

# Michael J. Cafarella

916 NW 63rd St.  
Seattle, WA 98107  
Tel: 206-257-9657  
email: [mjc@cs.washington.edu](mailto:mjc@cs.washington.edu)

Computer Science and Engineering  
University of Washington, Box 352350  
Seattle, WA 98195-2350  
Tel: 206-685-2034

---

## RESEARCH INTERESTS

---

Web-oriented research, using techniques drawn from databases and artificial intelligence.

---

## EDUCATION

---

Ph.D., Computer Science, **University of Washington, Seattle, WA** **Expected, 2009**  
Advisors: Dan Suciu and Oren Etzioni  
Dissertation: *Extracting and Managing Structured Web Data*

M.Sc., Computer Science, **University of Washington, Seattle, WA** **2005**

M.Sc., Artificial Intelligence, **University of Edinburgh, Edinburgh, Scotland** **1997**

B.A., Computer Science and History, **Brown University, Providence, RI** **1996**

---

## AWARDS/ACHIEVEMENTS

---

Fulbright Scholar to United Kingdom **1996**

William Gaston Prize for Excellence in Computer Science, Brown University **1996**

---

## PUBLICATIONS

---

### Conferences and Workshops

- [1] **Michael J. Cafarella**, Alon Y. Halevy, Nodira Khossainova: Operators for Web-Scale Data Integration. *Under Submission*.
- [2] **Michael J. Cafarella**: Extracting and Querying a Comprehensive Web Database. *Proceedings of the 4th Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, 2009*.
- [3] **Michael J. Cafarella**, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, Yang Zhang: WebTables: Exploring the Power of Tables on the Web. *34th International Conference on Very Large Databases (VLDB), Auckland, New Zealand, 2008*.
- [4] **Michael J. Cafarella**, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, Yang Zhang: Uncovering the Relational Web. *11th International Workshop on the Web and Databases (WebDB), Vancouver, Canada, 2008*.
- [5] **Michael J. Cafarella**, Christopher Re, Dan Suciu, Oren Etzioni, Michele Banko: Structured Querying of Web Text: A Technical Challenge. *Proceedings of the 3rd Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, 2007*.
- [6] **Michael J. Cafarella**, Dan Suciu, Oren Etzioni: Navigating Extracted Data with Schema Discovery. *10th International Workshop on the Web and Databases (WebDB), Beijing, China, 2007*.
- [7] Michele Banko, **Michael J. Cafarella**, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web. *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India, 2007*.
- [8] Luke McDowell, **Michael J. Cafarella**: Ontology-Driven Information Extraction with OntoSyphon. *Proceedings of the 5th International Semantic Web Conference (ISWC), Athens, GA, 2006*.

[9] **Michael J. Cafarella**, Oren Etzioni: A Search Engine For Natural Language Applications. *Proceedings of the 14th International World Wide Web Conference (WWW), Tokyo, Japan, 2005*.

[10] **Michael J. Cafarella**, Doug Downey, Stephen Soderland, Oren Etzioni: KnowItNow: Fast, Scalable Information Extraction from the Web. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Vancouver, Canada, 2005*.

[11] Oren Etzioni, **Michael J. Cafarella**, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Web-scale Information Extraction in KnowItAll (preliminary results). *Proceedings of the 13th International World Wide Web Conference (WWW), New York, NY, 2004*.

## Journals

[12] **Michael J. Cafarella**, Alon Halevy, Jayant Madhavan: Web-Scale Extraction of Structured Data. *SIGMOD Record 37(4), 2008*.

[13] **Michael J. Cafarella**, Edward Chang, Andrew Fikes, Alon Y. Halevy, Wilson C. Hsieh, Alberto Lerner, Jayant Madhavan, S. Muthukrishnan: Data Management Projects at Google. *SIGMOD Record 37(1), 2008*.

[14] Luke McDowell, **Michael J. Cafarella**: Ontology-Driven Unsupervised Instance Population. *Journal of Web Semantics 6(3), 2008*.

