## Ding Zhou

Research Scientist
Facebook Inc.
156 University Avenue
Palo Alto, California, USA

Email: dr.dingzhou@gmail.com
Homepage: http://www.cse.psu.edu/∼dzhou/

### Objective & Background Summary

Dr. Ding Zhou is a researcher and software engineer interested in application of machine learning and social network analysis to real-world business and Internet services. He authored 20+ research papers on influential international venues and 2 U.S. patents covering various machine learning topics on large-scale data. He has taken engineer and scientist positions at leading industrial companies including Facebook, Google, Yahoo!, NEC Labs. He is also a co-developer of the large open source search engine: CiteSeer$^x$ (http://citeseer.ist.psu.edu).

### Education

Ph.D. Computer Science and Engineering, Pennsylvania State University, USA, 2007.

B.S. Computer Science, Fudan University, China, 2004. *Summa Cum Laude*

### Industry Experience

*Research Scientist* — Jan 2008 - present
Facebook Inc. — Palo Alto, CA, USA

Design and develop advanced machine learning models for Internet advertising:

- *experiment with and develop keyword extraction tools to improve ads targeting coverage;*
- *research and develop social graph-based features in ads CTR estimation;*

Design and develop large-scale machine learning products on Hadoop/Hive:

- *develop an a/b test framework using Hive handling user log management (500GB/day);*
- *develop and improve the NewsFeed story ranking system (3B impressions/day);*
- *design and develop a Hadoop/Hive-based Web page de-duplication system;*
- *design and develop a recommendation system for social applications (60M users/day);*
- *design and develop a reputation system for social applications;*

Support data-driven corporate decision making.

*Research Intern* — Summer 2007
NEC Research America — Cupertino, CA, USA

Proposed a new method and developed a prototype system for document recommendation based on merging multiple graphs. The documents are represented in a low-dimensional feature space, combining techniques including directed graph Laplacian, dimension reduction and matrix factorization. Experiments showed 35% recommendation accuracy improvement on a real-world data set of $400K$ documents from CiteSeer.

*Engineer Intern* — Summer 2006
Google Labs — New York, NY, USA

Developed a production-ready automatic Optical Character Recognition (OCR) correction system for Google Books. The new system included functions of text alignment, language modeling, model training, and character-level auto correction. The system was implemented in C++.

## Ding Zhou

*Research Intern* Summer 2005
Yahoo! Search Sunnyvale, CA, USA
Experimented with and developed a prototype of automatic Web search result clustering system. Revised the prefix tree data structure and proposed a new way of search results real-time hierarchical clustering. The demo system was implemented in Python+Java.

*Software Engineer* Feb 2004 - Aug 2004
Porsche Asia-Pacific Shanghai, China
Led a small team and developed a Client/Server based enterprise resource management (ERP) system to optimize the company's business flow and manage the inventory of Porsche China sales office. The system was deployed in production and was implemented in Java + MySql.

### U.S. Patents

Determining User Affinity Towards Applications on a Social Networking Website
U.S. Patent pending, 2008.

Resource Management of Social Network Applications
U.S. Patent pending, 2008.

### Research Publications

Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement, Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein and Hongyuan Zha, In proceedings of *WWW'09: The 18th International World Wide Web Conference*, Madrid, Spain, 2009 (WWW 2009).

Learning Multiple Graphs for Document Recommendations, Ding Zhou, Shenghuo Zhu, Kai Yu, Xiaodan Song, Belle Tseng, Hongyuan Zha, C. Lee Giles In proceedings of *WWW '08: The 17th international conference on World Wide Web*, Beijing, China, 2008 (WWW 2008).

Exploring Social Annotations for Information Retrieval, Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, C. Lee Giles, In proceedings of *WWW '08: The 17th international conference on World Wide Web*, Beijing, China, 2008 (WWW 2008).

Discovering Temporal Communities from Social Network Documents, Ding Zhou, Isaac Councill, Hongyuan Zha, C. Lee Giles, In proceedings of *the 7th IEEE International Conference on Data Mining*, Omaha, Nebraska, USA, 2007, (ICDM 2007).

Co-Ranking Authors and Documents in a Heterogeneous Network, Ding Zhou, Sergey Orshanskiy, Hongyuan Zha, C. Lee Giles, In proceedings of *the 7th IEEE International Conference on Data Mining*, Omaha, Nebraska, USA, 2007, (ICDM 2007).

