

Student Research Symposium 2004

Luo Si

Federated Search in Uncooperative Environments

Conventional search engines such as Google or AltaVista are effective when an information source allows its contents to be crawled and indexed in a centralized database. However, a large amount of information cannot be crawled and searched by conventional search engines either due to intellectual property protection or frequent information update. This type of information is valuable. For example, hidden Web contents that can not be searched by conventional search engines have been estimated to be 2-50 times larger than the visible Web and are often created and maintained by professionals.

Federated search provides the solution of the search problem for the information that cannot be searched by conventional search engines. It includes three sub-problems: i) acquiring information about the contents of each information source (resource representation), ii) ranking the sources and selecting a small number of them for a given query (resource ranking), and iii) merging the results returned from the selected sources into a single ranked list (result-merging).

This work addresses federated search problems in uncooperative environments such as the Web where information sources can not be assumed to share their contents or use the same type of search engine. Empirically effective solutions have been proposed to the full range of federated search sub-problems such as new algorithms for information source estimation, resource selection and results merging.

Furthermore, a unified utility maximization framework is proposed to combine the separate solutions together to construct effective systems of different federated search applications. This is the first probabilistic framework for integrating the different components of a federated search system. The more unified view of federated search task provides a new opportunity to utilize available information. It enables us to configure individual components globally to get desired overall results of different applications, which is superior to the simple choice of combining individual effective solutions together in previous research.

This work advances the state-of-the-art of federated search. The more theoretical foundation, the better empirically results and the better modeling of real world

applications make the new research a bridge to turn federated search from a cool research topic to a much more practical tool.

Related references:

Si, L. & Callan, J. (2002a). Using sampled data and regression to merge search engine results. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Si, L. & Callan., J. (2003a). Relevant document distribution estimation method for resource selection. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Si, L. & Callan, J. (2003b). A Semi-Supervised learning method to merge search engine results. ACM Transactions on Information Systems, 21(4).

Si, L. & Callan, J. (2004). The effect of database size distribution on resource selection algorithms. In Distributed Multimedia Information Retrieval, LNCS 2924, Springer.

Si, L. & Callan, J. (2004). Unified Utility Maximization for Distributed Information Retrieval in Uncooperative Environments. In Proceedings of the 13th International Conference on Information and Knowledge Management, ACM.