第 30 卷 第 4 期
2015 年 8 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 30 No. 4
Aug. 2015

DOI:10.13364/j.issn.1672-6510.20140074

基于云平台的案例检索技术研究

熊聪聪，庞朝辉，王兰婷，耿世洁
（天津科技大学计算机科学与信息工程学院，天津 300457）

摘 要：针对大数据处理需求提出基于云平台的案例检索算法. 利用 MapReduce 技术改进了案例检索算法中常用的最近邻法，使其能够在多个服务器节点上并行执行，从而提高在海量数据情形下的案例检索速度. 实验表明：基于云平台的案例检索速度高于单节点检索，集群节点的数量对案例检索有一定的影响.

关键词：云平台；案例检索；MapReduce；最近邻法；改进
中图分类号：TP391.3 **文献标志码**：A **文章编号**：1672-6510(2015)04-0070-03

Case Retrieval Technology Based on Cloud Platform

XIONG Congcong，PANG Zhaohui，WANG Lanting，GENG Shijie
(College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract：A new case retrieval technique based on cloud platform was presented in order to process large amount of data. The KNN used in case retrieval was improved by using MapReduce technology, so that it can be executed in parallel across multiple server nodes and improve the speed of case retrieval among large amount of data. Experiments show that the case-based reasoning based on cloud platform is faster than that of the normal mode and the node number of the cluster has a certain effect on case-based reasoning.

Key words：cloud platform；case retrieval；MapReduce；KNN；improvement

云计算是一种基于互联网的计算方式. 通过这种方式，共享的软硬件资源和信息可以按需求提供给计算机和其他设备[1]. 云计算是以并行计算为核心技术，同时使用多种计算资源解决计算问题的过程. 通过并行计算集群完成数据的处理，再将处理的结果返回给用户，可以减少计算时间，提高系统的使用效率. 云计算实现了高效的并行计算与海量数据的管理，无疑是现今大数据时代的热门产业.

目前，无论是政府部门还是企业都将视角转向了云计算领域. 美国政府利用云计算技术建立了联邦政府网站，英国政府建立了国家级云计算平台(G-Cloud). 在我国，北京、上海、深圳、杭州、无锡等城市开展了云计算服务创新发展试点示范工作，以促进产业信息化[2]. 对政府用户而言，云计算能够提高办公效率、节约信息化成本，政府的推动同时也可以促进云计算产业的跨越式发展；对企业用户而言，企业可以利用云计算整合其现有的数据中心，实现对已有 IT 资源的充分利用，提高信息系统的效率和性能，加强经营决策的实时性.

CBR(case-based reasoning)技术[3]是通过重用或修改以前解决相似问题的方案来实现的. 随着互联网的飞速发展，数据量日益剧增，对于 CBR 的研究也要适应这一趋势. CBR 技术的一个典型的求解过程的基本步骤可以归纳为：案例检索(retrieve)、案例重用(reuse)、案例修正(revise)和案例保存(retain). CBR 解决问题的基本流程是利用目标案例的描述信息对案例库进行检索，得到与目标案例相类似的源案例，如果这个解答方案失败将对其进行调整，以获得一个能保存的成功案例，通过案例修正并保存可以获得一个新的源案例. 在案例推理过程中，案例表示、

收稿日期：2014–05–13；**修回日期**：2014–11–02
基金项目：国家自然科学基金资助项目(61272509)；天津市科技型中小企业技术创新资金项目(12ZXCXGX33500)
作者简介：熊聪聪（1961—），女，四川人，教授，xiongcc@tust.edu.cn.

案例检索和案例调整是案例推理研究的核心问题. 由于 CBR 技术对问题的解决是以经验知识为基础的，所以在应急事件处理、事件评估、医疗、企业管理等领域得到了广泛的应用.

常用的案例检索算法有知识引导法、神经网络法、归纳索引法和最近邻法. 其中，最近邻法是比较常见的一种检索算法[4]. 但是，目前对于案例检索算法的研究还停留在单节点检索，随着案例库中案例的增多，不管采用哪种算法都不能高效地对案例进行检索.

本文将案例检索中的最近邻算法与云计算平台进行结合，使得在海量数据的案例库中，可并行地对案例库进行检索，从而提高检索速度.

