August 2019

Tutorial: Are You My Neighbor?: Bringing Order to Neighbor Computing Problems

David Anastasiu
San Jose State University, danastasiu@scu.edu

Huzefa Rangwala
George Mason University

Andrea Tagarelli
University of Calabria

Recommended Citation

David Anastasiu, Huzefa Rangwala, and Andrea Tagarelli. "Tutorial: Are You My Neighbor?: Bringing Order to Neighbor Computing Problems" *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019). https://doi.org/10.1145/3292500.3332292

Tutorial: Are You My Neighbor? Bringing Order to Neighbor Computing Problems.

David C. Anastasiu*
San José State University
San José, CA, USA

Huzefa Rangwala
George Mason University
Fairfax, VA, USA

Andrea Tagarelli
University of Calabria
Rende (CS), Italy

ABSTRACT

Finding nearest neighbors is an important topic that has attracted much attention over the years and has applications in many fields, such as market basket analysis, plagiarism and anomaly detection, community detection, ligand-based virtual screening, etc. As data are easier and easier to collect, finding neighbors has become a potential bottleneck in analysis pipelines. Performing pairwise comparisons given the massive datasets of today is no longer feasible. The high computational complexity of the task has led researchers to develop approximate methods, which find many but not all of the nearest neighbors. Yet, for some types of data, efficient exact solutions have been found by carefully partitioning or filtering the search space in a way that avoids most unnecessary comparisons.

In recent years, there have been several fundamental advances in our ability to efficiently identify appropriate neighbors, especially in non-traditional data, such as graphs or document collections. In this tutorial, we provide an in-depth overview of recent methods for finding (nearest) neighbors, focusing on the intuition behind choices made in the design of those algorithms and on the utility of the methods in real-world applications. Our tutorial aims to provide a unifying view of "neighbor computing" problems, spanning from numerical data to graph data, from categorical data to sequential data, and related application scenarios. For each type of data, we will review the current state-of-the-art approaches used to identify neighbors and discuss how neighbor search methods are used to solve important problems.

CCS CONCEPTS

• **Information systems → Nearest-neighbor search**; • **Theory of computation → Nearest neighbor algorithms**.

KEYWORDS

nearest neighbor search, graph construction, sparse data.

ACM Reference Format:
David C. Anastasiu, Huzefa Rangwala, and Andrea Tagarelli. 2019. Tutorial: Are You My Neighbor? Bringing Order to Neighbor Computing Problems.. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332292

*Corresponding Author. Email: david.anastasiu@sjsu.edu

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332292

1 TUTORIAL OUTLINE

The tutorial (https://bit.ly/2VF5GFg) will first provide a formal definition of nearest neighbor search (NNS) and related problems and a summary of classical space partitioning-based approaches and their limitations. The importance of the NNS problem will be motivated by a series of application domains and methods that use NNS as a black-box kernel. The notion of neighbor is central for a plethora of problems and tasks in clustering, multi-label classification [11, 15, 17], anomaly detection, network/graph mining, recommender systems [1, 14, 19], bioinformatics and computational genomics [6, 9, 16], to name just a few. Depending on the domain, the notions of neighbor and neighborhood are key to enabling the modeling of a variety of phenomena; in the social network context, for example, such phenomena range from homophily effects to behavioral or state-transition dynamics, from information spread to information inference. Modeling such phenomena is essential for a number of classic problems in graph mining, such as community search, detection and evolution, link prediction, influence propagation, trust inference [12]. We will showcase the use of NNS methods as a key component in solving many of these problems.

The last few years have brought considerable advances in NNS methods. Some recent methods have focused on the discovery of new hash functions that, in the expectation, more closely relate objects [8, 13, 21]. Zhang and Zhang [21], for example, developed a hashing technique based on metric embeddings for edit distance that significantly outperforms all previous methods for DNA and other long string searches. Several works have shown that data-dependent hashing can outperform distribution-agnostic techniques. For metric spaces, a series of efficient approximate search methods rely on a navigable small world graph with nodes corresponding to the searched objects. Moreover, for objects best represented as sparse vectors, a variety of index traversal strategies achieve near-optimal search performance by leveraging the sparsity in the data and ignoring the majority of the object comparisons that are not similar enough [2–5, 13]. Anastasiu and Karypis, for example, developed several effective filtering techniques that leverage the Cauchy-Schwarz inequality applied to vector subspaces which enabled *exact* nearest neighbor graph construction methods to outperform even approximate solutions that were tuned to achieve at least 95% recall. Finally, some recent methods have expanded the scope of the search problem. Yu et al. [20] studied the case where the search has a limited computational budget, while Morales et al. [10] define a time-dependent similarity function for computing streaming similarity self-joins. In this tutorial, we will describe these and other recent state-of-the-art methods and provide guidance for choosing an appropriate method for finding neighbors in different domains.

