Files of a Feather Flock Together? Measuring and Modeling How Users Perceive File Similarity in Cloud Storage

Will Brackenbury
University of Chicago
wbrackenbury@uchicago.edu

Galen Harrison
University of Chicago
harrisong@uchicago.edu

Kyle Chard
University of Chicago
chard@uchicago.edu

Aaron Elmore
University of Chicago
aelmore@uchicago.edu

Blase Ur
University of Chicago
blase@uchicago.edu

ABSTRACT

Prior work suggests that users conceptualize the organization of personal collections of digital files through the lens of similarity. However, it is unclear to what degree similar files are actually located near one another (e.g., in the same directory) in actual file collections, or whether leveraging file similarity can improve information retrieval and organization for disorganized collections of files. To this end, we conducted an online study combining automated analysis of 50 Google Drive and Dropbox users' cloud accounts with a survey asking about pairs of files from those accounts. We found that many files located in different parts of file hierarchies were similar in how they were perceived by participants, as well as in their algorithmically extractable features. Participants often wished to co-manage similar files (e.g., deleting one file implied deleting the other file) even if they were far apart in the file hierarchy. To further understand this relationship, we built regression models, finding several algorithmically extractable file features to be predictive of human perceptions of file similarity and desired file co-management. Our findings pave the way for leveraging file similarity to automatically recommend access, move, or delete operations based on users' prior interactions with similar files.

CCS CONCEPTS

• **Information systems → Data management systems**.

KEYWORDS

cloud storage, Dropbox, Google Drive, data management, personal information management, user study

ACM Reference Format:
Will Brackenbury, Galen Harrison, Kyle Chard, Aaron Elmore, and Blase Ur. 2021. Files of a Feather Flock Together? Measuring and Modeling How Users Perceive File Similarity in Cloud Storage. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462845

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3462845

1 INTRODUCTION

Users spend a significant amount of time viewing, curating, and organizing collections of digital files [36, 73, 74]. Building on the success of recommender systems in other contexts [10, 29, 68, 79], researchers have recently developed a number of recommender systems to help users identify files they wish to retrieve [26, 35, 40, 50, 70]. Prior user-centered research has found that abstract notions of file similarity underpin how users view file organization and retrieval [8, 11, 21, 23, 51]. It is no surprise, then, that these recommender systems implicitly seem to rely on file similarity to make recommendations for file retrieval. For example, in Google Drive, if a user edits a document, the QuickAccess [26, 70] tool may suggest other files that were last modified at similar times.

However, these recent systems take a relatively narrow view of what it means for files to be similar. For instance, while most systems concretize similarity in terms of access patterns, we hypothesize that similarity of metadata and content features (e.g., filenames, objects recognized in images) might provide important signals to recommender systems. Furthermore, previous work focused almost exclusively on file retrieval, leaving open the question of whether a system that observes a user deleting or moving a file should also recommend that they delete or move (to the same place) similar files. We use the term **co-management** to describe this broader pattern of *managing similar files in similar ways.*

In this paper, we answer a series of complementary questions about conceptualizing co-management and file similarity more broadly than in prior work by conducting a two-part, online user study of 50 Google Drive and Dropbox users and their cloud accounts. The first part surveyed participants about how they used and organized their cloud accounts. After receiving participant consent, we also used the Google Drive and Dropbox APIs to analyze participants' accounts, collect metadata, and compute the similarity of pairs of files in the account in terms of eleven metadata and content features. Once this automatic processing had concluded, the participant returned for the second part of the study, answering survey questions about how they perceived the similarity between 18 pairs of files from their account, as well as whether they wanted to co-manage those files (i.e., find, move, or delete them together).

After describing related work (Section 2), we present our framework of file similarity and co-management (Section 3), unifying ideas from prior work. In this framework, we differentiate between **perceived similarity**, the degree to which users perceive files to be similar to each other in various ways, and **data similarity**, or the similarity of features that can be automatically extracted from

file access patterns, metadata, and contents. Section 4 describes our user study methodology, while Section 5 describes our user study participants and the contents of their accounts.

As our first research question, we wondered to what degree seemingly curated file repositories in consumer cloud storage stand to benefit from recommender systems. In Section 6, we examine the structure of participants' cloud accounts, focusing on where similar files are located. Echoing prior work [38, 39, 51, 57, 72], we found that some participants piled most of their files into a small number of folders (termed a **piler** hierarchy), while others organized their files into many folders with long chains of subfolders (termed a **filer** hierarchy). Our first key contribution came from analyzing the relative locations of pairs of files perceived by participants to be similar in these hierarchies. Intuition might have suggested that similar files would be located in the same directory, or perhaps in an adjacent directory. However, we found this not to be the case. Even in superficially organized filer hierarchies, pairs of files participants perceived as similar were located far away in the directory structure. We observed a similar result when looking at files' automatically extractable metadata and content features. As such, even the types of users whom prior work characterized as organized filers stand to benefit from automated recommendations about files that are inconvenient to find or that have been forgotten.

Our second research question was whether participants actually wanted to co-manage files they perceived as similar. In Section 7, we present correlations in our survey results, showing that participants did indeed desire to co-manage the majority of files they perceived as similar, whereas they wanted to co-manage only a small fraction of files they did not perceive as similar. Whereas existing tools [6, 34, 35, 50, 70] already leverage this result for finding and retrieving files, we show similar results for co-moving and co-deleting similar files, highlighting the need for broader co-management recommendations than are currently provided.

To lay the foundation for transitioning these insights to tools, our third research question investigated what metadata and content features are predictive both of whether humans perceive files as similar and whether they want to co-manage them. Existing tools focus on temporal information, such as files' last modification date or last access time, as a proxy for similarity [6, 34, 35, 50, 70]. While, as detailed in Section 8, our regression models did find temporal information to be predictive of both perceived similarity and desired co-management, we also identified metadata features (e.g., the similarity of filenames) and content features (e.g., the similarity of words used in a document or of objects recognized in images) as predictive. We conclude the paper in Section 9 by discussing how these insights can be incorporated into future tools that leverage broader notions of file similarity, beyond simply files' relative locations or temporal patterns, to help users co-manage similar files regardless of their relative locations in disorganized cloud accounts.

2 RELATED WORK

We first present prior work on how users organize file collections. While we describe the methods of the relevant research here, we further systematize that literature's key findings in Section 3. We then discuss prior efforts to build systems that aid in information retrieval in both the file-management and email contexts.

2.1 File Organization Behavior and Tools

Researchers first studied file organization in offices [44, 47, 51], later analyzing digital analogues [7, 8, 21]. These studies developed several frameworks that describe how humans categorize documents and files. Malone [51] and Kwasnik [44] asked eight and ten participants, respectively, to describe the organization of their office spaces. Barreau performed a similar study on seven managers' digital file collections [8]. Kwasnik [44] and Barreau [8] used these studies to develop frameworks describing how humans classify documents and files. Bergman et al. investigated whether such a framework (the "User-Subjective Approach") could drive the development of tools for managing file collections [11, 12]. Boardman and Sasse examined common folder classifications in file collections, augmenting these frameworks [21]. Brackenbury et al. later translated these ideas to the management of data lakes [23].

