基于改进关联规则算法的数据库关键词检索方法

孟　巍　张东宁　郭腾炫　宗振国　孔　鹏
（国网山东省电力公司营销服务中心（计量中心）济南 250000）

摘　要　传统的数据库关键词检索方法只考虑了关键词的二元关系，造成了数据库关键词特征匹配度低的问题。针对该问题，文中提出了基于改进关联规则算法的数据库关键词检索方法。为对数据进行初步的分析和处理，首先确定数据库信息特征的模糊空间，不仅需考虑关键词的二元关系，还需计算数据库信息关键词语义重要问题。通过关联规则数据结构分布重排，改进关联规则算法，实现数据库关键词检索。实验证明，基于改进关联规则算法的数据库关键词检索的特征匹配度较高，检索结果更加准确。

关键词： 关联规则算法；数据库；关键词；检索方法
中图分类号　TP311

Database Keyword Retrieval Method Based on Improved Association Rule Algorithm

MENG Wei,ZHANG Dongning,GUO Tengxuan,ZONG Zhenguo and KONG Peng
（State Grid Shandong Electric Power Company Marketing Service Center (Measurement Center),Jinan 250000,China）

Abstract　The traditional database keyword retrieval method only considers the binary relationship of keywords, resulting in the problem of low matching of database keyword features. Aiming at this problem, this paper proposes a database keyword retrieval method based on improved association rule algorithm. In order to analyze and process the data preliminarily, first determine the fuzzy space of database information features, not only need to consider the binary relationship of keywords, but also need to calculate the semantic importance of database information keywords. Through the rearrangement of association rule data structure, the association rule algorithm is improved to realize database keyword retrieval. Experiments show that the database keyword retrieval based on the improved association rule algorithm has a higher matching degree of features and more accurate search results.

Key words　Association rule algorithm,Database Keywords,Retrieval methods

0　引言

关联规则算法在数据库关键词检索中具有重要意义，其能发现数据集中的相关关系，这些关系可以用于指导关键词的检索和排序[1]。通过发现关键词之间的关联规则，人们可以更好地理解数据集的结构和语义信息，从而为用户提供更准确、相关和有用的检索结果。传统的关联规则算法在数据库关键词检索中存在一定的不足。首先，文献[2]的数据库关键词检索方法具有较高的计算复杂度，这可能导致在大型数据集上运行缓慢或无法处理。其次，文献[3]的数据库关键词检索方法通常只考虑关键词的二元关系，而忽略了关键词之间的上下文信息和语义关系[4]。这可能导致算法无法准确理解关键词之间的真正意义和关系，从而影响检索结果的准确性。通过对比传统方法，本文提出了一种基于改进关联规则算法的数据库关键词检索方法，旨在克服传统方法的不足，通过采用更高效的关联规则算法、考虑关键词的上下文信息和语义关系，确保其准确性，更好地满足用户需求，提高数据集的使用价值。

1　确定数据库信息特征的模糊空间

假设数据库中的资料库资讯为 V，其中含有 $[f_1,f_2,\cdots,f_n]$ 个资料库资讯特性。在每个资料库资讯特性中，皆可通过 $[f_{i1}\times f_{i2}\times\cdots\times f_{in}]$ 进行度量，再以 $[f_{i1},f_{i2},\cdots,f_{in}]$ 组成资料库资讯特性空间。然后，利用关键词中的模糊子集 $[a_{i1},a_{i2},\cdots,a_{in}]$ 求出资料库资讯的特征空间，并将其表示为 $[a_{i1}\times a_{i2}\times\cdots\times a_{in}]$[5]。图 1 显示了一个数据库信息特性的模糊空间。

