Advanced Engineering Informatics 30 (2016) 244–257

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Ontology-assisted provenance visualization for supporting enterprise search of engineering and business files

Saiful Khan [a,*], Urszula Kanturska [c], Tom Waters [c], James Eaton [c], René Bañares-Alcántara [a], Min Chen [b]

[a] Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK
[b] Oxford e-Research Centre, University of Oxford, 7 Keble Road, Oxford OX1 3QG, UK
[c] Laing O'Rourke, Admirals Park, Crossways, Dartford DA2 6SN, UK

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form 28 February 2016
Accepted 6 April 2016
Available online 22 April 2016

Keywords:
Enterprise search
Engineering document
Knowledge management
Information visualization
Provenance visualization

ABSTRACT

In many large engineering enterprises, searching for files is a high-volume routine activity. *Visualization-assisted search* facilities can significantly reduce the cost of such activities. In this paper, we introduce the concept of Search Provenance Graph (SPG), and present a technique for mapping out the search results and externalizing the provenance of a search process. This enables users to be aware of collaborative search activities within a project, and to be able to reason about potential missing files (i.e., false negatives) more effectively. We describe multiple ontologies that enable the computation of SPGs while supporting an enterprise search engine. We demonstrate the novelty and application of this technique through an industrial case study, where a large engineering enterprise needs to make a long-term technological plan for large-scale document search, and has found the visualization-assisted approach to be more cost-effective than alternative approaches being studied.

1. Introduction

With a growing amount of data in enterprises, managing and searching for information is becoming a challenge. In Information Retrieval (IR), the concept of *Enterprise Search* was introduced to encourage real-world application of IR, where an enterprise develops an integrated data infrastructure, and enables its stakeholders to have an effective search interface for information retrieval. The scope of enterprise search has been defined by a number of authors, e.g., Stenmark [1], Abrol et al. [2], and Hawking [3]. Many have argued the importance of enterprise search from the perspective of "the high cost of not finding information" [4]. This work focuses on using visualization to reduce the cost in the process of finding information.

It is estimated that the enterprise search industry is growing at 25% per year [5]. A tutorial in ACM SIGIR, 2010 [6] highlighted some of the key challenges in the area of enterprise search. Although these challenges are similar to Web search, many cannot be addressed by existing Web search solutions [6]. For example, the structure of enterprise or desktop data is different from that of Web data. The popular *ranking* technique (e.g., PageRank) is not effective in enterprises search [3]. Although the relevance feedback information can be used to learn and improve the rank of the search results in the enterprise search [7], such techniques often result in an extensive amount of search results but do not guarantee adequate recall and precision. Therefore, there is a need for effective designs of visualization and interaction [8]. In enterprise search, discovering information searched and found by experts and colleagues is often part of a search strategy [9,6]. Systematic support for such collaborative search effort is thus highly desirable.

File search is an elementary task in enterprise search. In a construction company, for example, a large number of files are captured and stored in each project. These files may be of types of reports, CAD files, presentations, scan copies, spreadsheets, art works, images, audios, videos and so on. It has been reported that stakeholders (e.g., engineers, managers, administrators) spend around 30% of time in file search [10–12]. However, searching through a large data infrastructure is a time consuming operation, as (i) search queries often return a long list of files, (ii) the list usually includes false positives, and (iii) the queries often miss some target files (false negatives).

Previous research have been focused on different mechanisms to improve the search time by enhancing the retrieval capability of enterprise search engines for engineering document [11,12], multi-topic documents [13], patents [14], mechatronic data [15],

* Corresponding author.
E-mail addresses: saiful.khan@eng.ox.ac.uk (S. Khan), UKanturska@laingorourke.com (U. Kanturska), TWaters@laingorourke.com (T. Waters), JEaton@laingorourke.com (J. Eaton), rene.banares@eng.ox.ac.uk (R. Bañares-Alcántara), min.chen@oerc.ox.ac.uk (M. Chen).

http://dx.doi.org/10.1016/j.aei.2016.04.003
1474-0346/

BIM document [16–18] and so on. However, existing search engines typically do not provide information about historical search activities. Even when it was done occasionally, the information would be provided in a log form, and it would be time-consuming to read and comprehend such data. In this work, we report a novel approach for using visualization to support collaborative search activities, utilizing both machine and human capabilities.

