用户体验视角下数字图书馆数据智能查询优化研究

王烁，牛庆银
（装甲兵工程学院，北京 100072）

摘要：针对传统数字图书馆数据智能查询技术一直存在响应时间长、效率低且查询准确性较差的问题，本文提出用户体验视角下数字图书馆数据智能查询优化研究。首先，介绍数字图书馆数据智能查询技术的进展及存在意义；其次，在用户体验视角下采用数据关联规则方法对数字图书馆数据进行挖掘与提取，并依据关系数据库的基本思想对关联规则提取计算进行描述；最后，按照权重值从大到小的顺序给出查询结果，从而实现智能查询优化。结果表明，本文提出的方法查询效率高，查询精度明显优于传统方法。

关键词：用户体验视角；关联规则；数字图书馆；智能查询
中图分类号：G250.76
DOI：10.3772/j.issn.1673-2286.2017.07.008

随着社会文明逐渐发展，知识信息呈海量增长。面对广泛传播的信息，人们对知识的需求更加迫切，其阅读需求也有更个性化的特性[1-2]。数字图书馆具有信息量庞大、容易存储、使用便捷等特点，因此，快速准确地搜索资源成为人们检索文献时的突出要求。当前数字图书馆资源普遍具有一定规模，数据智能查询成为数字图书馆发展中亟需解决的关键问题[3]。

当前，数字图书馆各类数字文献数量规模增长迅速。面对庞大的数据，如何对其进行智能查询成为当前数字图书馆领域的研究热点[4]。在关于数据智能查询的研究中，传统方法只适用于在单台机器中执行，扩展能力较差。并行与分布式技术处理已成为提高数字图书馆异常数字文献检索准确率的有效方法，并占据主导地位[5]。本文提出的用户体验视角下MapReduce 能为图书馆数据处理提供一个易于实现且功能强大的并行和分布式数据处理模型，这对于分析用户体验视角下高效的数据智能查询技术具有重要的应用价值。

1 数字图书馆数据智能查询技术进展

针对数据的智能查询，通常采用K近邻连接方法。因为K近邻连接查询在数据库中耗时较长，不适于在数据规模较大的情况，用户体验视角下的分布式K近邻连接方法应运而生[6]。

周屹等将Voronoi图和K近邻连接方法结合，对数据进行智能查询，通过Voronoi图对原始数据进行分割，将数据划分至相应分组中，所有分组对象均被分配至相同的云计算节点，在Reduce阶段通过K近邻连接方法对数据进行查询处理[7]。上述研究依据Voronoi图的数据查询方法在数据划分合理的情况下性能较高，若划分结果不合理，则大幅降低查询效率。

Abiteboul等在Hadoop分布式框架的基础上提出一种简单的K近邻连接方法，通过Z曲线法把多维空间kNN查询转换至一维空间查询，在云计算环境下实现高效数据智能查询[8]。该方法主要有三个MapReduce任务：第一个为读取HDFS中两张表R与S中的数据，同时建立表R与S的随机移动R_i与S_i，得到R_i与S_i的分割值；第二个是把R_i与S_i分配至对应的模块中，同时找到$r \in R$，$KNN(r, S)$的候选点；第三个是从候选集中得到k个最近邻点作为结果集。上述任务均需完成对HDFS的读写操作，而前两个任务产生的输出只是临时的中间结果，所以上述查询技术的效率较低。

上述K近邻连接方法均是在云计算环境的基础上实现的，其只是一种简单的数据处理方式，把数据处

理过程简化为不同阶段，约束云计算的表达效果[9]。在云计算环境中，所有任务输出均需写入磁盘，特别是Reduce任务需写入HDFS。赵丹[10]与喻宜等[11]提出的数据智能查询技术均包含3个串行MapReduce任务，所有MapReduce任务均需对HDFS进行读写，面对大量数据会产生更多开销，且降低查询效率。

聚类分析作为计算机领域的一个基础研究问题，在网络和生物信息等领域受到广泛关注[12]。聚类分析，即把海量数据中的数据对象依据特定准则分割为若干类别，使被分配至相同类的数据间距离较小，未被分配至相同类的数据间距较大。当前常用的聚类方法主要包括依据密度的聚类方法、依据划分的聚类方法、依据网格的聚类方法、依据层次的聚类方法和依据模型的聚类方法[13]。本文采用的关联规则方法因其易于实现且效率较高，常被应用于大规模数据聚类。

