Organization of digital resources as an original facet for exploring the quiescent information capital of a community

Guillaume Cabanac · Max Chevalier · Claude Chrisment · Christine Julien

Received: June 12, 2010 / Revised: July 7, 2011 / Accepted: November 15, 2011 /

Abstract Knowledge workers organize the documents they need for daily task achievement in their Personal Information Spaces (PISs). For a community, people's PISs constitute in-house value-added resources. Paradoxically, this information source is poorly exploited, as people tend to use external sources (e.g., the Web), although this is probably poorly appropriate in corporate context. This article tackles such information access issues in the common context. Our contribution consists in a faceted visual interface to explore various facets (points of view) of the information of a community, which remains quiescent otherwise. Besides common facets only based on information contents, we propose a new facet relying on the way users in a community manage and organize information. As a result, our approach exploits knowledge workers' efforts devoted to PIS management, turning them to profit for all by fostering mutual benefit between stakeholders. The proposed facet relies on an original organization-based similarity measure that we define and experiment.

Keywords Information System · Personal Information Space · Organization · Document · Faceted Search · Knowledge Worker · Visualization · Exploration

G. Cabanac · M. Chevalier · C. Chrisment · C. Julien
University of Toulouse
Computer Science Department
IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
E-mail: Guillaume.Cabanac@irit.fr, Max.Chevalier@irit.fr, Claude.Chrisment@irit.fr, Christine.Julien@irit.fr

1 Introduction

Information Systems (IS) today are the cornerstone of any community: corporations, organizations,[1] research laboratories, industries, online groups, and so on. Members of such communities definitively depend on their IS for achieving their tasks. They especially use it to access information embodied in documents, which may be located inside or outside the community. Once retrieved, each member stores and organizes the documents he/she considers useful regarding his/her activities in his/her Personal Information Space (PIS) built on filesystems or bookmarks, for instance. Their structure is mostly hierarchical, reflecting individuals' cognitive efforts spent on maintaining the PIS, which involves document insertions, deletions and moves. As a result, members' hierarchies truly are mines of information supporting their daily activities.

Nevertheless, such internal sources of information are too often neglected in favor of a unique external source: the Web. This situation mostly results from how difficult it is to access the document resources of one's community. Indeed, the documents filed by any individual remain out of reach for his coworkers unless he is allowed to modify the file permissions on his/her PIS and communicates the access paths to his/her colleagues. This situation is all the more paradoxical for documents retrieved from public information sources, which required long and tedious searches. Although introduced in the community, such public documents are unknown to most people because people are not able to reach them through the restricted PISs. Consequently, these internal document resources remain quiescent in the members' PISs. This has terrible consequences on communities' efficiency, as Lewis E. Platt, former CEO at Hewlett-Packard, criticized it: "If HP knew what HP knows, we would be three

[1] For clarity concerns, in the remainder of the article we keep the word 'organization' for referring to the way people classify documents.

times as profitable" [1]. Individuals, however, would make the most of their community documents if made available and highlighted by a win-win system, fostering mutual benefit between every member of the community.

This article tackles information access issues faced by any community. We designed a faceted visual interface for exploring the resources of any community. This contribution does not aim at either constituting a new document repository or feeding a corporate archive. On the contrary, we harness the community's existing IS without requiring any specific change. A point in favor of our approach lies in its unintrusive aspect: the way people work with documents is neither questioned nor changed. In practical terms, our proposed interface allows the visualization of documents and knowledge workers according to two main facets: their *topics* (extracted from document contents) or their *organization* (induced from document localizations within the PISs: two documents being all the more organization-based related since they are often classified in close folders through the PISs). These facets are displayed via several visualizations allowing the users to understand the many relations interweaving documents and knowledge workers.

The article is organized as follows. In Section 2, we discuss the reasons why documents introduced in an IS are poorly profitable for any community today. We claim, however, that such documents may turn to be valuable for the community. In this respect, we address in Section 3 the extraction of document organization from PISs, as a clue for learning how knowledge workers organize information. We then introduce organization-based similarity measures, and show in Section 4 their complementary role with respect to classical content-based similarity measures. In Section 5, both of these measures are considered for designing a faceted visual interface dedicated to the knowledge workers and managers of a community. It intends to increase the return on investment of the IS by harnessing the structured PISs, as illustrated in the scenario presented in Section 6. Then, we discuss the limitations of our approach in Section 7. Finally, Section 8 reviews related works before concluding this article by outlining research directions.

2 Exploring the document resources of a community: pitfalls and promises

This section highlights how dormant community members' PISs are. We hypothesize, however, that they can turn out to be highly profitable. We first review state-of-the-art approaches intending to put documents to advantage: manual and automatic dissemination, as well as techniques for information visualization. We then emphasize on their respective limitations, leading us to the designing of the proposed faceted visual interface.

2.1 Document hierarchies stored in an IS: valuable resources awaiting exploitation

In this section, we consider the links tying individuals to the documents they introduced into the community during their daily activities. We particularly focus on the *knowledge workers*: people who mainly work with information, and enrich it with value-added knowledge and materials to produce enhanced information [2, 3]. This corresponds to numerous occupations (e.g., engineers, scientists, managers, journalists). Sellen and Harper [4, p. 51] reported that they represented 31% of the US workforce in 1995. This proportion was intended to 'continue to increase significantly into the new millennium.' In practical terms, we are all becoming knowledge workers in so much as we increasingly work with information.

Feldman [5] estimated that knowledge workers devote between 15% and 35% of working time to searching for information. This task is highly cognitive since it requires one to put information needs into query terms. Then, these are submitted to several tools targeting the identified most appropriate information sources (e.g., search engine of a company's intranet). Lastly, search results must be analyzed for extracting the documents matching the user's information need. When he/she anticipates that a given document will be worthy for future use, he/she can decide to store it in his PIS. Actually he/she often manages multiple PISs, such as filesystems, Web bookmarks, and emails. Classically, a PIS is structured hierarchically as a tree of folders. The organization reflects how people split up their activities into tasks and subtasks, mostly for project planning purposes [6, 7]. Khoo et al. [8] commented on the frequency of hierarchies with three levels or more. This requires individuals to make much effort to identify the most appropriate target folder(s) for any given document, or to create a new folder whenever needed [9]. As a result, the PIS of each member comprises the documents he judged useful, organized so as to best achieve his daily activities. All members' PISs then constitute a profitable source of information for the community as a whole.

Paradoxically enough, the resources constituting the PISs do not benefit the entire community: by default, a PIS is accessible only to its owner. Therefore, the documents once found and kept by one member are not profitable to others, although some people may have similar interests and information needs. So these documents remain dormant in people's PISs, being the subject of repeated search efforts, sometimes in vain since 50% of the issued queries fail [5]. Ignoring nearby knowledge and skills leads to information recreation: a newly created report would consist of 90% preexisting information [5]. The next section comments on state-of-the-art sharing and disseminating techniques that one may consider in order to compensate for the aforementioned issues.

2.2 Limitations related to sharing and disseminating community documents

We detail the options offered to people for sharing and disseminating their documents within their community. Manual and automatic solutions are presented, emphasizing on their limitations for individuals, at a cognitive and at a motivational levels [1].

2.2.1 Manual sharing and dissemination of documents

There are many ways of sharing and disseminating documents within a community:

- By setting *read permission* on the filesystem folders. This sharing strategy is limited because one must identify all the potentially interested people and inform them of the path to the accessible folders.
- By creating a *shared networked folder* reachable by any member of the community. Each member then has to make an effort to contribute documents to this shared space. Its structure (in terms of file and folder names, and organization into subfolders) imposes a 'single thought' when maintained by a unique person, as the other people must subscribe to this idiosyncratic perception of the documents. This problem remains the same even with unsupervised classification techniques as proposed in [10, 11]. When document classification is left to the group as a whole, hoping to assist to the emergence of a consensual structure, each individual is nevertheless forced to subscribe to an organization scheme that may not correspond to his own perception. As a result, people have to adapt, which implies an increased cognitive load.
- By publishing them on a *wiki* or on the *intranet* of the community with software like MS SharePoint Services or Lotus Notes. This requires an effort from members who have to choose the most suitable categories for a given document, wondering where other people would search for it. This task is all the more difficult as intranets get increasingly large with the passage of time. A representative example may be IBM's, counting 5.5 million pages at least in 2006 [12]. Moreover, Feldman [5] estimates that 40% search attempts on major corporations' intranets fail, thus showing their poor effectiveness.
- By using *social bookmarking* [13] such as IBM's Dogear [14] or Nature's Connotea [15]. They allow the creation of a bookmark collection that can be shared with other users. Each bookmark is comprised of the URL of the document, a comment, and tags (i.e., descriptive words given by the individual). Then, browsing from tag to tag enables the exploration of the document collection. The main limitation of this approach relates to the various languages used, and the semantics of the tags because they are ambiguous: 'DB' may refer to 'database,' as well as 'Deutsche Bank,' for instance.

An alternative to sharing documents is to spreading them via emails or mailing lists. This active initiative consists in choosing the documents to send, then identifying potentially interested people. This requires efforts from the sender who has to anticipate his colleague's needs. Moreover, the recipients may well be overwhelmed when receiving documents from multiple coworkers on a daily basis. In order to reduce such efforts required from individuals manually, the automatic technologies presented in next section may be helpful.

2.2.2 Automatic sharing and dissemination of documents

Setting up an information filtering system is an alternative to manual document seeking and sharing. It intends to automatically recommend documents to people with respect to their information needs. This process requires to build profiles for both documents and users' needs. There is a plethora of criteria that may be considered for this purpose, as shown by Montaner et al. [16]. A possible option may represent document topics, and people's interests. The recommendation process then relies on a matching function between documents and user profiles. Its main limitation concerns the difficulty to model profiles, then to make them evolve as to best represent and satisfy the current expectations of the assisted user. Moreover, matching users with documents is also limited. It suffers from prominent issues, such as the need of a critical mass of end-users, the existence of the cold start problem (difficultly in issuing early recommendations), and the vocabulary problem [17] (e.g., dealing with synonyms, homonyms, stylistic devices).

Besides the sharing and disseminating approaches, knowledge workers may use various visualization systems presented in the next section for exploring the resources of communities.

2.3 Limitations of community documents visualization and exploration

For a user, exploring a document repository may imply the use of a faceted search system [18]. These allow users to filter documents using multiple criteria (facets) jointly, in order to quickly identify relevant documents and information. The idea is that 'any complex entity could be viewed from a number of perspectives or facets' [19]. The main advantages of such an approach are: facets are easily extensible and more expressive than other single document classification (e.g., hierarchy of concepts). Lots of work related to on-line services used faceted search and faceted classification in various contexts, such as Information Retrieval and Digital Library [20, 21, 22, 23]. The effectiveness of facets

for the search task was notably demonstrated by Kules et al. [24]. Unfortunately, these faceted search systems are mainly based on a textual display of information. As a result, the relations between information and their exploration in the document repository belonging to a community is rather difficult to unveil. For the same purpose, however, visual techniques have long been proposed as to visually explore document repositories. There is a wealth of literature about information visualization techniques and systems, as evidenced by various surveys [25, 26, 27]. Without attempting an exhaustive presentation, this section outlines real applications to illustrate the way users can explore document repositories. They use both document contents and metadata (e.g., file size and type).

Among *metadata*-based visualization approaches, Fekete and Plaisant [28] use the Tree-map visualization introduced by Johnson and Shneiderman [29] for representing a filesystem according to file size. Figure 1 shows interlocked boxes representing folders and files. Their color depends on file extension, and their size is proportional to the physical folder or file size. This visualization also shows the imbrication level of folders: the more a folder is imbricated, the darker it is.

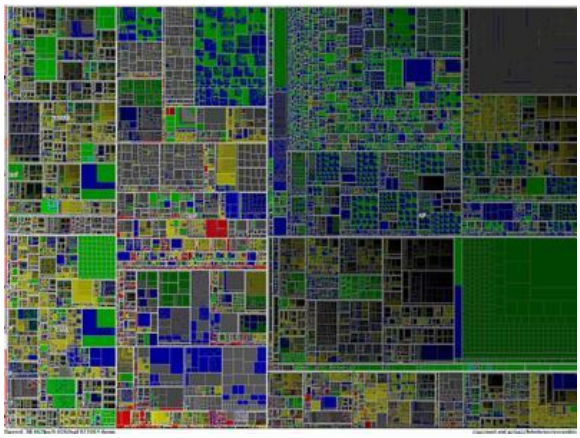

Fig. 1 Visualizing a document repository with a Tree-map [28].