Researchers have prototyped a number of automated tools to help users organize files and emails. In the domain of file management, Bergman et al. developed GrayArea, which provides a "deletion-lite" option [17]. In the same context of cloud storage we investigate, Bergman et al. developed a tool that nudged participants to save files to a suggested folder [18]. Segal and Kephart [66] created a similar tool for email, and Sinha and Basu [67] created a related tool for local file storage. However, none of these tools support such behavior beyond a file or email's initial "save" action. Researchers have also built tools to aid in the organization of other types of collections. For example, Segal and Kephart's MailCat suggests appropriate folders for an email [66]. Other tools group emails by topic [30] or by additional features [69]. In the context of collections of bookmarks, information about bookmarks' social context can aid organization and discovery [2, 54, 55].

2.2 Re-finding Information

Re-finding, or retrieving information a user has previously accessed but lost or forgotten about, is among the most common activities in information management [3]. As such, many tools and prototypes have supported re-finding. For example, Dumais et al.'s "Stuff I've Seen" facilitated information re-use by providing a unified index across types of information, adding contextual clues in the search interface [33]. Rhodes and Starner's Remembrance Agent used keyword searches to find similar emails and lines from text files [63]. Whittaker noted that, even without using specific tools, users typically organize collections to aid later re-finding actions [74]. Analyzing Office 365 logs, Xu et al. [78] identified how explanations impact adoption of document recommendations.

Temporal access patterns are often associated with information re-finding. In a field study of over 100 users, Jahanbakhsh et al. [40] found that users' frequency and types of prior interactions with documents influence later re-visitation patterns. Tata et al. [70] created and evaluated the QuickAccess tool for Google Drive, which provides time-based recommendations for information re-finding. As such, temporal features were among the many types we measured.

Several related tools support improved navigation to content of interest either via automatically provided shortcuts [5, 6, 50] or highlighting content likely to be accessed [35, 49]. Systems like Haystack [62], Stuff I've Seen [33], and various Semantic Desktop tools [27, 64, 65] instead enhanced an interface's search capability

Ours	Kwasnik [46], Barreau [8]	Bergman et al. [11]	Boardman and Sasse [21]	Brackenbury et al. [23]
Topic	Document Attributes	Subjective Classification Principle	Topic	The Data Itself
Creation Context	Situation Attributes / Document Attributes / Time	Subjective Context Principle	Project / Role	Origin
Derivation		Subjective Context Principle		Origin
Purpose	Disposition / Situation Attributes	Subjective Context Principle		Origin
	Order / Scheme			
	Document Attributes		Document Class	
	Value	Subjective Importance Principle		Current Characteristics

Table 1: Comparison of our framework for perceived similarity (left) with notions of similarity discussed in prior work.

to improve re-finding. In contrast, we look not just at file retrieval, but also co-management in moving and deleting files. Additionally, nearly all of these tools rely heavily on temporal access patterns, whereas we investigated many other metadata and content features.

While information re-finding could rely purely on search features, prior studies have found that users prefer standard file-management interfaces for navigation and re-finding [13, 14]. Teevan et al. [71], Bergman et al. [13], and Bergman et al. [16] conducted studies with semi-structured interviews, longitudinal measurement, and an in-lab study that identified a few reasons why navigation is preferred over search. They found that search has a higher cognitive burden. Furthermore, forming a search query requires a user to recall some context for the file without any aid. Teevan et al. [71] found that users navigate through file hierarchies using additional context gained at each step of navigation. We do not investigate search-related behavior, but this knowledge of users' navigation through file hierarchies informs our investigation of file hierarchy structure. Related to search, Civan et al. [28] and Bergman et al. [15] found that relying on file tags for information retrieval posed similar difficulties to search because tagging leaves files "placeless."

3 FRAMEWORK AND DEFINITIONS

Here, we define our notions of perceived similarity, data similarity, and co-management in the context of prior work. Toward building richer tools for information management, we empirically evaluate and quantify relationships among these concepts in Sections 7–8.

3.1 Perceived Similarity

We define *perceived similarity* as a user's subjective perception about how files may be similar or dissimilar. For example, users may *perceive* two documents to be similar if they were written by the same author or describe the same project. Prior work describes how many people use this idea to describe the organization of their files [8, 11, 21, 23, 44] and organize them for later retrieval [41, 74].

To evaluate whether users wish to manage similar files similarly, we focus on four dimensions of perceived file similarity synthesized from prior work [8, 11, 21, 23, 44]. Table 1 summarizes differences between our framework and prior work. Our framework includes:

- **Topic:** Two files are similar if they are about the same subject. Kwasnik [46] and Barreau's [8] frameworks described this concept as part of "Document Attributes," which included other items like "Author" and "Physical Form" (e.g., a spreadsheet printout). Topic also falls under Bergman et al.'s [11] "Subjective Classification Principle" (information with the same subject should be categorized together). **Example**: a photo of a dog and a document about dog grooming.
- **Purpose:** Two files are similar if they will likely be used for similar tasks or purposes. Purpose is a subset of "Situation Attributes" in Kwasnik [46] and Barreau's [8] frameworks, but also includes aspects of "Disposition," a user's intentions about whether to keep or discard the file. Bergman et al.'s [11] "Subjective Context Principle" also encompasses Purpose, as Purpose is part of the context when a file is saved. **Example**: a receipt and a W-2 form both saved for tax calculations.
- **Derivation:** Two files are similar if they are different versions of the same item, or if one "created" the other. Derivation is included under Bergman et al.'s [11] "Subjective Context Principle" given that a version of an item contains the same implicit context. Brackenbury et al. [23] discuss derivation as the "Provenance" component of "Origin." **Example**: a paper outline and the final version of that paper.
- **Creation context:** Two files are similar if they were created at the same time, by the same person, or in the same place. Kwasnik [46] and Barreau's [8] frameworks separate this across several categories as sub-attributes of "Source" ("Situation Attributes"), "Author" ("Document Attributes"), and "Time." **Example**: a poem authored at a writer's retreat and another person's poem written at the same retreat.

For three reasons, our framework does not include the attributes "Order" / "Scheme" (e.g., grouping, arrangement), "Document Attributes" (e.g., color, size), or "Value" (e.g., important, needs improvement) defined in other frameworks [11, 21, 23, 45]. First, in an empirical study, these aspects were some of the least common ways that interviewees described their file collections [45]. Second, "Document Attributes" and "Document Class" can more naturally be considered data similarity (defined later in this section), rather than perceived similarity. Third, "Order" and "Scheme" describe the organizational structure of a file collection, not perceived similarity. We use this synthesis of past work to guide our investigation of the relationship between similarity and co-management. Expanding or critically reevaluating past frameworks is not our focus. Prior work has investigated the reliability of these framework components, finding them to correspond to how users describe similarity [45].

3.2 Data Similarity

We define **data similarity** as comparisons of features that can be algorithmically extracted from files without human intervention. These features include **time** (e.g., last modified time), **metadata** (e.g., filename and size), and **content** features (e.g., text topics and objects identified in images). Table 2 lists all data similarity features we considered within these three categories. We hypothesized that data similarity could, at scale, help identify files perceived as similar.

