Paul Suganthan Gnanaprakash Christopher

Madison, WI - 53706   (608)-960-0666   paul.suganthan@gmail.com   linkedin.com/in/paulgc

Interests

Data Management, Data Integration, Data Science, Big Data, Machine Learning, Crowdsourcing.

Education

University of Wisconsin-Madison Aug 2012 - Jan 2018
Ph.D. Computer Sciences Advisor: Prof. AnHai Doan

- Dissertation: Toward Building End-to-End Entity Matching Solutions

M.S. Computer Sciences **GPA**: 3.94 / 4.0

College of Engineering Guindy, Anna University, India July 2008 - May 2012
B.E. Computer Science and Engineering **GPA**: 9.77 / 10.0

- Thesis: Search Engine Enhancement by Extracting Hidden AJAX Content in Web Applications

Work Experience

Software Engineer, Research and Machine Intelligence Mar 2018 - present
Google

- Solving problems in the intersection of data management and machine learning.

Research Assistant, CS. Dept. Aug 2012 - Jan 2018
University of Wisconsin-Madison

- Primary research focuses on developing techniques to scale execution of Entity Matching (EM) workflows containing rules, machine learning (ML), and crowdsourcing operations.
- Scale execution of ML models over the join of two tables using an RDBMS style approach.
- Help domain scientists perform EM by developing a scalable "hands-off" crowdsourced EM solution. System deployed as a service at CloudMatcher.io.
- Monitoring real-time events in Twitter using rules, knowledge base, and ML.

Open Source Developer Jan 2016 - Jan 2018

- Main developer of two Python packages providing tools for scalable string matching (*py_stringmatching* and *py_stringsimjoin*). Managed the end-to-end development and release process.
- Supervised graduate student contributors by reviewing code, and guiding them on best practices.
- Packages are currently being used at multiple organizations and in data science classes.

Software Engineering Intern, Ads Infrastructure May 2015 - Aug 2015
Google, Mountain View

- Building a Natural Language Interface to Databases. Worked on semantic analysis of the natural language query and generation of SQL.
- Developed a failure handling mechanism, which helps in handling ambiguous natural language queries by trying out different interpretations of the query.

Software Engineering Intern, Product Classification May 2014 - Aug 2014
Walmart Labs, Mountain View

- Worked on automatically generating rules for Product classification and optimizing the execution of such rules. Resulted in a SIGMOD 2015 industrial track paper.
- Developed an interactive tool to help analysts write, refine, and manage regex based classification rules.

Teaching Assistant, CS Dept. University of Wisconsin-Madison Jan 2015 - May 2015

- Course TA for the *Data Science* course.

Research Assistant, CS. Dept. May 2011 - Aug 2011
Simon Fraser University, Burnaby, Canada

- Inferring solvability of industrial CNF problems by using the tree width of graphs obtained from CNF problems

Publications

- *Smurf: String Similarity Joins Using Random Forest Conditions* (Under Submission)
  Paul Suganthan G. C., A. Akella, A. Doan
- *MatchCatcher: A Debugger for Blocking in Entity Matching*
  H. Li, P. Konda, Paul Suganthan G. C., A. Doan, EDBT 2018.
- *Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services*
  S. Das, Paul Suganthan G. C., A. Doan et. al., SIGMOD 2017.
- *CloudMatcher: A Cloud/Crowd Service for Entity Matching*
  Y. Govind, E. Paulson, A. Mukilan, Paul Suganthan G. C., et. al., BigDas Workshop, KDD 2017.
- *Human-in-the-Loop Challenges for Entity Matching: A Midterm Report*
  with many other authors, HILDA @ SIGMOD 2017.
- *Magellan: Toward Building Entity Matching Management Systems*
  P. Konda, S. Das, Paul Suganthan G. C., A. Doan et. al., VLDB 2016.
- *Magellan: Toward Building Entity Matching Management Systems over Data Science Stacks*
  P. Konda, S. Das, Paul Suganthan G. C., A. Doan et. al., VLDB Demo 2016.
- *Why Big Data Industrial Systems Need Rules and What We Can Do About It*
  Paul Suganthan G. C., C. Sun, K. Gayatri, H. Zhang, F. Yang et. al., SIGMOD 2015
- *Social Media Analytics: the Kosmix Story*
  with many other authors, IEEE Data Engineering Bulletin Sept. 2013.
- *AJAX Crawler*
  Paul Suganthan G. C., IEEE ICDSE 2012.

