

Towards Interactive File Browsing and Searching using Semantic Hierarchy

Pedro Teixeira, Daniel Gonçalves and Manuel J. Fonseca

Abstract. Navigation and browsing on a computer system is usually done using a hierarchically organized file system. However, this is not the most adequate method to search or locate a given file at a later time. To solve this, we propose a new approach for interactive browsing and searching of files, which takes advantage of the semantics associated to files. With this information extracted from files, we specify a semantic hierarchy, which is easier to understand and to memorize by users. We performed task analysis to understand how users browse, explore and search their files, and also to identify what file properties are more used to sort, search and remember files. From task analysis, we noticed that, excluding the name of the file, the file type is clearly the main property used, followed by the modification date and file size. Based on this information we created a semantic hierarchy, which we included in a prototype for interactive file browsing and searching. Testing showed that users could find their files faster and with fewer mistakes using our prototype than using Leap or the conventional browser.

Keywords: File browsing, File searching, Metadata, Semantic hierarchy, Visualization

1 Introduction

It is nowadays common for users to handle thousands of electronic documents. Unfortunately, the way in which those documents are organized makes this a cumbersome task. Indeed, documents are nothing more than files in a hierarchically organized file system. When trying to retrieve a file, users can resort to little more information than the file's location in the hierarchy. However, such a classification is fraught with problems. When storing a document, where to place it is often not a trivial decision. More than one place in the hierarchy (or no place at all!) might seem adequate. Also, what seems a good classification at one point in time might not be the one remembered at a later time. As such, finding the document can become impossible.

The aforementioned difficulties result from the fact that while users are handling documents, the computer handles files. A document is something a user remembers fairly well, as it was read or written for a reason, can have memorable contents, and was handled in a meaningful context. Files and hierarchical file systems have little relation with this, being more useful for the computer itself rather than for users.

In this paper, we present a novel approach, where users can browse and find documents, forgetting about directories and folders, concentrating in the documents themselves. To that end, we defined a semantic hierarchy created using the metadata associated to files (e.g. file type, modification date, size, author, subject, etc.), which are easier to use and to remember by users. In addition, we make navigation easier and more efficient by providing a user-centered visual navigation instead of a purely textual one.

In the remainder of the paper, we present and analyze some related works on file browsing, trying to understand what failed on trying to free users from the directory exploration. Section 3 presents the main results from task and user analysis, allowing us to understand what users know and do not know about files, creating a knowledge base for our solution. Next, we describe our solution based on a semantic hierarchy for interactive browsing and searching of files. In section 5 we describe the prototype developed. In section 6 we talk about the results obtained with testing, and finally in section 7 we present some conclusions and future work.

2 Related Work

To help users manage their files and documents several paradigms for file browsing and exploration were developed allowing the visualization of users' document collections in meaningful ways. These approaches, by moving away from the file system hierarchies, strive to convey an overall view of the users' documents. In this section we describe some of the solutions developed to tackle this problem.

PhotoMesa [1] is a solution for image browsing that relies on metadata (e.g. date, local, people, etc.) to group photos. It uses Treemaps [2] as the main visualization technique, flattening down folders and sub-folders to the same level. This lack of leveling increases the number of files shown at the same time, which is good for a small number of files but becomes unusable for larger collections (e.g. a file system). Although, PhotoMesa is a good browser for Photos, it is quite difficult to apply this kind of visualization and browsing to all file types.

FacetMap [3] is an application for browsing file systems, using Treemaps as the main visualization paradigm, and metadata from files (e.g. date, user, type of file, etc.) to define the exploration tree. This solution combines metadata, Treemaps and filtering to allow users to browse and search files in an easy way. Although, this idea of combining metadata and filtering is very interesting, the number of files that it can show at the same time is a drawback of the achieved solution.

