


FastLSH: A Similarity Search Engine on High-Dimensional Big Data

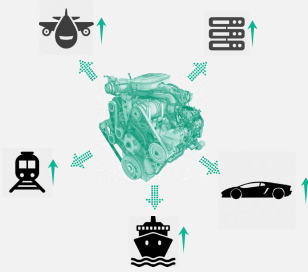
XU, Yaohai WONG, Nok Ching 

Similarity Search is an important component in Artificial Intelligence field. We proposed a flexible and high performance similarity Search Engine on High Dimensional Big data. The system can complete a query with one million base set in less than a hundred milliseconds.

The system aims at providing a **Fast, Flexible, Scalable and Extensible** solution for similarity search by its eight different packages. By providing a wide range of interfaces, user may pick the one best suits their own needs.

Objectives

- Developing a **Fast, Flexible, Scalable and Extensible** solution for similarity search
- Providing a visualization tool on big data to serves as a unique feature to assist data analysis
- Narrowing the technical gap by using Virtual Reality technology to visualize high dimensional data and demonstrate the algorithm in an interesting and attractive manner.



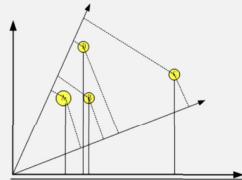
FastLSH is a powerful engine that can easily fit in and speed up various Artificial Intelligence models.

LSH algorithm

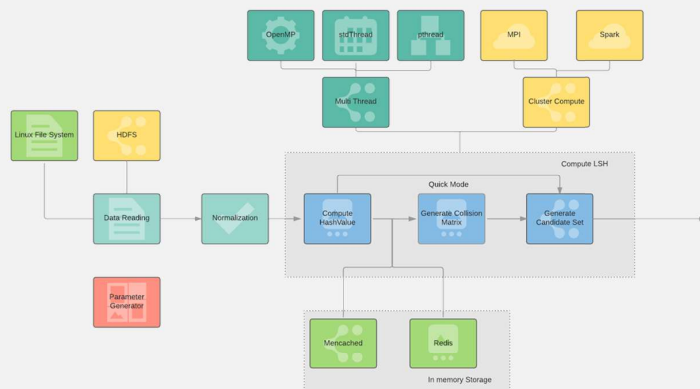
- Performing nearest neighbor search can be computational intensive and time consuming especially in high dimensional spaces.
- Numerous methods of approximate nearest neighbor search have been proposed like Locality sensitive hashing, Best bin first.
- The implementation of this project is based on a method of the LSH family, Collision Counting LSH, C2LSH, that is published by Gan et al. in 2012.
- This scheme has been found to be suitable in perform an estimation on the similarity between two points as well as providing range search for a query.
- This scheme has also proved to be guaranteeing on the query quality especially in high dimensional space.

LSH Definition A locality sensitive hash (LSH) is a hash family where similar items are more likely to collide. Formally, a hash family $\mathcal{H} = \{h : \mathcal{U} \rightarrow S\}$ is called (r_1, r_2, p_1, p_2) -locally sensitive if for all points $p, p' \in \mathcal{U}$,

1. if $d(p, p') \leq r_1$, then $\mathbb{P}[h(p) = h(p')] \geq p_1$.
2. if $d(p, p') > r_2$, then $\mathbb{P}[h(p) = h(p')] \leq p_2$.



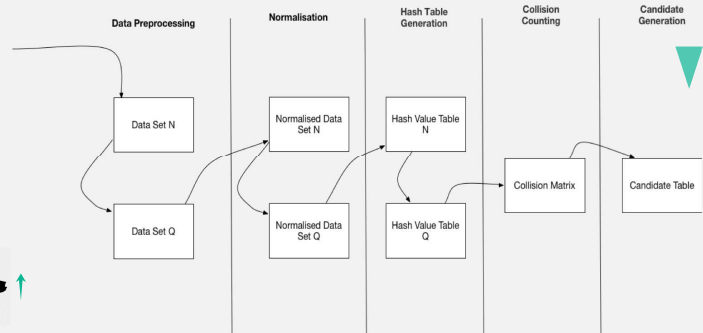
Note that this definition only makes sense if $r_1 < r_2$ and $p_1 > p_2$.









