软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(4):781−797 [doi: 10.3724/SP.J.1001.2013.04374]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

Web 数据源选择技术*

万常选[1,2], 邓 松[1,2], 刘喜平[1,2], 廖国琼[1,2], 刘德喜[1,2], 江腾蛟[1,2]

[1](江西财经大学 信息管理学院,江西 南昌 330013)

[2](数据与知识工程江西省高校重点实验室(江西财经大学),江西 南昌 330013)

通讯作者: 万常选, E-mail: wanchangxuan@263.net

摘 要: 在 Web 数据集成的过程中,如何从大量的 Web 数据源集合中选择合适数量的数据源,使得在满足特定查询需求的前提下尽可能地减少所需访问的数据源数量,同时保持返回数据结果的高质量,成为 Web 数据集成中的一个热点问题.以近十几年的研究实践为背景,介绍 Web 数据源选择的研究沿革及现状,并对 Web 数据源选择方法进行了归类.分别讨论了基于相关性的和基于质量的数据源选择的研究动机、研究方法和研究成果等,并对相关研究的目标、关键技术、优点和缺点进行了对比分析;最后展望了 Web 数据源选择未来的研究方向.

关键词: 数据集成;Web 数据源;文本;结构化与半结构化;源摘要

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 万常选,邓松,刘喜平,廖国琼,刘德喜,江腾蛟.Web 数据源选择技术.软件学报,2013,24(4):781−797. http://www.jos.org.cn/1000-9825/4374.htm

英文引用格式: Wan CX, Deng S, Liu XP, Liao GQ, Liu DX, Jiang TJ. Web data source selection technologies. Ruanjian Xuebao/Journal of Software, 2013,24(4):781−797 (in Chinese). http://www.jos.org.cn/1000-9825/4374.htm

Web Data Source Selection Technologies

WAN Chang-Xuan[1,2], DENG Song[1,2], LIU Xi-Ping[1,2], LIAO Guo-Qiong[1,2], LIU De-Xi[1,2], JIANG Teng-Jiao[1,2]

[1](School of Information and Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[2](Jiangxi Key Laboratory of Data and Knowledge Engineering (Jiangxi University of Finance and Economics), Nanchang 330013, China)

Corresponding author: WAN Chang-Xuan, E-mail: wanchangxuan@263.net

Abstract: In Web data integration, selecting data from a Web data source collection such that the specific query intents are satisfied while the number of accesses to data sources is minimized and the quality of returned results are guaranteed is a popular topic. In this paper, using the researches and practices in recent ten years as the background, the study focuses on the evolution and presents research in the area of Web data source selection and classifies Web data source selection methods. In addition, the paper discusses the research motivations, methods and results of relevance-based data source selection and quality-based data source selection. Moreover, the paper introduces the correlation research results and analyzes their destinations, key techniques, merits and demerits. Finally, some directions for future research are put forward.

Key words: data integration; Web data source; text; structured and semi-structured; source summary

Web 已成为一个拥有海量数据的信息源,许多应用领域迫切需要利用 Web 数据进行相关分析与挖掘,从中获取有用的知识.但是,Web 数据源具有自治、数据动态变化和数据不规范等特点,这使得有效利用 Web 上的信息成为一件十分具有挑战性的工作.为了有效利用 Web 中蕴藏的丰富且有价值的信息,需要有效地进行 Web 数

* 基金项目: 国家自然科学基金(61173146); 江西省高等学校科技落地计划(产学研合作)(KJLD12022); 江西省教育厅科技项目(GJJ12733, GJJ12732, GJJ11729)

收稿时间: 2012-09-06; 修改时间: 2012-12-03; 定稿时间: 2013-01-25

据集成.但是,由于 Web 规模巨大,如何有效地提高集成效率成为 Web 数据集成领域中的一个重要研究课题.

每个领域中都存在着大量的可供访问的 Web 数据源,每个数据源的查询接口也不尽相同.为了能够同时访问多个 Web 数据源,Web 数据集成系统必须对查询接口进行集成.当有了统一的访问接口后,如果只是把集成接口上的用户提交查询简单地转换成一个领域的每个 Web 数据源上的查询,显然是不可行的.因为这样操作存在以下问题: (1) 查询花费的代价太高;(2) 不是 Web 上每个数据源都能提供高质量的查询结果;(3) 由于 Web 数据源返回结果之间存在大量冗余,查询的数据源数量越多,冗余度也会越大.

基于以上原因,Web 数据源选择成为 Web 数据集成中的一个关键问题.把查询提交给很少量的数据源,但又要求返回的结果能够很好地满足用户的特定需求,是数据源选择的理想目标.针对不同的用户集成需求,Web 数据源选择方法各异.由于 Web 数据集成系统需提供与查询相关且高质量的检索结果给用户,因此研究人员主要依据数据源与查询的相关性以及数据源本身质量来进行 Web 数据源选择的相关研究.

Web 数据源主要可分为文本数据源和结构化与半结构化数据源两种类型.文本数据源通常可以被看作是一个由许多网页构成的“文件集”.结构化与半结构化数据源存储的是由多属性组成的现实世界的实体,其中,半结构化数据源存储的主要是 XML 数据.目前,基于数据源与查询相关性进行数据源选择的研究成果主要是针对以上两类数据源,前者的主要思路是把成熟的信息检索技术引入到文本数据源的选择过程中,后者的主要思路是通过挖掘蕴含在数据源中的结构化特征信息对数据源进行评价.

为了准确地选择出与用户查询最相关的相应数据源,就需要构建一个能够准确表征数据源内容的摘要.Web 数据源按合作关系可分为合作型与非合作型两种类型,前者可以自动提供其内容给用户而后者不会.非合作型数据源选择方法首先需要利用抽样、查询日志分析、接口分析等技术构建一个基于查询意图的数据源摘要,然后采用相关方法度量摘要与用户需求的距离.目前,大多数非合作数据源选择方案中用到的抽样技术均为随机抽样或主题抽样,少数文献在数据源选择过程中对抽样作了相应的改进[1,2].度量摘要与用户需求的距离,目前主要是采用信息检索中的常用距离判定方法.通常情况下,Web 数据源均为非合作型,因此合作型数据源选择方法局限性较大,其主要关注点在于如何有效地利用已知数据源信息构建精准摘要,该类方法主要为以后的 Web 数据源选择的研究起到相应借鉴的作用.

被选数据源质量较高就能够尽可能地降低 Web 数据集成中后续集成工作的难度,因此,尽管基于查询与数据源相关性进行源选择的研究较多,但是也有部分研究依据数据源本身质量选择相应数据源.这类工作不考虑具体查询词与数据源类型特点,主要思路是把源选择问题转换成多属性决策问题.

本文对现有的多种典型的 Web 数据源选择方法进行了总结和分析.按图 1 的分类结构,分别于第 1 节~第 3 节介绍基于数据源相关性的文本数据源选择方法、基于数据源相关性的结构化与半结构化数据源选择方法、基于数据源质量的数据源选择方法,并按照合作环境的不同以及方法特点进行分析与归纳.第 4 节对各种主要的 Web 数据源选择方法进行对比分析.第 5 节总结全文并对未来工作进行展望.

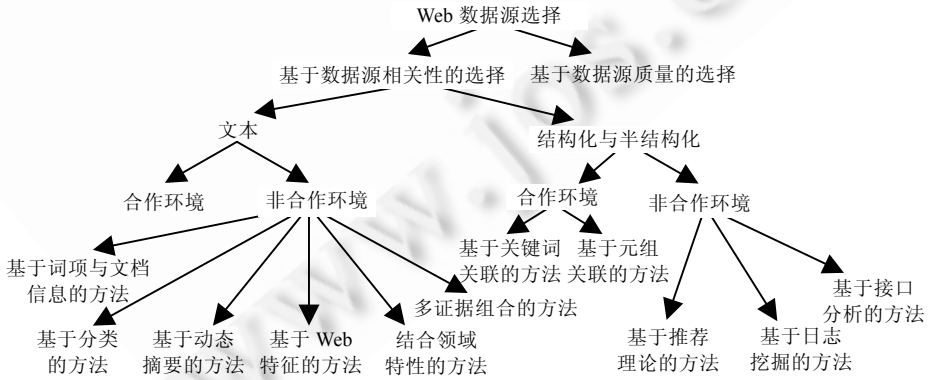


Fig.1 Classification of data source selection methods

图 1 Web 数据源选择方法分类

1 基于数据源相关性的文本数据源选择

在过去的十几年间,出现了较多的依据数据源与用户查询相关性进行文本数据源选择的高效方法,较好地满足了用户不同的查询需求.在进行以上数据源选择的时候,利用各种信息丰富数据源摘要使其满足用户的查询需求是关键.由于Web数据源通常是非合作的,因此已有的大多数数据源选择方法是基于非合作环境下的,只有很少量的数据源选择方法是基于合作环境的.不同方法适应于用户不同的需求,下面我们针对各种方法进行详细的分析与讨论.