PHLAT [4] is an interface for Windows Desktop Search that combines browsing and searching in one unique interface, through the creation of queries and filters. The user interface of PHLAT has the typical look from “Microsoft Office” applications, providing only text-boxes and checkboxes for users to filter, but not offering any

special visualization technique to give an overview of the returned results. This kind of interaction does not distinguish much PHLAT from common search applications, such as Live Search, Spotlight, Google Desktop, etc., which rely mainly on the name of the file and on its content.

Leap [5] is a Mac OS X application, which uses Spotlight [6] as basis to browse and search files. Leap premise is to forget the directory hierarchy and use file tags and file metadata to filter and organize results. However, due to the lack of relationship and organization between tags, users easily lose context and get lost in a collection of non-related metadata.

In addition to these solutions, there are others that try to create new ways for browsing and searching files, but they continue to be based on the directory hierarchy. Discovery [7] is a Treemap-based application, which shows the entire file system in the screen at the same time, using colors to distinguish different file types. StepTree [8] is basically similar but shows the information in 3D. Cone Trees [9] is a 3D visualization approach that shows a 3D tree of the system directories. Liquifile [10] is a file manager for Mac OS that besides showing the directory/file tree, draws circles to convey information about file size, and uses their position to inform about the creation date.

Looking at the majority of the related work, we can see that none of the above approaches succeeds in providing the user with a solution to browse and find documents and files, regardless of their location in the directory tree. Although, there are some solutions using metadata from files and tags, they do not take advantage of a metadata-oriented visualization; do not provide distinctive views for different file types; do not offer organized views for files sharing related metadata; and do not propose innovative and efficient filtering mechanisms.

Our approach will try to combine all of these requirements, to produce an efficient solution for file browsing and searching, based on the semantic information associated to files, and by providing a filtering and overview mechanism based on histograms. To that end, we first performed a user and task analysis to understand how users browse and search files using current applications, and what is the knowledge that users have about their files and their file system.

3 User and Task Analysis

To characterize potential users of our solution and to understand how they perform browsing and searching tasks presently, without our method, we decided to carry out task and user analysis. Additionally, we also collected information related to users' knowledge about files, and in particular concerning their metadata.

To that end, we created a questionnaire to be answered by potential users, with some experience with computers, like browsing files and surfing the web, but with different backgrounds and skills.

We received a total of 86 answers to our questionnaire, being 78 answered without our presence, and eight in our presence. With this in person questionnaires we tried to collect more information about the way users perform tasks, by taking benefit from the advantages of interviews, where we can ask more questions, explore relevant topics, do observation and take note of comments and suggestions.

3.1 Questionnaire analysis

Although, our questionnaire was composed of several questions, here we only present the most relevant, explaining their goals and the main conclusions achieved with them.

These first two questions had the purpose of finding if users use the OS default applications for browsing and searching or if they install any specific application to perform the tasks.

- Q1. Do you use a specific application to browse your files, or the OS default one?
- Q2. Do you use a specific application to find your files, or the OS default one?

In both questions 99% of the users answered that they use the OS default application.

The next question was posed to see how often people lose their files.

- Q3. How often do you find yourself looking for a file you don't know its location?

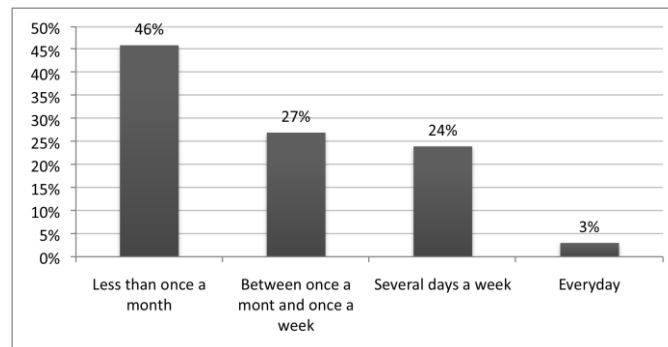


Fig. 1. How often users do not know the location of a file.