Module Relationship (Structure) of FastLSH

Performance Test

- Tests were being taken to measure the system performance following the normal execution procedure
- In Cluster & Multithread test, the test is only performed on the FastLSH-MPI and FastLSH-ESS by gradually increasing the threads used



The FastLSH Core System Execution Procedure

	 FastLSH ESS	 FastLSH Lib	 FastLSH Python	 FastLSH GPU	 FastLSH MPI	 FastLSH Spark
Data Preprocessing & Normalization Time(s)	20			12	23	24
Hash Table Generalization Time(s)	47			19	27	119
Collision Counting Time(s)	17			7	11	54
Candidate Generation Time(s)	0.1			0.1	0.2	0.6

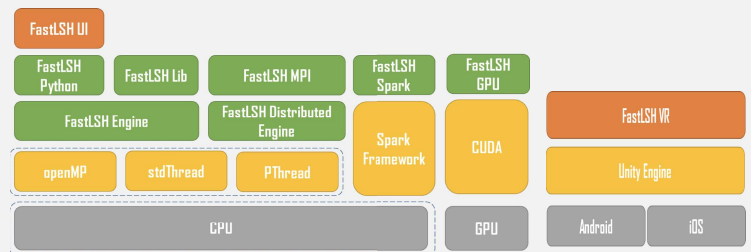
Performance test on execution time of each procedure

	FastLSH ESS	FastLSH Lib	FastLSH Python	FastLSH GPU	FastLSH MPI	FastLSH Spark
Query Time(s)	18			7	12	46
Query Time per single query	180ms			70ms	120ms	460ms

Performance test on the query time



- Both the multithread speedup and the cluster speedup stay close to the ideal line
- FastLSH-ESS line stay closer to its ideal line
- Overhead is shown to be larger in the FastLSH-MPI solution
- Cost is more expensive among nodes in the FastLSH-MPI solution
- Overhead will eventually higher than the performance gain as the cluster grows



The FastLSH System Stack

FastLSH Packages

Our big data similarity search explorer allows user to perform similarity search in many ways, including accessing it with web UI (FastLSH UI), using it as a C++/Python library (FastLSH Lib / FastLSH Python), entering Scala command (FastLSH Spark) as well as running it as a program on cluster (FastLSH MPI).

FastLSH-Lib

- FastLSH-Lib is a C++ library.
- FastLSH-Lib is importable to other C++ program.
- FastLSH-Lib provides the highest flexibility.
- It can make the FastLSH engine deeply and seamlessly integrated into the existing project.
- User can dive into the FastLSH engine, make changes per their needs or even tear it apart and redesign the workflow and the algorithm.
- FastLSH-Lib is made for development providing highest flexibility to allow user benefit most from the engine.

FastLSH-MPI

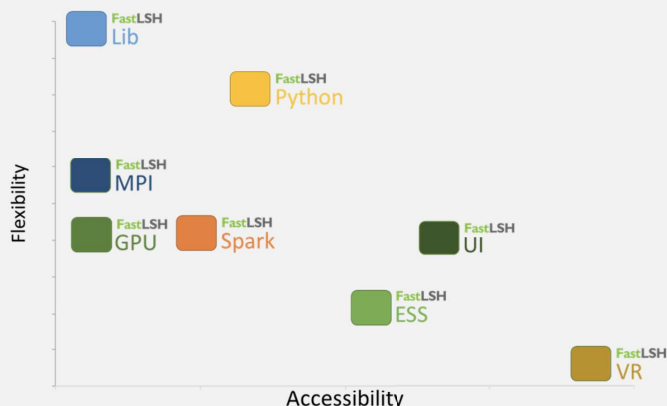
- FastLSH-MPI is a cluster computing version of FastLSH.
- It uses Message Passing Interface(MPI) to realize the communication among the nodes.
- It is a standalone package because the core engine is extensively changed.
- FastLSH-MPI makes the algorithm scalable among cluster with minimal communication cost.
- FastLSH-MPI intends to make the algorithm run in a cluster to provide high scalability and performance.

