C. Stary (Hrsg.): Mensch & Computer 2005: Kunst und Wissenschaft – Grenzüberschreitungen der interaktiven ART. München: Oldenbourg Verlag. 2005, S. 171-180

Using content analysis to support the noise detection and visualization of shared workspaces

Wolfgang Prinz, Baber Zaman
Fraunhofer FIT, RWTH Aachen, Germany

Abstract

Shared workspace systems provide virtual places for self-organized and semi-structured cooperation between local and distributed users. These systems are adopted by a large community over the past years and the volume of information managed by these systems is expanding rapidly. A problem that occurs frequently using these systems is the missing user support for the workspace organization and assistance for finding the right place for storing new contributions. This often results in weakly organized workspaces making it difficult to find documents. Current systems do not provide support for the identification of the homogeneity of the workspace content. Thus, support to find noisy documents and providing appropriate alternative locations is missing. This paper presents a solution that helps to detect such noise in workspaces and to provide alternative suggestions for the workspace organization. The visualization of the inter-folder similarity in workspaces helps users to find redundant folders and to reorganize the workspace hierarchy. The support for finding similar documents to a given document helps users to find related documents easily.

1 Introduction

In recent years shared workspace systems (Pankoke-Babatz and Syri 1996) have become a widespread application for the support of flexible and weakly structured cooperation in teams and communities. Typical examples for such systems are BSCW (Appelt 1999) (Fraunhofer FIT and OrbiTeam 2005), Hyperwave (Hyperwave AG 2005), Livelink (Open Text Corporation 2005) or MS-Sharepoint (Microsoft 2005). Application areas for these systems are the coordination of different kind of projects like lectures, exercises, diploma work, Industrial R&D projects, worldwide international research and development projects.

A shared workspace normally contains different kinds of information such as documents, pictures, and URL links to other Web pages, threaded discussions, or member profiles. The content of each workspace is represented as information objects arranged in a folder hierarchy. Awareness about the activities of the workspace members is provided by email notification or activity icons in the user interface of the system.

Since shared workspace systems do not impose a fixed structure on the workspace organization, each workspace can be organized according to the needs and requirements of the cooperating team. Most preferred structures for workspace organization are project structures (work packages, meetings) or organizational structures (departments, projects). Often structures that reflect both criteria are applied. However, the aim and intention of these structures is often not immediately visible to the users who share a workspace. Although workspace or folder descriptions can be used to describe the purpose of each workspace or folder, users are often confused about the hierarchy, resulting in the effect that they have problems in finding the adequate folder to which they can upload a new document or where they can find the appropriate information. Although the users cooperate through a shared workspace, they fail to use and understand it as a common information space (Bannon and Bødker 1997).

This paper discusses the design and implementation of a Semantic Workspace Organizer (SWO). SWO provides the functionality to detect noise in shared workspaces based on textual similarity of documents and it suggests more appropriate locations to which these documents can be moved. We consider documents which do not have content similarity with the rest of the documents in a folder as “noise”. Noise leads to a disorganization and fragmentation of the workspaces. The presented system provides the functionality to search for similar documents in a workspace for a given document from the file system. This feature makes it easy to find related work from workspaces. In addition, SWO provides a visualization of the inter-folder similarity for the whole workspace or selected folders. As the base system of our experiments, we have selected BSCW (Basic Support for Cooperative Work), a widely used shared workspace system (Appelt 1999) (Appelt 2001).

SWO applies text mining techniques on a workspace to find the intra-folder similarity of the documents and the inter-folder similarity among folders in different workspaces. The intra-folder similarity assists users to identify the documents, which do not match other documents and suggest alternative locations for these documents. The inter-folder similarity provides awareness about other projects and helps the users to find related contents. Users with similar interests can be found based on the contents they have created so far.

The expected benefit of this solution is a user-support for the reorganization of workspaces based on content similarity with improved consistency of the different workspaces and folders. This paper is structured into the following sections: the first section describes design requirements of the system. The second describes the resulting design and implementation of the system including its user interface, which is followed by the evaluation, conclusion and an outlook.