在图 1 中，用 A，B，C 表示数据库信息特性中一个模糊空间内的一条对应的线。根据 a_{i1}、a_{i2} 与 a_{i3} 的关系可知，

基金项目：国家电网有限公司科技项目（5700-202341299A-1-1-ZN）
作者简介：孟巍（1979—），本科，副高级工程师，研究方向为电力营销技术。

当资料库资讯的特性空间维度较高时，资料库资讯特性的模糊空间所占的区域也会随之缩小。反之，则资料库资讯特性模糊区会变大。

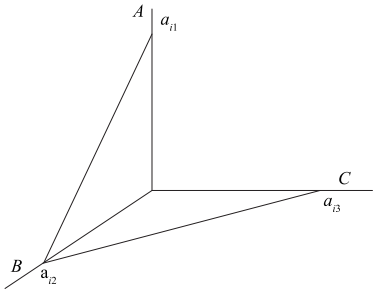

图1　模糊空间表示数据库的信息特性图

2　计算数据库关键词语义的重要性

在定义了数据库中的信息特性的模糊空间后，利用模糊语义距离来表示集成库中的信息。考虑到数据库信息关键字的语义重要性，有研究者提出了基于汉明距离的数据库信息关键字模糊语义距离计算方法[6]。设它为$Sim(x,y)$，可得式(1)：

$$Sin(x,y)=1-\frac{\sum_{i=1}^{n}\left(w\times\sum_{k}^{m}w_i\times\left|\mu(x)-\mu(y)\right|\right)}{\sum_{i=1}^{g}\left(w\times\sum_{k}^{m}w_i\times Max(\mu(x),\mu(y))\right)} \tag{1}$$

其中，n为数据库中关键词的语义特征矢量；i为综合数据库信息的属性数量(实数值)；w为平滑系数；m表示摩擦因素；k为综合数据库中关键字的重要权重系数；g为数据库中关键词的隶属度函数；μ为数据库中关键词的特征维度；x表示数据库信息一维异常汉明距离的横向坐标轴；y是数据库信息一维异常汉明距离的纵轴。

基于以上内容，将(1)式转化为(2)式：

$$Sin(x,y)=\frac{\sum_{i=1}^{n}\left(w\times\sum_{k}^{m}w_i\times Min(\mu(x),\mu(y))\right)}{\sum_{i=1}^{g}\left(w\times\sum_{k}^{m}w_i\times Max(\mu(x),\mu(y))\right)} \tag{2}$$

由(2)式可知，通过资料库资讯一维异常汉明距离的横坐标和坐标，可直接确定资料库资讯关键字在模糊语义距离属性中的特定位置，即模糊语义特征。在此基础上，根据其本质特性，对数据库中的信息(模糊特征值和数据库中的信息特征距离)进行计算，从而获得了一个完整的模糊语义数据库，如表1所列。

表1　数据库特征信息

数据库信息模糊特征数值	模糊语义特征描述	模糊语义距离属性	数据库信息特征距离
0.258	(0,2,0,8)	非负性	0.286
0.179	(0,1,0,9)	对称性	0.258
0.287	(0,2,0,7)	三角不等式	0.276

从表1可知，数据库信息的特征距离值在非负类型模糊语义距离时最大，在对称情形中，数值最小，在三角形的意义上，数值居中。在此基础上，有研究者提出了在数据库中使用对称模糊语义的距离属性来进行信息检索的方法[7]。在数据库中提取信息时，可能会出现查询条件不完备的情况，因此必须剔除不确定性，提高查询的准确率。另外，数据库中的特征间距普遍很小，可利用关键字的指数分布衰落度函数来计算数据库信息关键字的语义权重[8]，如式(3)所示：