1　案例检索算法的实现

1.1　案例的存储

1.1.1　案例的表示

CBR 技术中知识的表示偏于半结构化或者非结构化，其知识的表示是一个重要的问题. 本文采用本体的知识表示方式，利用构建工具 protege 进行本体的构建. 采用本体对案例库进行建模，能够为不同领域知识及规则提供描述框架及规范，构建易于扩展的术语词典，实现知识的统一描述和组织.

1.1.2　HDFS 存储

HDFS（Hadoop distributed file system）[5]以容错性好、可伸缩性强、代码开源等优势倍受关注，成为当前主流分布式文件系统之一. HDFS 是被设计成可以在大规模廉价机器上运行的分布式文件系统，其设计思想源自 GFS（Google file system）. 由于 Hadoop 平台上从节点可以随时扩充，且案例存储在云平台上，即 HDFS 文件系统上，使得案例库具有较好的横向扩展性，便于案例库的扩张与案例的存储.

1.2　案例的检索

1.2.1　案例检索算法

对于基于本体的表示方式，案例库中的案例包括案例的标识及各种属性. 本文提出的案例检索算法是根据最近邻的思想计算案例间的属性值的相似度，进行相似匹配[6].

以案例 X 和案例 Y 为例，它们的属性分别为 x_1，x_2，…，x_m 和 y_1，y_2，…，y_m. 根据各属性权值采用式（1）计算 X 与 Y 的相似度.

$$sim(X,Y)=\sum_{i=1}^{m}\omega_i sim(x_i,y_i) \tag{1}$$

式中：ω_i 为相应属性的权值，根据属性对案例的影响大小确定；$sim(x_i,y_i)$ 是案例 X 的第 i 个属性与相应 Y 的第 i 个属性的相似度.

根据案例的属性类型不同，相应的相似度计算方法有一定的区别：

对于确定的属性值（例如在农业生产中的温度，不同的温度对农作物产生不同的影响），不同属性间的相似度可由式（2）计算.

$$sim(x_i,y_i)=1-d(x_i,y_i)=1-\left|x_i,y_i\right|\Big/\left|\max_i-\min_i\right| \tag{2}$$

式中：$d(x_i,y_i)$ 是属性值间的相对距离；$\max_i$、$\min_i$ 是属性 i 的最大值和最小值.

对于不确定属性，即类型为布尔型的属性（例如天气是否下雨等），可由式（3）计算不同属性间的相似度.

$$sim(x_i,y_i)=\begin{cases}1 & x_i=y_i \text{或} x_i\subset y_i\\ 0 & \text{其他}\end{cases} \tag{3}$$

上述是一种基于最近邻算法的案例匹配算法，当案例库不大时可以及时地检索出与所给问题相似的案例的解决办法. 但是，当案例不断扩充，案例库增加至几百 GB 甚至 TB 以上时，这种做法就显得力所不及了. 因此，考虑在云平台上对算法进行改进，使得对海量数据的案例库检索仍然可以快速地返回结果.

1.2.2　算法改进

开源云计算平台 Hadoop 中的 MapReduce 是一个软件框架，基于它写出来的应用程序能够运行在由上千台商用服务器组成的大型集群上，并以一种可靠容错的方式并行处理 TB 级别的数据集. MapReduce 技术[7]最早由 Google 公司提出，是一种通过在大规模的廉价服务器集群上进行大数据处理的技术. MapReduce 是一种并行编程模型，运行在分布式文件系统之上，通过 map 和 reduce 操作分别进行数据的处理. MapReduce 模型简单，支持系统的扩展和高并发，是现阶段应用最多的大数据处理技术.

MapReduce 在工作时由 1 个主节点对集群进行控制，同时由 n 个从节点进行实际任务的处理. 在案例检索时，先将同一地理位置上的数据进行 map 操作，并对中间结果进行 combine 操作，将中间结果存储在本地的服务器上，这样就节省了数据传输的耗时. 然后，Reduce 节点根据 Master 节点提供的地理信息提取中间结果，再对这些中间结果提取进行 reduce 操作，完成数据的分析工作，获取数据中的知识，帮助完成决策的生成. MapReduce 的工作流程[8]

见图 1.