Content-based visualization approaches rely on the extraction and processing of the full text of documents. As a representative example, building a self-organizing map [30] requires the analysis of their contents prior to displaying their topics. Such maps are divided in areas representing labeled topics (e.g., 'courses' in the middle of Figure 2). The color gradient on the map varies according to the number of documents associated with each topic. In the same way as with a Tree-map, the user can focus on any area of the map to view a detailed sub-map of the selection. Using this technique for visualizing the PISs of a community would provide members with a global view of its topics. Boyer et al. [31] build on this idea, also offering people to identify owners of the selected documents, thus enabling one to browse documents as well as people in a mixed way. In addition, instead of displaying a unique dimension which characterizes documents (either size or topics in previous examples) systems such as DocCube [32] or Tetralogie [33] make it possible to analyze documents regarding several attributes corresponding to their metadata (size, creation date, authors) or their contents (topics). Such systems require to master data analysis and data mining skills, which is hardly the case for regular knowledge workers.

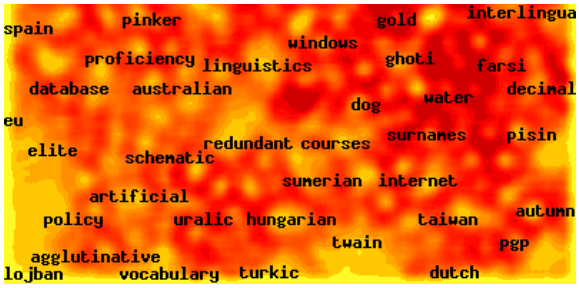

Fig. 2 Self-organizing map generated with the WEBSOM system [34].

This section outlined approaches providing global visualization of the documents owned by any community, mostly by exploiting their topics (extracted from their contents). Unfortunately, various other relations between information and knowledge workers are not considered. Due to the heterogeneity of users and their tasks, it would be worth to allow users to identify such alternative relations. For that purpose, we propose to merge the faceted search approach with visualization techniques. The underlying benefit is twofold. First, faceted search allows the users to see many facets of information without requiring data mining skills. Second, visualization techniques allow the users to encompass a large amount of information at the same time. Moreover, they can show relations among information pieces in a more understandable way. The faceted visual interface that we propose considers the various facets of information. In the remainder of the article, without any loss of generality, we focus on two facets: information contents, and the way knowledge workers organize it in the community for daily purposes. Indeed, there is no way for them, however, to identify documents most related to a given one. We hypothesize that documents filed together in a same folder (or close folders) by several knowledge workers share a common usefulness, even if they share no term in common. For example, a group of documents may invariably be used together by members carrying out daily tasks, even if they share no common contents. In such situations, content-based approaches fail to capture the implicit similarity between these documents. We claim that mining such associations reveals rich, semantic relations between (groups of) documents. The originality of our proposal is threefold:

1. It intends to make the community's IS profitable based on a win-win principle, by pooling and exploiting the existing PISs, which currently benefit their owners only. Moreover, it provides real-time access to the knowledge and documents of the community.
2. The proposed interface is designed for any kind of member in the community (and not restricted to data mining experts only): knowledge workers, as well as workforce managers from the Human Resources Management department, for instance.
3. Besides the exploitation of document contents, which is the classical Information Retrieval (IR) approach, we put forward the identification and use of implicit organization-based relations tying documents together, as knowledge workers do in their daily use.

Prior to presenting such a faceted visual interface in Section 5, we first define (in Section 3) and evaluate (in Section 4) the measures used for assessing how similar documents and individuals are (i.e., document-document, and user-user similarity measures).

3 Inter-document and inter-user similarity measures

In this section, we present the two similarity measures used for computing the various facets of the interface. We consider the classical content-based approach in Section 3.1. Then, we introduce our original organization-based approach in Section 3.2.

3.1 Classical content-based approach

Evaluating inter-document similarity is a core IR process, which usually depends on document contents [35, ch. 2].

First, terms are extracted from documents during the indexing process, which has been extensively investigated by the IR community [35, 36]. It is generally comprised of the four following steps: 1) segmenting the document contents by splitting it into words; it is dependent on the document format, 2) eliminating 'stop words' intends to exclude those which would not help an individual to distinguish a document from the rest of the documents in the collection. This is a language-specific step involving the knowledge of articles, determiners, and function words, among other categories. Then, 3) changing a word (possibly conjugated) into its canonical form is referred to as stemming (e.g., using the Porter's [37] algorithm for English). 4) The last step consists in counting the number of occurrences of each term appearing in each document.

Second, given the extracted terms, several mathematical models were proposed for computing the similarity of two given documents. The Vector Space Model [38] is the most prominent; each document is modeled as a vector in the vector space induced by the distinct terms of the document collection (i.e., documents from the PISs in the present case). Hence, any document d_i is represented as $\mathbf{d_i} = (w_i^1, \ldots, w_i^n)$, where each $w_i^j \in \mathbb{R}_+$ is the weight of the jth term in the document d_i, provided that n is the number of terms in the collection. This weight usually depends on two factors: the term's relative frequency tf_i^j in the document d_i, and the inverse of its frequency idf^j in the collection. Combining these two factors, such as $w_i^j = tf_i^j \cdot idf^j$ provides a numeric value, which is high if the term is frequent in the document and, at the same time, rare in the corpus. Finally, the similarity between two documents denoted d_1 and d_2 is computed by a function applied on the two vectors (e.g., $\cos(\mathbf{d_1}, \mathbf{d_2}) \in [0,1]$). The interested reader may refer to [36, ch. 6] for a detailed review of published variations on this general scheme.

As opposed to examining document contents for assessing their similarity, we propose in the next section to compare documents according to how knowledge workers organized them.

3.2 Proposed organization-based approach

In this section, we formalize an *inter-document organization-based measure* that is computed from individual hierarchies of documents. As people find and exploit interesting documents for realizing their tasks, they store them for various purposes. Their favorite organization is a hierarchy of folders [39] because it reflects how documents are related regarding people's tasks. When storing a document, the act of deciding which folder is best representative or even creating a new one from scratch is a highly cognitive task [9]. In spite of the involved cognitive efforts, hierarchical organization is appreciated because it allows individuals to keep documents under control [6]. Moreover, when a lexical-based measure is static because based on contents only, our organization-based measure is dynamic as it relies on evolving hierarchies. To sum up, hierarchical document organization conveys a high value that is mostly unexploited by the current approaches presented above. That is why we introduce a model for document organization, and the associated similarity measures in the following sections.

3.2.1 Modeling document organization with a multitree

In order to identify patterns of document organization, we need to represent the document hierarchy of several users into a unique data structure (excluding folders that users consider as 'miscellaneous'). Following previous works [40], we model these hierarchies using a *multitree* that groups together users' documents along with their paths, see Figure 3.

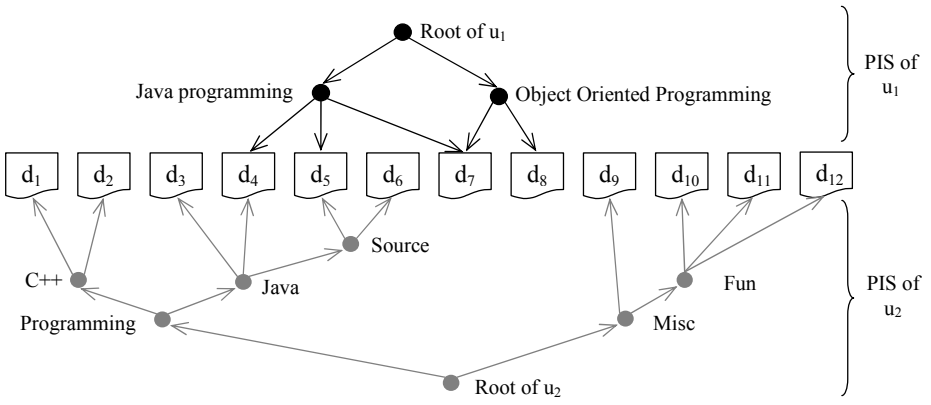

Fig. 3 A multitree comprised of two users' hierarchies (PISs).

Definition 1 A multitree $\mathscr{M} = \langle D, F, U, R_D, R_F, R_U \rangle$ is a sextuplet where $D = \{d_1, \ldots, d_n\}$ is a set of documents, $F = \{f_1, \ldots, f_m\}$ is a set of folders and $U = \{u_1, \ldots, u_l\}$ is a set of users. Moreover, we define the following relations:

- R_D is a binary relation on $D \times F$ that is called *document membership*. The pair $(d_i, f_j) \in R_D$ means that the document d_i is a direct child of the folder f_j.
- R_F is a binary relation defined on $F \times F$ that is called *folder membership*. The pair $(f_i, f_j) \in R_F$ means that the folder f_i is a direct child of the folder f_j.
- R_U is a binary relation defined on $U \times F$ that is called *root membership*. The pair $(u_i, f_j) \in R_U$ means that the user u_i owns the root of his hierarchy f_j.

Furthermore, we define $R_F^+ : F \to F$ as the function (1) that returns the direct parent folder p that contains a given folder f. If f is one of the roots of the multitree then $R_F^+(f) = \lambda$, where λ represents the null value

$$R_F^+(f) = p \mid \exists (f, p) \in R_F. \tag{1}$$

Definition 2 Let $\mathscr{G}$ be the graph associated with the multitree $\mathscr{M}$. A vertex of $\mathscr{G}$ is either a node (representing a folder) or a leaf (representing a document) while $R_D \cup R_F$ are edges of $\mathscr{G}$. A *path* from a root r to d is a sequence denoted $/r/f_1/f_2/\ldots/f_k/d$ such that $f_1 R_F r, f_2 R_F f_1, \ldots, d R_D f_k$. The direct descendant folder $f_1 \in F$ of the root r is called a *branch*. It is formally defined by the $b : F \to F$ function (2)

$$b(f) = \begin{cases} \lambda & \text{if } R_F^+(f) = \lambda \\ f & \text{if } b(R_F^+(f)) = \lambda \\ b(R_F^+(f)) & \text{else.} \end{cases} \tag{2}$$

Thanks to the multitree data structure, we compute an inter-document organization-based similarity. We detail how this is achieved in the following section.

3.2.2 Inter-document organization-based similarity

In this section, we detail an inter-document organization-based similarity which considers users' patterns of organization reflected by their document hierarchies (Definition 4). This depends on inter-folder similarity (Definition 3).

Definition 3 Following up works on URL similarity [41], we define the $\sigma_F : F^3 \to [0,1]$ function (3) that evaluates the organization-based similarity of two folders. The depth and number of common ancestors of two given folders are the two main criteria observed for evaluating their similarity, such that

$$\sigma_F(b, f_1, f_2) = 1 - \frac{s(f_1, m(f_1, f_2)) + s(f_2, m(f_1, f_2))}{s(f_1, b) + s(f_2, b) + 2}. \tag{3}$$

The $s : F^2 \to \mathbb{N}_+$ function[2] (4) returns the number of 'steps' (i.e., edges in the path from f_1 to f_2) that are assumed to be in a common branch b, such that

$$s(f_1, f_2) = |a(f_1) \ominus a(f_2)|. \tag{4}$$

To do that, we define the $a : F \to F$ function (5) that returns the set of ancestors of a given folder f (f included), such that

$$a(f) = \begin{cases} \emptyset & \text{if } f = \lambda \\ \{f\} \cup a(R_F^+(f)) & \text{else.} \end{cases} \tag{5}$$

Moreover, the $m : F^2 \to F$ function (6) returns the lowest common ancestor of two folders f_1 and f_2 (i.e., the folder that is an ancestor of both f_1 and f_2 and that has the greatest depth), such that

$$m(f_1, f_2) = f \mid \forall (f, f') \in (a(f_1) \cap a(f_2))^2 \\ \left(f \neq f'\right) \wedge \left(d(f) > d(f')\right). \tag{6}$$

This relies on the $d : F \to \mathbb{N}_+$ function that gives the depth of a folder.

Our purpose is to evaluate inter-document organization-based similarity. Remembering that documents of the multitree come from at least one user's hierarchy, we identify

[2] The $\ominus$ operator is the *symmetric set difference*, corresponding to the exclusive OR (XOR) in Boolean logic: $A \ominus B = (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$.

common patterns in document organization. For example, if people always classify a group of documents in a common folder or in the similar way, this means that people find them similar, for any reason related to their usage [6]. Thus, we have to observe repeated patterns of document organization: the more people organize the same collection of documents in the same manner, the more these documents are deemed to be organization-based similar.

