文章编号：2095-2163(2019)02-0148-04 中图分类号：TP 391 文献标志码：A

云存储中心多源文本主题融合模型研究

谌裕勇
（广东工业大学 华立学院，广州 511325）

摘 要：为了提高云存储中心多源文本主题信息资源的开发能力和调度能力，提高云存储中心多源文本主题检索效率，提出一种基于关联规则挖掘的云存储中心多源文本主题融合模型。构建云存储中心多源文本主题分布大数据模型，采用相空间重构方法进行大数据的特征分布式重建和融合聚类处理，提取云存储中心多源文本主题信息的关联规则特征量，采用相关性检测技术进行信息集成滤波，结合模糊聚类方法进行云存储中心多源文本主题特征分类处理，根据分类结果实现信息融合。仿真结果表明，采用该方法进行云存储中心多源文本主题信息融合的特征分类性较好，提高了云存储中心进行数据检索的准确率，数据召回性和检索效率等性能指标表现较好。

关键词：云存储中心；多元文本；主题；融合；聚类

Research on multi-source text topic fusion model in cloud storage center

CHEN Yuyong
（Huali College, Guangdong University of Technology, Guangzhou 511325, China）

【Abstract】 In order to improve the development ability and scheduling ability of cloud storage center multi-source text topic information resource and improve the retrieval efficiency of cloud storage center multi-source text topic, a multi-source text topic fusion model based on association rule mining is proposed. The big data model of multi-source text topic distribution in cloud storage center is constructed, then the feature distributed reconstruction and fusion clustering processing are carried out by using phase space reconstruction method, and the association rules feature quantity of multi-source text topic information in cloud storage center is extracted. The correlation detection technique is used for information integration filtering and the fuzzy clustering method is used to classify multi-source text features in cloud storage center. The information fusion is realized based on the classification results. The simulation results show that this method has better feature classification for cloud storage center multi-source text subject information fusion, and improves the accuracy of data retrieval in cloud storage center, data recall and retrieval efficiency.

【Key words】 cloud storage center; multi-source text; topic; fusion; clustering

0 引 言

随着云存储和云计算技术的快速发展，对云存储中心多源文本主题信息开发成为未来云存储和数据库建设的关键技术。随着数据资源规模的不断扩大，大量的云存储资源分布在云集成数据库系统中，通过云组合服务和大数据管理的模式，实现云存储资源共享，为了提高云存储系统的数据调度性能，需要对云存储中心多源文本主题进行融合处理，结合多媒体集成学习方法进行资源信息优化调度，提高主题信息资源的检索能力[1]。云存储中心多源文本主题信息表现为一组大数据，采用关联规则挖掘方法进行云存储中心多源文本主题资源信息整合，促进云存储中心多源文本主题信息检索效率的提升。

传统方法中，对云存储中心多源文本主题融合研究采用层次数据聚类方法，结合资源的聚类处理技术[2]，提取云存储中心多源文本主题信息的规则性关联特征量，采用向量量化编码方法实现计算资源的融合调度，取得了较好的调度效果[3]。文献[4]中，提出一种基于混合差分并行调度的云存储中心多源文本主题资源信息的整合算法，首先构建云存储环境下多媒体集成学习资源信息分布的数据结构和网格结构模型，采用资源信息流的样本聚类分析方法进行云存储环境下资源信息的属性归类处理，提高资源整合能力，但该方法计算开销较大，对云存储中心多源文本主题融合的实时性不好。针对上述问题，本文提出一种基于关联规则挖掘的云存储中心多源文本主题融合模型。首先构建云存储中心多源文本主题分布大数据模型，采用相空间重构方法进行大数据的特征分布式重建和融合聚类处理，提取云存储中心多源文本主题信息的关联规则特征量，然后采用相关性检测技术进行信息集成滤波，结合模糊聚类方法进行云存储中心多源文本主

作者简介：谌裕勇（1979-），男，硕士，讲师，主要研究方向：数据挖掘、数据分析、机器学习等。

收稿日期：2018-12-20

题特征分类处理，根据分类结果实现信息融合。最后进行仿真实验分析，展示了本文方法在提高云存储中心多源文本主题融合能力方面的优越性能。

1　云存储中心多源文本主题信息采样及特征分析

1.1　云存储中心多源文本主题信息资源采样

为了实现云存储中心多源文本主题融合模型的优化设计，采用统计分析方法进行云存储中心多源文本主题信息资源采集，对采集的云存储中心多源文本主题信息资源进行信息重构，构建云存储中心多源文本主题信息资源的特征信息流，采用线性回归分析模型和网格划分技术构建云存储中心多源文本主题信息资源的分布式结构模型[5]，用 x_{n-i} 表示云存储中心多源文本主题信息资源属性集的模糊分布自相关量，η_{n-j} 表示云存储中心多源文本主题信息资源属性特征向量的有限分布集，则云存储中心多源文本主题信息资源信息流重组模型表示为：

