Dr. Amit Awekar

Email: awekar@iitg.ac.in
Homepage: http://www.iitg.ac.in/awekar/
Phone: +91-361-258-2373
Address: Office Room Number 302, CSE Department, IIT, Guwahati, Assam, India 781039

Research Interests:

NLP: Structured representation of natural language text
Deep Learning: Data cleaning, Responsible Model compression
Data Mining: Incremental algorithms for dynamic datasets

Education:

Ph.D., Computer Science, North Carolina State University, Raleigh, NC, USA
Dissertation: *Fast, Incremental, and Scalable All Pairs Similarity Search*
Major area: Data Mining Advisor: Professor Nagiza F. Samatova
August 2005 – May 2010

M.Tech., Computer Science & Engineering, Indian Institute of Technology, Kanpur, Uttar Pradesh, India
Dissertation: *Selective Hypertext Induced Topic Search*
Major area: Data Mining Advisors: Professor Pabitra Mitra and Professor Harish Karnick
August 2003 – June 2005

B.E., Computer Engineering, Maharashtra Institute of Technology (Affiliated to Pune University), Pune, Maharashtra, India
August 1999 – June 2003

Honors and Awards:

2023	Best paper award in research track:
	ACM International Conference on Data Science and Management of Data
	Paper title: Surface Name Errors in Wikipedia
2011	IITG Microsoft Outstanding Young Faculty Award for one year from August 2011
2009	Certificate of Accomplishment in Teaching, North Carolina State University
2008	Outstanding Teaching Assistant Award, North Carolina State University
	Ranked among top 10 in over 500 teaching assistants
2006	Student Travel Grant for attending Workshop on Algorithms for Web Graph, Banff, Canada
2003	All India Rank 160 in Graduate Aptitude Test in Engineering
	Ranked among top-one percentile in over 37,000 students

Employment:

Indian Institute of Technology, Guwahati, India

10/2021– Present **Associate Professor**, Computer Science and Engineering
01/2011– 10/2021 **Assistant Professor**, Computer Science and Engineering

Indian Institute of Information Technology, Guwahati, India

08/2013–11/2013 **Guest Faculty**, Computer Science and Engineering

Maharashtra Institute of Technology, Pune, India

09/2010– 11/2010 **Visiting Assistant Professor**, Computer Engineering

North Carolina State University, Raleigh, NC, USA

08/2008– 12/2009 **Research Assistant**, Computer Science

08/2005– 05/2008 **Teaching Assistant**, Computer Science

Yahoo! Inc., Sunnyvale, CA, USA

01/2010– 02/2010 **Research Engineer**, Yahoo! Mail Anti-spam Team
05/2008– 07/2008, 05/2007 – 07/2007 **Summer Intern**, Yahoo! Mail Anti-spam Team
05/2006– 07/2006 **Summer Intern**, Yahoo! Research, Bangalore, India

Tata Institute of Fundamental Research, Pune, Maharashtra, India

05/2004– 07/2004 **Summer Intern**, Computational Mathematics Lab

Indian Institute of Technology, Kanpur, Uttar Pradesh, India

08/2003– 05/2005 **Teaching Assistant**, Computer Science and Engineering

Publications:

Google Scholar Profile DBLP Profile arXiv Profile

Summary by CORE Ranking: A* (3), A (8), Other (11)

Summary by publication venue: ECIR (4), WWW (2), CIKM (2), ACM CoDS (2), SIGIR (1), EACL (1), ACM HT (1), AKBC (1), Workshops (3), Others (5)

1. Surface Name Errors in Wikipedia
To appear in proceedings of the 10th ACM International Conference on Data Science and Management of Data (Mumbai, India January 4-7, 2023)
Anuj Khare, and **Amit Awekar**.
arXiv preprint:
DOI: https://doi.org/10.1145/3570991.3571043
Best short paper award

2. Scaling-up Mass Based Clustering
In proceedings of the 31st ACM International Conference on Information and Knowledge Management (Atlanta, USA October 17-21, 2022)
Nidhi Ahlawat, and **Amit Awekar.**
arXiv preprint:
DOI: https://doi.org/10.1145/3511808.3557691

3. Improving Relation Classification Using Relation Hierarchy.
In proceedings of the 27th International Conference on Natural Language & Information Systems (Valencia, Spain June 15-17, 2022)
Akshay Parekh, Ashish Anand, and **Amit Awekar**.
arXiv preprint:
DOI: https://doi.org/10.1007/978-3-031-08473-7_29

