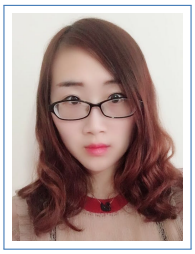
Yizhou Yan

+(1) 508-414-8049
yyan2@wpi.edu



Personal Info

Name: Yizhou Yan
Address: 87 Park Ave, APT 15, Worcester, MA
Mobile: +(1) 508-414-8049
Homepage: http://web.cs.wpi.edu/ yyan2
Email: yyan2@wpi.edu/yizhouyan9132@outlook.com

Education

2018.09-present **Visiting PhD Student in CSAIL**, *Massachusetts Institute of Technology*, Cambridge, MA, USA
2015.09-present **PhD in Computer Science**, *Worcester Polytechnic Institute*, Worcester, MA, USA
2013.09-2015.06 **M.E. in Software Engineering**, *Dalian University of Technology*, Dalian, China
2009.09-2013.06 **B.E. in Software Engineering**, *Dalian University of Technology*, Dalian, China
2009.09-2013.06 **B.A. in Japanese**, *Dalian University of Technology*, Dalian, China

Research Interests

My research focuses on ***big data analytics, data mining and machine learning***, in particular on algorithms and systems for scalable outlier detection from heterogeneous data sources including multi-dimensional data points, sequence data, and image data.

Publications

- **Conference Publications**
  1. **Yizhou Yan**, Lei Cao, Sam Madden, Elke Rundensteiner. "SWIFT: Mining Representative Patterns from Large Event Streams." (VLDB 2019)
  2. **Yizhou Yan**, Lei Cao, Elke Rundensteiner. "Scalable Top-n Local Outlier Detection." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM, Aug 2017.
  3. **Yizhou Yan**, Lei Cao, Caitlin Kuhlman, Elke Rundensteiner. "Distributed Local Outlier Detection in Big Data." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM, Aug 2017.
  4. **Yizhou Yan**, Lei Cao, Elke Rundensteiner. "Distributed Top-N Local Outlier Detection in Big Data." 2017 IEEE International Conference on Big Data. IEEE, Dec 2017.
  5. Lei Cao, **Yizhou Yan**, Caitlin Kuhlman, Qingyang Wang, Elke A. Rundensteiner, and Mohamed Eltabakh. "Multi-Tactic Distance-Based Outlier Detection." 2017 IEEE 33rd International Conference on Data Engineering (ICDE). IEEE, Apr 2017.
  6. Caitlin Kuhlman, **Yizhou Yan**, Lei Cao, Elke Rundensteiner. "Pivot-based Distributed K-Nearest Neighbor Mining." 2017 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Sep 2017.
  7. Yu Liu, Zhen Huang, **Yizhou Yan** and Yufeng Chen. "Science Navigation Map: an Interactive Data Mining Tool for Literature Analysis." *WWW' 15 Companion, Florence, Italy* May 18-22,2015
  8. Yu Liu, Zhen Huang, Jing Fang and **Yizhou Yan**. "An Article Level Metric in the Context of Research Community." *WWW'14 Companion, Seoul, Korea,* April 7-11, 2014.
- **Journal Article**

  Zhewen Shi, Yu Liu, **Yizhou Yan** and Xiaowei ZHAO. "A Hierarchical Community Detection Method in Complex Networks." *Journal of Computational Information Systems*, vol.9, no.24, pp. 9715-9724, 2013.
- **Manuscripts**
  1. **Yizhou Yan**, Lei Cao, Sam Madden, Elke Rundensteiner. "SeqDB: A database for Exploring Sequence Data." (In submission to SIGMOD 2019)
  2. Lei Cao, **Yizhou Yan**, Sam Madden, Elke Rundensteiner. "Efficient Discovery of Contextual Sequence Patterns." (In submission to VLDB 2019)
  3. Lei Cao, **Yizhou Yan**, Sam Madden, Elke Rundensteiner. "Outlier Detection from Image Data. " (In submission to ICLR 2019)

Posters

- Top-N Local Outlier Detection in Big Data, Graduate Research Innovation Exchange, WPI, 2017 (awarded final list).
- Distributed Local Outlier Detection in Big Data, Graduate Research Innovation Exchange, WPI, 2016 (awarded final list).

