一种基于标签体系的多用途快速检索算法的研究

周雅琳

（广东建设职业技术学院，广东 广州 510000）

摘 要 “标签体系 + 检索算法”是人们在进行资源管理，例如图书馆的馆藏资源管理、电子商务中商品管理、客户资源管理等资源管理工作中一种通用的解决问题的思路。本文介绍一种基于标签体系的快速检索算法，该算法可一定程度上解决多标签体系的场景下，对高维数据进行多关键字组合搜索引起的数据库进行大量关联运算的问题，提高检索效率，该算法和标签体系结合和形成一套高效的“标签体系 + 检索算法”解决方案，可应用于多种资源管理的场景。

关键词 标签体系 检索算法 高维数据索引 资源管理 高职院校图书馆

中图分类号：TP312 **文献标识码**：A **文章编号**：1007-0745(2021)05-0057-02

1 背景概述

“标签体系 + 检索算法”是人们在进行资源管理，例如图书馆的馆藏资源管理、电子商务中商品管理、客户资源管理等资源管理工作中一种通用的解决问题的思路。本文介绍一种基于标签体系的快速检索算法，该算法可一定程度上解决多标签体系的场景下，对高维数据进行多关键字组合搜索引起的数据库进行大量关联运算的问题，提高检索效率，该算法和标签体系结合和形成一套高效的“标签体系 + 检索算法”解决方案，可应用于多种资源管理的场景。在我们的前期论文《高职院校图书馆信息系统中的标签管理功能探讨》[1] 已对这一算法应用于高职院校图书馆的馆藏资源管理进行探讨，本文从另一角度进行更高抽象程度的梳理和探讨，以期待读者对这一解决方案和算法有更深的理解，能将这一解决方案和算法应用于更多的场景。

2 算法的数学基础

这一算法的数学基础是我们在 2003 年一份数据挖掘课程研究报告《一种基于神经生物学原理的多维数据索引算法》中提出的：公比为 2 的等比数列有一个重要的特性：数列中两个任意不完全相同的子列，两个子列各自的元素之和必定不相等，这一特性可用于高效的多维数据索引算法的实现；选择公比为 2 是为了各子列的元素和数尽量小，从而使算法能支持更多的维数和索引值。[2]

3 算法数学基础的证明

这里证明算法的数学基础，即证明数学命题：公比为 2 的等比数列，对于数列中任意两个不完全相同的子列，两个子列各自的元素之和必定不相等，这个命题在文献 [2] 中已给出严格证明，这里整理一个相对容易理解的证明过程如下：

（1）设有公比为 2 等比数列的子列 A(a1,a2,a3,a4,a5), 公比为 2 等比数列的子列 B（b1,b2,b3,b4,b5）, 两个子列不完全相同，不妨设从右往左逐个比对，第一对不同的元素是 a3 和 b3, 且 a3<b3, 同时 a5=b5,a4=b4。

（2）设有公比为 2 等比数列的子列 A1（1，2，4，8，...，a3），则子列 A1 的元素和大于或者等于子列 A 的子列 A2(a1,a2,a3) 的元素和。

（3）根据公比为 2 等比数列的性质，即使在 b3 最小、也就是 b3=2*a3 的情况下，也有子列 A1 的元素和等于 b3-1，从而得出在各种情况下，子列 A1 的元素和都小于 b3, 因此各种情况下，子列 A2 的所有元素和都小于 b3，进而小于 b1+b2+b3。

（4）综合（1）（2）（3），易得子列 A 的所有元素和小于子列 B 的所有元素和，子列 A 和子列 B 各自的元素之和不相等。

（5）综上所述，问题得以证明。

4 数据库设计要点

在数据库设计时，有这样一个“索引和数表”，每条记录的结构是：（记录 ID，资源 ID，索引和数）；在数据库中有一个“标签基本信息表”，每条记录的结构是：（标签 ID、标签种类、标签值、标签索引值）。

5 算法的关键步骤

（1）在设置标签体系的时候，每类标签的每一个标签值，例如文献 [1] 中提及的图书馆图书标签体系中，2019 年入馆教育标签体系中的“建筑工程技术”专业标签，在数据库中有这样一条记录与其对应（“标签 ID1”、“2019 年入馆教育”、“建筑工程技术”、“32”），其中 32 为公比为 2 的等比数列中的第 6 个数，标签体系中的每一条标签的记录中的“标签索引值”都和等比数列中的某一个数形成一一对应关系。

（2）以文献 [1] 中提及的图书馆图书的标签管理为例，在为图书打标签的时候，系统会在数据库中，为这本图书加入一条或多条“索引和数表”记录，结构为（记录 ID，

图书 ID，索引和数），例如，假设“图书 ID1”已经有索引值为“1”、“4”的两个标签，在打索引值为“32”的标签的时候，系统会在“索引和数表”中加入（“记录 ID101”、“图书 ID1”、“32”），（“记录 ID102”、“图书 ID1”、“33”），（“记录 ID103”、“图书 ID1”、“36”），（“记录 ID104”、“图书 ID1”、“37”）四条记录，分别代表标签“1”、“4”、“32”可能出现的四种新组合；每本图书搭每一个标签都执行上述算法步骤。