Definition 4 The $\sigma_D : D^2 \rightarrow [0,e]$ symmetric function (7) computes the organization-based similarity of two documents, provided that they are reachable from at least one branch in the multitree, such that

$$\sigma_D(d_1,d_2) = \frac{\exp\left({}^{u}/_{|U|}\right)}{|B|} \sum_{b \in B} \sigma_F(b, R_D^+(d_1,b), R_D^+(d_2,b)). \quad (7)$$

Since their insertion in a common branch is the consequence of human cognitive effort, we deduce that they share a common semantics according to their owner's need. In (7) the $R_D^+ : D \times F \rightarrow F$ function (8) returns the folder f that contains a given document d, provided that f is in the branch b, such that

$$R_D^+(d,b) = f \mid (\exists (d,f) \in R_D) \wedge (b \in a(f)). \quad (8)$$

Moreover, $B = b(d_1) \cap b(d_2)$ is a set of branches that both d_1 and d_2 have in common. Finally, u is the number of users that have a branch both containing d_1 and d_2. The $\exp\left({}^{u}/_{|U|}\right)$ factor models the fact that the more people classify two given documents in the same branch, the more these documents are organization-based similar. The rightmost part of (7) takes into account the average distance of folders containing the two documents.

3.2.3 Inter-user organization-based similarity

For computing the inter-user similarities, we build on the mega-document approach, as proposed in [42]. It consists of representing a user as a single document, made from the concatenation of his documents extracted from his PISs.

Definition 5 Let us consider the $d : U \rightarrow F^*$ function defined on R_D, R_F et R_U, which retrieves a user's documents. Prior to computing the organization-based similarity between the two users u_1 and u_2, we define the following sets:

- $D^{\cap} = d(u_1) \cap d(u_2) = \left\{d_1^{\cap}, \ldots, d_k^{\cap}\right\}$ contains the documents $d_i^{\cap}$ that u_1 and u_2 own in common.
- $D^{\ominus} = d(u_1) \ominus d(u_2) = \left\{d_1^{\ominus}, \ldots, d_l^{\ominus}\right\}$ contains the documents $d_i^{\ominus}$ owned by u_1 or (exclusive) by u_2, but not by u_1 and u_2 at the same time (note that these constitute the $D^{\cap}$ set).

From the organization-based inter-document similarity σ_D, we define the organization-based inter-user similarity (9) as the symmetric function $\sigma_U : U^2 \rightarrow \mathbb{R}_+$, such that

$$\sigma_U(u_1,u_2) = f\left(\sum_{i=1}^{k}\sum_{j=i+1}^{k} \sigma_D\left(d_i^{\cap}, d_j^{\cap}\right), \sum_{i=1}^{l}\sum_{j=i+1}^{l} \sigma_D\left(d_i^{\ominus}, d_j^{\ominus}\right)\right). \quad (9)$$

Notice that the sum operators (indexes i and j) are initialized with respect to the symmetric characteristic of the σ_D function (i.e., computing both $\sigma_D(x,y)$ and $\sigma_D(y,x)$ is unnecessary). In (9), the $f(x,y)$ function $f : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$ is increasing on x and on y. This intends to increase the σ_U value when organization relations are strong between the documents owned by both u_1 and u_2. It models the fact that two people are all the more organization-based related when they own the same documents, as well as when they organize them in a similar way. A potential instantiation is $f(x,y) = (y+1) \cdot e^{(x+1)}$, for favoring organization relations due to documents owned in common (x) with respect to those that are owned by only one person among the two people considered (y).

Unlike classical content-based approaches, the proposed organization-based measure does not require the extraction document contents. Moreover, it evaluates how closely individuals organize documents, instead of measuring how close their contents are (e.g., as in the Vector Space Model, see Section 3.1). We investigate the complementarity of these two approaches in the next section.

4 Evaluating the complementarity of content-based and organization-based similarities

In this section, we state and test the following hypothesis: organization-based similarities are different from classical content-based similarities (otherwise they are pointless and deserve no more attention). We validate our hypothesis on the example of Figure 3, and then on two larger test collections.

The two graphs in Figure 4 represent the computed similarities on the example of Figure 3. We may comment the observed differences [43]. On the one hand, we distinguish two separate document clusters on the organization-based view (Figure 4a). Each cluster gathers documents that are close to each other regarding the organization criterion. These clusters do not appear on the content-based view (Figure 4b) since all documents share at least one (frequent) word. This is likely to happen with large documents from a community. On the other hand, we notice that the most organization-related documents d_4 and d_5 are not so close regarding content-based similarities. Moreover, d_1 and d_{10} seem to be simi-

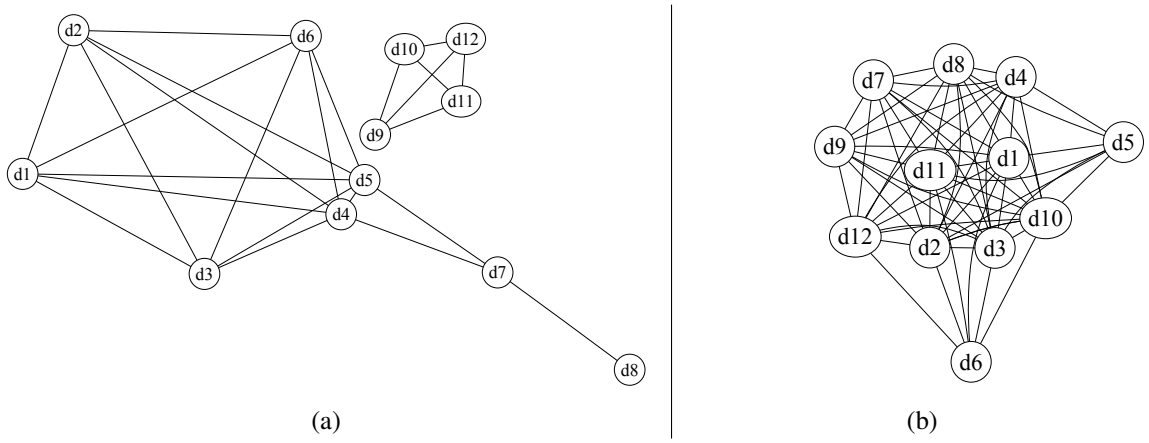

Fig. 4 Comparing inter-document *organization*-based (4a) similarities with *content*-based (4b) similarities. Edge length is inversely proportional to the similarity between the vertices representing documents d_1 to d_{12}.

lar regarding their contents whereas they are not used in the same way: they belong to distinct clusters in the organization-based similarities. To sum up, the proposed organization-based view enables people to discover yet unidentified document relations (i.e., by using content-based similarities only). These relations are valuable implicit knowledge.

We confirmed these empirical observations with two experiments completed on two test collections. We first detail the protocol, then report the results, and comment our findings.

4.1 Hypothesis: organization complements contents

We proposed a new inter-document similarity measure. For being useful and relevant, it should not be redundant compared to the classical content-based similarity measure. Therefore, the hypothesis to test is: organization-based similarity values are different from content-based similarity values.

4.2 Test collections

In order to check the aforementioned hypothesis, we designed an experiment that we realized with two test collections. First, we opted for the TREC OHSUMED dataset, as it is a standard benchmark used in IR that is publicly available. Second, we collected the PISs from a community of knowledge workers (i.e., real end-users).

4.2.1 TREC OHSUMED

Computing organization-based similarities require documents to be organized in a hierarchy. This led us to consider the OHSUMED test collection, as featured in TREC-9 [44]. It is 'a set of 348,566 references from MEDLINE[3], the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987–1991)'[45]. Every document was associated with at least one category of the MeSH[4] hierarchy. This information was used as query relevance judgments (qrels) in TREC. The OHSUMED test collection is separated in a training collection, and a testing collection. We retained only the sub-collection corresponding to cardiovascular diseases in MeSH. This sub-hierarchy has a depth of six nodes, and comprises 146 nodes overall. We thus conducted our experiment with the associated the $n = 4{,}974$ documents.

4.2.2 PISs of a community of knowledge workers

TREC OHSUMED [44] is made up of a single large hierarchy organizing many documents. In order to get closer to the PISs of a community, we intended to consider many hierarchies produced by several individuals for organizing the documents that they encountered or created. Thus, as a complementary test collection to TREC OHSUMED, we considered the PISs of a community. This comprised 14 knowledge workers involved in the group researching Information Systems in our lab. There were 10 males and 4 females, from 24 years old to 41 years old. Regarding their occupations, there were 1 intern, 2 masters students, 6 PhD students, and 5 professors. They agreed to disclose their bookmarks (as stored in their web browser) for the purpose of the present research. We indexed the $n = 4{,}176$ corresponding documents for computing content-based similarities. In addition, we modeled the 14 PISs as a multitree for computing organization-based similarities. On average, a PIS was comprised of 35 folders (median = 22) organized in a hierarchy 3 levels (median = 3) for storing 286 bookmarks (median = 145).

[3] Medical Literature Analysis and Retrieval System Online.

[4] Medical Subject Headings.

4.3 Methodology

Prior to comparing similarity measures, we computed them as follows for both test collections:

1. Regarding the content-based similarity, we indexed the n documents (see Section 3.1) from the considered test collection. Then, we instantiated each document as a vector in the Vector Space Model. Finally, we computed paired content-based similarities $s_C(d_i,d_j)$ according to $\cos(\mathbf{d_i},\mathbf{d_j}) \in [0,1]$.
2. Regarding the organization-based similarity, we instantiated the hierarchy corresponding to the n documents in the multi-tree data-structure (see Section 3.2.1). We computed paired similarities $s_O(d_i,d_j)$ with the function $\sigma_D \in [0,e]$ (Equation 7) that we normalized in the $[0,1]$ range. Normalization allows the comparison between s_C and s_O values afterwards.

According to the symmetrical property of the considered similarity measures, we computed content-based $s_C(d_i,d_j)$ and organization-based $s_O(d_i,d_j)$ similarities of the documents, such as $1 \leqslant i < j \leqslant n$. We obtained pairs such as $(s_C(d_i,d_j), s_O(d_i,d_j))$.

For comparing the results of the similarity measures, we used two statistical significance tests, namely Student's t test and Wilcoxon's signed rank test on paired samples [46]. These compute the p-value of significance, which is interpreted as follows: $p < \alpha$, where $\alpha = 0.05$ points out evidences of a statistically significant difference between the two tested samples. Such difference is higher when $p \to 0$. Moreover, Pearson's $r \in [-1,1]$ product-moment correlation coefficient allows the assessment of the correlation between the two samples. These are all the more similar when $r \to 1$.

4.4 Results

The following sections report our findings with respect to both test collections.

4.4.1 Experiments with TREC OHSUMED

We first plot the two similarity measures. Hull [46] underlines that plots enable the visual comparison of samples, allowing us to check the data distribution. In Figure 5, we show the distribution of s_C (i.e., content-based similarity measure). Similarity values do not vary much, since they all belong to $x \in [0.0, 0.2]$. We notice a concentration (40%) of the values around the zero value.

Organization-based similarity values are shown in Figure 6. Notice that, compared with Figure 5, the values are differently distributed. They are comprised in $y \in [0\%, 12\%]$ for organization-based and $y \in [0\%, 40\%]$ for content-based values. The s_O values are more diverse on the x axis contrary to s_C values.

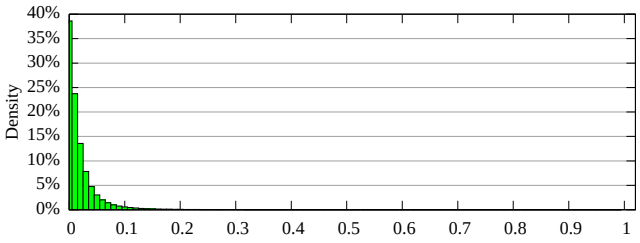

Fig. 5 Distribution of the s_C content-based similarity values.

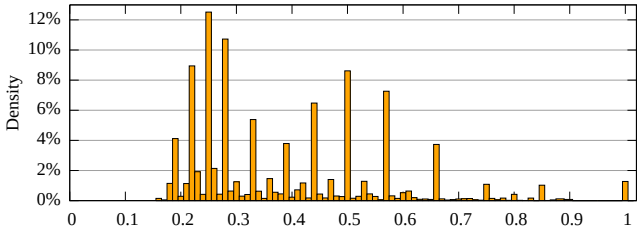

Fig. 6 Distribution of the s_O organization-based similarity values.

