基于自然语言语义的数据库目标检索研究①

Database Object Retrieval Based on Natural Language Query

张　金[1,2]　涂俊翔[1]　陈卓宁[1,2]　严晓光[1,2]

(1.华中科技大学机械科学与工程学院 湖北 武汉 430074;

2.武汉开目信息技术有限责任公司 湖北武汉 430223)

摘　要： 人们难以用在数据库中通常采用的SQL或者内容匹配等形式化查询语言准确、快捷表达其检索要求。虽然基于人类自然语言的数据库信息检索，无需用户具备任何基础培训和应用经验，但传统的基于语法分析的自然语言处理技术过于复杂，其实用性受到了很大的限制。本文利用贝叶斯网络在处理不确定性问题上的优势，提出一种基于扩展贝叶斯网的数据库目标检索模型来处理人类自然语言查询的模糊性、自由性，能有效并较准确地获得数据库中PDM管理的目标数据对象。

关键词： 信息检索 数据库 PDM 贝叶斯网 自然语言理解

1　引言

随着 PDM 的管理范围向产品全生命周期各层次扩展，越来越多的非专业人员包括管理、生产、采购等部门的人员需要一种易于掌握的界面去访问所需的信息。以 SQL（结构化查询语言）等为代表的形式语言是目前数据库查询界面的主要语言。对普通用户来说，SQL 等形式化语言既难学习，又难使用。并且在使用 SQL 等语言查数据库中的数据时，用户必须知道该数据库的数据模式。如果能采用人类自然语言进行查询将能大大方便这类用户的要求，同时减少 PDM 查询系统的人机界面的设计工作以及用户培训负担。

目前数据库的自然语言查询处理技术大多是以词性、句法结构等语法分析作为基础，需建立大量的语法规则和复杂的知识推理机制，其实用性受到了很大的限制[1,2]。文献[3]提出建立中文关键词模式库来映射查询语句的思想，但它只考虑查询语句中单个关键词与模式库内术语间的匹配，未考虑到同一查询语句中各关键词间的相互约束关系，因此难以处理汉语查询语句中关键词的语义模糊性，不能准确区分语义相近的数据对象。例如，查询语句中“编号”可能对应数据库中员工表中的编号属性也可能是指物料表中的码值属性。

贝叶斯网络是人工智能领域用于处理不确定性问题的重要方法，它作为概率模型的扩展已经应用于自然语言检索领域[4,5]。本文利用扩展的贝叶斯信念网络信息检索模型对数据库的汉语查询语句进行处理，它无需构建复杂的语法规则，能处理查询语句中各关键词之间的关联性，从而能更有效并较准确地获得用户所关心的 PDM 数据对象。

2　基于贝叶斯网络的信息检索模型

贝叶斯网络是一系列变量的联合概率分布的图形表示，它包含两个部分。一部分是由节点、有向弧所组成的有向无环图 (Directed Acyclic Graph，DAG)，网络中每一节点代表研究领域中的变量，节点之间的连接关系代表变量间的条件独立语义；另一部分是节点和节点之间的条件概率表（Conditional Probability Table, CPT），它量化了各变量间的相互

① 基金项目:国家 863 高技术研究发展计划资助项目(2006AA01Z187,2007AA040605)
收稿时间:2008-08-06

依赖关系。假设贝叶斯网中任意节点 x_i 的直接双亲节点集为 P_{ai}， x_i 的条件概率为 $P(x_i \mid P_{ai})$。对于顶点集合 $X=(x_1,x_2,...,x_n)$ 的联合概率分布可由下式计算：$P(X)=\prod_{i=1}^{n} P(x_i \mid P_{ai})$。