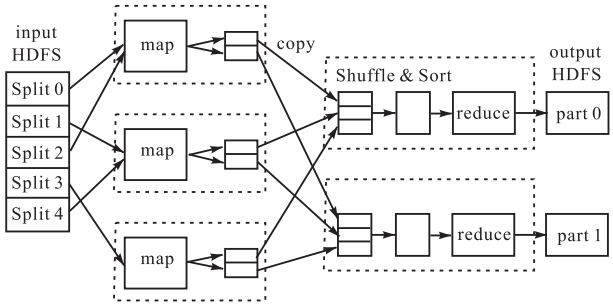

图 1　MapReduce 的工作流程

Fig. 1　Process of MapReduce

利用 MapReduce 技术改进的案例检索算法，其工作流程如下：

(1) 对案例库中的案例进行分片；

(2) Map 过程. 每个从节点对本地案例库中的案例进行分片处理. 输入的键值对为 (Case_ID，Case_Attri)，输出的键值对为 (Case_ID，Case_Attri)；

(3) Combine 过程，即案例间相似度的计算过程. 输入的键值对就是 map 过程的输出，输出的键值对是 (Case_ID，Case_Sim)；

(4) Reduce 过程. 根据案例的相似度从高到低对案例库中的案例进行排序. 输入的键值对为 Combine 过程的输出，输出的键值对是 (Case_Sim，Case_ID)；

(5) 最后提取出相似性最高的案例，为后续的案例的生成提供方案.

2　实　验

为验证在云平台上进行案例检索的可行性，分别在不同节点数的集群上进行实验.

2.1　实验环境

采用 8 台服务器，其中 1 台服务器作为主节点，7 台作为从节点. 每台服务器的软硬件环境均相同. 服务器配置见表 1.

表 1　集群节点配置

Tab. 1　Configurations of each node

项目	配置
CPU	Intel Xeon E5506，2.13 GHz，2 CORE
内存	2 GB，1 333 MHz
硬盘	SATA，200 GB
OS	Ubuntu 12.04
JDK	1.7.0_1
Hadoop	Hadoop-0.21.0

实验中将 8 台服务器的存储空间利用 Hadoop 的 HDFS 进行资源的虚拟，构建成一个大容量的虚拟资源池，将实验数据存储在虚拟资源池中. 实验时，分别用不同节点数的服务器对资源池中的数据进行检索计算.

2.2　实验结果

在实验中随机生成 2 个简单的数据集，数据量分别为 638 GB 和 1.31 TB，存储于资源池中. 数据集的样式见图 2.

```
4997 1.6 2.1 5.2 1.1 0.8 3.6 2.4 4.5
4998 1.0 1.1 1.2 2.1 0.3 2.3 1.4 0.5
4999 1.7 1.2 1.4 2.0 0.2 2.5 1.2 0.8
5000 1.2 1.8 1.6 2.5 0.1 2.2 1.8 0.2
5001 1.9 2.1 6.2 1.1 0.9 3.3 2.4 5.5
5002 1.0 0.8 1.6 2.1 0.2 2.3 1.6 0.5
5003 1.6 2.1 5.2 1.1 0.8 3.6 2.4 4.5
5004 1.7 1.2 1.4 2.0 0.2 2.5 1.2 0.8
5005 1.2 1.8 1.6 2.5 0.1 2.2 1.8 0.2
5006 1.9 2.1 6.2 1.1 0.9 3.3 2.4 5.5
5007 1.0 0.8 1.6 2.1 0.2 2.3 1.6 0.5
5008 1.6 2.1 5.2 1.1 0.8 3.6 2.4 4.5
```

图 2　部分实验数据集

Fig. 2　Partial experimental data set

两数据集分别在不同集群节点数量时的案例检索实验结果见图 3.

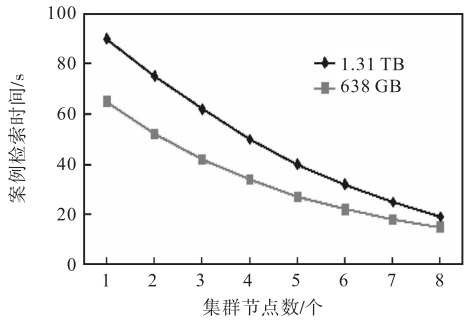

图 3　集群节点数量和数据量对案例检索时间的影响

Fig. 3　Effect of node number and data size on CBR

由图 3 可以看出：集群对 638 GB 的数据进行检索时，8 个节点比单节点要快 50 s；对 1.31 TB 的数据量进行检索时，8 个节点比单节点要快 71 s. 实验表明：集群节点数量对于案例的检索时间有一定的影响，对大数据量的数据进行检索，随着节点的增加，案例的检索速度会加快；同时，需要处理的数据量越大，这种效果表现的越明显.

3　结　语

本文提出一种基于云平台上的案例检索技术.

（下转第 77 页）