Feature	Files	Description
Time		
Last Modified	All	Logarithm of difference, in seconds, between the two files' last modified dates
Metadata		
Filename	All	Jaccard similarity of the list of bigrams (two-letter chunks) in the filenames
File Size	All	Logarithm of difference, in bytes, between the file size
Tree Distance	All	The number of steps to reach one file from the other when traversing the file hierarchy (represented as a tree)
Shared Users	All	Jaccard similarity of the lists of unique user IDs with whom the files have been shared
Contents		
File Contents	All	Jaccard similarity of chunks of the raw file contents using MinHash
Text Contents	Text	Cosine similarity between documents' Word2Vec [53] vector embeddings
Text Topic	Text	Cosine similarity of documents' Term Frequency Inverse Document Frequency (TF-IDF) vectors [77]
Table Schema	Spreadsheets	Jaccard similarity of the column names of spreadsheets, such as .xlsx, .csv, and .tsv files
Image Contents	Images	Jaccard similarity between unique objects recognized in images by object-detection algorithms [37]
Image Color	Images	Absolute difference between the average RGB values of each image

Table 2: The data similarity features we examined, the files to which they apply, and how we computed them. We cluster these features in three groups: time (the focus of the most closely related work [35, 50, 70]), file metadata, and file contents.

Prior work has postulated that data similarity can be used to identify similar items [24, 58, 62], yet did not fully test these claims. Prior implementations [6, 34, 35, 50, 70] have focused almost exclusively on time features, such as file-access patterns or recently accessed files. That said, tools like Haystack [62] do use text features for retrieval, yet they do so in the context of a user-defined query, rather than by comparing files. The seminal Remembrance Agent [63], which recommends other files that might be relevant, is most similar to how we envision the use of data similarity. However, the Remembrance Agent only uses text features. In short, we explore more diverse and comprehensive features than prior work.

3.3 Co-management

We refer to the pattern of managing similar files similarly as **co-management**. Supporting a user's ability to co-manage files has the potential to passively improve a user's file organization over time, similar to tools that identify the best folders for a user to save a new file or email [18, 66, 67]. We consider the following actions:

- **Find:** If a user accesses a file, they may also want to access another similar file.
- **Move:** If a user moves a file to another folder, they may also want to move another similar file to the same folder.
- **Delete:** If a user deletes a file, they may also want to delete another similar file.

We focused on Find, Move, and Delete actions because they are commonly studied and used in practice. The Find action relates to prior work that used recent file or folder accesses to provide shortcuts to similar files or folders [6, 34, 35, 50, 70]. To the best of our knowledge, Move and Delete have not previously been investigated. We did not investigate actions that are less common or less foundational for information management, such as renaming, creating symlinks, or copying files [15, 31, 56]. Future work could expand our co-management framework to evaluate those strategies.

4 METHODOLOGY

To answer our research questions, we conducted a two-part online user study. In Part 1, we asked participants about file management abstractly and performed an automated scan of their cloud account. In Part 2, we elicited participants' perceptions of the similarity between, and desire to co-manage, 18 pairs of files from their account. Our online appendix [1] contains our full survey instrument, as well as our full regression tables.

4.1 Recruitment and Part 1 Survey

We recruited participants on Prolific Academic [61], a recommended alternative [60] to the Amazon Mechanical Turk crowdsourcing marketplace. We required they be age 18+, live in the USA, and have completed 100+ tasks with 95% approval. We also required participants to have a Google Drive or Dropbox account that was at least three months old and had at least 100 files, including one shared file. Our institution's IRB approved our protocol. In the short Part 1 survey that followed, we asked general questions about participants' demographics and organization of their cloud account. This portion took 15 minutes on average. Compensation was $2.50.

4.2 File Processing

Once the participant authorized access to their cloud account, we used the Google Drive or Dropbox API to analyze their account, collect file metadata, and compute data similarity features. We extracted text from documents, as well as column headers from data tables. Using the Google Vision API [37], we also computed a color histogram, listed recognized objects, and extracted available text from images. To reduce computational costs, we only collected data similarity features pairwise on a stratified sample of 1000 files whose distribution of file types matched the underlying account's. For confidentiality, we hashed all human-readable information with a participant-specific salt that we discarded after processing.

Once processing was complete, we selected 18 pairs of files to show participants in Part 2. For each of the following criteria, we randomly chose pairs from all files which satisfied the criterion.

- 2 pairs had similar filenames (based on their bigrams)
- 2 pairs' filenames had a small Levenshtein edit distance
- 2 pairs had a similar set of shared users
- 2 pairs had a similar text topic (based on TF-IDF [77])
- 2 pairs had a similar table schema
- 2 pairs had similar image contents (in Google Vision [37])
- 1 pair was in the same directory (tree distance 0)

- 1 pair was located at tree distance 1
- 4 pairs were selected randomly

We added additional random pairs whenever an insufficient number of files matched any criterion above. Thresholds were set via pilot testing. Due to a coding error, the tree distance of some file pairs was calculated incorrectly during sampling, inadvertently excluding a small number of file pairs that otherwise might have been selected based on being in the same directory or at tree distance 1. This error was corrected prior to our data analysis and would only have impacted sampling for a few files matching a corner case.

4.3 Part 2

Once we finished processing a participant's files, we invited them to Part 2, a survey centered on these 18 pairs of files from their own account. For each file pair, in randomized order, we first asked the participant to describe both files in free text. We then asked them to describe in free text how they believed the files were similar or dissimilar. Next, we asked them to rate their agreement with a series of statements on five-point Likert scales ("strongly agree" to "strongly disagree," plus a "don't know" option). This series included statements about our four classes of perceived similarity (e.g., "I consider these two files to be similar in *Topic*"). It also included statements about our three types of co-management (e.g., "If I were searching for information, and I found one of these files to be relevant, I would also want to *see* the other file"). Part 2 took approximately one hour to complete. Compensation was $10.00.

4.4 Analysis Approach

We analyzed three types of data: (i) general metadata for all files in each participant's account; (ii) data similarity features computed pairwise for a representative sample of 1,000 files in each participant's account; and (iii) detailed survey responses from participants about 900 file pairs (50 participants × 18 pairs each). We report illustrative quotes from participants, but do not formally analyze them qualitatively. We used the 900 labeled file pairs to build mixed-effects ordinal logistic regression models with the four types of perceived similarity and three types of co-management as our dependent variables. Because the data was not independent, we included a random effect for each participant. The data similarity features were our independent variables. When a given data similarity feature was not applicable (e.g., the Image Contents feature does not apply when comparing a spreadsheet and an image), or in the rare cases when our extractor encountered an error (e.g., reading a malformed file), we filled missing values as 0 or 1 for similarity and distance features, respectively.

4.5 Limitations

We report on a convenience sample of crowdworkers that is not representative of any broader population. Despite efforts to communicate how our data collection respected the privacy of participants' accounts, privacy-conscious crowdworkers were unlikely to participate, further biasing our sample. Because we asked the same questions for each file pair, participants may have been prone to fatigue and inattention [48]. We mitigated this concern by iteratively shortening both multiple-choice and free-response sections through extensive pilot testing, as well as restricting the study to 18 file pairs.

	Min	Q_1	Median	Q_3	Max
Age of oldest file (days)	148	2,405	3,001	3,570	4,546
Total size (GB)	< 1	2	5	11	151
Total # files	123	298	541	1,445	17,081
(# images)	0	33	168	732	15,123
(# documents)	4	44	140	298	2,345
(# spreadsheets)	0	5	16	34	207
(# presentations)	0	0	3	10	152
(# web files)	0	0	0	2	2,453
(# media files)	0	6	41	129	5,532
(# other files)	0	7	25	87	10,060
Total # folders	3	9	27	95	3,185
Unique file extensions	5	12	15	24	75

Table 3: Characteristics of participants' cloud accounts.

