

Jonathan Siddharth

Email: siddharth.jonathan@cs.stanford.edu

<http://www.infolab.stanford.edu/~jonsid>

Ph: 650 704 2449

Objective:

To build products and services that solve problems relating to Information Retrieval, Web Search and Content Discovery using techniques from Machine Learning, Data Mining, Text Mining and Natural Language Processing.

Education:

Stanford University: 2007

MS in Computer Science with Distinction in Research, Artificial Intelligence Specialization
Best Masters Thesis Award in Computer Science, Stanford University (Christofer Stephenson Memorial Award for Graduate Research)

SVCE, Anna University: 2005

Bachelor of Engineering in Computer Science & Engineering (First Class with Distinction).
1st Rank out of Graduating batch of 133 students in Computer Science

Work & Research Experience:

Founder -Infoaxe Inc.: Web History Search, Social Network & Real-time Search Engine:

Jan-08 to Present

Infoaxe is a web browsing history powered Search and Discovery Engine with over 4M users from over 200 countries. Founded the company in 2008 along with my co-founder Vijay Krishnan. I co-designed and built several key technology components including the Infoaxe Web History Search Engine, Infoaxe Friend Share, Infoaxe Real-time Search Engine and the Infoaxe Recommendation Engine. Raised **\$3.9M** in funding for Infoaxe from Draper Fisher Jurvetson, Labrador Ventures, Band of Angels, Amidzad Partners and Stephen Oskoui(Smile Media).

Infoaxe Web History Search Engine:

Infoaxe's Web History Search Engine automatically saves and indexes every web page a user browses, across all the different computers and browsers she uses so that she can search previously visited web pages from any computer. I co-invented the Real-time Indexing system for Infoaxe's search engine that has been instrumental in scaling the backend to index 20M web pages a day. I also co-developed the Personalized ranking algorithms for the Infoaxe Search Engine that leverage implicit cues for ranking web pages like attention data, frequency and recency of page visits etc.

Jerusalem Post: Desktop: Making History <http://www.jpost.com/Home/Article.aspx?id=123038>

Infoaxe Friend Share:

Built a real-time passive sharing feature that lets users share pages they browse on specific sites with friends via an asymmetric Twitter-like Follower/Following model.

Infoaxe Real-time Search Engine:

Built a Real-time search Engine that leverages the aggregate attention data generated by the Web History Search Engine to build a fresher Web Search Engine that outperforms Google on several query categories like commercial (products, gadgets, deals, coupons etc), entertainment (movies, sitcoms, e-books) and current events. Launch covered by NY Times, CNN, TechCrunch etc. <http://www.nytimes.com/external/venturebeat/2009/11/20/20venturebeat-infoaxe--a-real-time-search-engine-that-does-53994.html>

Infoaxe Recommendation Engine:

Co-invented a recommendation engine for content that analyzes a user's web browsing data, attention data and social graph to recommend interesting and relevant pages to users based on his/her interests.

Infoaxe has been covered by major media outlets and technology blogs worldwide including the NY Times, CNN, BBC, ABC News, The Jerusalem Post, The Hindu, TechCrunch, VentureBeat, GigaOm etc. <http://www.infoaxe.com/press.html>

Scientist (Ranking & Search Relevance) - Powerset Inc. :Natural Language Search

(Acquired by Microsoft for rumored \$100M) Aug 07 to Jan-08

Co-designed the Ranking algorithm for Powerset's Natural Language Search Engine bringing Natural Language Understanding to Web Search. Built ranking features derived from deep Natural Language Processing on text including syntactic parsing (via PARC's Xerox Linguistic Environment (XLE)), semantic analysis, morphological analysis, lexical and ontological resources like WordNet, FreeBase and state of the art information retrieval techniques. Combined ranking features to formulate the ranking function. Led several statistical tests on Mechanical Turk to evaluate search relevance relative to competing search engines and iteratively improved the Ranking algorithm which outperformed Google, Yahoo & Live Web Search Engines on the AOL query set and other query sets for the Wikipedia corpus. Designed a Machine Learned Ranking function using Gradient Boosted Decision Trees. Made significant contributions to Powerset IP which helped the company's eventual acquisition by Microsoft for a rumored \$100M.