按照 Ribeiro-Neto 和 Muntz 所提出的贝叶斯网络信息检索模型[6](图 1)，节点 d_j 代表文档节点，节点 q 代表查询节点，节点 k_i 是与 q 、 d_j 相关联的术语节点。设 u 为所有 k_i 节点所组成的样本空间 U 的一个任意子集， $u \subseteq U$。信息检索的过程即可看作是计算相对概率 $P(d_j \mid q)$ 的过程。由贝叶斯规则和全概率公式可知：

$$P(d_j \mid q)=\eta \sum_{u} P(d_j \mid u)P(q \mid u)P(u) \qquad (1)$$

$P(d_j \mid q)$ 的值愈大，表示 d_j 与查询 q 的匹配程度愈高。反之亦然。 $P(d_j \mid q)$ 值大于规定阈值的文档 d_j 则被视为满足用户查询需求的目标文档。

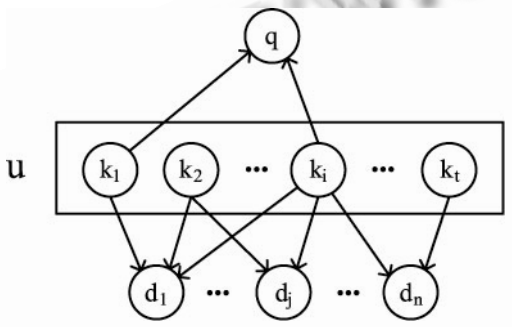

图 1　Ribeiro-Neto 和 Muntz 的贝叶斯网络信息检索模型

3　自然语言查询的数据库目标检索模型

结合数据库目标检索的特点，设计出基于扩展贝叶斯网络的信息检索模型如图 2 所示。本文后面各小节将详细描述模型各组成部分，并说明数据库目标的检索、推理过程。

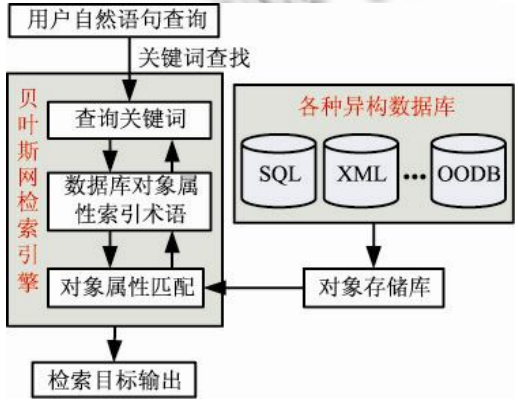

图 2　基于扩展贝叶斯网络的数据库目标检索

3.1　异构数据库的对象层抽象

PDM 平台下存在关系型数据库、对象型数据库以及 XML 数据库等多种异构数据库。为具通用性，我们将这些结构各异的数据库看成对象存储库，即为存储对象的集合 $C=\{o_1,o_2,...,o_n\}, n \geq 1$。对象由一组属性来描述，如 PDM 文档数据库中的文档对象具有文档编号、文档类别名、文档名称、文档编制时间、文档编制人等属性。具体对象由这些属性的取值惟一确定。有时具体对象的某一属性可以有多个取值，例如文档对象的编制人属性，这时我们称此属性为多值属性。多种属性的取值以集合的形式存在。

3.2　扩展的贝叶斯网络结构

用户查询的数据库对象是由查询语句中的一些关键词确定，这些关键词代表了 PDM 数据库对象的属性及其取值。计算机理解自然语言查询的过程即是这些关键词与数据库对象属性及其取值的匹配过程。按照上文所提的贝叶斯网络推理模型，我们构造出图 3 所示的贝叶斯网。模型中包括 3 类节点：查询节点 q，根节点， $k_i \in U=X \cup Y=\{x_1,...,x_r,...,x_m,y_1,...,y_s,...,y_n\}$， $(1 \leq i \leq m+n)$ 对象属性节点 $a_t(1 \leq t \leq u)$ 以及对象节点 $o_j(1 \leq j \leq v)$。查询节点 q 代表用户输入的自然语言查询语句。根节点中 $x_r(1 \leq r \leq m)$ 为 PDM 数据库对象属性的索引术语。它们由用户在查询中常使用的关键词组成，这些关键词表征了数据库对象的属性且在语义上与属性字段等价或相近。数据库对象属性的字段名必然是一个索引术语。索引术语是在调研众多的 PDM 用户的基础上根据领域专家的意见确定。根节点中 $y_s(1 \leq s \leq n)$ 为对象属性在数据库记录中的取值。对象属性节点 a_t 由可查询的数据库对象各属性字段的名称组成。对象节点 o_j 即为可查询的数据库目标对象。