[15] **Michael J. Cafarella**, Dan Suciu, Oren Etzioni: Structured Queries Over Web Text. *IEEE Data Bulletin (Special Issue on Web-Scale Data, Systems, and Semantics), 31(4):45-51, 2006*.

[16] Oren Etzioni, **Michael J. Cafarella**, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence 165(1), 2005*.

[17] Oren Etzioni, **Michael J. Cafarella**, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, Alexander Yates: Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI), San Jose, CA, 2004*.

## Non-Refereed Work

[18] **Mike Cafarella**, Doug Cutting: Building Nutch: Open Source Search. *ACM Queue 2(2): 2004*.

---

## INVITED TALKS

[1] *Structured Queries Over Web Text*, Database Seminar, University of Wisconsin, Madison, WI, May 2007.

[2] *Structured Queries Over Web Text*, Yahoo! Research, Santa Clara, CA, April 2007.

[3] *A Search Engine for Natural Language Applications*, UCLA Department of Electrical Engineering, Los Angeles, CA, December 2006.

[4] *KnowItAll*, Amazon, Inc., Seattle, WA, March 2006.

[5] *A Search Engine for Natural Language Applications*, Microsoft Research, Redmond, WA, April 2005.

[6] *KnowItAll Assessment*, Google, Inc., Mountain View, CA, May 2004.

---

## PROFESSIONAL SERVICE

[1] Program Committee for workshop on **Collaborative Web Tagging**, 2006.

[2] External Reviewer, **World Wide Web** conference (WWW), 2006.

[3] Program Committee for workshop on **Information Integration on the Web** (IIWeb), 2007.

[4] Reviewer, **SIGMOD Record**, Volume 36, Number 3, Sept 2007.

- [5] Reviewer, **SIGMOD Record**, Volume 36, Number 4, Dec 2007.  
[6] Reviewer, **Journal of Artificial Intelligence Research (JAIR)**, Volume 30, Sept-Dec 2007.  
[7] Reviewer, **Transactions on Database Systems (TODS)**, Volume 34, Mar 2009.

---

## RESEARCH EXPERIENCE

---

*WebTables*: Created an extraction system for tabular data on the Web. Led a research project at Google (which hosted the work) that involved three other graduate student interns. The WebTables extractor recovered *120 million* relational databases from a general Web crawl. These databases gave rise to several applications, including a structured-data search engine. By computing statistics over metadata from these databases, WebTables enabled several new database applications, including a *schema autocomplete* tool, and a tool for automatically computing attribute synonyms. The extracted databases are now used by several Google product and research groups.

*Octopus*: Served as lead graduate student in research project to create a data integration tool for extracted Web data. The scale of Web data makes it very appealing for integration and manipulation; for example, the user may want to compile a database of all the countrys computer science professors out of information that is scatterd across dozens of departmental websites. Octopus introduced several operators that make Web-scale integration substantially easier, such as relevance-ranking and clustering for data sources, and integration of sources that have no published metadata. Like WebTables, this project took place at Google.

*TextRunner*: Worked with one other graduate student to research and build a system that attempts to extract n-ary fact tuples from general Web text. For example, a biography about Einstein might yield [Einstein, was-born-in, 1875]. Unlike many other textual extractors, TextRunner does not rely on any user guidance for its extraction rules, making it suitable for processing the entire Web. Competing systems (*e.g.*, *Snowball* and *DIPRE*) obtain data for just a handful of domains at a time and thus cannot scale easily to the whole Web.

*Bindings Engine*: Created novel search engine to improve performance of natural language applications, which require use of very large Web text corpora. The Bindings Engine accepts keyword queries with embedded variables: for example, *cities such as <NounPhrase>*. Designed and implemented new variant of inverted index that reduced runtimes for Web extraction applications from hours to minutes.