Extracting Author Meta-data from Web, Shuyi Zheng, Ding Zhou, Jia Li, C. Lee Giles, In proceedings of *the 7th IEEE International Conference on Data Mining*, Workshop, Omaha, Nebraska, USA, 2007, (ICDM 2007).

IKNN: Informative K-Nearest Neighbor Pattern Classification, Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, C. Lee Giles, In proceedings of *the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Warsaw, Poland, 2007, (PKDD 2007).

Predicting User Clicks in Web Search, Ding Zhou, Levent Bolelli, Jia Li, C. Lee Giles, Hongyuan Zha, In proceedings of *the International Joint Conference on Artificial Intelligence*, Hyderabad,

India, 2007, (IJCAI 2007).

A Clustering Method for Web Data with Multi-Type Interrelated Components, Levent Bolelli, Seyda Ertekin, Ding Zhou and C. Lee Giles, In proceedings of *International World Wide Web Conference* Poster, Banff, Canada, 2007, (WWW 2007), Poster.

K-SVMeans:A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets, Levent Bolelli, Seyda Ertekin, Ding Zhou, C. Lee Giles, In proceedings of *ACM/IEEE International Conference on Web Intelligence*, San Jose, USA, 2007, (WI 2007).

Boosting the Feature Space: Text Classification for Unstructured Web Data, Yang Song, Ding Zhou, Jian Huang, Isaac Councill, Hongyuan Zha, C. Lee Giles, In proceedings of *the 2006 IEEE International Conference on Data Mining*, HK, China, 2006, (ICDM 2006).

Probabilistic Models for Discovering E-Communities, Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, Hongyuan Zha, In proceedings of *the 15th ACM International World Wide Web Conference*, Edinburgh, Scotland, 2006, (WWW 2006).

Rotational-Invariant L1-norm PCA for Robust Learning, Chris Ding, Ding Zhou, Xiaofeng He, Hongyuan Zha, In proceedings of the 23rd *International Conference on Machine Learning*, Pittsburgh, PA, 2006, (ICML 2006).

Topic Evolution and Social Interactions, How Authors Effect Research, Ding Zhou, Xiang Ji, Hongyuan Zha, Lee Giles, In proceedings of *the 15th ACM Conference on Information and Knowledge Management*, Arlington, VA, 2006 (CIKM 2006).

Multi-task Text Segmentation and Alignment by Weighted Mutual Information, Bingjun Sun, Ding Zhou, Hongyuan Zha, John Yen, In proceedings of *the 15th ACM Conference on Information and Knowledge Management*, Arlington, VA, 2006 (CIKM 2006), Poster.

A New Mallows distance based Metric for Comparing Clusterings, Ding Zhou, Jia Li, Hongyuan Zha, In proceedings of the 22rd *International Conference on Machine Learning*, Bonn, Germany, 2005, (ICML 2005).

Probabilistic models for discovering temporal semantic social networks, Lee Giles, Ding Zhou, In proceedings of the invited talk of the Temporal Network Workshop of *the 4th SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006 (KDD Workshop 2006).

Towards Discovering Organizational Structure from Email Corpus, Ding Zhou, Yang Song, Ya Zhang, Hongyuan Zha, In proceedings of the 4th *International Conference on Machine Learning and Applications, Los Angeles*, CA, U.S.A. 2005, (ICMLA 2005), published by IEEE.

CLINCH: Clustering Incomplete High-Dimensional Data for Data Mining, Zunping Cheng, Ding Zhou, Chen Wang, Wei Wang, Baile Shi, In proceedings of *the Seventh Asia Pacific Web Conference*, China, 2005, (APWeb 2005), published by Springer-Verlag publisher.

SUDEPHIC: Self-tuning Density-based Partition and Hierarchical Clustering, Ding Zhou, Zunping Cheng, Chen Wang, Haofeng Zhou, Wei Wang, Baile Shi, In proceedings of *the 9th International Conference on Database Systems for Advanced Applications*, Jeju Island, Korea, (DASFAA 2004), published by Springer-Verlag.

Tuning Parameters For Density-based Partition and Hierarchical Clustering, Zunping Cheng, Ding Zhou, Chen Wang, Haofeng Zhou, Wei Wang, Baile Shi, Journal of Research and Development of Computer Science (Chinese), National Core Journal of China.