4. Are Word Embedding Methods Stable and Should We Care About It?
In proceedings of the 32nd ACM Conference on Hypertext and Social Media (Dublin, Ireland August 30 – September 02, 2021)
Angana Borah, Manash Pratim Barman, and **Amit Awekar**.
arXiv preprint: https://arxiv.org/abs/2104.08433
DOI: https://doi.org/10.1145/3465336.3475098

5. Taxonomical Hierarchy of Canonicalized Relations from Multiple Knowledge Bases.
In proceedings of the 7th ACM IKDD CoDS and 25th COMAD (Hyderabad, India January 5-7, 2020)
Akshay Parekh, Ashish Anand, and **Amit Awekar**.
arXiv preprint: https://arxiv.org/abs/1909.06249
DOI: https://doi.org/10.1145/3371158.3371186

6. Mining Strengths and Weaknesses of Cricket Players Using Short Text Commentary.
In proceedings of the 18th IEEE International Conference on Machine Learning and Applications (Boca Raton, Florida, USA December 16-19, 2019)
Swarup Ranjan Behera, Parag Agrawal, Saradhi Vijaya V, and **Amit Awekar**.
DOI: https://doi.org/10.1109/ICMLA.2019.00122

7. Collective Learning from Diverse Datasets for Entity Typing in the Wild.
In proceedings of the 2nd International Workshop on EntitY REtrieval, co-located with CIKM 2019 (Beijing, China, November 03, 2019)
Abhishek, Amar Prakash Azad, Balaji Ganesan, Ashish Anand, and **Amit Awekar**.
arXiv preprint: https://arxiv.org/abs/1810.08782

8. Decoding the Style and Bias of Song Lyrics.
In proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (Paris, France, July 21-25, 2019)
Manash Pratim Barman, **Amit Awekar**, and Sambhav Kothari.
arXiv preprint: https://arxiv.org/abs/1907.07818
DOI: https://doi.org/10.1145/3331184.3331363

9. Fine-grained Entity Recognition with Reduced False Negatives and Large Type Coverage.
In Proceedings of the Automated Knowledge Base Construction Conference (Amherst, Massachusetts, USA, May 20-22, 2019)
Abhishek, Sanya Bathla Taneja, Garima Malik, Ashish Anand, and **Amit Awekar**.
arXiv preprint: https://arxiv.org/abs/1904.13178
DOI: https://doi.org/10.24432/C5QP4T

10. It’s Only Words And Words Are All I Have.
In Proceedings of the European Conference on Information Retrieval (Cologne, Germany, April 14-18, 2019)
Manash Pratim Barman, Kavish Dahekar, Abhinav Anshuman, and **Amit Awekar**.
arXiv preprint: https://arxiv.org/abs/1901.05227
DOI: https://doi.org/10.1007/978-3-030-15719-7_4

11. Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms.
In Proceedings of the European Conference on Information Retrieval (Grenoble, France, April 25-29, 2018)
Sweta Agrawal and **Amit Awekar.**
arXiv preprint: https://arxiv.org/abs/1801.06482
DOI: https://doi.org/10.1007/978-3-319-76941-7_11

12. On Low Overlap Among Search Results of Academic Search Engines.
In Proceedings of the International World Wide Web Conference (Perth, Australia, April 3-7, 2017)
Anasua Mitra and **Amit Awekar.**
arXiv preprint: https://arxiv.org/abs/1701.02617
DOI: https://doi.org/10.1145/3041021.3054265

13. Fine-Grained Entity Type Classification by Jointly Learning Representations and Label Embeddings.
In Proceedings of the Conference of European Chapter of the Association for Computational Linguistics (Valencia, Spain, April 3-7, 2017)
Abhishek Patel, Ashish Anand, and **Amit Awekar.**
arXiv preprint: https://arxiv.org/abs/1702.06709
DOI: https://doi.org/10.18653/v1/e17-1075

14. Faster K-Means Cluster Estimation.
In Proceedings of the European Conference on Information Retrieval (Aberdeen, UK, April 8-13, 2017)
Siddhesh Khandelwal and **Amit Awekar.**
arXiv preprint: https://arxiv.org/abs/1701.04600
DOI: https://doi.org/10.1007/978-3-319-56608-5_43