聚类分析面对的数据量越来越大，导致单个节点不能达到存储要求，依据单节点单进程的串行执行方法不能直接进行扩展。因此，需设计一种用户体验视角下的聚类方法来解决数据智能查询问题。

当前已有大量研究对算法关联规则方法进行改进。宋爱波等为满足关系数据库管理系统研究，提出一种依据磁盘的关联规则方法[14]；杨良斌为提高关联规则方法的查询精度，针对数据对象引入聚类方法[15]；Campbell等提出一种依据k-d树的数据结构方法，通过k-d树完成对数据空间的分割，将分割的左右单元看作独立单元并进行处理，降低计算量[16]。

上述数据智能查询方法都是在集中式单节点环境下实现的。在数据量逐渐增加的情况下，分布式并行计算形式被广泛关注。李建荣提出一种依据消息传递模型的方法[17]，主要对消息传输模式下方法的效率与扩展性进行改进，云计算框架利用计算机集群能力可有效解决大规模数据处理问题，被广泛应用。

2 数字图书馆数据智能查询优化的实现

2.1 数字图书馆数据智能查询目的

通过数据智能查询完成对图书馆文献信息资源的检索，其主要目的是为读者提供与其需求相匹配的文献资源以及个性化服务，特别是提高这类服务的精确度与便捷性。图书馆需向读者推送相关信息，而查询是读者得到所需信息的重要方式，是保障读者知情权的关键技术[18]。因此，图书馆查询系统一般在图书馆网站主页布局中占据重要位置，以便读者使用；而一项有效的数据智能查询技术是影响查询质量的关键。

2.2 结合关联规则的数据智能查询优化技术

随着图书馆资源建设数字化趋势的增强，以及由“重藏轻用”向“以用为主”的观念转变，图书馆数字资源的构成中除本地资源外，网络数据库资源日益增多，读者出于节约时间成本的需求，对各类数据的一站式检索服务提出更迫切的需要。

在用户体验视角下，依据MapReduce编程模型，即一种分布式计算函数，为数字图书馆数据处理提供一个易于实现的分布式数据处理模型。MapReduce由Map函数和Reduce函数组成。Map函数通过接受一个键值对，随机产生一组中间键值对。MapReduce框架会将Map函数产生的中间键值对中的相同值输送给一个Reduce函数来处理。Reduce函数接受一个键或一组相关的值，将这组值进行合并产生一组规模更小的值。

在MapReduce编程模型的技术上，运用关联规则提取方法用于数字图书馆数据查询，可提高数字图书馆资源的检索精确度，缩短响应时间。用户体验视角将数据处理过程看作一种数据流图并进行处理，通过数据流计算框架将用户体验视角下的查询任务分解成若干子任务，将若干存在依赖关系的任务进行拆分并重新整合，产生一个DAG任务，以降低计算开销。

在用户体验视角下，对数字图书馆数据的关联规则进行提取，依据关系数据库的基本思想对关联规则提取计算进行描述。查询请求模块主要对生成阅读量较大的图书历史记录按照读者翻阅时间排序，通过扫描具有相同“前驱图书”和“后继图书”的读者数量，根据阅读量较大的图书集读者总数，确定关联图书支持度；根据读者数量和具有相同“前驱图书”的读者阅读量，最终确定关联图书的置信度；根据图书关联挖掘的要求，建立事务数据表和关联规则表。其中，事务数据表用来存储数字图书馆的相关信息数据，包括读者证件号、图书编号、图书阅读量、图书借阅量和阅读时间等；关联规则表主要存储的是对事务数据库进行多次关联挖掘后，形成的具有强关联性的规则，主要包括前驱图书编号、后继图书编号、关联规则权重、支持度和置信度等重要数据。

