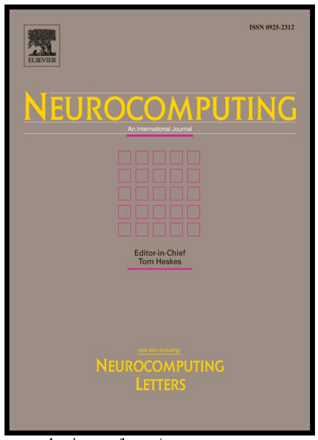
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

Bit Selection via Walks on Graph for Hash-based Nearest Neighbor Search

Ya Wang, Xianglong Liu, Danchen Zhang, Deqing Wang, Yuan Yang

PII: S0925-2312(16)30717-2
DOI: http://dx.doi.org/10.1016/j.neucom.2015.11.132
Reference: NEUCOM17329

To appear in: *Neurocomputing*

Received date: 14 September 2015
Revised date: 17 November 2015
Accepted date: 24 November 2015

Cite this article as: Ya Wang, Xianglong Liu, Danchen Zhang, Deqing Wang and Yuan Yang, Bit Selection via Walks on Graph for Hash-based Nearest Neighbor Search, *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2015.11.132

Bit Selection via Walks on Graph for Hash-based Nearest Neighbor Search

Ya Wang[a,b], Xianglong Liu[a,*], Danchen Zhang[c], Deqing Wang[a], Yuan Yang[b]

[a] *State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China*
[b] *School of Software, Beihang University, Beijing 100191, China*
[c] *School of Information Sciences, University of Pittsburgh, Pittsburgh 15260, U.S.A.*

Abstract

Recently hashing with multiple tables has become attractive in many real life applications, owing to its theoretical guarantee and practical success. To pursue the desired performance, usually great efforts are required on the hashing algorithm design for the specified scenario. Hash bit selection serves as a general method that can provide satisfying performance for different scenarios by utilizing existing hashing algorithms. In this paper, a novel bit selection framework via walks on graph is proposed to support both compact hash code generation and complementary hash table construction. It formulates the selection problem as the subgraphs discovery on an edge- and vertex- weighted graph, where the most desired subset corresponds to the frequently visited ones (bits/tables) in a Markov process. The framework is unified and compatible with different hashing algorithms. For compact code generation, it selects the most independent and informative hash bits using the Markov process over the candidate bit graph. For complementary hash table construction, it exploits the hierarchical authority relations among all candidate bits and separates them into a number of bit subsets as the candidate tables, from which multiple complementary hash tables can be efficiently selected. Experiments are conducted for two important selection scenarios, i.e., hashing using different hashing algorithms and hashing with multiple features. The results indicate that our proposed selection framework achieves significant performance gains over the naive selection methods under different scenarios.

Keywords: locality-sensitive hashing, bit selection, walk on graph, compact hash codes, complementary hash tables, Markov process

1. Introduction

Recent years have witnessed the explosive growth of visual data (image, video, etc.), which raises a great claim for fast nearest neighbor search (NNS). Compared to the conventional tree-based solutions like KD-Tree [1–3], hash-based ones have shown the promising performance in many applications like pose estimation [4], image search [5–8], active learning [9, 10], recommendation [11], graph search [12], etc., and thus attracted many attentions in the past decade. Locality-Sensitive Hashing (LSH) [13] first introduced the hash-based ANN research with the important concept named "locality sensitivity", which basically means that similar data points are hashed into similar codes. Since the binary hash codes can achieve compressed indexing and fast searching in a constant or sub-linear time, the following methods [5, 14**?** –25] mainly pursue binary hash functions to encode high dimensional data into binary codes, guaranteeing the locality sensitivity.

The random projection based LSH [14, 26] is able to efficiently produce hash bits and preserves the specific similarity metric like cosine, l_p-norm ($p \in (0, 2]$), etc. To further improve the code compactness, many well-learned hashing methods have been proposed to learn compact hash codes from the training data in the past few years. Among them, most methods place great efforts on the improvement of the coding quality and mutual complement of the hash functions. Generally, they archived this goal by using different hash types (*e.g.*, linear [19, 22**?** –25] and nonlinear [16, 20, 21, 27–29]), different settings (*e.g.*, unsupervised [5, 19, 21, 22, 25, 28, 30–32, 32**?** –34] and (semi-)supervised [8, 18, 20, 35]), and different optimizing techniques [7, 30, 33, 36–43]). The state-of-the-art hashing methods can achieve compressed storage and efficient computation using the compact codes. However, to satisfy the practical requirement for a balanced search performance, usually several hash tables are built that work together to cover more nearest neighbors. Till now, there are very a few related works regarding this problem [6, 37, 38, 41, 44–47], and meanwhile in practice most of these methods often require a huge number of tables without eliminating the table redundance.

Previous research have shown that to guaranteed efficient hash-based NNS, compact binary hash codes and complementary tables should be pursued, and many hashing methods have been proposed to focus on obtaining different well-worked hashing methods for different specific data sets and query scenarios. But for a specific application, designing a good hashing scheme is still time and labor consuming. A generic solution is highly desirable, so that can hash bits of high quality or complementary hash tables can be pursued adaptively for different applications. To address the problem, [48] first in-

*Corresponding author. Tel: +86 1082338094, Fax: +86 1082316736
Email address: xlliu@nlsde.buaa.edu.cn (Xianglong Liu)