Finally, we show in Figure 7 the distribution of the differences between content-based s_C values and organization-based s_O values. We notice that this distribution is neither centered on zero nor it fits a Normal distribution. Similar samples of data, however, would have fit a normal distribution since their paired differences are close to zero. As a result, the two tested similarity measures s_C and s_O are not identical. In addition, the $s_C - s_O$ differences are mostly negative. This means that $s_C < s_O$, which is in line with the observations commented beforehand.

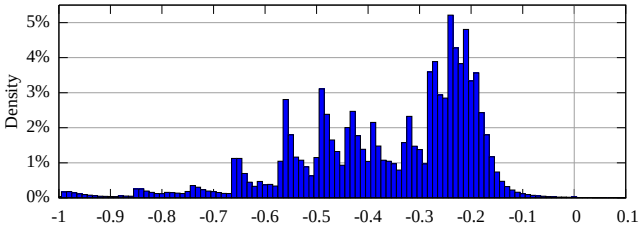

Fig. 7 Distribution of $s_C - s_O$ between the two similarity measures.

For complementing these observations, we show in Table 1 statistics related to the two series (s_C and s_O) and their differences ($s_C - s_O$). If the s_C and s_O were identical, the average and standard deviation of their differences would tend towards zero, which is not the case. The minimum organization-based similarity (0.166) is greater than zero, which is expected since the documents from the corpus are all in the same MeSH branch. Moreover, the range

of s_O is larger than the range of s_C because its standard deviation (0.172) is larger than the one of s_O (0.035).

Table 1 Statistics related to the three s_C, s_O, and $s_C - s_O$ series.

Variable	Average	StdDev	Minimum	Maximum
s_C	0.024	0.035	0.000	1.000
s_O	0.392	0.172	0.166	1.000
$s_C - s_O$	−0.368	0.170	−1.000	0.797

Regarding the hypothesis under experiment, both Student's and and Wilcoxon's tests conclude that there is a statistically significant difference between the two series ($p = 0.000$). Pearson's $r = 0.154$ coefficient also shows lack of correlation between organization-based and content-based similarity measures. These observations show that the two measures are neither identical nor similar, but they complement each other.

4.4.2 Experiments with folder hierarchies of end-users

In this section, we report the results of our study on the 14 PISs. The distribution of s_C is plotted in Figure 8. Similarity values do not vary much, since they belong to $x \in [0.0, 0.3]$. We notice a concentration (18%) of the values around the zero value. We notice that these content-based similarities vary more than those in the according graph for OHSUMED (Figure 5). This may be due to the topics of documents covering a single field in the test collection built from OHSUMED, whereas they cover a larger range of topics in the PISs.

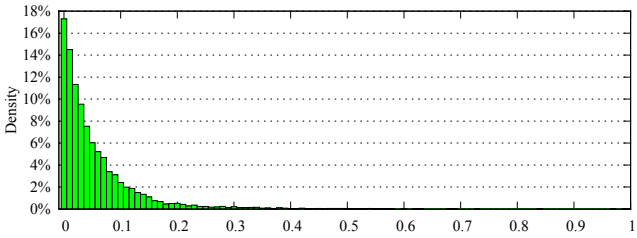

Fig. 8 Distribution of the s_C content-based similarity values.

Organization-based similarity values are shown in Figure 9. For ensuring similar conditions to the experiment with TREC OHSUMED, we only retained document pairs where $s_O(d_i, d_j) > 0$. There is again a difference between the distribution of these values and the content-based similarity values (Figure 8). Organization-based similarity values are comprised in $y \in [0\%, 40\%]$ *versus* $y \in [0\%, 18\%]$ for content-based values. The s_O values are more diverse on the x axis contrary to s_C values. A few x values are particularly frequent. These correspond to various frequent configurations of two documents with respect to their relatedness in the multitree: siblings, cousins, and so on.

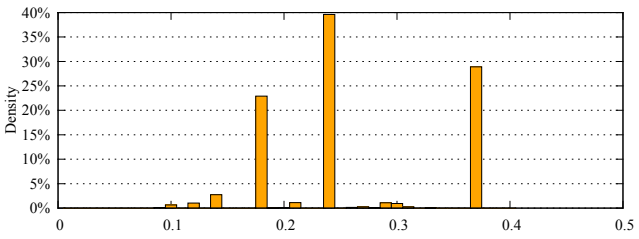

Fig. 9 Distribution of the s_O organization-based similarity values.

Finally, the distribution of the differences between content-based s_C values and organization-based s_O values is plotted in Figure 10. Again, we notice that this distribution is neither centered on zero nor fitting a Normal distribution. Similar samples, however, would have fit a normal distribution since their paired differences are close to zero. As a result, the two tested similarity measures s_C and s_O are not identical. In addition, the $s_C - s_O$ differences are mostly negative. This means that $s_C < s_O$, which is in line with the observations commented beforehand.

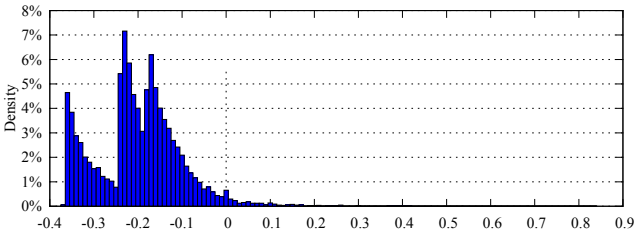

Fig. 10 Distribution of $s_C - s_O$ between the two similarity measures.

For complementing these observations, we show in Table 2 statistics related to the two series (s_C and s_O) and their differences ($s_C - s_O$). The minimum organization-based similarity (0.082) is greater than zero, which is expected since the documents from the corpus are all in the same PIS branch. Moreover, the range of s_O is larger than the range of s_C because its standard deviation (0.078) is larger than the one of s_O (0.048).

Regarding the hypothesis under experiment, both Student's and and Wilcoxon's tests conclude that there is a statistically significant difference between the two series ($p = 0.000$). Pearson's $r = 0.082$ coefficient also shows lack of correlation between organization-based and content-based similarity measures. These observations show that the two

Table 2 Statistics related to the three s_C, s_O, and $s_C - s_O$ series.

Variable	Average	StdDev	Minimum	Maximum
s_C	0.033	0.048	0.000	1.000
s_O	0.236	0.078	0.082	0.389
$s_C - s_O$	−0.006	0.040	−0.390	1.000

measures are neither identical nor similar, but they complement each other.

Complementary to the previous experiment with TREC OHSUMED data [44], this experiment with 14 PISs also supports the tested hypothesis: organization-based and content-based similarity yield different results, which may complement each other. We harness this complementarity in the faceted visual interface that we propose in the next section.

5 Faceted visual interface to explore the capital of a community

The proposed interface designed for community members is intended to meet operational and strategic requirements:

- The *operational* adjective refers to completing the various tasks assigned to each member. In this context, the interface provides a global view of the resources belonging to the community and the way they are organized. Exploration regarding their contents and their organization is supported. These two modalities build on similarity measures which complement each other. Using the proposed interface based on these two similarity measures, one can extract from the community PISs the documents most related or complementary to his own documents. Hertzum and Pejtersen [47] underline how useful it is for knowledge workers to search for documents for finding the associated people, and *vice versa*. As a way to meet this need, the interface allows one to weave between these two dimensions (people and documents).
- The *strategic* adjective refers to the activities carried out by the Human Resources Management department. In this context, the proposed interface exploits the PISs for displaying information related to the knowledge workers' activities involving documents. A use case may consist of identifying the documents used to achieve the tasks assigned to a given position. Considering this information will help to map key specialists with the relevant skills and knowledge, to constitute the suited group work for a given project, to identify emergent interests, and to limit turnover issues by anticipating skills to be renewed. In the same vein, Boyer et al. [31] proposed to make a map for the community, which contains document contents and relations among members of the community. We intend to complement their approach with the notion of document *organization*, allowing the identification of complementary links between documents according to how they were classified in the PISs.

The next section presents the faceted visual interface by first describing its various components giving access to the community resources (static aspect). Then, we formalize interactions offered to users (dynamic aspect), which make it possible to explore the community resources through the facets. Finally, we comment on implementation details: inter-document *content*-based and *organization*-based similarity measures, and visualization techniques.

5.1 Static aspect: representing the community resources

The faceted visual interface offers the exploration of two complementary corporate materials: documents, and community members. Using this interface, one can explore several items (a group of documents or people) or a single item (a document or a person). These exploration scenarios are supported by four *views*, as illustrated in Figure 11. Each view contains a major visualization displaying the selected item(s). It is complemented by three other visualizations giving complementary information about the item(s). These visualizations are called *facets*, since each one provides a specific point of view over the item under study.

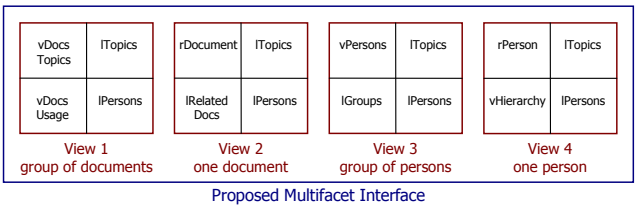

Fig. 11 Facets of the four views features in the faceted visual interface.

Considering each of the four views in turn, we list in Table 3 the available facets. We distinguished three kinds of facets for displaying information: visualization, list, and record. Note that each name of facet starts with the associated type first letter (v, l, or r).

The next sections detail each of the four views, describing the information displayed in each facet according to a concrete use case. For better understanding, we refer the reader to Figures 16-19 for an illustration of the views as they may be implemented.

5.1.1 View 1: representing a group of documents

The facets in view 1 (see also Figure 16) display the community documents gathered by topic (vDocsTopics) or by organization (vDocsUsage). These two modalities are respectively based on document contents and organization within

Facets		Dimensions			
Name	Description	Document		Person	
		Group *View 1*	Unity *View 2*	Group *View 3*	Unity *View 4*
rDocument	Record of a document		+		
rPerson	Record of a person				+
lRelatedDocs	List of related documents		+		
lTopics	List of topics	+	+	+	+
lGroups	List of groups			+	
lPersons	List of people	+	+	+	+
vDocsTopics	Document topics view	+			
vDocsOrganization	Organization-based view of documents	+			
vHierarchy	The document hierarchy of a person				+
vPersons	Representation of persons			+	

Table 3 Details of the *facets* comprised in the four *views* of the faceted visual interface.

the PISs. Intuitively, two documents are *content*-based similar if they share several terms. In addition, two documents are *organization*-based similar if they are often classified together in the PISs. Details about content-based, as well as organization-based similarity measures were given in Section 3. The user interacts with these facets by selecting one or more documents, being able to focus on them afterwards. The facet called lPersons displays the list of people who own the selected documents. It may be sorted by document number. Finally, the lTopics facet lists the topics related to the selected documents, in order of importance.

This view offers community members a representation of the topics extracted from the documents stored in the PISs. Focusing on a specific topic allows the visualization of people whose document topics match. Moreover, the documents classified with the chosen ones are also presented. These bring additional information related to the user's initial selection. As an evidence of knowledge workers' association of ideas, the vDocsOrganization facet allows us to make the most of the cognitive efforts spent by each knowledge worker.

5.1.2 View 2: representing one document

View 2 (see also Figure 18) displays the record of a document (rDocument) giving access to its title, contents, and absolute paths in the according PISs (i.e., `/home/userX/computer_science/relational_databases/indexing/bTree/lecture.pdf`, and `/home/userY/inventors/science/cs/Rudolf_Bayer/bio.html`. Note that the creation date of the document in each path is given. The document topics are listed in the lTopics facet; its owners are listed in the lPersons facet. Lastly, related documents (organized with the one under study) are listed in the lRelatedDocs facet, ordered by similarity of organization.

The facets of this view allow a given document topic to be displayed. They also give access to the people (knowledge workers) who chose to keep it in their PIS. The absolute path names provide extra evidence about how the given document is used by end-users for real tasks. As a user identifies the group of people interested in this document, it is then possible for him to explore their PIS for retrieving other interesting documents. The user may also get in touch with these knowledge workers directly. This use case meets knowledge workers' needs highlighted by Hertzum and Pejtersen [47]: finding documents for reaching people, and *vice versa*.

5.1.3 View 3: representing a group of persons

View 3 (see also Figure 19) represents in the vPersons facet a set of persons and the ties bounding them together, may they be organization- or content-based. This facet emphasizes link visualization. It is complemented by the lPersons facet that lists the displayed persons. Since community members are part of explicit groups (teams, work groups, committees, and so on) we use this information in the lGroups facet. It is comprised of names and member count for each distinct group corresponding to the visualized persons in vPersons (e.g., 'Sales dept. (12)'). Lastly, the lTopics facet lists the topics corresponding to the PISs owned by the displayed persons, sorted by number of associated documents.

