# 基于关键词的深度万维网数据库选择

范 举 周立柱
(清华大学计算机科学与技术系 北京 100084)

**摘 要** 该文提出一种基于关键词的深度万维网查询方法:用户用关键词的方式提交查询,该方法在线地选择能够反映查询意图并且提供高质量结果的万维网数据库.这种方法既避免了深度万维网数据抓取这一代价高、难度大的操作,又可支持多领域的数据库上的关键词查询,从而能够与现有的搜索引擎实现无缝集成.文中侧重于讨论基于关键词的数据库选择,从以下两个方面解决这一问题所涉及的挑战:(1)提出了一种度量关键词-领域属性关联的相关性模型,并设计了基于随机游动的算法从查询日志中发现潜在的关键词-属性关联;(2)给出了一种新的数据采样方法,并用于基于采样的数据库-查询的相关性模型中,最终解决深度万维网的数据库选择问题.在中文深度万维网真实数据集上的实验表明:提出的方法能够有效地选择与关键词查询相关的数据库,提供高质量的结果.

**关键词** 深度万维网;万维网数据库;关键词查询;领域选择;数据库选择
**中图法分类号** TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01797

## Keyword-Based Deep Web Database Selection

FAN Ju ZHOU Li-Zhu
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

**Abstract** This paper proposes a keyword-based Deep Web search method: Given keyword queries provided by users, the proposed method on-the-fly selects the databases capturing the query intent and providing high-quality data. The method, which is much more efficient than Deep Web crawling, can support keyword search over multiple-domain Deep Web databases, and thus can be smoothly integrated with the existing search engine architecture. In this paper, we focus on keyword-based Deep Web database selection, and study the research challenges that naturally arise in the proposed method. (1) We introduce an effective model to measure the relevance of database-domain attributes with respect to keyword queries, and propose a random-walk algorithm to compute the relevance from database query logs. (2) We develop a novel database sampling method for measuring the relevance of databases with respect to queries, in order to select relevant databases in the selected domains. We have implemented our methods on real data sets from the Chinese Deep Web. The experimental results show that our methods achieve high effectiveness.

**Keywords** deep Web; Web databases; keyword search; domain selection; database selection

## 1 引 言

深度万维网是万维网上十分重要的数据资源.它是指存储在万维网数据库(简称数据库)中,只能通过查询接口,即 HTML 表单被用户访问的数据资源.深度万维网中丰富的数据对用户获取高质量信息大有帮助.据统计,截止到 2007 年,共有 2500 万个万维网数据库[1];截止到 2011 年,共有 600 万个中文万维网数据库[2].此外,与表面万维网(即存在链接

收稿日期:2011-08-12;最终修改稿收到日期:2011-09-15.本课题得到国家自然科学基金重点项目"支持中文 Web 研究的基础设施建设和应用中的基本方法与关键技术"(60833003)资助.**范 举**,男,1984 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为万维网规模数据管理、结构化数据的查询推荐.E-mail: fanju1984@gmail.com. **周立柱**(通信作者),男,1947 年生,教授,博士生导师,中国计算机学会(CCF)高级会员,研究领域为信息系统、数据库系统、数字图书馆等.

关系的网页数据)相比,深度万维网的数据有良好的结构性.以智联招聘网①为例,它存储的是包含城市、公司等属性的结构化数据.

然而,获取深度万维网信息面临着一个巨大的挑战:由于被查询接口所隐藏,深度万维网数据难以被基于网页链接关系的搜索引擎抓取器获取,从而无法像表面万维网那样被大量地索引和便捷地查询.为了解决这一问题,现有的研究提供了以下两类解决策略.第1类称为深度万维网"表面化"[3],即对万维网数据库提交查询,抓取返回的结果网页,并添加到搜索引擎的索引中.这种方法的优势在于最大程度上利用了现有搜索引擎的体系结构,但也存在着以下两点局限:(1)深度万维网数据规模庞大、增长迅速,将其表面化的代价很大.此外,很多万维网数据库限制访问次数,这使得表面化变得十分困难.(2)将搜索结果以普通网页的形式索引,丧失了数据的结构性.第2类方法称为万维网数据库集成,即将同一"领域"数据库(数据模式相同或近似)的查询接口进行集成[4-5].用户在集成接口提交查询后,该方法对查询进行翻译,提交给数据库[6],并将返回的结果进行整合.这种方法避免了表面化带来的问题,但也存在以下局限:(1)它局限于单一的领域;(2)它不支持被广泛接受的关键词查询,因此难以与现有搜索引擎进行整合.

为了解决上述问题,本文提出一种基于关键词的深度万维网查询方法:面向多领域的万维网数据库,提供基于关键词的查询接口,自动选择与查询关键词相关的一个或多个领域以及领域内包含相关结果的数据库,最终整合并返回查询结果.该方法具有以下优势:与"表面化"的方法相比,避免了数据库抓取这一代价高、难度大的操作,在线地选择体现用户查询意图的数据库.与万维网数据库集成的方法相比,不再局限于单一领域,而且支持关键词查询.因此,该方法可以使深度万维网上的查询与现有的搜索引擎系统进行无缝集成,使信息获取变得十分便利.