2 TUTORS AND BIOGRAPHIES

David C. Anastasiu, Ph.D.: David C. Anastasiu is an Assistant Professor in the Department of Computer Engineering at San José State University. His research interests fall broadly at the intersection of machine learning, data mining, computational genomics, and high performance computing. Much of his work has been focused on scalable and efficient methods for analyzing sparse data, such as methods for identifying near neighbors, for searching related biochemical compounds, for characterizing how user behavior changes over time, and for personalized and collaborative presentation of Web search results. As a result of his algorithmic work in the area of Data Science, Prof. Anastasiu was awarded the Next Generation Data Scientist (NGDS) Award at the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA'2016). His work has been published in many top-tier conferences and journals, and he serves on the program committees of the most prominent IEEE and ACM data science-related conferences. His research is funded by NSF, Intel Labs, Flex, Infoblox, and NVIDIA Corporation. The tutorial will present material from a combination of his own research as it relates to efficient NNS methods and applications in traffic analytics and computational genomics.

Huzefa Ragwala, Ph.D.: Huzefa Rangwala is a Professor in the Department of Computer Science and Engineering, George Mason University. He received his Ph.D. in Computer Science from the University of Minnesota in 2008. His research interests include machine learning, learning analytics, bioinformatics and high performance computing. He is the recipient of the NSF Early Faculty Career Award in 2013, the 2014 GMU Teaching Excellence Award, the 2014 Mason Emerging Researcher Creator and Scholar Award, the 2013 Volgenau Outstanding Teaching Faculty Award, 2012 Computer Science Department Outstanding Teaching Faculty Award and 2011 Computer Science Department Outstanding Junior Researcher Award. His research is funded by NSF, NIH, NRL, DARPA, USDA and NVIDIA Corporation. The tutorial will present material from a combination of his own research as it relates to multi-instance learning and multi-task learning. He has presented well attended tutorials at SIAM SDM 2016, 2017 and ACM KDD 2017.

Andrea Tagarelli, Ph.D.: Andrea Tagarelli is an Associate Professor of Computer Engineering at the University of Calabria, Italy. He obtained his Ph.D. in Computer and Systems Engineering in 2006. His research interests include topics in data/text mining, machine learning, web and network science, information retrieval. He was program co-chair for the 2018 IEEE/ACM ASONAM conference, and co-organizer of workshops and a mini-symposium on clustering and other data-mining topics in premier conferences in the field (ECIR-16, ACM SIGKDD-13, SIAM DM-14, PAKDD-12, ECML-PKDD-11). He also presented well-attended tutorials on user behavior analysis and mining problems in social networks at WIMS-17, ACM UMAP-15, IEEE/ACM ASONAM-15. He is action editor for the Computational Intelligence Journal and associate editor for the Social Network Analysis and Mining Journal. The tutorial will present material from several of his works related to community detection and influence propagation.

ACKNOWLEDGMENTS

Part of this work was supported by NSF-1850557.

REFERENCES

[1] David C. Anastasiu, Evangelia Christakopoulou, Shaden Smith, Mohit Sharma, and George Karypis. 2016. Big Data and Recommender Systems. *Novática: Journal of the Spanish Computer Scientist Association* 240 (October 2016).

[2] David C. Anastasiu and George Karypis. 2014. L2AP: Fast Cosine Similarity Search With Prefix L-2 Norm Bounds. In *30th IEEE International Conference on Data Engineering (ICDE '14)*. 12.

[3] David C. Anastasiu and George Karypis. 2015. L2Knng: Fast Exact K-Nearest Neighbor Graph Construction with L2-Norm Pruning. In *24th ACM International Conference on Information and Knowledge Management (CIKM '15)*. 10.