2 User requirements

We started the requirement analysis with a user survey to identify the user needs of a support tool for the organization of shared workspaces. The survey has been distributed to 34 experienced and frequent users of BSCW. The results indicate that many users actually require a support tool that provides an indication of irrelevant documents in workspaces, that provides similarity of different folder contents and searching similar documents in the workspace system for given document. A brief summary of the results is given in Table 1.

Purpose	Option	Result
BSCW experience	more then 2 years	56 %
Frequency of BSCW use	daily	60 %
Frequency of document upload	daily	60 %
Problems finding the location for new documents	often Sometimes	32 % 38 %
Need to manage documents on content similarity	strongly recommended	77%

Table 1: Survey statistics (34 interviewee)

The survey results indicate that users often have problems to identify the appropriate location for a new contribution and that 77% of the users would appreciate a support for the organization of folders based on content similarity. Further discussions with users indicated that the application of a text mining approach can be useful to identify misplaced documents in shared workspaces. In the following section we describe the concept and implementation of the Semantic Workspace Organizer.

3 Semantic Workspace Organizer

In this section we describe the system architecture, followed by the textual analysis builder and the noise detection.

3.1 System Architecture

The system architecture is illustrated in Figure 1. The SWO system is designed as a client server application using the JAVA programming language. The SWO-Server explores the shared workspaces by performing a text analysis to examine the suggestion results. The information is represented using a Semantic Web Ontology. The SWO-client realizes the user interface accessing the server through web-services.

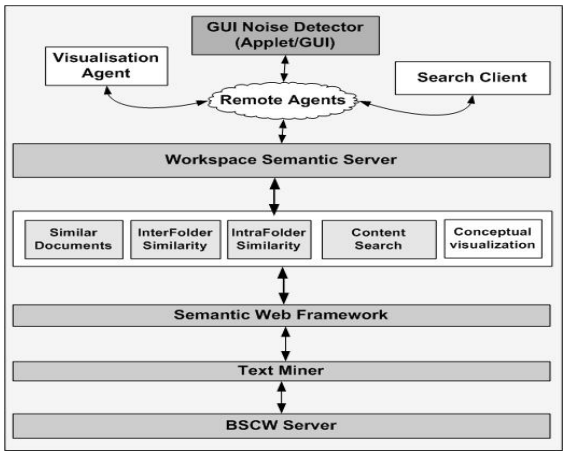

Figure 1: System architecture of the Semantic Workspace Organizer.

The primary component of the SWO-server is the text miner. The text miner component analyses the documents stored in the shared workspace to extract keywords and to build the index that is later used for text categorization and search. The semantic web framework component builds the ontology to support search and compare operations. The workspace semantic server realizes a server side component which is responsible for the communication with remote clients.

The use of semantic web methods in the SWO-server increases the modularity of the application since the ontology is machine readable and can be processed by any system. SWO-Server interfaces can be accessed through web-services, which are based on HTTP. The SWO realizes a pure semantic web application, where the SWO-Server builds the ontology for all underlying information and the SWO-Clients access this ontology and processes it.

3.2 Text Analysis and Noise Detection

Although the shared folders appear in a hierarchical structure in the users' workspaces, we can not consider this hierarchy as a true hierarchy of concepts. The folder hierarchies are not created using a conceptual structure based on the content of the folders. The user survey indicated that the folders are structured using organizational, project or other structures based on the purpose of the cooperation process for which they are used. Therefore, folders with the same or similar kind of content or not necessarily hierarchically ordered, but they may exist at different locations in the workspace. Thus the textual analysis of the folders must consider each folder as a separate class (category).

Many text categorization algorithms have been developed by the data mining community (Aas and Eikvil 1999). For our implementation we use the Apache Lucene Project (The Apache Jakarta Project 2005) to store the indexes of all the documents found in the shared

folders. This system was selected since it is easy to use and it provides a powerful support for searching.

