**Mitul Tiwari**

Email: mitultiwari@gmail.com

## Areas of Interest:

I have 10 years of experience in development of technologies and algorithms for internet applications. Over the years, I have worked on large-scale social recommendation systems, link prediction in social networks, search technologies, categorization of documents/query/tweets, information retrieval, search relevance, information extraction, query expansion, search spam detection, web mining, machine learning algorithms, data clustering, and classification algorithms. I have also worked on building high performance distributed systems for web crawling and caching. I have also developed and published articles on novel algorithms for distributed caching, network monitoring, multi-processor scheduling, web mining, etc.

## Education

**Ph.D. in Computer Science**. University of Texas at Austin, 2007
**M.S. in Computer Science**. University of Texas at Austin, 2003
**B.Tech. in Computer Science**. Indian Institute of Technology, Bombay, 2001

## Work Experience

**Engineering and Research Manager, LinkedIn, Mountain View, CA**, 02/2011 – present

Engineering and Research Manager 01/2014 – present
Tech Lead 03/2012 – present
Staff Research Engineer 07/2012 – 12/2013
Senior Research Engineer 02/2011 – 06/2012

I am a part of Search, Networks, and Analytics (SNA) group at LinkedIn, and work on data driven products. Leading a very talented bunch of data scientists and engineers for technological advancements in "People You May Know" (PYMK), releationship and growth relevance. PYMK is a large-scale recommender system that analyzes billions of edges to predict social connections. Extended PYMK's ability to bring new members on LinkedIn. Also, worked on "Related Searches", a search query recommender system and "Feed Relevance". Using Hadoop, PIG, Machine Learning algorithms, to analyze a ton of data to build these systems. Some of the contributions:

- More than 8 times increase in the number of connections from PYMK with new features and model improvement
- Built a new growth data product feature from scratch that is leading to significant membership growth
- Significant increase in the number of clicks for related searches through algorithm improvement
- Participated in eight hackday projects, winner of two hackday projects, and one hackday project on Linkedin labs so far. Three of the hackday projects incorporated in products
- Co-authored 8 papers related to recommender systems, link prediction, social network analysis, etc.
- Mentored three interns and co-mentored six other interns
- Organized reading group for about a year and hosted multiple talks at Linkedin
- Outreach through external talks and papers in: KDD, WWW, RecSys, CIKM, SIGIR, QCon, Web and Analytics Summit, Big data meetup, UT Austin, and USF
- A blog post on Linkedin's main blog and three blog posts on LinkedIn's Engineering blog

**Lead Member of Technical Staff, Kosmix, Mountain View, CA**, 10/2007 – 01/2011
I worked on building the next generation information retrieval platform. I work on query, tweets and document classification/categorization, text analysis, query parsing, query expansion, information extraction, and search relevance. I also work on building high performance, scalable systems for data fetching and caching.
**Tweet Classification**: In the last project at Kosmix, developed a tweets classification system based on vector space model. This classification system was developed for Tweetbeat.com to classify incoming tweets from Twitter fire-hose to trending events. Built using Java, Perl, Python, and Tokyo Cabinet.
**Document Categorization**: Worked on improving concept extraction from a given document (or tweet) to categorize the document among 10 million categories in the taxonomy. Also developed a framework to categorize hashtags present in tweets. Built using C++, Perl, Ruby, Cassandra, and MapUpdate.
**Related Topics**: In another project, developed a technique to analyze a given web page and figure out a set of related topics to explore. This technique was used by Kosmix's publisher solution product to provide links on web pages to explore related topics. Built using C++ and a taxonomy of 10 million nodes.
**Topic Page**: Developed back-end for Kosmix Topic pages. Built the infrastructure to mashup results from various APIs and results from search index. Developed techniques to analyze and score (1) search results from various sources using techniques such as Okapi scoring, (2) the most relevant type of information for the given query, (3) the importance of a web page. Also, developed query parsing framework to find out location in the query, important concepts in the query, and to expand query to get more results related to the given query. Filed a patent for the novel techniques developed for relevance of topic pages. Built using C++ and Perl.
**Concept Extraction**: Extracted 10 million music concepts (album, songs, artist) from MusicBrainz to add to our taxonomy of millions of nodes. Also created a module to show the music concepts on relevant topic pages. Built using C++, Python, Berkeley DB XML, and Mysql.
**Back-end Infrastructure**: Developed a high performance, scalable HTTP thread pool using Curl library to asynchronously fetch web pages at run-time. This infrastructure was used by Kosmix topic pages to fetch more than 15 million web pages everyday. Also, developed a scalable caching framework to cache the whole topic page except advertisements and parts of the topic pages. Built using C++, Curl, and Ehcache.

**Research Assistant, University of Texas at Austin**, 1/2004 – 8/2007
I worked on algorithms for emerging network applications. In particular, my research focussed on designing self-tuning algorithms for distributed resource allocation, especially for distributed caching, distributed monitoring, and multi-core scheduling problems.

**Summer Intern, Google Incorporation, Mountain View, CA**, 5/2004 – 8/2004
I was a part of the Search Quality group and worked on Automated Search Spam Filtering. I developed algorithms and heuristics for detecting spam in search results. I implemented my algorithms and performed experiments with a large amount ($\sim$ 13 TB) of crawled data, query logs, and toolbar data. I used more than 1000 machines to perform my experiments. Built using C++, MapReduce, Sawzall, and GFS.

