学校编码：10384　　分类号__　　密级__
学号：31520131153302　　UDC _

厦门大学

硕士学位论文

哈希二值码嵌入算法研究

Hashing binary code embedding algorithm

作者姓名 伍兆盖

指导教师姓名：纪荣嵘教授
专　业　名　称：计算机技术
论文提交日期：2016 年 4 月 19 日
论文答辩时间：2016 年 5 月 19 日
学位授予日期：2016 年 6 月日

答辩委员会主席：
评阅人：

2016　年　　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

近些年来随着移动和 PC 互联网等大规模数据的爆炸式增长，越来越多的研究人员开始对有效的大规模数据检索问题进行广泛而又深入的研究。通过暴力搜索直接比较查询点和数据库数据点之间的相似性显然是不可行的，因为对海量数据进行暴力搜索对计算和内存都有着极高的要求。针对这一问题，研究人员开始研究基于哈希算法的将原始数据空间中的高维特征数据转换为汉明空间中的低维哈希二值码，并通过计算原始数据在汉明空间中的哈希二值码之间的汉明距离来表征原始数据点之间的相似程度。采用哈希二值码嵌入算法有如下两方面的优势：（1）哈希二值码用二进制 0 1 序列表示，因此可有效地通过整数来存储，从而可大大降低原始海量数据的存储消耗；（2）可通过现代 CPU 内置的位运算方式快速计算哈希二值码之间的汉明距离，从而可实现对大规模数据的快速检索。

本文首先提出了一种通过在线学习方式的在线哈希二值码嵌入算法学习哈希函数。当数据规模较大时数据难以一次性载入到内存当中，这就造成哈希算法难以在实际中大规模使用。在线学习却没有这样的限制，因此本文借鉴了在线学习的优点，将在线学习应用于哈希函数的学习当中，通过在线学习的方式不需要在模型学习时一次性将数据全都载入到内存当中，而是每次通过一对新的数据特征向量选择性地更新模型参数。

本文还从数据的相对相似性角度出发提出了一种基于在汉明空间中保持数据相对顺序的亚线性顺序保持哈希二值码嵌入算法。现有的哈希算法大多通常只考虑数据点对之间的相似性，而本文则是把数据的相似性信息和数据的不相似性信息同时加入到目标函数中，通过在汉明空间中保持相似数据和不相似数据在原始空间中的相对相似性学习哈希函数。相较于基于数据点对相似性的哈希算法，基于相对相似性则加入了更多的监督信息，从而可更有效地学习哈希函数。

本文在 3 个知名的公开数据集上与 9 种主流哈希算法进行了对比。实验结果表明本文提出的在线哈希二值码嵌入算法和亚线性顺序保持哈希二值码嵌入算法在性能和效率上均优于其它算法。

关键词：哈希算法；在线学习；顺序保持

Abstract

The explosive growth in big data has attracted more and more attention in designing efficient indexing and search methods recently. The straightforward solution using exhaustive comparison is infeasible due to the prohibitive computational complexity and memory requirements for massive data. Therefore hashing is becoming increasingly popular for efficient retrieval for massive data. There are two advantages for hashing algorithm: firstly, we can use compact integer to represent binary codes so as to reduce the storage for large-scale data dramatically; secondly, we can take advantage of modern CPU built-in bit operation to calculate Hamming distance rapidly.

In this paper, we propose a hashing algorithm which learns hash function through online learning. Current data dependent hashing algorithms assume there are large amount of training data, however, it's difficult to load such scale data into memory and not applicable to large scale searching. However online learning doesn't constraint to such limit, so we train the data via online learning and it's no need to load the entire dataset into memory, and online learning is self-adaptive to new data points.

In this paper, we also propose a new hashing method which based on relative similarity, ranking based hashing algorithm. The traditional hashing algorithm only consider data pair similarity, however, we unite data similar and dissimilar message into object function, learning hash function by preserving similar data's similarity and dissimilar data's dissimilarity in Hamming space. Compared to traditional data pair similarity, relative similarity add more supervised information to hash function learning framework so we can get better hash function.

We have evaluated our algorithm on 3 well-known public datasets and compared with 9 state-of-the-art hashing algorithms. Experiments have shown our algorithms significantly outperform others.

Key words: Hashing Algorithm; Online Learning; Ranking Preserve.

