iHDViewer: A Visualization Tool for Tracking HD

Wenbo Wang[†], Xi Luo[†], Liangfu Lu[‡],Youyi Zheng[§†*]
[†]ShanghaiTech University
[‡]Tianjin University
[§]Zhejiang University
{wangwb, luoxi}@shanghaitech.edu.cn, {zyy}@cad.zju.edu.cn, {liangfulv}@tju.edu.cn

***Abstract*—Visualization techniques on hierarchical information have been proved their capability in presenting structured disk storage data. However, most existing techniques represented with simple visual methods such as pie chart or histogram , lacking the ability of grasping the usage of the computer. This paper presents a new interactive visualization technique named iHard Disk Viewer (iHDViewer) , which aims to display file details and their movement histories, as well as predict the user's activities through visualization results. This tool mainly utilizes four different view panels to display the distribution of files and file folders for a computer system. Specifically, A tree view explains the general distribution of the whole disk storage; A river trend view tracks the files moving history, which is implemented by cubic spline interpolator with restricted boundary conditions to ensure the trend stability and information accuracy; We also use the concept of coordinate extension to visualize the file moving events which supports more intuitive results; A filelist view panel shows the details of the selected files; while a circle view panel interactively displays the status of files and file folders. In addition, considering the importance of human visual perception system in the data analyzing processes, we utilize three color design principles to work through all visual methods. Lastly, this paper demonstrates the effectiveness and efficiency of our new idea to visually analyze computer FileSystems.**

***Index Terms*—Visualization, Cubic Spline Interpolation, Flow Trend Visualization, Interactive Visualization, Visual Perception**

I. INTRODUCTION

Nowadays, computer has become a necessity in our daily lives. Its importance has been publicly recognized either for working, studying, or entertaining. Especially in recent years, with the rapid development of IoT (Internet of Things), computers are designed to be capable of helping people to control home appliances and do surgeries. Although the improvements of computers utility bring various benefits, the drawbacks could not be ignored. As most of our personal information has been stored online since we were born, crimes related with digital data have become one of the major security problems in our society. For example, one type of criminals use illegal digital tools to get target's private information and then login their Internet banking to withdraw money; some criminals tend to store pornographic photos or videos on their computers, and then spread them online, which bring higher risks to other people's physical and psychological health growth. Therefore, it is important to have a consciousness of data security. However, if the crimes are sadly happened, it becomes critical for investigators to discover accurate evidences and catch the criminals.

In this 21st century, computer is one of the most frequently used crime-tools in forensic analysis. Until 2017, desktop drives had come in storage capacities of as much as 5TB as the market of computer storage capacities are required. While such large storage devices are involved in forensic investigations [**?**],from the data analyst' point of view, the higher the requirements of the computers are, the larger volume of data generated, and the harder for investigators to observe and discover crime clues, including: identifying suspicious files; tracking files history of movements; reaching files accurate destinations in a short time, etc..

Although the problems become harder than the past decades due to the gradually increaed data volume, forensic investigators did not stop discovering new ways to improve their investigation abilities. Data visualization is one of the techniques that have been utilized by the investigators into forensic investigations. Data visualization has been proposed to make creative and interesting graphics in working, and discovering complex data relationships. Most of today's visual analysis methods for computer filesystem are focusing on the representation of file structures and file details. Although some can visualize the relationships among files, they are classified by file attributes, for example: file extension, file size, date time etc.. And none of them considered the moving states of folders/files in a computer system, such as when the files are created and how they are operated in a certain time period, which are all critical information for information analysis.

Motivated by user requirements, we propose a new visualization tool named iHardDisk Viewer (iHD Viewer). This tool is not only capable of displaying the hierarchical structure of the large filesystem, it is also introduced a time based historical trend view to track and compare the activities of each file/file folder, which is the first time to use flow visualization in the analysis of large storage file system. A concept of file Split is also proposed to enable an intuitive file viewer for observing files' moving events, and a nested circle visualization with zoom in and zoom out interaction mechanism to largely improve the usability of this tool.

The paper is organized as follows. Section 2 discusses

*Corresponding author.

2375-9356/18/$31.00 ©2018 IEEE
DOI 10.1109/BigComp.2018.00057

the related works in details. Section 3 and 4 introduce the designing goals and implementation details; in section 5, we give a case study; while section 6 provides evaluation results. Section 7 gives a short discussion, and the final section concludes our work.

II. RELATED WORK

During the past decades, many visualization tools had already been implemented according to the requirements of displaying computer filesystem. The representative tools are: Free Disk Analyzer, Disk Investigator, JDiskReporter, TreeSize and Disktective. Specifically, Free Disk Analyzer and Disk Investigator display file details in text format; JDiskReporter, TreeSize and Disktective present classification results through pie chart; while WinDirStat, SpaceSniffer, Disk Space Fan display data with histograms. Some of these tools are popularly used in forensic research area. However, because these tools can only display the static information of each file, so if we consider the advantages and developments of visualization techniques in recent years, especially the developments of visualization techniques for hierarchical strutured data, such as the following three visualiztion techniques are the extentions of simple tree structure views: Tree Map [**?**], Cone Tree [**?**] and DOI tree [**?**], the representation methods of forensic tools can be improved.

