Research Statement

Jie Lu

Language Technologies Institute, School of Computer Science, Carnegie Mellon University

jielu@cs.cmu.edu http://www.cs.cmu.edu/~jielu

My general research interests lie in the use and development of information retrieval and machine learning techniques for effective and efficient adaptive information access and mining. I am particularly interested in content management and search in distributed environments such as enterprise networks, peer-to-peer file-sharing networks, and social networks. My other areas of interest include personalized search and domain-dependent (e.g., legal, medical) or task-specific (e.g, retail, travel) information processing and knowledge discovery.

Nowadays, most people find themselves engaging in information seeking activities in two different scenarios: i) content relevant to an information request is often located in distributed environments where new contents are constantly created and existing contents are frequently updated, and ii) multiple resources, evidence and criteria are required to provide high-quality results demanded by a specific and often well-defined information need. The former is viewed as a new territory for horizontal search (general-purpose search) which has so far primarily relied on information processing in centralized repositories; the latter has become the focus of vertical search (specialized search) which has gained increasing popularity over recent years. I am interested in both directions.

During the pursuit of my Ph.D. degree at Carnegie Mellon University, I have not only obtained a broad background in information retrieval, but have also accumulated extensive experience in several areas of this field. My dissertation research has focused on designing practical solutions for search in distributed environments [CIKM 2002, 2003; dg.o 2003; SIGIR 2004, 2006; ECIR 2005; JIR 2006], which can be applied to large-scale horizontal search in various applications. My work on user modeling [SIGIR 2006], genomic information retrieval [TREC 2006] and duplicate detection in large public comment datasets for rulemaking has provided me opportunities to explore new methods to use prior or domain knowledge and combine information from multiple sources for the problems of vertical search.

In addition to my own research, I have also worked with a Ph.D. student from the University of Duisburg-Essen in Germany on developing a peer-to-peer prototype for federated search of text-based digital libraries, and supervised a Master's student and an undergraduate student at Carnegie Mellon University on projects related to the design and implementation of software components for retrieval in distributed environments.

More details with regard to the research I have done are provided below, followed by an outline of my further research plans.

Current Research

My recent research work covers three distinct areas of information retrieval: federated search in peer-to-peer networks (dissertation research), genomic information retrieval, and automatic exact and near duplicate detection.

Federated Search

The centralized approach for search is popular in the World Wide Web due to its economic and administrative advantages. However, it is not appropriate for environments where copying contents to centralized repositories is not allowed (e.g., in the "Hidden Web") or not practical (e.g., in enterprise networks that lack the support of a central IT infrastructure). Federated search uses a single interface to support finding items that are scattered among a distributed set of information sources or services, which provides an effective, convenient and cost-efficient search solution for these environments. Peer-to-peer (P2P) networks become a potentially robust and scalable model for federated search over large numbers of distributed collections by integrating autonomous computing resources without requiring a central authority. However, prior work on P2P networks has mostly focused on search over document names, identifiers, or keywords from a small or controlled vocabulary. Little has been done on providing solutions to full-text federated search with relevance-based document ranking.

My dissertation research fills the gap by developing an integrated framework of network overlay, network evolution, and search models for full-text ranked retrieval in P2P networks. Multiple directory services maintain full-text representations of resources located in their network neighborhoods, and provide local resource selection and result merging services. The *network overlay* model defines a network structure that extends previous peer functionalities and integrates search-enhancing properties of interest-based locality, content-based locality, and small-world to explicitly support full-text federated search. The *network evolution* model provides autonomous and adaptive topology evolution algorithms to construct a network structure with desired content distribution, navigability and load balancing without a centralized control or semantic annotations [RIAO 2007, submitted]. The *network search* model addresses the problems of resource representation, resource selection, and result merging based on the unique characteristics of P2P networks, and balances between effectiveness and cost. For *resource representation*, contents in network neighborhoods are represented by full-text resource descriptions exponentially decayed with increasing distance [ECIR 2005; JIR 2006], and several pruning techniques are explored to reduce the sizes of full-text resource descriptions [CIKM 2002, 2003; dg.o 2003]. For *resource selection*, hubs select their hub neighbors based on neighborhood descriptions for effective query routing [ECIR 2005; JIR 2006], and consumers automatically construct user interest models based on search history to improve resource selection performance for future queries similar to past ones [SIGIR 2006]. For *result merging*, existing methods are extended to work effectively in P2P networks without global corpus statistics [SIGIR 2004]. The framework is a comprehensive and practical solution to full-text ranked retrieval in large-scale, distributed and dynamic environments with heterogeneous, open-domain contents.