**Summer Intern, Microsoft Corporation, Redmond, WA**, 6/2003 – 8/2003
I was a part of the MSN Search group. I designed and developed a WML proxy server to empower mobile device users to search and surf the web through MSN Search. I performed query analysis experiments with a large amount ($\sim$ 200 GB) of query and crawled data to investigate machine allocations in a data center. Built using C#, Openwave, XSLT, and WML.

**Teaching Assistant, University of Texas at Austin**, 8/2002 – 12/2003
Served as a TA for Advanced Operating Systems, a graduate level course, and for Software Engineering, a senior level undergraduate course.

## Relevant Skills
Java, C, C++, Perl, Python, R, Shell scripting, NodeJS/Express, Clojure, SQL. Familiar with Ruby.
Hadoop MapReduce, PIG, Hive, HDFS, Voldemort, Kafka, Cassandra, Mysql, Berkeley DB, Tokyo Cabinet.

## Awards and Honors

- Received Kosmonauts and Kosmix Kreed awards for innovative and team work towards building next generation information retrieval platform.
- Received MCD fellowship, University of Texas at Austin, 2001–2005.
- Placed in the top 0.05% of all examinees of IIT–JEE 1997, the Joint Entrance Examination for admission to IITs. I was ranked 49th among more than 100,000 students.
- Placed in the top 25 students selected for the gold medal of NSEP 1997, an examination conducted by the Indian Association of Physics Teachers (IAPT) in Physics.

## Publications

1. Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily, with Ashton Anderson, Jure Leskovec, Jon Kleinberg, and Daniel Huttenlocker. To appear in *the Proceedings of the 24th International World Wide Web Conference (WWW)*, May 2015.
2. The Browsemaps: Collaborative Filtering at LinkedIn, Lili Wu, Sam Shah, Sean Choi, Mitul Tiwari, Christian Posse. In *Proceedings of the 6th ACM RecSys Workshop on Recommender Systems and the Social Web*, October 2014.
3. Modeling Impression Discounting in Large-scale Recommender Systems , Pei Li, Laks V.S. Lakshmanan, Mitul Tiwari, Sam Shah. In *Proceedings of the 20th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, August 2014.
4. Structural Diversity in Social Recommender Systems, Xinyi Huang, Mitul Tiwari and Sam Shah. In *Proceedings of the 5th ACM RecSys Workshop on Recommender Systems and the Social Web*, October 2013.
5. Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach. Abhishek Gattani, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, AnHai Doan. In *Proceedings of the 39th International Conference on Very Large Data Bases (VLDB)*, August 2013.
6. Social Media Analytics: The Kosmix Story. With Venky Harinarayan, Anand Rajaraman, AnHai Doan, and others. In *IEEE Data Engingeering Bullettin*, 36 (3), 4-12.
7. Organizational Overlap on Social Networks and its Applications, Cho-Jui Hsieh, Mitul Tiwari, Deepak Agarwal, Xinyi (Lisa) Huang, and Sam Shah. In *Proceedings of the 22nd International World Wide Web Conference (WWW)*, May 2013.
8. Metaphor: a system for related search recommendations, Azarias Reda, Yubin Park, Mitul Tiwari, Christian Posse, and Sam Shah. In *Proceedings of the 21st International Conference on Information and Knowledge Management*, October 2012.
9. Social Networking in Developing Regions. Azarias Reda, Sam Shah, Mitul Tiwari, Anita Lillie, and Brian Noble. In *Proceedings of the International conference on Information and Communication Technologies and Development*, March 2012.
10. Online compression caching. C. Greg Plaxton, Yu Sun, Mitul Tiwari, and Harrick Vin. *Lecture Notes in Computer Science*, 5124 Springer, 2008, ISBN 978-3-540-69900-2.
11. Online aggregation over trees. C. Greg Plaxton, Mitul Tiwari, and Praveen Yalagandula. In *Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium*, March 2007.
12. Reconfigurable resource scheduling with variable delay bounds. C. Greg Plaxton, Yu Sun, Mitul Tiwari, and Harrick Vin. In *Proceedings of the 21st IEEE International Parallel and Distributed Processing Symposium*, March 2007.
13. Online hierarchical cooperative caching. Xiaozhou Li, C. Greg Plaxton, Mitul Tiwari, and Arun Venkataramani. *Theory of Computing Systems*, 39:851-874, 2006.
14. Reconfigurable resource scheduling. C. Greg Plaxton, Yu Sun, Mitul Tiwari, and Harrick Vin. In *Proceedings of the 18th Annual ACM Symposium on Parallelism in Algorithms and Architecture*, July 2006.
15. Online hierarchical cooperative caching. Xiaozhou Li, C. Greg Plaxton, Mitul Tiwari, and Arun Venkataramani. In *Proceedings of the 16th Annual ACM Symposium on Parallelism in Algorithms and Architectures*, June 2004.

16. Memex: A browsing assistant for collaborative archiving and mining of surf trails. Soumen Chakrabarti, Sandeep Srivastava, Mallela Subramanyam, and Mitul Tiwari. In *Proceedings of the 26th International Conference on Very Large Data Bases*, September 2000.
17. Using Memex to archive and mine community Web browsing experience. Soumen Chakrabarti, Sandeep Srivastava, Mallela Subramanyam, Mitul Tiwari. In *Proceedings of the 9th International World Wide Web Conference*, May 2000.

**Work Eligibility**: US Permanent Resident