Enterprise and Desktop Search

Pavel Dmitriev
Yahoo! Labs
Sunnyvale, CA, USA
dmitriev@yahoo-inc.com

Pavel Serdyukov
Delft University of Technology
Delft, The Netherlands
p.serdyukov@tudelft.nl

Sergey Chernov
L3S Research Center
Hannover, Germany
chernov@L3S.de

ABSTRACT

With the growing amount of information on users' desktops and increasing scale and complexity of intranets, Enterprise and Desktop Search are becoming two increasingly important Information Retrieval applications. While the challenges arising there are not completely different from those that the web community has faced for years, advanced web search solutions are often unable to address them properly. In this tutorial we give a research prospective on distinctive features of both Enterprise and Desktop Search, explain typical search scenarios, and review existing ranking techniques and algorithms.

Categories and Subject Descriptors: H. [Information Systems]: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval.

General Terms:
Algorithms, Measurement, Performance, Experimentation.

Keywords:
Enterprise search, desktop search, user feedback, exploratory search, expert search

1. ENTERPRIZE SEARCH

In [14] it is estimated that Enterprise Search industry is growing at 20% per year, and is expected to reach 2.55B by 2010. Despite its increasing practical importance, until now Enterprise Search has received relatively little attention from the web research community. Distinctive features of Enterprise Search stem from the fact that the social forces driving creation of content on the Web are different from those in the Enterprise. For example, analysis of the structure of the Enterprise Web [15] indicates that it is quite different from the well-known bow-tie structure of the public Web and PageRank is not as effective in the Enterprise Search setting. This and other differences result in unique challenges for crawling, indexing, and ranking components of a search engine. While the difficulty of Enterprise Search is a well acknowledged fact, there are relatively few research papers that attempt to study specific features that work and do not work in this area. In this part of the tutorial we highlight differences between Web and Enterprise Search and summarize existing approaches to deal with the Enterprise-specific search problems.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

1.1 User Feedback

While there are many problems that make Enterprise Search more difficult than Web Search, there are also some advantages. One such advantage is that there is typically no spam in the Enterprise. In fact, users are often inclined to cooperate with the search engine by providing their feedback and are not so concerned about privacy. Due to more manageable scale of the data, it is also much easier to utilize such feedback in the Enterprise setting than in the Web setting. This part of the tutorial is devoted to search algorithms which utilize explicit and implicit user annotations [10, 6], user browsing traces [4], query session analysis and eye tracking [5].

1.2 Exploratory Search

Since the quality of Enterprise Search is rather limited [22], exploratory and interactive user interfaces are necessary to acquire more feedback from users and eventually satisfy their information needs. While stressing on the need to summarize the search results to support exploratory search, we demonstrate the most representative approaches to visualize result summaries [16] and review techniques to stimulate interaction with the user. When explaining the utility of structured metadata for facets discovery, we put a special focus on the need to avoid information overload and present various methods for facet ranking and selection [24, 3]. Further, we describe methods to transform unstructured metadata (tags and tag clouds) into facet/value hierarchies using information extraction techniques and knowledge bases, such as Wordnet or Wikipedia. We also discuss the ways to structure search results and facilitate faceted browsing for documents with no metadata.

1.3 Expert Search

In the Enterprise people often search not only for relevant documents, but also for their colleagues that know something on the topic of their information need [17]. Expert finding is one of the most rapidly developing sub-domains in the Enterprise search world. We demonstrate typical use cases of expert finding problem and existing applications [21]. We continue with the introduction into the state of the art expert finding techniques: profile-based [2] and document-centric [1] methods. We also show how to take into account the exploratory behavior of the real-world search for experts and utilize the expertise evidence coming from indirectly related documents, social contacts [26] and sources available outside of the Enterprise, such as regional web pages, news and blogs [25].

2. DESKTOP SEARCH

While Enterprise Search spans activities of thousands of users across the intranet, a single user activity mostly occurs on a single desktop. Recent progress in Desktop Search technology allows any knowledge worker to have a personal search engine on top of his/her local documents. In this part of the tutorial we discuss Desktop Search goals and challenges, considering both industrial Desktop Search solutions and research prototypes. We also survey approaches to context detection, their application to ranking and evaluation strategies for Desktop Search.

2.1 Desktop Search in Industry and Academia

In this part of the tutorial we review several commercial systems, such as Yahoo, Copernic, Archivarius, Google, and Windows [20]. Later, we look at the systems that have been developed by researchers in industry and academia to support Personal Information Management. Relevant systems are either centered around tasks [11] or focused on advanced navigation and search functionality [12, 8].