1.1 非合作环境下基于词项与文档信息的选择方法

在早期的Web数据源选择的研究中,往往把数据源看成是一个大词袋或是一个大文档集,因此往往通过词项或文档摘要表征数据源内容.目前,已有的非合作环境下基于词项与文档信息的数据源选择方法可以分为以下3类:

1) 基于词项与逆文档集频率的选择方法

文献[3]依据传统信息检索中 *td-idf* 文档权重计算思想,把每个数据源当成一个大文档集,使用文档频率 *df* 代替词频,在贝叶斯推理网络理论的基础上提出了CORI数据源选择算法.CORI依据抽样词项频率以及词项的逆文档集频率组成的数据源摘要,计算每个数据源相对于特定查询的相似度得分.对于查询 q,数据源 c_i 得分记为 p,$p=b+(1-b)TI$,其中,

$$I=\frac{\log|c|+0.5}{|c|+1},T=\frac{df_t}{df_t+tf\_base+tf\_factor\frac{cw}{avcw}} \tag{1}$$

其中,df_t 是 c_i 中包含查询词的文档数量,$|c|$为数据源总数,cw 是 c_i 中词项数量,$avcw$ 是所有数据源的 cw 均值.公式(1)中其他的词是常量:b=0.4,tf_base=50,tf_factor=150.

如果使用高文档频率词进行查询抽样建立数据源摘要,CORI 选择效果会更佳[4].在数据源选择时,如果两个数据源的内容相差不大,数据源所采用的检索算法的效率就应该得到考虑.基于此,文献[5]考虑以上因素对CORI进行了相应的改进,取得了一定的效果.CORI经常被研究者用于进行比较,但D' Souza[6]通过大量的实验证明得到以下结果:(1) CORI 使用的相似度计算公式中的常量参数对数据集合比较敏感,因此针对不同数据源集合,如果均使用文中的标准参数值,CORI选择效果差别非常大;(2) 当候选数据源集合中有较多包含大量数据的数据源时,CORI基本不会把大数据源选入TOP-10数据源集合中.

2) 基于TOP-N文档数量的选择方法

为了解决上文中提到的CORI的显著缺陷,文献[7]在CORI基础上提出了一种结合数据源大小与抽样文档信息的ReDDE(相关文档分布评价)数据源选择算法.ReDDE 依据数据源所能提供的最相关文档数量评价数据源,在数据源 C_j 中与查询 q 相关文档数量用以下公式计算:

$$\hat{Rel}\_q(j)=\sum_{d_i\in C_{j\_samp}}p(rel|d_i)\times\frac{1}{N_{C_{j\_samp}}}\times\hat{N}_{C_j} \tag{2}$$

其中,$N_{C_{j\_samp}}$ 表示抽样文档数量,$\hat{N}_{C_j}$ 为依据文中概率公式估算的数据源总文档数量,$p(rel|d_i)$表示依据数据源 C_j 中返回结果在所有数据源返回的TOP-N结果中所占比例情况获取.

ReDDE的缺点在于比较偏爱大数据源,但在某些测试数据源集合中,如果出现包含少量的相关文档的大数据源,ReDDE也可以做出较为准确的选择.然而,ReDDE仍然没有解决CORI参数选取的难题.

3) 基于集合文档排序信息的选择方法

用户检索的目标可能是高召回率也可能是高准确率.对此,文献[8]提出采用统一数据源使用集成框架(UUM)来解决以上问题.该框架针对不同应用场合下的数据源选择采用不同的使用函数获取最优的结果.UMM框架中数据源选择算法的思路是:把所有抽样获取的文档组成一个文档集,同时计算数据源抽样压缩比,当比值小于100时,采用线性分段内插模型评价数据源中得分最高文档在抽样文档集合中的归一化的排序得分;反之,

使用 Logistic 模型.UMM 针对根据源提供者信息和发布日期人工构造的数据源测试集,返回结果文档的准确率比 ReDDE 和 CORI 略优,对于由数据源字母顺序构造的测试数据源集,UMM 比 ReDDE 和 CORI 算法在准确性上有较大的提升,超过 15%以上.UMM 需要通过人工训练的方式建立 Logistic 模型,因而在 Broker(集成接口)检索环境下是一种不可行的方案.

ReDDE,CORI 等数据源选择方法在面对多种测试集时性能差异很大.为了提升数据源选择算法的健壮性,文献[9]提出 CRCS(集中排序的数据源选择)方法.CRCS 方法的核心是利用集中取样文档的排序信息计算数据源的得分,该文采用了一种非线性的文档排名与文档得分的转换方法,具体转换公式如下所示:

$$R(D_j)=\alpha\exp(-\beta\times j) \tag{3}$$

其中,j 是数据源 S_i 中文档 D 在查询结果文档集中的排名.α,β是两个常量,取值范围在 1.2~2.8 之间,是实验的经验值.当获取了每个文档的转换得分之后,利用以下公式计算每个数据源的得分:

$$C_i=\frac{CH_i}{CH_{\max}\times|S_i|}\times\sum_{D\in S_i}R(D_j) \tag{4}$$

其中,CH_i 为数据源 S_i 的抽样估计数据量,$CH_{\max}$ 为测试集中最大数据源的数据量,$|S_i|$为数据源 S_i 摘要中抽样文档的数量.

CRCS 健壮性较好,面对多种测试集均有较为稳定的表现.CRCS 选择模型中用到了α,β两个参数,面对不同主题如何选择最合适的参数值,文中并没有给出科学的方法.

鉴于以上两种数据源选择算法中存在的问题,Thomas 在 UUM 算法的基础上提出了 SUSHI 内插值数据源选择法[10],该算法不需要进行数据训练工作.SUSHI 算法的主要思路是:抽样文档中每个单独文档都代表了数据源中一定数量的文档,因此可以按照实际抽样比例对抽样文档进行重新排序,再对每个数据源的抽样文档重新排序后的结果进行线性或 Logarithmic 拟合,获取每个数据源得分的内插值,然后依据内插值得分进行数据源选取.SUSHI 算法还存在以下问题有待改进:(1) 采用的曲线拟合方法较为粗糙,将来可以采用复杂拟合算法;(2) SUSHI 只用一条曲线拟合所有抽样文档得分,如果可以用两条曲线分别拟合相关与不相关文档得分,将可以进一步优化算法.

在实际的分布式环境中,存储在一个数据源中的数据往往有特定的主题或拥有者,因此,应该根据这个特点构造“组织化的数据源”作为测试数据源.文献[11]的实验结果表明,基于文档摘要的数据源选择方法优于基于词典的选择法.尽管如此,基于文档的数据源选择算法同样没有考虑 Web 数据源在万维网中所表现出来的特性,认为 Web 数据源仅由一个文档集合和一个文档检索引擎构成,因而具有较大的局限性.

1.2 非合作环境下基于分类的选择方法

在数据源选择过程中,通常通过提交若干抽样查询词,建立相应的源摘要.但是,当使用抽样方法建立数据源摘要时,若某个数据源的数据量较大,则通过抽样建立的摘要往往会丢失许多低频词信息,因此,建立的词典摘要往往是不完全的[12].不完全的数据源摘要将给数据源选择带来负面影响,尤其是当一个短查询中就含有低频词的时候.针对这个问题,早期的解决方法是进行查询词扩展,但该方法需要较多的人工交互操作[13].由于相同主题的数据源倾向于拥有相似的内容摘要,因此,这些数据源的内容摘要可以相互补充[1].

例如,数据源 CANCER.gov 摘要中没有包含 metastasis 这个词,而数据源 CancerBACUP 中该词出现次数为 3 569 次,由于 CANCER.gov 和 CancerBACUP 都是在同一个 Cancer 子类中,因此,并不是 CANCER.gov 不包含该词,而是在探测过程中丢失了这一内容.基于以上思想,Ipeirotis 提出了一种基于分层分类的数据源选择方法.该方法的思路是:首先对数据源依据主题进行层次分类[14,15],之后利用同一子类下数据源摘要互补的特性构建较为完整的子类摘要,然后选择出与查询最相关的子类[16],最后,在子类中获取与查询最相关的 K 个数据源.子类中获取 TOP-K 数据源的方法如下:(1) 如果最相关子类下数据源个数多于 TOP-K 数据源个数,则下层得分最高子类中 TOP-K 数据源均被选中,当被选中的数据源个数少于 K 时,则继续在同层采用平面选择策略选择出剩余的数据源;(2) 如果最相关子类下数据源个数少余 TOP-K 数据源个数,则在该层继续采用平面选择策略选择出剩余数据源.

为了进一步提高摘要的完整性,Ipeirotis 提出了 Shrinkage 方法及其使用策略[17].Shrinkage 策略的主要思想是:数据源进行层次分类之后,可以依据子类摘要与父类摘要以及子类与子类摘要之间的相关性信息修正数据源摘要中词项与词频数据的准确性,如图 2 所示.

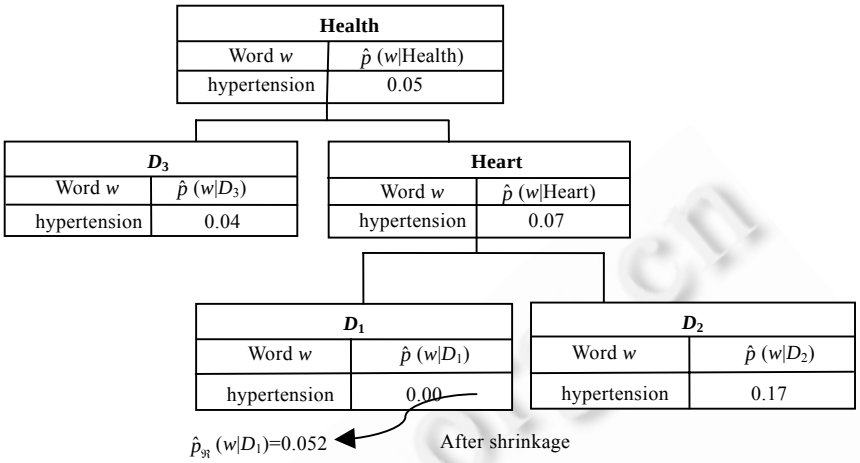


Fig.2 Shrinkage strategies in the data source selection

图 2 数据源选择中的 Shrinkage 策略

从图 2 可以得知,D_2 中"hypertension"是一个高频词,而 D_1 中该词没有出现.然而,D_1,D_2 均属于同一个子主题 Heart,因此可用统计理论从 $\hat{p}(\text{hypertension} \mid D_2)$ 推导出 $\hat{p}_{sh}(\text{hypertension} \mid D_1)$.