We chose to investigate personal file collections in cloud accounts because of the uniform and comprehensive APIs that Google Drive and Dropbox provide. Past work has noted that cloud accounts represent only part of a user's fragmented file collection [25], so our results may not generalize to other types of file collections. Notably, the types of files present, the organizational structure, and the usage context may all differ in local storage. Lastly, asking participants sequentially about perceived similarity and desired co-management may have biased them to identify similarity or co-management when they would not have done so otherwise. Future work should build on the lessons learned to investigate perceived similarity and co-management in a more naturalistic setting.

5 PARTICIPANTS AND THEIR ACCOUNTS

Here, we describe our participants and their cloud accounts.

5.1 Participant Demographics

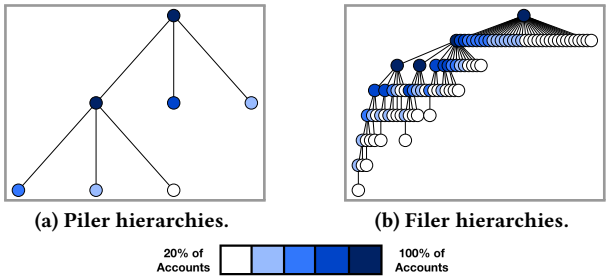
In total, 50 participants completed our full protocol. Among participants, 54.0% were female, 40.0% were male, and 6.0% were non-binary. The most common age range was 25–34 years old (48.0%) and the second most common was 18–24 years old (30.0%). Among participants, 26.0% had held a job or taken a course in computer science. For the study, 92.0% of participants used Google Drive, while 8.0% used Dropbox. Most participants (98.0%) reported using their service's web app to access their account. 60.0% reported using a mobile app, and 30.0% reported having automatic sync enabled. Participants reported being daily (34.0%), weekly (50.0%), or monthly users (14.0%) of their account; one participant chose not to respond. On average, participants estimated that their account contained 74.6% personal data and 25.4% professional data.

5.2 Participants' Cloud Accounts

Table 3 reports general characteristics of participant accounts. Our 50 participants collectively stored 119,388 files in their accounts. The median account was 8 years old and contained 5 gigabytes of data. Across accounts, we observed 341 unique file extensions. The most common file type was images (72,125 files), mostly .jpg (46,019) and .png (22,422) files. Second was a catch-all "other" category (14,729). The most common "other" file extension was flat (3,664), which is for database files, with .json (580) and .zip (402) as next most common. Documents (13,405) and media files (12,334) followed. Following trends observed in prior work on file collections [32],

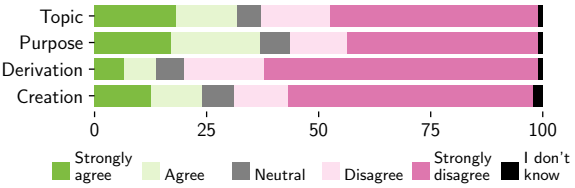
	Filer	Piler
e^{μ} # files	1,146	311
e^{μ} # folders	75	6
Mean files per folder	25	69
Mean depth	3.07	1.01
Mean breadth	9.13	2.37
e^{μ} # unique file extensions	26.00	11.00
e^{μ} # unique folders per extension	8.44	1.32
e^{μ} # unique extensions per folder	1.86	4.13

Table 4: Comparison of account characteristics by hierarchy type. e^{μ} is the adjusted mean of the distribution. *Depth* is the number of clicks needed to reach a file from the root. *Breadth* is the number of subfolders in a folder.

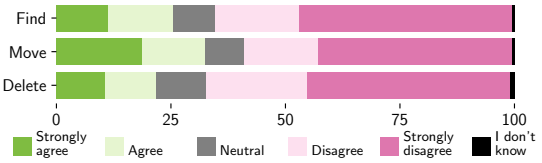


(a) Piler hierarchies. **(b) Filer hierarchies.**

Figure 1: The typical folder structure of piler and filer hierarchies. These trees merge participants' file structures, coloring nodes by the percentage of participants with that hierarchy type who had a node at that location. Nodes appearing for < 20% of participants with that hierarchy were pruned.



(a) Perceived similarity.



(b) Co-management

Figure 2: Participants' agreement that file pairs exhibited the four types of perceived similarity (top) or that the files should be co-managed in each of the three ways (bottom).

many characteristics were lognormally distributed, causing a large gap between the 75th percentile (Q_3) and the maximum value. We therefore report an adjusted mean (e^{μ} [32]) where appropriate. Due to sampling differences, we leave a comparison against the scale and structure of file collections in local storage to future work.

Participants were split on whether they considered their account well-organized: 36.0% reported that their account is well-organized, 40.0% disagreed, and 24.0% were neutral. Many well-organized participants justified their self-perception with their usage of folders (*"I name the folder of the topic what the photos or files fall under."*). They also reported strategies like organizing files by date. In contrast, some disorganized participants chose not to use folders (*"With text search and picture view I find it irrelevant"*), or reported difficulties doing so (*"I have folders I use to split up files, but I threw everything up there... and I have to go back and reorganize it"*).

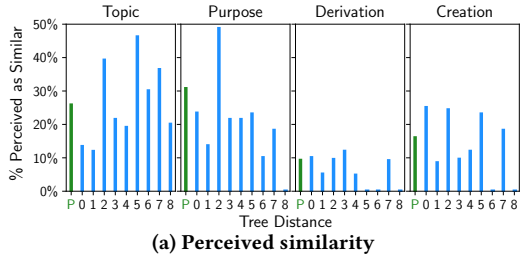
We examined the scale and structure of participants' accounts, as well as participants' free-text responses concerning organization, and found that participants' accounts naturally split into two groups matching those found in prior work [21, 38, 39, 51, 57, 72]. Participants had 50 folders on average, with a standard deviation of 531.9, while the median participant had 27 folders. Combining k-means clustering on the number of folders per participant with free-text responses yielded a cluster threshold of 10. We thus term accounts containing 10 or fewer folders **pilers** and accounts with over 10 folders **filers**. Among participants, 28.0% were pilers, while 72.0% were filers. Figure 1 visualizes the typical folder hierarchy for both classes. Each node in the tree represents a folder, colored proportional to the percentage of accounts that contained such a folder. We pruned all nodes that appeared in under 20% of accounts. As shown in Figure 1, piler hierarchies typically contained the root directory and one or two sub-folders. In contrast, most filer hierarchies contained many sub-folders and a few deeper branches. Table 4 further quantifies differences between piler and filer hierarchies. We investigate how these differences in hierarchy relate to similarity and co-management in Sections 6–7.

6 ACCOUNT ORGANIZATION

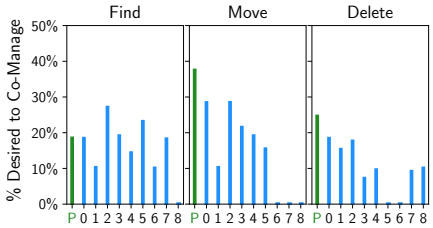
In this section, we present participant's overall responses about the perceived similarity and desired co-management of file pairs. We also investigate how pairs of files that participants perceived as similar, pairs of files that appeared similar in terms of data similarity features, and pairs of files that participants wanted to co-manage were distributed in the file hierarchy. If files were organized tightly by similarity, files that are similar would be located in the same folder, and files that are not similar would be located in different folders. We observed, however, that files that were similar in both participant perception and data characteristics, as well as files that participants wanted to co-manage, were distributed throughout the file hierarchy. This result highlights the need for recommender systems to help users co-manage files in cloud accounts.