（3）以文献 [1] 中提及的图书馆图书的标签管理为例，在取消某一标签时，参考上述过程易设计出从“索引和数表”减少相关记录的算法步骤。

（4）以文献 [1] 中提及的图书馆图书的标签管理为例，在检索图书时，系统根据用户所选择的标签，计算出“目标索引和数”，采用计算出的“目标索引和数”查找“索引和数表”中“索引和数”值和“目标索引和数”相等的记录，查找到的记录对应的图书就是符合标签组合要求的图书，这一算法过程只需查询一次数据库的表，就可以快速检索出符合用户标签要求的图书。

6 算法性能提升的重要原因

以文献 [1] 中提及的图书馆图书的标签管理为例，从上述算法关键步骤可以看出，算法搜索性能提升的重要原因是在查找符合多标签组合要求的图书时，这一算法只需查询一次数据库的一张表，可以快速检索出符合用户标签要求的图书，避免了数据库进行多次大量的关联运算；这一优势在图书总量大、标签体系复杂丰富的情况下特别明显。

从算法运行效率的时空分布情况看，算法是采用“打标签和取消标签时多花一点运算时间”换取“大量读者进行各种多标签组合检索检索时的高效率”的做法，对于实际业务情况来说，这种做法是合理的，有明显的效益。

7 算法应用场景探讨

在上述介绍算法过程中，为了方便读者理解，我们以图书馆中，馆藏图书的管理作为例子，实际上很多涉及资源管理的场景都可以用上述高效的“标签体系 + 检索算法”的解决方案，例如图书馆的馆藏资源管理、电子商务中的商品管理、客户资源管理等资源管理工作等。

7.1 算法在图书馆馆藏资源管理中的应用

在图书馆的馆藏资源管理中，主要可能用到的标签体系有：A 行业分类标签体系；B 入馆教育标签体系等。综合考虑各标签体系随时间的变化情况，结合上述算法，可形成一套高效的用于图书馆馆藏资源管理的解决方案。具体的结合详细方法可参考文献 [1] 中所述。这一解决方案在图书馆特色馆藏建设管理、参考咨询业务改进、入馆教育改进等业务工作中有重要用途 [3-5]。

7.2 算法在商品管理中的应用

在电子商务的商品管理中，主要可能用到的标签体系有 :A 商品用途标签体系；B 商品品牌标签体系；C 商品生产信息标签体系；D 商品存储信息标签体系；E 商品销售情况标签体系等。综合考虑各标签体系随时间的变化情况，结合上述算法，可形成一套高效的用于电子商务中商品管理的解决方案。

7.3 算法在客户资源管理中的应用

当前，在新一代信息技术不断发展的情况下，数字经济空前活跃，数字化转型升级已成为各企业面临的一个重要工程；在数字化转型工作中，对企业相关的各类客户通过客户信息系统进行管理和分类，掌握精准的客户画像，从而支撑实现“精准地把产品和服务销售给需要的客户”是一项重要的工作；在完成这一项重要工作的过程中，“标签体系 + 检索算法”是一个重要的工作工具。各企业可根据自身业务和客户群的特点，设计符合自身情况的标签体系，结合上述算法，实现对客户的高效管理和对目标客户的精准查找，助力自身数字化转型工作的进步。

7.4 算法在其他应用场景中的应用

除了上述三类举例的应用场景，本文提出的解决方案在各类资源管理类的场景中都有用武之地，有相对广阔的应用前景。

8 结语及工作展望

综上所述，本文在综合总结前期工作的基础上，提出了一种基于标签体系的多用途快速索引算法，介绍了算法的数学基础、算法数学基础的证明过程、算法实现过程中数据库的设计要点、算法的关键步骤、算法性能提升的重要原因、探讨了算法的应用场景。本文提出的算法和标签体系结合，形成一套“标签体系 + 检索算法”的解决方案，一定程度上解决多标签体系的场景下，对高维数据进行多关键字组合搜索引起的数据库进行大量关联运算的问题，提高检索效率，可用于多种资源管理场景；下一步工作可进一步拓展应用场景，让算法发挥更大作用。

参考文献：

[1] 周雅琳 . 高职院校图书馆信息系统中的标签管理功能探讨 [J]. 科学与财富 ,2020,11:3.
[2] 谢勤 . 一种基于神经生物学原理的多维数据索引算法 [Z]. 数据挖掘课程研究报告 ,2003.
[3] 周雅琳 , 谢勤 . 高职院校图书馆效益提升思路研究 [J]. 知识经济 ,2016,388(08):164.
[4] 周雅琳 , 谢勤 . 浅谈如何提高高职院校图书馆参考咨询服务 [J]. 卷宗 ,2016,06(02):38.
[5] 周雅琳 . 高职院校图书馆新生入馆教育内容体系改进方案研究—以广东建设职业技术学院为例 [J]. 广东教育 (职教),2020,08:25-26.