Applied Distributed Information Retrieval in Enterprise Search

Erwin Gunadi, Till Plumbaum, and Sahin Albayrak
Technische Universität Berlin
10587 Berlin, Germany
{firstname.lastname}@dai-labor.de

ABSTRACT

Distributed enterprise search as a special case of Distributed Information Retrieval (DIR) is characterized by the need to query multiple repositories in enterprise environments. However, in DIR research there is a lack of publicly available real-world datasets for evaluation purposes. As a result, there is a gap between insights gained from simulated environments and real-world investigations on distributed enterprise search. In this paper, we outline three fundamental issues based on our investigations of a large real-world distributed enterprise search system. We found that (1) the utilization of security features in enterprise repositories, (2) the adaptation of resource description and resource selection for enterprise model, and (3) repository grouping are fundamental real-world issues. We hypothesize that a better understanding of these issues will contribute to improve the distributed enterprise search and better support complex search tasks in enterprise environments. Based on our experience gained from a real-world system, we also outline needed steps to cope with these issues.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Distributed Systems

Keywords

distributed information retrieval; enterprise search, result aggregation

1. INTRODUCTION

Enterprise Search is an area of information retrieval which specifically addresses the information needs of enterprise users. Enterprise is defined as an organizational entity with an exclusive memberships of its users. A typical enterprise environment consists of multiple layers of access rights and different dedicated data repositories such as web servers, file servers, wikis, etc. Key differences to Web search are as follows: (1) Heterogeneous document types such as web pages: wiki, pdfs, emails, word documents etc. (2) Multiple document repositories: Documents are normally not held in a single file server or system. (3) Access restriction: hierarchies and roles rules for every document, and (4) Managed data generation process: Differing from web documents, which are created by individual entity, each enterprise defines their own document creation and update policy, which effectively valid for all of its members [9, 5, 13].

To address the above mentioned issues (1)-(4), the use of Distributed Information Retrieval (DIR) has been proposed [4, 5, 13]. DIR is a concept of managing different resources through a broker. A broker mediates between users and different sources, or repositories, to collect and combine search results. To accomplish this task three sub-problems need to be addressed: resource description, resource selection and result merging [2, 11]. With exception of the TREC Federated Web Search track [1] most of the DIR research have been based on synthetic test collections [11]. Due to the lack of appropriate datasets and the proprietary nature of enterprise data many improvements achieved in DIR are not directly suitable for distributed enterprise search [6, 5]. These factors prevent a further adoption of DIR improvements in distributed enterprise search and creates a gap between these two areas.

This paper has two main contributions. First, we bridge the gap between DIR and the enterprise context, by outlining three fundamental problems that have been widely ignored in the field of DIR research, but are mandatory to be considered in the enterprise context. Second, we propose ways to cope with these problems. The presented problems are (1) Utilization of security features in enterprise repositories, (2) Adaptation of resource description and resource selection for enterprise model, and (3) Repositories grouping. All our findings base on real-world experiences and insights we gain from the operation of a distributed enterprise search system at TU Berlin and the city's administration of Berlin with about 50.000 employees.

In the next Section we outline related works which discuss the integration of DIR in enterprise environment. In Section 3 we describe the architecture of our agent-based distributed enterprise search platform and its deployment. Section 4 details the open problems arise out from our experience. In Section 5 we conclude our paper.

2. RELATED WORKS

Various works have proposed DIR as a possible paradigm to implement enterprise search [5, 13, 12]. The main concept

ECIR Supporting Complex Search Task Workshop '15 Vienna, Austria
Published on CEUR-WS: `http://ceur-ws.org/Vol-1338/`.

[1] `https://sites.google.com/site/trecfedweb/`

of DIR is the usage of multiple resources, or repositories, using a broker-concept in order to satisfy a users' information need. This concept fits the main characteristic of an enterprise environment where normally multiple data repositories for different needs exists, such as web servers, file servers, wikis, etc. [9, 5, 13].

Works investigating how to secure enterprise search system are presented in [1, 14]. Bailey et al. [1] propose the implementation of document level security for enterprise search systems.They evaluated how on-the-fly security checks for each document during search result list building affect the search processing time. Zhou et al. [14] propose the usage of ontology-based user profiles to secure the search process. The ontology models the information search service and user role information, which can be maintained for different departments in an enterprise. It is still an open question about how security restrictions affects search result quality in a distributed enterprise search. Current work focuses more on performance issues than quality.