6.1 Analysis of Responses Overall

Figure 2a displays the distribution of participants' responses for the perceived similarity of the 900 file pairs they labeled. Each response was on a five-point Likert scale. Participants perceived file pairs as similar (responded "strongly agree" or "agree") in at least one of the four dimensions 46.9% of the time. Among our similarity dimensions, participants most often perceived pairs as similar in purpose (36.9% of pairs). Note that our stratified sampling approach was purposely biased to identify more similar file pairs. Among only the 417 file pairs selected randomly, participants perceived 39.0% as similar in at least one dimension, and 29.7% as similar in purpose. This proportion is likely closer to the underlying distribution. We



(a) Perceived similarity



(b) Co-management

Figure 3: The percentage of files participants perceived as similar (top) or desired to co-manage (bottom) either in piler (*P*) hierarchies or broken out by tree distance (*numbers*) in filer hierarchies. Both "strongly agree" and "agree" responses indicate similarity or co-management here. So that the percentages are meaningful, we only consider file pairs selected either randomly or based on tree distance, not those selected based on having similar features.

also note that perceived similarity differs significantly by dimension, ranging between 13.8% for derivation to 36.9% for purpose.

Figure 2b displays the distribution of participants' ratings about their desire to co-manage the 900 file pairs. The trends in perceived similarity hold here as well. Participants infrequently wanted to co-manage files, and the rates at which they did varied by the type of co-management. For randomly selected file pairs, participants desired to find, move, or delete files together for 19.7%, 26.6%, and 15.8% of file pairs, respectively, less than in our stratified sample.

6.2 Perceived Similarity in the File Hierarchy

Surprisingly, files that participants perceived as similar were often found in very different parts of the file hierarchy. Many of our analyses are based on **tree distance**, or the minimum number of transitions (to parent or child folders) to get from one folder to the other. Files in the same folder have tree distance 0, while files in adjacent folders have tree distance 1.

Figure 3a shows the distribution of file pairs perceived as similar with respect to tree distance in both piler and filer hierarchies. In filer hierarchies, 46.6% of file pairs that participants perceived as similar in at least one dimension had tree distance > 2, and 19.5% had tree distance ≥ 5. Participants frequently described files located far apart in the file hierarchy as very similar. For instance, for a file pair with tree distance 13, a participant wrote, *"These files are very similar. They are both songs that I like, by artists I like. They are a similar genre."* For all four types of perceived similarity, at least 92.5% of pairs at tree distance 2 were in "sibling" folders (i.e., the files' parent folders share the same parent). This organization pattern was described in prior work [32, 71] as a technique users employ to gradually filter into more fine-grained categories.

Because our stratified sampling targeted file pairs more likely to be similar than a random file pair, Figure 3 likely overestimates file similarity. We therefore examined the subset of file pairs that were sampled either randomly or only based on tree distance, finding similar trends. Of the file pairs sampled in this way that were perceived as similar in at least one dimension, 39.7% had Tree Distance > 2, while 19.8% had Tree Distance ≥ 5. In sum, we found that similar files are often not located in the same folder, and they are sometimes located quite far in the file hierarchy.

6.3 Co-management in the File Hierarchy

Figure 3b shows similar trends in participants' desire to co-manage files at different tree distances. Of file pairs that participants wanted to co-manage (find, move, or delete together), 40.0%, 37.7%, and 36.6%, respectively, had tree distance > 2. Of these files, 17.1%, 10.6%, and 14.9%, respectively, had tree distance ≥ 5. A participant described the similarity between files they wanted to move together (despite a tree distance of 7) as, *"They are both trainings but we need to keep them by month for our grant."* We also examined only the file pairs that were selected randomly, finding similar trends.

6.4 Data Similarity in the File Hierarchy

Finally, we explored the relationship between data similarity features and tree distance. We analyzed 11,653,450 pairs of data similarity features for all file types, with an additional 4,519,675 pairs for image similarity features, 4,262,444 for text similarity features, and 39,333 for table similarity. To our knowledge, this is the first large-scale analysis of data similarity in cloud storage. Figure 4 shows the relationship between data similarity and tree distance. We make two key observations. First, many more file pairs are dissimilar than are similar. Second, tree distance does not appear to correlate strongly with data similarity features. Intuitively, if users categorize files within a file hierarchy with similar files close to each other in the hierarchy, then one would expect to see median similarity decrease with tree distance. This does not occur here.

Taken together, these analyses emphasize that files that participants perceive as similar, files that participants wish to co-manage, and files that look similar in terms of algorithmically extractable features are all often located far apart in the file hierarchy. Potential explanations for the phenomenon itself include the following: the existence of distinct, but overlapping file hierarchies [22, 41]; the desire of users to categorize a file in multiple ways, but choosing one by necessity of the interface [14]; and the existence of partially categorized files [56]. We leave further investigation of root explanations to future work. Regardless, the dispersed locations of similar files will inhibit future retrieval without improved tools.

7 SIMILARITY IMPLIES CO-MANAGEMENT

We found that a file pair's perceived similarity strongly correlated with whether a participant wished to co-manage it. Specifically, participants wished to co-manage similar file pairs at a much higher rate than dissimilar pairs. Because participants expressed perceptions of similarity and desire to co-manage on 5-point Likert scales, we tried binarizing their preference in two ways: based on both

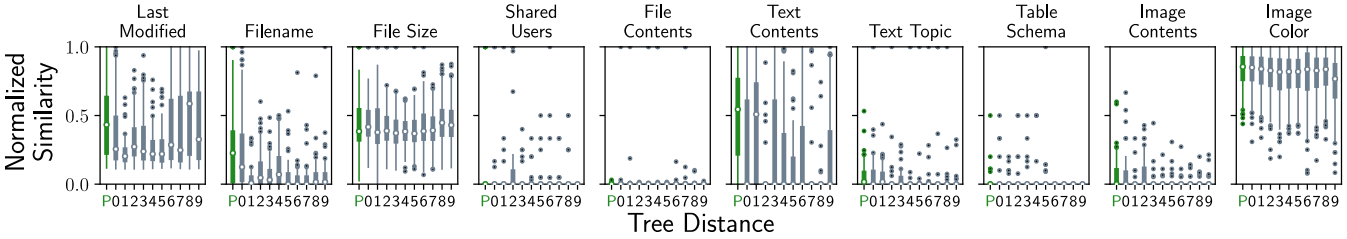
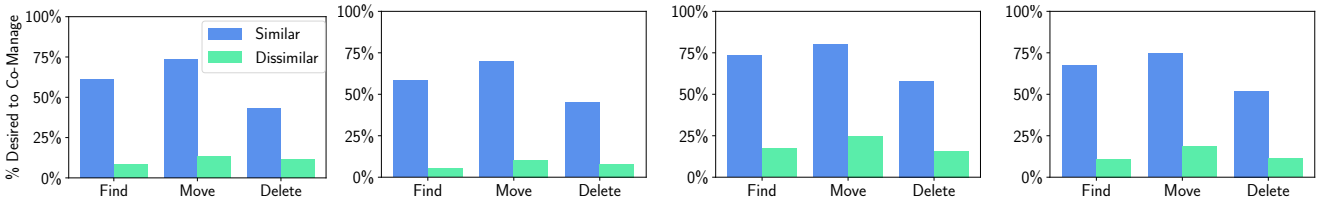
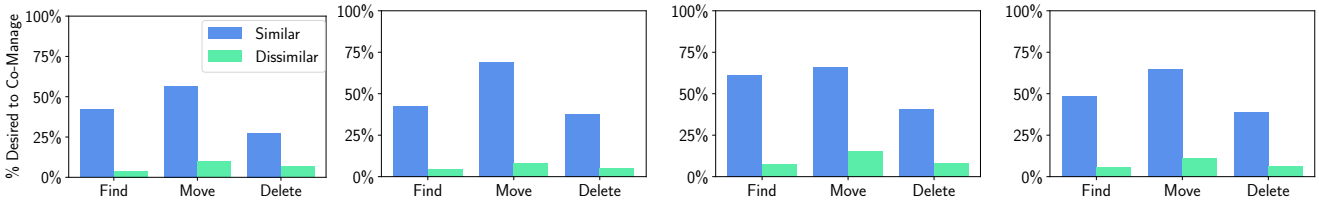