本文侧重于上述方法中的数据库选择问题,即给定一个关键词查询,(1)选择与查询相关的领域;(2)在选定的领域中选择数据库.为此,需要解决两点挑战:第一,关键词的灵活性导致难以判断用户的查询意图,从而难以选择与查询相关的领域.为了解决这一问题,本文提出了一种关键词-属性关联的相关性模型对领域进行排序,并设计了随机游动算法从查询日志中发现潜在的关键词-属性关联.第二,数据的隐藏性导致难以分析数据库的内容,从而难以选择与查询相关的数据库.为了解决这一问题,本文使用了一种基于采样的相关性模型对数据库进行排序,并提出了一种新型的采样方法,以获取用户当前最感兴趣的数据库记录.

简言之,本文的贡献在于:

(1)提出了基于关键词的深度万维网查询方法.

(2)研究了数据库领域与关键词查询的相关性模型,设计了一种随机游动算法从查询日志中发现关键词-属性的关联关系.

(3)研究了数据库内容与关键词查询的相关性模型,提出了一种新型的数据库采样方法.

(4)进行了真实数据集上的实验,验证了方法效果.

本文第2节描述问题定义和方法概览;进而在第3、4节讨论领域选择和数据库选择的技术要点;在第5节报告实验结果;第6节分析相关工作;第7节给出全文结论.

## 2 基于关键词的深度万维网数据库选择

### 2.1 问题定义

**数据**.本文主要研究多个异构的万维网数据库,并假设它们已经通过数据库集成技术[4-5]组织成若干个领域,且任一领域 $i$ 的数据模式可以表示为一个关系表 $R_i(A_1,A_2,\cdots,A_{m_i})$,其中 $A$ 为该领域的属性,领域中的数据库 $\{D_1,D_2,\cdots,D_{n_i}\}$ 的数据模式都可以对应到 $R_i$ 上.为了表述方便,在无歧义的情况下,$R_i$ 既可指示领域 $i$,也可指示该领域的关系表.

**查询及相关性**.本文侧重于讨论面向多领域万维网数据库的关键词查询 $K=\{k_1,k_2,\cdots,k_{|K|}\}$,其中 $k$ 为一个关键词.查询结果是与 $K$"相关"的一组数据库.相关性由函数 $\Phi(K,D)$ 表示,包含两个层面:

(1)"领域"相关性 $\Phi_S(K,R)$,衡量查询 $K$ 与数据库 $D$ 所在的领域关系表 $R$ 的相关程度.相关程度越大,领域相关性就越强;反之则越弱.

(2)"数据库"相关性 $\Phi_D(K,D)$,衡量数据库 $D$ 提供与查询 $K$ 相关的结果的能力.$D$ 包含的相关结果越多,数据库相关性就越强;反之则越弱.

**问题定义**.考虑多个领域,任一领域 $i$ 对应着一个关系表 $R_i$,并包含数据库集合 $\{D_1,D_2,\cdots,D_{n_i}\}$.给定关键词查询 $K$:(1)给出与查询相关的领域,并按照领域相关性 $\Phi_S$ 从高到低进行排序;(2)对于每个相关的领域,给出与查询相关的数据库,并按照数据库相关性 $\Phi_D$ 从高到低进行排序.

### 2.2 方法概览

本文提出了一个两阶段的数据库选择方法,如图1所示.第1阶段称为领域选择,即按照相关性

① http://www.zhaopin.com

$\Phi_S$对领域进行排序. 这个阶段的核心挑战在于：(1)关键词十分灵活，可能关联着多个领域的不同属性；(2)深度万维网的隐藏性导致预先获得关键词-属性关联关系的方法[7]不再可行. 为了解决这些问题，本文提出了查询日志挖掘的方法，包含线下的日志挖掘模块和线上的领域排序模块. 具体而言，在线下部分，日志挖掘模块对每个领域 $R_i$ 的查询日志进行挖掘，发现关键词-属性关联关系，并将不同领域的关联关系进行合并. 在线上部分，领域排序模块根据这些关联关系估计关键词查询 $K$ 与任一领域 $R_i$ 的相关性 $\Phi_S(K,R_i)$，并基于 $\Phi_S$ 将领域从高到低进行排序. 因此，在线下针对“单一”领域的日志进行挖掘，在线上则是回答用户“跨领域”的查询需求. 这部分内容将在第 3 节进行介绍.

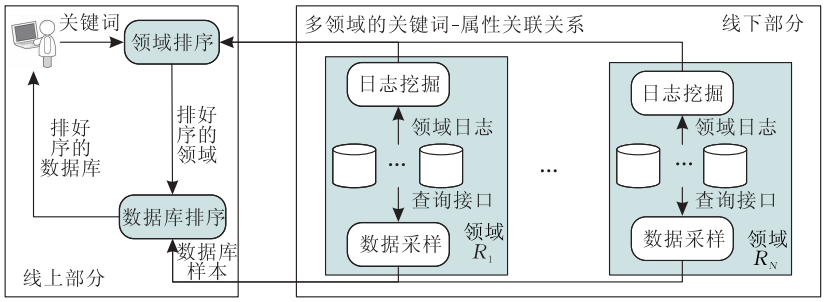



图 1　基于关键词万维网数据库选择方法概览

第 2 阶段称为数据库选择，即在选定的领域中，按照相关性 $\Phi_D$ 对数据库进行排序. 这部分的核心挑战是，由于数据隐藏性，难以估计数据库中相关结果的个数. 本文提出了基于数据采样的方法，包含线下的数据采样模块和线上的数据库排序模块. 具体而言，数据采样模块获得每个数据库的样本，数据库排序模块根据样本将选定领域的数据库按照数据库相关性 $\Phi_D$ 从高到低进行排序. 这部分将在第 4 节进行介绍.

