High-Dimensional Similarity Query Processing for Data Science

Jianbin Qin
Shenzhen University & Shenzhen Institute of Computing Sciences
Shenzhen, Guangdong, China
qinjianbin@szu.edu.cn

Wei Wang
Hong Kong University of Science and Technology
Kowloon, HKSAR, China
weiwcs@ust.hk

Chuan Xiao
Osaka University & Nagoya University
Suita, Osaka, Japan
chuanx@ist.osaka-u.ac.jp

Ying Zhang
University of Technology Sydney
Ultimo, NSW, Australia
Ying.Zhang@uts.edu.au

Yaoshu Wang
Shenzhen University & Shenzhen Institute of Computing Sciences
Shenzhen, Guangdong, China
yaoshuw@sics.ac.cn

ABSTRACT

Similarity query (a.k.a. nearest neighbor query) processing has been an active research topic for several decades. It is an essential procedure in a wide range of applications (e.g., classification & regression, deduplication, image retrieval, and recommender systems). Recently, representation learning and auto-encoding methods as well as pre-trained models have gained popularity. They basically deal with dense high-dimensional data, and this trend brings new opportunities and challenges to similarity query processing. Meanwhile, new techniques have emerged to tackle this long-standing problem theoretically and empirically.

This tutorial aims to provide a comprehensive review of high-dimensional similarity query processing for data science. It introduces solutions from a variety of research communities, including data mining (DM), database (DB), machine learning (ML), computer vision (CV), natural language processing (NLP), and theoretical computer science (TCS), thereby highlighting the interplay between modern computer science and artificial intelligence technologies. We first discuss the importance of high-dimensional similarity query processing in data science applications, and then review query processing algorithms such as cover tree, locality sensitive hashing, product quantization, proximity graphs, as well as recent advancements such as learned indexes. We analyze their strengths and weaknesses and discuss the selection of algorithms in various application scenarios. Moreover, we consider the selectivity estimation of high-dimensional similarity queries, and show how researchers are bringing in state-of-the-art ML techniques to address this problem. We expect that this tutorial will provide an impetus towards new technologies for data science.

ACM Reference Format:
Jianbin Qin, Wei Wang, Chuan Xiao, Ying Zhang, and Yaoshu Wang. 2021. High-Dimensional Similarity Query Processing for Data Science. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470811

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470811

1 TUTORIAL OUTLINE

This tutorial consists of five parts. The first part motivates the need for high-dimensional similarity query processing and introduces basic concepts. The second and third parts delve into query processing algorithms. The fourth part covers selectivity estimation algorithms. The fifth part discusses future directions and open problems. Presentation slides are available at https://szudseg.github.io/kdd21-tutorial-high-dim-simqp/.

1.1 Background and Preliminaries

We first introduce the applications of high-dimensional similarity query processing in data science and explains its increasing importance. Then we describe basic concepts: (1) data models and the way of which we convert raw data (text, images, video, etc.) to high-dimensional data; (2) similarity/distance functions, mainly Hamming distance for binary vectors and Euclidean distance and cosine similarity (angular distance) for real-valued vectors; (3) query types, i.e., search and join queries, or thresholded and top-k (k-NN) queries, depending on the dimension of categorization; (4) a summary of the solutions that will be elaborated in this tutorial.

1.2 Exact Query Processing

Exact query processing methods aim to find all the results that satisfy the similarity constraint. Researchers are interested in this type of solutions as it does not pose any uncertainty to the pipelines that apply similarity query processing as a component. Representative methods are based on trees [10], space partitioning [17, 19, 25] and dimensionality reduction [1, 4].

1.3 Approximate Query Processing

It is commonly believed that it is hard to compute the exact results of queries with a sub-linear cost due to the curse of dimensionality. Instead, computing approximate results is sufficiently useful for many practical problems, and these solutions empirically achieve significantly higher efficiency and scalability than exact ones.