Figure 4: Box plots depicting how each class of data similarity is distributed for all file pairs in participants' accounts. The box plot labeled *P* shows the distribution for all pairs in piler accounts. The remaining box plots represent the distribution in filer accounts at the tree distance specified by the label (e.g., "0" represents the distribution for file pairs in the same directory).



(a) Topic **(b) Purpose** **(c) Derivation** **(d) Creation Context**

Figure 5: How participants' desire to co-manage files correlated with their perceptions of files as similar in one of our four dimensions of perceived similarity. We binned "strongly agree" and "agree" responses as similar / to be co-managed.



(a) Topic **(b) Purpose** **(c) Derivation** **(d) Creation Context**

Figure 6: This figure is the same as Figure 5, but considers only "strongly agree" responses for similarity/co-management.

strong and mild preferences ("strongly agree" and "agree" responses indicated similarity/co-management, Figure 5) or only on strong preferences (only "strongly agree" responses, Figure 6). Comparing the figures, the correlation between similarity and co-management held regardless of preference strength. For example, among file pairs perceived as similar in creation context (strong and mild), participants wanted to co-move 74.7% of them, whereas they only wanted to co-move 18.4% of dissimilar file pairs. For strong preferences only, participants desired to co-move 64.6% of similar pairs, versus 11.1% of dissimilar pairs. The relationship between similarity and co-management was statistically significant regardless of similarity or co-management type (Spearman's rank correlation test, all $p < 0.001$). In the remainder of the paper we binarize based on both strong and mild preferences unless stated otherwise.

However, correlation between similarity and co-management was not perfect; participants also wished to co-manage some dissimilar file pairs. Among pairs that participants wished to co-find, 23.6%, 60.3%, 15.3%, and 36.2% were dissimilar in topic, derivation, purpose, and creation context, respectively. Some were similar in another dimension (*"Same student, but the content is much different"*), but many were explicitly dissimilar (*"They are dissimilar because File 1 is for dissertation and File 2 is for my job"*).

Overall, this evidence suggests that co-management tools based on perceived similarity and informed by data similarity might be able to identify files participants wish to co-manage and would not naturally discover. We discuss in Section 9 how this might inform the development of future tools for file management.

8 MODELING BASED ON DATA SIMILARITY

While the previous section highlighted the connection between perceived similarity and co-management, this insight is difficult to act on because perceived similarity is a "human" value. Thus, we built regression models to correlate algorithmically extractable features (data similarity) with perceived similarity and desired co-management. We found several features to be highly predictive.

Table 5 gives the odds ratios for our logistic regressions. These coefficients can be interpreted as the multiplicative increase in the probability that the response variable will be one level higher (e.g., "agree" to "strongly agree") for an increase of 1 in the data similarity value. All of our data similarity values are normalized to a $[0, 1]$ scale, and all distance metrics are turned into similarity metrics by subtracting their distance from the maximum value of 1. Therefore, the odds ratio is the multiplicative increase if a value has full similarity in that dimension, versus none.

	Perceived Similarity				Co-management		
	Topic	Purpose	Derivation	Creation	Find	Move	Delete
Data Similarity							
Last Modified	20.769***	11.422***	3.207**	17.075***	12.034***	12.618***	7.623***
Filename	3.978**	12.699***	18.094***	12.805***	10.885***	13.744***	4.286***
File Size	0.985	1.718	1.829	1.507	1.379	1.656	2.231*
Tree Distance	0.471	0.588	0.863	1.496	0.409	0.500	1.090
Shared Users	2.428**	2.874***	2.857**	3.124***	6.833***	7.218***	5.604***
File Contents	2.855**	3.473***	4.172***	3.703***	2.536**	2.027*	2.197**
Text Topic	3.072***	2.592**	2.315*	2.260*	1.707	2.588**	1.526
Table Schema	3.965	13.076*	1.845	2.393	2.905	1.951	3.993
Image Contents	36.777***	29.018***	8.938***	8.757***	13.085***	10.106***	2.767
Filer Hierarchy	0.884	1.255	1.078	2.036	1.385	0.704	0.789
Random Effects							
σ of random effect	1.095	0.680	1.519	1.140	1.409	1.251	1.416

Table 5: Our regression models showing odds ratios for data similarity features (* $p < .001$; ** $p < .01$; * $p < .05$).**

Some features, such as similarities in last modified times, are known to be predictive [6, 34, 35, 50, 70]. Others, such as image contents and filename, have rarely been used. Shared users, file contents, and text topic features were also statistically significant, but with smaller effect sizes. That tree distance was not a significant predictor matches evidence from prior sections. We also found no significant effect for whether a hierarchy was a piler or filer, suggesting that the importance of data similarity features may hold across both types of hierarchies. The size of the random effects indicates that individual variations between participants accounted for approximately half a point change in the mean Likert-scale rating of file pairs. This result suggests that user-specific features (e.g., personality, mood [52, 76]) may affect perceived similarity.

Many factors that were predictive of perceived similarity were also predictive of co-management. One exception was the image contents feature, which was not predictive of co-deletion, though this may be an artifact of our sample size. Future tools should leverage these features' predictiveness in supporting co-management.

9 DISCUSSION AND CONCLUSION

We investigated whether similarity can support co-management via an online study of 50 Google Drive and Dropbox users. We found that similar files were distributed across the file hierarchy, and that a user's perception of similarity between two files correlated with their desire to co-manage those files. We explored through regression analysis the ability of data similarity to predict perceived similarity and co-management. Last Modified, Image Contents, Filename, Shared Users, and Text Topic features were significant.

We believe our results demonstrate the need for, and feasibility of, similarity-driven co-management support. We propose the following design principles for future work on co-management:

Recommendations must work beyond retrieval. Though prior work has left recommendations for actions like movement or deletion under-explored, related work in email has shown that such operations are important to user workflows [3, 4, 59]. Beyond the usability benefits (improving previously identified issues [75]), supporting these operations can also improve privacy/security by ensuring unnecessary content is archived or removed [42, 43].

Recommendations must work across the hierarchy. Participants wanted to co-manage files located both close and far in the file hierarchy. Previous tools, such as Fitchett et al.'s enhanced finder interface [35], only highlighted file or folder icons in the current folder. Our results show that this leaves significant functionality untouched; users might overlook files in other folders. Therefore, interface enhancements like those used in BIGFile [50], with adaptive split-screen interfaces, are more appropriate for recommendations independent of file location. Recommendations could provide context and explanations, as in Xu et al.'s work [78].