Regarding the DIR algorithms Li et al. [7] evaluated the performance of various result merging algorithms on multiple enterprise repositories which are unique to each other. Li et al. argue that repositories in an enterprise context are not identical to each other and each of these repositories may have different size, document types, intended audience and administration control. These characteristics need to be explored in the context of DIR in enterprise search.

3. DISTRIBUTED ENTERPRISE SEARCH IN THE REAL-WORLD

The findings we present in this paper based on research cooperation with the service provider of the administration of Berlin, where we have deployed an agent-based distributed enterprise search system [3]. The system is currently used as the standard search platform for about 50.000 employees. The structure of Berlins' network confronted us with some challenges. Even though the whole city can be regarded as a closed organization with state officials as its employees, each of the city's districts maintains its own data management policy. This means a city district is a private network with access to all main data repositories such as city's own intranet, but without access rights to data repositories from other city districts. As opposed to the classic DIR setting these sub-networks cannot be served by a single main broker. It requires that each network area has its' own private broker and an extra broker installed in the main intranet. Another use case for an additional broker is the user desktop. To comply with the user privacy policy from the city's administration, local desktop files should not be externally accessible. Because of such restrictions we have build a local broker so that users can find their local desktop files. The interaction between the users, the search client and the multiple brokers is illustrated in Figure 1.

Figure 1 illustrates how the search client is used to contact all of the different brokers. It enables users to search in different network areas. In each of the network, multiple repositories are queried by the responsible broker. The desktop from the user also has a dedicated broker because local files should not be queried by an external broker.

4. OPEN PROBLEMS

In this section we describe following open problems derived from our investigations on the distributed enterprise search system for the administration of Berlin. We also suggest how these open problem can be further investigated in DIR research in enterprise context.

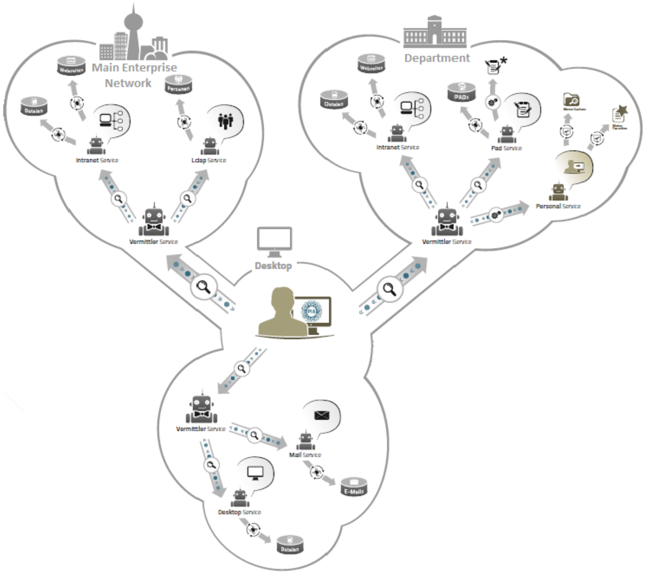

Figure 1: Multiple brokers serving different networks are contacted by a user through a search client

4.1 Utilization of security features in enterprise repositories

Until recently the research of DIR is based on the assumption that all documents are accessible to all users [11, 10]. However, in enterprise environments access to the documents is secured [1, 5, 13]. Depending on the access right each user may see different search results for the same search query. We argue that this property can be utilized as an essential feature for improving distributed enterprise search. Utilization means not only to comply with access restrictions, but also to improve resource selection and result merging algorithm by using the security information.

In order to accomplish this, DIR algorithms for different steps have to integrate the security information as a feature. For example, in DIR literature resource selection is responsible to select the resources with the most relevant documents. However, when a relevant repository is restricted for particular user or user group, these documents will not be shown. This can be mitigated when repositories, with more accessible documents, are higher prioritized even though they have less relevancy. In this case in navigational search tasks employees may get a better recall about the subject. To the best of our knowledge such behavior in a DIR setting is not yet researched. Creating suitable test collections for this purpose is needed to evaluate this essential feature.

4.2 Adaptation of resource description and resource selection for enterprise model

Li et al. [7] highlighted how repositories in an enterprise environment unique to each other are. Based on this fact Li et al. evaluate how these uniqueness may influence the result merging performance. The concern about the exploitation of unique features found in enterprise repositories should also be considered in the other sub-tasks of DIR: resource description and resource selection. Research in this area is needed in order to improve the application of DIR in Enterprise Search. Thus, it will help narrow the gap between these two research fields.