The models developed as integrated parts of the framework for full-text federated search in P2P networks can also be used in other applications. One application for the network overlay and evolution models is large-scale centralized search that uses multiple connected computing and storage resources (the server farm) to provide the services of a central authority. The network structure defined in the network overlay model can be used to organize these resources to provide regulated content distribution for more efficient query processing. The algorithms developed for the network evolution model can help to dynamically manage the resources in the face of constant changes in contents, requests and workload. The network evolution model and search models can also be applied to online social networks to automatically organize groups and communities and conduct search in this highly dynamic environment. The approach proposed in my dissertation for user modeling may benefit meta-search and personalized search applications because of its ability to distinguish long-term and short-term information needs and to apply different search strategies accordingly to optimize the overall search performance. The techniques of pruning resource selection indexes are useful for developing efficient solutions to other retrieval problems.

In addition to these models, I have also developed two P2P testbeds to provide the P2P research community common evaluation platforms for large-scale, full-text search in P2P networks.[1] The testbeds include 2,500 and 25,000 full-text digital libraries respectively, together with over one million queries. They have been used by other researchers to evaluate existing and new approaches to federated search [Renda and Callan, CIKM 2004; Klampanos et al., ECIR 2005; Castiglion and Melucci, ECIR 2007, submitted].

Genomic Information Retrieval

My work in genomic information retrieval focuses on combining multiple resources, evidence, and criteria to incorporate domain knowledge for query expansion and result ranking [TREC 2006]. The query expansion module improves existing techniques by using several term-weighting schemes to group and combine terms from different sources based on their characteristics, which proves to be more effective than the typical approach of treating expansion terms equally. For result ranking, different scoring criteria are used to evaluate evidence from document, passage, and term-matching granularities, which are further combined to produce a final ranking. Evaluation results show that result ranking by combining multiple scoring criteria and evidence consistently provides better performance compared with result ranking based on a single criterion. Our system was ranked 3rd (out of 30 participants) in its performance on passage retrieval for the Genomics Track of TREC 2006 [PCPsg* runs, TREC 2006 Genomics Track Overview]. Compared with other approaches, the main achievement of my work is its effective combination of information from various resources and aspects in multiple stages of retrieval for a domain-dependent application.

Automatic Duplicate Detection

The task of duplicate detection in large public comment datasets is to detect exact-duplicate and near-duplicate documents in comments made by the public about proposed federal regulations. Exact-duplicate and near-duplicate comments are typically created by copying and editing form letters provided by organized interest groups and lobbies. To utilize the domain knowledge about the creation process of duplicate documents, a new *fuzzy match* edit operation is introduced in my work to match sentences with minor word differences. The degree of fuzzy match between sentences is measured using traditional information retrieval techniques. A modified edit distance method is proposed to compare documents at the sentence granularity based on the edit operations of *substitution*, *insertion*, *deletion*, and *fuzzy match*. By combining the complementary strengths of a similarity-based approach commonly used in IR (flexibility and efficiency) and a string-based approach which measures the effort required to transform one document into another (accuracy), more effective and robust performance can be achieved for detecting near duplicates. This work is significant because it proposes a new method to better capture the characteristics of the data by integrating different approaches; it is also more effective than several alternative approaches.

