Enterprise Search in the Big Data Era: Recent Developments and Open Challenges

Yunyao Li
IBM Research - Almaden
yunyaoli@us.ibm.com

Ziyang Liu
NEC Laboratories America
ziyang@nec-labs.com

Huaiyu Zhu
IBM Research - Almaden
huaiyu@us.ibm.com

ABSTRACT

Enterprise search allows users in an enterprise to retrieve desired information through a simple search interface. It is widely viewed as an important productivity tool within an enterprise. While Internet search engines have been highly successful, enterprise search remains notoriously challenging due to a variety of unique challenges, and is being made more so by the increasing heterogeneity and volume of enterprise data. On the other hand, enterprise search also presents opportunities to succeed in ways beyond current Internet search capabilities. This tutorial presents an organized overview of these challenges and opportunities, and reviews the state-of-the-art techniques for building a reliable and high quality enterprise search engine, in the context of the rise of big data.

1. INTRODUCTION

Enterprises are increasingly relying on search as the primary means of information retrieval, which is mainly driven by two factors. On the one hand, as larger volumes and more varieties of information within an enterprise become available on-line, it is impossible to organize and maintain the content in a predefined hierarchical organization. On the other hand, as users are accustomed to retrieving any information they want on the Internet through search, they expect the same experience within an intranet.

While Internet search engines have been highly successful in content retrieval, enterprise search remains notoriously difficult. For instance, the intranet search engine deployed by IBM before 2011 returned no relevant results in its top 50 results for about 66% of user queries [4]. Not surprisingly, in many companies "nearly half of a knowledge worker's time is non-productive, spent gathering information, converting formats, unsuccessfully searching or recreating content that already exists" [1]. Direct application of techniques for Internet search often leads to poor results and causes loss of employee productivity [3].

This tutorial presents a comprehensive overview of recent work on enterprise search. Specifically, it will cover the following key topics, with an emphasis on new challenges that arose with the era of big data.

- Enterprise search: overview
- Backend data analysis: data collection, within-document analysis, cross-document analysis, building effective search indexes
- Frontend query processing: query analysis, result representation
- Search quality management: utilization of feedback, annotations, and search administration tools
- Case study: intranet search at IBM
- Open challenges and research opportunities

Due to the challenges and significant benefits of building a reliable enterprise search engine, enterprise search has been a topic of keen attention among researchers and practitioners who produced a large number of research and industrial papers in major database conferences. The Enterprise Search Summit has been held every year since 2004. Research labs such as IBM and MSR have been also actively working in this area in the past several years. One well received related tutorial was given in WWW 2010 [2] with a focus on the search quality aspect. In this tutorial we will incorporate more topics related to enterprise search as outlined in Section 2, as well as emerging topics in this area over the past four years with the rise of big data. We will also present a case study on IBM intranet search based on a state-of-the-art big data platform. We will describe by concrete examples how recent research on the various topics have contributed towards a better enterprise search engine that is currently powering IBM intranet search.

2. TUTORIAL OUTLINE

2.1 Data Analysis at Backend

The goal of the backend is to extract and derive as much information as possible from the source data so that high quality search results can be returned for various user queries with as little runtime computation as possible. The emergence of big data as a platform for enterprise search brings in both opportunities and challenges by enabling sophisticated data analysis techniques that are not possible before.

Data collection. In enterprise search, the data can come from a variety of sources and in a variety of forms. Many data sources in the enterprise lie in specialized silos that require special push and pull techniques to retrieve. Some are protected by various access control mechanisms. The refresh rates for various data sources may differ greatly. A plain web crawler may only be able to collect fractions of the enterprise data, and only a tiny fraction of the varieties of data sources. We will describe various mechanisms to overcome the data collection challenges with big data solutions.

Proceedings of the VLDB Endowment, Vol. 7, No. 13

Local analysis. Extracting various pieces of information from individual documents, referred to as *local analysis*, is a non-trivial task that demands great effort in research and development on its own right. It must handle a variety of enterprise document types, and incorporate various specific dictionaries and auxiliary data. It must leverage the rich metadata in enterprise documents, such as publication date, applicable geography, associated divisions and products, and provide a uniform view of document metadata for a heterogeneous collection of documents. We will review a number of different approaches such as rule-based methods and document filtering based methods.

Global analysis. Information across documents are analyzed in global analysis, using techniques such as PageRank and its descendants. Direct application of PageRank to enterprise search is not effective, but a large amount of useful information can be obtained through other global analysis techniques. For example, enterprise contents often contain an excessive number of duplicate documents and documents organized into specialized hierarchies. We will discuss various techniques required in a concerted effort to extract the relevant information suitable for search indexes. We will present later on in the case study on how these techniques fit into a big-data-based solution.