$$P=(D|T)=ae^{-a(c-o)} \tag{3}$$

其中，D是数据库关键词，T是检索时间，a是数据库中的关键词项数目，e是数据库的先验概率，c是信息在数据库中的时间，o为建立数据库的时间。

假设关键词对数据库信息元组的重要性权重为K，如式(4)所示：

$$K=Pn\cdot\sum_{j=1}w(a)\cdot IR(a,k) \tag{4}$$

其中，j为数据库信息的模糊语义距离属性，其默认值通常为1，a为数据库信息元组的重要性，I表示关键字在数据库中的重要程度，R表示关键词的得分。

根据(4)式，可以得到数据库关键词的语义重要程度。

3　改进关联规则算法

为解决在扩展的文本库中存在的簇特征不一致的问题，本文在传统的基于虚拟数据分布的数据库关键词检索技术的基础上，提出了一种新的数据库关键词检索方法。该方法利用关联规则的分布关系，在对有限带宽中心进行预先提取的基础上，建立了一个实体关联知识库，并对该知识库进行有效的引导与重组[9]。其表达式如式(5)所示：

$$Q_{rev}(T)=\frac{\frac{1}{N-T}\sum_{i=1}^{N-T}(x_i-x_{i+T})^3}{\left[\frac{1}{N-T}\sum_{i=1}^{N-T}(x_i-x_{i+T})^2\right]^{\frac{3}{2}}} \tag{5}$$

其中，Q_{rev}为重组后的知识库，N为语义信息，T为频繁项集，x为项集。

设由语义信息构成的具有长度K可以被划分为16个32比特密钥$K_0,K_1,\cdots,K_{15}$的异质数据序列，基于遗传算法，假定一个关联规则条目q_i具有n_i个最接近项目$d_i=(d_{i1},d_{i2},\cdots,d_{in_i})$[10]，得到二叉树$W$，如式(6)所示：

$$W(q_i,d_{ij})=\frac{g(q_i,d_{ij})\times\log_2[f(d_{ij})+1]}{\sum_{j=1}^{n_i}\left\{g(q_i,d_{ij})\times\log_2[f(d_{ij})+1]\right\}} \tag{6}$$

当一组最佳特性解出现在t_0时刻的一个分布数据库系统中，则具有一个标准公式，如式(7)所示：

$$\xi=2\rho_{\max}\lambda_{\max}(q_i)n\delta^2KT \tag{7}$$

其中，ξ为最优特征解集合，ρ为支持度，λ为置信度，δ为关联规则。

通过数据库流程对特征序列进行预处理。在关联规则数据结构的分布重组过程中，通过对系统功能的模糊减数簇迁移算子$h_i(t)$的研究以及对信息流的查询扰动$n_{pi}(t)$的研究，得出了第一代减法聚类离散度，如式(8)所示：

$$X_{ri}(t) = X(t)*h_i(t) + n_{pi}(t) \tag{8}$$

其中，$X(t)$表示数据库信息项。

当计算信息流减法簇多层簇中心卷积时，进行关联规则数据结构分布重排，从而得到式(9)：

$$h(t) = f(d_{ij}) \sum_{i=1}^{M} h_i(t)*h_i(-t) \tag{9}$$

其中，$f(d_{ij})$为关键词d_{ij}的度数。

通过以上步骤，可实现对关联算法的改进。

4 实现数据库关键词检索

确定数据库信息搜索时的优先权后，在进行数据库信息搜索时，就可利用搜索数据库信息的模糊相似性来抽取关键词，然后提取含有关键字词的元组，从而对得到的数据库信息进行模糊相似度计算。在使用关键词搜寻资料库资讯时，可将使用者搜寻最频繁的关键字作为使用者的资料库资讯特性，并撷取资料库资讯模糊相似度。假设可以用等式来表示搜索数据库信息的模糊相似性，如式(10)所示：

$$V = \frac{\sum_{i=1}^{n}\left(w \times \sum_{k}^{m} K \times Min(\mu(x),\mu(y))\right)}{\sum_{i=1}^{g}\left(w \times \sum_{k}^{m} K \times Max(\mu(x),\mu(y))\right)} - 10(\mu^2) \tag{10}$$

