Subhabrata Mukherjee

Contact Information

Seattle, USA
WA 98109

Voice: (+1) 206-465-0093
E-mail: subhabrata.mukherjee.ju@gmail.com
WWW: https://people.mpi-inf.mpg.de/∼smukherjee/

Research Interests

Information Extraction, Natural Language Understanding, Probabilistic Graphical Models and Generative Models, Representation Learning, Fact-checking, Recommendation Systems

Education

Max Planck Institute for Informatics, Saarbrücken, Germany

Ph.D., Computer Science and Engineering, **November 2013 - March 2017**

- Dissertation Topic: "Probabilistic Graphical Models for Credibility Analysis in Evolving Online Communities". Developed machine learning models that exploit the joint interaction between multiple factors (e.g., textual content, social network, latent topics etc.) to extract high quality knowledge from (noisy) user-generated content.
- Advisor: **Prof. Gerhard Weikum**
- Dissertation Committee: Prof. Gerhard Weikum, Prof. Jiawei Han, Prof. Stephan Günnemann, Prof. Dietrich Klakow
- Grade: **summa cum laude**
- Honors: **SIGKDD Doctoral Dissertation Runner-up Award** (one of the top-3 best doctoral dissertations world-wide in data mining, SIGKDD 2018)

Indian Institute of Technology (IIT - Bombay), Mumbai, India

M.Tech., Computer Science and Engineering, July 2012

- Dissertation Topic: "Adaptation of Sentiment Analysis to New Linguistic Features, Informal Language Form and World Knowledge"
- Advisor: **Prof. Pushpak Bhattacharyya**
- CGPA 9.62, Maximum 10.0

Jadavpur University, Kolkata, India

B.Tech., Computer Science and Engineering, July 2010

- CGPA 8.73, Maximum 10.0

Professional Experience

Amazon, Product Graph (Machine Learning Science), Seattle, USA

Machine Learning Scientist **October 2017 - now**

Working on building the Amazon Product Knowledge Graph — the authoritative knowledge base for every product in the world. Building large-scale machine learning and deep learning models for open information extraction and representation learning. Specific projects include:

- Developed *OpenTag* (KDD 2018) bringing deep learning and active learning together for state-of-the-art imputation and entity extraction improving Catalog coverage by ∼50% in production.
- Developed *OpenKI* (NAACL 2019) for integrating web-scale open information extraction and knowledge bases with relation inference and neighborhood encodings.
- Common sense knowledge discovery and embedding with extreme multi label (XML) classification from text and knowledge bases.

Max Planck Institute for Informatics, Saarbrücken, Germany

Researcher **March 2017 - September 2017**

Worked on credibility analysis, recommender systems, and influence networks.

Google Research, Machine Learning and Intelligence, Mountain View, California USA

Intern **August 2015 - December 2015**

Worked on semantic annotation of large-scale datasets (audio, video, web-tables, map-reduce job logs etc.) with Knowledge Graph to improve Google Dataset Search (GOODS).

IBM Research Lab, Human Language Technologies, Delhi, India
Research Engineer **October 2012 - October 2013**
Worked on unsupervised domain ontology construction from corpus to improve Question-Answering systems (Watson); self-assist systems as virtual call center agents to guide a customer in performing various domain-dependent tasks; intent classification of voice queries on mobile devices; generative models for personalized recommendation.

Teaching Experience

Indian Institute of Technology (IIT-Bombay), Mumbai India

Teaching Assistant for CS 725 (Foundations of Machine Learning), CS 626 and CS 460 (Natural language Processing and the Web), CS 101 (Computer Programming)

Academic Service (Organizer, Reviewer, PC)

Organizer: SIGKDD 2019 Workshop on “Truth Discovery and Fact Checking: Theory and Practice”, Domain Specific Speech and Language Understanding Workshop, Amazon Machine Learning Conference (AMLC 2018); Knowledge Graphs: Construction, Management and Querying, Semantic Web Journal (Editorial Board Member)
PC: SIGKDD 2019 (Research, Applied Data Science Tracks), National Science Foundation (NSF), Amazon Research Awards (ARA 2017, 2018), Amazon Machine Learning Conference (AMLC 2018), Humanizing Artificial Intelligence (IJCAI 2018, 2019), Natural Language Interfaces for Web of Data (ISWC 2018, 2019), Exploiting AI for Data Management Systems (SIGMOD 2018), Interactive Data Exploration and Analytics (SIGKDD 2017), Social Aspects in Personalization and Search (ECIR 2018)
Reviewer: PLOS One, ACM Transactions on Knowledge Discovery from Data (TKDD), IEEE Transactions on Knowledge and Data Engineering (TKDE), Information Systems (Journal), Data Mining and Knowledge Discovery (DAMI), Artificial Intelligence (Journal), IEEE Transactions on Computational Social Systems (TCSS), Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Journal of Web Semantics, Journal of Human-Computer Studies