Shrinkage 策略不是在所有的情况下都有效,它只适合在数据源相对某查询得分较为不确定的情况下才使用.如果由抽样文档建立的摘要已包含了数据源中较高比例的文档,在这种情况下摘要已经足够完整,那就不应该使用 Shrinkage 策略.另外,当查询语句中每个查询词在一个摘要中出现的概率接近于在所有抽样文档中出现的概率,或者当每个查询词出现的概率与总抽样文档中出现的概率都相差很大时,使用 Shrinkage 策略只能获取非常有限的增益甚至起反作用.因此,文献[17]依据查询词得分方差自动判定在特定情况下是否采用 Shrinkage 策略,以提高源选择算法的有效性.

已有的数据源选择方法只考虑了自身的特征,并没有考虑数据源之间的关联,针对抽样摘要不完整的问题,文献[18]也提出了一种基于联合概率分类模型的数据源选择方法.该方法首先为每个数据源建立一个多特征的 Logistic 模型,然后利用查询与数据源 Logistic 模型的关联进行数据源排序,最后利用联合分类模型重排 TOP-10 个与查询相关的数据源.

基于分类的数据源选择方法,由于摘要的完整性有较大提高,选择的准确性也有较大的提高,是目前文本数据源选择方法中性能较为优良的,但该方法的前提是对数据源要先进行分类聚集,因而前期计算工作量较大.

1.3 非合作环境下基于动态摘要的选择方法

由于数据源中的数据内容随时间的改变而改变,因此,数据源摘要也应该是动态变化的;其次,用户对检索结果的准确性需求也是不一致的,如果源摘要是固定的,那么准确性也就是固定的.为了解决以上问题,就必须采用动态数据源选择法,目前,动态源选择方法可分为以下几类:

(1) 实时摘要法.基本思路就是[19,20]:首先把查询词提交给中间集成接口,集成接口把查询提交给每个数据源,获取每个数据源返回的排名位于前五的文档,然后计算每个文档的得分,包含得分最高的几个文档的数据源被选择;

(2) 动态学习法.基本思路是[21]:当用户查询时,从数据源中选择子集进行查询,然后计算结果网页与查询词的相似度,采用加权均值法调整该数据源的相似度,随着系统的运行,数据源与查询词的相似度不断被动态调整

以反映数据源的实际情况,从而为数据源选择做出判断依据;

(3) 动态探测法.当数据源摘要被预先建立之后,如果用户愿意花更多的时间等待更为准确的 TOP-*K* 返回结果,那么传统的数据源选择方法就无能为力了.基于此,Liu 等人[22]提出了基于动态探测的 DPro 数据源选择法.DPro 方法的思路是:首先依据抽样建立数据源 PRD(概率相关性分布模型),然后依据数据源操作独立性假设选择出期望准确度最高的*K*个数据源子集,如果该准确度高于用户给定值,则算法停止;反之,探索更多数据源重新计算相应的期望值.为了减少算法运算量,DPro 采用了贪心探测算法,该方法如用在 bGLOSS[23]等数据源选择算法上,可以显著地提高返回结果的准确性;

(4) 查询选择法.Cetintas[24,25]认为,经典的数据源选择方法忽视了一些有用的动态信息,例如用户查询的价值,因为对于一个实际的搜索引擎来说,已有的用户查询为当前的数据源选择提供了非常准确的信息.基于此,作者提出了一种新的数据源选择方法.该方法的核心思想是:利用已有的查询和数据源的相似度以及已有查询与特定用户查询的相似度计算数据源的得分.基于用户查询的选择方法,其性能比 ReDDE 等经典数据源选择算法有较大提升,但是依赖于用户的参与和长时间的查询信息的积累.类似的研究还有 Puppin[26]提出的基于查询与文档聚类的数据源选择方法,该方法首先通过元搜索接口收集用户查询,然后把收集的查询提交给每个数据源获取每个查询的 TOP-*N* 结果文档进行聚类,依据文档聚类结果得到相应的查询聚类,最后,采用 TF-IDF 技术计算特定查询与查询聚类的相关性以进行数据源选择.以上方法在仅仅选择一个相关数据源时,准确度比 CORI 提高了 11%~15%.

采用实时摘要法进行数据源选择,其效果优于 CORI,但需要更多的流量用于下载前 *N* 篇文档,因此反应时间明显增加.动态学习法比较倾向于选择出信息量大的数据源,因其相似度变化较快、选择机会较多.该算法的缺点是随着系统的运行,一些数据源的相关度会逐渐降低,被选择的机会越来越小[21].动态探测法可以较好地满足用户不同查询准确度的需求,但是如果采用抽样算法获取每个数据源的 PRD,计算量太大.为了尽可能地减少运算量,文献[22]提出的 PRD 获取方法是建立在绝对误差独立与相对误差独立的假设基础上的.然而,上述的两个假设并不是在大多数情况下都成立,因而,如何依据少量抽样获取准确的 PRD 是一个值得研究的方向.

1.4 非合作环境下基于Web特征的选择方法

为了降低实验难度,大多数的 Web 数据源选择方法利用人工数据源代替真实的 Web 数据源.随机抽取或按年表抽取 TREC 测试集中的文档形成一个文档集,再结合一个文档查询引擎,就可以形成一个“人工 Web 数据源”.真实的 Web 数据源不仅包含着大量的页面文档信息,而且具有丰富的网络特性.因此,进行 Web 数据源选择时应充分挖掘 Web 数据源的网络特性.

页面之间的链接信息可以反映出每个页面的重要程度,基于以上思想,文献[27]依据文档与查询的归一化相关度得分以及归一化的文档重要性得分建立数据源评价模型,取得了较好的效果.在实际 Web 环境下,部分数据源提供搜索接口,而部分没有搜索接口,所以,仅采用元搜索方法进行数据源的选择是不现实的.基于以上原因,文献[28]提出了一种混合式的源选择算法,算法的主要思想是:(1) 对于有查询接口的数据源采用通用的源选择算法,如 CORI,ReDDE;(2) 对于没有接口的数据源利用链接锚文本信息替代文档信息,根据替代文档排名进行 Web 数据源的选择.

结合数据源的 Web 特征提出的数据源选择方法,比基于“人工数据源”的选择方法的适用性更强,但还存在以下问题:(1) 在真实网络环境下,网页爬虫可能遭遇网络机器人的排斥,因此很多网页不能被索引;(2) 数据源索引的维护以及更新工作量巨大.如何破解以上瓶颈还有待于进一步的研究.

1.5 非合作环境下结合领域特性的选择方法

大部分数据源选择方法不关注领域特性,然而许多领域下的数据源都有其相应的特点,如何抓住领域特点,提升数据源选择的效率,是一个值得深入研究的方向.目前,研究人员已对专利领域、博客领域以及 P2P 领域的数据源选择展开了相关研究[29−31],表 1 总结了各种结合领域特性的数据源选择方法的相关技术、采用的评价标准以及相应的实验结论.

Table 1 Methods comparison of field-based data source selection

表 1 基于领域的数据源选择方法比较

领域	领域特点	源选择技术	评价标准	结论
专利	每个 PTO 有具体的主题,但文档在集合中比较分散	采用 KL,CORI 已有技术	TOP-*N* 数据源中相关文档数量	依据主题构成的专利数据源集合,KL 效果略优于 CORI,但对于没有组织的数据源集合,准确性低于 CORI
博客	一个博客存在多个主题	基于聚类的分布式检索技术提出 Pseudo-Cluster 选择方法	NDCG, MAP, P@10	Pseudo-Cluster 结合“话题相似度”与“话题明晰度”两个惩罚因子,优于 UMM 算法
P2P	数据源数据相互依赖,查询结果的重叠度较高	CORI+Overlap	召回率	相同召回率下,数据源选择数量相对于 CORI 显著减少

以上结合领域特性的数据源选择方法主要是依据领域数据源中存储数据的特点进行数据源的选择.在不同的领域环境下,用户的查询需求各异,因此,未来的研究可以进一步结合特定领域下用户的查询特点进行数据源的选择,以提升用户的满意度.

1.6 非合作环境下多证据组合的选择方法

在 Web 数据源的相关研究中,每个数据源选择算法选用的证据都有其特定意图.为了能够较为客观地选择出集成所需的数据源,研究者通常采用多个证据来度量数据源的有效性.

文献[32]较早地提出了一种多证据组合式的数据源选择算法,认为在进行数据源选择的过程中,除了需要考虑检索相关性之外,还要考虑检索期望质量和文档数、检索过程花费以及文档交付费用.其作者把源选择问题转换成了计算最优化方案的问题,提出了相应的 DTF 数据源选择方法.DTF 算法在大部分情况下其性能优于 CORI[33].文献[34]把 DTF 与 CORI 相结合,使得当用户提交长查询的时候,基于 TOP-*N* 的集成检索结果的准确性较 CORI 和 DTF 有所改进.但当用户提交短查询时,其性能较 CORI 和 DTF 有所下降.文献[35−37]对 DTF 进行了相应的改进,考虑了更多的数据类型,如姓名、年龄、图像等,针对姓名、年龄数据采用基于布尔谓词的相似性判别方法,针对图像数据提出了基于直方图的相似度评价模型.此外,Arguello[38]也提出了相应的多证据组合式的数据源选择算法,该算法使用的证据分为以下 3 类:(1) 集合文档特征信息;(2) 具体查询的主题特征;(3) 查询点击记录信息.通过这些证据的组合,把数据源选择问题转换成机器分类与学习的问题.