Looking at the results from Fig.1 we can see that we have around 50% of optimistic users who almost always know where their files are. However, if we do some simple calculations to combine the different answers, we can say that on average people lose

files 3.7 times a month. In our opinion, this value is relevant enough for us to try a new approach on file browsing and searching, to help users finding their files.

Questions Q4 and Q5 were created to identify which are the most relevant metadata, i.e. what are the characteristics of files that users know, use, and remember.

- Q4. When you are browsing a folder, which characteristics do you use to sort the files?
- Q5. When you can't find a file's location, which file characteristics do you remember?

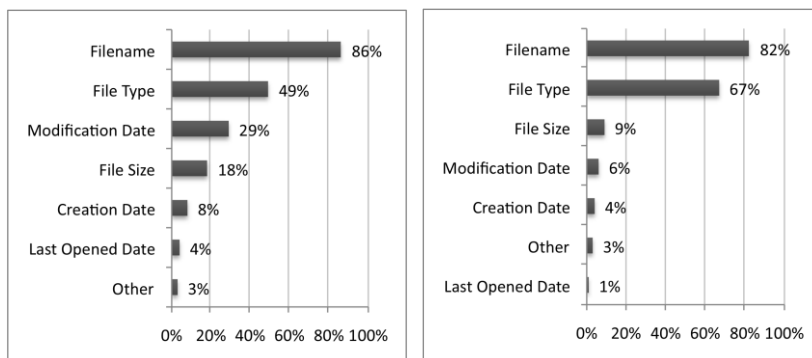


Fig. 2. Most relevant characteristics about a file remembered by users. On the left we have answers to question Q4, while on the right we have answers to question Q5.

As we can see, the filename is the most important characteristic of a file that users use and remember most. We believe that this is due mainly to the use of the default file manager based on the directory hierarchy that “forces” users to memorize file names. Typically, it would be practically impossible, for a common user, to find their files if they do not know their names.

When it comes to metadata we can see that the file type is the characteristic people need/remember the most, followed by modification date and file size.

The next question was made to find out if users know the most common file types.

- Q5. Which file types do you know?

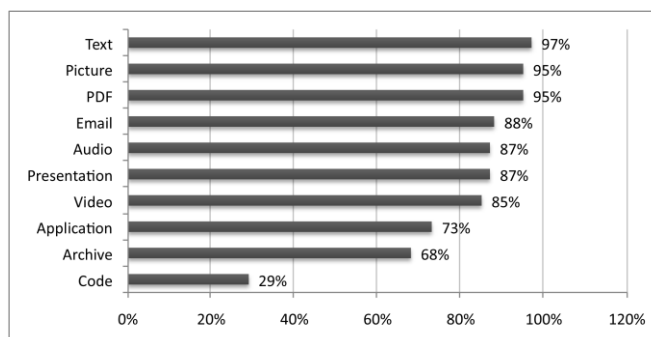


Fig. 3. File types known by users.

From Fig. 3 we can observe that almost all users know Text, Picture, PDF, Email, Audio, Presentation and Video files. However, not everyone knows Archive files and most people don't know what code files are. It was interesting to notice that although some people do not know what application files are, 100% of the Mac Users knew this file type.

During the in person questionnaires, we asked users to give examples of file types they knew. Most of them gave examples of applications and file extensions that were correctly related to the given file types.

Since pictures are one of the most used media file (74% of the users stated that they use their computer to organize and see Photos), we decided to include this question in the questionnaire to identify what are the most relevant metadata from photos, and what is its order of importance.

Q6. Imagine you want to find a photo that you lost, what do you remember about it?