FastLSH-Spark

- FastLSH-Spark is a Spark implementation of the LSH.
- Functional programming style is adopted so the code is concise and clean.
- The access to Spark makes the FastLSH easy scalable.
- User can change and modify the workflow or the algorithm. It can also be plugged into the existing spark project.
- FastLSH-Spark intends to make FastLSH accessible to popular distributed data processing engine for its scalability.

FastLSH-GPU

- FastLSH-GPU is a CUDA implementation of FastLSH.
- It is empowered by the SIMD feature of GPU.
- The calculation is being run by thousands of threads and be processed simultaneously.
- Threads initiated can be scaled up as the dataset scales
- User can modify the program to fit their computation.



The accessibility and flexibility of each FastLSH Package

Everything is Open Source, find us on <https://github.com/FastLSH>

If you are a Data Miner.....

For Everyone

FastLSH-ESS

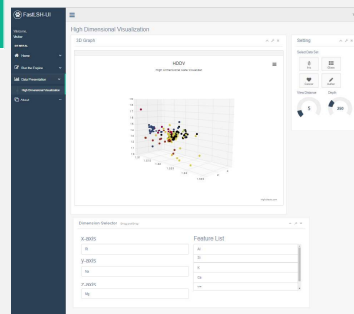
- FastLSH-ESS is the basic interface of FastLSH.
- ESS stands for essential.
- It is an instant application.
- It only performs the essential workflow of FastLSH to help user to locate highly similar points.

FastLSH-Python

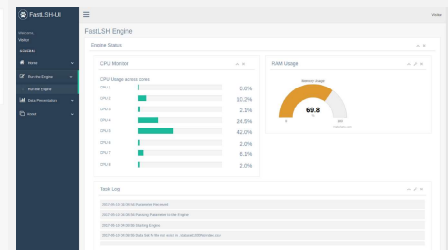
- FastLSH-Python is a Python module.
- It has the feature of both high accessibility and flexibility. The core of the module remains the C++ implementation. It serves as a bridge between lower level and higher level language.
- FastLSH-Python can be easily plug into tons of existing projects.
- FastLSH-Python intends to provide high level interface to serve more data science projects.

FastLSH-UI

- FastLSH-UI is a web interface of FastLSH.
- It provides a graphic interface for the core engine.
- Users can use the high-performance engine by the well organized and user friendly web interface.
- Users can concentrate on designing their model without being stopped by technical bar.
- Data visualization is also provided as a unique feature to assist data analysis.
- FastLSH-UI intends to provide a user-friendly interface with low technical bar for easy data analytic access.



FastLSH-UI -- High Dimension data visualizer



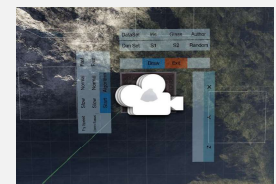
FastLSH-UI Monitoring the FastLSH status through control panel

FastLSH-VR

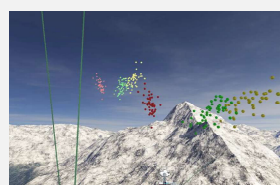
- FastLSH-VR is a Virtual Reality demonstration project for promotion and education purpose.
- Data points are placed in a 3D coordination system.
- User can move in the virtual environment by simply pressing the button and looking at the direction user wants to move.
- Animation simulated algorithm will be represented in virtual space to illustrate each step performed by the FastLSH package.
- Users will feel that they are immersing inside the data.
- The demonstration is implemented with Unity Engine.
- The virtual reality is achieved by Google Cardboard, a commodity VR viewer.
- The app supports both Android and iOS. By placing their phone inside the cardboard, they can enjoy the LSH algorithm demonstration without hassles.
- It requires no technical skill or any background knowledge.



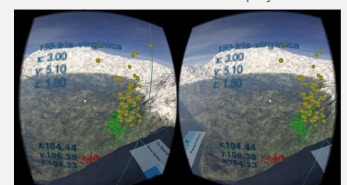
User feels standing on a flying carpet , fly around and observe the data points.



Control panels hovering around the user, allowing dataset selection and dimensions selection, changing speed and real-time statistic display



FastLSH-VR demonstrates the LSH projection



Rendered VR view of FastLSH-VR