu Z. Estimating size of search engines in an uncooperative environment. In: Proc. of the 2nd Int'l Workshop on Web-Based Support Systems 2004 (WSS 2004). Beijing: IEEE Computer Society Press, 2004. 81−87.

[14] Shokouhi M, Zobel J, Scholer F, Tahaghoghi SMM. Capturing collection size for distributed non-cooperative retrieval. In: Proc. of the 29th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR 2006). Seattle: ACM Press, 2006. 316−323.

凌妍妍(1985－),女,安徽黄山人,硕士生,主要研究领域为 Deep Web 数据集成.

刘伟(1976－),男,博士生,主要研究领域为 Deep Web 数据集成,Web 数据抽取.

孟小峰(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据集成,XML 数据管理,移动数据管理.

© 中国科学院软件研究所 http://www.jos.org.cn