15. Batch Incremental Shared Nearest Neighbor Density-Based Clustering Algorithm for Dynamic Datasets.
In Proceedings of the European Conference on Information Retrieval (Aberdeen, UK, April 8-13, 2017)
Panthadeep Bhattacharjee and **Amit Awekar.**
arXiv preprint: https://arxiv.org/abs/1701.09049
DOI: https://doi.org/10.1007/978-3-319-56608-5_50

16. Incremental Shared Nearest Neighbor Density-Based Clustering.
In Proceedings of the ACM International Conference on Information and Knowledge Management (San Francisco, USA, October 27-November 01, 2013)
Sumeet Kumar Singh, and **Amit Awekar.**
DOI: https://doi.org/10.1145/2505515.2507837

17. Mutual Exclusion Rule Mining from Transaction Databases.
First Indian Workshop on Machine Learning (IIT Kanpur, India, July 1-2, 2013)
Hardik Modi, and **Amit Awekar**.

18. Parallel all pairs similarity search.
In Proceedings of the International Conference on Information and Knowledge Engineering (Las Vegas, Nevada, USA, July 18-21, 2011)
Amit Awekar, and Nagiza F. Samatova.

19. Incremental all pairs similarity search with Reduced I/O Overhead.
In Proceedings of the International Conference on Information and Knowledge Engineering (Las Vegas, Nevada, USA, July 13-17, 2009)
Amit Awekar, Nagiza F. Samatova, and Paul Breimyer.

20. Incremental all pairs similarity search.
In Proceedings of the Third Workshop on Social Network Mining and Analysis, Held in Conjunction with the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Paris, France June 28, 2009)
Amit Awekar, Nagiza F. Samatova, and Paul Breimyer.

21. Fast matching for all pairs similarity search.
In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Milan, Italy, September 15-18, 2009)
Amit Awekar, and Nagiza F. Samatova.

22. Selective hypertext induced topic search.
In Proceedings of the 15th international Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006)
Amit Awekar, Pabitra Mitra, and Jaewoo Kang.

Funded Projects:

Structured Representation of Biomedical Text
Status: Ongoing (March 2023 – February 2024)
Budget: 2,500,000 rupees
Collaborator: Dr. Ashish Anand
Funding: Eli Lilly India Private Limited
Deliverables: Deep Learning models for biomedical entity recognition and linking

Efficient Deep Learning Models for Underwater Exploration
Status: Ongoing (March 2023 – February 2026)
Budget: 3,500,000 rupees
Collaborator: Dr. Ashish Anand
Funding: IIT Guwahati Technology Innovation and Development Foundation
Deliverables: Methods for responsible compression of Deep Learning models

Addressing the Bottlenecks of Peer Review Systems
Status: Completed (September 2019 – March 2021)
Budget: 1,400,000 rupees
Collaborator: Dr. Ashish Anand
Funding: Digiscape Tech Solutions Limited
Deliverables: Modules for peer review system for scientific paper publication

Algorithms for Graph Similarity Self-Join
Status: Completed (June 2018 – June 2021)
Budget: 660,000 rupees
Funding Agency: Science and Engineering Research Board, DST
Deliverables: Algorithms for finding near duplicates in graph datasets

Kanimuni: Set of Educational Games for School Students
Status: Completed (April 2015 – March 2020)
Budget: 2,000,000 rupees
Collaborators: Dr. Prasad Bokil, Dr. Sheetal Gokhale, and Dr. Srinivasan (IIT Guwahati)
Funding Agency: Design Innovation Center, IIT Guwahati
Deliverables: Web based and stand-alone educational games

Infrastructure for Mining Collaborative Knowledge Repositories.
Status: Completed (September 2011 – August 2013)
Budget: 500,000 rupees
Funding Agency: Start up Grant, IIT Guwahati
Deliverables: Toolkit and APIs for mining Wikipedia and other open collaborative knowledge repositories

Teaching:

- Introduction to Computing: Spring 23, 18, 12
- Data Structures: Monsoon 21, 20, 19, 18, 13
- Database Management Systems: Monsoon 11, 06, Summer 11, Spring 22, 20, 19, 17, 16, 15
- Automata, Grammar, and Computability: Spring 08
- Algorithms (CSE Minor): Spring 14, 13
- Data Mining: Spring 21, Monsoon 17, 16, 15, 14, 12
- Natural Language Processing: Monsoon 22, 20
- Mathematics for Computer Science: Summer 13
- Short term courses and workshops
  - Applied Deep Learning (co-organized with Dr. V. Vijaya Saradhi): Monsoon 2022
  - Deep Learning for NLP (co-organized with Dr. Ashish Anand): Monsoon 2019
  - Society and the Web (co-organized with Dr. Ranbir Singh): Monsoon 2011

