LSHBOX 0.9
User Manual

Zhifeng Xiao, xiaozf@gmail.com

Gefu Tang, tanggefu@gmail.com

Yang Long, 20288ly@sina.cn

June 30, 2015

Chapter 1

Introduce

Locality-Sensitive Hashing (LSH) is an efficient method for large scale image retrieval, and it achieves great performance in approximate nearest neighborhood searching.

LSHBOX is a simple but robust C++ toolbox that provides several LSH algrithms, in addition, it can be integrated into Python and MATLAB languages. The following LSH algrithms have been implemented in LSHBOX, they are:

- K-means Based Double-Bit Hashing (KDBQ)
- Double-Bit Hashing (DBQ)
- Spectral Hashing (SH)
- Iterative Quantization (ITQ)
- Random Hyperplane Hashing
- LSH Based on Random Bits Sampling
- LSH Based on Thresholding
- LSH Based on p-Stable Distributions

There are two repositories for compilation and performance tests, they are:

- LSHBOX-3rdparty: 3rdparty of LSHBOX, it is for compilation
- LSHBOX-sample datasets: datasets for performance tests

Part of the code depends on the C++11, So I think your compiler should support this feature.

We hope that there are more people that join in the test or contribute more algrithms.

Please feel free to contact us if you have any questions.

Chapter 2

Compilation

LSHBOX is written in the C++ programming language. And it also can be easily used in many contexts through the Python and MATLAB bindings provided with the library.

In order to make LSHBOX simple and easy to use, the library don't need to compile. You only need to add the include directory or modify the program search path, then you can use this library directly in C, C++, Python or MATLAB.

You can use CMAKE to build some tools for the test of this library.

In some cases, if you want or need to compile it by yourself with Python and MATLAB, please delete the comment of the last two lines in file "CMakeLists.txt", and you will find the compiling progress of python must rely on Boost library or some part of this library. For more detailed information, you can view the document "./python/README".

During compilation, create a new directory named "build" in the main directory, then choose a appropriate compiler and switch to the "build" directory, finally, execute the following command according to your machine:

- Windows

```
cmake -DCMAKE_BUILD_TYPE=Release .. -G"NMake Makefiles"
nmake
```

- Linux

```
cmake ..
make
```

Chapter 3

Usage

This chapter contains small examples of how to use the LSHBOX library from different programming languages (C++, Python and MATLAB).

- **C++**

```
/**
 * @file itqlsh-test.cpp
 *
 * @brief Example of using Iterative Quantization LSH index for L2 distance.
 */
#include <lshbox.h>
int main(int argc, char const *argv[])
```

```cpp
{
    typedef float DATATYPE;
    std::cout << "LOADING DATA ..." << std::endl;
    lshbox::timer timer;
    lshbox::Matrix<DATATYPE> data("audio.data");
    std::cout << "LOAD TIME: " << timer.elapsed() << "s." << std::endl;
    std::cout << "CONSTRUCTING INDEX ..." << std::endl;
    timer.restart();
    std::string file = "itq.lsh";
    bool use_index = false;
    lshbox::itqLsh<DATATYPE> mylsh;
    if (use_index)
    {
        mylsh.load(file);
    }
    else
    {
        lshbox::itqLsh<DATATYPE>::Parameter param;
        param.M = 521;
        param.L = 5;
        param.D = data.getDim();
        param.N = 8;
        param.S = 100;
        param.I = 50;
        mylsh.reset(param);
        mylsh.train(data);
    }
    mylsh.save(file);
    std::cout << "CONSTRUCT TIME: " << timer.elapsed() << "s." << std::endl;
    lshbox::Matrix<DATATYPE>::Accessor accessor(data);
    lshbox::Metric<DATATYPE> metric(data.getDim(), L1_DIST);
    unsigned K = 10;
    unsigned Q = 10;
    lshbox::Scanner<lshbox::Matrix<DATATYPE>::Accessor> scanner(
        accessor,
        metric,
        K,
        std::numeric_limits<float>::max()
    );
    std::cout << "RUNING QUERY ..." << std::endl;
    timer.restart();
    for (unsigned i = 0; i != Q; ++i)
    {
        std::cout << "----- QUERY " << i + 1  << " -----" << std::endl;
        scanner.reset(data[i]);
        mylsh.query(data[i], scanner);
```

```
        std::vector<std::pair<unsigned, float> > result;
        result = scanner.topk().getTopk();
        for (auto it = result.begin(); it != result.end(); ++it)
        {
            std::cout << it->first << ", " << it->second << std::endl;
        }
        std::cout << "Frequency: " << scanner.cnt() << std::endl;
    }
}
```