Documents which have not been uploaded to the right location are considered as noise. For example if a document related to one project-topic is uploaded to another topic-folder. This misplacement of contributions often creates usability problems in shared workspaces. It makes it difficult to find the relevant information by other users.

For our approach we consider documents which have very low content similarity to other documents in a folder as noise or misplaced documents. Noise detection is based on finding the intra-folder similarity of documents in a folder. The intra-folder similarity measures the similarity of all documents with each other and finally finds the similarity of each document to its folder by making an average of similarities to other documents as shown in Eq 1.

All the documents and their similarities are visualized using graphs. Graph nodes are document names and their edges are similarity percentages. The similarity of two documents is not symmetric. To represent the similarity of two documents on the edge, we use the average of similarity scores between two documents and edges represent individual similarity as well as average similarity. Results of the Equation 1, which brings similarity of document to its folder are named as noise score. The auxiliary function similarity($document_i$,$document_k$) finds the similarity of documents based on their content using results of Apache Lucene.

$$noiseScore(document_i) = \frac{1}{k-1} * \sum_{k=files\_in\_folder \& k \neq i} similarity(document_i, document_k) \quad \text{-- Eq 1}$$

The documents which have a noise score less then a predefined similarity quotient e.g. 20% (configurable by the user), are considered as noise. The system provides suggestions (folders having similarity more than the current folder) for these documents Upon user approval, documents are either moved or linked to the new location.

The inter-folder similarity provides an overall view of folders with their similarities. The inter-folder similarities are illustrated using graphs, where nodes are folders in the workspace and edges are similarity between folders. The inter-folder similarity is calculated by comparing all documents of given folders with each other and the similarities are determined by making an average of the document similarities. Equations 3 and 4 show how the inter-folder similarity is calculated.

$$folderScore(folder_x, folder_y) = \frac{1}{i} * \sum_{i=files\_of\_folder\_x} docSimilarity(document_i, folder_y) \quad \text{-- Eq 2}$$

$$docSimilarity(document_i, folder_y) = \frac{1}{k} * \sum_{k=files\_in\_folder\_y} similarity(document_i, document_k) \quad \text{-- Eq 3}$$

4 The SWO user interface

This section describes the user interface elements of the application and the provided functionality. The current implementation provides user interfaces for the inter-folder and the intra-folder similarity. It also provides user interfaces for a content-based search and to find similar documents based on content similarity. Standard functions such as workspace browsing and document views are also included. The user interface is organized in three main panes: the workspace browser, a content panel and a history panel. The content panel contains different interfaces for document view, the inter/intra-folder similarity and searching.

This similar documents view provides the functionality to find similar documents in all shared workspaces to which a user has access based on content similarity. After the user has identified a local document, this document is compared to all documents in the shared workspaces using the semantic web framework. Similar documents are shown in a list indicating the matching score. This functionality is useful to find related document as well as users who have contributed similar content.

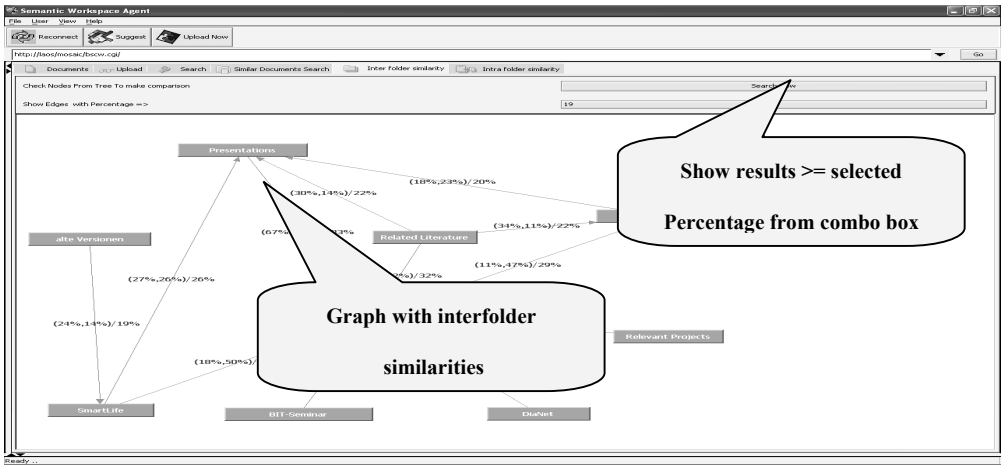

Figure 2: Inter- folder similarity graph view of SWO-client.