利用多证据组合法进行数据源选择,在一定程度上改善了仅依据查询相关性进行数据源选择带来的偏颇.以上文献中提出的数据源选择方法还存在以下问题:(1) 算法复杂度显著增加,如何平衡算法的效果与计算效率是值得进一步研究的问题;(2) 证据的适用性不强,每个用户的集成需求是不同的,因而应依据不同用户的需求进行证据的选取.

1.7 合作环境下的选择方法

在数据源选择的初期研究阶段,部分研究者通常假设通过采用 STARTS[39]协议使得每个数据源可以直接提供自己的内容摘要.Gravano 基于该思想,先后提出了 3 种数据源选择方案[40]:第 1 种是适用于布尔查询检索系统的 bGIOSS 数据源选择方法;第 2 种是适用于向量空间检索系统的 vGIOSS[41]数据源选择算法;最后一种是更适合于分布式环境的 hGIOSS 算法.基于布尔查询检索模型构建 bGIOSS,其前提是假设查询词满足独立分布的条件,依据独立分布公式估算相关文档个数,以此进行数据源的评价.vGIOSS 选择算法依据 f_i(数据源中包含特定词的文档数量)和 W_i(数据源所有文档中特定词的总权重)选择有价值的数据源.考虑到数据源文档中词项重要性的方法还有 Yuwono[42]提出的基于 CVV(线索有效性方差)的数据源选择算法.文献[42]认为,若一个词项对于数据源区分度越大,则其重要性也就越高.CVV 方法的局限性较强,当查询词为小数量且仅出现在少数数据源中的生僻词时,该方法完全不适用[43].hGIOSS 可以作为 vGIOSS 的上层部件,用户进行查询时,首先提交查询词给 hGIOSS,hGIOSS 把服务器上的源摘要作为文档处理,从而对数据源进行排序.在 Gravano 提出以上技术之后,随后出现的 CORI 数据源选择方法[6]的准确度较其提高了 10%以上,且性能较为稳定[44].

复杂的查询往往可以获得更高的准确性和召回率,所以很多性能优异的检索系统都采用了查询扩展技术.

检索系统处理复杂查询的效率远低于简单查询,为了有效提升基于复杂查询的数据源选择效率,文献[45]提出了一种基于轻量查询(简短且处理效率较高的查询)探测的数据源选择方法.如果以上方法能够被广泛地应用于集成检索系统中,有利于形成一个标准的数据源访问协议,可以进一步扩展 STARTS 协议.

由于在现实环境中数据源是异构的,而且一个数据源可能包含多个主题.如果不分主题地去匹配数据源中的词项,则将导致数据源选择的失败.例如,一个数据源中包含水果苹果的文章和苹果电脑的文章,如果仅去匹配苹果两词,显然是不对的.基于此,文献[46]提出 KL 数据源选择算法,KL 是一种基于文档聚集和语言模型的分布式检索技术.KL 采用 *K*-Means 文档聚集算法用于按主题组织数据源集合,语言模型用于表征主题和有效地为一个查询选择相应的主题集.本方法的主要缺点在于不能动态地创建一个新的主题,且聚类算法复杂度较高.

在真实的 Web 环境下,数据源不会主动地向用户提供内容摘要.以上的研究关注于如何在已经得知数据源摘要的情况下,对摘要信息合理地利用,以帮助数据源的选择,该类数据源选择算法对以后的研究起到了相应的借鉴作用.此外,鉴于搜索引擎公司的强大号召力,相关企业在数据源选择的过程中,要求各数据源提供所需的各种信息,因此进一步提高数据源选择的准确性也是未来的一个研究方向.

2 基于数据源相关性的结构化与半结构化数据源选择

结构化 Web 数据源在 Web 数据源中所占比例较大,随着 Web 技术的日益成熟,结构化与半结构化 Web 数据源选择受到人们越来越多的关注.目前,该类数据源的选择研究还处于起步阶段.结构化与半结构化 Web 数据源存储的是由若干属性组成的现实世界的实体,因此可以充分利用其查询接口上各属性的特征值和结构化记录之间的关联性进行数据源的选择.当前,结构化与半结构化数据源选择主要围绕合作、非合作两种环境特点展开相关研究,我们针对两种环境下的数据源选择方法进行了归纳分类,并在下文进行分析与讨论.

通常情况下,结构化与半结构化的 Web 数据源大多为非合作数据源,目前主要有 3 种方法用于相关数据源的选择.

1) 基于推荐理论的方法

文献[47]认为,用户期待的集成结果要和用户意图相关就应是可信的和重要的.基于以上目标,提出了通过构建基于元组相似度的"数据源推荐图"进行数据源选择的方法.

图 3 所示为 3 个数据源的推荐图结构举例,每个节点代表一个数据源.每条边代表一个数据源对指向的另一个数据源的推荐度.例如,边($S_1 \rightarrow S_2$)代表 S_1 对 S_2 的推荐度.具体的推荐度值依据以下公式计算获取:

$$A_Q(S_1,S_2)=\sum_{q\in Q}\frac{A(R_{1q},R_{2q})}{|R_{2q}|} \tag{5}$$

其中,$A_Q(S_1,S_2)$为基于抽样查询集合 Q 获取的 S_1 对 S_2 的推荐度.Q 为抽样查询词集合,R_{1q},R_{2q} 分别为向 S_1,S_2 提交查询 q 后所得结果集合,$A(R_{1q},R_{2q})$表示结果集中相似度.

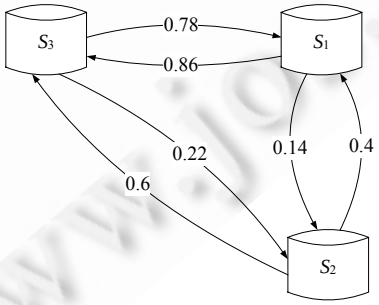


Fig.3 Recommended graph structure of the three data sources

图 3 3 个数据源的推荐图结构

为了克服抽样偏颇,进一步利用以下公式平滑总体推荐权重:

$$w(S_1 \to S_2) = \beta + (1-\beta) \times \frac{A_Q(S_1, S_2)}{|Q|} \tag{6}$$

其中,β为经验值,文中取为 0.1.在进行数据源选择之前,首先通过抽样计算每个候选数据源的推荐得分,然后选取得分最高的N个数据源,再采用Google Base搜索引擎向这些数据源提交查询词获取相应的查询结果.该方法相比于CORI等算法,在查询准确性上有较大提升.

2) 基于日志挖掘的方法

在数据源选择的过程中,用户反馈的信息有着非常重要的价值.Yu[48]较早地关注了结构化数据源的选择问题,提出了一种基于 TOP-N 查询的结构化数据源选择算法.该方法通过对用户提交的查询记录进行分析,建立Freq-Q(常用查询词集合)和 Infreq-Q(非常用查询词集合),然后依据 Freq-Q 与其对应的最匹配的元组记录,建立相应的数据源直方图摘要.算法实验结果表明,该查询算法是有效的.基于关键字的查询十分灵活,且查询词中隐含着丰富的结构化的模式信息.文献[49]提出了一种基于关键字的深网数据源选择方法,核心思路是:(1) 采用查询日志挖掘策略建立基于 KA(关键字-领域属性)关联的源相关性模型对候选领域进行排序;(2) 利用基于KA 关联和频率的采样建立数据源摘要.对结构化较强的领域,该方法具有较好的准确率;但对于结构化较弱的领域,该方法准确度低于传统的基于文档频率的源选择方法.

以上基于推荐理论和基于日志挖掘的两种方法均需要依靠用户查询词记录进行数据源摘要的构建,因此需要相关的搜索引擎公司提供相应的数据支持.

3) 基于接口分析的方法

满足不同需求的 Web 数据源往往有着不同的查询接口,结构化与半结构化数据源查询接口属性信息为数据源的选择提供了相应的依据.文献[50]采用本体来映射结构化数据源的查询表单属性,进行结构化数据源选择.类似的研究还有 Mihaila 提出的基于 SCQD(源内容与数据质量)模型的数据源选择方法[51].以上方法均依据结构化查询接口属性信息表征数据源与查询词的相关性,存在着一定的片面性,因为查询接口信息并不能准确地代表数据源中的数据内容.

文献[52]认为,在某些领域中,单一的数据源不能提供给用户满意的检索结果.例如,给定基因名称 ERCC6,如果用户需要找出这个基因上的 DNA 序列多态性以及比对这个基因与其他非人类哺乳动物同源基因的相似性,这就涉及到多种类型的数据库.由于以上查询中许多关键字之间的关系只有对多个相互依赖的 Web 数据源进行查询后才可以获取,因此,应该以一种合适的顺序查询多个数据源.该文针对以上问题,提出了基于多数据源相互依赖图的结构化数据源选择算法,该方法的核心在于给每个相关数据源建立 3 种类型的关联.文中利用$R(MI,OI,O,C)$表征一个深网数据源,其中,MI 为数据源接口的必填属性,OI 为选填属性,O 为输出属性,C 为限制条件.假设有两个深网数据源 R_1 和 R_2,则建立以下 3 种类型的依赖关联:(1) 当 $O_1 \cap MI_2 \neq \varnothing$时,$R_1$ 的查询输出可以用于填写 R_2 中必填的输入接口;(2) 当 $O_1 \cap OI_2 \neq \varnothing$时,$R_1$ 的查询输出可以用于填写 R_2 中可选的输入接口;(3) R_1 的可选输入属性同时也是输出属性的一部分,能够作为 R_2 的必填输入或者可选输入.当为数据源建立了以上关联之后,再采用基于领域本体的数据源排序算法进行数据源的选择.该方法可以有效地支持用户提交的实体-属性查询以及实体-实体查询.