假设数字图书馆数据库中记录图书的集合用T=

$\{T_1, T_2, \cdots, T_n\}$进行描述，其中$n$描述总记录个数；将待查询字段构成一个集合，用$I=\{I_1, I_2, \cdots, I_m\}$进行描述，其中$m$代表待查询记录数量。每条记录$T_i$中均含有$I$的一个子集。假设$X$、$Y$代表$I$的子集，同时$X \cap Y=\varnothing$，当$T_i$中不仅含有$X$，还含有$Y$时，则称$X \geqslant Y$在$T_i$中成立。若$X \geqslant Y$在$T$的$s$条记录成立，则称$X \geqslant Y$在$T$中的支持度为$s$，即公式（1），其可信度$c$的计算见公式（2）所示。

$$s=\frac{|\{T_i|T_i\text{中含有}X, Y\}|}{|T|}\times 100\% \quad (1)$$

$$c=\frac{|\{T_i|T\text{中含有}X, Y\}|}{\{T_i|T\text{中含有}X\}}\times 100\% \quad (2)$$

若支持度与可信度均高于用户设定阈值，则$X \geqslant Y$，即为T的一个关联规则。

分析上述过程可知，关联规则提取可通过查找支持度超过阈值的数据项集合；以及从数据项集合中挑选出可信度高于阈值的数据项，将获取的数据项集合看作检索的关联规则两个过程实现[19]。

结合数字图书馆查询特征，本文在分析图书馆数据基础上，将title（图书题目）、h1-h6（各级标题的文本标签）、强调类标记看做统计对象。针对检索关键数据，利用其在不同标记信息下出现的频率和权值对其和相关图书的关联程度进行加权处理。

读者使用数字图书借阅关联规则数据库进行智能查询是一种循环递归查询过程，其具体查询步骤包括：（1）使用初始“前驱图书”在图书借阅关联规则数据库中查询“后继图书”；（2）若能查询出结果，则在数据库中继续智能查询，至查询不出任何结果为止。

在采用关联规则对数字图书馆中的数据进行查询时，引入组关键词的概念。组关键词是在每次检索和查询时挖掘出的一组查询词，第一组关键数据即已知关键数据库。将通过关联规则法获取的每组关键数据看作下一轮循环时的查询关键数据进行数据查询[20]。第i组关键词的出现频率加权计算公式为（3）。

$$f_{key_{ij}}=w_{title}f_{title}+w_h f_{h_{ij}}+w_{emp}f_{emp_{ij}} \quad (3)$$

式（3）中，w_{title}、w_h、w_{emp}分别用于描述title、h1-h6、强调类标记的权重。

为有效判断不同标记在数字图书馆数据中的关键程度，需满足条件：$w_{title}>w_h>w_{emp}$，同时$w_{title}+w_h+w_{emp}=1$。f_{title}、$f_{h_{ij}}$、$f_{emp_{ij}}$分别用于描述在第j个页面中，第i次查询后获取的一组关键数据在以上标记中出现的频率和，即公式（4）。

$$P_j=\sum_{i=0} W_{key_i} f_{key_{ij}} \quad (4)$$

P_j用于描述数字图书馆第j个页面的关联程度，W_{key_i}用于描述第i次搜索获取的数字图书馆关键数据权重，$f_{key_{ij}}$用于描述第i次搜索中第k个关键数据在该数字图书馆数据中依据标记信息的加权频率。

针对数字图书馆中的不同网页，依据标记信息与关键词求得数字图书馆不同网页信息关联度，并用各标记信息出现次数加权与各关键词权重积累加和进行表示，如公式（5）和（6）。

$$awconf=\frac{P(\text{A\&B})}{P(\text{A})} \quad (5)$$

$$W_{key}=\frac{\sum W_{key_i}}{n_{relation}}\times awconf \quad (6)$$

$\sum W_{key_i}$用于描述数字图书馆数据中所有关键数据的权重之和，$n_{relation}$用于描述关联规则数字图书馆数据中项目数量，$awconf$用于描述得到的关联规则置信度。

针对用户检索的关键词，按照权重值从大到小的顺序给出查询结果，从而实现对数字图书馆数据的查询。

3 实验结果与分析

本文将待查询的数字图书馆数据信息进行重组，使其存在连续读取特性，并依据该连续性对信息进行保存，以增强读取速率。数字图书馆读者对信息进行查询时，通常需要跨库检索才能获得较全面的检索结果集。