Locality Sensitive Hashing. Locality sensitive hashing (LSH) is a data-independent hashing approach with probabilistic guarantees on the worst-case performance [9]. It relies on a family of hash functions that map similar objects to the same hash codes with higher probability than dissimilar objects. Plenty of solutions have been proposed. Recent development focuses on supporting various similarity measures [29] and space-efficient indexing [22, 28].

Learning to Hash. Learning to hash (L2H) a data-dependent approach that maps original data to another (often Hamming) space by exploiting the data distribution. The underlying principle is to preserve the similarity information within an appropriate neighborhood. Additional heuristics and optimizations are often added to further reduce the information loss caused by the mapping or increase generalization to unseen data. Recent advancements feature deep learning in both supervised and unsupervised manner [3, 8, 13, 21]. Another line of methods is based on product quantization [11], with the unique ability to handle billions of objects.

Partition-based Methods. Methods in this category can be deemed as dividing the high-dimensional space into multiple disjoint regions. Partition is often carried out in a recursive way, so the index is represented by a tree or a forest. Notable methods are based on pivoting [27], hyperplane [2, 20], or compact partitioning such as cluster [7] or Voronoi diagram [16].

Neighborhood-based Methods. These methods construct a proximity graph where nodes represent objects and edges connect nearby objects. The main idea is to perform a search for similar objects atop the proximity graph. These methods achieve top accuracy and speed trade-off in empirical evaluations [12]. Notable methods include k-NN graph [5], hierarchical navigable small world [14], and navigating spreading-out graph [6].

1.4 Selectivity Estimation

Selectivity estimation outputs the approximate number of data objects that satisfy a selection criterion. Due to its use in density estimation, outlier detection, image retrieval, and query optimization, selectivity estimation for high-dimensional data has received considerable attention recently. Representative solutions are sampling [26] and kernel density estimation [15]. A recent trend is to formalize it as a regression task and utilize ML methods [23, 24].

1.5 Future Opportunities

We highlight a number of promising directions for future research: (1) It is interesting to explore ML models as approximate solutions (e.g., learning to index or learning to sample). (2) Answering composite queries (e.g., conjunctive queries) over multiple attributes will receive more attention. (3) Another direction is to develop efficient algorithms for query processing in data management systems, where ML, CV, and NLP techniques can help improve the quality.

2 PREVIOUS EDITIONS

The previous edition of this tutorial appeared at VLDB 2020 [18]. The new edition focuses on data science related applications (e.g., classification, regression, anomaly detection, and recommender systems). In addition, the new edition features the following new materials: (1) a thorough discussion on the use of similarity query processing in various application scenarios (e.g., the role of similarity queries in the entire workflow and the selection of algorithms), (2) more data models and a broader range of related works, and (3) more recent technical advancements and future trends.

Acknowledgements This work was supported by Guangdong Basic and Applied Basic Research Foundation 2020B1515120028, Guangdong Pearl River Recruitment Program of Talents 2019ZT08X603, Shenzhen Continuous Support Grant 20200811104054002, HKUST Red Bird Visting Scholar Program, JSPS Kakenhi 17H06099, 18H04093, and 19K11979, and ARC FT170100128 and DP210101393.