Invited / Award Talks

1. **Deep Learning for Knowledge Extraction and Integration to build the Amazon Product Graph**, Knowledge Graph Conference, Columbia University, USA, 2019 (Upcoming)
2. **Modeling Joint Interactions for Information Extraction**
   - Microsoft Research AI (MSR AI), Redmond, USA, 2019
   - Google AI, New York, USA, 2019
3. **Probabilistic Graphical Models for Credibility Analysis in Evolving Online Communities**
   - SIGKDD Doctoral Dissertation Award Talk, London, UK, August 2018
   - MIT Media Lab, Cambridge, USA, December 2016
   - Amazon, Seattle, USA, December 2016
   - Bell Labs, Cambridge, UK, November 2016
   - IBM Research Lab, Zurich, Switzerland, August 2016

Workshops / Tutorials

- **SIGKDD 2019** Workshop on “Truth Discovery and Fact Checking: Theory and Practice” with Qi Li (UIUC), Cong Yu (Google Research) and Jiawei Han (UIUC)
- **SIGKDD 2018** Tutorial on **Fact Checking: Theory and Practise** with Xin Luna Dong (Amazon), Christos Faloutsos (Amazon/CMU), Xian Li (Amazon), Prashant Shiralkar (Amazon)
- **SIGKDD 2018** Tutorial on **Graph and Tensor Mining for Fun and Profit** with Xin Luna Dong (Amazon), Christos Faloutsos (Amazon/CMU), Andrey Kan (Amazon), Jun Ma (Amazon)

Publications

[DBLP] [Google Scholar]

2019

- Dongxu Zhang, **Subhabrata Mukherjee**, Colin Lockard, Xin Luna Dong, Andrew McCallum, OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference,

NAACL 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics (Acceptance rate: 31%)

- **Subhabrata Mukherjee** and Stephan Günnemann, GhostLink: Mining Latent Influence Networks for Influence-aware Item Recommendation, **WWW 2019**: ACM International Conference on World Wide Web (Acceptance rate: 18%)

2018

- Kashyap Popat, **Subhabrata Mukherjee**, Andrew Yates, Gerhard Weikum, DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning, **EMNLP 2018**: Conference on Empirical Methods in Natural Language Processing (Acceptance Rate: 26%)
- Guineng Zheng, **Subhabrata Mukherjee**, Luna Dong, FeiFei Li, OpenTag: Open Attribute Extraction from Product Profiles, **KDD 2018**: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Acceptance rate: 8%)
- Kashyap Popat, **Subhabrata Mukherjee**, Jannik Stroetgen and Gerhard Weikum, CredEye: A Credibility Lens for Analyzing and Explaining Misinformation, **WWW 2018**: ACM International Conference on World Wide Web
- **Subhabrata Mukherjee**, Emanuele Coviello, Xin Luna Dong and Fabian Moerchen, MusicVersionTagger: Music Version Extraction of Tracks from Catalog Meta-data, **AMLC 2018**: Amazon Machine Learning Conference

2017

- **Subhabrata Mukherjee**, Probabilistic Graphical Models for Credibility Analysis in Evolving Online Communities, **PhD Dissertation**, Saarland University, 2017
- **Subhabrata Mukherjee**, Kashyap Popat and Gerhard Weikum, Exploring Latent Semantic Factors to Find Useful Product Reviews, **SDM 2017**: SIAM Conference on Data Mining (Acceptance rate: 26%)
- Kashyap Popat, **Subhabrata Mukherjee**, Jannik Stroetgen and Gerhard Weikum, Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media, **WWW 2017**: ACM International Conference on World Wide Web

2016

- **Subhabrata Mukherjee**, Stephan Günnemann and Gerhard Weikum, Continuous Experience-aware Language Model, **KDD 2016**: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Acceptance rate: 8.9%)
- **Subhabrata Mukherjee**, Sourav Dutta and Gerhard Weikum, Credible Review Detection with Limited Information using Consistency Features, **ECML-PKDD 2016**: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Acceptance rate: 28%)
- Kashyap Popat, **Subhabrata Mukherjee**, Jannik Stroetgen and Gerhard Weikum, Credibility Assessment of Textual Claims with Web Evidence, CIKM 2016: ACM International Conference on Information and Knowledge Management

2015

- **Subhabrata Mukherjee**, Hemank Lamba and Gerhard Weikum, Experience-aware Item Recommendation in Evolving Review Communities, **ICDM 2015**: IEEE International Conference On Data Mining (Acceptance rate: 18.1%)
- **Subhabrata Mukherjee** and Gerhard Weikum, Leveraging Joint Interactions for Credibility Analysis in News Communities, **CIKM 2015**: ACM International Conference on Information and Knowledge Management (Acceptance rate: 17.9%)