The inter-folder similarity view (Figure 2 and 3) visualizes the similarity of given folders using directed graphs. Each node represents a folder and edges indicate the similarity of the folders. The similarity is shown in percentages using the format (source similarity, target similarity/average similarity). The user interface provides functions to filter the folders with certain percentage by selecting the required minimal similarity value from a combo box. Tests have shown that a value of 25 % leads to a good selection resulting also in a comprehensible graph display. Therefore the number of edges linked to the various boxes may differ. The initial display of the interface places the folders with the highest number of edges in the centre of the canvas. Users can rearrange the different boxes to create a view that fits their needs best.

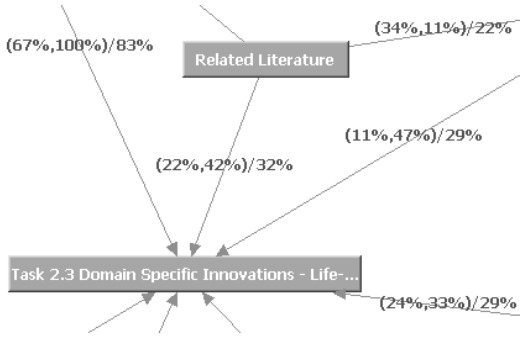

Figure 3: Detail of the inter- folder similarity graph.

The intra-folder similarity view (Figure 4) visualizes the similarity of all documents with each other in a single folder. Documents are represented as graph nodes and their similarity is shown on edges. The user interface also displays the accumulative similarity of each document with all other documents.