This view helps in identifying the topic interests characterizing any group of persons, being either explicit (a team stated in the community chart) or tacit (people getting on well, having lunch together, etc.). Consequently, any knowledge member can identify and explore his team's topics. This may be of great help when it comes to introducing a new partner, who has to adapt quickly and to assimilate the common topics of the hosting team. Similarly, the knowledge of a team's main topics as extracted from its members'

PISs may help the Human Resources Management department in issuing a job record. Such information may be especially profitable when creating or renewing a position in a corporate context.

5.1.4 View 4: representing one person

The record of a person (rPerson) in view 4 (see also Figure 17) is comprised of the following information: identity (name and surname) and the groups he/she is affiliated to. The hierarchical display of his/her documents as structured in his/her PIS is available through the vHierarchy facet. The list of the deduced topics is displayed in the lTopics facet, in alphabetical order or in decreasing order of importance. Lastly, the lPersons facet lists the persons who share similar topics, or who use (filed) documents in the same way as the one represented in this fourth view.

A real use case, for a given user, may consist in viewing his/her own record to identify the knowledge workers close to him/her, regarding either document topics or document organization. Then, reading their records allows him/her to know their favorite topics. The further exploration of their PIS contents, according to both topics and structure, must improve the users' perception of nearby skills and knowledge, which remains hidden and implicit otherwise.

5.2 Dynamic aspect: exploring the community resources

The proposed interface displays the resources of a community regarding the four aforementioned views. This section defines two kinds of interaction (namely *inter-view* and *intra-view* interactions) offering the user ways for exploring this information capital.

Whatever the view, *intra-view* interaction aims at passing the user selection of items in a facet to the three other facets in the same view. For instance, selecting a set of topics associated with a person (facet lTopics of view 4, see also Figure 17) allows the immediate identification of these topics within the person's PIS (vHierarchy facet). In addition, one can view the people who share the same topics as the selected person (lPersons facet). Alternatively, the user can issue a query with keywords and boolean operators for selecting the corresponding items. In practical terms, each facet highlights the items corresponding to the resulting selection with a suitable layout (specific color, font weight, and so on). Intra-view interaction allows the identification of a given item in all the facets constituting a view, knowing that these facets offer further information compared to what is already extracted from the PISs.

The second kind of interaction, called *inter-view*, allows one to browse from one view to another. In practical terms, the interface replaces the currently displayed view by another one which better meets the needs expressed by the user's action on the former view. For instance, selecting a person in the vPersons facet of view 3 (see Figure 19) leads to the displaying of view 4 (see Figure 17) because it brings additional information about the selected person. In this way, the various actions carried out on facets allow the proper exploration of community resources.

We modeled the dynamics of the proposed interface in with a state diagram represented in Figure 12. The four states correspond to the four views detailed in Table 3. A transition from a state s_1 to another state s_2 is triggered by actions on a facet of the view corresponding to s_1. Details of these actions are given on labels associated with each transition (arrow) tying two states together. When multiple actions are possible, they are separated by a comma. An action is denoted $s(f)$ where s represents a selection and f a facet. More precisely, a multiple selection is denoted 'm,' the selection of a single item is denoted 's,' and '*' denotes a multiple or single selection. For instance, the label 'm(lPersons), *(rPerson)' between view 4 and view 3 means that a multiple selection in the list of persons m(lPersons) or a single selection in the person's record rPerson leads to the displaying of view 3.

A global view of the proposed faceted visual interface with possible interactions through its views is given in Figure 13, which synthesizes the static (Table 3) and dynamic (Figure 12) aspects. The various links connecting views together show the interactive aspect of the interface. This enhances browsing within the community resources. The next section details how the information displayed in each facet is extracted or created.

5.3 Designing the proposed faceted visual interface

Implementing the proposed faceted visual interface required us the modeling of input data from which content-based and organization-based similarities are computed. These are then displayed in the facets of the four views of the interface.

5.3.1 Modeling the IS components for our approach

The faceted visual interface does not intend to constitute a new source of information, but to enable one to explore and analyze knowledge workers' PISs. Our approach is thus based on using the existing community IS for extracting data related to the documents of its members. Personal folders and files are excluded from the processing for privacy concerns. Knowledge workers may declare some items as confidential by inserting a specific character string in its name (e.g., 'personal'). The strings to use can be defined at the community level. As a result, only non personal folders and files are considered. The UML conceptual model in Figure 14 shows the required data for computing the facets identified in Table 3. Once extracted, they are periodically

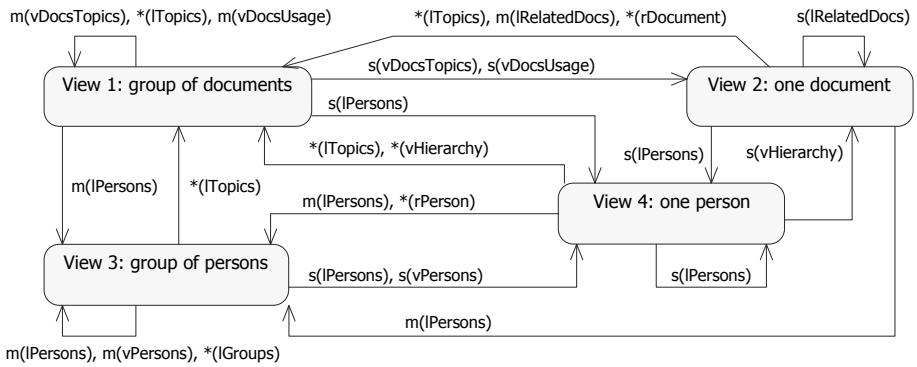

Fig. 12 UML State diagram representing the dynamic aspect of the faceted visual interface.

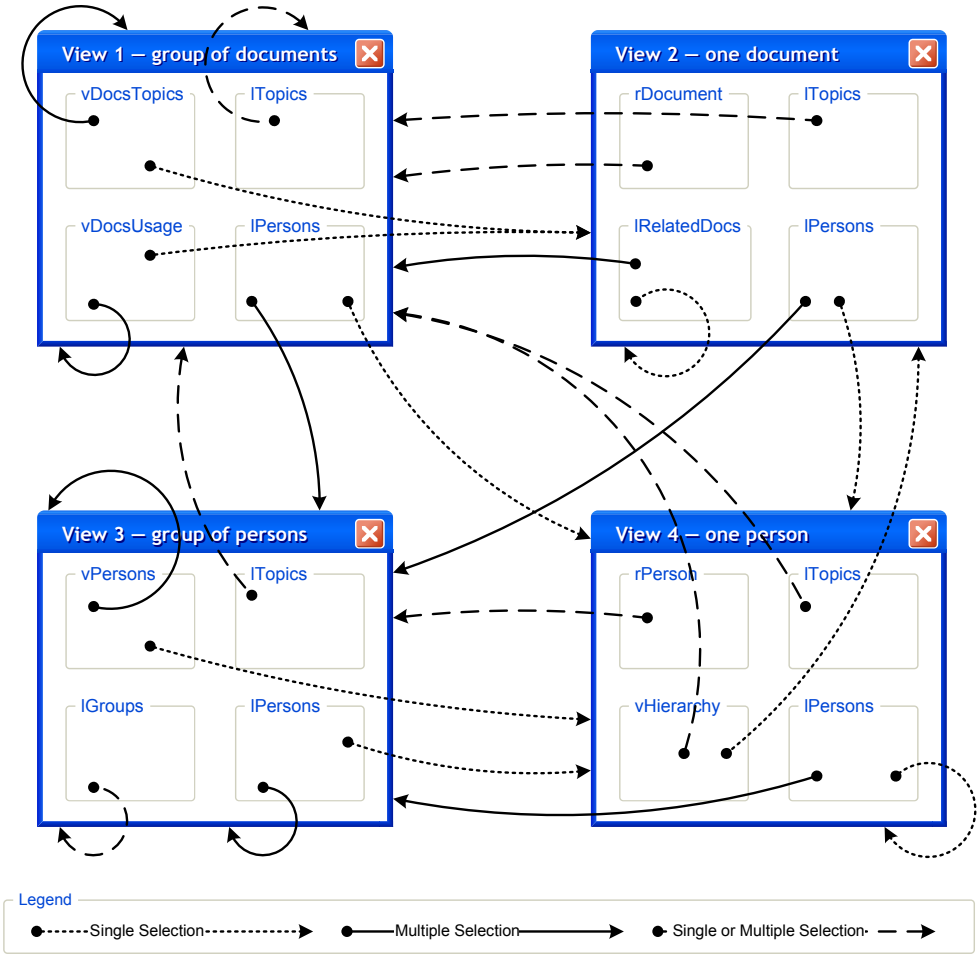

Fig. 13 Synthesis of the static and dynamic aspects of the proposed faceted visual interface.

updated for reflecting the current activities of the community. The final result of the indexing process, detailed in Section 3.1, is stored in the Index class, linked to the association between the Document and Term classes. Thus, the count attribute represents how many occurrences of a given term are present in a document.

The next section presents the twofold use of documents. On the one hand, we exploit document contents (represented by the Term and Index classes) and inter-document content-based similarity metrics (simC method). On the other hand, we benefit from how documents are organized (Folder class) through the simO method, reflecting the use of documents.

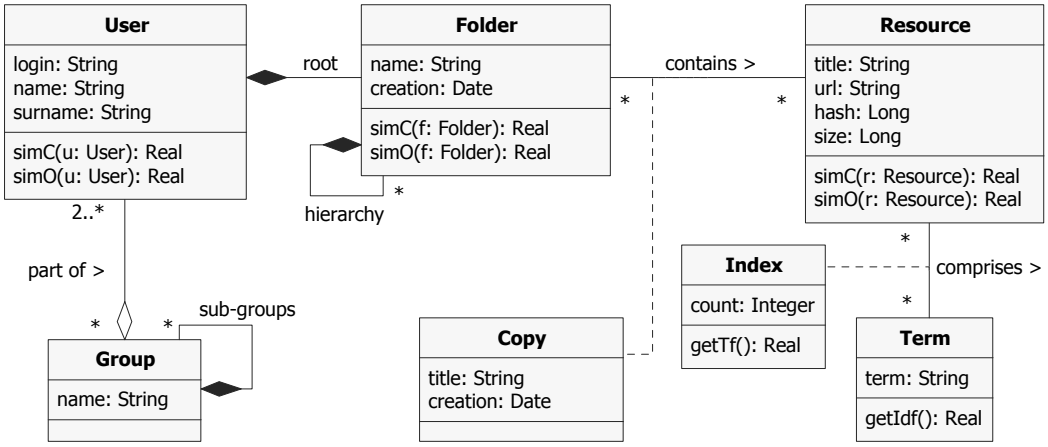

Fig. 14 UML Class Diagram modeling the data exploited by the faceted visual interface.

5.4 Visualization techniques for representing documents and people

As underlined in Section 2.3, the information visualization literature is very rich. There is a plethora of visualization techniques and tools as presented in [25, 26, 27]. We had, however, to choose appropriate visualizations for reaching the proposed aim (i.e., offering knowledge workers a global view of the community resources). Two main criteria were considered for that purpose. First, the adopted visualizations must allow the displaying of items regarding their similarity (topic or organization-based). Second, they must be able to display numerous items, especially if the community under study has many members. This is related to the scalability of the system. The contribution of this article is not related to the visualizations we adopted, but rather to the joint exploitation of content- and organization-based similarities for the proposed faceted visual interface. As a result, the adopted visualizations that we present in this section are for illustrative purposes; they may be reconsidered according to several criteria specific to each community under study.

We adopt a graph-based visualization for displaying organization links between documents or persons. This favors the identification of document groups jointly used (i.e., forming a connected graph). Nodes represent documents or persons; they are connected by edges whose lengths are inversely proportional to the similarity of the two connected documents. These edges are labeled with the absolute path extracted from the PISs where they come from; this helps individuals understand how they are linked together. A force-directed placement algorithm [48, 49] is used to draw the graph taking into account organization-based similarities. For instance, Figure 15 shows an illustrative (dezoomed) graph computed for the experiment reported in [43] and showed in Figure 4.