**示例 1.** 考虑招聘和产品两个领域，其领域属性分别是招聘(城市，公司)和产品(名称，厂商). 在线下部分，日志挖掘模块根据每个领域的日志发现可能的关键词-属性的关联关系，如：“北京”关联着招聘领域的城市；“微软”关联着招聘领域的公司，以及产品领域的厂商等等. 在线上部分，考虑查询：“北京 微软”，领域排序模块根据挖掘出的关联关系计算查询与两个领域的相关性，并判断招聘领域的相关性大于产品领域. 对于每个相关的领域，数据采样模块估计数据库中相关结果的个数；数据库排序模块据此给出数据库的排序.

## 3　领域选择

本节介绍领域选择的相关方法. 第 3.1 小节讨论领域相关性 $\Phi_S$ 模型；第 3.2 和第 3.3 小节给出利用日志发现关键词-属性关联关系的模型和算法.

### 3.1　领域相关性模型

领域相关性模型用于度量关键词查询 $K$ 与任一领域关系表 $R$ 的相关性 $\Phi_S(K,R)$. 本文提出一种基于关键词-属性关联关系的方法. 一般情况下，用户用关键词指示领域属性的实例，如一个想搜索北京微软公司工作的用户会输入查询“北京 微软”. 因此，直观地讲，查询关键词与领域属性实例的关联性越大，该查询与领域的相关性就越强.

形式化地，定义关键词-属性的关联关系如下.

**定义 1**(关键词-属性关联，简称 KA 关联). 考虑关键词 $k\in K$ 和属性 $A\in R$，KA 关联 $m_{k,A}$ 存在，如果 $k$ 是属性 $A$ 的值，关联程度表示为概率 $P(m_{k,A})$.

假设用户的关键词不存在歧义性，即不存在关键词 $k$，同时与关系表 $R$ 中的不同属性 $A_i$ 和 $A_j$ 相关 $(i\neq j)$，则 $k$ 与 $R$ 的相关性可以用最大 KA 关联来估计，$\Phi_S(k,R)=\max\limits_{A\in R}\{P(m_{k,A})\}$. 进而，将所有关键词与关系表的相关性相加，得到

$$\Phi_S(K,R)=\sum_{k\in K}\max_{A\in R}\{P(m_{k,A})\} \tag{1}$$

### 3.2　关键词-属性关联的挖掘模型

式(1)中的核心在于计算关键词 $k$ 与属性 $A$ 的 KA 关联程度 $P(m_{k,A})$. 现有方法采用数据统计的策略，如计算关键词出现在属性值中的频率[7]. 然而，该策略不适用于深度万维网数据隐藏的场景. 本文提出一种查询日志挖掘的策略. 该策略基于这样的观察：尽管关键词查询十分灵活，但查询往往隐含着一些结构化的模式.

表 1 提供了招聘领域的查询日志示例，如查询 $Q_1$ 和 $Q_4$ 对应着“公司＋招聘”的模式，其中“公司”为领域属性，“招聘”为关键词. 现有研究将这种模式称为查询模板(简称模板)[8]. 查询模板反映了用户在某一领域数据库上用关键词表达查询意图的习惯模式，对挖掘 KA 关联大有帮助. 例如，根据“公司＋招

聘”模板，可以推测 $Q_4$ 中的“谷歌”可能为一家公司. 然而，根据模板推测 KA 关联并非轻而易举：(1) 简单地利用模板可能引入错误，例如根据“城市＋招聘”模板会把 $Q_4$ 中的“谷歌”误判为一个城市；(2) 产生模板本身需要依托 KA 关联. 例如，给定了“谷歌”与公司有关，才可将 $Q_4$ 归纳为“公司＋招聘”模板. 因此，KA 关联的程度与模板的质量相互影响：KA 关联越可靠，产生模板的质量就越高；模板的质量越高，推测的 KA 关联也就越可靠. 为了形式化地表述这种关系，首先给出以下定义.

**表 1 招聘领域关键词-属性关联关系挖掘示例**

| 查询 | |
|---|---|
| $Q_1$. 微软 招聘 | $Q_4$. 谷歌 招聘 |
| $Q_2$. 北京 微软 | $Q_5$. 北京 谷歌 |
| $Q_3$. 北京 招聘 | |
| 种子关联 | 推测的关联 |
| $m_1^s$. 微软-公司 | $m_3$. 谷歌-公司 |
| $m_2^s$. 北京-城市 | $m_4$. 谷歌-城市 |
| 招聘-领域关键词 | $m_5$. 谷歌-‘谷歌’ |
| 查询模板 | |
| $T_1$. 公司＋招聘 | $T_4$. 城市＋城市 |
| $T_2$. 城市＋公司 | $T_5$. 城市＋谷歌 |
| $T_3$. 城市＋招聘 | |

**定义 2**(查询模板). 给定领域 $R(A_1,A_2,\cdots,A_m)$ 和关键词全集 $W$，模板 $T$ 定义为 $\langle w_1,w_2,\cdots,w_{|T|}\rangle$，其中任一 $w_i\in W\cup\{A_1,A_2,\cdots,A_m\}$，且至少存在一个 $w_j\in\{A_1,A_2,\cdots,A_m\}$.