Professional Service:

Program Committee Member

- AAAI: 20
- ACM IKDD CODS-COMAD: 23, 22, 21, 20, 18 (Co-Chair: Young Researchers' Symposium)
- IEEE International Conference on Tools with Artificial Intelligence: 17, 16 (Area Chair: Data Mining)
- ACM Conference on Hypertext and Social Media: 14
- International World Wide Web Conference (Demo track):, 14
- Indian International Conference on Artificial Intelligence, Bangalore, India: 11
- Second Warm-up Workshop for World Wide Web 2011 Conference, Kolkata, India: 10
- Symposium for Graduate Research, North Carolina State University, Raleigh, NC, USA: 09

Reviewer

- Transactions on Knowledge and Data Engineering, IEEE: 2016-2022
- Science and Engineering Research Board: 2017-2019
- Transactions on ICT, Computer Society of India: 2015-2017
- Defence Science Journal, DRDO, 2016
- International Journal on Artificial Intelligence Tools, World Scientific: 2016, 2015
- International Conference on Computer and Communication Technology, Allahabad, India, 2011
- International Conference on Parallel Processing, Vienna, Austria, 2009

Invited Talks and Presentations:

- Algorithms for web-scale problems, QIP short term course on data structures and algorithms, IIT Guwahati, July 2011

Students

Ph.D.

Rohit Raj Rai: Compression of Deep Learning Models
(July 2022 - present)

Nidhi Ahlawat: Isolation Forest based unsupervised learning algorithms for large and dynamic datasets (December 2018 - present)

Akshay Parekh: Understanding and mitigation of noise in crowd-sourced relation classification data (December 2017 – January 2023, co-advising with Professor Ashish Anand, Thesis submitted, First job: Eli Lilly)

Abhishek: Multi-domain fine-grained entity recognition
(December 2015 – July 2020, co-advised with Professor Ashish Anand, First job: Faculty Member, BITS Pilani)

M.Tech.

2022-23
Rishab Deo Singh: Mining dynamic graphs
Kishore M: Entity recognition in biomedical texts
Sama Rohith Reddy: High performance algorithms for near duplicate detection in graph datasets
Ashish Dev:

2021-22
Anuj Khare: Surface Name Error correction in Wikipedia across multiple languages (First job: VMware)
Anil Singh: Comparative Analysis of Datasets and Models for Fine-grained Entity Recognition (First job: IBM)
Akshat Jain: Comparative Analysis of Models and Datasets for Biomedical Question Answering (First job: Mercedes-Benz)

Stuti Priyambda: Near Duplicate Detection in Labelled Graph Datasets (First job: Flipkart)

2020-21
Kapil Kukreja: Surface Name Errors in Wikipedia: Identification and Correction (First job: Microsoft)
Rahul Vats: Using vertex-edge overlap as a proxy for graph edit distance to find near duplicates in graph datasets (First job: Oracle)
Manish Gupta: Similarity computation for short text data (First job: Paytm)

2019-20
Ayush Jaiswal: Near-duplicate detection across multiple graph datasets (First job: Oracle)
Kushal Kumar Dey: Visualization scientific papers (First job: Microsoft)

2018-19
Priya Badchariya: Near-duplicate detection in graph datasets using vertex-edge overlap similarity measure (First job: SAP Labs)
Pammi Sairam: Near-duplicate detection in graph datasets using sequence similarity measure (First job: Oracle)
Divyam Lamiyan: Analysis of hate speech in Twitter and Instagram (First job: ThoughtSpot)

2017-18
Abhinav Anshuman: Deep learning-based models for English song lyrics mining (First job: Dell)
Aditya Gaurav: Analysis of social media presence of various armed forces (First job: Cisco)

2016-17
Kavish Dahekar: Analysis of English song lyrics over last five decades (First job: SAP Labs)
Vinayak Jadhav: WayOut: An educational game for learning directions (First job: SAP Labs)
Nihal Jain: Mining frequent disjunctive itemsets (First job: Huawei)
Adish: Perception management for Indian Army using social network analysis
Pawan Singre: Data infrastructure for social network analysis (First job: Agility E Services)