深网数据库拥有者通常会对用户提交的查询类型、返回元组的数量设定一定的限制,因此,传统的随机抽样技术效果不佳,影响到相关数据源选择方法的准确性.鉴于结构化与半结构化深网接口属性的特点,文献[53]针对利用具有布尔、类别、数字类型属性单表存储数据的深网数据源的抽样技术进行了相关研究,提出了基于查询空间的随机游走数据源抽样技术.该方法首先以宽泛的查询开始随机下钻,然后迭代地重复下钻操作,同时增加随机被选的谓词,直至得到一个有效的查询.文献[53]中提出的抽样方法可以通过查询接口有效地获取数据源的均匀样本集,但却存在两个问题:(1) 随机下钻时遇到下溢节点会提前终止;(2) 所采用的拒绝抽样技术会丢失短随机下钻结果.文献[54]分别采用基于回溯的随机下钻技术以及加权抽样技术,较好地解决了文献[53]中存在的以上两个问题,从而可以获取一个代表性更强的数据源样本集作为数据源摘要,这将进一步提高数据源选择的效果.

结构化与半结构化数据源中的数据有着丰富的语义信息,如果能够构建一个准确表征数据源内容及语义的源摘要,将可以极大地提高数据源选择的准确性.由于此类研究刚刚兴起,针对半结构化数据源选择主要采用基于关键字关联的方法,针对结构化数据源选择则采用基于元组关联的方法,具体分析如下:

1) 基于关键字关联方法

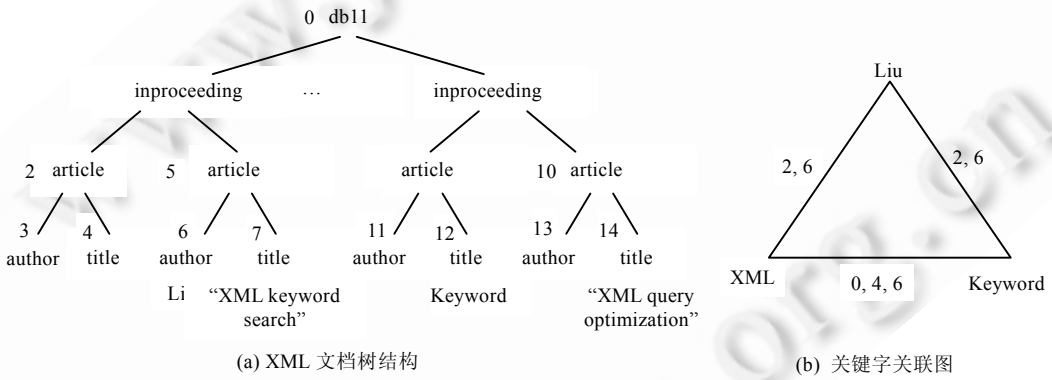
XML 已成为表示和交换信息的重要格式,选择对于查询有用的 XML 数据源,是建立高效的信息集成系统的关键问题.如果直接把 IR 中的通用技术用到 XML 数据源选择过程中是不合适的,因为 XML 文件树不仅具有丰富的结构化语义信息,而且关键字之间有明确的关联关系.如果不考虑以上 XML 的数据特性,进行查询的相关性判定显然是不合适的.例如,对于如图 4(a)所示的 XML 文档树,如果一个用户需要查询作者姓“Liu”且题目名为“XML keyword”的相关信息,返回节点 6、节点 7 组成的结果显然比返回节点 6、节点 14 和节点 12 组成的结果更有意义.基于以上思想,文献[55]提出了一种基于 K-Graph(关键字关联图)的数据源选择算法.该方法的核心便是计算包含在两个节点中的关键字的距离,具体公式如下所示:

$$score(k_i,k_j,n_i,n_j)=\frac{weight(n_i,k_i)+weight(n_j,k_j)}{dist(n_i,n_j)+1} \tag{7}$$

其中,k_i,k_j 为 n_i,n_j 两个节点中分别包含的关键字,$dist(n_i,n_j)$为两个节点间的路径长度.可以依据 XML 树建立相应的查询关键字关联图 K-Graph,图中每条边上的权重,代表包含两个关键字的节点的路径长度,两个关键字之间可能有多条路径,因而一条边上可以有多个权值.$weight(n_i,k_i)$代表了节点 n_i 中关键字 k_i 的重要性:

$$weight(n_i,k_i)=\log_2\frac{N}{N_{k_i}},$$

其中,N 是 XML 树中元素总数, N_{k_i} 为包含关键字 k_i 的 XML 元素的个数.基于图 4(a)所示的 XML 文件树,查询关键字关联图 K-Graph 如图 4(b)所示.



(a) XML 文档树结构 (b) 关键字关联图

Fig.4 Structured semantic information of XML documents

图 4 XML 文档的结构化语义信息

类似的 XML 数据源选择研究还有朱冠胜等人[56]提出的基于关键字频率与路径长度评分模型的数据源选择方法.

2) 基于元组关联的方法

在分布式检索系统中有效地选择数据源,一个通用的步骤是依据词项频率和逆集合频率(ICF)为关系数据库建立摘要.然而,直接采用上面的方法为结构化数据库建立摘要是不合适的,因为存在下面两个问题:(1) 关系数据库表中的数据是被规范化了的,所以关键字统计信息不能真实地度量在关系数据库中某个关键字的重要性;(2) 由于连接次数越多,证明该数据源与查询词的关联性越低,因此需要考虑为了获取出现全部关键字的查询结果而进行连接操作的次数.基于以上原因,Yu 等人[57]提出了一种基于关键字查询的结构化数据源选择的方法.该方法充分利用关键字之间的有用信息,通过 KRM(关键字关系矩阵)表征每一个结构化数据源中的摘要,依

据 KRM 得分对数据源进行排序.KRM 由数据源内容关系图与结构关系图构成,如图 5 所示.其中,t 表示数据源中的元组,k 为查询关键字.图 5(a)表示了关键字是否在相应元组中出现的情况,图 5(b)展示了元组之间是否存在关联,两个元组若有主外键关联,则它们联系的得分为 1,否则为 0.

$$D=(d_{i,j})_{m\times n}=\begin{Bmatrix} & t_1 & t_2 & \dots & t_n \\ k_1 & 1 & 0 & \dots & 0 \\ k_2 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ k_m & 1 & 0 & \dots & 0 \end{Bmatrix} \qquad T=(t_{i,j})_{n\times n}=\begin{Bmatrix} & t_1 & t_2 & \dots & t_n \\ t_1 & 0 & 1 & \dots & 1 \\ t_2 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ t_n & 1 & 0 & \dots & 0 \end{Bmatrix}$$

(a) 数据源内容关系图 (b) 数据源结构关系图

Fig.5 Keyword relationship matrix for KRM

图 5 关键字关系矩阵 KRM

如果能够从搜索引擎公司获取大量的用户查询数据用于相关分析与挖掘,基于日志挖掘的方法可以获得较好的选择效果.在不能获取以上数据的情况下,可以采用基于接口分析的方法;但当多个数据源接口信息相近时,该方法将具有较大的局限性.在真实的 Web 环境下,往往很难预先获取准确的数据源的索引以及数据源的关联信息.基于 XML 语义特征以及关键字矩阵的数据源选择算法均是建立在数据源提供合作的前提下的,如何结合最新的深网抽样技术建立一个包含丰富结构化信息的数据源摘要,将是一个富有挑战性的问题.

3 基于数据源质量的数据源选择

与用户查询相关的数据源质量参差不齐,一旦低质量的数据源被选入用于集成检索的候选数据源集合,会给 Web 集成后续操作,如数据转换、实体关联、数据融合等,带来较多的困难,因此,有必要研究相应的高质量数据源选择方法.

为了对数据源进行有效的质量评估,通常需要建立数据源质量评价模型 Q-mode=$\{D,W,Q,S\}$,其中,$D=\{d_1,d_2,\dots,d_m\}$代表一组数据源集合,$W=\{w_1,w_2,\dots,w_n\}$代表一组数据源维度集合,$Q=\{q_1,q_2,\dots,q_n\}$为维度得分集合,$S=\{s_1,s_2,\dots,s_m\}$为数据源得分集合.维度以及相应计算公式一般依据作者经验设置,维度权重则依据用户偏好设定或通过 SVM 等方法训练获取,依据加权后维度得分总和进行数据源排序.