以前面提到的模板“公司＋招聘”为例，“公司”是领域 $R$ 的一个属性名称，“招聘”为 $W$ 中一个关键词. 基于模板的定义，给出查询-模板对齐的定义.

**定义 3**(模板-查询对齐，简称 QT 对齐或对齐). 给定日志中的关键词查询，$Q=\langle k_1,k_2,\cdots,k_{|Q|}\rangle$ 以及查询模板 $T=\langle w_1,w_2,\cdots,w_{|T|}\rangle$，QT 对齐 $L(Q,T)$ 存在，如果：(1) 长度相同，$|Q|=|T|$，且 (2) 查询关键词唯一地对应到模板中的属性或关键词，即 $\forall k_i\in Q$，如果 $w_i$ 为关键词，则 $k_i=w_i$；如果 $w_i$ 对应着属性 $A_j$，则有 $P(m_{k_i,A_j})>0$.

QT 对齐体现了依托于 KA 关联的查询到模板的归约关系. 例如，如果存在 KA 关联：微软-公司，则可以产生查询“微软 招聘”与模板“公司＋招聘”的 QT 对齐. 特别地，分别记 KA 关联 $m$ 和模板 $T$ 产生对齐 $L$ 为 $m\rightsquigarrow L$ 和 $T\rightsquigarrow L$.

基于以上定义，本文提出了 KA 关联、QT 对齐和模板之间关系的概率模型如下. 首先计算 KA 关联程度 $P(m)$，考虑 $m$ 产生的所有 QT 对齐，得到

$$P(m)=\sum_{m\rightsquigarrow L}P(m\mid L)\cdot P(L) \tag{2}$$

其中，$P(m|L)$ 表示 $m$ 在 $L$ 中的重要程度. 考虑均匀性假设，估计该概率 $P(m|L)=\frac{1}{|\{m|m\rightsquigarrow L\}|}$.

$P(L)$ 表示对齐 $L$ 中查询 $Q$ 与模板 $T\rightsquigarrow L$ 的紧密程度. 直观上讲，紧密程度取决于两方面因素. 其一是模板的质量：质量越高的模板，相应的对齐就越紧密；其二是 $L$ 中所有 KA 关联的程度. 由此可看出，对齐反过来又决定于 KA 关联，即

$$P(L)=\alpha_1P(T)+\alpha_2\sum_{m\rightsquigarrow L}P(L\mid m)P(m) \tag{3}$$

其中，$\alpha_1$ 和 $\alpha_2$ 为参数，且 $\alpha_1+\alpha_2=1$；$P(L|m)$ 表示 $L$ 在 $m$ 产生的所有对齐中的重要程度. 考虑均匀性假设，可以估计.

$P(T)$ 为模板 $T$ 的质量，取决于 $T$ 产生的所有 QT 对齐的紧密程度：这些对齐越紧密，则模板就越可信，质量就越高，即

$$P(T)=\sum_{T\rightsquigarrow L}P(T\mid L)\cdot P(L) \tag{4}$$

其中 $P(T|L)$ 表示 $L$ 对 $T$ 的影响. 考虑均匀性假设，可以估计 $P(T|L)=\frac{1}{|\{T|T\rightsquigarrow L\}|}$.

### 3.3 关键词-属性关联的挖掘算法

从上述模型可见，KA 关联 $m$、QT 对齐 $L$ 以及模板 $T$ 概率的计算是迭代进行的. 因此，笔者提出一个弱监督的随机游动(Random Walk)算法. 基本思想是：预先给定一些 KA 关联作为种子，并设定其关联程度为 1；建立“关联-对齐-模板”的三分图(简称 MLT 图)；利用随机游动的框架和式(2)、(3)、(4)迭代地计算 $P(m)$、$P(L)$ 和 $P(T)$.

首先定义 MLT 图如下.

**定义 4**(MLT 图). MLT 图定义为无向图 $G(V,E)$，其中结点集合 $V$ 包含关联结点、对齐结点和模板结点，即 $V=V_m\cup V_L\cup V_T$；边集 $E$ 包含关联-对齐边，即 $e=\langle v_m,v_L\rangle$，$v_m\in V_m$，$v_L\in V_L$ 以及对齐-模板边，即 $e=\langle v_L,v_T\rangle$，$v_L\in V_L$，$v_T\in V_T$.

表 1 给出了招聘领域的查询日志 $Q_1,Q_2,\cdots,Q_5$，并假设用户指定两个 KA 关联 $m_1^s,m_2^s$ 以及关键词“招聘”(即认为“招聘”不与属性关联). 由于“谷歌”可能对应到城市、公司以及非属性(普通的关键词)，猜测关联 $m_3,m_4,m_5$，进而归纳出模板 $T_1,T_2,\cdots,T_5$ 和 QT 对齐 $L_1,L_2,\cdots,L_9$，最终创建 MLT 图(图 2).

下面给出基于 MLT 图的 KA 关联挖掘算法，如图 3 所示. 首先，根据种子设置图上节点的初始值：种子节点的概率设置为 1，其余设置为 0. 然后，分别根据式(2)、(3)和(4)对非种子的关联结点 $m$、对齐结点 $L$ 和模板节点 $T$ 进行更新，直到停止条件(如迭代次数限制)得到满足.