Summer Internship: Yahoo! Search Relevance and Machine Learned Ranking Group,

Yahoo! Inc.: *Summer '06*

A. Search Regionalization (Region/Language Mixing in Search Relevance)

Added automatic search regionalization to Yahoo! Search (Inktomi). Encoded a ranking feature that captured whether a user was interested in local content or global content. Performed Query classification on Clickstream data to identify Query classes where search users preferred local content to global content and used that to train a new ranking algorithm. Tests indicated a 3% gain in DCG (Discounted Cumulative Gain). Resulted in huge gains in search relevance in several international markets for Yahoo! Search.

B. Context Driven Ranking:

Improving search relevance by obtaining more 'context' (contextually related words) automatically for the search query, weighting it appropriately and using it to improve search relevance on the Discounted Cumulative Gain metric. Eg. For the search query "photography",

contextually related words would be "pictures", "camera", "film" etc. The presence of these contextually related words in a document is scored positively for relevance to the query.

Research Assistant -Stanford Infolab, Computer Science Department, Stanford University:

Fall 06, Winter 06, Spring 07

(Advisors: Dr.Andreas Paepcke and Prof. Hector Garcia Molina)

a. Language Model Based Ranking for IR:

Worked on a suite of methods to improve recall and precision for Text based Information Retrieval by using a generative model of the Postings list.

b. Web Page Classification:

Worked on Bayesian and SVM based classification of news articles as part of the Web Sociologists Workbench Project at Stanford to help Web Sociologists mine the web for collecting aggregate statistics of online media coverage of special events like the California Special Election, Hurricane Katrina etc.

c. Web Page duplicate detection

Designed an algorithm for near duplicate detection in extremely large Web Page Collections. Improved on existing state of the art methods like shingling and I-Match by more than 24% in combined precision and recall and is faster by a factor of 3 at execution time. Published at SIGIR'08 and is currently an open source Sourceforge project. SpotSigs has been added to the computer science curriculum for a course at Duke University.

Research Assistant -Project Based Learning Lab, Stanford University:

(Advisor: Dr.Renate Fruchter)

Fall 05, Winter 05, Spring 06

a. Hierarchical Enterprise Search:

Built a Hierarchical Text Search Engine for the enterprise to allow searching for results at various granularities and which also leverages the information inherent in the nodes in the hierarchy for ranking of results.

b. Hierarchical Search Results Visualization with a Map Metaphor:

Implemented a TreeMap Visualization of the search results for the above Search Engine to allow visualization of the results at multiple levels of granularity by allowing users to zoom in and zoom out to facilitate easier navigation.

c. Community Image Sharing Engine:

Implemented an Image sharing System that lets users share images on their Desktop with other users currently logged in on different computers.

Other Research Projects at Stanford:

- a. SQUINT: A Support Vector Machine (SVM) based approach to score sections of a web page based on relevance to a query to let users navigate quicker to the most relevant section of a web search result.
- b. Turkalator: A Statistical Machine Translation System for Turkish that accounts for the unique features of an agglutinative language.