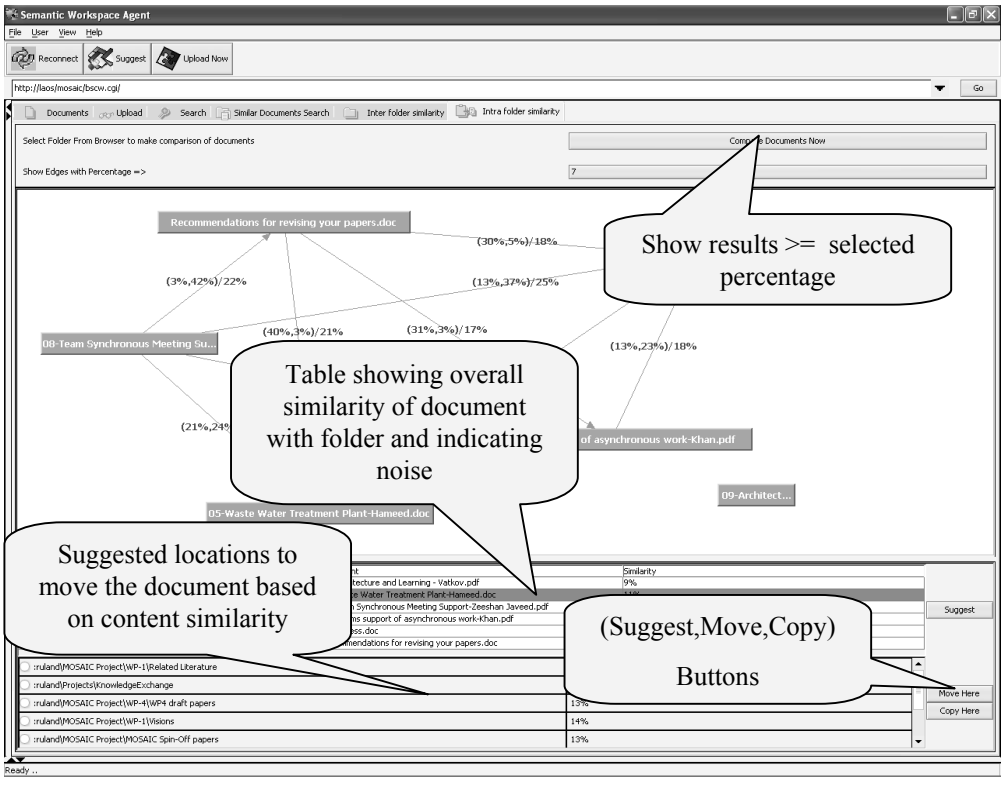

Figure 4: Intra-folder similarity graph view of SWO-client

If the document similarity is less than the defined similarity coefficient (e.g. 20 %), it is marked as noise and visualised using a different colour scheme. Pushing the suggest button finds better locations for the selected document and displays the results in the suggestion panel. The move button moves or links the document to the suggested location.

5 Evaluation

We have evaluated the system by distributing the system to five expert users of the shared workspace system BSCW. The test scenario consisted of different shared workspaces. Some of these workspaces were known to the users while others were new to the users. Each evaluation session took between 20-40 minutes including the discussion. Due to the limited number of users we will not provide any statistical analysis of the feedback results, but we report on qualitative findings derived from the evaluation sessions with the test users.

The graphical representation of the inter-folder similarity and intra-folder similarity was highly appreciated. Users argued that this is an intuitive representation for the similarity scenario. Spontaneously they grabbed specific boxes to rearrange the graph around the folder of interest. Users agreed that the inter-folder and intra-folder similarity displays are useful tools for the workspace (re)organization. The information provided by the systems was considered suitable to identify folders and documents that should be either moved to other locations or merged with existing folders.

In some instances the documents identified by SWO as noise documents where not considered as misplaced documents by the users. This occurs mainly in folders that represent organisational structures instead of content structures. Therefore the proposal to automatically reorganise documents or workspaces was not accepted by the users. Several users argued that the tool is also a suitable knowledge management tool. The inter-folder similarity can be used to estimate the overlap of contents among different projects. Therefore relevant projects can be identified from which knowledge can be gained.

Finding similar documents was highly appreciated. Users argued that the idea can be extended to provide a list of related or similar documents based on a combination of different criteria like documents created by the same or other users, other documents in the same folder or documents having similar activity pattern. Within the context of knowledge management and expert finding (Ackerman, Pipek et al. 2003) users argued that this functionality is also valuable to identify users who have contributed on similar topics.

In summary the users provided a positive response to the functionality of the application and they wished to continue using the system also in the future. However, it was also made clear that users would prefer an integration of these functionalities into the standard user interface of BSCW. This demand is caused by the requirement that users do not wish to switch between different systems for different functionalities. Therefore the integration of selected SWA functions into the standard BSCW interface is currently considered.

6 Related work and conclusion

A lot of research on document mining and categorization has been conducted in the information retrieval community while the CSCW community has not really integrated that work in their considerations to a large scale. One exception are group-based recommendation systems (Konstan, Miller et al. 1997), the MILK system (Agostini, Albolino et al. 2003), or the AWAKE system (Novak, Fleischmann et al. 2004) that uses text mining approaches to provide related information to the documents a user is currently focusing on.

Other approaches consider the process and workflow context of a document to support knowledge-based document retrieval (Celentano, Fugini et al. 1995). The work presented in this paper aims at a proactive provision of information to support the organization of shared information. This concept is also used by the FXPAL Bar (Billsus, Hilbert et al. 2005) or by Calvin (Bauer and Leake 2001) to provide information to users based on their current tasks and interest.

In this paper, we present a concept of using content analysis of information available in shared workspaces to generate the inter-folder similarity and the intra-folder similarity. The intra-folder similarity assists users in finding overlap between different folders and projects. The inter-folder similarity leads to detection of noise in workspaces and assists users in finding better locations with higher content similarity. Initial user feedback shows evidence that the approach actually supports users in the process of finding and moving noise documents to appropriate locations. Finding a related set of documents to current work can be achieved by using the "find similar documents" functionality of the system. The system searches for all documents which have high content similarity to a given document.

