

Research Statement

My main research interests are in information retrieval, the subject of using computational modeling and techniques to facilitate users in their search, information processing and decision making process. I drive my research with a deep understanding of every aspect of the search process, drawing insights from retrieval theory, as well as areas such as user behavior and natural language understanding. I verify the insights with data analyses, sometimes at a large scale, and carry them out as statistical inference or structured retrieval models. Guiding the analysis and usage of data with a deep understanding also creates new problems for information retrieval and related fields to solve.

I try to do impactful research. My thesis research is the first to quantitatively study the *vocabulary mismatch* problem in retrieval, which leads to effective ways of predicting whether a query term is likely to mismatch relevant documents, and a number of theoretically motivated interventions that significantly improve retrieval using the mismatch predictions. Another topic of interest is *structured retrieval* enabled by advanced query languages and diverse document structure. It is my belief that any effective retrieval technique will find its root in the retrieval model. The advanced structured retrieval models provide a versatile and solid basis for various search technologies and applications to build on. My research applies structured retrieval in several human language technology applications. For example, in *question answering*, effective structured queries are formulated based on the semantic structure of the question and answer sentences (Zhao and Callan 2009a). In *intelligent tutoring*, structured retrieval enables the tutoring system to efficiently identify reading material that matches students' interests and reading levels, significantly reducing teachers' efforts and promoting students' interest in learning a second language (Heilman et al 2008). A growing number of applications will find structured retrieval a necessary tool.

Making the search needs of these applications explicit as structured queries has also brought back new insights to the most basic yet important application – the *ad hoc retrieval* task. It is the new insights from structured retrieval along with a deep understanding of retrieval models that led to my new approach to the term mismatch problem, opening up new directions to solve mismatch and improve retrieval.

Modeling Term Mismatch

Term or vocabulary mismatch is the problem where query terms fail to appear in relevant documents. It has been a problem since the beginning of library science. Techniques such as the classification system of library books (to some extent), pseudo relevance feedback and latent semantic indexing all aimed to, but struggled in solving vocabulary mismatch. Given a topic and its relevant document set, my thesis research models the likelihood of term mismatch as the proportion of relevant documents that do not contain the term t , i.e. $P(\bar{t}|R)$. This has enabled the first large scale quantitative analysis of term mismatch in retrieval, leading to effective ways of predicting $P(\bar{t}|R)$. My research also points out the role term mismatch plays in probabilistic retrieval models: The complement of $P(\bar{t}|R)$, $P(t|R)$ which measures term recall, is one of the two class-conditional probabilities that determine the Bayesian optimal term weight for term t – the Robertson Spärk-Jones weight. The other class-conditional probability is that of a term appearing in non-relevant documents, which has led to the traditional idf term weighting. The central role term mismatch plays in retrieval models suggests several theoretically motivated methods to use $P(t|R)$ predictions to improve retrieval.

Mismatch modeling, formulated as $P(t|R)$ prediction, is an important problem in basic retrieval models that lacked good solutions.

My research identified factors that may cause mismatch and framed $P(t|R)$ prediction as a standard machine learning problem using features that model some of the most important contributing factors, such as replaceability (synonyms of the query term appearing in relevant documents instead of the original term) and abstractness (abstract query terms replaced by more specific terms in relevant documents). Given the features and query terms from training topics with known relevance, a statistical regression model is learnt and used to predict $P(t|R)$ for test topics. Predicted $P(t|R)$ term weights consistently improve retrieval by 15-25%, with high statistical significance (Zhao and Callan 2010).

My research designed effective query dependent features, but these features require feedback retrieval and LSI computation, which made the method unsuitable for real time response scenarios. To understand the query dependent nature of $P(t|R)$ and to speed up its prediction, my thesis research studied the causes of the cross-topic variation of $P(t|R)$ for the same term (Zhao and Callan 2012a). It showed that query-dependent features are needed for the successful prediction of $P(t|R)$, even when the occurrences of the term in different topics share the same word sense. At the same time, we observed that for many term occurrences, the cross-topic variation of $P(t|R)$ is relatively low. Thus efficient $P(t|R)$ prediction methods were designed using historic occurrences of the same term in training queries with known relevance to predict $P(t|R)$ in a test topic. The query term coverage of this history-based prediction method depends on the set of query terms used for training. To improve coverage, a semi-supervised learning method was designed, which uses a small training set with relevance judgements to bootstrap a large number of queries (from e.g. a query log) with no relevance information.

