

海量信息检索挖掘及视觉三维展现方法仿真

孙钟德^{1,3}, 蒋 薇², 刘巧根², 顾承华²

(1 扬州市气象局, 江苏 扬州 225009; 2 江苏省气象局, 江苏 南京 210008;

3 扬州大学 工学院, 江苏 扬州 225009)

摘 要: 为了提高网络中海量信息检索挖掘的效率和准确率, 提出一种基于分布式处理技术的结构化的海量信息检索挖掘及视觉三维展现方法. 利用分布式检索系统的 MapReduce 引擎, 能够实现海量图像信息相似性度量的检索挖掘; 利用 MapReduce 检索挖掘模型中的映射和约简操作过程, 能够对图像相似性的度量进行运算, 并对特征相识度相同的图像进行合并; 利用 SIFT、HOG 和颜色矩这 3 个特征参数构成的图像特征向量, 构建与图像特征对应的视觉和语义词汇, 实现检索结果的三维可视化展示. 仿真结果证明, 改进算法进行高海量信息检索挖掘耗时短、准确率高.

关键词: 结构化; 海量信息; 视觉展示

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2016)09—0137—04

Structured Huge Amounts of Information Retrieval Mining Technology and Visual Display of 3D Technology

SUN Zhong-de^{1,3}, JIANG Wei², LIU Qiao-gen², GU Cheng-hua²

(1 Yangzhou Meteorological Administration, Yangzhou 225009, China;

2 Jiangsu Meteorological Administration, Nanjing 210008, China;

3 College of Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract: In order to improve the efficiency and accuracy of massive information retrieval and mining in the network, a new method based on distributed processing technology is proposed. Using distributed retrieval system of MapReduce engine, to achieve mass image information measure of searching for similar; using the MapReduce retrieval by mining model mapping and reduction operation process, can the image similarity metric operations, and the feature met with image merging; using sift and hog and color moments of the three characteristic parameters of image feature vector construct 3D visualization and image features corresponding to visual and semantic vocabulary, achieve the retrieval result display. The simulation results show that the improved algorithm is time-consuming, high accuracy and high accuracy.

Key words: structured; huge amounts of information; visual display

1 引言

网络环境下的海量信息资源主要为语义复杂、表现形式多样、结构各异、以数字资源为主体的信息资源^[1-2], 在实际的检索挖掘过程中需要根据信息资源之间的关联和规律进行可视化展示^[3], 为人们提供网络信息资源内容的共享^[4]. 因此, 如何对网络中结构化的海量信息资源进行准确检索和挖掘, 获取目标信

息资源之间的关联模式, 并对检索挖掘结果进行可视化展示^[5], 逐渐成为相关领域的研究重点^[6].

目前, 网络海量信息检索挖掘与视觉展示主要利用并行关联规则^[7]、Apriori 算法^[8]、数据特征提取方法进行. 由于网络数据库的结构存在着异构性、动态性和复杂性的特点, 因此传统算法进行网络中结构化的海量信息检测挖掘与视觉展示的过程中, 会存在着巨大的冗余计算, 造成检索挖掘的准确率

和效率不高,难以为人们提供准确的视觉展示^[9-10].

本文对此提出一种基于分布式处理技术的网络结构化的海量信息检索及视觉展示三维方法.

2 海量信息检索挖掘与视觉展示的原理

2.1 海量信息的特征提取

在网络中的数据库中,海量信息资源之间存在着很大的关联性,目标信息资源与其他信息资源之间的区别不明显,难以将直接目标信息资源检索出来.以网络数据库中的图像检索挖掘为例,为了对网络中海量的图像信息资源进行准确检索挖掘,需要将图像映射到频域空间.设置任意图像 I 的大小为 $M \times N$,其矩形网格 $S = \{(m, n) : 1 \leq m \leq M, 1 \leq n \leq N\}$ 的二维阵列为 $\{f(m, n) : (m, n) \in S\}$,其中, $(m, n) \in S$ 为网格中像素的坐标, $f(m, n)$ 为像素灰度.