You can get the sample dataset 'audio.data' from http://www.cs.princeton.edu/cass/audio.tar.gz, if the link is invalid, you can also get it from LSHBOX-sample-data.

You can run the following program for a complete test in C++:

- tools/create_test_data.cpp
- tools/scan_run.cpp
- tools/rbslsh_test.cpp
- tools/rhplsh_test.cpp
- tools/thlsh_test.cpp
- tools/psdlsh_test.cpp
- tools/shlsh_test.cpp
- tools/itqlsh_test.cpp
- tools/dbqlsh_test.cpp
- tools/kdbqlsh_test.cpp

- **Python**

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
# pylshbox_example.py
import pylshbox
import numpy
# prepare test data
float_mat = numpy.random.rand(100000, 192)
float_query = float_mat[1,:]
unsigned_mat = numpy.int32(float_mat*5)
unsigned_query = unsigned_mat[1,:]
# Test rbsLsh
rbs_mat = pylshbox.rbslsh()
rbs_mat.init_mat(unsigned_mat.tolist(), '', 521, 5, 20, 5)
result = rbs_mat.query(unsigned_query.tolist(), 1)
indices, dists = result[0],result[1]
# Test rhpLsh
rhp_mat = pylshbox.rhplsh()
```

```
rhp_mat.init_mat(float_mat.tolist(), '', 521, 5, 6)
result = rhp_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
# Test thLsh'
th_mat = pylshbox.thlsh()
th_mat.init_mat(float_mat.tolist(), '', 521, 5, 12)
result = th_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
# Test psdlsh with param.T = 1
psdL1_mat = pylshbox.psdlsh()
psdL1_mat.init_mat(float_mat.tolist(), '', 521, 5, 1, 5)
result = psdL1_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
# Test psdlsh with param.T = 2
psdL2_mat = pylshbox.psdlsh()
psdL2_mat.init_mat(float_mat.tolist(), '', 521, 5, 2, 0.5)
result = psdL2_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
# Test shLsh
sh_mat = pylshbox.shlsh()
sh_mat.init_mat(float_mat.tolist(), '', 521, 5, 4, 100)
result = sh_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
# Test itqLsh
itq_mat = pylshbox.itqlsh()
itq_mat.init_mat(float_mat.tolist(), '', 521, 5, 8, 100, 50)
result = itq_mat.query(float_query.tolist(), 2, 10)
indices, dists = result[0],result[1]
```

- **MATLAB**

```
% lshbox_example.m
% prepare test data
dataset = rand(128,100000);
testset = dataset(:,1:10);
% Test rhplsh
param_rhp.M = 521;
param_rhp.L = 5;
param_rhp.N = 6;
[indices, dists] = rhplsh(dataset, testset, param_rhp, '', 2, 10)
% Test thlsh
param_th.M = 521;
param_th.L = 5;
param_th.N = 12;
[indices, dists] = thlsh(dataset, testset, param_th, '', 2, 10)
% Test psdlsh with param_psdL1.T = 1
```