质量是一个综合的概念,无法用单一的维度来进行度量,所以,数据源质量评价的首要任务是确定数据源质量的维度.Wang 等人在调研的基础上提出了一个数据质量的概念框架[58],其中包含了 15 个质量维度,后续研究选取的数据质量维度往往来自于这一框架.如,Naumann[59]选取了 3 个数据源质量评价指标:易理解性、数据长度和易使用性,然后采用 DEA(数据包络)方法直接识别出一组有效的数据源,因而不需要考虑评价指标权值选取的问题.该方法的缺点在于 DEA 的复杂度太高,仅适用于少量的数据源的选择.类似的研究还有余伟等人[2]提出的基于数据质量的 Deep Web 数据源选择方法,该方法首先提出了改进的分层抽样和雪球抽样方法用于抽样获取数据源质量评价数据,然后根据经验在 Wang 等人提出的 15 个质量维度中选择了 6 个适用于评价深网质量的维度建立源质量评价模型,同时,依据用户主观评价反馈获取相应的维度权重,最后,依据质量模型计算各数据源得分.

Aboulnaga 等人[60]在设计μBE 数据集成系统中提出了基于集成效用数据源选择方法.μBE 围绕 3 个方面评价数据源质量:数据源模式在受约束条件下相互匹配程度、数据源中数据特征(数据量、覆盖度、冗余度)以及数据源本身的特征(延时、可靠性、费用、权威性).μBE 通过迭代地解一系列的受限优化问题来找出适合集成的数据源,具体的受限优化公式如下所示:

$$\arg\max_{S\subseteq U}(Q(S))=\sum_{i=1}^{|F|}w_iF_i(S) \tag{8}$$

其中,U 是候选数据源集合;S 为待集成的相应数据源;$|S|\leqslant m$,m 为用户指定的数据源个数;$Q(S)$为数据源总得分;

$F_i(S)$为具体的数据源评价维度,例如数据源模式在受约束条件下的相互匹配程度、数据量等;$|F|$为维度数量;w_i为维度权重,值可以由用户依据偏好给定.每次迭代开始时,用户可以对该方案进行反馈,如指定新的约束用于基于聚类的模式匹配、为中间集成模式添加属性、为质量维度设置新的权重,甚至指定新的质量维度.如此重复,直到用户满意为止.类似的研究还有鲜学丰等人[61]提出的基于迭代的 Web 数据源选取和集成的方法,该方法的核心在于增益函数,即评价一个新数据源加入到集成系统中可能带来的增益.

以上考虑数据源质量的选择方法是不考虑主题特性的,依据经验选取统一质量维度,所以在不同主题下选择准确性较不稳定.基于此,文献[62]首先通过用户反馈建立推荐数据源集合与拒绝数据源集合,然后利用两个集合中数据源在各维度上得分平均值的相差度以及取值范围的重叠度选择核心质量维度,最后依据核心维度建立各主题数据源的质量评价模型.针对不同主题的数据源选择,该方法的准确性较高,且运算量较小.

以上数据源选择方法的目的在于选择出某个领域下的高质量的数据源,且不关注数据源的具体类型.然而用户往往不仅关心某个数据源在某领域下的质量排名,而且同样关注该数据源与查询的相关度及其所能提供的与其查询需求相关的数据的质量.Mihaila[51]在这方面进行了初步研究,利用查询接口属性信息表征源内容,在结构化与半结构化数据源的选择过程中,不仅考虑源与查询的相关性,还考虑数据源的完整性、新旧程度以及更新粒度等质量参数.结合数据源与用户查询的相关性,为每个用户选择个性化的质量维度建立数据源选择模型,是进一步值得研究的方向.

4 分析比较

数据源选择问题作为 Web 数据集成中的热点问题,得到较多学者的关注.本文总结并分析了目前主要的 Web 数据源选择技术,并对各种技术进行了分类,指出了各种技术的着眼点、优缺点、关键技术以及检索质量.表 2 列出了目前主要的 Web 数据源选择方法.从表 2 中可以看出:针对于文本数据源,采用抽样文档排序信息及分层聚类等方法可以取得较好的效果;针对于结构化与半结构化数据源的选择要获得高质量的检索结果,突破口在于结构化与半结构化语义信息的挖掘与使用.

Table 2 Comparison of methods for Web data source selection

表 2 各种 Web 数据源选择方法对比

方法名称	分类	着眼点	优点	缺点	关键技术	检索质量
CORI[3]	基于词项与抽样文档的“文档集合”的选择	非合作环境下开创性的方法	理论依据成熟	模型参数对数据集合较敏感	*td-idf* 相似度评价模型	较低
ReDDE[7]		数据源大小的影响	大数据源的准确性较高	偏爱大数据源	基于 CORI 的改进	中等
CRCS[9]		抽样文档排序信息的重要性	健壮性较好	参数设定没有科学方法	文档排名与得分转换	中等
SUSHI[10]		抽样文档的代表性	不需要数据训练,不存在参数设置问题	曲线拟合方法较粗糙	抽样文档重排及拟合	高
Shrinkage[17]	分层分类的源选择	相同主题的数据源拥有相似的内容摘要	摘要的完整性有较大提高	聚类工作量较大,算法复杂	概率统计	较高
JPCM[18]	联合概率分类	相似数据源对应相同的查询	摘要的完整性有较大提高	概率模型复杂,运算量较大	联合概率分布模型	高
Dpro[22]	动态源选择	用户指定 TOP-*N* 准确度	可以适合多种准确度需求	迭代工作量较大	概率相关性分布模型	高
SourceRank[47]	结构化源的选择	结果应可信与重要	考虑到信息的价值	未考虑结构语义	数据源推荐图	中
KA[49]		基于日志挖掘关键字与领域关联	结构化强的领域准确度高	结构化弱的领域准确度低	关键字-领域属性图	较高
KRM[57]		结构化元组的关键字检索	充分考虑元组之间的语义	需要数据源合作	KRM	高
K-Graph[55]		XML 关键字检索	考虑了 XML 元素之间的语义	需要数据源合作	K-Graph	高

针对不同类型数据的数据源选择,业界已提出很多不同的方法,这些方法可以在某些情况下进行相互借鉴和补充,以进一步提高用户集成检索的满意度.例如,基于分层分类以及联合概率分类的文本数据源选择技术可以较好地解决非合作型文本数据源摘要完整性的问题,非合作型结构化与半结构化数据源选择时同样需要解决以上问题.以上方法可以提供一些解决思路;动态文本数据源选择技术Dpro可以满足用户指定TOP-*N*检索准确度的需求且检索质量很高,而结构化与半结构化数据源选择中还没有提出相应技术,值得进一步加以研究;结构化与半结构化数据源选择涉及到多数据源合作提供集成检索结果的问题,该问题在文本数据源选择中同样具有现实意义,研究人员可以借鉴相应思路以及依据文本数据的特点展开相关研究.

5 总结与展望

随着Web数据源在网络上大量且不断地涌现,对Web数据源进行相关集成的研究已成为一个非常迫切的问题.为了降低为获取高质量数据而需要付出的对Web数据源的访问代价,数据源选择有着重要的应用价值.数据源选择的研究方兴未艾,研究人员已经在该领域开展了大量的研究工作,在Web数据源选择方法上也有一些阶段性研究成果.本文回顾与总结了近十几年来国际上在该领域的主要研究成果,综述了在Web数据集成的环境下数据源选择技术的研究现状,总结并对比分析了各种数据源选择技术的研究目标、研究方法、关键技术、优点和缺点等,指出了仍然存在的问题和将来可能的解决方法.

Web数据集成是一个新兴的研究领域,包含了很多需要解决的问题.总的来说,Web数据源选择技术的研究还是当前的研究热点,尤其是结构化与半结构化的深网数据源的选择研究目前还处于起步阶段,仍然有着大量的关键问题有待深入、细致的研究.基于本文的讨论,我们认为Web数据源选择领域主要还有如下有待研究的问题,希望对该领域的其他研究人员有所启发:

(1) 支持新的查询需求.近年来,随着微博以及各种交流空间等社区平台的快速崛起,使得人们在互联网上的行为变得日趋复杂,任务搜索应运而生,它将用户搜索需求转化为用户下达的搜索任务,以用户为中心,生成一个用户要完成的特定任务,然后给用户推送一些极具实用价值的数据信息.因此,Web数据源选择方法如何紧密结合下一代搜索技术与理论,建立起符合任务搜索需求的源摘要及源选择模型,将是一个前瞻性的问题;

(2) 当前,多媒体数据源数量日益增大,因此,多媒体数据源的选择同样值得研究.文献[63]依据多媒体数据源的特点首次提出了以图像、声音模版构建数据源摘要进行源选择的方法,其核心在于采用聚簇的方法构建层次化模版图以及为每个数据源构建与相应模版对应的概率分布图,该方法取得了较好的选择效果.由于近年来图像检索技术发展迅猛,已产生很多新的高效的检索方法,因此,如何结合新的技术,如基于图论的图像检索方法,进行数据源选择,是一个值得关注的研究方向;

(3) 挖掘不可穷举属性信息.针对于结构化与半结构化数据源的选择,现有的工作是在查询接口的数字属性和离散属性上进行特征概括,对源的选择起到一定促进作用,但还未从根本上解决问题,下一步研究工作要能够对非数字的不可穷举属性进行特征概括[64].结构化数据源中往往存在非结构化的内容数据,例如,亚马逊购书网站中图书记录尽管是结构化的,但是图书评论及内容摘要信息是文本化的,且对顾客的购买起到重要的导向作用.因此,如何挖掘结构化数据源中的这部分信息进行数据源的选择,是一个值得研究的课题;

(4) 基于抽样的结构化信息挖掘.结构化与半结构化数据源中往往包含着丰富的结构化信息,已有的研究成果是建立在数据源提供合作的前提下获取关键字关联规则及索引结构信息的,但在实际的Internet环境中,这个前提是很难存在的.因此,如何通过抽样获取Web数据源的结构化信息,将是一个需要解决的问题;

(5) 动态摘要的建立.Web数据源时常更新,内容不断丰富,以前未涉及到的主题在经过一段时间的更新后,数据源可能已经完成覆盖,以前内容不够丰富的话题也有可能变成某个数据源的热点话题.因此,有必要建立起动态的数据源摘要.如何增量更新抽样查询词及识别所需更新的摘要内容,是需要研究的内容.