On the other hand, today's visual efficiencies tend to decrease when the user requirements grow in seeking relationships, and visualization techniques also bring good results in displaying data relationships. W.Wang, Mao.Huang and Liangfu.Lv [**?**] had used data classification methods on six data attributes to display the file relationships through parallel coordinates, which provides a clear results. But their method was limited to the relationships among single data files. Therefore, current visualization approaches for filesystem analysts are required to visualize not only the general structure of the contents and details of files, but to know the activities over each file.

A widely accepted information tracking method is flow visualization, which uses river stream metaphor to explain data. The usefulness of this idea has been certified in many research branches. The first branch is about large document collections, which users could compare theme changes in one set of documents to those in another set, and at the same time, the visual results conduct a variety of long-term stories to motivate the text generation. The representative works are ThemeRiver [**?**], Event River [**?**] and CiteRiver [**?**]. Specifically, in document collections; The second branch is the evolution of social media topics, its main works are listed as: OpinionFlow [**?**], EvoRiver [**?**], RoseRiver [**?**], TextFlow [**?**]; The third branch is text semantic analysis, and ThemeDelta [**?**] is a representative work in recent years. In detail, Text semantic analysis and topics evolution in social media are two similar research areas, there is no strict division between these two areas, we classify them according to the contents to be analyzed. For the former is more focused on the contents of the text, such as the words, sentences, or paragraphs, and then relate them through the ranking of the ordered words in each text; while the latter pays more attention to the attributes of the text, such as a label, which is often used to explore the profound insights of evolving topics.Some other research areas are also benefited from the idea of river pattern. Such as LoyalTracker [**?**], a tool designed for search engines that using the river stream changes to discover the customer loyalty and switching behavior from massive data sets in business intelligence.

Unlike the above works, this paper addresses the problem of better understanding and analyzing the behavior of computer users/owners, especially to improve the working efficiencies of investigators in their investigation tasks to find useful clues.

III. SYSTEM DESIGN

A well formalized visualization tool shall maximize the capabilities to perceive and understand complex and dynamic data. Such as information overview, pattern discovery and searching. In this section, we discuss the design goals, main architecture and visual methods.

A. Design Goals

Nowadays, digital data is overwhelmingly generated. Hierarchical structured information are getting more complex to be analyzed as thousands or even millions of elements and relationships might be covered. Therefore, two basic design goals are required to design a proper file system analyzation tool: Firstly, the users shall be able to find objects quickly and easily with the powerful assistance of human perceptual system; Secondly, display information simply and meaningfully, where "simple" means that it is easy to be comprehended by a beginner in the area of visual analysis, and "meaningful" represents the tool can provide results' accuracy and rich contents.

B. System Overview

iHardDisk Viewer (iHD Viewer) Fig.11 is started with a display of the filesystems hierarchical structure in text format with shaded colour and a layer control. The darker the colour, the higher the layer of this folder; the lighter the colour, the lower the layer of this folder. This iHD Viewer is also able to track the changing events of each folder and differentiate their states at various time period through three view panels. The first view panel is called packed circle view panel, which is using circles to represent folders, and the inside files are displayed by their names. This view panel visually summarizes the distribution of all files in a disk drive storage; The second panel is called river trend view panel, which is using the shape of river streams to display the moving frequencies of all files,either moving in or moving out, and the width of each stream represents the files changing frenquency, whereas the color of each stream represents different file folders; The third panel named ring view panel, which aims for comparing the changing frequencies among file folders, it is a supplementary result for analysts to connect it with river trend result. In addition, iHDViewer supplies a "contents + detail" display by

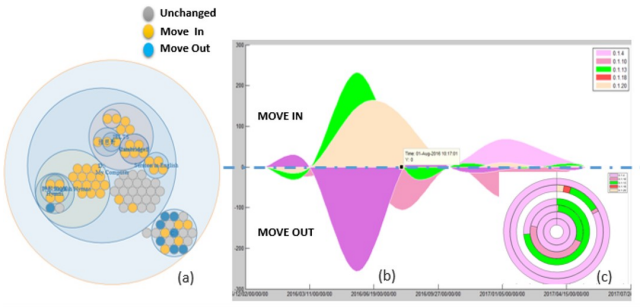

Fig. 1. iHDViewer, a visualization tool for visually monitoring computer files. (a) Overview a hard disk storage. The large circle represents different folders in a computer, while each small bubble insided the large circle represents different files. Colors represent the different changing status of files: Yellow represents the file is created, Blue represents the file is modified, and Gray represents the file has not changed during that time period. (b) Compare the movements over all files. The color strips represent the moving activities among folders; The trends in the positive direction of X-axis represent the files are created, while the trends in negative direction of X-axis describe the files are modified. (c) Summarization of the file movement activities. Each ring represents a selected time period, and each color represents the selected folder.

