Shared Files – The Scalability Perspective / Tamar Israeli

Abstract

File sharing is an important way to distribute information within a group. This activity allows a group of people to access a particular file and grants them special rights, such as the ability to view, edit, or delete the file. When a group of two or more people starts collaborating, they typically face a dilemma of how to share the files they create together. Groups need to choose between storing the shared files in a common repository (e.g. using cloud-based services such as Google Drive and Dropbox) or distributing the files as email attachments and then storing them in personal repositories. If they choose a common repository, the group first needs to agree on the file organization method, which we refer to as Group Information Management (GIM). If they choose email distribution and personal repositories, each person in the group can organize the files in his/her own way, and we refer to this as PIM (Personal Information Management).

There are good theoretical arguments for the use of each one of these methods: GIM supporters argue that managing personal information requires more work: collaborators must independently manage their own personal collection of shared files, thus duplicating files, time and cognitive effort. Furthermore, there may be significant problems involved in retrieving, managing, and reconciling different versions of a document when multiple versions are distributed by email to multiple participants (Ducheneaut & Bellotti, 2001; Whittaker, Bellotti, & Gwizdka, 2007). PIM supporters point to problems in Group Information Management. Teams cannot agree on a common organizational scheme, making it difficult to retrieve information organized by others (Berlin et al., 1993; Lutters, Ackerman, & Zhou, 2007; Rader, 2009; A. Voida et al., 2013).

Contradictions were found between two studies that examined GIM retrieval efficiency. Bergman et al. found that PIM retrieval was significantly more successful than GIM. This explains why participants preferred to share files via email (Bergman, Whittaker, & Falk, 2014). Massey et al. examined the effectiveness of group information management among work teams in the organization. Their participants felt they successfully managed their shared repository using a few simple strategies and did not experience problems when retrieving shared files (Massey, Lennig, et al., 2014).

Library of Information Science
Bar-Ilan University, Ramat-Gan, Israel 5290002-טל' 972
Email: Ruthi.Tshop@biu.ac.il

הספריה למדעי המידע
אוניברסיטת בר-אילן, רמת גן 5290002
Tel. 3-5318163

www.is.biu.ac.il/library

One of the explanations for the contradiction between these findings is that Massey et al's research was done in a low workload environment. Participants were grouped in small teams; they were familiar with the tasks and responsibilities of the other team members and used relatively small active repositories. However, the researchers expressed doubt as to whether the simple strategies used to successfully organize and retrieve files would prove efficient under more complex conditions. This gave rise to the question: Is GIM scalable?

Scalability is the ability to perform efficiently when facing growing amounts of work, or its enlargement potential in order to accommodate such growth (Bondi, 2000). A system is considered scalable if it works effectively under an increased load. Scalability is a major problem in information retrieval and was the basis for this study. We aimed to study the effect of magnitude variables (such as collection size, the number of file versions, the number of people that collaborate, working hours) on retrieval success and efficiency, and which sharing-method - PIM or GIM withstanding size problems better, and therefore is more scalable?

Several articles suggested that the magnitude of file collection can hinder retrieval and cause a scalability problem. However, to the best of our knowledge, the effect of collection size on retrieval success and efficiency has never been systematically evaluated. Using an experiment, we tested the magnitude effects of each variable on file retrieval and compared the two sharing methods.

Another aim was to examine the coping strategies used and whether they mitigate the negative impact of magnitude on shared file retrieval.

To answer these questions, we examined variables whose magnitude could negatively impact file retrieval. These include variables related to collection size (number of files and folders), variables related to the target file (number of versions, number of collaborators, days since recent retrieval, folder depth and physical distance between participants), variables related to workload (subjective perception of workload, work hours, and number of email messages sent and received), personality related variables and background variables. Our dependent variables were: failure rate, retrieval time and percentage of retrievals with misstep/s. Our hypothesis was that PIM would be more scalable than GIM in retrieval success and efficiency as workload increases.

We used a mixed method research design. In the quantitative part we used an experiment and a questionnaire. Our 289 participants were tested using their own shared files on their own computers by using designated software. Participants were asked to retrieve selected files from a list of shared files they previously accessed. Retrieval success and efficiency were examined, comparing the two methods, PIM and GIM under different load conditions. The experiment was accompanied by a questionnaire with questions related to subjective and objective workloads, personality and cognitive components of the participant, and background variables. Additionally, qualitative research was conducted using semi-structured interviews with twenty teachers and academics. During the interviews, participants were asked about their sharing method preferences, the problems they encounter when working with shared files, and their coping strategies.

