

## Trevor Strohman

---

Google  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
strohman@cs.umass.edu  
<http://ciir.cs.umass.edu/~strohman>

---

### Research interests

Information retrieval, data processing, algorithms, databases, systems research.

---

### Education

**Ph.D. in Computer Science** December 2007

University of Massachusetts, Amherst, MA.

Passed portfolio (candidacy requirements) with distinction, December 2005.

Dissertation: *Efficient Processing of Complex Features for Information Retrieval*

Advisor: W. Bruce Croft

**M.S. in Computer Science** September 2005

University of Massachusetts, Amherst, MA.

Master's project: *Optimization Strategies for Complex Queries*

Advisor: W. Bruce Croft

**B.S. in Computer Science** December 2003

California Polytechnic State University, San Luis Obispo, CA

*Summa cum laude* (3.96/4.0)

GRE: 650v/800q/800a, GRE Computer Science: 870

Senior project: *Shared-tail Database Architecture*

---

### Awards

UMass Portfolio Distinction (2005)

J.L. Moore Ph.D. Fellowship (2003, 2004, 2006)

Department Scholarship (2002)

Outstanding Undergraduate, Computer Science (2001)

Chevron Academic Scholarship (2000)

---

**Software****Galago** not yet available

Galago is an open-source search engine that is the basis of my dissertation work. The main purpose of Galago is to build binned indexes, which are custom indexes based on a particular query formulation. However, Galago also supports a subset of the Indri functionality, making it simultaneously useful as an educational tool. It incorporates a MapReduce-style pipeline, which makes for an extremely flexible and scalable indexer. The complete implementation is in Java, but a C++ query processing engine can be used for high throughput query loads.

**Indri** <http://www.lemurproject.org/indri>

Indri is a search engine based on language modeling and the inference network model. It incorporates an extensive query language, developed by Don Metzler, Howard Turtle and W. Bruce Croft, which enables highly flexible ranking. It also supports retrieval using a cluster of machines for extra speed. Indri is used by many universities as a basis for information retrieval research, but it is also used in commercial settings, and is the basis of a supported commercial product.

**Imprint** <http://www.ampersandbox.com/imprint>

Imprint is an envelope and label printing product for Mac OS X. I designed, wrote, and marketed this product over a three year period, and sold over 1000 copies.

**Scholar** <http://ciir.cs.umass.edu/~strohman/code>

Scholar is a program for Mac OS X that decomposes two-column research papers into a single, scrolling column reminiscent of an online news article. This is a particularly efficient way to read conference papers on a laptop computer screen.

**T-Tree** no longer available

T-Tree is the physical layer of an ISAM database. It includes a B+-Tree and a record manager implementation. The library is multithreaded, multiplatform, and has a dynamic cache system that responds to system load, making it suitable for both small-footprint clients and large server situations. T-Tree was a component of a commercial client backup application.

**IR Eval** <http://ciir.cs.umass.edu/~strohman/code>

The ireval toolkit is a set of Java classes that evaluates the output of an information retrieval system. It serves the same purpose as the popular trec\_eval tool, except ireval can be easily embedded into a larger system (like a self-training system).

---

**Books****Search Engines: Information Retrieval in Practice**

Croft, W. B., Metzler, D. and Strohman, T., 2009.

---

**Conference papers****Efficient Document Retrieval in Main Memory**

SIGIR 2007: Strohman, T., Croft, W. B.

**Optimization Strategies for Complex Queries**

SIGIR 2005: Strohman, T., Turtle, H., Croft, W. B.