Future Research

For future research, I am open to all problems related to search and text mining. I am particularly interested in working on content management and search in distributed environments, personalized search, and domain-dependent or task-specific information access. A few of my specific interests are described below.

[1] http://www.cs.cmu.edu/~callan/Data

Search in Distributed Environments

Content-based federated search in a broad range of distributed environments, such as social networks and corporate environments, remains an open problem. I am interested in adapting the solutions developed in my dissertation to more domains, applications, and types of information needs.

One way to adapt the proposed framework to social networks and other distributed environments with explicitly or implicitly formed communities is to allow users within a community to share their search experiences about the usefulness of different information sources with regard to their interests (relevance feedback). Currently, the defined network overlay doesn't require users' collaborative efforts for federated search. However, some lightweight information sharing among users with similar interests may greatly improve the chance of locating relevant contents quickly without significantly increasing cost. I plan to extend the current framework to incorporate collaborative search and filtering by automatically organizing users into interest-based groups and allowing user models to be shared within each group. The approach I developed that uses pseudo relevance feedback to learn user interest models can be adapted to work together with the methods that incorporate implicit relevance feedback in environments where it is hard to get explicit feedback from users.

To bring federated search to the next level, important content-independent features should be incorporated into the developed framework. In my dissertation research, the performance of federated search is mostly optimized in terms of search accuracy measured in precision/recall and efficiency measured in the percentage of information sources reached by query messages. Since the framework is sufficiently flexible to allow more factors to be considered in measuring search performance, I will explore a utility-based approach to substitute for the similarity-based approach used in search and topology evolution. The utility function can combine multiple factors such as accuracy, efficiency, authority, reliability, latency, monetary cost, etc., which will certainly make full-text federated search more practical and useful in a wider range of distributed environments.

Personalized Search

Personalized search has become an active research area. It constructs user models from search history to adapt search to the future information needs of individuals. The weakness of a pure history-based approach is that the search performance of new information requests that are not related to previous search may be unreliable. The adaptive user modeling approach I developed may remedy this problem because it has the ability to automatically distinguish different types of information needs and select the appropriate search strategy accordingly. Therefore, I am interested in applying it to personalized search and studying its effectiveness. I believe that personalized search can benefit from dynamic search strategy selection based on different types of information needs.

Domain-Dependent and Task-Specific Information Access

Specialized search provides high-quality results for domain-dependent and task-specific information access, and greatly complements general-purpose search. What intrigues me most in specialized search is its potential to incorporate knowledge about domains/tasks to better capture the characteristics of the data and users, which can lead to considerably improved performance. I am interested in working towards general frameworks to incorporate and integrate information from multiple sources such as contents, prior knowledge, and external resources. The frameworks will include more sophisticated techniques than simple forms of combination for information integration, especially for the cases where information is represented in a wide variety of forms, or implicit dependencies exist between different pieces of information. One important lesson I have learned from my previous work is that understanding the characteristics of the domain, task, and data first and developing techniques accordingly is far more important and effective than mechanically applying theoretically sound models without detailed data analysis. I will continue to use this general approach in designing the most appropriate methods for domain-dependent and task-specific solutions.

Methods developed for domain-dependent tasks often use specialized knowledge resources. In some cases these resources can be created by text-mining of large or well-organized corpora. I am mostly interested in mining entity relations that are embedded in unstructured text contents. I plan to conduct research on learning and extracting typical relational patterns between entities in specific domains (e.g., genes, diseases, symptoms, and medicines in biomedical domain) for active knowledge discovery, focusing on semi-supervised or unsupervised methods that require few training data. Furthermore, I will work on adapting and improving the techniques developed for simple entities to more complicated ones with multiple attributes or facets, which I believe will benefit many domain-dependent applications. Again, careful data analysis and attention to details goes a long way toward building the best solutions to particular problems.