In addition, we opted for two other visualizations for displaying the document topics: the aforementioned self-organizing maps [30] and the hierarchical visualization, which is familiar to users. As shown in Figure 2, self-organizing maps emphasize on various topics existing in a document collection, as well as on their relative importance according to the number of documents per topic. In addition, we proposed to represent a set of documents by computing a hierarchy of folders organized by topic using the hierarchical agglomerative clustering algorithm [50]. This creates a binary tree whose depth is $n-1$ for n documents. For usability concerns, we reduce this deep hierarchy thanks to the slicing algorithm with parameterizable depth, as introduced in [51]. Finally, the resulting folders are labeled according to the most representative terms (i.e., highest $tf \cdot idf$ values) extracted from the documents they contain.

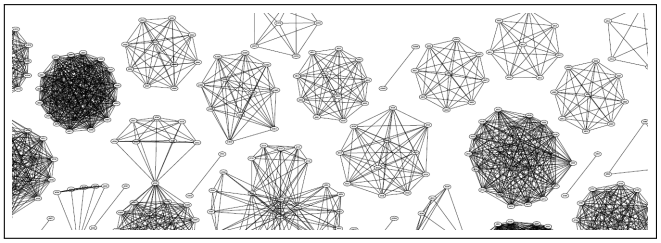

Fig. 15 Sample of the organization-based graph of documents drawn for the experiment we reported in [43].

6 Implementing the faceted visual interface & use case

The faceted visual interface we designed was implemented as a single window comprising four tabs: one for each view, as shown in Figures 16-19. Hence, the user constantly sees the active view. The interface was implemented in Java 6 with the Swing standard widget library. Data displayed in the interface comes from a relational database run on a Oracle server version 11. The dedicated Web page[5] allows any-

[5] `http://www.irit.fr/~Guillaume.Cabanac/MultiFacettes`

one to launch this software. It gives access to the resources of a demo community for illustrating the following scenario. Pierre, a knowledge worker, recently joined a Java development team. As a newcomer, he uses the faceted visual interface to gather materials on this subject. At start up, he gets view 1 in Figure 16, which displays the computed hierarchy of the documents owned by the members of his community (top left facet). These documents are also displayed according to their organization-based similarities (bottom left facet). Eager to first reach the documents judged as worth keeping by his team leader, Pierre accesses this person's record (Figure 17). It is comprised of his identity (Jean Dupont), the groups he is affiliated to, the topics extracted from his documents, his PIS structure, as well as the people sharing similar topics. Then, Pierre spots the JAVA folder in Jean's PIS; he focuses on Jean's document entitled *Liveconnect Java Javascript*. As a result, Pierre accesses view 2 (Figure 18) showing the document record with detailed title, URL, topics, and paths within the users' PISs where it was found. The documents jointly used with the selected one as well as their owners' logins are also presented to Pierre. Selecting two people on the fourth facet refreshes the interface for displaying view 3 (Figure 19). This enables Pierre to view both the topics and groups shared among the two chosen individuals. This exploration goes on and on, as Pierre steps from one view to another, and so on.

7 Discussion

Throughout this article, we argue in favor of designing an unintrusive interface, requiring no adaption from community members. We emphasized on this requirement as to limit users' resistance to change. However, their participation is desirable when they wish to keep some documents private, or restrain their access to a limited group of individuals for instance. Contrary to such a participation intending to reduce the amount of accessible resources, other users may like to grade their documents (e.g., on a 5-point scale). Such a feedback—possibly expressed as annotations anchored to documents—may well be used for highlighting the document passages marked as the most interesting by the community.

Regarding PIS exploration, we should comment on an important point: although providing real-time access to knowledge members and documents of a community, the proposed faceted visual interface does not directly allow expertise identification. Actually, even if someone keeps many documents about a given subject, this does not make him an expert yet. This observation only provides evidence about this person being interested in this subject. On the other hand, a true expert may not need to store in his PIS the documents he either assimilated long ago or he could find very easily in a different way [3].

In addition, a study of motives behind paper document archival shows that *building a legacy* comes second, just after the ability to re-find a document [52]. Hypothesizing that these motives apply to PISs, the proposed interface allows reasonable document sharing without requiring owners to change the way they work. It is likely that such a system based on a win-win principle (knowledge workers benefit from each other's PISs) will foster community members' active participation, as they will realize that their cognitive efforts spent on PIS management truly benefit the community in the broadest sense.

8 Related Works

In this article, we tackled the following research question: how can a community benefit from the documents painstakingly collected by its members with the passage of time? We claimed that these documents remain dormant in the hard drives of organizational members. Our contribution intends to turn this dormant capital into its full potential for the benefit of the community as a whole. We discuss its major characteristics in this section, and relate them to previous works from the literature.

8.1 Personal Information Management

Personal Information Management (PIM) 'refers to both the practice and the study of the activities people perform in order to acquire, organize, maintain and retrieve information for everyday use.' [53, 54]. Indeed, knowledge workers rely on multiple information repositories, such as electronic documents, emails, web pages. These are also referred to as information 'silos' [55]. PIM systems provide users with a unified way to access information extracted from these silos. Several PIM systems were developed, as evidenced by the list of 120+ systems provided at `http://pim.famnit.upr.si/pim/pimtools.html`. From the user perspective, there are two major access methods that are discussed in [56]:

1. Navigation 'which exploits structures the user has set up for retrieval and involves incremental manual traversal of these structures' [56].
2. Search as 'a more indirect way to find information, where the user generates textual labels that refer to the name of information item, one of its attributes or its contents' [56]. Besides common search engines capabilities, Haystack [57] promotes interactions between any user and the information he uses to adapt the information retrieval process. More recently, such kind of approaches have been developed as desktop search systems [55, 58].

Regarding those two methods, the Phlat [59] system is an example that empowers the user by providing both of

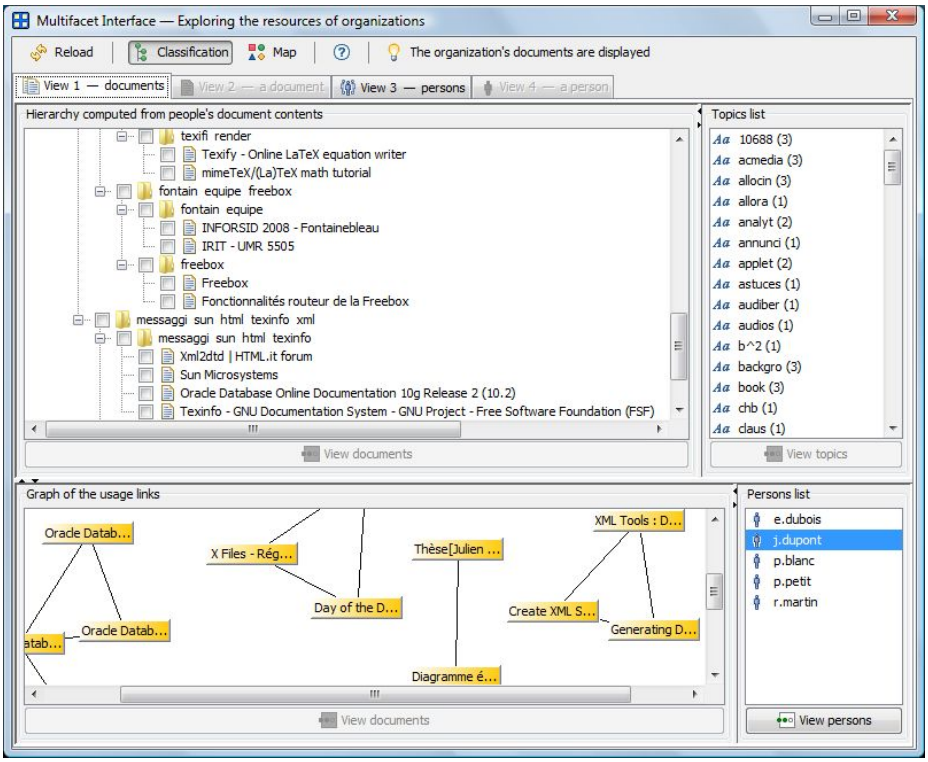

Fig. 16 View 1 of the faceted visual interface showing the community documents and members.

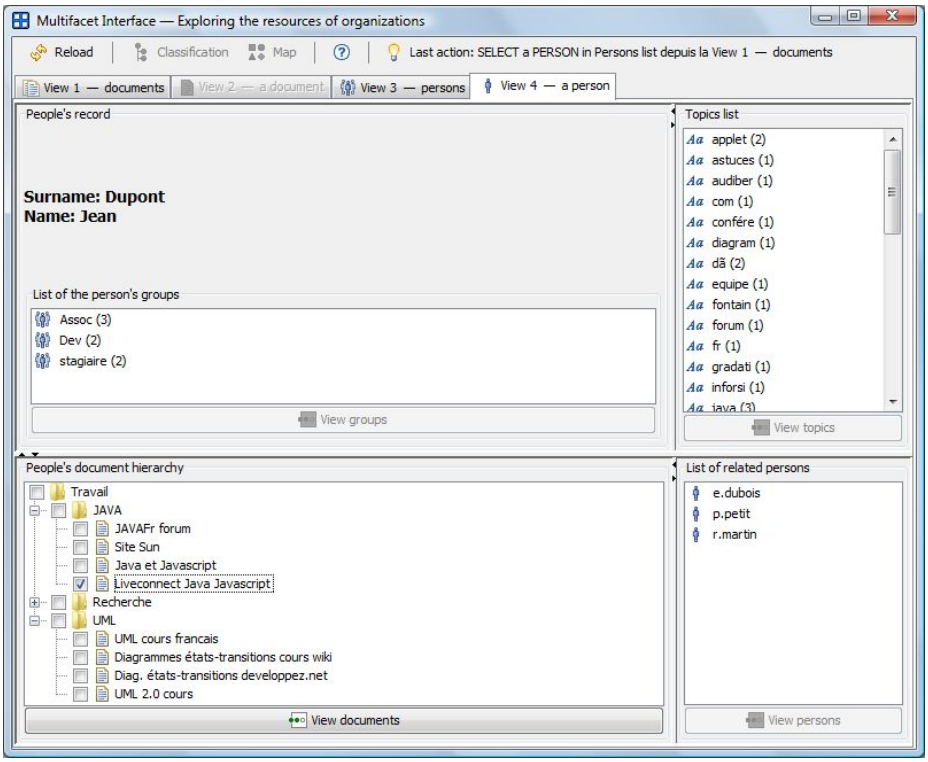

Fig. 17 View 4 of the faceted visual interface showing the record of the member called 'Jean Dupont' and related information (e.g., related persons, topics)

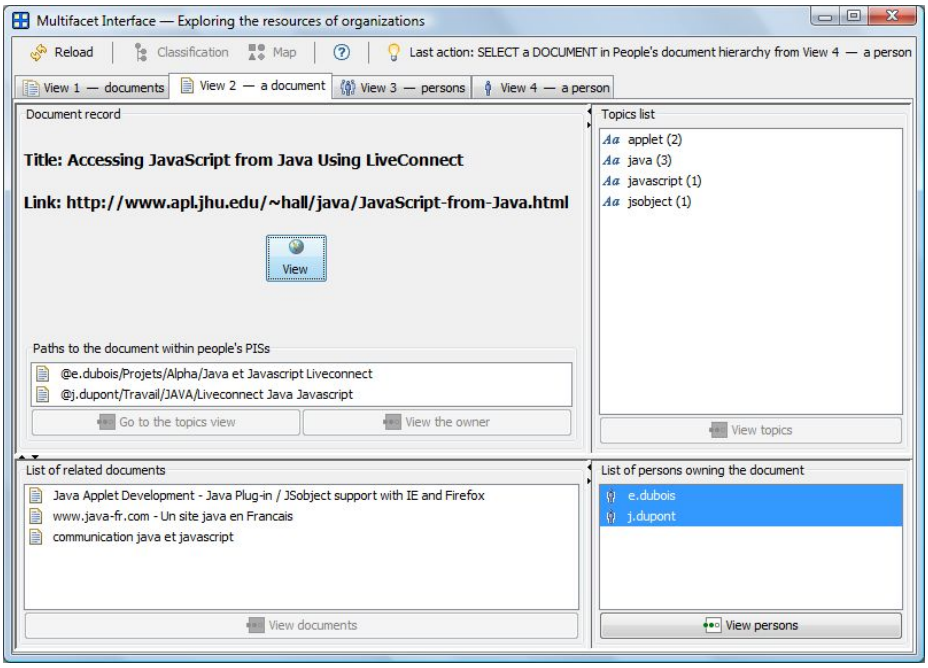

Fig. 18 View 2 of the faceted visual interface showing the record of a selected document and related information (e.g., document owners).