目 录

摘要.....I

Abstract.....II

第一章 引言.....1

1.1 研究背景与意义.....1
1.2 哈希算法研究现状.....4
1.3 本文工作及组织结构.....8

第二章 相关工作.....11

2.1 在线核哈希算法.....11
2.1.1 哈希函数.....11
2.1.2 目标函数.....12
2.1.3 目标函数优化.....13
2.2 排序保持哈希算法.....14
2.2.1 哈希函数.....14
2.2.2 目标函数.....15
2.2.3 目标函数的优化.....15
2.3 本章小结.....17

第三章 在线哈希二值码嵌入.....19

3.1 哈希函数.....19
3.2 目标函数.....20
3.3 汉明损失函数.....21
3.4 目标函数优化.....21
3.5 带外扩展.....23
3.6 实验设计与分析.....24
3.6.1 数据集与实验设置.....24
3.6.2 对比算法.....27
3.6.3 实验结果与分析.....29
3.7 本章小结.....37

第四章 亚线性顺序保持哈希二值码嵌入.....39

4.1 哈希函数.....39
4.2 目标函数.....40
4.3 目标函数优化.....41
4.4 带外扩展.....43
4.5 实验设计与分析.....44
4.5.1 数据集与实验设置.....44

4.5.2 对比算法.....44
4.5.3 实验结果与分析.....45
4.6 本章小结 54

第五章 总结与展望 57

5.1 总结 57
5.2 展望 58

[参考文献].....59

致谢.....65

读研期间发表论文 67

Content

Abstract I

1 Introduction 1

1.1 Research background and significance 1
1.2 Hashing algorithm research status 4
1.3 Organization of this paper 8

2 Relative hashing algorithm 11

2.1 Online kernel hashing algorithm 11
2.1.1 Hashing function 11
2.1.2 Objective function 12
2.1.3 Optimization of objective function 13
2.2 Ranking preserve hashing algorithm 14
2.2.1 Hashing function 14
2.2.2 Objective function 15
2.2.3 Optimization of objective function 15
2.3 Conclusion 17

3 Online hashing binary code embedding 19

3.1 Hashing function 19
3.2 Ojbective function 20
3.3 Hamming loss function 21
3.4 Optimization of objective function 21
3.5 Band extension 23
3.6 Experiments and analysis 24
3.6.1 Datasets and settings 24
3.6.2 Comparing algorithms 27
3.6.3 Experiment results and analysis 29
3.7 Conclusion 37

4 Sublinear ranking preserve hasing binary code embedding 39

4.1 Hashing function 39
4.2 Objective function 40
4.3 Optimization of objective function 41
4.4 Band extension 43
4.5 Experiments and analysis 44
4.5.1 Datasets and settings 44
4.5.2 Comparing algrorithm 44
4.5.3 Experiments and analysis 45
4.6 Conclusion 54

5 Conclusion and future work 57

5.1 Conclusion 57
5.2 Future work 58

[Reference] 59

Acknowledgement 65

Publishment 67

第一章 引言

1.1 研究背景与意义

最近二十年以来随着移动互联网和 PC 互联网、大数据、移动端和 PC 端社交网络媒体以及其它计算机技术的高速发展，信息获取途径的多样化和广泛化，数据表现出高速增长、快速变化和规模巨大等特点。现如今万维网有着超过三亿的网站，包含了超过一万亿的网页数量。推特和新浪微博每天发出的推文和微博数目超过一亿条，雅虎每天信息交换数量超过三十亿。除了文本数据规模愈来愈大之外，图片分享网站 Flickr 存储了超过五十亿张图片，然而这个数字仍然在以每分钟 3000 张照片的速度增加。国际知名视频共享网站 YouTube 在每分钟内都有超过 100 多个小时的视频内容上传和发布。由于如此高速激增的海量数据从而使得现代信息技术必须面对如何高效储存、处理和搜索海量数据的严峻挑战。然而实际上和存储消耗相比，在海量数据库中检索用户查询相关的内容被证明是一个更大的挑战。除了广泛使用的基于文本结构的商业搜索引擎如谷歌、百度和必应之外，基于内容的图像检索（Content Based Image Retrieval，CBIR）在过去的十几年受到了研究人员的广泛的关注[1]。基于图像内容的图像检索系统并非依赖于类似文本检索中的关键词索引结构标签，而是通过提取其非结构化数据的内容特征向量（如图像的 SIFT[2]、SURF[3]和 GIST[4]等视觉特征描述子）来对数据对象进行索引和查询。基于内容的图像检索系统为了能够快速响应用户的视觉搜索需求而必须首先能构建高效的索引结构并从中快速返回检索结果。

最近邻检索也被称之为相似性检索或相似项线索，最近邻检索的目的是从海量数据库中检索出和用户查询距离小于给定阀值的数据点或者找出不超过给定数目的距离最小的数据点，这些检索出的数据点被称之为用户搜索查询的最近邻。在海量数据库中进行精确的最近邻检索时，用户查询和数据库中的所有数据点计算它们之间的相似性的计算代价是巨大的。作为更好的替代方法，近似最近邻检索被证明是更加有效的检索方式，且近似最近邻检索在大多数通常的实际应用中能满足足够精度的检索需求，因而国内外许多研究者们对近似最近邻检索做了大量的深入研究。

