Research Statement
Hui Fang

Various kinds of text information, such as web pages, emails, instant messages, literature and blogs, are being continuously produced everywhere in the world in every possible way. Such a huge amount of information overwhelms users, and poses significant challenges in text information management. Search engines are by far the most useful tools to help users find information, and have become more and more essential in all aspects of our life. It is estimated that 6 billion user queries were submitted to search engines in October, 2006 alone. Clearly, the accuracy of search engines significantly affects our productivity and our quality of life.

My primary research interest is information retrieval (IR), which is the underlying science of search engines. The accuracy of a search engine is mainly determined by its retrieval model, and developing effective retrieval models is arguably the most important research problem in IR. Although retrieval models have been studied for decades, the current retrieval models are still far from satisfactory. In particular, there is no way to predict the performance of a retrieval function analytically. As a result, heavy parameter tuning is always necessary for any state of the art retrieval function to achieve optimal performance. Furthermore, current retrieval models are also limited in supporting a user to find other types of information than documents. I have a broad interest in developing general retrieval models that can effectively help users find various types of information from all kinds of data collections. In my current research, I proposed and developed a novel axiomatic framework for studying and developing retrieval models that can potentially lead to major breakthroughs in developing new retrieval models. I also developed new retrieval models for retrieving relevant entities (e.g., persons) as opposed to documents, and proposed useful techniques for solving various problems in bioinformatics.

Current Research

Axiomatic Approaches to Information Retrieval (Dissertation Work): The key challenge in developing retrieval models is to model relevance appropriately. Over the decades, many text retrieval models have been proposed, but all of them model relevance indirectly through converting the problem of relevance modeling to some other problems such as vector space algebra or probabilistic models. Unfortunately, such indirect modeling of relevance prevents us from understanding what makes a retrieval function perform well. Thus, we have to rely on extensive, labor-intensive experiments to tune retrieval parameters and optimize the empirical performance of a retrieval function.

In my dissertation, I proposed a novel axiomatic approach to modeling relevance directly through term-level retrieval constraints. The axiomatic framework makes it possible to predict the empirical performance of a retrieval function analytically, and to derive more robust and effective retrieval models. Specifically, I analyzed how a reasonable retrieval function should behave, and formalized the necessary properties of a reasonable retrieval function as retrieval constraints. These constraints provide a good explanation for many empirical observations about retrieval functions, and make it possible to evaluate a retrieval function analytically rather than relying on extensive experiments as before ([ACM SIGIR04 Best Paper]). Furthermore, I derived new retrieval functions using the axiomatic approach. The main idea is to search for retrieval functions that satisfy all the reasonable retrieval constraints. The derived retrieval functions are shown to be more robust and less sensitive to the parameter setting than the existing retrieval functions with comparable optimal performance ([ACM SIGIR05]). Recently, I have exploited the axiomatic approach to derive retrieval models that support semantic term matching as opposed to syntactic
term matching commonly used in most existing retrieval models ([ACM SIGIR06]).

**Entity Retrieval:** Most search engines return documents (e.g., web pages), but users also look for other types of information. For example, a user may want to find persons who are knowledgeable about a given topic. Such an information need exists in many real-world applications, including, e.g., assigning submissions to PC members of a conference for reviewing and routing complaints to staff responsible for the corresponding problems in customer service. It is clear that a regular retrieval model can not be applied to such non-traditional retrieval tasks directly. In my recent study ([ECIR07]), I developed new retrieval models for such entity retrieval problems that aim to find relevant entities (e.g., persons) for a given query over both unstructured data (i.e., text collections) and structured data (i.e., XML fragments).

**Bioinformatics:** Bioinformatics is an exciting research area that allows me, as a computer scientist, to contribute my knowledge and expertise to solve the practical problems posed or inspired by the management and analysis of huge amounts of biological data. My work in this area includes (1) new clustering techniques and evaluation methodology for microarray data analysis ([ISMB04 poster, Master thesis]); (2) statistical methods for motif function prediction ([Applied Bioinformatics]); (3) robust retrieval models for biological literature ([TREC03, TREC05]).

**Future Directions**

The potential impact of breakthroughs in retrieval models is huge. We are far from understanding what makes a retrieval function perform well for a particular retrieval task. The axiomatic approach provides a promising new framework to achieve this goal. For my future research, I plan to further explore more robust and effective retrieval models, and extend these models to different domains, such as bioinformatics.

**Optimal Retrieval Models:** Developing an optimal retrieval function has huge impact, because it will improve the accuracy of every search engine. The proposed axiomatic framework has a great potential for developing new optimal retrieval functions, that are robust and effective for different applications, through searching for retrieval functions that satisfy all the reasonable constraints. In my current work, I searched only the neighborhood of existing retrieval functions. In the future, I plan to explore a general strategy to systematically search for an optimal retrieval function in the whole function space.

**Query Adaptive Retrieval Models:** Many studies in IR show that no single retrieval model is able to return satisfactory results for every query. The main reason is that the existing retrieval models fail to adjust their scoring functions dynamically based on queries. I plan to extend my work on the axiomatic approach to design retrieval models that are able to automatically adapt to different queries based on their characteristics and inherent difficulty. Furthermore, the current search engines provide little support for a user to refine queries when the search results are poor. I plan to study interactive support to help users formulate better queries when necessary.

**Opinion Retrieval Models:** A common information need is to find opinions about a given topic. The Web is a valuable resource of opinions including various types of information, such as web pages, mailing lists and blogs. Such a subjective information collection poses new challenges to IR techniques because traditional retrieval models are designed to rank documents based only on their contents. I plan to develop retrieval models that consider both contents and opinions of
documents. Such retrieval models are also useful in many other real world applications, such as content-based web advertising, where a set of documents sharing the same topic and same opinion should be identified for a given document.

**Bioinformatics:** The advent of new technologies in biology has resulted in an overwhelming increase in biological literature. One way to utilize the knowledge in literature is to develop effective retrieval models to help biologists collect information about a particular topic. Although the problem can be regarded as a standard retrieval problem, the definition of “relevance” is unique, and a specialized search engine targeting at the domain would be necessary to optimize retrieval performance. I believe that axiomatic retrieval models provide unparalleled opportunities for incorporating a specific definition of relevance appropriate for biology literature search, and I plan to develop effective axiomatic retrieval models for this particular domain.

Biologists need to not only collect published information about a given topic, but also mine unknown knowledge from literature and biological data such as genome sequences, protein interaction, and microarray expression data. Both retrieved and mined information can be combined with biologists’ knowledge to form novel hypothesis and make new discoveries. Data mining techniques have been applied to many biological problems to discover unknown knowledge, but existing techniques usually produce a huge number of results, which often require biologists to spend tremendous time to go through the results. Thus, it is extremely important to prioritize or further automatically refine the results produced by data mining techniques. My previous study proposed new evaluation methods to prioritize the results produced by a microarray clustering method. I plan to apply text mining techniques to automatically verify and refine the data mining results, which can help biologists better understand the mining results.

My research directions are related to many other areas of computer science and engineering, such as natural language processing, machine learning, human-computer interaction, data mining, databases, and bioinformatics. I have worked with, and will continue collaborating with researchers from multiple disciplines.