CiteSeer$^X$: A Scalable Autonomous Scientific Digital Library, Huajing Li, Isaac Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, C. Lee Giles, In proceedings of *the 1st International Conference on Scalable Information Systems*, Hong Kong, China, 2006, (InfoScale 2006), published by IEEE.

CiteSeerX: Next-Gen CiteSeer. Isaac Councill, Huajing Li, Levent Bolelli, Yang Song, Ziming Zhuang, Jian Huang, Yang Sun, Ding Zhou, Wang-Chien Lee, Anand Sivasubramaniam, and C. Lee Giles. The 2nd International Conference on Open Repositories. Poster. San Antonio, TX, USA. January 2007.

### Academic Research Experience

**Research Assistant, Georgia Institute of Technology, Sep 2006 – Dec 2007.**

• *Clustering Temporal Social Documents*: In this project, temporal social networks are observed where community structures are discovered. The social networks include heterogeneous relationship among users, documents, and social events, observed over time. We propose a new constrained spectral graph partitioning to combine the prior knowledge about a graph into partitioning. The temporal communities are discovered via threading the discovery of static communities at each time period. The problem is formulated as a Quadratically Constrained Quadratic Programming (QCQP) problem which we then convert into an eigenvector problem. We compare the proposed approach with an extension of *normalized cut* and an approach based on semi-definite programming, obtaining significant accuracy improvements.

• *Learning Social Network Flows to Rank*: The goal of this project is to rank social network actors by learning the information flow. We pursue the goal via modeling the network flow in social networks using the documents the actors share and act on. The network flow is computed by solving a quadratic programming (QP) problem with both *edge-wise* and *aggregate* constraints corresponding to various types of implicit preferences. Experiments are carried out on a real-world dataset extracted from CiteSeer site, attaining significant improvements compared with , citation count, collaboration PageRank, citation PageRank.

• *Improving Information Retrieval by Tags*: The goal of the project is to leverage the manual efforts that users have made on documents to improve information retrieval. We propose to combine the modeling of social annotations with the language modeling-based methods. We propose a unified framework to combine the modeling of social annotations with the language modeling-based methods for information retrieval. The proposed approach consists of two steps: (1) we seek to discover topics in the contents and annotations of documents while categorizing the users by domains; and (2) we enhance document and query language models by incorporating user domain interests as well as topical background models. Differences in user domain expertise are also considered when combining the discovered user domain interests. We experiment with a sample collection of del.icio.us data. Our results show 22% improvements in DCG scores over the traditional approaches and 11% over counterpart approaches, compared with language modeling-based IR and an extension of it using Latent Dirichlet Allocation (LDA).

**Research Assistant, Pennsylvania State University, Jan 2005 – Dec 2007.**

• *Co-Ranking on Graph Couples*: This work seeks to rank connected heterogeneous entities by their centrality on graphs. In particular, two heterogenous graphs are connected via a third bipartite graph, one of which is the social network and the other is the document citation network. We propose a new method for co-ranking actors and documents respectively in their social networks and citations networks. The new method seeks to leverage the authorship relationship between actors and documents, by which the rankings of actors and documents mutually depend on each other. A new co-ranking framework is introduced based on coupling random walks in two different graphs. Empirical evaluation made on a CiteSeer dataset demonstrates significant improvements in actor ranking quality compared with other traditional author evaluation metrics, including, number of publications, number of citations, and PageRank.

• *Topic Evolution and Social Interactions*: In this collaboration with NEC Labs, we propose a method for discovering the dependency relationships between the topics of documents shared in social networks using the latent social interactions. We seek to measure the dependency among topics in documents. By viewing the evolution of topics as a Markov chain, we estimate a Markov transition matrix of topics by leveraging social interactions and topic semantics. The discovered topic relationship is used for clustering topics into meta topics.

• *Mining CiteSeer Click-through*: In this project, the meta-log data of CiteSeer is analyzed, seeking to improve the ranking quality for queries of typical kinds. The project consists of two main parts: (1) the engineering part of this project is a socket-level logging framework with backend database support. The logger/logging clients are implemented in Java; (2) machine learning on CiteSeer logs for predicting Web search clicks. A new conditional probability hierarchy is learned from the co-occurrences of terms and documents. The semantic structures are discovered from queries as hierarchies. The posterior probability of a document given the query is then estimated along this hierarchy.

