

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/>

Education

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

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

Objective & Background Summary

Interested in a research and development position that implements machine learning, social network analysis and social text analysis to real business applications.

Researcher on a wide range of machine learning topics including: *link analysis and community discovery*, *social recommendations*, *ranking and co-ranking*, and *user click-through mining*. Engineer or scientist on projects from major industrial companies including Yahoo!, Google, NEC Labs, and Facebook; Co-developer of the large open source project: CiteSeer^x (<http://citeseer.ist.psu.edu>).

Industry Experience

<i>Research Scientist</i> Facebook Inc. Develop and improve the news feed ranking system; Design and develop large scale machine learning products based on Hadoop/Hive; Design and develop the reputation, recommendation, and ranking systems of social applications on Facebook platform; Research machine learning issues in large online social networks; Support data-driven corporate decision making.	Jan 2008 - present Palo Alto, CA, USA
<i>Research Assistant</i> NEC Research America -project: NEC-graph-fusion.	Summer 2007 Cupertino, CA, USA
<i>Engineering Intern</i> Google Labs -project: Google-lm-ocr.	Summer 2006 New York, NY, USA
<i>Research Assistant</i> Yahoo! Search -project: Yahoo-search-cls.	Summer 2005 Sunnyvale, CA, USA
<i>Software Engineer</i> Porsche Asia-Pacific -project: Porsche-erp.	Feb 2004 - Aug 2004 Shanghai, China
<i>Research Assistant</i> Shanghai Database Research Center -project: Fudan-dm.	Feb 2003 - Aug 2004 Shanghai, China

Ding Zhou

Academic Experience

Research Assistant, Professor Hongyuan Zha, Georgia Institute of Technology, Sep 2006 – Dec 2007.
– project names: Gatech-sna-flow, Gatech-ir-tags, Gatech-sna-cls.

Research Assistant, Professor C. Lee Giles, Pennsylvania State University, Jan 2005 – Dec 2007.
– project names: PSU-click, PSU-sna-citeseer, PSU-corank.

Sponsored Participant, Graduate Summer School, University of California, Los Angeles, Jul 2005.
– summer school seminars.

Research Assistant, Professor Hongyuan Zha, Pennsylvania State University, 2004 – 2005.
– project names: PSU-cls-measure, PSU-sna-email.

Research Assistant, Prof. Wei Wang, Prof. Baile Shi, Fudan University, China, 2003 – 2004.
– project names: Fudan-dm.

Teaching Assistant, Scientific Computing, Pennsylvania State University, Spring, 2005.

Teaching Assistant, Algorithm Design and Analysis, Pennsylvania State University, Fall, 2004.

Patents

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

Resource Management of Social Network Applications,
U.S. Patent (pending);

Research Publications

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).

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).

Learning 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.

Ding Zhou

K-SVMMeans: 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).

R1-PCA: Rotational Invariant L1-norm Principal Component Analysis for Robust Subspace Factorization, 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 22nd *International Conference on Machine Learning*, Bonn, Germany, 2005, (ICML 2005).

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.

Ding Zhou

CiteSeer^X: A Scalable Autonomous Scientific Digital Library, Huajing Li, Isaac Council, 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 Council, 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.

Research Services

Program Committee, ACM International Conference on World Wide Web, 2009;
Program Committee, AAAI International Conference on Weblogs and Social Media, 2009;
Program Committee, ACM Intl Conference on Information and Knowledge Management, 2008;
Program Committee, ACM SIGKDD Int'l Workshop on Social Network Mining and Analysis, 2008;
Program Committee, ACM International Workshop on Search in Social Media, 2008;
Reviewer, International Journal of Computational Intelligence Research, 2008;
Reviewer, IEEE Transactions on Multimedia, 2008;
Reviewer, IEEE Transactions on Systems, Man, and Cybernetics, 2008;
Reviewer, Communication of the ACM, 2007;
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;

Ding Zhou

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

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.
Finalist on IBM Research Fellowship, United States, 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.
Research Assistant Award, Information Sciences and Technology, Penn State, 2005-2006.
Teaching Assistant Award, Dept of Computer Science and Engineering, Penn State, 2004-2005.
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; 6 years in C++;
Experiences in PHP, JSP, ASP, XML, Visual Basic, MySQL, Oracle, Microsoft SQL Server;
Strong learning and communication skills.