My thesis research also investigated applying $P(t|R)$ prediction as a fine grain diagnostic tool to guide query expansion to focus on terms that need expansion most (Zhao and Callan 2012b). Such a term-level diagnostic tool has never before been available to the retrieval system. This research confirmed that terms with low $P(t|R)$ are the ones that need expansion. Using predicted $P(t|R)$ to guide expansion to the 2 terms with the lowest predicted $P(t|R)$ is enough to achieve close-to-optimum performance. The optimal performance happens when all query terms are carefully expanded.

One important discovery in using query expansion to solve vocabulary mismatch is that Boolean conjunctive normal form (CNF) structured queries, which are widely used by practitioners, but not as much in the research community, are more effective than the dominant bag of word expansion with the same set of *high quality* expansion terms.

In summary, my thesis research focuses on the probabilistic modeling of term mismatch and its effect on retrieval. It is approached with data analyses on hundreds of topics and with constant clarifications of our knowledge about the mismatch phenomena. Effective features have been discovered during the process and used in a supervised prediction model to predict $P(t|R)$. As suggested by theory, using predicted $P(t|R)$ as term weights significantly improves retrieval accuracy. Analyses on the cross-topic variation of $P(t|R)$ led to more efficient prediction methods. Besides term weighting, focusing expansion on the query terms with the lowest $P(t|R)$ is another intervention suggested by the theory. The resulting CNF structured queries were found to outperform the dominant bag of word expansion.

Structured Retrieval

Structured queries use query operators such as phrase, window, AND, OR, AND-NOT, field-term or field-field containment that restrict matching and/or combine evidence. Human language technology (HLT) applications are beginning to be built on top of structured retrieval engines, so that different querying strategies can be easily tested, and effective strategies that suit the application can be discovered. However, structured retrieval is rarely found to outperform its keyword counterpart, and is usually slower. The main difficulty is to find the right query structure for each application.

In the first few years at CMU, I worked on structured retrieval extensively. I formulated effective *structured queries* for different applications (Zhu et al, 2007; Zhao and Callan 2009a, 2009b, 2010), built *structured retrieval models* and *evaluated* them on structured documents (Zhao and Callan 2008). I worked on statistical *smoothing* to solve term level mismatch in short fields, *translation models* to address field level mismatch between the semantic role structure of the question and that of the answer sentences (Zhao and Callan 2009a), and on the *efficient evaluation* of structured queries.

Ad hoc retrieval is probably one of the most difficult problems for structured retrieval to show an improvement over the well tuned traditional bag of word models. Yet my thesis research finds that the use of term weighting and Boolean CNF style queries surpasses existing keyword baselines. For many applications such as question answering, patent retrieval or intelligent tutoring, the most effective strategy has been a problem-by-problem approach, which first understands the needs of the applications in a top-down manner, and then informs the decision of choosing the right kind of queries.

Corpus Development & Going Large Scale

Throughout my research, I have been involved in the development of several different data collections for different research projects.

For computer-assisted language learning, reading material is needed which has to match the keywords targeted for learning and the preferences of the individual students (e.g. topical interests and reading levels). To allow for personalization and to filter out poor quality documents, tens of millions of Web pages need to be crawled, processed and filtered. I maintained the whole system, but focused on the active querying of a search engine to only crawl documents that will likely match the tutoring needs.

The ClueWeb09 project aimed to create a domain-general collection of 1 billion high quality Web pages that covers the top 10 most used languages on the Web, and simulates the first tier of a commercial Web search engine. The crawl was done in a high-quality page first crawling order. I worked on seeding the crawl with high PageRank pages from an existing large corpus, OPIC approximation for PageRank scores, language identification, language distribution control and final PageRank computation.

To create an accurate and comprehensive knowledge base from the Web, as part of the Read-The-Web project, very narrowly focused batches of thousands of Web pages are needed as evidence to verify hypotheses derived from the knowledge base. I identified areas of the knowledge base that need improved coverage most, and used targeted querying and crawling to discover pages that would likely improve these areas.

In many of these efforts I used Condor or Torque to schedule jobs on computer clusters, and parallel computing frameworks like MapReduce (Hadoop) or Scope (from Microsoft, which has a SQL like script language). I find parallel computing useful for data processing and retrieval evaluation at a large scale.

Supporting the Research Community

For the benefit of both my own research and the academic environment that makes it possible, part of my research effort has been to contribute back to the research community. I have contributed structured retrieval model and efficiency improvements back to the Lemur Project's Indri search engine, which is an open source search engine widely used by the research community. Every month, I also spend some time answering questions from Lemur Project's user community. The ClueWeb09 dataset is another such effort. The resulting ClueWeb09 dataset has been licensed by over 200 research groups worldwide and used in over 6 TREC evaluation tasks as well as NTCIR evaluations.