2014

- **Subhabrata Mukherjee**, Gerhard Weikum and Cristian Danescu-Niculescu-Mizil, People on Drugs: Credibility of User Statements in Health Communities, **KDD 2014**: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Acceptance rate: 14.6%)
- **Subhabrata Mukherjee**, Jitendra Ajmera and Sachindra Joshi, Domain Cartridge: Unsupervised Framework for Shallow Domain Ontology Construction from Corpus, **CIKM 2014**: ACM International Conference on Information and Knowledge Management (Acceptance rate: 20.8%)

- **Subhabrata Mukherjee**, Gaurab Basu and Sachindra Joshi, Joint Author Sentiment Topic Model, **SDM 2014**: SIAM Conference on Data Mining (Acceptance rate: 15.4%)
- **Subhabrata Mukherjee** and Sachindra Joshi, Author-Specific Hierarchical Sentiment Aggregation for Rating Prediction of Reviews, LREC 2014: Language Resources and Evaluation Conference
- **Subhabrata Mukherjee** and Sachindra Joshi, Help Yourself: A Virtual Self-Assist Agent, WWW 2014: ACM International Conference on World Wide Web (Demo)
- **Subhabrata Mukherjee**, Jitendra Ajmera and Sachindra Joshi, Unsupervised Approach for Shallow Domain Ontology Construction from Corpus, WWW 2014: ACM International Conference on World Wide Web (Poster)

2013

- **Subhabrata Mukherjee** and Sachindra Joshi, Sentiment Aggregation using ConceptNet Ontology, **IJCNLP 2013**: International Joint Conference on Natural Language Processing (Acceptance rate: 23.4%)
- **Subhabrata Mukherjee**, Ashish Verma and Kenneth W. Church, Intent Classification of Voice Queries on Mobile Devices, WWW 2013: ACM Conference on World Wide Web (Poster)
- **Subhabrata Mukherjee**, Gaurab Basu and Sachindra Joshi, Incorporating Author Preference in Sentiment Rating Prediction of Reviews, WWW 2013: ACM International Conference on World Wide Web (Poster)

2012

- **Subhabrata Mukherjee** and Pushpak Bhattacharyya, Sentiment Analysis in Twitter with Lightweight Discourse Analysis, **COLING 2012**: International Conference on Computational Linguistics (Acceptance rate: 16%)
- **Subhabrata Mukherjee** and Pushpak Bhattacharyya, YouCat : Weakly Supervised Youtube Video Categorization System from Meta Data & User Comments using WordNet & Wikipedia, **COLING 2012**: International Conference on Computational Linguistics (Acceptance: 16%)
- **Subhabrata Mukherjee** and Pushpak Bhattacharyya, Feature Specific Sentiment Analysis for Product Reviews, **CICLING 2012**: International Conference on Intelligent Text Processing and Computational Linguistics (Acceptance rate: 28.6%)
- **Subhabrata Mukherjee** and Pushpak Bhattacharyya, WikiSent : Weakly Supervised Sentiment Analysis Through Extractive Summarization With Wikipedia, **ECML-PKDD 2012**: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Acceptance rate: 23.7%)
- Balamurali A.R., **Subhabrata Mukherjee**, Akshat Malu and Pushpak Bhattacharyya, Leveraging Sentiment to Compute Word Similarity, GWC 2012: Global WordNet Conference
- **Subhabrata Mukherjee**, Akshat Malu, Balamurali A.R. and Pushpak Bhattacharyya, TwiSent: A Multi-Stage System for Analyzing Sentiment in Twitter, CIKM 2012: ACM International Conference on Information and Knowledge Management (Poster)

Book Chapters

Sentiment Analysis of Reviews, In Encyclopedia of Social Network Analysis and Mining (ESNAM) (Second Edition), Springer New York, 2017.

Achievements

KDD Ph.D. Dissertation Runner-up Award (one of the top-3 best doctoral dissertations world-wide in data mining, SIGKDD 2018)

Graduated *summa cum laude* from Max Planck Institute for Informatics, Germany, 2017

Invited for Microsoft Research PhD Summer School, Cambridge, UK, 2015

Student Travel Award for SIAM Data Mining (SDM), 2014

IMPRS (International Max Planck Research School) Scholarship for PhD, 2013

Member of the group IBM Watson Solutions that has been recognized with the 2013 North America

Frost & Sullivan Award for New Product Innovation, 2013

10 pointer in Semester 3 & 4 at IIT Bombay, and overall rank within Top 5 in the department, 2012

Honors (undergraduate) in Computer Science, Jadavpur University, 2010

99.90 percentile in the all-India GATE Entrance Exam (2010), and 99.81 percentile in the all-India WBJEE Entrance Exam (2006)

Received the Meritorious Children Award from Indian Oil Corporation, Kolkata, 2004

Runners-up at the Inter-School Chess competition held at Don Bosco (2003), and the annual Chess competition held at South Point High School for 2 yrs

References

Available upon request.