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-i} + \sum_{j=0}^{M_{MA}} b_j \eta_{n-j}, \tag{1}$$

其中，a_0 为统计数据的采样幅值，b_j 为云存储中心多源文本主题信息资源的最优关联规则分布属性。采用分段样本统计分析方法进行云存储中心多源文本主题信息资源的联合关联互信息特征分析[6]，云存储中心多源文本主题信息资源的标量时间序列为 $x(t)$，$t = 0,1,\cdots,n-1$，结合模糊信息特征分析方法，采用相关的数据分析和信息采集技术，分析反映主体资源信息的相关性指标，得到主题信息分布的有限集合为：

$$X = \{x_1, x_2, \cdots, x_n\} \subset R^s, \tag{2}$$

结合融合数据聚类模型，得到云存储中心多源文本主题信息资源的关联相关性特征提取结果为：

$$C(l) = \sum_{j=1}^{k} \sum_{k=1}^{n_j} (\| x_k^j - A_j(L) \|)^2, \tag{3}$$

在大数据处理环境下，云存储中心汇聚了大量的多源信息资源[7]，在模糊聚类中心，得到云存储中心多源文本主题特征的二元语义特征映射描述为：

$$\theta: S \to S \times [-0.5, 0.5], \tag{4}$$

$$\theta(s_i) = (s_i, 0), s_i \in S. \tag{5}$$

设实数 $\beta \in [0,T]$ 为相似度，将关联指标参量加载到信息处理模块，采用关联规则挖掘方法[8]，实现信息采样和特征提取。

1.2　相空间重构与特征提取

构建云存储中心多源文本主题分布大数据模型，采用相空间重构方法进行大数据的特征分布式重建，当多源文本主题信息分布聚类中心的相对距离满足 $\| C(l) - C(l-1) \| < \xi$，得到云存储中心多源文本主题信息资源的聚类迭代式为：

$$A_j(L+1) = \frac{1}{n_j} \sum_{i=1}^{k} X_i^j \quad j = 1,2,\ldots,k, \tag{6}$$

设 (s_k, a_k) 和 (s_1, a_1) 为云存储中心多源文本主题信息资源融合节点之间的模糊贴近度矢量，采用相空间重构方法进行特征重组[9]，相空间重构模型为：

$$\max Z = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij} c_{ij}, \tag{7}$$

$$st = \sum_{j=1}^{m} x_{ij}, \tag{8}$$

$$st = \sum_{i=1}^{m} x_{ij}, \tag{9}$$

$$x_{ij} = 1, \tag{10}$$

$$st = 0, \text{or}, 1, \tag{11}$$

其中，$x_{ij} = 1$ 表示云存储中心多源文本主题信息资源融合的回归系数，提取云存储中心多源文本主题信息的关联规则特征量，得到云存储中心多源文本主题信息资源属性分类评估约束因子为：

$$ind(P) = \left\{ \begin{matrix} (x,y) \in U^2 \mid a(x) = a(y), \\ \forall a \in P \end{matrix} \right\}, \tag{12}$$

计算云存储中心多源文本主题信息资源的模糊关联度特征，得到信息融合的检测统计分析模型表达式为：

$$TTD = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k + \delta, \tag{13}$$

其中，TTD 表示关联规则集，在数据融合的相空间中，得到云存储中心多源文本主题信息大数据挖掘后输出为：

$$X_p(u) = s_c(t) e^{j2\pi f_0 t} = \frac{1}{\sqrt{T}} rect\left(\frac{t}{T}\right) e^{j2\pi(f_0 t + Kt^2)/2}. \tag{14}$$

其中，$s_c(t)$ 表示多源文本主题信息的并行调度集，由此提取云存储中心多源文本主题信息的关联规则特征量，根据特征提取结果进行信息融合聚类处理。

2　云存储中心多源文本主题融合模型优化

2.1　关联规则挖掘模型

在上述构建了云存储中心多源文本主题分布大数据模型和采用相空间重构方法进行大数据的特征分布式重建处理的基础上，进行云存储中心多源文本

主题融合模型的优化设计,本文提出一种基于关联规则挖掘的云存储中心多源文本主题融合模型,提取云存储中心多源文本主题信息的关联规则特征量,采用多特征的静态拟合方法进行信息流重组[10],则资源分布集合的优先级属性可以表示为 $P(n_i)=\{p_k \mid pr_{kj}=1,k=1,2,\cdots,m\}$。采用并行调度的关联规则挖掘方法进行云存储中心多源文本主题大数据挖掘,得到资源信息流的分组关系为:

$$Q^w=\sum_{k\in R_w}F_k^w,\ w\in W, \tag{15}$$

$$V_a=\sum_{w\in W}\sum_{k\in R_w}\delta_{ak}^w F_k^w,\ a\in A, \tag{16}$$

$$F_k^w\geqslant 0,\ k\in R_w, w\in W, \tag{17}$$

采用多元信息融合方法,进行云存储中心多源文本主题信息流的自适应分配,得到资源信息流为:

$$flow_k=\{n_1,n_2,\cdots,n_q\},q\in N, \tag{18}$$

其中,q 表示多个节点重组下的云存储中心多源文本主题信息流集合,n_q 表示负载,云存储中心多源文本主题信息关联规则挖掘输出为:

$$u_i=\frac{1}{N}\sum_{i=1}^{N}u_i=\frac{1}{MN}\sum_{m=1}^{M}\sum_{i=1}^{N}x_{mi}. \tag{19}$$

根据关联规则挖掘结果采用分组样本回归分析方法进行主题信息融合。

2.2 信息融合滤波

给定云存储中心多源文本主题信息资源融合的相关因子,分别是 $a_1,a_2,\cdots,a_k$,在云存储中心多源文本主题信息资源分布结构模型下,以 β 为边界条件,得到云存储中心多源文本主题信息资源融合的拓展外延 M^β:

$$M^\beta=\{x\mid x\in M,|f(x)\cap Y|/|Y|\geqslant\beta,0\leqslant\alpha\leqslant\beta\leqslant 1\}, \tag{20}$$

采用 $U(t)=\sum_{M\in E}P[M]$ 表示云存储中心多源文本主题信息资源融合主体的信任度属性状态集合,$A_{st}\subseteq P\times T$,构建云存储中心多源文本主题信息资源模糊指派调度集合,采用相关性检测技术进行信息集成滤波,结合模糊聚类方法进行云存储中心多源文本主题特征分类处理,C 均值聚类模型为:

$$L_\xi=\begin{cases}|f(x)-y|-\xi, & |f(x)-y|\geqslant\xi,\\ 0, & |f(x)-y|<\xi.\end{cases} \tag{21}$$

由此得到资源融合的模糊函数为:

$$f(x)=\sum_{i=1}^{l}(a_i+a_i^*)k(x-x_i)+b \tag{22}$$

计算云存储中心多源文本主题信息资源的模糊关联度特征,采用 C 均值聚类方法进行大数据融合处理,优化的模型可表达为:

$$\min F=R^2+A\sum_i\xi_i$$

$$\text{s.t.: } \|\phi(x_i)-o\|^2\leqslant R^2+\xi_i \text{ and } \xi_i\geqslant 0,i=1,2,\cdots, \tag{23}$$

$$\max\sum_i\alpha_i K(x_i,x_i)-\sum_i\sum_j\alpha_i\alpha_j K(x_i,x_j)$$

$$\text{s.t.: }\sum_i\alpha_i=1 \text{ and } 0\leqslant\alpha_i\leqslant A,i=1,2,\cdots, \tag{24}$$

由于 $\sum_i\alpha_i=1,K(x_i,x_i)=1$,云存储中心多源文本主题特征分类的优化模型为:

$$\max(1-\sum_i\sum_j\alpha_i\alpha_j K(x_i,x_j))$$

$$\text{s.t.: }\sum_i\alpha_i=1 \text{ and } 0\leqslant\alpha_i\leqslant A,\ i=1,2,\cdots. \tag{25}$$

其中,$K(x_i,x_j)=e^{\frac{-\|x_i-x_j\|^2}{2\sigma^2}}$,$\sigma$ 值越小,收敛性越好,可见,设计的资源融合模型是稳健收敛的。

3 仿真实验分析

为了测试本文方法在实现云存储中心多源文本主题融合和检索中的应用性能,进行仿真实验,实验中分析软件为 Excel 2007 和 SPSS19.0,相关参数为:$Q=200$,$c_1=30$,$c_2=10$,$c_r=2$,$\mu_1=\mu_2=0.01$,$\rho_1=\rho_2=0.01$,$\delta=0.8$,云存储中心多源文本主题分布的相关性统计分析结果见表 1。

表 1 云存储中心多源文本主题分布的相关性统计分析结果

Tab. 1 The correlation statistical analysis results of multi-source text topic distribution in cloud storage center