Research Assistant -Directed Undergraduate Research, SVCE, Anna University: (Advisor: Prof. Srinivasan) (2002-05)

- a. Autonomous Navigation of Vehicles using Neural Networks
- b. Neural Network based Evidence Combination for Medical Diagnosis
- c. Swarm Intelligence for Task Allocation in Grid Computing
- d. Task Allocation Algorithms for Mesh Computing

Some Relevant Stanford Courses:

- a. Machine Learning (CS229) – Prof. Andrew Ng.
- b. Text Retrieval and Web Search (CS276) – Prof. Chris Manning and Dr. Prabhakar Raghavan.
- c. Natural Language Processing (CS224N) – Prof. Chris Manning.
- d. Probabilistic Models in Artificial Intelligence (CS228) – Prof. Daphne Koller.
- e. Computational Methods for Data mining (CME340) – Prof. Sep Kamvar.
- f. Transaction Processing and Distributed Databases (CS347) – Prof. Hector Garcia-Molina.
- g. Independent Research – Prof. Hector Garcia-Molina and Dr. Andreas Paepcke.

Scholarships & Awards:

- | | |
|----------------|--|
| 2007 | Christofer Stephenson Memorial Award for Graduate Research – Awarded for Best Masters Thesis in Computer Science, Stanford University. |
| 2007 | Masters in Computer Science with Distinction in Research – Departmental Honor in Computer Science, Stanford University. |
| 2006 | Research Assistantship Award by Stanford Infolab, Department of Computer Science at Stanford University – Full tuition waiver and Stipend. |
| 2005 | Research Assistantship Award by Project based learning Lab at Stanford University – Full tuition waiver and Stipend. |
| 2005 | Bharat Petroleum Scholarship - Awarded to 15 students in India for Graduate studies. |
| 2004 | Computer Science Departmental Scholarship - 1st in the CS Department out of 133 students. Tuition Fee Waiver. |
| 2005 | Merit Award- 1st Rank in 8th Semester , SVCE, Anna University. |
| 2005 | Best Paper Award for Senior thesis Research Project - A Comprehensive Architectural Framework for Task Management in Scalable Computational Grids using Swarm Intelligence. (Abacus'05 National level Tech Symposium at Anna University). |
| 2004 | Merit Award- 1st Rank in 6th Semester , SVCE, Anna University. |
| 2002-05 | CAT Prize- 1st Rank in Continuous Assessment tests, Semesters 5,6,7,8 SVCE, Anna University. |
| 2001 | Merit Scholarship – Admission to SVCE, Anna University. |
| 2001 | National Scholarship by State Government of TamilNadu , Std. 12. |
| 1996-01 | Lucas TVS Scholarship - for each of the years from 1996 to 2001 for placing first among Lucas TVS Employees in High School and Undergraduate studies. |
| 1999 | Best Outgoing Student Award by Lions Club , Std. 10. |
-

Other Academic Honors:

2005	1st Rank among graduating batch of 133 , in the CS Department, SVCE, Anna University.
2005	1st Rank, 8th Semester , SVCE, Anna University.
2004	1st Rank, 6th Semester , SVCE, Anna University.
2002-05	1st Rank, Continuous Assessment Tests , Semesters 5, 6, 7,8 SVCE, Anna University.
2001	Top 1% in Higher Secondary Examination , Std. 12.
2000	1st Rank out of 150 students, Common Examination , Std. 11.
1999 - 01	1st Rank, High School - Mathematics, Physics, Chemistry, Biology, Eng.

Publications:

1. SpotSigs: robust and efficient near duplicate detection in large web collections. Theobald, M., Siddharth, J., and Paepcke, A. 2008. In *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Singapore, Singapore, July 20 - 24, 2008). SIGIR '08. ACM, New York, NY, 563-570. DOI= <http://doi.acm.org/10.1145/1390334.1390431>
2. "SpotSigs: Near Duplicate Detection in Web Page Collections". Masters Thesis Report, Department of Computer Science, Stanford University. Won the Christofer Stephenson Award for Best Masters Thesis.
3. "Sentient Autonomous Vehicle using Advanced Neural net Technology", Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems - CIS 2004 organized by IEEE Systems, Man, and Cybernetics Society, ISBN 0-7803-8644-2, Dec 1-3,2004,Singapore.
4. "A Minimal Fragmentation Algorithm for Task Allocation in Mesh-Connected Multicomputers", Proceedings of the IEEE International Conference on Advances in Intelligent Systems - Theory and Applications -AISTA 2004 in cooperation with IEEE Computer Society, IEEE Press, ISBN 2-9599-7768-8,15-18 Nov 2004, Luxembourg, Western Europe.
5. "Knowledge Discovery in Clinical databases with Neural Network Evidence Combination", Proceedings of 2nd IEEE International Conference on Intelligent Sensing and Information Processing - ICISIP 2005 organized by University of Melbourne and co-sponsored by IEEE Engineering in Medicine and Biology Society, IEEE Press, ISBN 0-7803-8840-2, pp.512, Jan 4-7, 2005, Chennai, India.
6. "A System for Power-aware Agent-based Intrusion Detection in Wireless Ad Hoc Networks", Proceedings of Springer Verlag LNCS, 2005 International Conference on Computer Networks and Mobile Computing - ICCNMC 2005, Zhangjiajie, China, Aug 2005.
7. "A Comprehensive Architectural Framework for Task Management in Scalable Computational Grids"- Best Paper Award - Abacus'2005, an Annual National level technical symposium conducted by the Computer Science and Engineering Association of the Department of Computer Science and Engineering (DCSE), College of Engineering, Anna University, March 14-15,2005, Chennai, TN, India.

8. “A Two Tier Neural inter-network based approach to Medical Diagnosis using k-Nearest Neighbor Classification for Diagnosis pruning.” Proceedings of International Conference on Bio-medical Engineering, ICBME’05, Singapore endorsed by IEEE Engineering in Medicine & Biology Society & the IEEE.

Technical Skill Set:

Programming Languages: Java, C++, C.

Scripting Languages: Shell Scripting, JSP, PHP, Javascript, Perl, Ruby.

Databases and related skills: MySQL, Oracle Postgres, JDBC.

Assembly Languages: 8085, MASM (8086), interfacing with input/output devices.

Internet: HTML, XML, XSL, DHTML, CSS.

Other Skills & Software familiar with: XML Beans, C#, Lucene, Nutch, Heritrix, Matlab, R.

Most recently used: Java, JSP, MySQL running on Linux. Work with Apache Web Server and Tomcat as the Servlet container.

Invited Talks:

- Spoke on Panel at Silicon Valley Code Camp 2010 titled “Creating the next Google” (<http://www.siliconvalley-codecamp.com/Sessions.aspx?ForceSortBySessionTime=true&AttendeeId=426>)
- Spoke on Panel on “Entrepreneurship for Engineers” at Stanford University.
- Spoke at Semantic SIG organized by SD Forum presenting Infoaxe.
- Gave a talk on the Architecture of a modern Web Search Engine at Kruzade’09 – National Technical Symposium organized by PSG College of Technology, Coimbatore, India.
- Gave a talk on Web Search Engine Design and Challenges at SVCE, Anna University in Jan ’10 (<http://cs.svce.ac.in/images/MrSiddharthJonathan/index.html>)

Other Responsibilities and Positions held:

- **Computer Science Masters Admissions Committee**, Stanford University.
 - **International Student Advisor, Computer Science** – Bechtel International Center, Stanford University
 - **Computer Science Masters Student Advisor** –Stanford University
 - **IEEE Officer** – Corporate Liaison, Stanford IEEE Chapter. Organized regular Tech Talk Series for Graduate and Undergraduate Students in Electrical Engineering and Computer Science at Stanford University. Organized “I don’t know to CEO” an Entrepreneurship conference at Stanford University with leading founders and CEOs from Silicon Valley.
 - **Stanford India Association** – Core Group Member.
 - **Mentor for Research Group on Search Engines and Information Retrieval** – Computer Science Department, SVCE, Anna University
 - **Class Representative** - Computer Science Department, SVCE, Anna University
 - **Organizing Committee, Interrupt’03** – National level Technical Symposium at SVCE, Anna University.
-