A Generic Method for Accelerating LSH-based Similarity Join Processing

SPEAKER **Ms YU Chenyun**

PhD Student
Department of Computer Science
City University of Hong Kong
Hong Kong

DATE 12 April 2017 (Wednesday)
TIME 5:00 pm - 5:30 pm
VENUE CS Seminar Room, Y6405, 6th Floor
Yellow Zone, Academic 1
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong

ABSTRACT

Locality sensitive hashing (LSH) is an efficient method for solving the problem of approximate similarity search in high-dimensional spaces. Through LSH, a high-dimensional similarity join can be processed in the same way as hash join, making the cost of joining two large datasets linear. By judicially analyzing the properties of multiple LSH algorithms, we propose a generic method to speed up the process of joining two large datasets using LSH. The crux of our method lies in the way which we identify a set of representative points to reduce the number of LSH lookups. Theoretical analyses show that our proposed method can greatly reduce the number of lookup operations and retain the same result accuracy compared to executing LSH lookups for every query point. Furthermore, we demonstrate the generality of our method by showing that the same principle can be applied to LSH algorithms for three different metrics: the Euclidean distance (QALSH), Jaccard similarity measure (MinHash), and Hamming distance (sequence hashing). Results from experimental studies using real datasets confirm our error analyses and show significant improvements of our method over the state-of-the-art LSH method: to achieve over 0.95 recall, we only need to operate LSH lookups for at most 15% of the query points.

This paper has been published in the IEEE Transaction on Knowledge and Data Engineering (TKDE) and will be presented at the TKDE poster session of the 33rd IEEE International Conference on Data Engineering (ICDE 2017), April 19-22, 2017, San Diego, California, United States of America.

Supervisor: Dr Sarana Nutanong
Research Interests: data processing, query optimization, large-scale machine learning

All are welcome!

In case of questions, please contact Dr NUTANONG Sarana at Tel: 3442 7762, E-mail: snutanon@cityu.edu.hk, or visit the CS Departmental Seminar Web at http://www.cs.cityu.edu.hk/news/seminars/seminars.html.