相关系数	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1	0.432	0.654	0.456	0.654	0.464
x_2	0.432	1	0.232	0.478	0.454	0.333
x_3	0.323	0.334	1	0.598	0.443	0.524
x_4	0.445	0.654	0.221	1	0.622	0.564
x_5	0.587	0.321	0.243	0.454	1	0.564
x_6	0.521	0.343	0.834	0.432	0.634	1

根据表 1 的云存储中心多源文本主题分布相关性检测结果进行关联规则挖掘,得到挖掘结果如图 1 所示。

分析图 1 得知,本文方法能准确挖掘云存储中心多源文本主题信息关联规则项,从而提高信息融合能力,测试不同方法进行文本主题信息融合处理后的召回率,得到对比结果如图 2 所示。分析图 2 得知,采用本文方法进行云存储中心多源文本主题信息融合的特征分类性较好,提高了云存储中心进行数据检索的准确率,数据召回性较好。

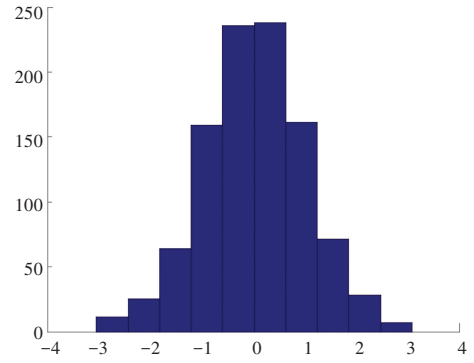

图 1　云存储中心多源文本主题信息关联规则挖掘结果

Fig. 1　Mining results of multi－source text topic information association rules in cloud storage center

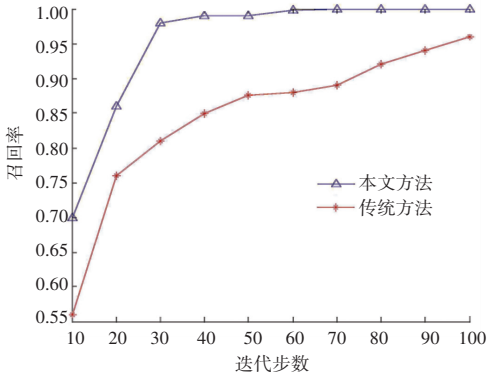

图 2　召回率对比

Fig. 2　Recall rate comparison

4　结束语

结合多媒体集成学习方法进行资源信息优化调度，提高主题信息资源的检索能力，本文提出一种基于关联规则挖掘的云存储中心多源文本主题融合模型。构建云存储中心多源文本主题分布大数据模型，采用相空间重构方法进行大数据的特征分布式重建和融合聚类处理，提取云存储中心多源文本主题信息的关联规则特征量，采用相关性检测技术进行信息集成滤波，结合模糊聚类方法进行云存储中心多源文本主题特征分类处理，根据分类结果实现信息融合。研究得知，采用本文方法进行云存储中心多源文本主题信息融合的特征分类性较好，提高了云存储中心进行数据检索的准确率，数据召回率较高。

参考文献

[1] 廖大强. 面向多目标的云计算资源调度算法[J]. 计算机系统应用, 2016, 25(2):180-189.

[2] 徐立洋, 黄瑞章, 陈艳平, 等. 基于狄利克雷多项分配模型的多源文本主题挖掘模型[J]. 计算机应用, 2018, 38(11): 3094-3099,3104.

[3] 高悦, 王文贤, 杨淑贤. 一种基于狄利克雷过程混合模型的文本聚类算法[J]. 信息网络安全, 2015(11):60-65.

[4] 易利容,王绍宇,殷丽丽,等. 基于多变量 LSTM 的工业传感器时序数据预测[J]. 智能计算机与应用,2018,8(5):13-16.

[5] 王伟,胡长武,郭栋,等. 一种面向云构软件的云操作系统[J]. 计算机科学,2017,44(11):33-40.

[6] FERCOQ O, RICHTÁRIK P. Accelerated, parallel and proximal coordinate descent[J]. SIAM Journal on Optimization, 2015, 25(4):1997-2023.

[7] 刘测,韩家新. 面向新闻文本的分类方法的比较研究[J]. 智能计算机与应用,2018,8(5):38-41.

[8] 王树恒,吐尔根·依布拉音,卡哈尔江·阿比的热西提,等. 基于 BLSTM 的维吾尔语文本情感分析[J]. 计算机工程与设计,2017,38(10):2879-2886.

[9] 郑娜,王加阳. 不完备序信息系统的证据特征及属性约简[J]. 计算机工程与应用, 2018, 54(21): 43-47,200.

[10]李涛,王次臣,李华康. 知识图谱的发展与构建[J]. 南京理工大学学报(自然科学版),2017,41(1):22-34.