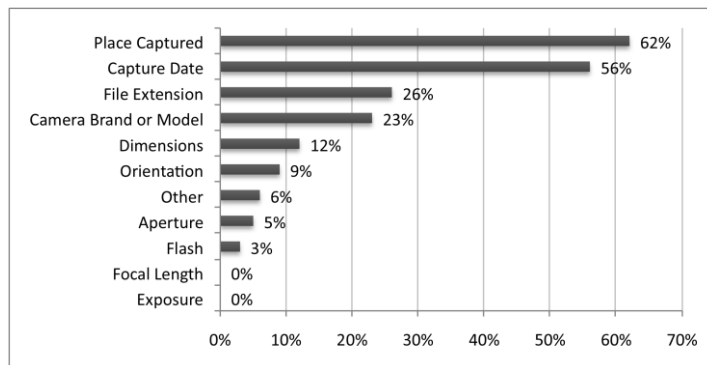


Fig. 4. Most relevant information that users remember about photos.

Since music has already a well-defined selection of relevant metadata and an appropriate hierarchy used in almost all main mp3 applications and devices (e.g. Artist, Album, Genre, etc.) we did not include any question for music.

3.2 Main Outcomes

From the above data collected during task analysis, we can conclude that our users have some experience using computers and know the name of almost all common file types. Additionally, we observed that users consider the File Type as the most important metadata, followed by Modification Date and File Size, when browsing and searching for files. Finally, we verify that users found themselves looking for a file that they do not know where it is, three to four times a month.

4 Proposed Solution

We propose a new paradigm for interactive file browsing and searching, independently of their location, through the replacement of the directory structure by a semantic hierarchy, created using the metadata associated to files. We take advantage of the knowledge that users have about File Types, making it the main filter of our browsing and searching mechanism. Our approach explores also the relationship between files of the same type to make exploration and finding easier and faster. Additionally, and since our solution performs a first filter by file type, we can provide different views for different file types, making the presentation of files richer and consequently easier for users to recall, instead of remember, what they are looking for. Finally, we combine this with a histogram view to give an overview of the existing files and to allow users to perform fast filtering actions.

4.1 Semantic Hierarchy

The semantic hierarchy is based on the fact that different file types have different metadata, so for each file type we can have different properties that can be combined. The main purpose of this hierarchy, unlike the other, is to filter files instead of organizing them. Our semantic hierarchy has a fixed number of three levels, making the browsing and searching action faster by reducing the number of potential clicks that users have to perform for locating a file.

At the top of our hierarchy we have three main nodes, identified during task analysis: file type, modification date and file size (see Fig. 5-left). Although currently we only have three nodes, in the future we can add more, if needed. The next level of the file type node has the different file types has sons (see Fig. 5-center). When the user selects a type of file, the system performs a filtering, showing only the files of that type. Additionally, it presents the next level of the hierarchy, composed by the specific metadata associated to this file type (see Fig. 5-right). Modification date will have sons, such as, “Today”, “Yesterday”, “Past Week”, etc., and file size will have several intervals for size, for instance “1 MB – 5 MB”. Then, users can use these metadata nodes to filter the list of files.

Our solution chooses the more appropriated visualization to the selected file type. For instance, if we select pictures it will show thumbnails, however, if we select music it will show the name, artist, album, genre, etc. of the music.

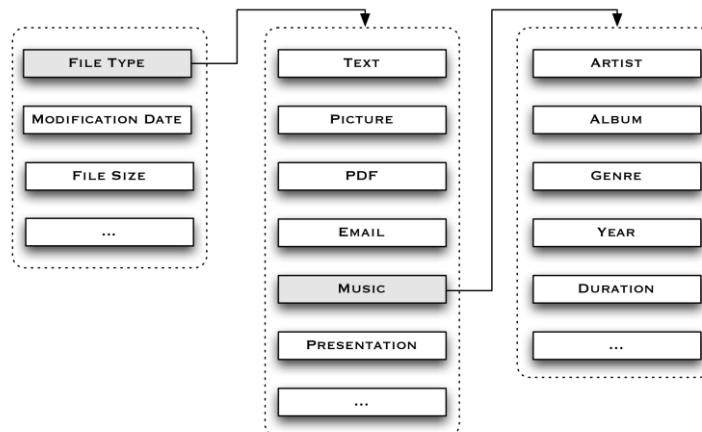


Fig. 5. Semantic Hierarchy, showing three levels for File Type and Music.