- 
- Short papers
- Recommending Citations for Academic Papers**  
SIGIR 2007: Strohman, T., Croft, W. B., Jensen, D.
- Indri: A language model-based search engine for complex queries**  
Strohman, T. Metzler, D., Turtle, H., Croft, W.B.  
International Conference on Intelligence Analysis (IA), 2005.
- 
- Other papers
- Indri: Lessons Learned From Three Terabyte Tracks**  
TREC 2006: Metzler, D., Strohman, T. and Croft, W.B.
- Low Latency Index Maintenance in Indri**  
Strohman, T., Croft, W.B.  
SIGIR Workshop on Open Source Information Retrieval (OSIR), 2006.
- Indri at TREC 2005: Terabyte Track (Notebook Version)**  
TREC 2005: Metzler, D., Strohman, T., Zhou, Y. and Croft, W.B.
- UMass Robust 2005 Notebook: Using Mixture of Relevance Models for Query Expansion**  
TREC 2005: Metzler, D., Diaz, F., Strohman, T., Croft, W.B.
- Indri at TREC 2004: Terabyte Track**  
TREC 2004: Metzler, D., Strohman T., Turtle H., and Croft, W.B.
- UMass at TREC 2004: Notebook**  
TREC 2004: Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Metzler, D., Strohman, T., Turtle, H., and Wade, C.
- 
- Technical reports
- Recommending Citations for Academic Papers**  
IR-466 (2006): Strohman, T., Croft, W. B., Jensen, D.
- Custom Object Layout for Garbage Collected Languages**  
TR-06-06 (2006): Novark, G., Strohman, T., Berger, E.
- Indri: A language model-based search engine for complex queries (extended version)**  
IR-407 (2005): Strohman, T. Metzler, D. Turtle, H., Croft, W.B.
- Dynamic Collections in Indri**  
IR-426 (2005): Strohman, T.
- 
- Presentations
- Efficient Document Retrieval in Main Memory**  
Strohman, T.  
SIGIR 2007, Amsterdam, Netherlands, July 2007.

## Flexibility and Efficiency in Text Retrieval Systems

Strohman, T.  
Endeca, Cambridge, Massachusetts, May 2007.

## Using the Lemur Toolkit for Information Retrieval

Strohman, T. and Ogilvie, P.  
SIGIR Tutorial, 2006.

## Resume (Re)search: Present and Future

Strohman, T.  
Monster Worldwide, Maynard, MA, August 2006.

## Optimization Strategies for Complex Queries

Strohman, T.  
SIGIR 2005, Salvador, Brazil, August 2005.

---

## Work experience

STAFF SOFTWARE ENGINEER, Google February 2008 - present  
Lead a project to improve the efficiency and latency of the search engine. Worked on a search index builder refactoring project. Extensively analyzed query results caching strategies based on query logs.

Languages: C++, Python, Sawzall, Go, JavaScript, Java

CONSULTANT, Lexalytics February 2007  
Added space reclamation and index merging to Indri.  
Languages: C++

CONSULTANT, Monster Worldwide June 2006 - December 2007  
Created demo applications and proposed methods for improving a component of the monster.com search system.  
Languages: PHP, C++, C#, XSLT

RESEARCH ASSISTANT, UMass Amherst September 2003 - December 2007  
Created Galago, an system for building custom indexes for efficient retrieval.  
Primary developer of Indri, a search engine that combines language modeling and the inference network framework.  
Languages: C++, Python, Java, PHP, C#, Groovy

PRESIDENT, Ampersandbox January 2003 - present  
Developed Imprint, a label and envelope printing program for Mac OS X.  
Languages: Objective-C (primary), Python

SENIOR SOFTWARE ENGINEER - Veritas Software June 1998 - September 2002  
Developed components of NetBackup Professional, a client/server backup product for Windows clients with Windows and Solaris servers.

Developed T-Tree, the physical layer of an ISAM database system.  
Developed tools to analyze T-Tree databases and fix inconsistencies  
Designed next-generation architectures for future desktop backup products  
Optimized compute-intensive routines in x86 assembly language  
Wrote code to store only changes between versions of files on a backup server, and to recombine multiple change files in a single pass.  
Languages: C++ (primary), C#, Perl, Visual Basic, x86 assembly

SOFTWARE ENGINEER, genX Software August 1997 - June 1998  
Created a cross-platform (Linux/Windows) server for a fast-paced 3D game  
Modeled continuous player motion for more accurate player collision detection  
Languages: C (primary), Perl, x86 assembly

SOFTWARE ENGINEER, WaveQuest Software January 1997 - August 1997  
Worked on network architecture for a networked Windows game  
Wrote path-finding code for a combat simulation game  
Languages: C

---

Professional  
activities

Grant Review Panelist, National Science Foundation 2007  
Program Committee, SIGIR 2009-2010  
Program Committee, WWW 2010  
Reviewer, SIGIR 2007-2009  
Reviewer, Transactions on Information Systems