```
param_psdL1.M = 521;
param_psdL1.L = 5;
param_psdL1.T = 1;
param_psdL1.W = 5;
[indices, dists] = psdlsh(dataset, testset, param_psdL1, '', 1, 10)
% Test psdlsh with param_psdL2.T = 2
param_psdL2.M = 521;
param_psdL2.L = 5;
param_psdL2.T = 2;
param_psdL2.W = 0.5;
[indices, dists] = psdlsh(dataset, testset, param_psdL2, '', 2, 10)
% Test shlsh
param_sh.M = 521;
param_sh.L = 5;
param_sh.N = 4;
param_sh.S = 100;
[indices, dists] = shlsh(dataset, testset, param_sh, '', 2, 10)
% Test itqlsh
param_itq.M = 521;
param_itq.L = 5;
param_itq.N = 8;
param_itq.S = 100;
param_itq.I = 50;
[indices, dists] = itqlsh(dataset, testset, param_itq, '', 2, 10)
% Test dbqlsh
param_dbq.M = 521;
param_dbq.L = 5;
param_dbq.N = 4;
param_dbq.I = 5;
[indices, dists] = dbqlsh(dataset, testset, param_itq, '', 2, 10)
% Test kdbqlsh
param_kdbq.M = 521;
param_kdbq.L = 5;
param_kdbq.N = 4;
param_kdbq.I = 5;
[indices, dists] = kdbqlsh(dataset, testset, param_itq, '', 2, 10)
```

Have you ever find the empty string used in the Python and MATLAB code? In fact, they can be used to save the index through pass a file name. Like the following, you will find the next query speed faster than the first, because there is no re-indexing.

- **Python**

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
```

```python
# pylshbox_example2.py
import pylshbox
import numpy
import time
# prepare test data
float_file = 'audio.data'
float_query = numpy.random.rand(192)
# Test itqLsh
# First time, need to constructing index. About 1.5s.
start = time.time()
itq_file = pylshbox.itqlsh()
itq_file.init_file(float_file, 'pyitq.lsh', 521, 5, 8, 100, 50)
result = itq_file.query(float_query.tolist(), 2, 10)
print 'Elapsed time is %f seconds.' % (time.time() - start)
# Second time, no need to re-indexing. About 0.05s.
start = time.time()
itq_file2 = pylshbox.itqlsh()
itq_file2.init_file(float_file, 'pyitq.lsh', 521, 5, 8, 100, 50)
result = itq_file2.query(float_query.tolist(), 2, 10)
print 'Elapsed time is %f seconds.' % (time.time() - start)
```

- **MATLAB**

```matlab
% lshbox_example2.m
dataset = rand(128,500000);
testset = dataset(:,1:10);
% Test itqlsh
param_itq.M = 521;
param_itq.L = 5;
param_itq.N = 8;
param_itq.S = 100;
param_itq.I = 50;
% First time, need to constructing index. About 10s.
tic;
[indices, dists] = itqlsh(dataset, testset, param_itq, 'itq.lsh', 2, 10);
toc;
% Second time, no need to re-indexing. About 2s.
tic;
[indices, dists] = itqlsh(dataset, testset, param_itq, 'itq.lsh', 2, 10);
toc;
```

Chapter 4

Algorithm

LSHBOX is based on many approximate nearest neighbor schemes, and the following is a brief description of each algorithm and its parameters.

- **Locality-Sensitive Hashing Scheme Based on Random Bits Sampling**

✧ Reference

P. Indyk and R. Motwani. Approximate Nearest Neighbor - Towards Removing the Curse of Dimensionality. In Proceedings of the 30th Symposium on Theory of Computing, 1998, pp. 604-613.

A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), 1999.

