

Ambiguity and Context-Aware Query Reformulation

Hui Zhang

School of Library and Information Science, Indiana University

Abstract

An effective information retrieval (IR) interface should work like a consultant to assist users in fulfilling their information needs despite any lack of knowledge. Some progress has been made towards this goal (e.g., query expansion, relevance feedback); however, the problem of query ambiguity is still poorly addressed in spite of its importance to human-computer interaction and IR.

There are two major forms of ambiguity in natural language: semantic and syntactic. The source of semantic ambiguity comes from the meaning of words whereas the source of syntactic ambiguity comes from the construction of the sentence. Because most of the queries submitted to an IR systems are short [9], semantic ambiguity is prevalent in IR. In addition to natural language, another major source of ambiguity comes from the user's intention. For example, a query such as *DNA* could indicate user information needs on any of the following topics: health care, law enforcement, or biology.

Most of the research on query disambiguation relies on existing knowledge resources such as dictionaries and WordNet for definitions of word meanings (e.g., [4, 6, 10]). The limitation of these approaches is that the information contained in the knowledge resources may be inappropriate or outdated for the tasks. Schutze [7] proposed an unsupervised approach to induce word sense from term clustering. However, the computing cost of his approach is high, which makes it unfeasible for analyzing large collections even with modern hardware.

Providing support for end users through query editing and reformulation is one of the core functions of an IR interface. The fact is, however, that a large number of initial queries are unspecific and incomplete with respect to the user's information need because he is in anomalous state of knowledge (*ASK*) [2]. A recent study estimated that 16% of the queries submitted to a web search engine were ambiguous [8]. In addition to the number of ambiguous queries, IR systems are ineffective in handling these ambiguous queries effectively. As a result, many of the query suggestions made by an IR system are misleading because of the ambiguity of the query. To overcome the limitations of previous methods, I propose a set of recommendations based on the following three tasks:

1. Establish word meanings by harvesting and clustering query contexts from user sessions in a query log;
2. Disambiguate user's intentions and polysemous query words based on the contexts of both the query and the document;
3. Assist diversity and exploratory search with context-aware query reformulation.

Task 1 is a preparatory task that can be done off-line. However, Tasks 2 and 3 are at the heart of a search interface and have significant impact on retrieval performance and user satisfaction. In this presentation, I will discuss issues, methods, and preliminary results for each of the tasks.

Reference:

- [1] Agirre, E. and Edmonds, P. Word Sense Disambiguation: Algorithms and Applications. Springer, 2007.
- [2] Belkin, N. J., Oddy, R. N. and Brooks, H. M. ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38, 2 (1982), 61-71.
- [3] Guo, J., Xu, G., Li, H. and Cheng, X. A unified and discriminative model for query refinement. In Proceedings of the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (Singapore, 2008). ACM.
- [4] Liu, S., Yu, C. and Meng, W. Word sense disambiguation in queries. Proceedings of the 14th ACM international conference on Information and knowledge management (2005), 525-532.
- [5] Pass, G., Chowdhury, A. and Torgeson, C. A picture of search. ACM New York, NY, USA, City, 2006.
- [6] Sanderson, M. Word sense disambiguation and information retrieval. Springer-Verlag New York, Inc. New York, NY, USA, 1994.
- [7] Schutze, H. and Pedersen, J. Information retrieval based on word senses. Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (1995), 161-175.
- [8] Song, R., Luo, Z., Wen, J.-R., Yu, Y. and Hon, H.-W. Identifying ambiguous queries in web search. In Proceedings of the Proceedings of the 16th international conference on World Wide Web (Banff, Alberta, Canada, 2007). ACM.
- [9] Spink, A., Wolfram, D., Jansen, M. B. J. and Saracevic, T. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52, 3 (2001), 226-234.
- [10] Voorhees, E. M. Using WordNet to disambiguate word senses for text retrieval. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (1993), 171-180.
- [11] Wang, X. and Zhai, C. Mining term association patterns from search logs for effective query reformulation. In Proceedings of the Proceeding of the 17th ACM conference on Information and knowledge management (Napa Valley, California, USA, 2008). ACM.
- [12] Xu, J. and Croft, W. B. Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (1996), 4-11.