References:

[1] Ipeirotis PG, Gravano L. Distributed search over the hidden Web: Hierarchical database sampling and selection. In: Bernstein PA, Ioannidis YE, Papadias RRD, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). San Francisco: Morgan Kaufmann Publishers, 2002. 394−405.

[2] Yu W, Li SJ, Wen LJ, Tian JW. Ranking of deep Web sources based on data quality. Journal of Chinese Computer Systems, 2010, 31(4):641−646 (in Chinese with English abstract).

[3] Callan JP, Lu ZH, Croft W. Searching distributed collections with inference networks. In: Fox EA, Ingwersen P, Fidel R, eds. Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'95). New York: ACM Press, 1995. 21−28. [doi: 10.1145/215206.215328]

[4] Callan J, Connell M. Query-Based sampling of text database. ACM Trans. on Information Systems (TOIS), 2001,19(2):97−130. [doi: 10.1145/382979.383040]

[5] Craswell N. Methods for distributed information retrieval [Ph.D. Thesis]. Canberra: The Australian Nation University, 2000.

[6] D'Souza D, Zobel J, Thom J. Is CORI effective for collection selection? An exploration of parameters, queries, and data. In: Bruza P, Moffat A, Turpin A, eds. Proc. of the 9th Australasian Document Computing Symp. Melbourne, 2004. 41−46. http://ww2.cs.mu.oz.au/~alistair/adcs2004/

[7] Si L, Callan J. Relevant document distribution estimation method for resource selection. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2003). New York: ACM Press, 2003. 298−305. [doi: 10.1145/860435.860490]

[8] Si L, Callan J. Unified utility maximization framework for resource selection. In: Grossman DA, Gravano L, Zhai CX, Herzog O, Evans DA, eds. Proc. of the 13th ACM Conf. on Information and Knowledge Management (CIKM 2004). Washington: ACM Press, 2004. 32−41. [doi: 10.1145/1031171.1031180]

[9] Milad S. Central-Rank-Based collection selection in uncooperative distributed information retrieval. In: Amati G, Carpineto C, Romano G, eds. Proc. of the 29th European Conf. on IR Research. Heidelberg: Springer-Verlag, 2007. 160−172. [doi: 10.1007/978-3-540-71496-5_17]

[10] Thomas P, Shokouhi M. SUSHI: Scoring scaled samples for server selection. In: Allan J, Aslam JA, Sanderson M, Zhai CX, Zobel J, eds. Proc. of the 32nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2009). New York: ACM Press, 2009. 419−426. [doi: 10.1145/1571941.1572014]

[11] D'Souza D, Thom JA, Zobel J. Collection selection for managed distributed document databases. Information Processing and Management, 2004,40(3):527−546. [doi: 10.1016/S0306-4573(03)00008-6]

[12] Huang SM, Yen DC, Yang LW, Hua JS. An investigation of Zipf's law for fraud detection. Decision Support Systems, 2008,46(1): 70−83. [doi: 10.1016/j.dss.2008.05.003]

[13] French JC, Powell AL, Gey F, Perelman N. Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. In: Paques H, Liu L, Grossman D, eds. Proc. of the 10th Conf. on Information and Knowledge Management (CIKM 2001). New York: ACM Press, 2001. 199−206. [doi: 10.1145/502585.502619]

[14] Gravano L, Ipeirotis PG, Sahami M. QProber: A system for automatic classification of hidden-Web databases. ACM Trans. on Information Systems (TOIS), 2003,21(1):1−41. [doi: 10.1145/635484.635485]

[15] Ipeirotis PG, Gravano L, Sahami M. Probe, count and classify: Categorizing hidden Web databases. In: Aref WG, ed. Proc. of the 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2001). New York: ACM Press, 2001. 21−24. [doi: 10.1145/375663.375671]

[16] Ipeirotis PG, Gravano L. Classification-Aware hidden-Web text database selection. ACM Trans. on Information Systems (TOIS), 2008,26(2):1−66. [doi: 10.1145/1344411.1344412]

[17] Ipeirotis PG, Gravano L. When one sample is not enough: Improving text database selection using shrinkage. In: Weikum G, König AC, Deβloch S, eds. Proc. of the 2004 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2004). New York: ACM Press, 2004. 767−778. [doi: 10.1145/1007568.1007655]

[18] Hong D, Si L, Bracke P, Witt M, Juchcinski T. A joint probabilistic classification model for resource selection. In: Crestani F, Marchand-Maillet S, Chen HH, Efthimiadis EN, Savoy J, eds. Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2010). New York: ACM Press, 2010. 98−105. [doi: 10.1145/1835449.1835468]

[19] Abbaci F, Savoy J, Beigbeder M. A methodology for collection selection in heterogeneous contexts. In: Proc. of the Int'l Conf. on Information Technology: Coding and Computing (ITCC 2002). Washington: IEEE Computer Society Press, 2002. 529−535. [doi: 10.1109/ITCC.2002.1000443]

[20] Rasolofo Y, Abbaci F, Savoy J. Approaches to collection selection and results merging for distributed information retrieval. In: Proc. of the 10th Conf. on Information and Knowledge Management (CIKM 2001). New York: ACM Press, 2001. 191−198. [doi: 10.1145/502585.502618]

[21] Duan QL, Yang RG, Hua SQ. Method for database selection of deep Web based on the dynamic learning algorithm. Journal of Zhengzhou University (Natural Science Edition), 2010,42(1):5−8 (in Chinese with English abstract).

[22] Liu VZ, Luo RC, Chu WW. Dpro: A probabilistic approach for hidden Web database selection using dynamic probing. In: Özsoyoglu ZM, Zdonik SB, eds. Proc. of the 20th Int'l Conf. on Data Engineering (ICDE 2004). Washington: IEEE Computer Society Press, 2004. 1−12.

[23] Gravano L, Garcia-Molina H, Tomasic A. The effectiveness of GIOSS for the text database discovery problem. In: Snodgrass RT, Winslett M, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD'94). New York: ACM Press, 1994. 126−137. [doi: 10.1145/191843.191869]

[24] Cetintas S, Si L, Yuan H. Learning from past queries for resource selection. In: Cheung DWL, Song IY, Chu WW, Hu XH, Lin JJ, eds. Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009). New York: ACM Press, 2009. 1867−1870. [doi: 10.1145/1645953.1646251]

[25] Cetintas S, Yuan H. Using past queries for resource selection in distributed information retrieval. Technical Report, 11-012, West Lafayette: Purdue University, 2011.

[26] Puppin D, Silvestri F, Laforenza D. Query-Driven document partitioning and collection selection. In: Li JZ, Lee WC, Silvestri F, eds. Proc. of the 1st Int'l Conf. on Scalable Information Systems (InfoScale 2006). New York: ACM Press, 2006. 34−41. [doi: 10.1145/1146847.1146881]

[27] Yu C, Meng W, Wu WS, Liu KL. Efficient and effective metasearch for text databases incorporating linkages among documents. In: Aref WG, ed. Proc. of the 2001 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2001). New York: ACM Press, 2001. 187−198. [doi: 10.1145/376284.375684]

[28] Hawking D, Thomas P. Server selection methods in hybrid portal search. In: Baeza-Yates RA, Ziviani N, Marchionini G, Moffat A, Tait J, eds. Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2005). New York: ACM Press, 2005. 75−82. [doi: 10.1145/1076034.1076050]

[29] Larkey LS, Connell ME, Canllan J. Collection selection and results merging with topically organized U.S. patents and TREC data. In: Proc. of the 9th Conf. on Information and Knowledge Management (CIKM 2000). New York: ACM Press, 2000. 282−289. [doi: 10.1145/354756.354830]

[30] Seo J, Croft WB. Blog site search using resource selection. In: Shanahan JG, Amer-Yahia S, Manolescu I, Zhang Y, Evans DA, Kolcz A, Choi KS, Chowdhury A, eds. Proc. of the 17th Conf. on Information and Knowledge Management (CIKM 2008). New York: ACM Press, 2008. 1053−1062. [doi: 10.1145/1458082.1458222]

[31] Bender M, Michel S, Triantafillou P, Weikum G, Zimmer C. Improving collection selection with overlap awareness in P2P search engines. In: Baeza-Yates RA, Ziviani N, Marchionini G, Moffat A, Tait J, eds. Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2005). New York: ACM Press, 2005. 15−19. [doi: 10.1145/1076034.1076049]

[32] Fuhr N. A decision-theoretic approach to database selection in networked IR. ACM Trans. on Information Systems, 1999,17(3): 229−249. [doi: 10.1145/314516.314517]

[33] Nottelmann H, Fuhr N. Evaluating different methods of estimating retrieval quality for resource selection. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval. New York: ACM Press, 2003. 290−297. [doi: 10.1145/860435.860489]

[34] Nottelmann H, Fuhr N. Combining CORI and the decision-theoretic approach for advanced resource selection. In: McDonald S, Tait J, eds. Proc. of the 26th European Conf. on IR Research. Heidelberg: Springer-Verlag, 2004. 138−153. [doi: 10.1007/978-3-540-24752-4_11]