利用下述公式能够对图像 I 进行离散傅里叶变换:

$$F(u, v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) e^{-j(2\pi/N)(mu+nv)} \quad (1)$$

式中, $1 \leq u \leq M, 1 \leq v \leq N, j = \sqrt{-1}$.

利用下述公式能够对图像在空间频域 (u, v) 能量进行计算:

$$E_{u,v} = |F(u, v)|^2 \quad (2)$$

利用下述公式能够对反映图像能量分布的频域熵进行计算:

$$T = - \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |F(u, v)|^2 \log |F(u, v)|^2 \quad (3)$$

T 越大,表明图像的细节越丰富,特征提取的效果就越好.

在网络数据库中海量图像信息资源检索挖掘的过程中,设置检索挖掘区域的阈值为 ϵ ,大于 ϵ 的图像为检索挖掘的目标图像,小于 ϵ 的图像为非目标图像.在对某一幅图像进行检索挖掘时,需要分别计算出目标区域与背景区域的空间频率能量 $E(O)$ 和 $E(B)$,则它们相对应的频率熵分别为 $T(O)$ 和 $T(B)$,利用下述公式能够对图像熵函数进行描述:

$$T(I) = T(O) + T(B) \quad (4)$$

当上述函数获得最大值 $T_{\max}$ 时,即可实现图像特征的最佳提取.经过特征提取后,能够得到海量图像资源中图像的特征,为图像资源的检索挖掘提供了准确依据.

2.2 海量信息检索结果的可视化展示

在网络数据库中,由于图像的颜色、形状、纹理等视觉特征存在着较强的主观性,难以建立统一的

标准,同时,人类理解的图像特征存在着较强的主观性,造成检索挖掘的效果不够理想.

设置检索挖掘的图像函数为 $f(x, y)$,经过傅里叶变换后得到的函数为 $F(u, v)$,对图像的傅里叶功率进行计算,并用计算结果表示图像的纹理特征,傅里叶功率谱能够准确描述图像中纹理特征的分布.设置

$$F(u, v) = R(u, v) + jI(u, v) \quad (5)$$

则图像中功率谱的幅度能够用下述公式进行计算:

$$|F(u, v)| = \sqrt{R^2(u, v) + I^2(u, v)} \quad (6)$$

图像中功率谱的相位能够用下述公式进行计算:

$$\bar{\omega}(u, v) = \arctg \frac{I(u, v)}{R(u, v)} \quad (7)$$

图像中功率谱的能量能够用下述公式进行计算:

$$E(u, v) = |F(u, v)|^2 = R^2(u, v) + I^2(u, v) \quad (8)$$

图像频谱能够反映图像的灰度分布,是图像的重要特征.因此,根据图像的频谱特征能够完成海量图像信息检索挖掘结果的可视化展示.

3 基于分布式处理技术的信息检索及视觉展示三维方法

3.1 分布式海量信息检索挖掘系统

Hadoop 能够在复杂网络环境下对异构性数据库海量信息中进行快速、并行检索挖掘目标信息,利用其分布式检索系统(Hadoop Distributed File System, HDFS)的 MapReduce 引擎,能够实现海量图像信息(TB 级以上)相似性度量的检索挖掘,整个检索挖掘过程产生的数据都存储于分布式系统的数据库 Hbase 中,并将若干 PC 机作为从属节点进行图像信息的分布式管理,提高了图像信息检索挖掘的效率.

HDFS 是 Hadoop 系统的核心部分,由于其具有分布式存储模式和流数据访问模式的特点,因此可以对网络中的海量信息进行快速检索挖掘. HDFS 为主/从结构形式,如图 1 所示.将其中一台 PC 设置为主服务器(即主节点),则其余的 PC 为从服务器(即从属节点).