Indexing. The results of data analysis are used to build search indexes, which are then used by the frontend to serve queries. A naive approach to indexing such as the default ones provided by the Lucene index engine will do poorly for enterprise search due to the rich set of jargons, metadata and context that is absent in general natural language. We will present the challenges and later in the case study describe a highly adaptive approach based on Lucene, which is well suited to enterprise search.

2.2 Serving User Queries at Frontend

The search engine frontend serves two roles: (1) to analyze and understand user queries and to translate them into queries suitable for the indexes; (2) to organize and rank the results and present them effectively to the user.

Query analysis. Users often find it difficult to come up with the optimal query keywords. There are an array of proposed solutions suitable to enterprise environments where data have heterogeneous structures: (1) query cleaning for handling misspelled keywords, synonyms and non-quantitative keywords (e.g., "small"); (2) query refinement for queries that are too general, as well as polysemy. (3) generating and selecting query forms for users to specify precise queries without mastering structured query language; (4) query interpretation for automatically generating multiple precise queries from vague user query. We will review and compare these query analysis techniques, as well as techniques for obfuscating query intentions and suppressing sensitive aggregates in enterprise search for privacy reasons.

An important query type in enterprise search, expert search, which searches for people with specific expertise, presents additional challenges compared with retrieving a specific document. For example, expertise information may need to be discovered from problem solving histories; a problem may need to be collaboratively solved by multiple experts; the search algorithm should take into account the social information, e.g., the location/department of the expert, how well two people know each other, etc. We will discuss several expert/solution search approaches in this tutorial.

Result representation. Result representation is important for enterprise search. The results may be from different sources and have different structures, and users may issue queries of different types (informational, navigational, expert etc.), thus the optimal representation of different query results may differ. In this tutorial we will discuss topics of result representation including result ranking, which requires different techniques than those used by Web search engines, result diversification, result summary, result differentiation, result categorization, and access control based on user privilege.

2.3 Search Quality Management

Search quality is crucial for enterprise search, as it directly affects the productivity of the company. In enterprises it is usually relatively easier to get user feedbacks and annotations. A number of search algorithms that utilize them have been proposed, which improves the quality of result scoring, filtering as well as entity disambiguation. In addition, enterprise search deployments are typically managed by administrators who are domain experts but not search experts. This tutorial will cover recent efforts on helping the administrators translate their knowledge of the specific content and search needs of the domain into tuning the underlying engine.

2.4 Case Study: Intranet Search at IBM

We will present a case study on IBM intranet search and describe by concrete examples how various techniques discussed in the tutorial have contributed towards a better enterprise search engine currently powering IBM intranet search. We will also discuss the opportunities and challenges of building such a engine based on a state-of-the-art big data platform.

2.5 Open Challenges and Promising Research Opportunities

We will discuss two major open challenges for future research on enterprise search.

Scalable Deep Data Analysis Recent development in enterprise search has shown that deep data analysis beyond traditional simple keywords is crucial for better enterprise search. One open challenge is how to leverage big data technologies to incorporate more sophisticated data analysis techniques such as entity resolution to support enterprise search in a scalable fashion. Examples of other challenges include how to build large scale real-time indexes to ensure that the latest data content is searchable and how to automatically identify different query types and apply the appropriate search algorithms, ranking functions and result presentations.

Scalable Search Quality Management In spite of recent advancements, enterprise search engines remain largely managed in an ad-hoc fashion. One major challenge is how to enable automatic evaluation and monitoring of the engines with minimum manual involvement. Another interesting direction is how to leverage crowdsourced solutions for better search quality in the context of enterprise search.

To summarize, this tutorial overviews a multitude of problems in enterprise search, and discusses relevant, state-of-the-art techniques. We hope that it will help researchers, enterprise search developers and architects, as well as corporate stakeholders gain insight and better contribute to the field.

3. REFERENCES

[1] Enterprise Findability Without the Complexity. Google white paper. http://goo.gl/aFpSD0.

[2] P. Dmitriev, P. Serdyukov, and S. Chernov. Enterprise and Desktop Search. In *WWW*, pages 1345–1346, 2010.

[3] D. Hawking. Challenges in Enterprise Search. In *ADC*, pages 15–24, 2004.

[4] H. Zhu, S. Raghavan, S. Vaithyanathan, and A. Löser. Navigating the Intranet with High Precision. In *WWW*, pages 491–500, 2007.