

# AKANKSHA BAID

---

1210 W. Dayton St, Madison, WI 53706.  
(510) 676-2139 • baid@cs.wisc.edu • <http://www.cs.wisc.edu/~baid>

## EDUCATION

---

### **Ph.D.(in progress), Computer Sciences,**

Advisor: Prof. Jeffrey Naughton,

Thesis: *Toward Scalable Keyword Search over Relational Data*

University of Wisconsin, Spring 2011 (*expected*).

### **B.S., Computer Science and Applied Mathematics,**

California State University, East Bay, May 2005.

Summa Cum Laude with Honors

## RESEARCH INTERESTS

---

Database systems and applications.

## SYSTEMS SKILLS AND EXPERIENCE

---

- *Languages:* C, C++, Java, Perl, Python, Shell Scripting, XML, XSLT, SQL
- *Operating Systems:* MS Windows, Linux, UNIX
- *Software Packages:* IBM DB2, SQL Server, Postgres SQL, Lucene, Eclipse, JDK

## PUBLICATIONS

---

- [1] Exploring Non-Answers in Keyword Search over Structured Data. (*In preparation*)
- [2] Akanksha Baid, Ian Rae, Jeixing Li, AnHai Doan and Jeffrey Naughton. Toward Scalable Keyword Search over Relational Data. **VLDB 2010**.
- [3] Akanksha Baid, Ian Rae, AnHai Doan and Jeffrey Naughton. Toward Industrial-Strength Keyword Search over Relational Data. **ICDE 2010**.
- [4] Eric Chu, Akanksha Baid, Xiaoyong Chai, Anhai Doan and Jeffrey Naughton. Combining Keyword Search and Forms for Ad Hoc Querying of Databases. **SIGMOD 2009**.
- [5] AnHai Doan, Jeffrey F. Naughton, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, Ba-Quy Vuong. The Case for a Structured Approach to Managing Unstructured Data. **CIDR 2009**.
- [6] AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, Ba-Quy Vuong. Information Extraction Challenges in Managing Unstructured Data. **SIGMOD Record, Winter 08, Special Issue on Managing Information Extraction**.
- [7] Alkis Simitsis, Akanksha Baid, Yannis Sismanis, Berthold Reinwald. MCX: Multidimensional Content Exploration. **VLDB 2008**.
- [8] Akanksha Baid, Andrey Balmin, Heasoo Hwang, Erik Nijkamp, Jun Rao, Berthold Reinwald, Alkis Simitsis, Yannis Sismanis, Frank van Ham. DBPubs: Multidimensional Exploration of Database Publications. **VLDB 2008 Demo**.
- [9] Eric Chu, Akanksha Baid, Ting Chen, AnHai Doan, and Jeffrey F. Naughton. A Relational

Approach to Incrementally Extracting and Querying Structure in Unstructured Data. **VLDB 2007**.

## PATENT APPLICATIONS

---

[1] A. Baid, B. Reinwald, A. Simitis, Y. Sismanis. System, method, and apparatus for multidimensional exploration of content items in a content store. Filed by IBM Corporation in 2009.

## EXPERIENCE

---

### **Sep 2007 - present,**

Research Assistant with Prof. Jeffrey Naughton

*Dept. of Computer Sciences, UW-Madison*

My thesis work revolves around developing techniques and algorithms for better querying of both structured data and unstructured data. In the course of my research I have focused on building end-to-end deployable solutions for keyword search over relational data. At Wisconsin I also work closely with Prof. AnHai Doan.

### **Summer 2007,**

Research Intern with Yannis Sismanis

*IBM Almaden, San Jose, CA*

During my internship at IBM Almaden I worked on building a system and designing scalable algorithms for effectively analyzing and exploring unstructured data by combining keyword search with OLAP-style aggregations, navigation, and reporting. This work led to two papers in VLDB 2008. IBM has also filed a patent for the same.

### **Summer 2006,**

Research Intern, with Hans Granqvist

*Weblogs.com, Mountain View, CA*

Designed and developed a blog analysis tool that analyzes the 3000 pings/sec that Weblogs.com received. The work involved working in a team of two people and developing production level code that was deployed on a weekly basis.

