Curriculum Vitae

陳弘軒 **Hung-Hsuan Chen, Ph.D.**
Researcher, Computational Intelligence Technology Center,
Industrial Technology Research Institute, Hsinchu, Taiwan
工業技術研究院 巨量資料中心 研究員

CONTACT INFORMATION
http://dr-hhchen.appspot.com/
hhchen1105@gmail.com

SUMMARY

I am interested in investigating and applying **scalable** techniques of **big data mining** and **social network** analysis to various domains, such as **cloud services and web applications**, **digital libraries**, and **information retrieval**. I have co-authored **over a dozen refereed research papers** that have received **over a hundred citations**. I consistently collaborate with industrial corporations, such as Alcatel Lucent, Dow Chemical, and Google to bring research results into practice. I have developed several publicly available Internet services. Recent projects include **CSSeer** (http://csseer.ist.psu.edu/), an expert recommender system for computer scientists that mines the CiteSeerX digital library, and **CollabSeer** (http://collabseer.ist.psu.edu/), a collaborator recommender system for computer scientists based on users' research interests and their previous coauthoring behaviors. I have made contributions to Open Source Software Projects, such as NetworkX, SeerSuite (the basis of the CiteSeerX digital library), and JUNG (Java Universal Network/Graph Framework). I am familiar with state-of-the-art **software developing** techniques, such as **unit testing**, **test driven development**, **MVC (model-view-controller) model**, and distributed source code version management tools (e.g., **Git**). I am also highly experienced with **MapReduce** distributed programming model for big data processing.

EDUCATION

Ph.D. Computer Science and Engineering, The Pennsylvania State University, University Park 2008 - 2013
M.S. Computer Science, National Tsing Hua University 2004 - 2006
B.S. Computer Science, National Tsing Hua University 2000 - 2004

RECENT HONORS

Highest F1-score and highest precision, the Competition of Plagiarism Detection (Source Retrieval), the Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2014
Best Paper Award, College of Engineering Research Symposium, The Pennsylvania State University 2013
Highest F1-score, the Competition of Plagiarism Detection (Source Retrieval), the Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) 2013
Invited to Amazon PhD Research Symposium, selected out of over 250 PhD students to present research works at Amazon's headquarters. **The acceptance rate is single digit percentage** 2013
Travel Award, Special Interest Group on Management of Data (SIGMOD) 2013
Travel Award, International Conference on Healthcare Informatics (ICHI) 2013

Travel Award, International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP) 2012

Refereed Publications

The full text of these papers can be downloaded at: http://dr-hhchen.appspot.com/
* In Computer Science, conference papers are typically formal publications[1], and good conferences are usually more competitive than journals[2]. A good rule of thumb is that **the best conferences are sponsored by ACM**[3].

2014

Hung-Hsuan Chen, Madian Khabsa, C. Lee Giles. The Feasibility of Investing of Manual Correction of Metadata for a Large-Scale Digital Library. *International Digital Libraries Conference (DL)*, 2014.

Zhaohui Wu, Jian Wu, Madian Khabsa, Kyle Williams, Hung-Hsuan Chen, Wenyi Huang, Suppawong Tuarob, Sagnik Ray Choudhury, Alexander Ororbia, Prasenjit Mitra, C. Lee Giles. Towards Building a Scholarly Big Data Platform: Challenges, Lessons and Opportunities. *International Digital Libraries Conference (DL)*, 2014.

Kyle Williams, Hung-Hsuan Chen, C. Lee Giles. Classifying and Ranking Search Engine Results as Potential Sources of Plagiarism. *ACM Symposium on Document Engineering (DocEng)*, 2014.

Kyle Williams, Hung-Hsuan Chen, C. Lee Giles. Supervised Ranking for Plagiarism Source Retrieval. *International Conference and Labs of the Evaluation Forum (CLEF)*, 2014. (**Highest F1-score and highest precision in Source Retrieval task of Plagiarism Detection at PAN 2014**)

Jian Wu, Kyle Williams, Hung-Hsuan Chen, Madian Khabsa, Douglas Jordan, C. Lee Giles. CiteSeerX: AI in a Digital Library Search Engine. *Proceedings of the 26th Innovative Applications of Artificial Intelligence Conference (IAAI)*, 2014.

Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernandez-Ramirez, Hung-Hsuan Chen, Zhaohui Wu, C. Lee Giles. CiteSeerX: A Scholarly Big Dataset. *Advances in Information Retrieval - 36th European Conference on IR Research (ECIR)*, 2014.

2013

Hung-Hsuan Chen, C. Lee Giles. ASCOS: an Asymmetric Network Structure COntext Similarity Measure. *ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013.

Hung-Hsuan Chen, David J. Miller, C. Lee Giles. The Predictive Value of Young and Old Links in a Social Network. *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks (DBSocial)*, 2013.

Kyle Williams, Hung-Hsuan Chen, Sagnik Ray Choudhury, C. Lee Giles. Unsupervised Ranking for Plagiarism Source Retrieval. *International Conference and Labs of the Evaluation Forum (CLEF)*, 2013. (**Highest F1-score in Source Retrieval task of Plagiarism Detection at PAN 2013**)

[1] Steve Lawrence. Online or invisible. *Nature* 2001/05

[2] Bertrand Meyer, Christine Choppy, Jørgen Staunstrup, Jan van Leeuwen. Research Evaluation for Computer Science. *Communications of the ACM* 2009/04

[3] Michael Ernst. Choosing a venue: conference or journal? http://homes.cs.washington.edu/~mernst/advice/conferences-vs-journals.html

Hung-Hsuan Chen, Pucktada Treeratpituk, Prasenjit Mitra, C. Lee Giles. CSSeer: an Expert Recommendation System based on CiteSeerX. *Proceedings of the 13th Annual ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2013 (poster).

Hung-Hsuan Chen, Liang Gou, Xiaolong (Luke) Zhang, C. Lee Giles. Towards the Discovery of Diseases Related by Genes Using Vertex Similarity Measures. *International Workshop on Data Mining for Healthcare (DMH)*, 2013.

2012

Hung-Hsuan Chen, Yan-Bin Ciou, Shou-De Lin. Information Propagation Game: a Tool to Acquire Human Playing Data for Multi-Player Influence Maximization on Social Networks. *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012 (demo).

Sumit Bhatia, Cornelia Caragea, Hung-Hsuan Chen, Jian Wu, Pucktada Treeratpituk, Zhaohui Wu, Madian Khabsa, Prasenjit Mitra, C. Lee Giles. Specialized Research Datasets in the CiteSeer^X Digital Library. *D-Lib Magazine*, July/August 2012.

Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, C. Lee Giles. Predicting Recent Links in FOAF Networks. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP)*, 2012.

Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, C. Lee Giles. Discovering Missing Links in Networks Using Vertex Similarity Measures. *Proceedings of the ACM Symposium on Applied Computing (SAC)*, 2012.

2011

Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, C. Lee Giles. CollabSeer: A Search Engine for Collaboration Discovery. *Proceedings of the 11th Annual ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2011.

Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, C. Lee Giles. Capturing Missing Links in Social Networks Using Vertex Similarity. *Proceedings of the 6th ACM International Conference on Knowledge Capture (K-CAP)*, 2011.

∼2010

Liang Gou, Xiaolong Zhang, Hung-Hsuan Chen, Jung Hyun Kim, C. Lee Giles. Social network document ranking. *Proceedings of the 10th Annual ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2010.

Liang Gou, Hung-Hsuan Chen, Jung Hyun Kim, Xiaolong Zhang, C. Lee Giles. SNDocRank: Document Ranking Based on Social Networks. *Proceedings of the 19th ACM International World Wide Web Conference (WWW)*, 2010 (poster).

Liang Gou, Hung-Hsuan Chen, Jung Hyun Kim, Xiaolong Zhang, C. Lee Giles. SNDocRank: a Social Network-Based Video Search Ranking Framework. *ACM International Conference on Multimedia Information Retrieval (MIR)*, 2010.

