Distributed Enterprise Search using Software Agents

Erwin Gunadi, Michael Meder, Till Plumbaum, Frank Hopfgartner, and Sahin Albayrak
Technische Universität Berlin
TEL 14, Ernst-Reuter-Platz 7
10587 Berlin, Germany
{firstname.lastname}@dai-labor.de

ABSTRACT

In this paper we introduce a distributed information retrieval system using agent-based technology. In this multi-agent system, each agent has its own specific task and can be used to handle a specific document repository. The system is designed to automatically comply with access restriction rules that are normally enforced in companies. It is used in the administration offices of the German capital city Berlin where it serves as a testbed for further research on aggregated search in an enterprise environment with roughly 50,000 employees.

Keywords

distributed information retrieval, enterprise, software agents

1. INTRODUCTION

Enterprise environments face scattered and unstructured information, as data is created in different formats across different places [7]. This makes searching in such environment challenging, since established search algorithms and techniques cannot easily be transferred to an enterprise scenario [7, 8, 10]. Hence, research on the development of appropriate desktop and enterprise systems is required. Hawking [2, 3] and Mukherjee et al. [7] identify various challenges that arise from such environments: First of all, heterogeneous document types are created, which are stored in multiple document repositories. Moreover, different access restrictions need to be considered, i.e., not everyone is allowed to access all data in an enterprise. While some data is available for every employee (e.g., a company directory service or the company's web pages), more sensitive data (e.g., employees' personal data, blueprints of top secret prototypes, etc.) might be accessible by a minority only. Moreover, the sensitive nature of data available in an enterprise raises specific demands on data security [2]. In particular, using centralized indices can be seen as a major security risk for enterprise search systems.

In this demo paper, we introduce an enterprise search system with distributed indices which addresses the data accumulation task of enterprise search systems. The framework incorporates the idea of data mining agents, a technique which has been successfully employed to create data warehouses [6]. We use autonomous agents for every task in the data accumulation and indexing activity, i.e., each agent provides core services that cover a specific part in the backend. Complex tasks such as crawling and indexing a file server is achieved by combining the corresponding agents, i.e., the autonomous agents form a community to provide a joint service in creating search engine capabilities. When multiple data repositories (collections) need to be indexed we use these agent communities to build a distributed search engine. Search request are handled by broker agents that verify users' identity and their access rights using the enterprise's directory access constraints that are defined using Lightweight Directory Access Protocol (LDAP).

The paper is structured as follows. In Section 2, we review state-of-the-art research related to enterprise search. Section 3 provides an overview of our system. Section 4 concludes this paper.

2. RELATED WORK

Existing enterprise search systems can be classified into two categories: (1) systems that create one centralized index and (2) systems that create many distributed independent indices. A centralized index can be used when it is possible to crawl all of the relevant data sources into a single index structure. However, since in most cases information is stored at different locations and due to physical constraints such as geographical location, low bandwidth connections and administration restrictions, gathering data in one search index is not always feasible [3]. Another important feature of enterprise search system is that it has be able to handle security and rights management issues [7, 2].

We address the issue of building distributed independent indices by applying software agents. Jennings et al. [5] provide a detailed comparison of the agent and other software engineering paradigms. One of the advantages of this concept is that we can model each agent to handle different unique tasks. A similar approach has been introduced by Zhou et al. [11], who, however, relies on ontologies to model user access. Our system is more flexible since its user access is managed automatically by exploiting the existing access rights saved in LDAP.

Note that currently, there are already some commercial enterprise search products available that provide distributed search to some extent. Following recent revelations in the news though, it remains unclear whether such proprietary off-the-shelf systems really comply with strict information privacy rules, especially when outside of US legislation. Therefore, government agencies e.g., in Germany are advised to rely on alternative systems that allow them to maintain control over their data. Addressing this issue, our system is currently trialed in the administration offices of a large European capital city, allowing us to study open research challenges in the field using real users in real context.