If we consider the documents produced or manipulated by a single user, instead of the documents contained in a folder we can use our SWO server to find the inter-user similarities. Thus users can be compared based on the contents they have produced and it may lead to find users with common interests. Just like finding similar documents to a given document, a user search can be performed by providing a document and asking for all users who have produced documents similar to that one.

Shared workspaces contain a lot of useful information, which is not yet used in a way, as it could be used. Currently workspaces are primarily a collection of objects in a shared folder hierarchy. This paper presents an approach to make workspaces and their organization more transparent. After the successful realisation of the concept in an external client to BSCW, the next steps will focus on the integration of selected functionalities in the standard BSCW system and user interface.

References

Aas, K. and L. Eikvil (1999). Text categorisation: A survey, Technical report, Norwegian Computing Center.

Ackerman, M., V. Pipek and W. Wulf, Eds. (2003). Sharing Expertise: Beyond Knowledge Management. Cambridge, MA, The MIT Press.

Agostini, A., S. Albolino, G. D. Michelis, F. D. Paoli and R. Dondi (2003). "Stimulating knowledge discovery and sharing[28] (abstract only) " SIGGROUP Bull. 24 (1): 14-14

Appelt, W. (1999). WWW Based Collaboration with the BSCW System. SOFSEM'99, Milovy, Czech Republic, Springer Lecture Notes in Computer Science 1725, 66-78.

Appelt, W. (2001). What Groupware Functionality do Users Really Use? Proceedings of the 9th Euromicro Workshop on PDP 2001, Mantua, IEEE Computer Society, Los Alamitos.,

Bannon, L. and S. Bødker (1997). Constructing Common Information Spaces. ECSCW'97: Fifth European Conference on Computer Supported Cooperative Work, Lancaster, UK, Kluwer Academic Publishers,

Bauer, T. and D. B. Leake (2001). Exploiting Information Access Patterns for Context-Based Retrieval. International Conference on Intelligent User Interfaces, IUI-01, ACM-Press,

Billsus, D., D. M. Hilbert and D. Maynes-Amizade (2005). Improving Proactive Information Systems. Intelligent User Interfaces, IUI'05, San Diego, California, USA, ACM-Press, 159-165.

Celentano, A., M. G. Fugini and S. Pozzi (1995). "Knowledge-based document retrieval in office environments: the Kabiria system " ACM Trans. Inf. Syst. 13 (3): 237-268

Fraunhofer FIT and OrbiTeam, (2005), BSCW - Basic Support for Cooperative Work, http://bscw.fit.fraunhofer.de

Hyperwave AG, (2005), Hyperwave, http://www.hyperwave.com

Konstan, J., B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Rield (1997). "GroupLens: Applying Collaborative Filtering to Usenet News." Communications of the ACM 40(3): 77-87.

Microsoft, (2005), Microsoft Sharepoint, http://www.microsoft.com/sharepoint/

Novak, J., M. Fleischmann, W. Strauss, M. Wurst, K. Morik, C. Kunz and J. Ziegler (2004). Verbindung heterogener Experten-Communities durch die Entdeckung, Visualisierung und Nutzbarmachung von stillem Wissen – das AWAKE Projekt. Konferenzband der Konferenz Mensch und Computer 03, Teubner Verlag,

Open Text Corporation, (2005), Livelink, http://www.livelink.com

Pankoke-Babatz, U. and A. Syri (1996). Gemeinsame Arbeitsbereiche: Eine neue Form der Telekooperation. Herausforderung Telekooperation: Fachtagung Deutsche Computer Supported Cooperative Work, Stuttgart-Hohenheim, Springer, 51-68.

The Apache Jakarta Project, (2005), Apache Lucene, http://lucene.apache.org/java/docs/index.html