2.2 User Context

The important feature of Desktop Search is that the user context is readily available. Such context can be represented by location, activity, time, season, emotional state, etc. Modern machine learning methods allow to recognize complex user activities and classify them into high-level tasks [23]. They use signals coming from the user's interaction with local files and switches between application windows and web page visits. In this part of the tutorial we show how being informed about the current user activities, a search engine is able to predict which desktop ranking algorithm is the most appropriate for the user at the query time [9]. We present the evaluation of these systems against industrial solutions and discuss open problems.

2.3 Evaluation

Experimental evaluation is a long-standing challenge in Desktop Search. The traditional Cranfield evaluation methodology, as adopted by TREC, cannot be directly applied to Desktop Search. It is highly complicated by privacy concerns for personal data, idiosyncratic user behavior, different levels of computer-literacy, variety of information tasks, and non-repeatability of the experiments. In the last part of the tutorial we discuss ideas for Desktop Search evaluation [18, 13, 7, 19]. We conclude the tutorial with the summary of open questions and promising research directions in Enterprise and Desktop Search.

3. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the Conference on Research and Development in Information Retrieval*, pages 43–50, 2006.

[2] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *Proceedings IJCAI-2007)*, pages 2657–2662, 2007.

[3] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *WSDM '08: Proceedings of the conference on Web Search and Data Mining*, pages 33–44, New York, NY, USA, 2008. ACM.

[4] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *WWW '08: Proceeding of the World Wide Web conference*, pages 51–60, New York, NY, USA, 2008. ACM.

[5] G. Buscher. Attention-based information retrieval. 2007.

[6] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1227–1236, New York, NY, USA, 2009. ACM.

[7] S. Chernov, G. Demartini, E. Herder, M. Kopycki, and W. Nejdl. Evaluating personal information management using an activity logs enriched desktop dataset. In *Proceedings of the 3rd Workshop on Personal Information Management*, 2008.

[8] P.-A. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle++ : Semantically enhanced searching and ranking on the desktop. In *European Semantic Web Conference, Budva, Montenegro, 11-14.06.06*, 2006.

[9] P.-A. Chirita and W. Nejdl. Analyzing user behavior to rank desktop items. In F. Crestani, P. Ferragina, and M. Sanderson, editors, *SPIRE*, volume 4209 of *Lecture Notes in Computer Science*, pages 86–97. Springer, 2006.

[10] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *WWW '06: Proceedings of the World Wide Web conference*, pages 811–817, New York, NY, USA, 2006. ACM.

[11] A. N. Dragunov, T. G. Dietterich, K. Johnsrude, M. McLaughlin, L. Li, and J. L. Herlocker. Tasktracer: a desktop environment to support multi-tasking knowledge workers. In *IUI '05: Proceedings of the conference on Intelligent user interfaces*, pages 75–82, New York, NY, USA, 2005. ACM.

[12] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03: Proceedings of the conference on Research and development in informaion retrieval*, pages 72–79, New York, NY, USA, 2003. ACM.

[13] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *SIGIR '07: Proceedings of the conference on Research and development in information retrieval*, pages 23–30, New York, NY, USA, 2007. ACM Press.

[14] G. Grefenstette. Enterprise search trends and challenges. In *CHORUS Final Conference*, 2009.

[15] D. Hawking. Challenges in enterprise search. In *ADC '04: Proceedings of the Australasian Database Conference*, pages 15–24, Darlinghurst, Australia, Australia, 2004.

[16] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[17] M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. *Inf. Process. Manage.*, 36(5):761–778, 2000.

[18] D. Kelly. Evaluating personal information management behaviors and tools. *Commun. ACM*, 49(1):84–86, 2006.

[19] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *Conference on Information and Knowledge Management, Hong Kong, China, 2-6.11.09*, 2009.

[20] C.-T. Lu, M. Shukla, S. Subramanya, and Y. Wu. Performance evaluation of desktop search engines. In *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on*, pages 110–115, Aug. 2007.

[21] M. T. Maybury. Expert finding systems. Technical Report MTR06B000040, MITRE Corporation, 2006.

[22] R. Mukherjee and J. Mao. Enterprise search: Tough stuff. *QUEUE Magazine*, 2004.

[23] T. Rattenbury and J. Canny. Caad: an automatic task support system. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 687–696, New York, NY, USA, 2007. ACM.

[24] S. B. Roy, H. Wang, G. Das, U. Nambiar, and M. Mohania. Minimum-effort driven dynamic faceted search in structured databases. In *CIKM '08: Proceeding of the Conference on Information and Knowledge Management*, pages 13–22, New York, NY, USA, 2008. ACM.

[25] P. Serdyukov and D. Hiemstra. Being Omnipresent to Be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding. In *FCHER'08: Proceedings of the SIGIR'08 Workshop on Future Challenges in Expertise Retrieval*, 2008.

[26] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM '08: Proceeding of the Conference on Information and Knowledge Management*, pages 1133–1142, New York, NY, USA, 2008.