Awards and Certifications

2018.05 Honorable Mention for Teaching Assistant of the Year Award

2018.05 Best Teaching Assistant award in Computer Science

2017.11 Microsoft PhD Fellowship Program, awarded final list

2017.08 KDD 2017 student travel grant

2017.05 ICDE 2017 student travel grant

2014.10 Excellent Postgraduate Award (Top 5%)

2009.09-2012.06 Learning Merit Scholarship (Top 15%, twice; Top 5%, once), Individual Scholarship (Top 10%, once)

2012.02 Honorable Mention in ICM, USA

Technical Skills

Java, Python, C, C++, Scala, Matlab

Distributed Systems (Hadoop, Spark), MySQL, SQL Server, Oracle

Research Experiences

2018.04-2018.10 **SeqDB: A Database for Exploring Sequence Data.**
Presented a system that efficiently supports a rich variety of pattern exploration requests with varying pattern semantics and input parameter settings; provided a sequence pattern query language (PQL) that enables the users to issue pattern-related query and mining requests; designed an innovative index structure that succinctly summarizes sequence data using a small set of sequential patterns.

2017.12-2018.04 **Detecting Outliers from Image Data.**
Proposed the first approach that automatically discovers outliers from such image data. Our proposed strategy separates outliers from inliers without relying on any labeled outlier based on the idea that the intermediate states produced in the classification process can be effectively leveraged to serve as the outlierness measure of each image. We further enhance the outlier detection capacity of deep neural forest by introducing a new architecture that isolates the decision trees in the forest and a regularization on the loss function to penalize the large entropy routing decisions.

2017.05-2017.11 **Mining Representative Patterns from Large Event Streams.**
Proposed a new stream pattern mining semantics, called MDL-based Representative Pattern semantics (*MRP*). Then proposed an algorithm that continuously produces this representative pattern set upon event arrival as well as the batch update strategy that scales to high velocity event streams. Besides, the patterns mined in the history are compactly maintained by leveraging the multiple time granularity tilted-time window to efficiently monitor the changes of patterns and their trends.

2017.02-2017.07 **Typical and Outlier Pattern Mining in IoT Sequence Data.**
Proposed a new system called TOP to make sense of sequences by finding Typical and Outlier Patterns in IoT sequence data. Specifically, designed new frequent pattern semantics called Contextual Bi-frequent patterns (CBF patterns) supported by TOP. Developed customized algorithms for mining both CBF patterns as well as outliers that violate CBF patterns.

2016.06-2017.02 **Scalable Top-n Local Outlier Detection.**
Proposed the first scalable Top-N local outlier detection approach called TOLF. The key innovation of TOLF is a multi-granularity pruning strategy that quickly prunes most points from the set of potential outlier candidates. Designed customized density-aware indexing structure that effectively supports the pruning strategy, while accelerating the kNN search.

2015.11-2016.06 **Distributed Local Outlier Detection in Big Data.**
Proposed the first distributed solution for the Local Outlier Factor (LOF) method, namely DLOF, which is scalable to terabyte level data.

2015.09-2015.11 **Multi-Tactic Distance-Based Outlier Detection.**
Proposed the first distributed distance-based outlier detection approach using the MapReduce-based infrastructure, called DOD. Our experimental study confirms the efficiency of DOD and its scalability to terabytes of data – beating the baseline solutions by a factor of 20x.

Experiences

2018.05-2018.08 Internship at Microsoft Research (Redmond)

2015.09-2018.05 Teaching Assistant at Worcester Polytechnic Institute

2016.01-Present Research Assistant at Worcester Polytechnic Institute

2016.01-Present External Reviewer for SIGMOD (2017,2018), TKDE (2017), ICDE (2017), VLDB(2017), KDD(2018)

Reviewer for TKDD (2018), Journal of Distributed and Parallel Databases.

2015.04-2015.07 Internship at Baidu