学校编码：10384　　分类号＿＿＿密级＿＿＿
学号：23220131153333　　UDC ＿＿＿

厦门大学

硕 士 学 位 论 文

局部敏感哈希改进算法研究

Research on Improved Locality Sensitive Hashing Algorithm

岑伟

指导教师姓名：缪克华 副教授

专 业 名 称：模式识别与智能系统

论文提交日期：2016 年 月

论文答辩时间：2016 年 月

学位授予日期：2016 年 月

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

“大数据”时代给数据检索带来了新的挑战，相似性检索显得尤为重要。局部敏感哈希算法是相似性检索中最流行的一种，该算法是建立在哈希的基础上的一种近似最近邻算法，它能将检索时间复杂度缩减到线性。与其它基于 Tree 的数据结构相比，局部敏感哈希算法能较好的处理数据在高维空间中的检索问题。注意到高维数据利用该算法检索得到候选集后需要进行相似度计算，而这一部分所耗费时间的占整个数据检索时间的比重非常大，所以该算法在处理大规模数据时的性能仍然需要提高。

针对局部敏感哈希算法，考虑到数据维数和数据量的急剧增加而检索时间还不能满足需求情况下，提出了在不同场合下两种改进的局部敏感哈希算法并应用于图像，本文在局部敏感哈希算法的基础上做了以下工作：

1. 对利用局部敏感哈希算法查找前 k 个最相似（Top-k）的问题，提出了一种基于次数排序的局部敏感哈希算法（ST-LSH）。该算法将唯一化前的索引号按出现次数进行排序来输出，避免了相似度计算花费大量时间的问题。实验结果显示，与改进前的算法相比，在保证识别率基本不变的情况下，改进后的算法大大减少了数据检索时间。

2. 在利用局部敏感哈希算法进行数据类别的查找过程中，为了弥补容易受噪声点的影响，提出了一种基于 k 近邻分类（k-nearest neighbor）局部敏感哈希算法（KNN-LSH）。KNN-LSH 充分利用了唯一化后的候选集的信息，利用候选集唯一化后的索引号所属的类别进行 k 近邻分类。即统计当前候选集所属类别的信息，将查询对象归为所属类别最多的那一类，从而避免了大量的相似度计算。实验结果证明，与改进前的算法相比，KNN-LSH 对分类问题的识别率要好于传统方法，具有稳定性，同时也缩短了数据检索时间。

关键字：相似性检索；局部敏感哈希；高维数据；

Abstract

"Big Data" brings new challenges to data retrieval, and similarity search is particularly important. Locality sensitive hashing algorithm is one of the most popular algorithms in similarity search, it is an approximate nearest neighbor algorithm that based on the establishment of hashing, it reduces the complexity of retrieval time to linear. When compared with other data structure algorithms that based on the Tree , locality sensitive hashing can handle data better in high-dimensional space retrieval. Noting that after getting retrieved candidate set for high dimensional data, there will need similarity computation, this part takes a big proportion of time-consuming in the whole retrieval time, so the algorithm performances still needs to improve when dealing with large-scale data.

For locality sensitive hashing algorithm, taking into account the data dimension and a sharp increase in the amount of data retrieval time still can't satisfy the demand, put forward two improved locality sensitive hashing algorithms in different applications and applied them to the image, this paper do the following work:

1. When using locality sensitive hashing algorithm to find the k most similar (Top-k) objects, put forward a locality sensitive hashing algorithm based on the sort of times (ST-LSH). In order to avoid the problem large number of similarity computation, this algorithm outputs data according to the number of occurrences before unique the data. Experimental results show that compared with traditional locality sensitive hashing, improved algorithm ensures the recognition rate nearly unchanged, and greatly reduces data retrieval time.

2. When using locality sensitive hashing algorithm to find data label , in order to compensate its susceptible to noise points, propose locality sensitive hashing algorithm based on k-nearest neighbor classifier (k-nearest neighbor) (KNN-LSH) . KNN-LSH takes full advantage of the information after unique the candidate set, use categories of candidate set after unique technology to k-nearest classification. which means statistic the information of current candidate set belongs to a certain category,

the query object is classified to the largest number of the category , thus avoiding a lot of similarity computation. Experimental results show that when compared with the pre-modified algorithm, KNN-LSH classification performs better recognition rate than the traditional methods, with stability, but also shorten the data retrieval time.

Keywords： similarity search; locality sensitive hashing; high-dimensional data;

目 录

Contents