4.2 Browsing vs Searching

One of the main drawbacks of existing applications for searching, such as, Spotlight, Windows Search, Google Desktop, etc., is that they only focus on searching, discarding the presentation of results. Typically, results are handle by the file manager of the operating system, or are presented as a list of file names. This list is suitable when we have few results, but it makes difficult to locate a file if the list is very long. Our approach overcomes this by providing two mechanisms. One for searching that uses keywords (using Spotlight functionalities) and/or metadata associated to files, and another for browsing and exploring the content of the file system using the semantic hierarchy, described above.

In summary, users can at any time explore their files by navigating through the semantic hierarchy, or searching for a specific file using filters, while visualizing the particular metadata associated to each file.

4.3 Histogram

Our approach includes also a new method with a dual functionality, the histogram. First, it provides an overview of the existing files, according to a continuous metadata; second, it offers an interactive mechanism for filtering based on selection, more flexible and powerful than the ordinary text-based mode. The histogram represents continuous metadata, such as, modification date, file size, number of pages, music duration, etc., in the form of a chart, where the x axis is the metadata value and the y axis represents the number of files (see Fig. 7). With this representation we can see how files are distributed along the metadata, getting an overview of our files. Additionally, the histogram allows the specification of an interval of values that is used to filter the results. For example in Fig. 7 we can see the grey area that represents a selection, meaning that the user only wants to see documents with 13-17 pages.

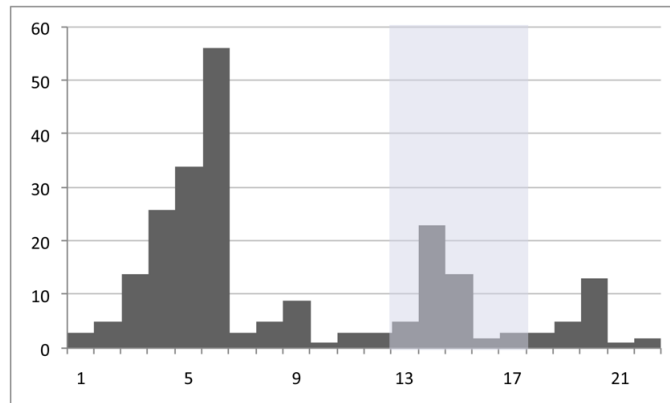


Fig. 7. Histogram for PDF Files, where the horizontal axis represents the number of pages. The grey area corresponds to the filtering of files containing between 13 and 17 pages

4.4 Interaction Scenarios

To better explain how our approach works and to stress its advantages over other solutions, we will now present three simple tasks of browsing and exploration, in four different applications. The first is a typical file manager (Mac OS Finder/Windows Explorer); the second is a searching application (Spotlight/Windows Search); the third is the Leap [5] application that also uses a semantic hierarchy; and finally our solution.

Task	Check which songs are from the 80's	Find a presentation from which I do not remember the file name neither the location on disk	See all portrait photos from 2007 and 2008
File Manager	Not possible.	Not possible.	Not possible.
Searching Application	Create a complex filter by using several menus, combo-boxes and text-inputs, to select just audio files and to define the limits for the 80's.	Use some keywords that possibly are in the presentation, create a filter to select just presentations, using a set of menus, combo-boxes and text-inputs.	Create a complex filter by using several menus, combo-boxes and text-inputs, to select image files, the years and to define the type of orientation.
Leap	Not available. Leap does not consider audio files as documents.	Select File extension (not file type), use some keywords or tags associated to the file.	Not possible.
Our Solution	Select the file type as Music, then go to the histogram and select the desired time interval.	Select Presentation file type, filter by number of slides, title or author.	Select Image as file type, then go to the histogram and select the years, and then select the orientation.