*TGEN*: Created domain-independent tool that synthesizes a relevant relational schema given a set of extracted triples. For example, running an information extractor over a Web site about movies might yield triples such as [Frenzy, was-made-in, 1972], [Hitchcock, directed, Vertigo], and others. Unfortunately, the extracted facts are usually noisy and incomplete, so a good schema is usually difficult to find. When tested on a focused crawl of nutrition-related Web pages, TGEN found many high-quality relations, on topics such as *vitamins/minerals, foods, health authorities, diseases, etc.*

*KnowItAll*: Collaborated with several other students on an unsupervised system that extracts high-quality facts from general natural-language text on the Web. KnowItAll employs user-input extraction phrases; for example, using the extraction phrase *cities such as X*, where *X* is a noun-phrase, allows KnowItAll to compile lists of cities. The system then assesses the quality of extracted data using statistics computed over Web search engine hitcounts. Demonstrated high performance and recall across five domains, and automatically obtained more than 50,000 facts in a single run.

---

## PROFESSIONAL EXPERIENCE

---

**Graduate Research Assistant, University of Washington, Seattle** **2003 - Present**  
*Advisors: Oren Etzioni and Dan Suciu*

Research assistant for multiple projects in Web information extraction and systems for managing extracted data. Work includes the above-listed *TextRunner*, *Bindings Engine*, *TGEN*, and *KnowItAll* projects.

**Research Intern, Google, Inc.** **2007 - 2008**  
*Manager: Alon Halevy*

Research intern in the structured data research group. Lead researcher on the *WebTables* and *Octopus* projects, listed above.

**Co-Creator, Hadoop and Nutch Open Source Projects** **2002 - 2007**

Co-creator and engineer for two open source projects: Nutch, a web search engine; and Hadoop, a suite of cluster computing tools. Hadoop is widely popular and is deployed at Yahoo! (on over 10,000 CPU cores), Facebook, NYTimes.com, and the Google-IBM academic cluster (shared by MIT, CMU, Stanford, UC Berkeley, University of Washington, and University of Maryland).

**Engineer, Tellme Networks** **2000 - 2002**

Designed and wrote code for the largest and most popular voice browser. Designed first version of Call Control XML (now a W3C standard).

**Engineer, Marimba Corporation** **1998 - 2000**

Designed and wrote code for a large scale software deployment system.

---

TEACHING EXPERIENCE

**TA, Advanced Internet and Web Services (CSE454), Univ. of Wash. Winter 2004, Fall 2006**

Sole teaching assistant for two quarters of class on search engine design and information retrieval algorithms. Designed class programming assignments, graded student homeworks and programming assignments, and supervised independent group final projects.

**Seminar Organizer, University of Washington, Seattle** **Fall, 2006**

Organized weekly graduate seminar to read and discuss foundational database papers.

**TA, Department of Computer Science, Brown University** **1993 - 1996**

Undergraduate teaching assistant for introductory programming classes. Assisted in designing and grading programming assignments.

---

SOFTWARE

[1] **Hadoop**, co-creator. A widely-used open-source software suite of cluster computing tools. Includes a distributed filesystem and an implementation of the Map/Reduce framework. Popular in both industry and academia.

[2] **Nutch**, co-creator. An open-source search engine, including a Web crawler and search query processing system. Used for research and teaching, and deployed at Oregon State University.

---

REFERENCES

**Oren Etzioni**

Professor, Department of Computer Science and Engineering.  
University of Washington  
Box 352350  
Seattle, WA 98195-2350  
(206)-685-3035  
etzioni@cs.washington.edu

**Alon Y. Halevy**

Head, Structured Data Research Group, Google, Inc.  
1600 Amphitheatre Parkway  
B43 11DB  
Mountain View, CA 94043  
(650)-253-2574

halevy@google.com

**Dan Suci**

Professor, Department of Computer Science and Engineering.

University of Washington

Box 352350

Seattle, WA 98195-2350

(206)-685-1934

suciu@cs.washington.edu

---