在从属服务器中,对图像信息检索挖掘的任务进行分配、协调与存储等操作.所有待检索挖掘的数据根据分布位置的不同可以分成元数据和数据这两类,其中元数据分布在主服务器中, Hadoop 对两类数据进行交互式操作.将数据划分为多个数据的子块,然后对所

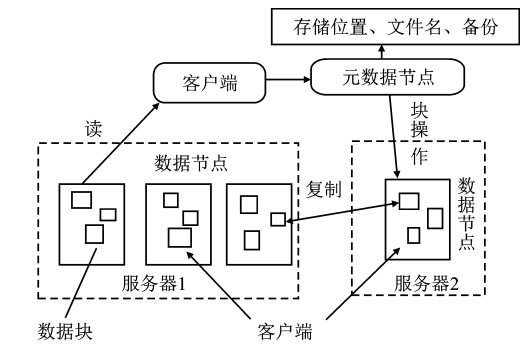


图 1 HDFS 结构形式

有数据的子块进行备份,实现数据子块的分布存储.将数据子块作为信息检索挖掘的基本单元,可以使各个子服务器之间的信息处理过程更顺畅.

HDFS 具有海量信息数据的检索挖掘能力,主要原因在于:(1)流数据的访问方式为单输入多相应的方式,使信息的各个操作过程更迅速;(2)信息检索挖掘的能力与单个子服务器的性能基本无关,无需配置高性能的 PC 机,降低了系统运行成本.

3.2 并行处理的信息检索挖掘过程

MapReduce 检索挖掘模型包含两个操作过程,分别为映射(Map)和约简(Reduce),这两个过程为并行计算过程,在海量的图像信息检索挖掘中的作用是对图像相似性的度量进行运算,并将检索挖掘结果发送给用户.映射的作用是对图像特征的相似度进行计算,简约的作用是对特征相识度相同的图像进行合并.因此,映射的操作过程是海量信息检索挖掘的关键过程.网络中海量图像信息检索挖掘的并行处理流程如图 2 所示.

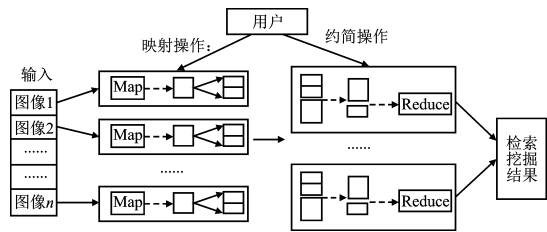


图 2 海量图像信息的并行检索挖掘流程

MapReduce 能够实现网络中海量图像信息的并行检索挖掘,对海量图像信息检索挖掘的相关定义如下:

基于 MapReduce 的结构化海量图像信息检索挖掘算法

输入: HDFS 分布储存的海量图像信息

输出: 图像信息的检索挖掘结果

Map 函数

(1)初始化参数. Key: 图像特征的位置属性. Value: 图像的特征属性;

(2)利用图像特征的相似度函数对图像特征进行计算;

(3)利用检索挖掘规则对匹配的相似度进行计算;

(4)输出中间结果.

Reduce 函数

(1)调用 Map 函数的计算结果, key: 图像特征相似度, value: 图像序列;

(2)若计算结果 key 相同,则对其 value 进行合并处理;

(3)对图像的相似度计算结果按照从大到小的顺序对约简结果进行排序;

(4)输出图像的检索挖掘结果.

3.3 三维视觉展示方法

网络中结构化的海量信息检索挖掘的关键是两幅图像相似度的判断,传统的检索方法主要依赖于图像的颜色、纹理等底层特征,但由于存在着语义鸿沟,因此传统算法无法满足海量图像信息检索挖掘的需要.为了突破语义鸿沟,需要利用视觉展示的三维技术对检索挖掘结果进行展现,视觉展示三维技术能够有效突破语义鸿沟的影响,可以准确地反映图像信息的结构信息,并将视觉展示的三维信息作为图像检索挖掘的必要条件.

视觉展示三维技术的核心是,利用 SIFT、HOG 和颜色矩这三个特征参数构成的图像特征向量,通过对图像进行平移、旋转、遮挡等操作方法,构建与图像特征对应的视觉和语义词汇,快速检索挖掘出图像中与语义存在较强关联性的视觉规则.视觉展示三维技术的示意图能够用图 3 进行描述.