Recommendations must work for both Piler and Filer hierarchies. As in the previous point, highlighting icons would likely be inappropriate in a piler hierarchy. There are likely to be many files in a single folder, and highlighted icons would not be sufficiently visible or provide context. On the other hand, in filer hierarchies, files are likely to be further apart, and it would be important to provide context on where co-managed files live (e.g., showing a visualization of the file hierarchy). Tools implementing co-management must support both types of contextual feedback.

Recommendations must use features beyond access patterns. Access patterns are a highly informative feature. In fact, many prior studies and tools restrict the scope of recommendations to recently accessed files [19, 20, 70]. However, users have difficulty retrieving older or infrequently accessed files [75], which are of interest to users [40]. Access patterns are unlikely to be an informative feature for these files. Future work should direct energy toward extracting predictive metadata and content features from users' cloud accounts in order to improve tooling.

Furthermore, future work should identify what context users need to evaluate recommendations. For example, a user might not accept a co-movement recommendation if they forget what files are in the destination folder. This mirrors issues with tag-based file systems, where a lack of spatial context impedes usability [9, 15, 28]. Examining different forms of context for each action (movement, deletion, etc.) will be an important component of engineering co-management support. Our own future work will use these lessons to build and evaluate a co-management tool for cloud storage.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. CNS-1801663 and OAC-1835890, as well as the CERES Center for Unstoppable Computing. We thank Sophie Welber, Valerie Zhao, and Michelle Aninye for assisting with data analysis and visualization.

REFERENCES

[1] 2021. Online appendix. https://www.blaseur.com/papers/sigir21-appendix.pdf

[2] David Abrams, Ron Baecker, and Mark Chignell. 1998. Information archiving with bookmarks: Personal web space construction and organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[3] Tarfah Alrashed, Ahmed Hassan Awadallah, and Susan Dumais. 2018. The lifetime of email messages: A large-scale analysis of email revisitation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*.

[4] Tarfah Alrashed, Chia-Jung Lee, Peter Bailey, Christopher Lin, Milad Shokouhi, and Susan Dumais. 2019. Evaluating user actions as a proxy for email significance. In *Proceedings of the World Wide Web Conference*.

[5] Xinlong Bao and Thomas G. Dietterich. 2011. FolderPredictor: Reducing the cost of reaching the right folder. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 1 (2011).

[6] Xinlong Bao, Jonathan L. Herlocker, and Thomas G. Dietterich. 2006. Fewer clicks and less frustration: Reducing the cost of reaching the right folder. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*.

[7] Deborah Barreau and Bonnie A. Nardi. 1995. Finding and reminding: File organization from the desktop. *ACM SigChi Bulletin* 27, 3 (1995), 39–43.

[8] Deborah K. Barreau. 1995. Context as a factor in personal information management systems. *Journal of the American Society for Information Science* 46, 5 (1995).

[9] Yael Benn, Ofer Bergman, Liv Glazer, Paris Arent, Iain D. Wilkinson, Rosemary Varley, and Steve Whittaker. 2015. Navigating through digital folders uses the same brain structures as real world navigation. *Scientific Reports* 5, 1 (2015).

[10] James Bennett, Stan Lanning, et al. 2007. The Netflix prize. In *Proceedings of the KDD Cup and Workshop*.

[11] Ofer Bergman, Ruth Beyth-Marom, and Rafi Nachmias. 2003. The user-subjective approach to personal information management systems. *Journal of the American Society for Information Science and Technology* 54, 9 (2003), 872–878.

[12] Ofer Bergman, Ruth Beyth-Marom, and Rafi Nachmias. 2008. The user-subjective approach to personal information management systems design: Evidence and implementations. *Journal of the American Society for Information Science and Technology* 59, 2 (2008), 235–246.

[13] Ofer Bergman, Ruth Beyth-Marom, Rafi Nachmias, Noa Gradovitch, and Steve Whittaker. 2008. Improved search engines and navigation preference in personal information management. *ACM Transactions on Information Systems* 26, 4 (2008).

[14] Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, and Ruth Beyth-Marom. 2013. Folder versus tag preference in personal information management. *Journal of the American Society for Information Science and Technology* 64, 10 (2013), 1995–2012.

[15] Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, and Ruth Beyth-Marom. 2013. Tagging personal information: A contrast between attitudes and behavior. In *Proceedings of the 76th ASIS&T Annual Meeting*.

[16] Ofer Bergman, Maskit Tene-Rubinstein, and Jonathan Shalom. 2013. The use of attention resources in navigation versus search. *Personal and Ubiquitous Computing* 17, 3 (2013), 583–590.

[17] Ofer Bergman, Simon Tucker, Ruth Beyth-Marom, Edward Cutrell, and Steve Whittaker. 2009. It's not that important: demoting personal information of low subjective importance using GrayArea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[18] Ofer Bergman, Steve Whittaker, and Yaron Frishman. 2019. Let's get personal: The little nudge that improves document retrieval in the cloud. *Journal of Documentation* (2019).

[19] Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2010. The effect of folder structure on personal file navigation. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2426–2441.

[20] Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2012. How do we find personal files? The effect of OS, presentation & depth on file navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[21] Richard Boardman and M. Angela Sasse. 2004. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[22] Richard Boardman, Robert Spence, and M. Angela Sasse. 2003. Too many hierarchies? The daily struggle for control of the workspace. In *Proceedings of the 10th International Conference on Human-Computer Interaction*.

[23] Will Brackenbury, Rui Liu, Mainack Mondal, Aaron J. Elmore, Blase Ur, Kyle Chard, and Michael J. Franklin. 2018. Draining the Data Swamp: A Similarity-based Approach. In *Workshop on Human-In-the-Loop Data Analytics*.

[24] Sergio Canuto, Thiago Salles, Thierson C. Rosa, and Marcos A. Gonçalves. 2019. Similarity-based synthetic document representations for meta-feature generation in text classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[25] Robert Capra and M.A. Perez-Quinones. 2006. Factors and evaluation of refinding behaviors. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[26] Suming Jeremiah Chen, Zhen Qin, Zachary Teal Wilson, Brian Lee Calaci, Michael Richard Rose, Ryan Lee Evans, Sean Robert Abraham, Don Metzler, Sandeep Tata, and Mike Colagrosso. 2020. Improving recommendation quality at Google Drive. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[27] Paul-Alexandru Chirita, Stefania Costache, Wolfgang Nejdl, and Raluca Paiu. 2006. Beagle++: Semantically enhanced searching and ranking on the desktop. In *Proceedings of the European Semantic Web Conference*.

[28] Andrea Civan, William Jones, Predrag Klasnja, and Harry Bruce. 2008. Better to organize personal information by folders or by tags?: The devil is in the details. *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008).

[29] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*.

[30] Gabor Cselle, Keno Albrecht, and Rogert Wattenhofer. 2007. BuzzTrack: Topic detection and tracking in email. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*.

[31] Jesse David Dinneen and Charles-Antoine Julien. 2019. The ubiquitous digital file: A review of file management research. *Journal of the Association for Information Science and Technology* (2019).