2013-14
Kunj Kothari: Incremental mutual exclusion rule mining (First job: Cognizant)
Prayag Surendran: Open source toolkit for Wikipedia mining (First job: Myntra)
Shailesh Prajapati: FP tree-based algorithms for mutual exclusion rule mining (First job: Oracle)

2012-13
Hardik Modi: Mutual exclusion rule mining in transaction datasets (First job: Microsoft)
Goutam Das: Scalable APIs for Wikipedia mining (First job: Cisco)

2011-12
Apurba Paul: Classifying online question answering discussions as open or resolved (First job: Oracle)
Manoj Singh Chauhan: Data management APIs for Wikipedia mining (First job: CDOT)

B. Tech.
2022-23
Suryansh Singh
Tanishq Katare and Keshav Chourasiya
Swastika Gupta

2021-22

Aryan Chauhan and Rishikesh Songra: Language models for grammatical evaluation
Khandesh Sai Lokesh and Jagana Vineeth: Analysis of datasets for fine-grained entity recognition

2020-21

Ritam Majumdar: Analysis of errors in Wikipedia surface names (First job: BNY Mellon)
Anubhav Tyagi: Graph based image retrieval (First job: Goldman Sachs)
Bhavnick Singh: Trends in the use of social media by national armed forces

2019-20

Yagyansh Bhatia and Akhil Chandra Pnachumarthi: Code search engine using embedding methods
Arpan Konar and Ayush Sanjay Agarwal: Automated paper-reviewer matching system
Prashanth Ravichandar: Categorizing the errors in Wikipedia surface names

2018-19

Srikar Paruchuru and Chandan Reddy: Hierarchical model for graph representation learning
Kushal KSVS and Dharmesh Chourasiya: Data dependent dissimilarity computation in dynamic datasets
Nityanad Rai and Abhinav Bollam (co-advising with Professor S. K. Bose): Efficient graph similarity search using graph edit distance

2017-18

Sambhav Kothari: Synthetic dataset generation for fine-grained entity mining (First job: Bloomberg)
Yash Pote: Speeding up neural network training by identifying redundant training examples (First job: National University of Singapore)
Akash Dupare: Voca, An educational social game to improve language skills (First job: Honeywell Technology Solutions)
Nitish Garg: Limitations of existing algorithms for graph similarity self-join (First job: DE Shaw)

2016-17

Sweta Agrawal: Deep learning for detecting cyberbullying across multiple social media platforms(First job: Adobe)
Pritam Sarkar: MagMates: An educational game for learning magnetism (First job: Medlife)
Rahul Kumar Gond: Pologono: An educational game for learning shapes

2015-16

Shriraj Bhardwaj: ChimieRush: A social game for learning periodic table (First job: Adobe)
Parag Adhau: Frequent item set mining using node sets (First job: Snapdeal)
Siddhesh Khandelwal: Heuristics for speeding up k-means clustering (First job: Research Assistant, IISc)
Pulkit Arora: Analysis of edit history of Wikipedia in Indian languages (First job: Microsoft)

2013-14

Rishikesh Ghewari: Anytime algorithms for association rule mining (First job: Ebay)
Pydi Prasanna: Independence rule mining (First job: Samsung)

2012-13

Snehlata: Incremental association rule mining (First job: Microsoft)
Sumeet Kumar Singh: Incremental shared nearest neighbor density based clustering (First job: Microsoft)

Last updated on September 2022

N. Vishnu Teja: Incremental ROCK-Robust Clustering Algorithm for Categorical Attributes (First job: Goldman Sachs)

2011-12
Dhruv Sharma: Near duplicate entity detection in text databases (First job: MS, UC Irvine)
Sumit Raj: Similarity search in time series databases (First job: MS, University of Minnesota)
Chinmaya Poswalia: Search engine for IITG intranet (First job: Amazon)

Department and Institute Service

- Department Research Area coordinator for Machine Learning
- Department Admission Committee (September 2017 - March 2020)
- Department Undergraduate Program Committee Member (July 2015 to October 2017)
- Department Post-graduate Program Committee Member (July 2022- Present, April 2013 - July 2014)
- Department Timetable Coordinator (March 2022 - Present)
- Department Library (July 2012 - December 2014)

Last updated on September 2022