[35] Nottelmann H, Fuhr N. Decision-Theoretic resource selection for different data types in MIND. In: Callan J, Crestani F, Sanderson M, eds. Proc. of the ACM SIGIR 2003 Workshop on Distributed Information Retrieval. New York: ACM Press, 2003. 43−57. [doi: 10.1007/978-3-540-24610-7_4]

[36] Callan J, Crestani F, Nottelmann H, Pala P, Shou XM. Resource selection and data fusion in multimedia distributed digital libraries. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval. New York: ACM Press, 2003. 363−364. [doi: 10.1145/860435.860502]

[37] Nottelmann H, Fuhr N. The MIND architecture for heterogeneous multimedia federated digital libraries. In: Callan J, Crestani F, Sanderson M, eds. Proc. of the ACM SIGIR 2003 Workshop on Distributed Information Retrieval. New York: ACM Press, 2003. 112−125. [doi: 10.1007/978-3-540-24610-7_9]

[38] Arguello J, Callan J, Diaz F. Classification-Based resource selection. In: Cheung DWL, Song IY, Chu WW, Hu XH, Lin JJ, eds. Proc. of the 18th Conf. on Information and Knowledge Management (CIKM 2009). New York: ACM Press, 2009. 1277−1286. [doi: 10.1145/1645953.1646115]

[39] Gravano L, Chang KCC, Garcia-Molina H, Paepcke A. STARTS: Stanford proposal for Internet meta-searching. In: Peckham J, ed. Proc. of the ACM Int'l Conf. on Management of Data (SIGMOD'97). New York: ACM Press, 1997. 207−218. [doi: 10.1145/253260.253299]

[40] Gravano L, Garcia-Molina H, Tomasic A. GIOSS: Text-Source discovery over the Internet. ACM Trans. on Database Systems, 1999,24(2):229−264. [doi: 10.1145/320248.320252]

[41] Gravano L, Garcia-Molina H. Generalizing GIOSS to vector-space databases and broker hierarchies. In: Dayal U, Gray PMD, Nishio S, eds. Proc. of the 21st Int'l Conf. on Very Large Databases (VLDB'95). San Francisco: Morgan Kaufmann Publishers, 1995. 78−89.

[42] Yuwono B, Lee DL. Server ranking for distributed text retrieval systems on the Internet. In: Topor R, Tanaka K, eds. Proc. of the 5th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'97). Singapore: World Scientific Press, 1997. 41−49.

[43] Craswell N, Bailey P, Hawking D. Server selection on the World Wide Web. In: Proc. of the 5th ACM Conf. on Digital Libraries (DL 2000). New York: ACM Press, 2000. 37−46. [doi: 10.1145/336597.336628]

[44] French JC, Powell AL, Callan J, Viles CL, Emmitt T, Prey KJ, Mou Y. Comparing the performance of database selection algorithms. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'99). New York: ACM Press, 1999. 238−245. [doi: 10.1145/312624.312684]

[45] Hawking D, Thistlewaite P. Methods for information server selection. ACM Trans. on Information Systems, 1999,17(1):40−76. [doi: 10.1145/297117.297123]

[46] Xu J, Croft WB. Cluster-Based language models for distributed retrieval. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'99). New York: ACM Press, 1999. 254−261. [doi: 10.1145/312624.312687]

[47] Balakrishnan R, Kambhampati S. SourceRank: Relevance and trust assessment for deep Web sources based on inter-source agreement. In: Srinivasan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R, eds. Proc. of the 20th Int'l Conf. on World Wide Web (WWW 2011). New York: ACM Press, 2011. 227−236. [doi: 10.1145/1963405.1963440]

[48] Yu C, Philip G, Meng WY. Distributed top-*N* query processing with possibly uncooperative local systems. In: Freytag JC, Lockemann PC, Abiteboul S, Carey MJ, Selinger PG, Heuer A, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB 2003). San Francisco: Morgan Kaufmann Publishers, 2003. 117−128.

[49] Fan J, Zhou LZ. Keyword-Based deep Web database selection. Chinese Journal of Computers, 2011,34(40):1797−1804 (in Chinese with English abstract).

[50] Wang Y, Zuo WL, He FL, Wang X, Zhang AQ. Ontology-Assisted deep Web source selection. Computer Science for Environmental Engineering and EcoInformatics, 2011,159(2):66−71. [doi: 10.1007/978-3-642-22691-5_12]

[51] Mihaila GA, Raschid L, Vidal ME. Using quality of data metadata for source selection and ranking. In: Suciu D, Vossen G, eds. Proc. of the 3rd Int'l Workshop on the Web and Databases (WebDB 2000). Heidelberg: Springer-Verlag, 2000. 93−98.

[52] Wang F, Agrawal G, Jin RM. A system for relational keyword search over deep Web data sources. Technical Report, Columbus: The Ohio State University, 2008

[53] Dasgupta A, Das G, Mannila H. A random walk approach to sampling hidden databases. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2007). New York: ACM Press, 2007. 629−640. [doi: 10.1145/1247480.1247550]

[54] Dasgupta A, Jin X, Jewell B, Zhang N, Das G. Unbiased estimation of size and other aggregates over hidden Web databases. In: Elmagarmid AK, Agrawal D, eds. Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2010). New York: ACM Press, 2010. 855−866. [doi: 10.1145/1807167.1807259]

[55] Nguyen K, Cao J. K-Graphs: Selecting top-*k* data sources for XML keyword queries. In: Hameurlain A, Liddle SW, Schewe KD, Zhou XF, eds. Proc. of the 22nd Int'l Conf. on Database and Expert Systems Applications (DEXA 2011). Heidelberg: Springer-Verlag, 2011. 425−439. [doi: 10.1007/978-3-642-23088-2_31]

[56] Zhu GS, Huang H, Yang WD. Keyword search based XML data source selection. Journal of Chinese Computer Systems, 2012, 33(6):1183−1188 (in Chinese with English abstract).

[57] Yu B, Li GL, Sollins K, Tung AKH. Effective keyword-based selection of relational databases. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2007). New York: ACM Press, 2007. 139−150. [doi: 10.1145/1247480.1247498]

[58] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996,12(4):5−33.

[59] Naumann F, Freytag JC, Spiliopoulou M. Quality-Driven source selection using data envelopment analysis. In: Chengalur-Smith IN, Pipino L, eds. Proc. of the 3rd Int'l Conf. on Information Quality (ICIQ'98). Cambridge: MIT, 1998. 137−152.

[60] Aboulnaga A, El Gebaly K. μBE: User guided source selection and schema mediation for Internet scale data integration. In: Chirkova R, Dogac A, Özsu MT, Sellis TK, eds. Proc. of the 23rd Int'l Conf. on Data Enginering (ICDE 2007). Washington: IEEE Computer Society Press, 2007. 186−195. [doi: 10.1109/ICDE.2007.367864]

[61] Xian XF, Zhao PP, Yang YF, Xin J, Cui ZM. Efficient selection and integration of hidden Web database. Journal of Computers, 2010,5(4):500−507.

[62] Deng S, Wan CX, Liu XP, Liao GQ. Selection of deep Web data sources based on user feedback. Journal of Chinese Computer Systems, 2012,33(11):2367−2371 (in Chinese with English abstract).

[63] Chang W, Sheikholeslami G, Wang J, Zhang AD. Data resource selection in distributed visual information systems. IEEE Trans. on Knowledge and Data Engineering, 1998,10(6):926−946. [doi: 10.1109/69.738358]

[64] Liu W, Meng XF, Meng WY. A survey of deep Web data integration. Chinese Journal of Computers, 2007,30(9):1475−1489 (in Chinese with English abstract).

附中文参考文献:

[2] 余伟,李石君,文利娟,田建伟.基于数据质量的 Deep Web 数据源排序.小型微型计算机系统,2010,31(4):641−646.

[21] 段青玲,杨仁刚,华松青.基于动态学习的 Deep Web 数据源选择算法.郑州大学学报(理学版),2010,42(1):5−8.

[49] 范举,周立柱.基于关键词的深度万维网数据库的选择.计算机学报,2011,34(10):1797−1804.

[56] 朱冠胜,黄浩,杨卫东.XML 关键字检索系统的数据源选择.小型微型计算机系统,2012,33(6):1183−1188.

[62] 邓松,万常选,刘喜平,廖国琼.基于用户反馈的深网数据源选择.小型微型计算机系统,2012,33(11):2367−2371.

[64] 刘伟,孟小峰,孟卫一.Deep Web 数据集成研究综述.计算机学报,2007,30(9):1475−1489.

万常选(1962－),男,江西新建人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据管理,XML 信息检索,金融数据挖掘,情感计算.

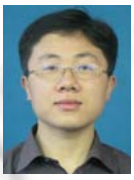
E-mail: wanchangxuan@263.net

廖国琼(1969－),男,博士,教授,CCF 高级会员,主要研究领域为数据库,数据挖掘,物联网数据管理.

E-mail: liaoguoqiong@163.com

邓松(1982－),男,博士生,讲师,主要研究领域为 Web 数据管理,数据挖掘.

E-mail: daonicool@sina.com



刘德喜(1975－),男,博士,副教授,CCF 高级会员,主要研究领域为 Web 数据管理,XML 信息检索,自然语言处理.

E-mail: dexi.liu@163.com

刘喜平(1981－),男,博士,讲师,主要研究领域为 Web 数据管理,XML 信息检索与挖掘.

E-mail: lewislxp@gmail.com

江腾蛟(1976－),女,博士生,讲师,主要研究领域为 Web 数据管理,XML 信息检索,情感计算.

E-mail: tj_jiang@163.com