Future Directions

I believe all effective retrieval techniques are either motivated by the retrieval model or could and should be built into the retrieval model. My planned research will in one direction explore the current retrieval models, and in another direction expand the realm of search by extending its use to even wider scenarios.

Ad hoc retrieval and Conjunctive normal form (CNF) expansion queries

Ad hoc retrieval is a very basic form of the retrieval problem where the system is only given a user query and a collection of documents to rank. My thesis research showed that term mismatch is a significant problem in ad hoc retrieval, and also that in both theory and practice, a general and effective solution to the mismatch problem is to use Boolean CNF structured queries to expand each query term with its synonyms. However, only expert searchers like lawyers or librarians are good at creating effective CNF expansion queries, and it still takes lots of iterations and effort to arrive at the right query. My research will aim to make CNF queries easier to use for the search experts as well as ordinary users through 1) robust retrieval modeling, 2) automatic identification of high quality synonyms or sources to discover high quality synonyms, and 3) novel interfaces to facilitate user interaction.

Currently, CNF queries treat all synonyms the same, so that whenever a false synonym is included, false positives can populate the top ranks and damage retrieval. Proper synonym weighting can be used to modulate the synonyms, so that the adverse effect of false synonyms will be kept at a minimum, even when they are included by the user or an automatic algorithm. Current automatic algorithms suggest candidate synonyms from sources such as thesauri or pseudo relevance feedback. The kind of data analyses based on relevance information used in my thesis suggests new ways of identifying high quality synonyms. One such method is to automatically identify high quality sources where good synonyms may come from, so as to facilitate users in their synonym identification effort. Sometimes, even when domain experts think highly of an expansion term, if no one manually checks the result lists, the expanded query can still return many false positives. Because manual examination of rank lists is a daunting task, novel user interfaces that compare rank lists and reveal the exact effects of an expansion term on the rank list will be important for effective user interaction, especially in the manual formulation of CNF queries.

Structured retrieval modeling

Structured retrieval modeling is an important extension of the current ad hoc retrieval models. Current structured retrieval systems (e.g. Indri) offer minimum support to use information from related fields

outside the scoring context, yet many search applications need this ability. For example, in sentence retrieval, co-references may require named-entity fields of surrounding sentences to be accessed to retrieve sentences that only mention "the president" instead of "Barack Obama". In Web search, inlink anchor texts from other pages are found useful for the retrieval of the current page. Similarly, on social network sites like Facebook or LinkedIn, users may care about what others say about a particular person.

Smoothing in statistical language models is one method which uses information from outside the field that is being scored, which would merge word distributions from other sentences or sentences from other documents into the language model of the current sentence. However, such specialized solutions do not provide the flexibility needed for the many different use cases. It is probably best to allow users to use the query language to access related fields or documents, so that users or system developers can utilize these fields in a way that makes the most sense to them. For example, a reference-search query on LinkedIn may look for people who can connect the searcher to someone else with a certain skill, whereas a Facebook search may look for a ranked list of people whose friends say they are easy going. My research will aim to enhance the capability and efficiency of the structured retrieval engines, as well as develop tools for users to effectively create structured queries.

Extending search – computer facilitated human problem solving

Zooming out of the retrieval model world, search itself is still largely in the growth stage. On the one hand, ordinary users are given access to the Web and Web search engines 24x7, especially with the recent expansion in the smart phone market. On the other hand, people accomplish tasks and make decisions constantly. When the user is in the middle of a series of actions or decisions, or needs to make a quick decision or a seemingly small decision unaware of its significance, or has a limited user interface like a mobile phone, she may proceed with gut feelings, skipping a search even when it can help. Thus, to better facilitate the human user, and make search easier to access, future systems will be able to understand the user, to *predict when* the user may have a search need, to infer the *content of the need*, and *initiate a search*, without the user typing or speaking into the search box. This proactive search capability will smooth the process of initiating a search from a user need, and extend the application of search and artificial intelligence to serve much wider areas of human needs.

Currently, personal assistants that run on mobile devices, such as Siri or Vlingo, smooth the search process by using contextual information of the user to personalize search. However, they still rely on the user to tell them what to do. Recommendation systems do actively push results to the user. However, existing systems tend to overload the user with irrelevant advertisements and recommendations, regardless of the user's needs. This is typically incentivized by possible purchases from the user, without realizing that the user may learn to ignore all recommendations if they are often irrelevant, at which point, it becomes suboptimal for all parties. Given the current situation, academia is probably the best place to carry out the research of understanding and predicting users' potential search needs, and industry will likely follow suit if success is demonstrated.