Selected Project List:

NEC Project: Learning for Recommendations by Fusing Multiple Graphs (NEC-graph-fusion)

This project addresses the document recommendation problem by learning document similarities from multiple graphs sharing a same subset of nodes. The document similarities are captured by a low-dimensional feature space where the relative positions of documents implicitly determine the similarities between documents. The low-dimensional feature space is learned from a directed graph, a bipartite graph, and multiple-class labels (conceptually another bipartite graph with special properties), which represent the most common graph types in reality. Different objective functions are defined for the three types of graphs, combining techniques including the Laplacian on directed graphs, graph embedding and matrix factorization. Experiments are carried out on the citation, authorship, and venue graphs from CiteSeer. Recommendation quality is measured by the precision in predicting citations given several starting citations.

Clustering Temporal Social Documents (Gatech-sna-cls)

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 eigen-vector problem. We compare the proposed approach with an extension of *normalized cut* and an approach based on semi-definite programming, obtaining significant accuracy improvements.

Co-Ranking Graphs Couples (PSU-corank)

This work seeks to rank connected heterogeneous entities by their centrality on graphs. In particular, two heterogeneous 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.

Learning Social Network Flows to Rank (Gatech-sna-flow)

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. The various constraints employed for the QP problem are derived from the actions of social actors as well as the temporality of them. 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 (Gatech-ir-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).

Google Project: OCR Correction by Machine Learning (Google-lm-ocr)

Optical Character Recognition (OCR) is the process of machine recognition of printed characters, usually involving analysis of image features. Due to the lack of control of the OCR engine, we seek to correct the systematic OCR errors posteriorly using machine learning, primarily based on texts. An OCR correction system is developed, including functions of alignment, language modeling, model training, and correction testing. In the training step, the system aligns paired text documents (ground-truth and OCR-ed text); learns the probabilistic transition probabilities from one text streams to another. In the testing step, the system combines the letter-level and term-level language models to estimate the confidence of word being correct. Corrections are made automatically with high confidence. The project is implemented in C++.

NEC Project: Topic Evolution and Social Interactions (NEC-sna-topic)

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.

Yahoo! Project: search result clustering (Yahoo-search-cls)

The *search result clustering* is a summer project at Yahoo! summer, 2005. An interactive demo system is developed on Yahoo search results. We organize Web search results real-time into a topic hierarchy, where a topic is the labeled with words for easy browsing. We revise a prefix word tree structure with a document list at each node so as to generate the base clusters. Then we greedily pick base clusters hierarchically in a way that maximize the *coverage* while maintaining the *distinctiveness*. Then these base clusters are merged bottom-up with certain trimming on the way. The system is implemented in Python+Java.

Social networks in CiteSeer (PSU-sna-citeseer)

This project explores the co-authorship social networks in CiteSeer and use it to explain the topic dynamics in CiteSeer documents. We correlate the discovered topics in CiteSeer with the latent social networks and seek to rank authors by their impact on such topic evolutions. We measure the properties of topics that tend to grow in different patterns. We perform the clustering of topics using spectral graph partitioning obtaining the topic hierarchies.

Mining CiteSeer Click-through (PSU-click)

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.

Comparing Clustering Results (PSU-clc-measure)

In this project, we study the comparing of clustering results. We propose a measure for comparing clustering results to tackle two issues insufficiently addressed by existing methods: (a) taking into account the distance between cluster representatives when assessing the similarity of clustering results; (b) constructing a unified framework for defining a distance based on either hard or soft clustering and ensuring the triangle inequality under the definition. Our measure is derived from a complete and globally optimal matching between clusters in two clustering results. It is shown that the distance is an instance of the Mallows distance metric between probability distributions in statistics.

Social Network Analysis in Emails (PSU-sna-email)

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 for 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.

Porsche Project: ERP Database System (Porsche-erp)

As an IT instructor at Porsche China Office, I design and lead a team in developing the Client/Server based enterprise resource management system (ERP). We integrate various data formats from distributed dealers over China. The system manages various types of information based on hybrid of DBMS and Intra-net file sharing. The programming language used is Java + MySQL.

Large-scale data mining (Fudan-dm)

The project TETRIS is a national key project funded by NSF. I participated in the project and (1) developed the clustering module in the open platform based on Web Services integration; (2) carried out research in clustering techniques for large-scale and high-dimensional data. My proposal of several new clustering algorithms were accepted in proceedings of two international conferences. The project is implemented in Java.

References

Available upon request.