### **Jan 2003 - May 2005,**

Research Assistant with Prof. Hilary Holz

*California State University, Hayward, CA*

As a research assistant for the Adaptive Hypermedia and Assistive Technologies Lab, I worked on an automated, machine-learning based Adaptive Unix tutorial that presented material to students based on their competence with UNIX like systems.

## SELECTED RESEARCH PROJECTS

---

### **Exploring Non-Answers in Keyword Search over Structured Data**

Keyword search over structured data (KWS) returns the relationships between the users keywords based on key-foreign key joins in the underlying structured data. This is done by mapping the keyword query to many structured queries, each of which corresponds to a relationship. Our goal is to also display non-answers to the user. Unlike non-answers in IR or for a single structured query, we were faced with the problem of evaluating and explaining non-answers for hundreds or even thousands of structured queries in response to a potentially ambiguous keyword query. We built a system which utilizes a lattice based structure for efficiently determining non-answers and the underlying cause of the non-answer.

### **Scalable Keyword Search (KWS) over Structured and Semi-Structured Data**

Like IR systems, KWS systems accept keywords as input, however instead of documents they return relationships between the user's keywords based on the structure in the underlying database. We found that current KWS solutions often require the solution of sub-problems

that are NP-complete. Consequently, the time spent on these problems often grows very quickly as the data size, the query size, the complexity of the schema, or the size of the answer space increases. In this work we leveraged template based querying and indexing techniques to bypass the performance bottle neck in KWS systems. We built a working system and conducted a user study to evaluate the effectiveness of our solution.

### **Combining Keyword Search and Forms for Ad Hoc Querying of Databases**

Database systems are hard to query for users uncomfortable with a formal query language. To address this problem, form-based interfaces and keyword search have been proposed; while both have benefits, both also have limitations. We investigated combining the two approaches. Specifically, we proposed to take as input a target database and then generate and index a set of query forms offline. At query time, a user with a question to be answered issues standard keyword search queries; but instead of returning tuples, the system returns forms relevant to the question. The user may then build a complex structured query with one of these forms and submit it back to the system for evaluation. We addressed challenges that arise in form generation, keyword search over forms, and ranking and displaying forms.

### **Multidimensional Content eXploration (MCX)**

In this work we combined keyword search with OLAP-style aggregation, navigation, and reporting. We explored how unstructured data or data with limited metadata should be organized in a well-defined multidimensional structure, so that sophisticated queries can be expressed and evaluated. The metadata provided traditional OLAP static dimensions that were combined with dynamic dimensions discovered from the analyzed keyword search result, as well as measures for document scores based on the link structure between the documents. We also provided means for multidimensional content exploration through traditional OLAP roll-up and drill-down operations on the static and dynamic dimensions, solutions for multi-cube analysis and dynamic navigation of the content. We implemented and deployed a system (DBPubs) over the database publication domain at IBM Almaden.

### **A Relational Approach to Incrementally Extracting and Querying Structure in Unstructured Data**

There is a growing consensus that it is desirable to query over the structure implicit in unstructured documents, and that ideally this capability should be provided incrementally. However, there is no consensus about what kind of system should be used to support this kind of incremental capability. We explore using a relational system as the basis for a workbench for extracting and querying structure from unstructured data. As a proof of concept, we applied our relational approach to support structured queries over Wikipedia. We show that the data set is always available for some form of querying, and that as it is processed, users can pose a richer set of structured queries. We also provide examples of how we can incrementally evolve our understanding of the data in the context of the relational workbench.

---

## TEACHING EXPERIENCE

Served as teaching assistant for the following courses: Numerical Analysis, Introduction to Algorithms, Analysis of Software Artifacts and Graduate Operating Systems.

---

## SELECTED ACTIVITIES AND HONORS

- External reviewer for ICDE 2008, ICDE 2010, VLDB 2010
- Computer Science department High Achievement award 2005
- Mathematics department High Achievement award 2005

---

## REFERENCES

Available upon request