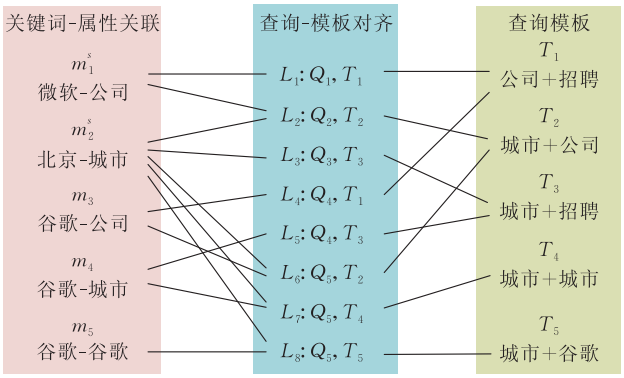



图 2　表 1 中招聘领域查询的 MLT 图

**输入**：图 $G(V_m \cup V_L \cup V_T, E)$，种子关联集合 $V_m^s$
**输出**：$V_m$ 中任一 $m$ 的关联程度 $P(m)$

1. $V_m^s$ 中任一 $m^s$ 设置 $P(m^s)=1$；其余节点概率设为 0；
2. 对于任一节点 $v \in V_m \cup V_L \cup V_T$，如下进行迭代：
   如果 $v \in V_m - V_m^s$，根据式(2)更新对应的 $P(m)$
   如果 $v \in V_L$，根据式(3)更新对应的 $P(L)$
   如果 $v \in V_T$，根据式(4)更新对应的 $P(T)$；
3. 如果停止条件满足则终止，否则转步 2.

图 3　基于 MLT 图的关键词-属性挖掘算法

**示例 2.** 利用图 2 给出算法示例(设 $\alpha_1=\alpha_2=0.5$). 初始情况下，设置 $P(m_1^s)=1$ 和 $P(m_2^s)=1$；其余设置为 0. 迭代 5 次后，可得到 $P(m_3)=0.62$、$P(m_4)=0.58$、$P(m_5)=0.55$，由此可判断“谷歌”是公司的概率比较大.

**算法实现的讨论.** 万维网用户查询的数量是巨大的，而且如果把所有的查询关键词都与所有可能的属性进行枚举，那么 KA 关联数目将非常多，因此这个 MLT 图的规模很大.

为了实现有效的计算，笔者使用了基于数据库的策略：将 MLT 图存储在关系数据库中. 进行迭代计算时，通过 SQL 语句依次访问每个节点及其相邻节点，并将更新后的分值写回数据库. 在此过程中，可以建立索引结构和内存中的缓冲结构对节点和相邻节点进行快速访问. 该策略的优点是简单且容易实现，缺点是需要频繁地访问数据库，效率不高.

在后续的工作中，笔者计划采用迭代式 MapReduce 框架，如 Haloop[9]. 该框架将所有迭代任务抽象成以下公式：$R_{i+1}=R_0 \cup (R_i \times L)$，其中 $R_0$ 是初始数据，$R_i$ 是第 $i$ 次迭代的结果，$L$ 是每次迭代不变的数据. Haloop 引入了 loop-aware 的任务调度，并优化了缓存、索引等. 与基于数据库的策略相比，Haloop 可以大大提高计算效率.

# 4　数据库选择

本节介绍如何在选定领域中选择与关键词查询相关的数据库. 第 4.1 小节给出基于采样的数据库相关性模型；第 4.2 小节讨论数据采样算法.

## 4.1　数据库相关性模型

直观上讲，如果查询 $K$ 在数据库 $D$ 中命中的记录越多，二者的相关性 $\Phi_D(K,D)$ 就会越大，即

$$\Phi_D(K,D)=|D_K|,\ \forall d \in D_K \text{ 满足查询 } K.$$

然而，正如第 2.2 小节提到的，由于数据是隐藏的，在没有提交查询 $K$ 之前，无法准确得到 $K$ 命中的记录个数. 为了解决该问题，现有方法提出了采样的策略[10-11]，基本思想是：采样出数据库中的一部分作为数据库样本 $D'$，计算查询 $K$ 在 $D'$ 中命中的记录个数 $\Phi_D(K,D')$；假设 $D'$ 中的一条记录可代表 $D$ 中 $|D|/|D'|$ 条结果，估计数据库相关性如下：

$$\hat{\Phi}_D(K,D)=\Phi_D(K,D')\cdot\frac{|D|}{|D'|} \tag{5}$$

通过式(5)进行估计需要解决两个问题：其一是获得数据库 $D$ 的大小，可采用现有方法解决[12]，篇幅有限，此处不进行展开. 其二是获取数据库 $D$ 的样本 $D'$，将在下一小节进行介绍.

## 4.2　数据采样方法

在理想情况下，一个随机且无偏的样本 $D'$ 可以满足式(5)的估计要求. 然而，实际情况下，由于无法直接访问数据库 $D$ 中的数据，只能采用基于查询的采样(QBS)框架[13]获取有偏的样本. 该框架的基本思想是：选择一系列的关键词查询提交到数据库 $D$ 上，将每个查询 $K$ 返回的 Top-$N$ 条结果(通常取 $N=4$)放入样本库 $D'$ 中，直到停止条件得到满足(后文会对停止条件进行讨论).