表1为读者查询需求表，其中Q_n代表读者查询信息，D_n代表数字图书馆中的数据库，R_n代表数据库记录，“1”代表提出查询申请，“0”代表未提出查询申请。

表 1 读者查询需求表

数据库		查询申请			
		Q_1	Q_2	Q_3	Q_4
D_1	R_1	1	1	0	1
	R_2	0	0	1	0
	R_3	0	0	0	1
D_2	R_4	0	0	0	0
	R_5	1	1	0	0
	R_6	1	1	1	1

由表1可知，所有请求均需依次对数据库D_1和D_2进行一次信息查询操作，即一个申请需进行2次查询，平均查询次数是2。

利用数据重组令查询具有连续读取的特性，分析表2可知，查询请求Q_1—Q_3在重组后只需1次查询操作，查询请求Q_4需2次操作，对数据进行重组后，平均1.25次即可完成一个查询操作，说明连续读取特性能提高指定文献资源的查询效率。

表 2 查询信息重组表

数据库		查询申请			
		Q_1	Q_2	Q_3	Q_4
D_1	R_1	0	0	1	0
	R_2	0	0	0	1
	R_3	0	0	0	0
D_2	R_4	0	1	0	1
	R_5	1	1	0	0
	R_6	1	1	0	1

针对数字图书馆中三种不同的查询类型，分别采用关联规则方法、K近邻连接方法和K-Means方法进行查询，查询响应时间比较结果如表3所示。

表 3 三种方法查询响应时间比较结果

查询类型	实 例	最大记录数/个	关联规则方法响应时间/秒	K近邻连接方法响应时间/秒	K-Means方法响应时间/秒
单一关键词	检索1	2 314	14.20	29.30	35.40
关键词组合	检索2	1 789	0.86	8.25	10.16
主题	检索3	2 816	0.54	9.36	12.87

K近邻连接方法时查询单一关键词的响应时间最高为29.30秒，而K-Means方法响应时间达35.40秒，相比K近邻连接方法增加6.10秒；而本文所提出的关联规则方法的查询响应时间为14.20秒，相比K近邻连接方法、K-Means方法分别降低15.10秒、21.20秒，且关键词组合、主题查询响应时间也均有一定程度降低，表明采用本文提出的关联规则方法具有一定优势。

在对文献资源的检索结果进行评价的过程中，查全率与查准率一般被看作两个相互矛盾的衡量指标。在一定范围内，查准率越高，查全率越低，而在查全率逐渐提高的情况下，会影响查询效率。

在进行数字图书馆文献资源检索时，针对各个查全率水平，利用公式（7）对三种查询方法的查准率进行处理。

$$\overline{P}(\mathrm{r})=\sum_{i=1}^{Nq}\frac{P_i(\mathrm{r})}{Nq} \tag{7}$$

其中，$\overline{P}(\mathrm{r})$用于描述查全率是r时的平均查准率，Nq用于描述查询总次数，$P_i(\mathrm{r})$用于描述查全率是r时第i个查询的查准率。

针对不同查询类型，对关联规则方法、K近邻连接方法和K-Means方法在不同查全率水平级下的查准率进行统计分析（见表4）。在保证相同查全率下，关联规则方法的查准率相比K近邻连接方法和K-Means方法有明显提升，且随着查全率增加，噪声数据被有效抑制，关联规则方法的查准率保持在较高水平，整体性能较强。

表 4 不同查全率水平级下三种方法查准率平均值比较结果

%

查全率	关联规则方法	K近邻连接方法	K-Means方法
10	83.2	80.7	81.5
20	77.8	62.2	36.7
30	69.3	50.4	52.8
40	53.8	32.6	41.6
50	51.5	28.6	32.5
60	49.9	25.1	29.3
70	39.3	23.2	20.5
80	31.2	17.9	16.6
90	26.7	14.3	13.8

4 结语

针对传统的查询方法一直存在查询效率低的问题，在用户体验视角下，提出基于关联规则法的数字图书馆数据智能查询优化方法，并将其应用于数字图书馆文献资源检索查询。通过关联规则挖掘技术实现数据的智能查询，结果表明，将该方法应用于数字图书馆数据的查询能有效提高查询效率和精度。

参考文献

[1] 张凯,郭健栖.图书馆主题大数据调查及前瞻性构想——基于百度

指数的分析[J].中国图书馆学报,2016,42(6):51-66.