As we can see from the three scenarios above, our approach offers a more efficient and easy way for browsing and searching files, due to the use of contextual metadata about the type of file. We take advantage of the selected file type and present specific information associated to it. While we do this, the other solutions ignore this relationship making the browsing and searching of files more complex and slower.

5 Prototype

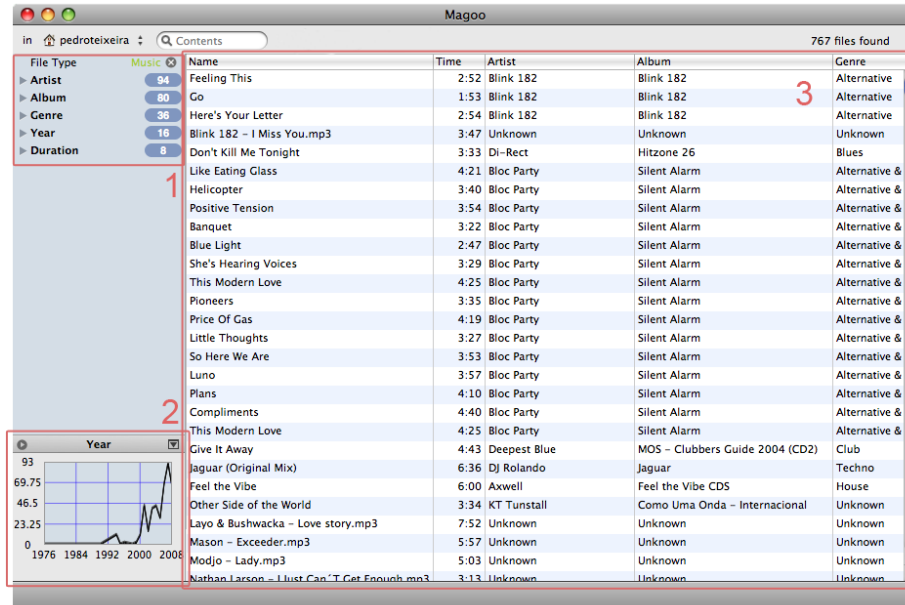


Fig. 8. Prototype for browsing and searching files using semantic hierarchy. Area 1 presents the semantic hierarchy, 2 shows the histogram and 3 the list of files that satisfy the criteria selected.

In Fig. 8 we show the user interface of our prototype, highlighting the three main areas. Area 1 presents the semantic hierarchy, where users can first select by File Type, Modification Date or File Size. After selecting one of this options the system shows the next level of the hierarchy. In the current example, the user selected “Music” as File Type, and the system showed the main metadata associated to Music (Artist, Album, Genre, Year, Duration) indicating the number of occurrences for each characteristic. Area 2 illustrates the histogram, showing the number of music per year. With this type of visualization, users can get an idea of the distribution of music by years and they can also select one temporal region (as illustrated in Fig. 9) to filter results.

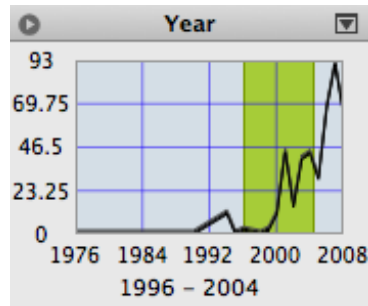


Fig. 9. Histogram representation for the Year of the music. The green area represents the selected time interval (1996-2004).

The visualization and filtering using the histogram can be applied to any continuous metadata. In the case of music we can have histograms for year and duration. Finally, area 3 lists all the files that satisfy the different selections and filters applied by the user. The way files are represented depends of their type. In this example files are music, so the system lists all metadata that are pertinent about music, namely, name, duration, artist, album and genre. If we are listing images, it will display thumbnails; if it were PDF files it will present the name and the number of pages and so one for the different file types. Thus, for each type of file we use the specific knowledge associated to each type of file to select the best visualization method.