QBS 框架中的核心问题是查询的选择，即选择最优的查询，使样本库 $D'$ 的“质量”最高. 在衡量样本库的质量方面，现有方法[10-11,13]一般基于数据的统计信息，如关键词频率等，试图尽可能地获得高频词对应的数据记录. 然而，由于深度万维网发展迅速，数据更新频繁，用户往往对当前的热门数据更感兴趣. 因此，本文提出一种新型的衡量样本库 $D'$ 质量的指标：样本库 $D'$ 覆盖住“热门”数据记录的能力. 由于查询日志最能反映用户当前的查询兴趣，因而考虑从日志中选取“高质量”的关键词进行采样. 一种简单地衡量关键词质量的方法是它在日志中的频率，即包含该关键词的查询的个数：关键词越频繁，其用于采样的质量就越高. 然而，一些高频词的“区分度”不大，未必能够命中用户最感兴趣的记录. 例如，招聘日志的高频词为“招聘”、“求职”等，它们获取高质量记录的能力有限. 本文提出了一种新型的采样方法：选择与领域属性关联紧密的关键词进行采样(图 4). 算法的输入是第 3 节挖掘的 KA 关

联,按照关联程度对相应的关键词进行排序,并依次地提交到数据库中进行采样(第1~2行),将关键词返回的 Top-$N$ 的结果添加到样本库 $D'$ 中(第3行).如果停止条件得到满足,则算法终止,返回样本库 $D'$,否则继续选择关键词采样.

**输入**:(1) KA 关联集合 $M=\{m_1,m_2,\cdots\}$;
(2) 整数 $N$
**输出**:样本库 $D'$

1. 从 $M$ 中选择最大的 KA 关联 $m$;
2. 使用 $m$ 的关键词进行采样;
3. 返回 Top-$N$ 的结果,将返回的结果加入 $D'$;
4. $M=M-\{m\}$;
5. 若满足停止条件,返回 $D'$,否则转步 2.

图 4 基于 KA 关联的数据采样方法

**示例 3**. 以图 2 为例,考虑挖掘出的 KA 关联:"北京-城市"、"微软-公司"、"谷歌-公司".按照关联程度大小,依次使用对应的关键词进行采样.

**停止条件的讨论**. QBS 框架[13]对停止条件进行了讨论:当样本规模 $|D'|$ 达到一定数量,则停止采样算法,并通过实验来选择停止的样本规模,一般为500篇文档.本文也采用类似的停止条件,并在第5节比较不同样本规模下,不同采样方法的准确程度.

## 5 实 验

本节给出真实数据集上的实验结果.程序由 JAVA 语言实现,所有的实验运行于 Ubuntu 服务器:Intel Core 2 Quad X5450 3.00 GHz 处理器,4 GB 内存.

### 5.1 实验设置

**数据集**. 本文选择了中文深度万维网热门的5个领域(见表2).每个领域选择了若干万维网数据库,如招聘领域选择了智联招聘网①、前程无忧网②.因篇幅有限,此处省略领域数据库的完整列表.

表 2 实验数据集

| 领域 | #数据库 | 查询日志规模 |
|---|---|---|
| 招聘(城市 公司 职位) | 2 | 13652 |
| 影视(名称 明星) | 8 | 38824 |
| 歌曲(名称 歌手 专辑) | 5 | 100000 |
| 高考(省份 高校 专业) | 4 | 6596 |
| 游戏(名称) | 3 | 3260 |

领域的查询日志提取自 Sogou 浏览器③用户访问日志,即匿名用户历史上浏览过的 URL.提取查询日志的方法如下:(1)使用2010年6月的浏览日志;(2)选择同一领域数据库的 URL,解析通过 GET 方式传递的关键词查询;(3)为了便于分析模板特性,只保留至少含有两个关键词的查询.领域模式、数据库和查询日志的统计信息见表2.

**实验评测及对比方法**.

(1) KA 关联挖掘效果评测. 考察从单一领域查询日志中挖掘(第3.3节)的 KA 关联是否准确:对于任一领域,使用 Sogou 输入法细胞词库作为标准 KA 关联④,从中随机选出一定比例作为种子进行挖掘,用余下的作为测试,衡量推测的 KA 关联的准确率.

(2) 领域选择效果评测. 考察领域选择模型(第3.1节)是否有效,即能否选择出"跨领域"的查询最相关的领域:将查询日志的一部分(如20%)用于测试,余下的按照领域进行 KA 关联的挖掘.测试时,根据模型给出测试查询 Top-1 的领域,并与其真实领域进行比较,计算准确率.此处考虑对比模型 TF-IDF 方法,即将同一领域的查询日志看成一篇大的"文档",计算关键词在该领域的频率(TF)以及包含该关键词的领域的个数(DF),进而使用标准的 TF-IDF 模型计算查询与领域的相关性.

(3) 数据库采样效果评测. 考察采样方法(第4.2节)能否获得高质量的样本:笔者下载了狗狗影视⑤的 14 277 条记录作为数据库 $D$.分别采用基于频率的方法(对比方法)和基于 KA 关联的方法得到样本库 $D'$.随机选择50个查询,在 $D$ 中得到真实的相关记录个数 $|D_K|$,通过样本库估计相关记录个数 $|D'_K|$,进而计算相对误差率:$\dfrac{\mathrm{abs}(|D_K|-|D'_K|)}{|D_K|}$.