[2] 完颜邓邓,高峰.澳大利亚高校图书馆研究数据管理服务的调查分析[J].图书与情报,2015(3):71-76.

[3] DUROCHER S,SHAH R,SKALA M,et al.Linear-Space data structures for range frequency queries on arrays and trees[J]. Algorithmica,2016,74(1):1-23.

[4] 李善青,赵辉,宋立荣.基于大数据挖掘的科技项目查重模型研究[J].图书馆论坛,2014,34(2):78-83.

[5] LIM J,PARK S,SEO K,et al.Reverse k-nearest neighbor query processing method for continuous query processing in bigdata environments[J].Lancet,2014,14(10):454-462.

[6] 陈涛,夏翠娟,刘炜,等.关联数据的可视化技术研究与实现[J].图书情报工作,2015,59(17):113-119.

[7] 周屹,杨泽雪.空间数据库中的线段K近邻查询研究[J].计算机工程与应用,2015,51(18):131-134.

[8] ABITEBOUL S,BOURHIS P,VIANU V.Highly expressive query languages for unordered data trees[J].Theory of Computing Systems,2015,57(4):927-966.

[9] WEBER G M.Federated queries of clinical data repositories: scaling to a national network[J].Journal of Biomedical Informatics, 2015,55:231-236.

[10] 赵丹.基于DBS的园林植物数据查询分析系统的设计与研究[J].电子设计工程,2016,24(18):31-33.

[11] 喻宜,吕志来,齐国印.分布式海量时序数据管理平台研究[J].电力系统保护与控制, 2016,44(17):165-170.

[12] 茹文,忻展红.图书馆借阅数据分类信息的关联性研究[J].北京邮电大学学报(社会科学版),2016,18(1):14-19.

[13] COMMEAN P K,RATHMELL J M,CLARK K W, et al.A query tool for investigator access to the data and images of the national lung screening trial[J].Journal of Digital Imaging,2015, 28(4):439-447.

[14] 宋爱波,万雨桐,贡欢,等.海量多维数据的存储与查询研究[J].计算机工程与应用,2016,52(13):25-31.

[15] 杨良斌.数据挖掘领域研究现状与趋势的可视化分析[J].图书情报工作,2015(S2):142-147.

[16] CAMPBELL W S,PEDERSEN J,MCCLAY J C,et al.An alternative database approach for management of SNOMED CT and improved patient data queries[J].Journal of Biomedical Informatics, 2015, 57(C):350-357.

[17] 李建荣.基于数据挖掘的移动用户个性化推荐系统研究与设计[J].现代电子技术,2016(22):59-63.

[18] 张红莉.大数据环境下文献传递营销工作的思考[J].图书馆工作与研究,2014(12):126-128.

[19] 谷权峰.面向中美百万册数字图书馆的图书资源管理系统[J].计算机工程与科学,2010,32(4):146-150.

[20] PEUTE L W,DE KEIZER N F,JASPERS M W.The value of retrospective and concurrent think aloud in formative usability testing of a physician data query tool[J].Journal of Biomedical Informatics,2015,55:1-10.

作者简介

王炼，女，1977年生，硕士，馆员，研究方向：图书馆信息服务，E-mail：wanglian1399@163.com。
牛庆银，男，1970年生，硕士，教授，硕士生导师，研究方向：数学与计算机应用。

Research on Digital Intelligent Query Optimization of Digital Library Based on User Experience

WANG Lian, NIU QingYin
(The Academy of Armored Force Engineering, Beijing 100072, China)

Abstract: According to the data of traditional digital library intelligent query technology has long response time, low efficiency and poor accuracy of query problem, proposed user experience research on query optimization in data from the perspective of intelligent digital library. Firstly, this paper introduces the development of intelligent digital library data query technology and the meaning of existence; secondly, using the method of data association rules in user experience from the perspective of mining and extraction of digital library data, and based on the basic idea of the relational database to extract association rules to calculate description; finally, according to the weight values are given in ascending order from big to small the query results, so as to realize intelligent query optimization. Experimental results show that the proposed method has high query efficiency and better query accuracy than traditional methods.

Keywords: User Experience Perspective; Association Rules; Digital Library; Intelligent Query

（收稿日期：2017-05-09）