在给定的数据库里面检索相似相关的数据本质上是最近邻检索问题[2]。在现实应用场景中，暴力搜索直接将搜索查询和海量数据库中的所有样本直接进行线性比较，然而线性时间复杂度对于当今海量数据库来说显然是不可接受的。在实际的大规模检索系统中除了遇到检索数据规模大的问题之外，还有一个广泛存在的问题是数据的特征向量维度高，这被称之为维度灾难[5][6]。这是因为现代人们日常使用的数据往往包含更多更丰富的信息，其特征向量的维度通常有成千上万维，典型的如音频，图形图像和视频等多媒体数据。因此暴力搜索不仅计算复杂度高，将原始空间中的海量数据加载到内存中亦是一个关键的性能瓶颈。然而在实际应用中通常近似最近邻检索（Approximate Nearest Neighbors, ANN）就足以能够满足日常大多数的检索应用需求，近似最近邻搜索的主要思想是从海量搜索数据库搜索出的最近邻搜索结果中容许一定程度的相似性误差，并通过某种可接受的方式放宽对搜索出的最近邻搜索结果的严格限制，并以此降低了检索空间和提高了检索性能和效率。基于树的索引方法例如 KD 树[7]，ball 树[8]，metric 树[9]，vantage point 树[10]等在过去的几十年里曾是非常流行的搜索索引解决方案，但是基于树的索引方法往往要求更多的数据存储空间，有时甚至比数据本身占用空间还大。此外，随着检索数据特征向量维度的增大，基于树的索引方法的检索性能急剧下降[11]。为了应对高维数据问题，近些年来有研究人员提出了一种称为乘积量化的方法，乘积量化的基本思想是通过对子空间分解技术来编码高维数据向量[12][13]。

基于树的索引方法通常对检索数据进行递归的划分，然而基于哈希算法的索引方法则是重复地对整个数据集进行划分并且由此从每一次划分当中生成一个哈希函数。基于哈希算法的二分划分将原始输入空间中的数据映射到离散的哈希二值码空间，这个离散的哈希二值码空间被称之为汉明空间。最终每个原始数据样本被表示成一串 01 哈希二值码序列。哈希二值码嵌入算法的基本原理是通过相同的哈希函数将数据库数据点和查询点映射为等长的哈希二值码，然后通过计算搜索 query 和海量数据库数据点的哈希二值码之间的汉明距离，并以此汉明距离来表征搜索 query 和海量数据库数据点之间的相似程度，并将汉明距离最小的部分检索结果作为搜索 query 的近似最近邻。本文的目的则是研究如何高效地学习哈希函数和哈希二值码。近些年来基于哈希算法的索引方法正成为广泛使用的

近似最近邻检索解决方案，尤其是在大规模图像检索，模式匹配和目标检测等应用领域。

基于哈希二值码嵌入算法的索引结构有如下两个方面优势：（1）哈希二值码由二进制 01 序列表示，因此可以有效地通过紧凑的整数形式来存储，从而可大大降低原始海量数据的存储消耗；（2）可通过现代 CPU 内置的位运算方式快速地计算哈希二值码之间的汉明距离，从而可实现对大规模数据的高效检索。图 1-1 左图表示通过哈希二值码嵌入算法将原始数据空间中的数据点映射为汉明空间中的哈希二值码之后，再以哈希二值码作为索引关键字构建倒排索引表，将每一个数据库数据点的哈希二值码作为索引 key，具有相同哈希二值码的原始数据点则对应到同一个哈希二值码 key 的链表当中。图 1-1 右图表示基于哈希二值码嵌入算法的单表在线搜索过程，通过将数据库映射为哈希二值码的相同的哈希函数将查询映射为哈希二值码，并以此哈希二值码和倒排索引表中各索引项计算小于给定汉明距离的索引项，并对返回索引项的哈希桶中返回检索结果再做一次原始空间距离重排，重排后的结果作为最终的检索结果。图 1-2 表示基于哈希二值码嵌入算法的多表在线搜索过程，在每个倒排索引表中的检索过程和单表检索相同，多表检索则是将多个单表的检索结果合并，并以此提高检索的召回率。