[4] David C. Anastasiu and George Karypis. 2017. Efficient identification of Tanimoto nearest neighbors; All Pairs Similarity Search Using the Extended Jaccard Coefficient. *International Journal of Data Science and Analytics* 4, 3 (Nov 2017), 153–172.

[5] David C. Anastasiu and George Karypis. 2017. Parallel cosine nearest neighbor graph construction. *J. Parallel and Distrib. Comput.* (2017). https://doi.org/10.1016/j.jpdc.2017.11.016

[6] Wout Bittremieux, Pieter Meysman, William Stafford Noble, and Kris Laukens. 2018. Fast open modification spectral library searching through approximate nearest neighbor indexing. *bioRxiv* (may 2018), 326173. https://doi.org/10.1101/326173

[7] Antonio Caliò, Roberto Interdonato, Chiara Pulice, and Andrea Tagarelli. 2018. Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks. *IEEE Trans. Knowl. Data Eng.* 30, 12 (2018), 2421–2434. https://doi.org/10.1109/TKDE.2018.2820010

[8] Y. Cao, H. Qi, W. Zhou, J. Kato, K. Li, X. Liu, and J. Gui. 2018. Binary Hashing for Approximate Nearest Neighbor Search on Big Data: A Survey. *IEEE Access* 6 (2018), 2039–2054.

[9] Helen N. Catanese, Kelly A. Brayton, and Assefaw H. Gebremedhin. 2018. A nearest-neighbors network model for sequence data reveals new insight into genotype distribution of a pathogen. *BMC Bioinformatics* 19, 1 (12 Dec 2018), 475. https://doi.org/10.1186/s12859-018-2453-2

[10] Gianmarco De Francisci Morales and Aristides Gionis. 2016. Streaming Similarity Self-join. *Proc. VLDB Endow.* 9, 10 (June 2016), 792–803. https://doi.org/10.14778/2977797.2977805

[11] Nathan LaPierre, Mohammad Arifur Rahman, and Huzefa Rangwala. 2016. CAMIL: Clustering and Assembly with Multiple Instance Learning for phenotype prediction. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 33–40.

[12] Y. Li, J. Fan, Y. Wang, and K. Tan. 2018. Influence Maximization on Social Graphs: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (Oct 2018), 1852–1872.

[13] Yuliang Li, Jianguo Wang, Benjamin Pullman, Nuno Bandeira, and Yannis Papakonstantinou. 2019. Index-based, High-dimensional, Cosine Threshold Querying with Optimality Guarantees. In *22nd International Conference on Database Theory (ICDT), March 26-29* (2019-03-13). http://db.ucsd.edu/wp-content/uploads/2018/12/ICDT_2019_paper.pdf

[14] Xia Ning, Christian Desrosiers, and George Karypis. 2015. *A Comprehensive Survey of Neighborhood-Based Recommendation Methods*. Springer US, Boston, MA, 37–76. https://doi.org/10.1007/978-1-4899-7637-6_2

[15] Mohammad Arifur Rahman, Nathan LaPierre, and Huzefa Rangwala. 2017. Phenotype Prediction from Metagenomic Data Using Clustering and Assembly with Multiple Instance Learning (CAMIL). *IEEE/ACM transactions on computational biology and bioinformatics* (2017).

[16] Huzefa Rangwala, Anveshi Charuvaka, and Zeehasham Rasheed. 2014. Machine learning approaches for metagenomics. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, 512–515.

[17] Yukihiro Tagami. 2017. AnnexML: Approximate Nearest Neighbor Search for Extreme Multi-label Classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 455–464.

[18] Andrea Tagarelli, Alessia Amelio, and Francesco Gullo. 2017. Ensemble-based community detection in multilayer networks. *Data Min. Knowl. Discov.* 31, 5 (2017), 1506–1543. https://doi.org/10.1007/s10618-017-0528-8

[19] Jun Xu, Xiangnan He, and Hang Li. 2018. Deep Learning for Matching in Search and Recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 1365–1368.

[20] Hsiang-Fu Yu, Cho-Jui Hsieh, Qi Lei, and Inderjit S. Dhillon. 2017. A Greedy Approach for Budgeted Maximum Inner Product Search. In *Advances in Neural Information Processing Systems 30 (NIPS'17)*. 5459–5468.

[21] Haoyu Zhang and Qin Zhang. 2017. EmbedJoin: Efficient Edit Similarity Joins via Embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 585–594.