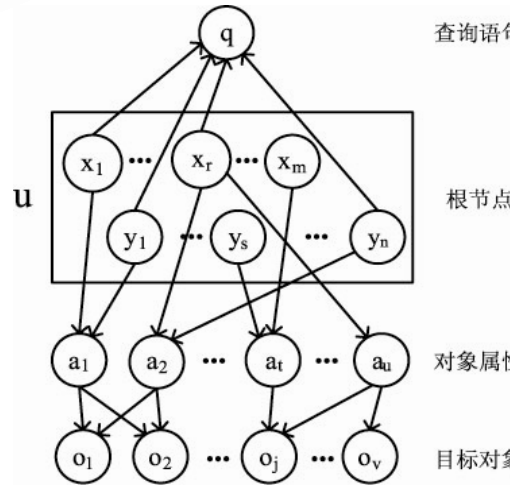

图 3　检索模型的扩展贝叶斯网络结构

我们不对查询语句作复杂的句法分析和词语切分，而是在查询语句 q 中查找与 k_i 匹配的字符串。如果在查询语句中找到 k_i，则自节点 k_i 有一条边指向查询节点 q，表示 k_i 是组成查询的一个术语 q。如果有多条边从根节点指向查询节点，则表示查询目标将由 q 中的多个关键词共同决定。此即是同一查询语句中各关键词语义上的相互约束关系在贝叶斯网络中的表现。若自根节点 x_r 有一条边指向对象属性节点 a_t，表示 x_r 是对象属性 a_t 的一个索引术语；自根节点 y_s 有一条边指向对象属性节点 a_t，表示 y_s 是对象属性的一个取值。从此结构中可以看出，该贝叶斯网络不仅考虑了任一对象属性 a_t 同其语义相近的索引词之间的匹配关系，同时考虑了 a_t 同其取值之间的匹配关系，从而提高了目标检索的准确性。

3.3 自然语言查询目标的推理过程

确定自然语言查询的数据库目标的过程即是计算 $P(o_j|q)$ 值的过程。$P(o_j|q)$ 的值越大，表示 o_j 是用户查询目标对象的可能性越大。$P(o_j|q)$ 反之亦然。值大于规定阈值的数据库对象则被视为满足用户查询的数据库目标。由贝叶斯规则和全概率公式可知：

$$P(o_j|q)=\alpha\sum_u P(o_j|a_t)P(a_t|u)P(q|u)P(u) \qquad (2)$$

为确定 $P(u)$ 的值，可假设样本空间 U 的各任意子集 u 等概率发生，即 $P(u)=(1/2)^{m+n}$。条件概率 $P(o_j|a_t)$、$P(q|u)$ 的计算同 Ribeiro-Neto 和 Muntz 的检索模型，参见文献[6]。$P(a_t|u)$ 表征了给定的子集 u 与对象 a_t 属性的匹配程度。$P(a_t|u)$ 值越大则 u 与 a_t 越匹配。反之亦然。依据矢量空间模型计算 $P(a_t|u)$ 如下：

$$P(a_t|u)=\frac{\sum_{i=1}^{m+n} w_{it}\times w_{iu}}{\sqrt{\sum_{i=1}^{m+n} w_{it}^2}\times\sqrt{\sum_{i=1}^{m+n} w_{iu}^2}} \qquad (3)$$

其中 w_{iu} 为术语 k_i 在 u 中的权重，当 $k_i\in u$ 时，$w_{iu}=1$，否则 $w_{iu}=0$；w_{it} 为术语 k_i 在属性 a_t 中的权重，对权重 w_{it} 的不同计算方法将得到不同的检索策略。

4 应用实例

目前该技术应用于开目 PDM 对象管理器中。对象管理器是企业协同管理平台的一部分，它对企业的物料(包括零件、部件、产品)、文档、组织、设备、客户等各种信息进行统一的对象化管理。在管理器中实现基于语义导航功能能大大方便各类用户的查询。表 1 列出了某企业的对象管理器的部分查询对象及其可查询属性。