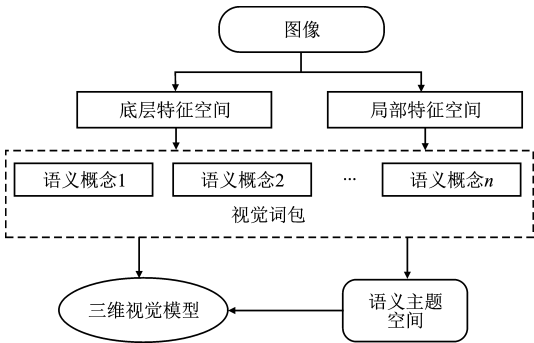


图 3 视觉展示三维技术的示意图

利用聚类算法对图像特征向量进行分类,将一个视觉单词设置为一个聚类中心,利用视觉单词的集合

描述一幅图像,利用潜在狄里克雷分配主题模型对图像与视觉单词之间的相关程度进行计算,选取相关程度较高的视觉单词作为图像的视觉展示模型。

设置图像 I , 其视觉单词的序列为 z , 视觉单词的集合为 $D = \{d_1, d_2, \dots, d_n\}$, 其中 d_i 为图像的视觉单词, n 为视觉单词的数目, 则视觉单词与图像主题之间的相似度能够用下述公式进行计算:

$$\text{Rdelevance}(z, D) = \max\{P(d_i | z(\frac{1}{n}P(d_i | z)))\} \tag{9}$$

根据上面阐述的方法,能够对网络中结构化的海量图像信息检索挖掘的结果进行三维视觉展示,实现了图像信息的准确检索挖掘。

4 实验结果及分析

4.1 实验环境设置

为了验证本文提出的海量信息检索挖掘及视觉三维展现方法有效性,需要相关实验验证。实验在 matlab7.1 环境下进行,利用传统算法作为对照。在实验中,设置 PC 机的数目为 1 台,其中 1 台作为主服务器(主节点),其余 9 台作为从服务器(从属节点),每台 PC 机的配置为: Intel (R) Core (TM) 2 DUO T9800 双核 2.93 GHz, 8 GB DDR3 1066 内存, 500 GB 硬盘, Windows7 64 位操作系统。在网络中随机选取 10 万幅图像作为测试图像集,对不同算法下图像检索挖掘的效率和准确率进行比较。

4.2 不同算法实验结果对比及分析

实验中分别利用传统算法和改进算法进行海量图像的检索与挖掘,不同算法的实验结果分别能够用图 4 和图 5 进行描述。

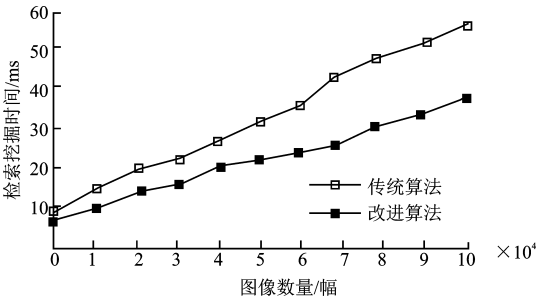


图 4 不同算法检索挖掘的时间

分析图 4、5 可知,利用改进算法进行海量图像信息的检索挖掘,检索挖掘的效率和准确率更高,这是由于传统的图像检索挖掘方法只能在一台 PC 机上进行,虽然在图像数量较低的时候能够保持一定的检索挖掘效率,但是一旦图像的数量达到一定的