✧ Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Binary code bytes
    unsigned N;
    /// The Difference between upper and lower bound of each dimension
    unsigned C;
};
```

Implementation

```
#include <lshbox/rbslsh.h>
```

✧ Notice

According to the second assumption in the paper, all coordinates of points in P are positive integer. Although we can convert all coordinates to integers by multiplying them by a suitably large number and rounding to the nearest integer, but I think it is very fussy, What's more, it often gets criticized for using too much memory when in a larger range of data. Therefore, it is recommended to use other algorithm.

- **Locality-Sensitive Hashing Scheme Based on Random Hyperplane**

✧ Reference

Charikar, M. S. 2002. Similarity estimation techniques from rounding algorithms. In Proceedings of the Thiry- Fourth

Annual ACM Symposium on theory of Computing (Montreal, Quebec, Canada, May 19 - 21, 2002). STOC '02. ACM, New York, NY, 380-388. DOI= http://doi.acm.org/10.1145/509907.509965

✧ Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Binary code bytes
    unsigned N;
};
```

✧ Implementation

```
#include <lshbox/rhplsh.h>
```

- **Locality-Sensitive Hashing Scheme Based on Thresholding**

✧ Reference

Zhe Wang, Wei Dong, William Josephson, Qin Lv, Moses Charikar, Kai Li. Sizing Sketches: A Rank-Based Analysis for Similarity Search. In Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems . San Diego, CA, USA. June 2007.

Qin Lv, Moses Charikar, Kai Li. Image Similarity Search with Compact Data Structures. In Proceedings of ACM 13th Conference on Information and Knowledge Management (CIKM), Washington D.C., USA. November 2004.

✧ Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Binary code bytes
    unsigned N;
    /// Upper bound of each dimension
    float Max;
```

```
    /// Lower bound of each dimension
    float Min;
};
```

✧ Implementation

```
#include <lshbox/thlsh.h>
```

● **Locality-Sensitive Hashing Scheme Based on p-Stable Distributions**

✧ Reference

Mayur Datar , Nicole Immorlica , Piotr Indyk , Vahab S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, Proceedings of the twentieth annual symposium on Computational geometry, June 08-11, 2004, Brooklyn, New York, USA.

✧ Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Index mode, you can choose 1(CAUCHY) or 2(GAUSSIAN)
    unsigned T;
    /// Window size
    float W;
};
```

✧ Implementation

```
#include <lshbox/psdlsh.h>
```

● **Spectral Hashing**

✧ Reference

Y. Weiss, A. Torralba, R. Fergus. Spectral Hashing. Advances in Neural Information Processing Systems, 2008.

✧ Parameters

```
struct Parameter
{
```

```
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Binary code bytes
    unsigned N;
    /// Size of vectors in train
    unsigned S;
};
```

- Implementation

```
#include <lshbox/shlsh.h>
```

- **Iterative Quantization**

- Reference

Gong Y, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(12): 2916-2929.

- Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Binary code bytes
    unsigned N;
    /// Size of vectors in train
    unsigned S;
    /// Training iterations
    unsigned I;
};
```

- Implementation

```
#include <lshbox/itqlsh.h>
```

- **Double-Bit Quantization Hashing**

Reference

Kong W, Li W. Double-Bit Quantization for Hashing. In AAAI, 2012.

Gong Y, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(12): 2916-2929.

Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Number of projection dimensions
    ///corresponding to 2*N binary code bytes for each vector
    unsigned N;
    /// Training iterations
    unsigned I;
};
```

Implementation

```
#include <lshbox/dbqlsh.h>
```

- **K-means Based Double-Bit Quantization Hashing**

Reference

Zhu, H. (2014). K-means based double-bit quantization for hashing. Computational Intelligence for Multimedia, Signal and Vision Processing (CIMSIVP), 2014 IEEE Symposium on (pp.1-5). IEEE.

Parameters

```
struct Parameter
{
    /// Hash table size
    unsigned M;
    /// Number of hash tables
    unsigned L;
```

```
    /// Dimension of the vector, it can be obtained from the instance of Matrix
    unsigned D;
    /// Number of projection dimensions
    ///corresponding to 2*N binary code bytes for each vector
    unsigned N;
    /// Training iterations
    unsigned I;
};
```

- Implementation

```
#include <lshbox/kdbqlsh.h>
```

- Notice

According to the test, Double-Bit Quantization Hashing and Iterative Quantization performance very good and are superior to other schemes. Iterative Quantization can get high query accuracy with minimum cost while Double-Bit Quantization Hashing can achieve better query accuracy.