表 1　数据库查询对象及属性举例

对象类	可查询属性举例
图文档	图档代号、当前持有者、安全级别、名称、类型、最高版本、备注、创建时间、最后修改时间、创建人
零部件	零部件代号、父零部件代号、装入数量、零部件名称、零部件分类、材料
员工	姓名、性别、入职时间、离职时间
产品	名称、型号、规格、性能

图 4 是该企业的对象管理器的浏览、导航界面。图 4(a)是用户浏览企业对象分类树时的显示界面。这时用户未输入任何查询语句，系统未对目标对象进行条件过滤。界面的列表框列出了属于 E1141(载货汽车)的所有零部件。如果用户给出查询：q = “属于外购件的部件”。这时系统不对该语句作词性、句法结构等方面的语法分析，而是将其与系统内建的贝叶斯网根节点层的术语进行完全匹配分析。由于索引术语“外购件”、“部件”符合匹配要求，它们各有一条边（有向弧）会指向该查询节点 q；在另一方向，这两个根节点层术语又通过对象属性层指向目标对象层的各数据库具体对象。通过这些有向弧关联关系并利用式 2 可计算出数据库中各具体对象在当前查询下的条件概率 $P(o_j|q)$ (这时只需计算界面查询树中当前查询节点下各具体对象的 $P(o_j|q)$，见图 4(a))。$P(o_j|q)$ 的值表征了数据库具体对象 o_j 与当前查询 q 的关联程度。$P(o_j|q)$ 的值越大，表示 o_j 是用户查询目标对象的可能性越大。反之亦然。图 4(b)列出了在用户查询后 $P(o_j|q)$ 值大于阈值的目标对象。这些对象按 $P(o_j|q)$ 值的大小以降序排列显示。

表 2 列出的是管理器导航功能正确性测试的结果。系统所采用的测试语句是根据调查得来的，具有一定的代表性。表中误差率是系统处理自然语言查询所返回结果中漏选的正确对象与误选的错误对象的数量总和与所有正确查询对象的数量之比。对导航系统而言，该查询误差在可接受的范围内。

(a)　自然语句查询输入前的界面　　(b) 自然语句查询输入后的显示结果

图 4　集成扩展贝叶斯网推理技术的企业对象浏览、导航界面

表 2　系统正确性测试

语句字数(以汉字字符计算)	测试语句数	误差率
1~15	250	9.3%
16~30	250	12.7%

注：基于效率的考虑，单次查询语句的长度限制在 30 个汉字之内

5　结束语

本文对 PDM 系统中各种异构数据库进行对象化抽象，在此基础上建立了基于扩展贝叶斯网络的 PDM 数据库查询模型。该模型能很好地并较简单地处理自然语句查询中的模糊语义问题，从而能更有效并较准确地帮助用户快速检索到所需的目标数据对象。该技术已经在开目 PDM 的对象管理器中得到了实际应用，对 PDM 朝着智能化方向的发展有着积极意义。

参考文献

1 刘开瑛,郭丙炎.自然语言理解.北京:科学出版社,1991.

2 孟小峰,王珊.数据库自然语言查询系统 Nchiql 中语义依存树向 SQL 的转换.中文信息学报,2001,15(5):40－45.

3 张连蓬,刘国林,江涛,李云岭,季民.受限自然语言查询在 GIS 中的应用.测绘学院学报, 2002,19(4):283－285.

4 Calado P, Silva AS, Laender AHF, Ribeiro-Neto B, Vieira RC. A Bayesian network approach to searching web database through keyword-based queries. Information Processing and Management, 2004, 40(5): 773－790.

5 Kyoung-Min K, Jin-Hyuk H, Sung-Bae C. A semantic Bayesian network approach to retrieving information with intelligent conversational agents. Information Processing & Management, 2007, 43(1): 225－236.

6 Ribeiro-Neto B, Muntz R. A belief network model for IR. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York: ACM Press, 1996. 253－260.