Liang Gou, Jung Hyun Kim, Hung-Hsuan Chen, Jason Collins, Marc Goodman, Xiaolong Zhang, C. Lee Giles. MobiSNA: a Mobile Video Social Network Application. *ACM Workshop on Data Engineering for Wireless and Mobile Access (MobiDE)*, 2009.

Hung-Hsuan Chen, Kuan-Ta Chen, Cheng-Chun Tu. A User-Centric Framework for Computing Applications' Network Robustness. *ACM Special Interest Group on Data Communications (SIGCOMM)*, 2008 (poster).

Chen-Lung Chan, Shih-Yu Huang, Hung-Hsuan Chen, Wei-Hao Tung, Jia-Shung Wang. An Application-Level Multicast Framework for Large Scale VOD Services. *IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2005.

Research Projects

CSSeer 2012 - 2014

Chief developer of CSSeer, an expert recommender system for computer scientists based on the CiteSeerX digital library.

- CSSeer automatically extracts topic terms from $+1,500,000$ research papers.
- CSSeer recommends experts and compiles related terms (mainly in Computer Science) based on a user submitted query term.
- The framework is shipped to Dow for internal expert discovery.
- URL: `http://csseer.ist.psu.edu/`
- Source: `https://github.com/hhchen1105/expertseer/`

CollabSeer 2009 - 2012

Chief developer of CollabSeer, a potential collaborator recommender system for computer scientists based on the CiteSeerX digital library.

- CollabSeer includes $+1,300,000$ computer science related documents and $+300,000$ unique (disambiguated) authors.
- CollabSeer recommends potential collaborators based on the querist's research interests and previous coauthoring behaviors.
- URL: `http://collabseer.ist.psu.edu/`

MobiSNA 2008 - 2011

Co-developed MobiSNA, a multimedia digital library for mobile phone and portable device users.

- MobiSNA improves search experience by a new document ranking mechanism that integrates textual relevanece, user's interests, and her/his friends' interests.
- URL: `http://mobisna.ist.psu.edu/`

Comic Layout Generator 05/2008 - 07/2008

- Developed the first prototype of the layout generator for a comic generation system, an automatic platform to summarize game players' actions and interactions in a video game.

The Application's Network Robustness Evaluator 11/2007 - 05/2008

- The robustness of network applications was quantified in terms of their ability to handle network errors (e.g., network delay and loss) based on users' departure decisions.

DTV/MHP integrated program, EPG sub-program 2004 - 2006

- Co-developed a personalized TV program recommender system based users' previous watching behaviors and various other features.
- Co-developed a 3-dimensional browsing interface for an electronic program guide, which enables more information to be displayed on a limited TV screen.

Microsoft Windows CE .NET curriculum subject 07/2003 - 12/2003

- Improved the default memory management feature of Windows CE such that the required memory space of each process could be dynamically allocated and the number of simultaneous processes could be larger than the original constraint (32).

HTTP Load Balancer 03/2003 - 05/2003

- Co-developed a load balancer to actively detect or predict loads in each back-end web server instead of passively balancing loads by using Round Robin.
- System throughput increases linearly with the number of back-end web nodes.

Research and Working Experience

Researcher, Computational Intelligence Technology Center, Industrial Technology Research Institute 2014 - present

RA, Information Sciences and Technology, The Penn State University 2008 - 2014

- Developed generic techniques for expert recommendation for scientific documents.
- Developed several algorithms to find the relevance level between different objects.
- Proposed several methodologies to improve the ranking algorithms for search engines.
- Developed the keyphrase extracting component of the open source search engine CiteSeerX (+600K unique users/month, +4M ingested documents in 2014), which automatically crawls, ingests, and indexes scientific documents from the Internet.
- Detected computer-generated fake papers in the CiteSeerX digital library.
- Analyzed user behaviors (e.g., downloading, searching, page transitions) of the CiteSeerX users from logs (+3 billion log entries).

RA, Computer Science and Information Engineering, National Taiwan University 08/2011 - 01/2012

- Investigated the information propagation problem in which multiple parties compete with each other to maximize their influence or minimize the competitors' influence in a social network.