[32] Jesse David Dinneen, Charles-Antoine Julien, and Ilja Frissen. 2019. The scale and structure of personal file collections. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[33] Susan Dumais, Edward Cutrell, Jonathan J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[34] Stephen Fitchett and Andy Cockburn. 2012. Accessrank: Predicting what users will do next. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[35] Stephen Fitchett, Andy Cockburn, and Carl Gutwin. 2014. Finder highlights: Field evaluation and design of an augmented file browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[36] Qin Gao. 2011. An empirical study of tagging for personal information organization: Performance, workload, memory, and consistency. *International Journal of Human-Computer Interaction* 27, 9 (2011), 821–863.

[37] Google. 2019. Google Vision API. https://cloud.google.com/vision/.

[38] Sharon Hardof-Jaffe, Arnon Hershkovitz, Hama Abu-Kishk, Ofer Bergman, and Rafi Nachmias. 2009. Students' organization strategies of personal information space. *Journal of Digital Information* 10, 5 (2009).

[39] Sarah Henderson and Ananth Srinivasan. 2009. An empirical analysis of personal digital document structures. In *Proceedings of the Symposium on Human Interface*.

[40] Farnaz Jahanbakhsh, Ahmed Hassan Awadallah, Susan T. Dumais, and Xuhai Xu. 2020. Effects of past interactions on user experience with recommended documents. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*.

[41] William Jones, Ammy Jiranida Phuwanartnurak, Rajdeep Gill, and Harry Bruce. 2005. Don't take my folders away!: Organizing personal information to get things done. In *Proceedings of the CHI '05 Extended Abstracts on Human Factors in Computing Systems*.

[42] Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. 2018. Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[43] Mohammad Taha Khan, Christopher Tran, Shubham Singh, Dimitri Vasilkov, Chris Kanich, Blase Ur, and Elena Zheleva. 2021. Helping users automatically find and manage sensitive, expendable files in cloud storage. In *Proceedings of the 30th USENIX Security Symposium*.

[44] Barbara H. Kwasnik. 1989. How a personal document's intended use or purpose affects its classification in an office. In *ACM SIGIR Forum*, Vol. 23. 207–210.

[45] Barbara H. Kwasnik. 1991. The importance of factors that are not document attributes in the organization of personal documents. *Journal of Documentation*. (1991).

[46] Barbara H. Kwasnik. 1992. The role of classification structures in reflecting and building theory. *Advances in Classification Research Online* 3, 1 (1992), 63–82.

[47] Mark W. Lansdale. 1988. The psychology of personal information management. *Applied Ergonomics* 19, 1 (1988), 55–66.

[48] Paul J. Lavrakas. 2008. *Encyclopedia of survey research methods*. Sage Publications.

[49] Bongshin Lee and Benjamin B. Bederson. 2003. *Favorite folders: A configurable, scalable file browser*. Technical Report. University of Maryland.

[50] Wanyu Liu, Olivier Rioul, Joanna Mcgrenere, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2018. BIGFile: Bayesian information gain for fast file retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[51] Thomas W. Malone. 1983. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems* 1, 1 (1983), 99–112.

[52] Charlotte Massey, Sean TenBrook, Chaconne Tatum, and Steve Whittaker. 2014. PIM and personality: What do our personal file systems say about us?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.
[54] David R. Millen, Jonathan Feinberg, and Bernard Kerr. 2006. Dogear: Social bookmarking in the enterprise. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[55] David R. Millen, Meng Yang, Steven Whittaker, and Jonathan Feinberg. 2007. Social bookmarking and exploratory search. In *Proceedings of the 10th European Conference on Computer-Supported Cooperative Work*.
[56] Kyong Eun Oh. 2012. What happens once you categorize files into folders? *Proceedings of the American Society for Information Science and Technology* (2012).
[57] Kyong Eun Oh. 2017. Types of personal information categorization: Rigid, fuzzy, and flexible. *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1491–1504.
[58] Michael Oppermann, Robert Kincaid, and Tamara Munzner. 2020. VizCommender: Computing text-based similarity in visualization repositories for content-based recommendations. *arXiv:2008.07702* (2020).
[59] Soya Park, Amy X. Zhang, Luke S. Murray, and David R. Karger. 2019. Opportunities for automating email processing: A need-finding study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[60] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.
[61] Prolific. 2019. https://www.prolific.co/.
[62] Dennis Quan, David Huynh, and David R. Karger. 2003. Haystack: A platform for authoring end user semantic web applications. In *Proceedings of the International Semantic Web Conference*.
[63] Bradley Rhodes and Thad Starner. 1996. *Remembrance Agent: A continuously running automated information retrieval system.* Technical Report. AAAI.
[64] Leo Sauermann, Gunnar Aastrand Grimnes, Malte Kiesel, Christiaan Fluit, Heiko Maus, Dominik Heim, Danish Nadeem, Benjamin Horak, and Andreas Dengel. 2006. Semantic desktop 2.0: The gnowsis experience. In *Proceedings of the International Semantic Web Conference*.
[65] Markus Schröder, Christian Jilek, and Andreas Dengel. 2019. Interactive concept mining on personal data. *arXiv:1903.05872* (2019).
[66] Richard B. Segal and Jeffrey O. Kephart. 1999. MailCat: An intelligent assistant for organizing e-mail. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.
[67] Debmalya Sinha and Anupam Basu. 2012. Gardener: A file browser assistant to help users maintaining semantic folder hierarchy. In *Proceedings of the 4th International Conference on Intelligent Human Computer Interaction*.
[68] Brent Smith and Greg Linden. 2017. Two decades of recommender systems at Amazon.com. *IEEE Internet Computing* 21, 3 (2017).
[69] John C. Tang, Eric Wilcox, Julian A Cerruti, Hernan Badenes, Stefan Nusser, and Jerald Schoudt. 2008. Tag-it, snag-it, or bag-it: combining tags, threads, and folders in e-mail. In *Proceedings of the CHI '08 Extended Abstracts on Human Factors in Computing Systems*.
[70] Sandeep Tata, Alexandrin Popescul, Marc Najork, Mike Colagrosso, Julian Gibbons, Alan Green, Alexandre Mah, Michael Smith, Divanshu Garg, Cayden Meyer, et al. 2017. Quick access: Building a smart experience for Google drive. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[71] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[72] Francesco Vitale, Izabelle Janzen, and Joanna McGrenere. 2018. Hoarding and minimalism: Tendencies in digital data preservation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
[73] Roger Whitham and Leon Cruickshank. 2017. The function and future of the folder. *Interacting with Computers* 29, 5 (2017), 629–647.
[74] Steve Whittaker. 2011. Personal information management: From information consumption to curation. *Annual Review of Information Science and Technology* 45, 1 (2011).
[75] Steve Whittaker, Ofer Bergman, and Paul Clough. 2010. Easy on that trigger dad: A study of long term family photo retrieval. *Personal and Ubiquitous Computing* 14, 1 (2010), 31–43.
[76] Steve Whittaker and Charlotte Massey. 2020. Mood and personal information management: How we feel influences how we organize our information. *Personal and Ubiquitous Computing* 24, 1 (2020), 695–707.
[77] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems* 26, 3 (2008).
[78] Xuhai Xu, Ahmed Hassan Awadallah, Susan T. Dumais, Farheen Omar, Bogdan Popp, Robert Rounthwaite, and Farnaz Jahanbakhsh. 2020. Understanding user behavior for document recommendation. In *Proceedings of The Web Conference*.
[79] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*.