6 Testing

We evaluated our solution with user testing and a questionnaire. For the testing we measured the time users took to perform a series of tasks in our solution, in Leap and in the OS default search application (Finder). The questionnaire was created to find out user's opinion about the three applications and to validate the most important choices we made in our solution. The Test results validated our solution, not only because of the time values obtained (see Fig. 10), that were better than the two other applications, but also because of the questionnaire answers, that sustained our decisions and the satisfaction of the users with our solution.

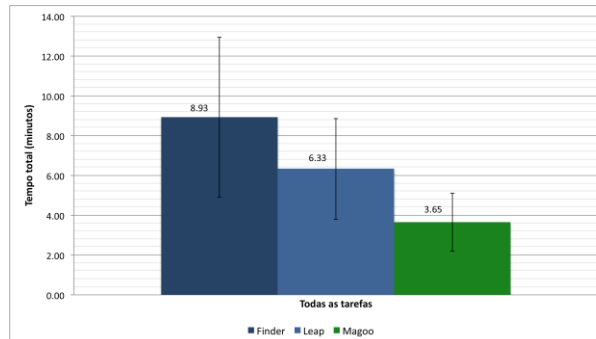


Fig. 10: Time (in minutes) it took to perform all the presented tasks. The Green column is our solution

7 Conclusions and Future Work

In this paper we presented a new approach for interactive browsing and searching of files based on a semantic hierarchy created from the metadata extracted from them. Our solution hides the traditional directory structure, allowing users to forget about the location of files and concentrating only on pertinent information about their files, which are easier to remember and to understand.

Our solution takes advantage of the specific metadata associate to each type of file for presenting the next levels of the semantic hierarchy and to decide which visualization mechanism is used to show the list of files resulting from the filtering. This is a major advantage of our approach, since existing solutions (e.g. Leap) do not explore this relationship between metadata, leading users to situations where they easily get lost and out of context.

To achieve this solution we first performed user and task analysis to understand how users perform tasks related to browsing and searching for files, and what is the knowledge that users have about their files and their file system. We noticed that users are unable to find files three to four times a month and that they consider the File Type as the most significant metadata for exploring and finding files.

References

1. Bederson, B. B. PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps. In: 14th Annual ACM Symposium on User interface Software and Technology (UIST'01), pp. 71--80, ACM, (2001).
2. Shneiderman, B., Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)*, vol. 11, 1, pp. 92—99, (1992)
3. Smith, G., Czerwinski, M., Meyers, B., Robbins, D., Robertson, G., and Tan, D. S., FacetMap: A Scalable Search and Browse Visualization. *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, 5 (Sep. 2006)

4. Cutrell, E., Robbins, D., Dumais, S., and Sarin, R., Fast, flexible filtering with phlat. In: SIGCHI Conference on Human Factors in Computing Systems (CHI '06), pp. 261—270, ACM, (2006).
5. Ironic Software, Inc. Leap, <http://www.ironicsoftware.com/leap/index.html>
6. Apple – Mac OS X – Leopard Sneak Peek – Spotlight, <http://www.apple.com/macosx/leopard/spotlight.html>
7. Baudel, T., Browsing through an information visualization design space. In: Extended Abstracts on Human Factors in Computing Systems (CHI'04), pp. 765--766, ACM, (2004).
8. Bladh, T., Carr, D. A., and Scholl, J., Extending tree-maps to three dimensions: A comparative study. LNCS vol. 3101, pp. 50--59, (Mar 2005).
9. Robertson, G. G., Mackinlay, J. D., and Card, S. K., Cone Trees: animated 3D visualizations of hierarchical information. In: SIGCHI Conference on Human Factors in Computing Systems (CHI'91), pp. 189--194, ACM, (1991).
10. Waldeck, C., Liquid 2D Scatter Space for File System Browsing. In: Ninth international Conference on information Visualization, pp. 451—456, IEEE Computer Society, (2005).