

The amount of information in the world is growing exponentially, increasing tenfold every five years. My Ph.D. research focuses on keyword search within relational databases because relational databases store much of the world's information but cannot be accessed via traditional search techniques. Relational keyword search is considerably more complex than searching unstructured text due to the numerous relationships among tuples that match disjoint sets of search terms: information from different relations must be joined to create coherent search results. My work to date emphasizes the evaluation of existing search techniques, which has led me to several methods that improve the quality of search results. In future work, I anticipate developing new algorithms to mitigate the scalability and performance problems endemic to existing approaches. My research addresses critical problems that must be overcome before relational keyword search will be suitable for real-world retrieval tasks.

### **Evaluation of Relational Keyword Search Strategies**

Standardized evaluation has been at the heart of information retrieval (IR) evaluation since the 1960s. My creation of an evaluation benchmark (published at CIKM '10) specifically designed for relational keyword search fills a void noted by numerous researchers during the past decade. My evaluation benchmark follows the precedent of the IR community for the evaluation of retrieval systems whereas previous evaluations had not applied these best practices. Using my benchmark, I've determined that many of the claimed improvements in the literature are actually implementation artifacts or the result of experimental design decisions. I've also shown that the scalability and performance of existing search techniques are not acceptable for databases with millions of tuples. The immediate value of my work is twofold: 1) it provides researchers with objective experiments to validate proposed improvements, and 2) it clearly identifies a direction for future research by exposing areas where improvement is critical.

The design of the evaluation benchmark was non-trivial, requiring me to investigate two related research questions. First, what would users search for if relational keyword search systems were available? Realistic evaluation depends upon realistic queries. My classification of queries in existing search logs is the first in the context of relational keyword search. In my work, I propose a technique to synthetically generate evaluation query workloads and show that these queries are representative of real user queries. Second, why do existing evaluations arrive at different conclusions? Researchers inevitably face a myriad of experimental design decisions but have little understanding of how they affect their experimental results. My work in this area establishes the impact that query and result semantics, the database schema, and the use of database subsets instead of the original database all have on experimental results.

### **Ranking**

Users depend upon search engines to identify the most relevant of the millions of web pages that match a query. Relational keyword search systems have a similar challenge due to the many relationships among tuples that contain search terms. At KEYS '10, I demonstrated the benefits of adapting previous research from the IR community to keyword search in databases; the novelty of this work stems from its departure from the two families of scoring functions that have persisted for nearly a decade. For CIKM '11, I used machine learning to determine the importance of previously proposed ranking factors. My work revealed that intuitively good schemes perform poorly when ranking marginally relevant results, and my novel ranking scheme significantly outperforms them at high recall levels. By showing that computationally expensive scoring factors are unnecessary, my research also enables novel schemes that improve scalability and performance.

## Impact

The impact of my work is threefold. First, keyword search in databases enables existing data repositories to be used more effectively. Scientists use databases to store their data, and the general public unconsciously interacts with databases through popular websites; improved search techniques will benefit both of these diverse groups. Second, relational keyword search alleviates the need for programmers to develop custom search interfaces for different databases: when a database is created, a keyword search interface will be immediately available much like standardized application programming interfaces (APIs) allow a database to be accessed from a variety of programming languages. Third, many existing applications will benefit from my work. Electronic medical records are revolutionizing health care so imagine the benefits if doctors had instant access to previously hidden relationships between patients. For example, an outbreak of staph infections could be quickly traced to the common relationship among patients like sharing a particular hospital room. The intelligence community would also benefit from improved decision support systems as analysts would have access to a greater amount of data and the relationships between different nuggets of information. Another application is expert recommender systems that integrate information from users' social networks to improve recommendations for potential collaborators. Hence, the impacts of my research are significant in their ability to transform how society uses existing data repositories.

## Research Plan

“If you don't know where you are going, any road will get you there.” ~ Lewis Carroll

Robust empirical evaluation is a cornerstone of my research. Good evaluation methodology is critical to ensure research stays focused on the overarching goal and progresses incrementally toward it. My research goal is to make relational keyword search practical for existing data repositories so other scientists and the general public can use information more effectively. My work on evaluation undergirds this objective by ensuring that experimental results accurately reflect the improvements of proposed techniques. My accomplishments to date show that much work remains, particularly reducing techniques' memory footprint and execution time.

I anticipate focusing heavily on the scalability and performance of relational keyword search strategies during the next few years. My research will investigate

- minimizing memory consumption by inferring information from the underlying database,
- indexing schemes to improve overall performance, and
- the tradeoffs between scalability and performance.

My previous work on ranking suggests how to alleviate a number of current bottlenecks. For example, inexpensive graph traversal algorithms can be substituted for more expensive heuristics without sacrificing retrieval quality. Integrating this research with an open-source database management system will improve both scalability and performance by reusing the highly optimized database indexes and query processing schemes that already exist. Given my previous successes at CIKM '10 and '11, I anticipate continuing to publish my work at that venue; my work to improve scalability and performance is also applicable to larger database conferences such as VLDB and SIGMOD.

As a complementary research agenda, I will also investigate computer science education. One area I will focus on is object-oriented design because many students appear to lack this fundamental skill even after completing their undergraduate education. Given the widespread adoption of object-oriented programming paradigms in industry, improving undergraduate education in this area will significantly benefit the workforce.

I will also investigate how to emphasize software dependability throughout the computer science curriculum because the world increasingly depends upon software to manage critical infrastructure but this software often cannot be trusted. I believe that we must emphasize dependability early if we will ever build dependable commodity software. My well-received presentation at SIGCSE '10 on improving the coverage of security when teaching web programming suggests that this venue is an ideal place to publish my educational research.

NSF grants have supported me at both the undergraduate and graduate levels, and I anticipate receiving funding from NSF's Information Integration and Informatics program for my continued research in relational keyword search. I believe the practical nature of my research will appeal to students at all levels, and I look forward to including undergraduates in my research program. Solving real-world problems has also been shown to increase the involvement of historically under-represented students in computer science. NSF's Transforming Undergraduate Education in STEM (TUES) program would be an ideal source of funding for education research, given the scope of the issues I will investigate.