• *Social Network Analysis in Emails*: We analyze the community structures in the Enron email dataset in this project. We look into the learning of user communication patterns in terms of text. We discover the formation of the information carriers, email texts in this case, and the inferences that could be made about the hidden *community structure*. We propose two generative Bayesian graphical models social network community detection. A revision of Gibbs-sampling is devised to train the models. We compare this work with various text mining work based on generative modeling, including Latent Dirichlet Allocation (LDA), author-topic model, unigram mixture model, etc.

**Research Services**

Program Committee, ACM International Conference on Data Engineering, 2010;

Reviewer, ACM SIGIR International Conference on Information Retrieval, 2009;

Program Committee, ACM International Conference on World Wide Web, 2009;

Program Committee, Second International Workshop on Social Computing, Behavior Modeling, and Prediction, 2009;

Program Committee, AAAI International Conference on Weblogs and Social Media, 2009;

Program Committee, ACM International Conference on Information and Knowledge Management, 2008;

Program Committee, ACM SIGKDD International Workshop on Social Network Mining and Analysis, 2008;

Program Committee, ACM International Workshop on Search in Social Media, 2008;

Reviewer, International Journal of Computational Statistics and Data Analysis, 2006;

Reviewer, International Journal of Data Mining and Knowledge Discovery, 2006;

External Reviewer, Nature, 2006;

External Reviewer, ACM Transactions on Information Systems, 2005;

External Reviewer, ACM International Conference on Research and Development in Information Retrieval, 2005, 2006, 2007, 2008;

External Reviewer, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, 2007;

External Reviewer, ACM International World Wide Web Conference, 2006, 2007, 2008;

External Reviewer, ACM/IEEE Joint Conference on Digital Libraries, 2005, 2006, 2007;

External Reviewer, International Joint Conferences on Artificial Intelligence, 2005, 2006;

External Reviewer, ACM International Conference on Information and Knowledge Management, 2005, 2006, 2007, 2008;

External Reviewer, AAAI International Conference on Artificial Intelligence, 2006, 2007;

External Reviewer, SIAM International Conference on Data Mining, 2006, 2007;

**Invited Talks**

Data Analysis at Facebook, Joint Statistical Meetings, Denver, Colorado, Aug 2008;

Learning Graphs for Recommendation, hosted by University of California, Santa Cruz, May 2008;

Machine Learning Problems at Facebook, hosted by PSU Graduate Symposium, Feb 2008;

Exploring Social Annotations for Information Retrieval, at WWW 2008;

Graph Fusion for Recommendations, at WWW 2008;

Co-Ranking Authors and Documents, at ICDM 2007;

Temporal Community Discovery in Heterogeneous Networks, at ICDM 2007;

Recommendations by Learning from Multiple Graphs, at NEC Research, Sep 2007;

Knowledge discovery in social networks, hosted by Cornell University, Mar 2007;

Learning to predict user clicks, at IJCAI 2007;

Topic Evolution and Social Interactions, at CIKM 2006;

Probabilistic models for community discovery, at WWW 2006;

OCR correction using language modeling, poster, hosted by Google New York, Sep 2006;

A new search result clustering algorithm, hosted by Yahoo! Search, Aug 2005;

Comparing clusterings based on Mallows Distance, at ICML 2005;

Density-based clustering via discriminant weighting, at DASFAA 2004;

**Recent Awards and Grants**

Q2 Facebook Hackathon Winner: most innovative award, May 2008.

Travel Grant Award to the International Conference on Data Mining, Nov 2007.

Travel Grant Award to the Intl. Joint Conference on Artificial Intelligence, Jan 2007.

Nomination to IBM Research Fellowship by Penn State (2/252), 2006.

Nomination to Microsoft Research Fellowship by Penn State (3/252), 2006.

Fred A. & Susan Breidenbach Fellowship, Penn State, 2004-2005.

Travel Grant to the graduate summer school, Institute for Pure & Applied Mathematics, University of California, Los Angeles, 2005.

University Thesis Award, Fudan University, 2004.

Renmin Fellowship, Computer Science, Fudan University, 2001 - 2004.

Promising Student Award, Government of Changsha City, China, 2000.

**Skills**

5 years in Python; 6 years in Java; 7 years in C++;

Experiences in PHP, JSP, HTML/XML, MySQL, Hadoop/Hive, Linux Systems;

Strong learning and communication skills.

**References**

Available upon request.