Software Engineer Intern, Google 05/2010 - 08/2010

- Developed a potential customer discovery platform in C++ for Google AdSense based on parametric-based machine learning modeling. The system is on top of the MapReduce framework to handle user clicking logs and user profiles.

RA, Institute of Information Science, Academia Sinica 11/2007 - 07/2008

- Quantified the robustness of network applications in terms of their ability to handle network errors.

RA, Computer Science, National Tsing Hua University 09/2004 - 07/2006

- Proposed a distributed algorithm to efficiently discover frequent items in distributed data streams. This is particularly useful in mining typical patterns for large amounts of continuous data, such as the logs of websites and telecommunication systems.

TA, Operating Systems, National Tsing Hua University 09/2005 - 01/2006

- Designed and graded three projects for 100+ students using the NachOS operating system.

Invited Talks

Talks in addition to those involved in the conference publications above.

Gaining values from big data – using digital libraries and complex networks as examples. The Department of Computer Science, The Rochester Institute of Technology 2014

ExpertSeer: a keyphrase based expert recommender for digital libraries. College of Engineering Research Symposium, The Pennsylvania State University 2013

ASCOS: an asymmetric similarity measure based on network topology. Network Science Seminar, The Pennsylvania State University 2013

Mining experts and each author's expertise from a digital library. Amazon PhD Symposium, Amazon 2013

The challenges of aggregated search: using expert search as an example. SIG Comp Seminar, The Pennsylvania State University 2013

ExpertSeer: a keyphrase based recommendation framework for a digital library expert discovery. Graduate Exhibition, The Pennsylvania State University 2013

Ranking authors in a search engine with and without social network influence. Guest Speaker of IST 441: Information Retrieval and Search Engines, The Pennsylvania State University 2013

CollabSeer: a search engine for collaboration discovery. Graduate Exhibition, The Pennsylvania State University 2012

Integrating social influence to search engines. Guest Speaker of IST 441: Information Retrieval and Search Engines, The Pennsylvania State University 2011

Professional Service

Reviewer, IEEE Transactions on Knowledge and Data Engineering (TKDE) 07/2014
Reviewer, Physica A: Statistical Mechanics and its Applications 05/2014
Reviewer, Journal of Information Science and Engineering (JISE) 01/2014, 07/2013
Reviewer, ACM International Conference on World Wide Web (WWW) 2014
Reviewer, International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP) 2013
Sub-reviewer, ACM Transactions on Information Systems (TOIS) 05/2014
Sub-reviewer, Digital Libraries (DL) 2014
Sub-reviewer, IEEE Intelligent Systems 07/2013
Sub-reviewer, International Conference on Theory and Practice of Digital Libraries (TPDL) 2013
Sub-reviewer, ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2013, 2012, 2011, 2010
Sub-reviewer, ACM International Conference on Research and Development in Information Retrieval (SIGIR) 2014, 2013, 2012, 2011
Sub-reviewer, ACM International Conference on World Wide Web (WWW) 2013, 2012, 2011
Sub-reviewer, ACM International Conference on Information and Knowledge Management (CIKM) 2012

Sub-reviewer, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2012
Sub-reviewer, International Conference on Machine Learning (ICML) 2012
Sub-reviewer, ACM International Conference on Knowledge Discovery and Data Mining (KDD) 2011

Technical Skills

Programming Languages: Python (expert), C/C++ (proficient), Java (proficient), R (proficient), MATLAB (fair), PHP (fair), C# (prior experience)

Tools/Packages: Apache Solr/Lucene, MySQL, Git, Hadoop, LaTeX

Operating Systems: UNIX/Linux, MS-Windows

Misc.

My other professional pages:
Google Scholar http://scholar.google.com/citations?user=T29tmA8AAAAJ
DBLP http://www.informatik.uni-trier.de/~ley/pers/hd/c/Chen:Hung=Hsuan.html
ACM http://dl.acm.org/author_page.cfm?id=81440600313
LinkedIn http://www.linkedin.com/in/hhchen
GitHub https://github.com/hhchen1105/