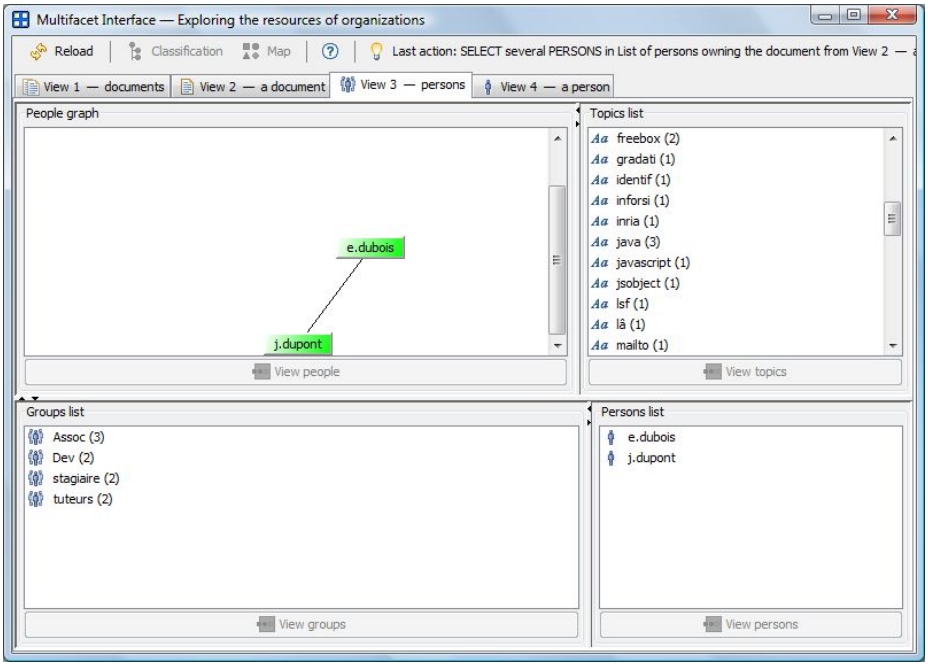

Fig. 19 View 3 of the faceted visual interface showing the owners of the selected document and related information (e.g., groups, topics).

these methods through a 'smooth continuum of exploratory search.'

Compared with PIM, our contribution aims at providing a user with an exploratory search feature targeting his/her own information extracted from his/her PIS. This feature is akin to the aforementioned navigation method, where the 'view' concept in our contribution acts as a step in the user's navigation process. In this article, we demonstrated how to exploit digital documents. However, this proposal can be generalized to any type of information present in users' silos.

8.2 Group Information Management

Group Information Management (GIM) 'refers to the practice and the study of the individual actions performed to sup-

port group activity' [60]. In short, GIM is PIM extended to groups. Indeed, an information-based task is rarely achieved by a single knowledge worker. As a result, several people usually work on shared materials. The need to share these materials across a group is highlighted in [53]. GIM systems reduce issues encountered by users who usually share documents and information, as introduced in Section 2.2.1 and discussed in [61, 62].

Compared with GIM, our contribution also promotes information sharing between members of a community. With GIM systems, users explicitly define access rights to their materials, as a way to ensure user control over their own material, which is considered as an important feature in [60]. In our contribution, we intend to empower the community. Thus, we hypothesized that people are willing to share information with other members based on a win-win principle. As a consequence, all information extracted from PISs is shared among community members by default. To satisfy the information control and privacy requirements, in our approach, users can hide materials from the community by using a specific naming scheme, as explained in Section 5.3.1. To sum up, our approach fosters sharing first and foremost, while giving users the choice to keep some materials private.

8.3 Faceted access to information

Personal or Group Information Management systems rely on several facets, such as:

1. Document contents according to terms [63] or ontologies [32].
2. Hierarchical organization of documents [64].
3. Document consumers [59] and their relationships through social networks [65].
4. Temporal dimension of information [59].
5. Document annotations [66].

In order to provide a comprehensive access to information, several facets are usually combined together in a same application, as shown in [64, 59, 67]. In order to highlight relations between information items, facets commonly build on similarity measures. These measure the matching between various kinds of materials, such as:

- Documents through their contents (mostly stemming from the literature in IR [36]), links [68], or annotations [69].
- Users, mostly stemming from literature in Recommender Systems [e.g., 70].

Our contribution features several combined facets related to documents, users, and their relationships. We emphasized on rendering the relations between facets (e.g., documents and the people using them are shown together). We defined two original similarity measures based on the organization of documents in Section 3. These measures allow the computation of inter-document and inter-user similarities. They may be used as complementary evidence with respect to other usual similarity measures for building facets.

8.4 Visual information access

Personal and Group Information Management systems deal with large amounts of information. Since textual visualizations are limited — they do not allow scalable and effective exploration of materials — systems implement 2D and 3D visualization techniques. Much research has been achieved in the Information Visualization field for that purpose. Regarding materials present in silos, visualizations handle emails as proposed by Hubmann-Haidvogel et al. [67] or summarized by Kjun[6], and hierarchies of documents [e.g., 29], [7] for instance. We refer the interested reader to [26] for a comprehensive coverage of Information Visualization.

Our contribution relies on existing visualization techniques. We do not pretend to contribute a new visualization technique. The goal is rather to provide complementary visualizations that are combined for the user to achieve effective exploration of materials.

8.5 Dynamic queries

In the literature, systems handling 'dynamic queries' [71, 55] support users by offering direct interaction with data. For each interaction engaged by the user (e.g., selection, zoom, filter change), the system displays its result immediately. This fosters incremental exploration of datasets. This dynamic feature is promoted in 'synchronized visualizations' like in [67, 65, 72]. These allow the repercussion of any manipulation made through one visualization to the other ones. For instance, selecting an item in one visualization provokes its selection in all other displayed visualizations.

The proposed approach supports dynamic queries when the selection of the user leads to a possible change of facet (e.g., in Figure 13, selecting a single user in View 3 leads to the displaying of View 4, which represents the person's record). Furthermore, our approach supports synchronized visualizations between facets or even views. This promotes the incremental exploration of datasets.

[6] http://pim.famnit.upr.si/blog/index.php?/archives/117-More-than-21-ways-to-visualize-and-explore-email-inbox.html

[7] http://pim.famnit.upr.si/wiki/index.php/Visualizing_large_hierarchies

9 Conclusion and future work

Knowledge workers in modern communities own Personal Information Spaces (PISs) where they organize useful documents related to their daily activities. The hierarchical structure is commonly used, reflecting individuals' association of ideas as they split up and organize their information space. Indeed, grouping documents in a same folder means they are used together for a given task. Actually, PISs are mines of information created incrementally, as knowledge workers keep finding new nuggets of information. Although the contents of a PIS may correspond to the needs of several knowledge workers (e.g., affiliated with a common team) it is accessible to its owner only. In addition, spreading documents either manually or automatically through recommender systems suffers from many issues: implied cognitive overload, accuracy of the representation of user needs since they evolve constantly, and so on. This may contribute to explain why people favor external information sources for information seeking tasks (e.g., the Web). However, the PISs seem more appropriate for the activities of any community since their contents stem from its members' positive relevance feedback.

In order to better promote the investment by knowledge workers managing their PIS, we proposed to give access to these valuable information resources. We defined a faceted visual interface that merges faceted search and visualization techniques in order to concurrently display the manifold relations between information and knowledge workers. This intends to makes the most of the heterogeneity of users and their tasks. To do that, we first defined organization-based inter-document and inter-user similarity measures. We then showed that these are complementary. Finally, we exploited these sources of evidence in the designed faceted visual interface, which allows the exploration of a community documents and knowledge workers. The displayed information is extracted from the community Information System, more precisely from its members' PISs. Our proposal builds on a win-win principle: community members' cognitive efforts are turned profitable for the whole community; in return anyone can explore the collective resources and find relevant documents regarding his activities. Targeting knowledge workers as well as the Human Resources Management department, we underlined how a classical exploration based on document contents is worth completing with a organization-based exploration realized through the interface facets.

A short term direction may consist in evaluating the benefits of the proposed faceted visual interface in a real community for validating our approach. First, we envisage experimenting this interface with a research team from our laboratory. The feedback from scholars who are experts in their domains, and from newcomers such as Masters students will provide a qualitative evaluation similar to Millen and Fontaine's [73] reported results. Second, we will go into such preliminary observations in greater depth thanks to quantitative experiments.

A second direction consists of considering the temporal dimension for visualizing community resources. Some tasks require a fresh domain knowledge (e.g., consulting, technology forecasting, business intelligence) whereas others require a solid background and long-term knowledge of the domain (experience with a technology, retrospective about a domain). As a result, highlighting the real use/organization of documents classified in PISs would allow one to draw a distinction between resources and topics which (re)emerge, and those progressively abandoned.

A third direction consists in fitting the four design principles suggested by [74]. As matter stands, our contribution supports two among these four principles:

- *Integrate analytical and browsing oriented ways of exploration.* This is supported by the browsing feature of the proposed interface.
- *Provide views to different dimensions of an information space.* This is supported by the combination of the various views and facets of the proposed interface.
- *Support various ways of formulating an information need.* As a first step towards this goal, a search engine may be added to the proposed interface. Another direction consists in allowing users to express their information needs by providing a sample of (non)relevant documents.
- *Make search a pleasurable experience.* Improving this point would require the selection of adapted/pleasant visualizations to explore the quiescent information capital of a community.

In the long term, we plan to automatically identify the groups of individuals. Identifying and exploring social relations as well as similarities between the topics, and the use of documents (e.g., printing, sending) are many indicators that will be exploited for that purpose. Another direction concerns the use of many other sources of evidence (e.g., human annotations or tags from silos) as to assess information quality and relevance. this would lead to better identifying the several relations existing between users and documents.

References

1. Pamela J. Hinds and Jeffrey Pfeffer. Why Organizations Don't "Know What They Know": Cognitive and Motivational Factors Affecting the Transfer of Expertise. In Mark S. Ackerman, Volker Wulf, and Volkmar Pipek, editors, *Sharing expertise: Beyond knowledge management*, chapter 1, pages 3–26. MIT Press, Cambridge, MA, USA, 2003. ISBN 0262011956.

2. Peter Ferdinand Drucker. *Landmarks of tomorrow: A report on the new "post-modern" world*. Transaction Publishers, 1959.
3. Alison Kidd. The marks are on the knowledge worker. In *CHI'94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 186–191, New York, NY, USA, 1994. ACM. ISBN 0-89791-650-6. doi: 10.1145/191666.191740.
4. Abigail J. Sellen and Richard H.R. Harper. *The myth of the paperless office*. The MIT Press, Cambridge, MA, USA, 2003. ISBN 026269283X.
5. Susan Feldman. The high cost of not finding information. *KM World magazine*, 13(3), March 2004. URL `http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9534`.
6. William Jones, Ammy Jiranida Phuwanartnurak, Rajdeep Gill, and Harry Bruce. Don't Take My Folders Away!: Organizing Personal Information to Get Things Done. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1505–1508, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-002-7. doi: 10.1145/1056808.1056952.
7. William Jones. How People Keep and Organize Personal Information. In Jones and Teevan [75], chapter 3, pages 35–56. ISBN 978-0-295-98755-2.
8. Christopher S.G. Khoo, Brendan Luyt, Caroline Ee, Jamila Osman, Hui-Hui Lim, and Sally Yong. How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Inf. Res.*, 11(2), January 2007. URL `http://informationr.net/ir/12-2/paper293.html`.
9. James Rucker and Marcos J. Polanco. Siteseer: personalized navigation for the Web. *Commun. ACM*, 40(3): 73–76, 1997. ISSN 0001-0782. doi: 10.1145/245108.245125.
10. Harris Wu and Michael D. Gordon. Collaborative Filing in a Document Repository. In *SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference*, pages 518–519, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009099.
11. Harris Wu, Michael D. Gordon, and Kurt DeMaagd. Document Co-Organization in an Online Knowledge Community. In *CHI'04: CHI'04 extended abstracts on Human factors in computing systems*, pages 1211–1214, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-703-6. doi: 10.1145/985921.986026.
12. Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita. Using Annotations in Enterprise Search. In *WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 811–817, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135900.
13. Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005. doi: 10.1045/april2005-hammond.
14. David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social Bookmarking in the Enterprise. In *CHI'06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-372-7. doi: 10.1145/1124772.1124792.
15. Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social Bookmarking Tools (II): A Case Study – Connotea. *D-Lib Magazine*, 11(4), April 2005. doi: 10.1045/april2005-lund.
16. Miquel Montaner, Beatriz López, and Josep Lluís de la Rosa. A Taxonomy of Recommender Agents on the Internet. *Artif. Intell. Rev.*, 19(4):285–330, 2003. doi: 10.1023/A:1022850703159.
17. George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987. ISSN 0001-0782. doi: 10.1145/32206.32212.
18. Giovanni Maria Sacco and Yannis Tzitzikas. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*. Springer Publishing Company, Inc., 2009. ISBN 978-3-64202-358-3.
19. Barbara H. Kwasnik. The role of classification in knowledge representation and discovery. *Library Trends*, 48(1):22–47, 1999.
20. R. B. Allen. The role of classification in knowledge representation and discovery. *Electronic Publishing*, 8 (2-3):247–257, 1995.
21. Jonathan Koren, Yi Zhang, and Xue Liu. Personalized interactive faceted search. In *WWW'08: Proceeding of the 17th international conference on World Wide Web*, pages 477–486, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367562.
22. Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *CHI'03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 2003. ACM. ISBN 1-58113-630-7. doi: 10.1145/642611.642681.
23. Einat Amitay, David Carmel, Nadav Har'El, Shila Ofek-Koifman, Aya Soffer, Sivan Yogev, and Nadav Golbandi. Social search and discovery using a unified approach. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1211–1212, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: http://doi.acm.org/10.1145/1526709.