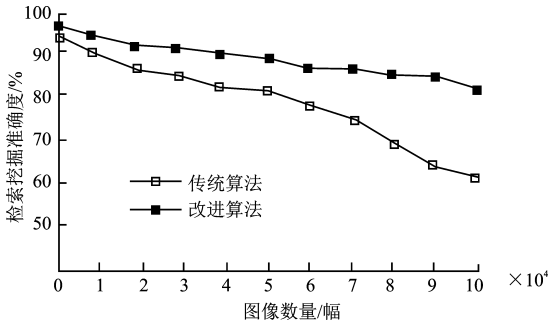


图 5 不同算法检索挖掘的准确度

级别后,单台 PC 机的检索挖掘操作过程就需要消耗大量的时间,极大地降低了图像检索挖掘的效率。同时,传统算法难以应对网络中图像数据库异构性与复杂性的特点,造成图像检索挖掘的准确率大幅度减低,因此,传统方法难以取得理想的实验效果。

而改进算法能够取得海量图像信息检索挖掘的理想结果,这是由于改进算法能够利用其分布式检索系统的 MapReduce 引擎,对海量图像信息进行相似性度量的检索挖掘,并且整个检索挖掘过程产生的数据都存储于分布式系统的数据库 Hbase 中,将若干 PC 机作为从属节点进行图像信息的分布式管理,提高了图像信息检索挖掘的效率。在进行检索挖掘结果的展示方面,改进算法能够利用聚类算法对图像特征向量进行分类,将一个视觉单词设置为一个聚类中心,利用视觉单词的集合描述一幅图像,利用潜在狄里克雷分配主题模型对图像与视觉单词之间的相关程度进行计算,并选取相关程度较高的视觉单词作为图像的视觉展示模型,实现了检索挖掘结果的三维视觉展示。因此改进算法能够取得理想的实验效果。上述实验充分体现出改进算法在网络结构化的海量信息检索挖掘方面的有效性。

为了进一步验证改进算法的优越性,对不同算法检索挖掘的漏检率进行了统计,不同算法的检索挖掘的漏检率能够用图 6 进行描述。

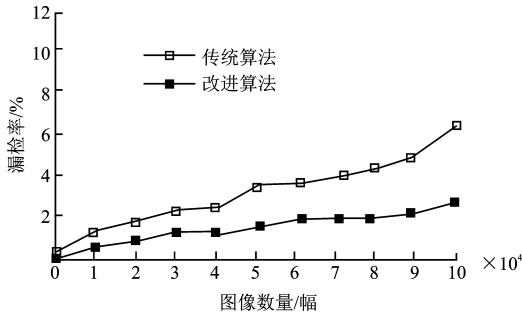


图 6 不同算法的漏检率

以提取网络流量的不同特征分量,不仅要以简化网络流量建模过程,并且能够对网络流量进行平滑和消噪,有利于提高网络流量的预测精度。

(3) 采用 RVM 避免了其他模型的,提高了网络流量预测结果的拟合精度,泛化能力也更优,具有广泛的应用前景。

参考文献:

- [1] 邱婧,夏靖波,吴吉祥. 网络流量预测模型研究进展[J]. 计算机工程与设计, 2012, 33(3): 865-869.
- [2] Wang R, Liu Y, Yang Y X, et al. Solving the applied classification problem of P2P traffic via optimized support vector machines[C]// Sixth International Conference on Intelligent Systems Design and Applications. Jinan, China: IEEE, 2006: 534- 539
- [3] Wang W S, Jin J J, Li Y Q. Prediction of inflow at three gorges dam in Yangtze River with wavelet network model [J]. Water Resources Management, 2009, 23(13): 2791-2803.
- [4] 周德懋,李舟军,康荣雷. 基于时间序列分析的网络流量预测模型研究[J]. 现代电子技术, 2009, 32(8): 115-117.
- [5] 董梦丽,杨庚,曹晓梅. 网络流量预测方法[J]. 计算机工程, 2011, 37(16): 98-100.
- [6] 高波,张钦宇,梁永生. 基于 EMD 及 ARIMA 的自相似网络流量预测[J]. 通信学报, 2011, 32(4): 47-56.