REFERENCES

[1] S. Berchtold, C. Böhm, and H. Kriegel. The pyramid-technique: Towards breaking the curse of dimensionality. In *SIGMOD*, pages 142–153, 1998.
[2] E. Bernhardsson. Annoy at github https://github.com/spotify/annoy, 2015.
[3] D. Cai, X. Gu, and C. Wang. A revisit on deep hashings for large-scale content based image retrieval. *CoRR*, abs/1711.06016, 2017.
[4] K. Chakrabarti and S. Mehrotra. Local dimensionality reduction: A new approach to indexing high dimensional spaces. In *VLDB*, pages 89–100, 2000.
[5] W. Dong, M. Charikar, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *WWW*, pages 577–586, 2011.
[6] C. Fu, C. Xiang, C. Wang, and D. Cai. Fast approximate nearest neighbor search with the navigating spreading-out graph. *PVLDB*, 12(5):461–474, 2019.
[7] K. Fukunaga and P. M. Narendra. A branch and bound algorithms for computing k-nearest neighbors. *IEEE Trans. Computers*, 24(7):750–753, 1975.
[8] N. Gao, M. Wilson, T. Vandal, W. Vinci, R. R. Nemani, and E. G. Rieffel. High-dimensional similarity search with quantum-assisted variational autoencoder. In *KDD*, pages 956–964, 2020.
[9] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, pages 604–613, 1998.
[10] M. Izbicki and C. R. Shelton. Faster cover trees. In *ICML*, pages 1162–1170, 2015.
[11] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, 2011.
[12] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin. Approximate nearest neighbor search on high dimensional data - experiments, analyses, and improvement. *IEEE Trans. Knowl. Data Eng.*, 32(8):1475–1488, 2020.
[13] J. Lu, V. E. Liong, and J. Zhou. Deep hashing for scalable image search. *IEEE Trans. Image Processing*, 26(5):2352–2367, 2017.
[14] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, 2020.
[15] M. Mattig, T. Fober, C. Beilschmidt, and B. Seeger. Kernel-based cardinality estimation on metric data. In *EDBT*, pages 349–360, 2018.
[16] G. Navarro. Searching in metric spaces by spatial approximation. *VLDB J.*, 11(1):28–46, 2002.
[17] M. Norouzi, A. Punjani, and D. J. Fleet. Fast exact search in hamming space with multi-index hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1107–1119, 2014.
[18] J. Qin, W. Wang, C. Xiao, and Y. Zhang. Similarity query processing for high-dimensional data. *PVLDB*, 13(12):3437–3440, 2020.
[19] J. Qin, C. Xiao, Y. Wang, W. Wang, X. Lin, Y. Ishikawa, and G. Wang. Generalizing the pigeonhole principle for similarity search in hamming space. *IEEE Trans. Knowl. Data Eng.*, 33(2):489–505, 2021.
[20] P. Ram and K. Sinha. Revisiting kd-tree for nearest neighbor search. In *KDD*, pages 1378–1388, 2019.
[21] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen. Unsupervised deep hashing with similarity-adaptive and discrete optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):3034–3044, 2018.
[22] Y. Sun, W. Wang, J. Qin, Y. Zhang, and X. Lin. SRS: solving c-approximate nearest neighbor queries in high dimensional euclidean space with a tiny index. *PVLDB*, 8(1):1–12, 2014.
[23] Y. Wang, C. Xiao, J. Qin, X. Cao, Y. Sun, W. Wang, and M. Onizuka. Monotonic cardinality estimation of similarity selection: A deep learning approach. In *SIGMOD*, pages 1197–1212, 2020.
[24] Y. Wang, C. Xiao, J. Qin, R. Mao, M. Onizuka, W. Wang, and R. Zhang. Consistent and flexible selectivity estimation for high-dimensional data. *CoRR*, abs/2005.09908, 2020.
[25] R. Weber, H. Schek, and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *VLDB*, pages 194–205, 1998.
[26] X. Wu, M. Charikar, and V. Natchu. Local density estimation in high dimensions. In *ICML*, pages 5293–5301, 2018.
[27] P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *SODA*, pages 311–321, 1993.
[28] B. Zheng, X. Zhao, L. Weng, N. Q. V. Hung, H. Liu, and C. S. Jensen. PM-LSH: A fast and accurate LSH framework for high-dimensional approximate NN search. *PVLDB*, 13(5):643–655, 2020.
[29] E. Zhu, F. Nargesian, K. Q. Pu, and R. J. Miller. LSH ensemble: Internet-scale domain search. *PVLDB*, 9(12):1185–1196, 2016.