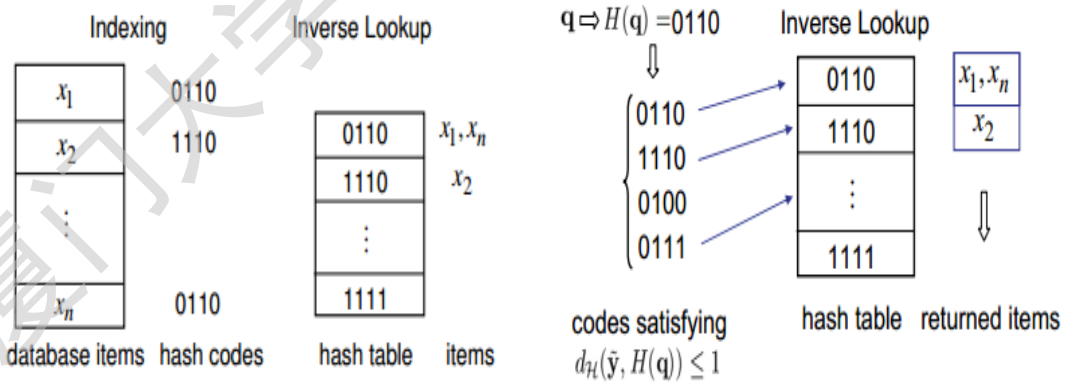

图 1-1 基于哈希二值码嵌入算法的倒排索引构建和单表在线查询

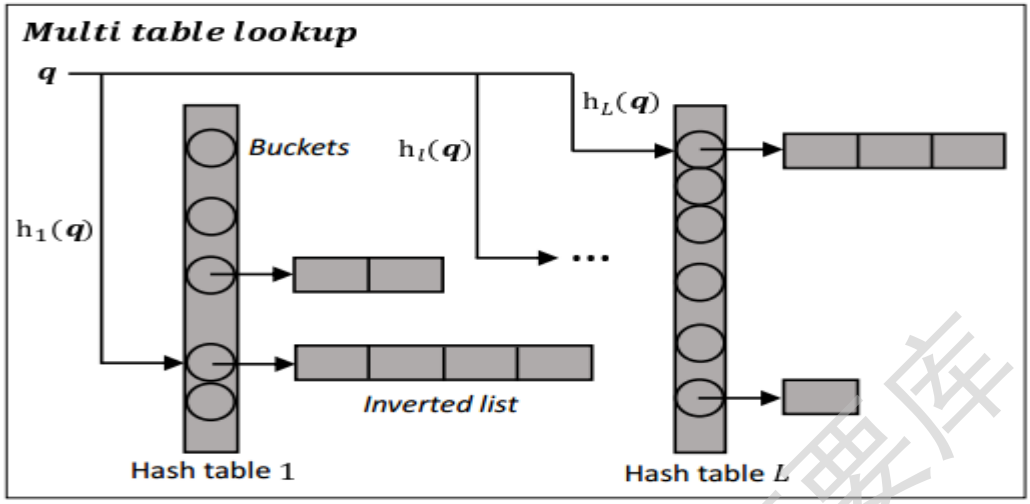

图 1-2 基于哈希二值码嵌入算法的多表查询

因为哈希二值码嵌入算法在计算和存储上的优势，近些年来哈希二值码嵌入算法广泛地应用于图形图像、视频和文本挖掘等机器学习领域。例如图像检索[14][15]，移动视觉搜索[16]，图像块匹配[17]，图像分类[18]，人脸识别[19][20]，目标跟踪[21]，文本和图像去重[22][23]，快速文本检索[24]等。通过哈希二值码嵌入算法将原始空间中的高维实数特征向量变换到汉明空间中的低维离散哈希二值码的过程中会有很多信息丢失。为了使得对原始数据特征向量做哈希映射之后，检索的准确率和召回率仍然能够保持到一个较高的水平，国内外越来越多的研究人员开始研究并研究出了大量基于机器学习的哈希二值码嵌入算法。

1.2 哈希算法研究现状

哈希二值码嵌入算法可以分为随机型哈希二值码嵌入算法和学习型哈希二值码嵌入算法。在随机型哈希二值码嵌入算法中，最著名的当属基于随机投影的局部敏感哈希二值码嵌入算法（Locality Sensitive Hashing，LSH）[25]。局部敏感哈希二值码嵌入算法是数据独立的哈希二值码嵌入算法，其哈希函数的学习不依赖于数据集，局部敏感哈希二值码嵌入算法使得原始空间中相似的数据点做哈希二值码嵌入处理后能以较大概率投影到超平面的同一侧，而对于原始空间中不相似的数据点则以较大概率投影到超平面的不同侧。尽管局部敏感哈希二值码嵌入算法能给检索带来亚线性的时间复杂度，但是基于随机投影方法的局部敏感哈希二值码嵌入算法亦有其不足之处，首先，为了取得预期的搜索准确率局部敏