### 5.2 实验结果

**KA 关联挖掘实验**. 实验选用不同比例的词库作为种子进行训练,并计算推测的 KA 关联的准确率.

图5给出了实验结果:推测的 KA 关联在各领域的准确率均处于55%~88%之间,这说明挖掘算法可以提供比较准确结果.其中,准确率最高的是"高考"领域:在75%~88%之间,其原因在于该领

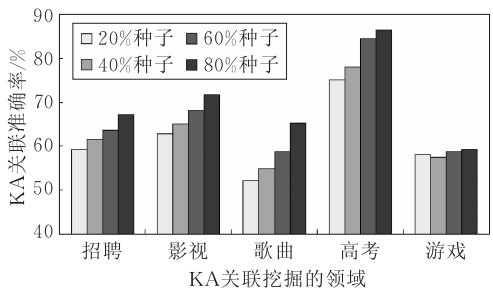



图 5 各领域 KA 关联挖掘效果评测

① http://zhaopin.com
② http://51job.com
③ http://www.sogou.com
④ 细胞词库包含若干属性及属性的取值,如"明星"属性包含"刘德华"、"周润发"等6599个取值.
⑤ http://movie.gougou.com

域的查询结构性较强，多为“省份 高校 专业”的模式，因此根据模板猜测 KA 关联的效果就比较好；准确率最低的是“游戏”领域，其原因有二：第一是游戏的名称纷杂，由词库生成的种子覆盖查询关键词的比例比较低；第二是游戏领域的查询结构性较弱，用户一般只输入游戏名称.

总结起来，通过查询模板进行 KA 关联挖掘的效果主要取决于领域查询的结构性：查询越能够归约为几类典型模板，KA 关联的挖掘效果就越好.

**领域选择实验**. 实验选择一定比例的查询日志进行训练：(1) 对于 TF-IDF 方法，统计关键词的频率；(2) 对于本文提出的模型，挖掘 KA 关联. 选择余下的一些查询进行测试. 如果模型给出的 Top-1 领域与测试查询真实的领域相同则记为 1，否则记 0，最后计算平均的准确率.

图 6 给出了实验结果. 为了更好地对模型进行比较，将查询按照真实的领域进行了分类. 可以看出，对于招聘、影视、歌曲和高考 4 个领域的查询，本文提出的模型均优于 TF-IDF 模型，即能更好地估计出查询所在的领域. 其中最显著的是影视领域，提出的模型准确率提高了 30%. 原因在于：模型考虑了查询与领域典型模板的相关性，可以减少因只考虑频率信息而引入的错误. 如：由于“长沙”在高考领域的频率远远大于招聘领域，因而“长沙 销售”可能会被 TF-IDF 模型误判为高考领域. 如果考虑了查询模板，则可判定该查询与“城市 职位”的模板相关性很高，因而更可能属于招聘领域. 另一方面，对于“游戏”领域的查询，提出模型的准确率不如 TF-IDF 的方法，其原因在于：游戏领域的模式比较简单(仅名称一个属性)，推测 KA 关联和模板的准确率不高(图 5)，因为对领域选择的帮助不大.

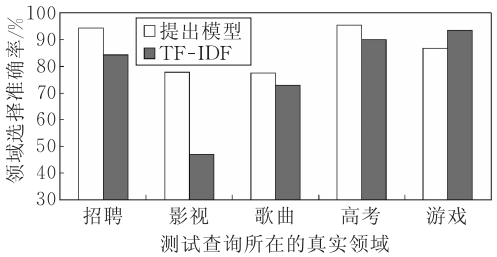



图 6　领域选择准确率的对比

总结起来，对于结构性较强的领域(即属性较多)，基于查询模板的方法能够提高领域选择的准确率；对于结构性较弱的领域，基于频率的方法比较好.

**数据库采样实验**. 实验采用基于 KA 关联的采样和基于频率的采样获取数据库样本，进而比较它们对相关记录估计的影响. 图 7 给出了实验结果：基于 KA 关联的采样可以获得比较低的误差，例如：当选取 3000 的样本时，平均误差率大概是基于频率方法的一半. 实验结果说明：仅凭关键词在查询中的频率难以判断它对于采样的重要程度；本文提出的基于 KA 关联的方法优先使用能够与领域属性对应的关键词进行采样，可以获得较好的数据库样本，降低相关记录估计的错误率.

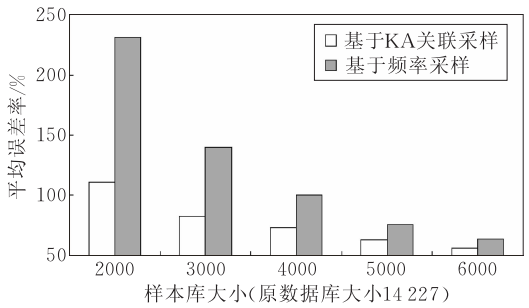



图 7　数据库样本对相关结果估计的影响

## 6　相关工作

深度万维网信息检索的研究主要分为两类：深度万维网“表面化”[3] 和万维网数据库集成[4-6]. 前者试图离线地将万维网数据库的网页内容抓取下来；后者局限于某一应用领域提供“垂直化”的服务. 与这两类方法相比，本文提出了基于关键词的深度万维网查询方法，一方面避免了数据库内容抓取这一代价高、难度大的操作，另一方面不再局限于某一领域，而是提供跨领域的关键词查询方式.

现有研究[8] 也使用了查询模板的概念，但该方法侧重于在点击数据(Click-Through)中发现模板与某一领域的关联度，并假设知道所有的关键词-属性关联关系(KA 关联)——该假设在深度万维网的场景下是不成立的. 尽管使用了模板的概念，本文侧重于假设知道部分 KA 关联(种子)，挖掘出更多的 KA 关联和模板，并估计它们的质量.