- [7] 姚奇富,李翠凤,马华林,等. 灰色系统理论和马尔柯夫链相结合的网络流量预测方法[J]. 浙江大学学报(理学版), 2007, 34(4): 396-400.
- [8] 党小超,郝占军. 基于改进 Elman 神经网络的网络流量预测[J]. 计算机应用, 2010, 30(10): 2648-2652.
- [9] 王俊松,高志伟. 基于 RBF 神经网络的网络流量建模及预测[J]. 计算机工程与应用, 2008, 44(13): 6-11
- [10] 罗赞骞,夏靖波,王焕彬. 混沌 支持向量机回归在流量预测中的应用研究[J]. 计算机科学, 2009, 36(7): 244-246
- [11] 刘渊,王鹏. 融合小波变换与贝叶斯 LS-SVM 的网络流量预测[J]. 计算机应用研究, 2009, 26(6): 2229-2231.
- [12] 张文金,许爱军. 混沌理论和 LSSVM 相结合的网络流量预测[J]. 计算机工程与应用, 2013, 49(15): 101-104.
- [13] 王升辉,袁正定. 结合多重分形的网络流量非线性预测[J]. 通信学报, 2007, 28(2): 45-50.
- [14] 张颖路. 基于遗传算法优化支持向量机的网络流量预测[J]. 计算机科学, 2008, 35(5): 177-179.

作者简介:

黄 震 男, (1980-), 硕士, 讲师. 研究方向为无线传感器网络、智能算法. E-mail: 195146501@qq. com.

钟锡武 男, (1993-). 研究方向为 Zigbee 无线传感器网络.

(上接第 140 页)

分析图 6 可知, 利用改进算法进行海量图像信息的检索挖掘, 能够保持更低的漏检率, 实验结果充分体现出改进算法的优越性。

5 结束语

提出一种基于分布式处理技术的结构化的海量信息检索挖掘及视觉三维展现方法. 由于采用了分布式结构的检索挖掘方法, 让主服务器与各个从服务器之间相互协作, 并行处理, 提高了各个服务器之间的数据流量, 降低了运行的冗余数据量, 提高了检索挖掘的效率, 并且通过视觉三维技术对检索挖掘结果进行更好的展示。

参考文献:

- [1] 沈宝国,蒋超峰,孙明. 基于单目视觉的轴类零件 3D 重构技术[J]. 机械设计与制造工程, 2015, 44(2): 17-20.
- [2] 杨彦波,刘滨,祁明月. 信息可视化研究综述[J]. 河北科技大学学报, 2014, 35(1): 91-102.
- [3] 王向军,卞越新,刘峰. 远距离三维坐标测量中双目视觉系统结构参数的优化[J]. 光学精密工程, 2015, 23(10): 2902-2908.

- [4] 王颖,韩静文,金翠云. 圆结构光视觉三维点管道缺陷检测及重构[J]. 应用光学, 2014, 35(3): 441-445.
- [5] 王颖,张瑞. 管道内表面圆结构光视觉三维测量系统[J]. 红外与激光工程, 2014, 43(3): 891-896.
- [6] 张锦华,乔蕊,汤跃明. 考虑重叠特征分类的图像检索模型仿真[J]. 计算机仿真, 2015, 32(8): 444-447.
- [7] 李荣,徐燕华. 基于视觉信息的图像特征提取算法研究[J]电子设计工程, 2016, 9(24): 188-190.
- [8] 谢世斌,韩志宏,余刚. 3D 立体图文的制作与展示技术[J]. 数码影像时代, 2014, 11(12): 101-103.
- [9] 刘建兰,朱丽. 基于计算机三维视觉的物流产品防伪算法研究[J]. 物流技术, 2014, 33(1): 354-356.
- [10] 张顺岚,莫建文. 基于双目视觉的三维人脸识别算法[J]. 电视技术, 2014, 38(9): 214-217.

作者简介:

孙钟德 男, (1980-), 高级工程师. 研究方向为计算机信息网络与应用气象. E-mail: 16148258@qq. com.

蒋 薇 女, (1983-) 硕士, 工程师. 研究方向为气候预测与气候变化.

刘巧根 男, (1970-), 工程师. 研究方向为应用气象与气象服务.

顾承华 男, (1965-), 工程师. 研究方向为应用气象.