Everything is Filed under 'File' – Conceptual Challenges in Applying Semantic Search to Network Shares for Collaborative Work

Dirk Ahlers Mahsa Mehrpoor
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no, mahsa.mehrpoor@ntnu.no

ABSTRACT

Lots of professional collaborative work relies on shared networked file systems for easy collaboration, documentation, and as a joint workspace. We have found that in an engineering setting with tens of thousands of files, usual desktop search does not work as well, especially if the project space is huge, contains a large number of non-textual files that are difficult to search for, and is partially unknown by the users due to information needs reaching into previous years or projects. We therefore propose an approach that joins content and metadata analysis, link derivation, grouping, and other measures to arrive at high-level features suitable for semantic similarity and retrieval to improve information access for this case of professional search.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Design, Documentation

Keywords

Collaboration, File System, Network Shares

1. INTRODUCTION

The most common content management system or digital library is arguably a shared file system on a network share that is used as a collaborative workspace. It does not provide actual management functions, but is often the easiest way for users to collaborate on large amounts of files. The challenge is that it is not always a very well structured library. We find no content-focused metadata and apart from textual documents, there is a huge number of varied, non-textual, partly proprietary file types that are hard to index and that have no explicit relations between them. This is not new, but we want to explore what this means in a professional environment where access to information is vital, yet not always properly supported. We share thoughts on solutions to these issues informed by an ongoing project that aims to provide improved information access by deploying semantic recommendation and search solutions.

Our scenario is shared folders or network shares that are used as shared file systems by a group of engineers in a collaborative professional environment [1, 5]. The engineers form a multidisciplinary group with many different domains of expertise with respective documentation as well as specific file types. However, these issues are more general and may also occur in other types of storage systems, such as intranets, digital libraries, or online collaboration tools, as well as in other collaborative settings. Apart from retrieving known files within the shared project workspace, an issue is retrieving files from previous projects that might have solved similar problems, but have almost no personnel overlap with the current project team. This makes knowledge transfer much more difficult as finding files or even getting a good overview is time-consuming. This is different from the local desktop search scenario, which often deals with re-finding of information. In the professional search scenario, on the other hand, routine operations are getting an overview or finding something in a non-personal unfamiliar storage.

We aim to solve this with a solution that supports exploration, search, and recommendation tasks on the corpus. In the following, we discuss the background of the scenario, highlight interesting issues and discuss potential approaches to develop semantic search to collaborative file system workspaces.

2. SCENARIO

The approach includes common document indexing, inferring links and grouping from both textual and non-textual documents, similarity measure, and using recommendation approaches to generate item or workflow-based recommendations, supporting individual views of relevance within the file system.

Contrary to content management systems (CMS) or digital libraries, there is no or very limited user-provided metadata annotated to files. Directly available metadata is mostly related to the file system storage, not to the content, so there is no deeper semantics directly available. More specific, we have two types of sources available, similar to desktop search. First, on the filesystem-level, there are file and

HT '15 September 1-4, 2015, Guzelyurt, TRNC, Cyprus
ACM 978-1-4503-3395-5/15/09
DOI: http://dx.doi.org/10.1145/2700171.2791046.

directory names as well as filesystem metadata, such as creation or modification date, size, owner etc. Second, there is the application-level that depends on the support of the file types and can include metadata as well as file content. For simple file types such as text, no metadata is available, the file is pure textual content. For application-specific formats, specific metadata can be available. Some is mandatory, such as the length of a video, other is optional, such as the title of a presentation. This depends on the support for the file format inside the operating system and the indexer of the search engine. Without application-specific adapters, the content can be undecipherable binary data. We will therefore extend available search tools.

In the engineering case we are looking at, we have additional issues that a lot of the 3D files are binary, but even if we have a parser, there is little textual information inside them that's available and useful for search. A simple keyword search is insufficient in easily bringing up relevant files in this scenario; therefore we aim to use a recommender systems approach. We aim to automatically build an index that contains additional derived metadata, grouping, links, and similarity information.