基于数据采样的数据库选择模型提出于元搜索领域[10-13]. 其基本思想是通过数据库样本来估计原数据库的相关结果个数. 与这些方法相比，本文根据深度万维网的应用场景，提出了 QBS 框架[13] 下的一种新型查询选择方法，有助于衡量数据库与当前流行查询的相关性.

## 7　结束语

本文提出了一种基于关键词的深度万维网查询

方法,并重点研究了数据库选择问题.具体而言,提出了基于关键词-属性关联的领域相关性模型,设计了基于查询日志的关键词-属性关联挖掘算法;提出了基于数据采样的数据库相关性模型,设计了一种新型的数据采样方法.在中文真实数据集上进行了实验,实验结果表明了提出方法的有效性.

**致 谢** 对 Sogou 公司的资助和数据支持表示感谢!

## 参 考 文 献


[1] Madhavan J, Cohen S, Dong X, Halevy A, Jeffery S, Ko D, Yu C. Web-scale data integration: You can afford to pay as you go//Proceedings of the CIDR. Asilomar, USA, 2007: 342-350

[2] Liu Yu-Kui, Zhou Li-Zhu, Fan Ju. Research on the present status of Chinese deep web database. Chinese Journal of Computers, 2011, 34(2): 360-370(in Chinese)
(刘玉奎, 周立柱, 范举. 中文深度万维网数据库的现状研究. 计算机学报, 2011, 34(2): 360-370)

[3] Madhavan J, Ko D, Kot L, Ganapathy V, Rasmussen A, Halevy A. Google's deep web crawl. PVLDB, 2008, 1: 1241-1252

[4] He H, Meng W, Yu C, Wu Z. Automatic integration of Web search interfaces with wise integrator. VLDB Journal, 2004, 12: 256-273

[5] He B, Zhang Z, Chang K C-C. Knocking the door to the deep web: Integrating web query interfaces//Proceedings of the SIGMOD. Paris, France, 2004: 913-914

[6] Zhang Z, He B, Chang K C-C. Light-weight domain-based form assistant: Querying Web databases on the Fly//Proceedings of the VLDB. Trondheim, Norway, 2005: 97-108

[7] Fan J, Li G, Zhou L. Interactive SQL query suggestion: Making databases user-friendly//Proceedings of the ICDE. Hannover, Germany, 2011: 351-362

[8] Agarwal G, Kabra G, Chang K C-C. Towards rich query interpretation: Walking back and forth for mining query templates//Proceedings of the WWW. Raleign, USA, 2010: 1-10

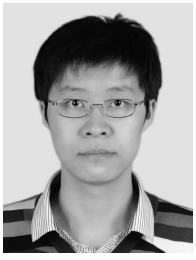
[9] Bu Y, Howe B, Balazinska M, Ernst M D. HaLoop: Efficient iterative data processing on large clusters. PVLDB, 2010, 3(1): 285-296

[10] Si L, Callan J P. Relevant document distribution estimation method for resource selection//Proceedings of the SIGIR. Toronto, Canada, 2003: 298-305

[11] Thomas P, Shokouhi M. Sushi: Scoring scaled samples for server selection//Proceedings of the SIGIR. Boston, USA, 2009: 419-426

[12] Shokouhi M, Zobel J, Scholer F, Tahaghoghi S M M. Capturing collection size for distributed non-cooperative retrieval//Proceedings of the SIGIR. Seattle, USA, 2006: 316-323

[13] Callan J, Connell M. Query-based sampling of text databases. ACM Transactions on Information Systems (TOIS), 2001, 19(2): 97-130

**FAN Ju**, born in 1984, Ph. D. candidate. His research interests include keyword search over structured data, query suggestion over structure data and Web-scale data management.

**ZHOU Li-Zhu**, born in 1947, professor, Ph. D. supervisor. His research interests include information system, ontology-based search and representation, database system, digital library, etc.

**Background**

Due to its large scale and high data-quality, Deep Web has become an extreme popular data source on the Web. Deep Web refers to the contents that are stored in the Web databases and can only be accessed by querying through search interfaces, i.e., HTML forms. Deep Web is playing a very important role to help Web users retrieve large scale high-quality information. There have been 25 000 000 Web databases to 2007[1], and 600 000 Chinese Web databases to 2011[2]. Compared with surface Web (Web pages with linked relationship), the data on deep Web is usually well structured. However, the wealth of data available on the Deep Web could not be retrieved by existing search engines, because the data is hidden behind the search-interfaces. In order to address the problem, two types of methods have been proposed. The first method, which is called Deep Web surfacing[3], pre-submits queries to Web databases and crawls the result pages. Although it can be well integrated with existing search-engine architecture, the method has difficulties when crawling the pages due to the large scale of data and access limitations, and it loses the structure of the data. The second type of method, which is called Web database integration, provides a vertical access way for databases in an application domain by interface integration[4-5] and result merging. While avoiding crawling Deep Web contents, this method can only allow users search databases in one domain and could not support keyword search, which is well accepted by Web users.

In this paper, we propose a keyword-based Deep Web search method: Given keyword queries provided by users, the proposed method on-the-fly selects the databases capturing the query intent and providing high-quality data. The method, which is much more efficient than Deep Web crawling, can support keyword search over multiple-domain Deep Web databases, and thus can be smoothly integrated with the existing search engine architecture. This paper focuses on keyword-based Deep Web database selection, and studies several research challenges, such as domain selection and database selection.