Research Projects

Smurf: Scaling Up String Similarity Joins Using Random Forest Conditions
@ University of Wisconsin Madison, with AnHai Doan and Aditya Akella

String similarity joins (SSJs) find strings from two given sets A and B that refer to the same real-world entity. Most current SSJ works consider only join conditions that are a single predicate, e.g., $editdist(x, y) < 3$. In this work, we show that it is possible to create many predicates even when we only have strings to work with, and that we can combine these predicates to form complex meaningful join conditions, which can significantly improve SSJ accuracy. Specifically, we consider complex join conditions such as a random forest ML model. Our key technical contribution is a solution to execute a random forest efficiently over large A and B using a RDBMS style approach.

Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services
@ University of Wisconsin Madison, with Sanjib Das and AnHai Doan

Many works have applied crowdsourcing to entity matching (EM). While promising, these approaches are limited in that they often require a developer to be in the loop. To address this problem, a recent work has proposed Corleone, a solution that crowdsources the entire EM workflow, requiring no developers. While promising, Corleone is severely limited in that it does not scale to large tables. We propose Falcon, a solution that scales up the hands-off crowdsourced EM approach of Corleone, using RDBMS-style query execution and optimization over a Hadoop cluster. Specifically, we define a set of operators and develop efficient implementations. We translate a hands-off crowdsourced EM workflow into a plan consisting of these operators, optimize, then execute the plan. These plans involve both machine and crowd activities, giving rise to novel optimization techniques such as using crowd time to mask machine time.

Magellan: Toward Building Entity Matching Management Systems
@ University of Wisconsin Madison, with Pradap Konda, Sanjib Das, AnHai Doan, et al.

Most current EM works focus only on developing matching algorithms. We argue that far more efforts should be devoted to building EM systems. We present a new kind of EM system, Magellan, which is novel in four important aspects. (1) It provides step by step how-to guides that tell users what to do in each EM scenario. (2) It provides tools to help users do these steps. (3) These tools are built on top of the data analysis and Big Data stacks in Python, allowing Magellan to borrow a rich set of capabilities in data cleaning, IE, visualization, learning, etc. (4) Magellan provides a powerful scripting environment to facilitate interactive experimentation and quick "patching" of the system.

Rule Management in Big Data Industrial Systems
@ University of Wisconsin Madison and WalmartLabs, with Chong Sun, Frank Yang, and AnHai Doan

Big Data industrial systems that address problems such as classification, information extraction, and entity matching very commonly use hand-crafted rules. Today, however, little is understood about the usage of such rules. In this work we explore this issue. Our main conclusions are (1) using rules (together with techniques such as learning and crowdsourcing) is fundamental to building semantics-intensive Big Data systems, and (2) it is increasingly critical to address rule management, given the tens of thousands of rules industrial systems often manage today in an ad-hoc fashion.

Talks

Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services 2017
ACM SIGMOD

Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services 2017
Wisconsin Database Group Seminar

Constructing an Interactive Natural Language Interface for Relational Databases 2016
Wisconsin Database Group Seminar

Building a Natural Language Interface to Databases 2015
Google

Why Big Data Industrial Systems Need Rules and What We Can Do About It 2014
Wisconsin Database Group Seminar

Rule Management in Big Data Industrial Systems 2014
Walmart Labs

Relevant Courses

Database Systems, Data Models and Languages, Machine Learning, Natural Language Processing, Distributed Systems, Computer Networks, Operating Systems, Artificial Intelligence.

Software Skills

- Languages & Tools: Python, Java, C++, Cython, HTML, JavaScript, SQL, Git.
- Data Science Tools: Pandas, Scikit-learn, Numpy, Matplotlib, Dask, Anaconda.
- Experience in Map-Reduce framework — Hadoop.

Service

- External Reviewer, SIGMOD 2018

Academic Achievements

- David DeWitt Fellowship, UW Madison 2012 – 2013
- Best Outgoing Undergraduate Student, CEG 2008 – 2012
- Alumni Golden Jubilee Award for Proficiency in English 2008 – 2009

References

Available upon request.