3. STRATEGIES/METHODOLOGY

There are several disadvantages of a file system compared to the Web, but also some unique features that can be used. In a way, it is similar to single domain site search. However, there are no links between documents, there is a different structure and organization than theInformation Architecture of a Web site and there is a huge amount of relevant yet non-textual files. Also, there is much less frequent interaction between users and documents that in a normal recommender system setting. For example, collaborative filtering could be less useful due to few users, with highly different requirements and roles and the added cold-start problem.

We propose to complement the usual full-text indexing and similarity with additional metadata and annotations, especially for non-textual or no-textual-content files, with features that can be derived from the organization of the file system and its use. A related approach aims to replace directory location with semantic tags [3] and others, such as the Nepomuk project aim for semantic desktops [4]. Another option is to infer a graph structure of semantic links in a file system. It was shown that simple features such as content overlap, file name overlap, and name reference (a filename mentioned in another file) can be used for ranking [2]. The file system can be thought of as providing important context for the files [6], an approach we want to follow up on. The general approach is to automatically generate high-level annotations from low-level features. Annotations and classifications can partially be derived automatically from content, metadata, and file and path names.

Files could be conceptually grouped (i.e. Excel calculation, Word documentation, CAD 3D model) without there being a direct content similarity. Yet, there exists the possibility to group by some other measure of similarity. Examples are the location in the file system, shared paths, similar file names, being in the same folder as other relevant documents, access by same user or group, or, if content is available, mention of entity/product/part names, or general content similarity. This measure can be weighted by the number and type of other files in that directory. For example, a folder that is deep in the hierarchy and has a rather specific name and contains only 3 files, we can assume that they all belong to the same concept. We may find additional features such as similar names, matching parts of names (example, drawing, sketch, 3d). Yet, in a folder filled with hundreds of pdfs, we cannot assume a relation of all of them to each other and need to weight name and content similarity higher. Additional contextual similarity can be based on derivations, such as backup files derived from a file or standard groupings from programming languages or compiler runs that produce a predictable set of files. Some files are only important in conjunction with others, or are superseded by others. In other cases, related documents are distributed over different places in the hierarchy, for example one general design document, a detailed file about the actual model, a separate folder for programming of electronic components, purchases in the finance folder, and related documents from the previous project that has a similar part. Initial work on these features looks promising.

From previous interviews with about ten stakeholders we understand that people mostly work in their individual domains. General documents that are used by more team members are often easier to find and it is clearer where they are. Thus, this will help to search the long tail of project and workspace files.

4. CONCLUSION

Shared file systems are easily mounted in a local system and are easy to use with all available tools and programs. Thus, file systems are used as a workspace as well as a documentation of finished results. In most cases, document management systems are used for finished work, while users are still doing their work locally. The lines get more blurred with online tools or systems such as SharePoint that allow office applications to direct open from and save to a web system or other tools such as Google Drive or Microsoft OneDrive. We expect to be able to use our approaches there as well, as they strongly depend on the file structure that users set up to organise their files as well as files' contents. Our future work concerns a deeper analysis of the contents and distributions of shared file systems, including analyses of the performance of the different parts of our approach on file sets drawn from professional workspaces.

5. REFERENCES

[1] D. Ahlers and M. Mehrpoor. Semantic social recommendations in knowledge-based engineering. In *Social Personalization Workshop 2014*. CEUR, 2014.

[2] D. Bhagwat and N. Polyzotis. Searching a File System Using Inferred Semantic Links. Hypertext '05, 2005.

[3] O. Eck and D. Schaefer. A semantic file system for integrated product data management. *Advanced Engineering Informatics*, 25(2), 2011.

[4] T. Groza, S. Handschuh, et al. The Nepomuk project - On the way to the social semantic desktop. In *I-Semantics' 07*. JUCS, 2007.

[5] M. Mehrpoor, J. A. Gulla, D. Ahlers, K. Kristensen, S. Ghodrat, and O. I. Sivertsen. Using process ontologies to contextualize recommender systems in engineering projects for knowledge access improvement. In *ECKM2015*, 2015.

[6] C. A. N. Soules and G. R. Ganger. Connections: Using context to enhance file search. SOSP '05. ACM, 2005.