3. SYSTEM OVERVIEW

In order to accumulate data from distributed sources, we rely on autonomous software agents that can be combined to build joint services. We propose to provide autonomous agents for every task in the data accumulation and indexing activity, i.e., each agent provides a core service that covers a specific part in the back-end. Example agents include crawling agents based on Nutch or search agents based on Solr. The combination of these agents forms a distributed search architecture that individually crawls and indexes all of the available repositories and create multiple independent search verticals. Search requests are received by a broker agent via REST-APIs and the broker agent will delegate these requests to the various search agents. An example of the combination of these agents is shown in Figure 1.

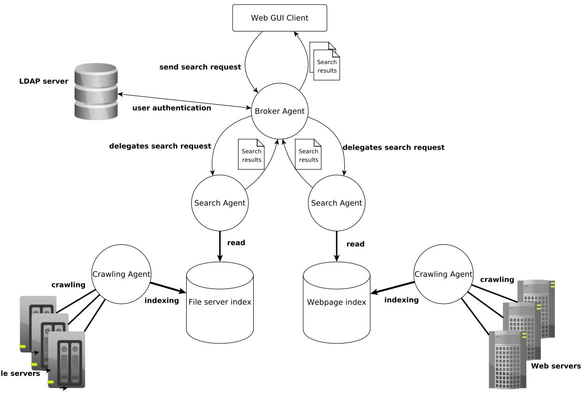

Figure 1: Broker Agent delegates search requests to different search agents on indices prepared by crawling agents

Due to the diverse nature of the collections, some collections such as company external websites do not require any authentication while data repositories or wiki pages require it. The broker agent guarantees that the company's folder restriction rights are respected by communicating with the company's LDAP server.

For organization and implementation of these agents and their communication, we use the JIAC (Java-based Intelligent Agent Componentware) framework [4]. JIAC is a software-agent framework that has been optimized for large-scale applications and services. Agents are arranged on platforms, allowing the arrangement of agents that belong together with the control of at least one manager. Further, JIAC supports encrypted communication and user authorization.

4. CONCLUSION AND FUTURE WORK

In this demo, we introduce a document crawling and indexing system which can be applied in a desktop and enterprise search scenario. As we have outlined, the typical IT infrastructure of enterprise intranets requires flexible data aggregation methods that are able to handle distributed sources. We propose the use of autonomous software agents which communicate via a common agent framework. Advantages of this approach are the agent's flexibility in handling distributed infrastructures as well as the easy possibilities to expand the framework's services further, i.e., by combining existing agents or by adding new software agents.

The introduced approach is currently used as the data aggregation method of a Desktop and Enterprise search system of the administration of Berlin with roughly 50,000 employees, thus assisting a large user base in their professional information gathering task. As future work, we aim to study novel approaches for federated search [1, 9] and analyze interaction logs of these users to further study open research questions, e.g., on content personalization and aggregated result ranking. A demo of the system can be accessed on `http://pia-demo.dai-labor.de/`

Acknowledgement

We acknowledge the support of Christian Scheel, Steven Winter, Jens Meiners, and Arif Oksal. The system has ben developed under the umbrella of IDBB, a research collaboration between DAI-Labor and ITDZ Berlin.

5. REFERENCES

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

[2] D. Hawking. Challenges in enterprise search. *Proceedings of the 15th Australasian database conference - Volume 27*, 2004.

[3] D. Hawking. Enterprise Search. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 641–684. Addison-Wesley, 2nd edition, 2010.

[4] B. Hirsch, T. Konnerth, and A. Heßler. Merging Agents and Services — the JIAC Agent Platform. In *Multi-Agent Programming: Languages, Tools and Applications*, pages 159–185. Springer, 2009.

[5] N. R. Jennings and M. Wooldridge. Agent-oriented software engineering. *Artificial Intelligence*, 117:277–296, 2000.

[6] M. Klusch, S. Lodi, and G. Moro. Agent-Based Distributed Data Mining: The KDEC Scheme. In *AgentLink*, pages 104–122, 2003.

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

[8] J. Peng, C. Macdonald, B. He, and I. Ounis. A study of selective collection enrichment for enterprise search. In *CIKM*, pages 1999–2002, 2009.

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

[10] W. Zheng, H. Fang, C. Yao, and M. Wang. Search result diversification for enterprise data. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1901–1904, 2011.

[11] L. Zhou. Multi-agent based distributed secure information retrieval. In *CMC'10*, volume 1, pages 76–79, 2010.