File search is an explorative process, during which a user progressively refines search criteria by observing and learning from successes and errors in the earlier search results. Devising effective search criteria is a complex cognitive process, which depends on many factors, such as individual users' knowledge about the files to be searched, their familiarity of the organization of the file repository, the critical nature of the search tasks, the likelihood of missing files or multiple copies of the same files, and so on. Visualization-assisted search tools have been found to be effective in supporting search activities, e.g., assisting in iteratively reformulating queries and visualizing search results [19].

In an industrial search operation for engineering and business documents, search tasks are often performed by a team over a period. In order to enable such tasks to be performed efficiently and effectively, it is necessary for team members to access and share the *provenance information* [20] about the historical and ongoing activities related to a task. For complex provenance records, visualization can provide an enabling tool for users to observe provenance information at a glance [21], improving the cost-effectiveness in collaborative search tasks, especially in an industrial setting.

This work was conducted in the context of an industrial application, where intensive file search is part of an industrial process and is to be performed by a team on a daily basis. The background requirements were defined several years ago for the construction industry. By 2016, for a large number of building projects, a construction company must pass on a collection of documents to the client when a building is completed [22], enabling search and reuse of the digital assets. As the total number of files related to a building project could easily be in tens or hundreds of thousands, and many are distributed among different teams and stakeholders in the company, the need for file search in a large distributed file system becomes inevitable.

As part of a feasibility study led by a large construction enterprise, we developed a prototype system that provides visualization-assisted search capabilities. The objective is to enable the enterprise to assess the merits of visualization in comparison with two other approaches, database without visualization and search engine without visualization. This paper presents the visualization-assisted search system, and reports our experience in delivering a technical solution for a challenging industrial problem.

In this work, we addressed several requirements that cannot currently be met by existing enterprise search systems (e.g., Lucene [23], Solr [24], and Indri [25]). These requirements include:

(a) It is desirable to capture, retrieve, and visualize the provenance records of project-based search processes that feature continuing and collaborative search activities lasting for days or weeks. To accomplish this:
   - We have introduced a graph structure called *Search Provenance Graph (SPG)* for recording the search provenance in individual projects and for enabling provenance visualization.
   - We have developed a combined glyph-graph visual representations for visualizing SPGs in a focus + context manner.

(b) It is necessary to couple a visualization interface with a search engine, and to have a common semantic framework for the knowledge base in a search engine and analytical processing in visualization. For example, to measure the similarity between two search records, it is essential to ensure that the same measurement is used by the search engine and provenance visualization. Because it is not feasible to access the knowledge base within a commercial search engine, we had to develop a research search engine. For this purpose,
   - We have developed a prototype *enterprise search engine* with *visualization-assisted search capabilities* for supporting collaborative search activities, query formulation and reformulation, and identification of false positives and false negatives.
   - We have developed *multiple ontologies* as a knowledge base for supporting the search activities as well as for computing semantic similarity in constructing and updating SPGs for provenance visualization. Ontologies stores concepts and instances, along with their relations and properties as a knowledge-base [26].

(c) It is beneficial to evaluate the proposed technical solution in the context of a real world application. To achieve this:
   - We have worked closely with a management team in the industrial partner to compare three approaches for supporting enterprise search operations, and assisted the industrial partner in making a strategic plan for meeting the new requirements to be implemented in 2016. Our visualization-assisted approach compared favorably with alternative approaches under investigation.

The rest of the paper is organized as follows. In Section 2 we present an overview of the related literature. In Section 3 we describe the main motivation behind this work, and a real world industrial scenario. In Section 4 we provide a formal definition of the proposed SPG. In Section 5 we briefly describe an enterprise search engine developed to enable the implementation of SPG and the visualization facilities. In Section 6 we propose an algorithm to compute the similarity between queries in SPG. In Section 7 we present glyph-based visualization and provenance visualization techniques developed for this system. In Section 8 we report an evaluation of the visualization facilities by the domain experts. In Section 9 we offer our concluding remarks.

2. Related work

Our work builds mainly on two areas of research *(a) Information Retrieval (IR):* search strategies, query reformulation, and query similarity measurement, and *(b) Visualization:* visualization-assisted search, provenance visualization, ontology-assisted visualization, and glyph-based visualization.

2.1. Search-related visualization

Many research papers have been published by the visualization and IR communities on navigation and exploration of the Web information space.

VisGets [19] was developed to visualize different Web search attributes, to assist in formulating search queries and to visualize search results using interactive geographic maps and tag clouds. Applications of VisGets to a variety of Web data collections were demonstrated in [27]. An empirical study on interactive and visual exploration of the Web using VisGets was reported in [28]. Fluid Views [29] allows users to view Web search results geographically and temporally in dual layers. PivotPaths [30] allows users to