Our 289 participants conducted 1,557 file retrievals. Of these files, 70% were PIM files and 30% were GIM files. Results indicate that PIM retrieval was more successful and efficient than GIM. Almost all variables related to the collection size and the target file were found to negatively affect retrieval: An increase in the number of files and folders negatively affected retrieval time. This finding is consistent with the intuitive assumption that multiple items make it harder to locate them. Other variable that affected retrieval time was multiple versions. File versions are created in the workflow. When the versions multiply it becomes difficult to spot the latest one. Use of different devices and work environments also contributes to multiple versions. Results indicate that keeping more than a single version of a file increases retrieval time. The number of collaborators was positively related to the percentage of failures. The more collaborators there are, the greater the need for coordination. We found that when there were more than five collaborators, the percentage of retrieval failures increased sharply. The number of days since last retrieval affected retrieval efficiency. Studies in cognitive psychology have shown that memory fades over time. It is easier to retrieve files that are currently in use than older files. However, the results showed that participants successfully retrieved files that had not been accessed for a long time, and that the time interval that may substantially increase the likelihood of failure was not days or weeks but rather months.

As opposed to the file related variables, most of the participant related variables (workload, personality variables, and background variables) had little effect on retrieval. Of these variables, the participants' age had a negative effect on retrieval time. Among the

personality traits we examined (such as *order tendency*, *memory*, *sociability* and *pressure resistance*) only one personality trait (*need for control*) affected retrieval.

Our study revealed several indications that PIM is more scalable when facing magnitude problems than GIM: (a) Having *file versions* had a significantly negative affect on retrieval time for GIM files but not for PIM files; (b) *subjective workload* magnitude leads to more retrievals with mistake/s for GIM files than for PIM files; and (c) decreased *need to be in control* resulted in higher retrieval time for GIM files but not PIM files; (d) the *number of emails sent and received* affected GIM retrieval failure more than PIM; and (e) the *number of days elapsed* since the target file was last retrieved affected retrieval time for GIM more than PIM.

Consistent with previous studies, we found that participants preferred folder navigation over search for their files retrieval, and that retrieval times using navigation were shorter. A surprising finding regarding the retrieval method was that male search percentages were nearly three times higher than for female, even though search requires verbal thinking which according to research literature is more characteristic of women.

Although the GIM retrieval problem is well-known in the research literature (e.g. Berlin et al., 1993; Capra et al, 2014), we found very little awareness of this in the study. Although many of our participants had just experienced the GIM retrieval problem during the experiment, they didn't report it. Accordingly, there was little use of coping strategies to address these problems.

The fact that magnitude has a different effect on PIM and GIM retrievals explains the contradiction between Massey et al's findings and those of Bergman et al's. Both sharing methods work reasonably well under low load conditions, but when there is an increase in the values of different variables, PIM is more scalable and can bear greater loads than GIM. This can be explained by the fact that in GIM, other people save the files according to their organization methods and the user finds it difficult to duplicate this method in order to access the files (Berlin et al., 1993; Lutter et al., 2007; Rader, 2009; Voida et al., 2013). When additional problems such as large collections, multiple collaborators, and long time since the last retrieval are added, the two issues interact and the memory of the file location fades faster than in PIM, where each user organizes his files in his/her own way.

These findings - that PIM is more scalable then GIM - have practical implementations for organizations that tend to grow over time. For example, a startup company usually begins

with a small number of founders working closely together. They may not find it particularly difficult to keep their files in a shared repository using GIM. However, these organizations tend to grow rapidly. When this happens, there is a rapid increase in the number of files, folders, collaborating members and number of old files. At this point, it will be harder for workers to retrieve their files, but they may not be aware of the problem and in any case, it is difficult to change established work practices. Our research findings indicate that it is worthwhile for a rapidly growing organization to consider using the PIM sharing method rather than the GIM one, as it scales up better when facing magnitude problems.

MMS Numbder: 990026144850205776

Library of Information Science
Bar-Ilan University, Ramat-Gan, Israel 972-טל'
Email: Ruthi.Tshop@biu.ac.il
הספריה למדעי המידע
אוניברסיטת בר-אילן, רמת גן 5290002
Tel. 3-5318163
www.is.biu.ac.il/library