其中，μ为关键词搜索频率。

将式(10)作为一个最后基于关键词的搜索数据库信息的表达式，并基于关联规则搜索关键词特征向量的位置，以搜索数据库信息—模糊相似程度，并输出具有最大值的数据库信息—模糊相似度。资料库资讯的模糊相似性愈高，说明搜寻的结果愈符合关键词。至此，以改进关联规则算法为基础的数据库关键词检索方法已基本实现。

5 实验

5.1 实验准备

本实验在高性能计算机上开展，具备大内存、高速硬盘和分布式计算环境，并应用了最新版本的Python或Java编程语言及相关的数据挖掘和机器学习库。同时，采用了MySQL或PostgreSQL等数据库管理系统来存储和管理数据集。实验的关键在于选择具有代表性、多样性和足够规模的数据集，并使用多核高频的处理器和至少8GB内存的计算机。此外，使用SSD或高性能的机械硬盘可以加快数据读写速度，提高实验效率。为进行数据库关键词检索实验，本文将文献[5]、文献[10]中的方法与本文方法进行对比实验，选取1 000个真实序列集作为实验对象，再分别用3种数据库关键词检索方法对序列集进行检索，并记录数据库关键词特征匹配数。

5.2 实验及结果分析

通过对比实验，验证了本文算法在数据库中的应用效果。通过对网络文本数据库中的关联规则进行最小支撑阈值的计算，得到minsup p_1 = 30.0%。在此基础上，开展数据库关键词搜索，对数据库中的规则流进行结构重新排序，进行关联特征的挖掘，并对数据库中的关键词进行不同的搜索，获得的特征匹配结果如图2所示。

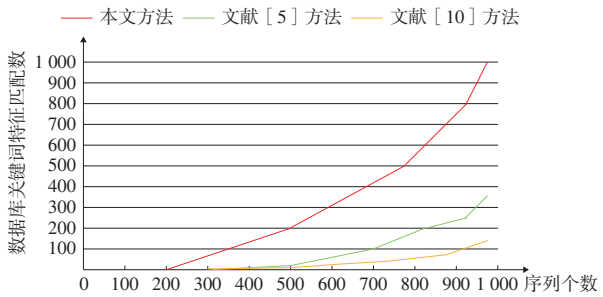

图2 数据库关键词特征匹配结果(电子版为彩图)

由图2可以看出，利用本文基于关联规则算法建立的关键词检索方法，得到的特征匹配数量是文献[5]方法、文献[10]方法的2倍之多。与常规的关键词检索方法相比，本文方法在特征匹配度上有很大的优势，具备更高的数据库关键词检索精度，提高了特征匹配度。

6 结语

研究基于改进关联规则算法的数据库关键词检索方法是一项具有挑战性和实用性的任务。通过深入探讨关联规则算法在数据库关键词检索中的应用，本文提出了改进后的关联规则算法，以提高检索的准确性。通过不断的优化和完善，该方法将在数据挖掘和信息检索领域发挥更大的作用。

参考文献

[1] 刘宁，牛佳乐，郑剑，等.基于向量空间模型的信息资源关键词智能检索工具的研究[J].自动化技术与应用，2023，42(10)：105-107，161.

[2] 张伟涛，米吉提·阿不里米提，艾斯卡尔·艾木都拉，等.基于深度神经网络的资源匮乏语言语音关键词检索[J].现代电子技术，2022，45(11)：68-72.

[3] 何亨，蒋俊君，冯可，等.多云环境中基于属性加密的高效多关键词检索方案[J].计算机科学，2021，48(S2)：576-584.

[4] 宋志平，吾尔尼沙·买买提，库尔班·吾布力，等.基于层次匹配的维吾尔文关键词图像检索[J].计算机工程与设计，2022，43(12)：3461-3467.

[5] 周倩，戴华，盛文杰，等.云环境下可验证关键词密文检索研究综述[J].计算机科学，2022，49(10)：272-278.

(下转第279页)