As an example use case of such exploitation, in an enterprise environment a common repository type is file servers. Documents from file servers, which are relevant for search queries, are stored in a hierarchy of directories. We can include these directory names as a part of the resource description of a repository. This means content of a resource description includes not only sampled documents but also directory names. These directory information can be used as important terms that improve the resource selection algorithm. Such exploitation is yet to be investigated in distributed enterprise search context.

4.3 Repositories grouping

One of the challenges from our enterprise environment setting is the need of multiple brokers to handle different networks. Even though the multiple brokers can be seen as a technical feature, it introduces a new perspective in handling multiple repositories. The ability of repositories grouping open the possibility to boost repositories with similar types as a group. The application of boosting for theme specific repositories is being actively researched in distributed web search task context [8]. In this paper, the authors investigated how news sources can be ranked and placed in the web search result. In our deployment scenario we found that search queries about specific law and regulations are common. In this case, having a theme specific group of repositories means documents comes from the law-themed repositories receive higher rank than documents from non law-themed ones.

By having groups of repositories, result merging techniques may rank not only based on the repository rank but also on repositories-group rank. To accomplish this the subtasks of DIR must be adapted to broker context, namely broker description and broker selection. When a particular group of repositories is highly relevant for a search query, the gained broker ranking may be used to highlight a group of repository in the search result page.

5. CONCLUSION

Due to the availability of heterogeneous repositories, previous works have proposed the application of DIR in enterprise search [5, 11, 13]. However, the improvements achieved in DIR research are rarely investigated in real-world scenarios, especially in enterprise environments [12]. Recent works show that further research in the application of DIR in realworld settings is needed in order to close the gap between DIR research and its' real-world application in enterprises. In this paper, we introduced three issues based on our experience from a real distributed enterprise setting.

For future works, we need to investigate how security information can be utilized for the different DIR sub-tasks. This also applies for the unique features found in enterprise repositories, such as file paths and directories. It permits us to better understand the application of DIR in enterprise search, thus, bridges the gap between these two research areas. Also more effort building appropriate test collections for a real-world DIR use case, like the work from Nguyen et al. [10], has to be done for evaluating distributed enterprise search. This helps to adapt available techniques for different tasks in DIR (resource description, resource selection and result merging) for enterprise use cases. Another investigation is needed on how to accommodate repositories grouping by building multiple brokers. Similar to news integration in web search [8], group of repositories can be differently ranked for incoming queries. The ranking result may then be used to highlight a group of repository in the presentation of search result page (SERP) in enterprise search.

6. REFERENCES

[1] P. Bailey, D. Hawking, and B. Matson. Secure search in enterprise webs: tradeoffs in efficient implementation for document level security. *CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006.

[2] J. Callan. Distributed information retrieval. In *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.

[3] E. Gunadi, M. Meder, T. Plumbaum, C. Scheel, F. Hopfgartner, and S. Albayrak. Distributed enterprise search using software agents. In *Proceeding AAMAS '14*, pages 1623–1624, Paris, France, 2014.

[4] D. Hawking. Challenges in enterprise search. In *ADC '04 Proceedings of the 15th Australasian database conference*, volume 27, pages 15–24, 2004.

[5] D. Hawking. Enterprise Search. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 641–684. Addison-Wesley, 2010.

[6] L. Jie, S. Lamkhede, R. Sapra, E. Hsu, H. Song, and Y. Chang. A unified search federation system based on online user feedback. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, page 1195, New York, New York, USA, 2013. ACM Press.

[7] P. V. Li, P. Thomas, and D. Hawking. Merging algorithms for enterprise search. *Proceedings of the 18th Australasian Document Computing Symposium on - ADCS '13*, pages 42–49, 2013.

[8] R. McCreadie, C. Macdonald, and I. Ounis. News vertical search: when and what to display to users. In *SIGIR '13: 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 253–262, 2013.

[9] R. Mukherjee and J. Mao. Enterprise search: Tough stuff. *Queue*, 2(2):36–46, 4 2004.

[10] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1874–1878, 2012.

[11] M. Shokouhi and L. Si. Federated Search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[12] P. Thomas. To what problem is distributed information retrieval the solution? *Journal of the American Society for Information Science and Technology*, 63(7):1471–1476, July 2012.

[13] M. White. *Enterprise Search.* O'Reilly Media, Inc., 2012.

[14] L. Zhou. Multi-agent based distributed secure information retrieval. In *CMC'10*, pages 76–79, 2010.