1526933.

24. Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. What do exploratory searchers look at in a faceted search interface? In *JCDL'09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 313–322, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-322-8. doi: 10.1145/1555400.1555452.
25. Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000. ISSN 1077-2626. doi: 10.1109/2945.841119.
26. Chaomei Chen. *Information visualization: Beyond the horizon*. Springer, 2nd edition, May 2006. ISBN 184628340X.
27. YunYun Yang, Lucy Akers, Thomas Klose, and Cynthia Barcelon Yang. Text mining and visualization tools – impressions of emerging capabilities. *World Patent Information*, 30(4):280–293, December 2008. doi: 10.1016/j.wpi.2008.01.007.
28. Jean-Daniel Fekete and Catherine Plaisant. Interactive Information Visualization of a Million Items. In *INFOVIS'02: Proceedings of the IEEE Symposium on Information Visualization*, pages 117–124, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1751-X. doi: 10.1109/INFVIS.2002.1173156.
29. Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *VIS'91: Proceedings of the 2nd conference on Visualization*, pages 284–291, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press. ISBN 0-8186-2245-8. doi: 10.1109/VISUAL.1991.175815.
30. Teuvo Kohonen. *Self-organizing maps*. Springer-Verlag, Secaucus, NJ, USA, 3rd edition, 2001. ISBN 3540679219.
31. Marc Boyer, Marie-Françoise Canut, Max Chevalier, André Péninou, and Florence Sèdes. Cartographie de l'organisation : une approche topologique des connaissances. In *EGC'07 : actes des 7^e journées Extraction et Gestion des Connaissances*, volume RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, pages 557–568. Cépaduès, 2007.
32. Josiane Mothe, Claude Chrisment, Bernard Dousset, and Joel Alaux. DocCube: Multi-dimensional visualisation and exploration of large document sets. *J. Am. Soc. Inf. Sci.*, 54(7):650–659, 2003. doi: 10.1002/asi.10257.
33. Gilles Hubert, Josiane Mothe, Anis Benammar, Taoufiq Dkaki, Bernard Dousset, and Said Karouach. Textual Document Mining Using a Graphical Interface. In *HCI'01: Proceedings of the 9th international conference on Human Computer Interaction*, volume 1, pages 918–922. Lawrence Erlbaum Associates, August 2001. ISBN 0-8058-3607-1.
34. Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the WEBSOM method. *Inf. Sci.*, 163(1-3):135–156, 2004. ISSN 0020-0255. doi: 10.1016/j.ins.2003.03.017.
35. Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern information retrieval*. ACM Press/Addison-Wesley, 1999. ISBN 0-201-39829-X.
36. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. ISBN 978-0521865715.
37. Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. doi: 10.1108/eb046814.
38. Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18 (11):613–620, November 1975. doi: 10.1145/361219.361220.
39. David Abrams, Ron Baecker, and Mark Chignell. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. In *CHI'98: Proceedings of the conference on Human factors in computing systems*, pages 41–48, New York, NY, USA, 1998. ACM Press. ISBN 0-201-30987-4. doi: 10.1145/274644.274651.
40. Max Chevalier, Claude Chrisment, and Christine Julien. Helping People Searching the Web: Towards an Adaptive and a Social System. In *ICWI'04: Proceedings of the 3rd International Conference WWW/Internet*, pages 405–412. IADIS, 2004.
41. Michel Jaczynski and Brigitte Trousse. WWW Assisted Browsing by Reusing Past Navigations of a Group of Users. In Barry Smyth and Padraig Cunningham, editors, *EWCBR*, volume 1488 of *LNCS*, pages 160–171. Springer, 1998. ISBN 3-540-64990-5. doi: 10.1007/BFb0056330.
42. Claus-Peter Klas and Norbert Fuhr. A new Effective Approach for Categorizing Web Documents. In *Proceedings of the 22th BCS-IRSG Colloquium on IR Research*, April 2000.
43. Guillaume Cabanac, Max Chevalier, Claude Chrisment, and Christine Julien. An Original Usage-based Metrics for Building a Unified View of Corporate Documents. In Roland Wagner, Norman Revell, and Günther Pernul, editors, *DEXA'07: Proceedings of the 18th International Conference on Database and Expert Systems Applications*, volume 4653 of *LNCS*, pages 202–212. Springer, September 2007. ISBN 3-540-74467-3. doi: 10.1007/978-3-540-74469-6_21.
44. Ellen M. Voorhees. Overview of TREC 2001. In *TREC'01: Proceedings of the 10th Text REtrieval Conference*, 11 2001.

45. NIST. README file for TREC-9 Filtering Track Collections. `http://trec.nist.gov/data/filtering/README.t9.filtering`, 2001.
46. David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference*, pages 329–338, New York, NY, USA, 1993. ACM Press. ISBN 0-89791-605-0. doi: 10.1145/160688.160758.
47. Morten Hertzum and Annelise Mark Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. *Inf. Process. Manage.*, 36(5):761–778, 2000. ISSN 0306-4573. doi: 10.1016/S0306-4573(00)00011-X.
48. Peter Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.
49. Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, 1991. ISSN 0038-0644. doi: 10.1002/spe.4380211102.
50. N. Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Inform. Stor. Retr.*, 7(5):217–240, 1971. doi: 10.1016/0020-0271(71)90051-9.
51. Yoëlle S. Maarek and Israel Ben-Shaul. Automatically Organizing Bookmarks per Contents. *Computer Networks and ISDN Systems*, 28(7-11):1321–1333, 1996. doi: 10.1016/0169-7552(96)00024-4.
52. Joseph 'Jofish' Kaye, Janet Vertesi, Shari Avery, Allan Dafoe, Shay David, Lisa Onaga, Ivan Rosero, and Trevor Pinch. To Have and to Hold: Exploring the Personal Archive. In *CHI'06: Proceedings of the conference on Human Factors in computing systems*, pages 275–284, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-372-7. doi: 10.1145/1124772.1124814.
53. William Jones and Harry Bruce. A Report on the NSF-Sponsored Workshop on Personal Information Management, Seattle, WA, 2005. Technical report, 2005.
54. Willian Jones. *Keeping Found Things Found: The Study and Practice of Personal Information Management*. Morgan Kaufmann Publishers, November 2007.
55. Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR'03: Proceedings of the 26th ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860451.
56. Steve Whittaker. Personal information management: From consumption to curation. *Annu. Rev. Inform. Sci. Technol.*, 45, 2011.
57. Eytan Adar, David Karger, and Lynn Andrea Stein. Haystack: per-user information environments. In *CIKM'99: Proceedings of the 8th international conference on Information and knowledge management*, pages 413–422, New York, NY, USA, 1999. ACM. ISBN 1-58113-146-1. doi: 10.1145/319950.323231.
58. Paul Chirita, Rita Gavriloaie, Stefania Ghita, Wolfgang Nejdl, and Raluca Paiu. Activity based metadata for semantic desktop search. In Asunción Gómez-Pérez and Jérôme Euzenat, editors, *ESCW'05: Proceedings of the 2nd european Semantic Web conference*, volume 3532 of *LNCS*, pages 199–213. Springer, 2005. doi: 10.1007/11431053_30.
59. Edward Cutrell, Daniel Robbins, Susan Dumais, and Raman Sarin. Fast, flexible filtering with Phlat – personal search and organization made easy. In *CHI'06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 261–270, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7. doi: 10.1145/1124772.1124812.
60. Wayne G. Lutters, Mark S. Ackerman, and Xiaomu Zhou. Group Information Management. In Jones and Teevan [75], chapter 14, pages 236–248. ISBN 978-0-295-98755-2.
61. Tara Whalen, Elaine Toms, and James Blustein. File sharing and group information management. In *PIM'08: Proceedings of the workshop on Personal Information Management*, 2008.
62. Thomas Erickson. From PIM to GIM: personal information management in group contexts. *Commun. ACM*, 49(1):74–75, January 2006. ISSN 0001-0782. doi: 10.1145/1107458.1107495.
63. Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In Evangelios Simoudis, Jiawei Han, and Usama Fayyad, editors, *KDD'96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 238–243, Menlo Park, California, 1996. AAAI Press.
64. Keith Andrews, Wolfgang Kienreich, Vedran Sabol, Jutta Becker, Georg Droschl, Frank Kappe, Michael Granitzer, Peter Auer, and Klaus Tochtermann. The InfoSky visual explorer: Exploiting hierarchical structure and document similarities. *Info. Vis.*, 1(3–4):166–181, 2002. doi: 10.1057/palgrave.ivs.9500023.
65. Florian Evequoz and Dennis Lalanne. Personal information management through interactive visualizations. In *Infovis-DC'07: Proceedings of the Doctoral Colloquium of IEEE Information Visualization Conference*, pages 158–160. IEEE, 2007.
66. Stephan Bloehdorn, Olaf Goerlitz, Simon Schenk, and Max Voelkel. TagFS – tag semantics for hierarchical file systems. In *I-KNOW'06: Proceedings of the 6th International Conference on Knowledge Management*,

September 2006.

67. Alexander Hubmann-Haidvogel, Arno Scharl, and Albert Weichselbraun. Multiple coordinated views for searching and navigating web content repositories. *Inf. Sci.*, 179(12):1813–1821, May 2009. ISSN 0020-0255. doi: 10.1016/j.ins.2009.01.030.
68. Glen Jeh and Jennifer Widom. SimRank: a measure of structural-context similarity. In *KDD'02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775126.
69. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. FolkRank: A ranking algorithm for folksonomies. In *FGIR'06: Proceedings of the workshop on Information Retrieval*, pages 111–114, 2006.
70. Marta Millan, Maria Trujillo, and Edward Ortiz. A collaborative recommender system based on asymmetric user similarity. In Hujun Yin, Peter Tino, Emilio Corchado, Will Byrne, and Xin Yao, editors, *IDEAL'07: Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, volume 4881 of *LNCS*, pages 663–672. Springer, 2007. doi: 10.1007/978-3-540-77226-2_67.
71. Christopher Ahlberg, Christopher Williamson, and Ben Shneiderman. Dynamic queries for information exploration: an implementation and evaluation. In *CHI'92: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 619–626, New York, NY, USA, 1992. ACM. ISBN 0-89791-513-5. doi: 10.1145/142750.143054.
72. Josiane Mothe, Claude Chrisment, Taoufiq Dkaki, Bernard Dousset, and Daniel Egret. Information mining: use of the document dimensions to analyse interactively a document set. In *ECIR'01: Proceedings of the European Colloquium on IR Research*, pages 66–77. BCS, April 2001.
73. David R. Millen and Michael A. Fontaine. Improving Individual and Organizational Performance through Communities of Practice. In *GROUP'03: Proceedings of the international conference on Supporting group work*, pages 205–211, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-693-5. doi: 10.1145/958160.958192.
74. Jens Gerken, Mathias Heilig, Hans-Christian Jetter, Sebastian Rexhausen, Mischa Demarmels, Werner A. König, and Harald Reiterer. Lessons learned from the design and evaluation of visual information-seeking systems. *Int. J. Digit. Libr.*, 10(2-3):49–66, December 2009. ISSN 1432-5012. doi: 10.1007/s00799-009-0052-6.
75. William Jones and Jaime Teevan. *Personal information management*. University of Washington Press, WA, USA, 2007. ISBN 978-0-295-98755-2.