

Research Statement

Jinyoung Kim

January 2, 2011

My research interests include all areas of information retrieval and human-computer interaction. As my thesis work, I am focusing on challenges in providing an effective search system for personal information management (PIM).

Having finished my seventh semester in the MS/Ph.D program in UMass Computer Science department, I am starting to establish myself as an Information Retrieval researcher, making several contributions in the area of my research. In the remainder of this statement I will describe my research experience and interested projects for internship in detail.

Semi-structured Document Retrieval Advised by Professor W. Bruce Croft, I worked on the problem of XML document retrieval. As a part of this work, we published the paper ‘A Probabilistic Retrieval Model for Semi-structured Data’ [6] in ECIR’09. This work was motivated by the observation that users’ queries have implicit structure that can be exploited for retrieval, we suggested a novel probabilistic retrieval model for XML document collection (PRM-S) that automatically finds the structure of a user’s keyword query.

For instance, given a query ‘meg ryan romance’ for the IMDB collection, we first infer that the user may have meant *actor* by ‘meg ryan’ and *genre* by ‘romance’. With this mapping of query-terms and document fields, the initial keyword query is transformed into a structured query, resulting in a dramatic performance gain over competitive baselines.

Retrieval Model for Desktop Search After this initial work, we worked on the problem of searching personal document collections, widely known as desktop search. Since the most conspicuous problem in studying such collections is privacy issue, we developed a novel technique by which one can generate and validate reusable test collections for desktop search. Our method of generating pseudo-desktop collection as well as the results of retrieval experiments was presented in SIGIR’09 Workshop [3] and CIKM’09 [4].

We followed up this research by devising and evaluating a suitable retrieval model for desktop search. This work was published in SIGIR’10 [5]. In this retrieval model, we suggested and evaluated a retrieval model for desktop search. In the retrieval model, type-specific results (rank list and type score) are merged into a final rank. We focused on the type prediction part, suggesting a novel type prediction method that exploits the structure of documents, and experimented with a machine learning approach where we train a multi-class classifier that will score each collection against a given query.

Since we needed a large number of queries and relevance judgments to train such a model, we performed a user study by developing a human-computation game called

DocTrack that helped us to gather a sufficient amount of user data in a realistic setting. Using the queries collected with the game, we showed that the type prediction method we suggested improve performance significantly, which was published in SIGIR'10 [5].

Associative Browsing of Personal Information As an alternative means of accessing personal information, we are currently investigating associative browsing, which is intuitive for the user and complementary to other methods of personal information access, such as keyword search. In [1], we proposed an associative browsing model of personal information in which users can navigate through the space of documents and concepts (e.g., person names, events, etc.).

Instability of Web Search Results I worked at Microsoft Bing Search during the summer of 2010 with research scientist Vitor Carvalho. We investigated the instability in web search results, focusing on undesirable changes in top results of major search engines over time. This instability is mostly due to the dynamic nature of the web and the complex architecture of commercial search engine,

In the project, we aimed to understand the degree and nature of these changes, and introduce a technique by which we can control undesirable instability. We found that all major search engines suffer from instability issues. We then devised a technique for reducing up to 30% of the instability while still significantly improving overall NDCG for the affected queries. The first part of this work will be presented at ECIR'11[2] and the second part is currently under submission to the World Wide Web conference (WWW'11).

Research Projects of Interest Through this proposed internship, while continuing my efforts in information retrieval research, I aim to focus my attention to areas which involves a more close interaction with the user. This include personalized search, exploratory search, interactive IR and so on, which are typically referred to as Human-Computer Information Retrieval (HCIR). Depending on the project, I am also open to other problems of information retrieval and human-computer interaction.

Given the increasing complexity of information needs and the lack of such support in traditional search engines, I believe there are many significant research problems in that area. Also, since many of tasks in HCIR involves modeling rich user profile and complex interaction, many of research issues in HCIR are highly relevant to my thesis work.

Thank you for considering my application.

References

- [1] Jinyoung Kim, Anton Bakalov, David A. Smith, and W. Bruce Croft. Building a semantic representation for personal information. *Proceedings of CIKM '10, 18th ACM International Conference on Information and Knowledge Management*, pages 1741–1744, 2009.
- [2] Jinyoung Kim and Vitor R. Carvalho. An analysis of time-instability in web search results. In *Proceedings of ECIR'11: 33rd European Conference on Information Retrieval*. ACM, 2011.
- [3] Jinyoung Kim and W. Bruce Croft. Building pseudo-desktop collections. In *SIGIR Workshop on the Future of IR Evaluation*. ACM, 2009. workshop.

- [4] Jinyoung Kim and W. Bruce Croft. Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM '09: 18th ACM International Conference on Information and Knowledge Management*, pages 1297–1306. ACM, 2009.
- [5] Jinyoung Kim and W. Bruce Croft. Ranking using multiple document types in desktop search. In *Proceedings of SIGIR '10: 33rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 2010.
- [6] Jinyoung Kim, Xiaobing Xue, and W. Bruce Croft. A probabilistic retrieval model for semistructured data. In *Proceedings of ECIR '09: 31st European Conference on Information Retrieval*, pages 228–239. Springer, 2009.