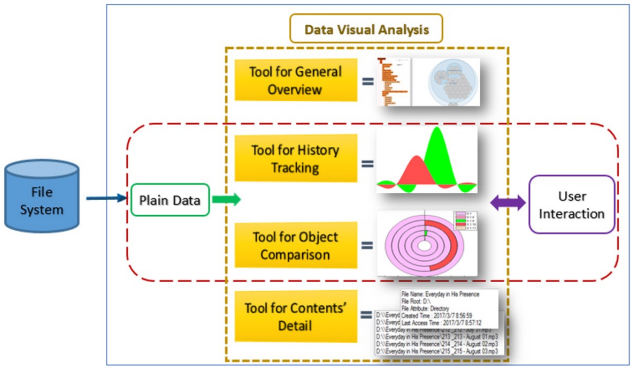

Fig. 2. The system interface

a list view method , the users can directly see the target files' attributes: including file path file size, file name, file extension, etc.. Generally, these four view panels are well-coordinated to help users understand the structure of the system, and also discover desired contents. Fig.2 is a system interface of iHD Viewer.

IV. IMPLEMENTATION DETAILS

We have implemented a visualization tool to track system files activities. This section describes the implementation details.

A. River Trend Calculation

The idea of discerning information patterns (relationships or trends) through a river metaphor was proposed by Susan Harvre, Elizabeth Hetzler,Paul Whitnev and Lucy Nowell [**?**]. They implemented a proof-of-principle prototype of ThemeRiver. Each stream represents an individual theme; the strength of a theme is the collective strength of the selected themes and is calculated by the number of documents that contain the theme word. Based on this idea of discovering relationships among complicated data, some researchers extended the method to convey more information through analyzing the river diversion and river stream pattern. For example, M. Ghoniem [**?**] constructed a keyword hierarchy and bundle the currents according to the hierarchy; while W.W.Cui and S.Liu [**?**] proposed that every river stream was composed by several sub trends, and they used data mining and visualization integration techniques to explore relationships from the splitting and merging of the river sub-trends. The advantages of using river metaphor to observe information are clear from these previous works.

Therefore, in our paper, we apply the idea of river metaphor to represent the large data volume in hard disk drives. Specifically, considering the requirements of data accuracy and computational complexity for filesystem analysts, we choose cubic spline interpolation [**?**] to present the moving events of each file in a file system. The main reason is that Cubic Spline Interpolation is one of the most popular interpolation algorithms, it uses the method of segmentation to avoid higher order polynomials, and at the same time the stability can also be remained. In addition, the results of each segment is only depend on the next curves, which is also suitable for file system analysis.

In this section, we first describe the data, and then introduce how we apply cubic spline interpolation to display files moving history.

1) Data Details: All data are collected in pairs, where the first parameter is referred to time and the second parameter is referred to the associated magnitude. Each trend stream represents a file folder, which is available for users to choose. And the streams are differentiated by various colors. The width of each river trend represents the strength of each folder, which is calculated by the number of occurrences of the files created/modified in each time interval. Specifically, we denote data as: $(t_i, r_{i,j})$, where $i = 1, ..., n$, $j = 1, ..., m$. t_i indicates successive time (can either be monthly/daily, or /hours/minutes/seconds); $r_{i,j} >= 0$, which is a measure of to what extent the files are affected the j_{th} folder directories at the $t_i th$ time. The more numbers of files are created or modified in a folder at time t_i, the higher affections of this folder are. The river flow gives a smoothly varying presentation of the data where their values are stacked.

2) Algorithm: Interpolation approach theoretically provides better visual effects for the changes of discrete data. So it is a desirable method to describe our filesystems moving historical events. In our paper, we use cubic spline interpolation method rather than line interpolation or nearest neighbor interpolation because the curves drawed by this method are more smooth.

To create the best fit and pleasing river flows for a series of data points, we denote the flow by r, which is used as an interpolator. In this subsection, we firstly explain the method-constraints, which is using for getting better visual flow trends, then descript the methods in details.

- Mathematical Constraints

We are looking for an $S(t_i)$ which is the most close to the value of original data points at the segment $[t_i, t_{i+1}]$. According to the definition of an interpolator, a desired interpolator requires: (1) $f(t_i) = r_i$, which is to ensure the continuity of the connection between adjacent curves; (2) $f(t_i) >= 0$, which is to remain each river trend being positive to provide a better display.

Particularly, to get the desired r_i , we need to set some constraints for the interpolators [**?**], in our paper, the constraints we set are listed below :

Constraint 1: This can be defined as $S(t_0) = S(t_n)$, which is using for curve connection among all segments.

Constraint 2: The interpolator also needs to be smooth at the curve joints, defined as $S'(t_0) = S'(t_n)$.

Constraint 3: The curves requires the concavity and convexity of the joints are consistent, defined as $S"(t_0) = S"(t_n)$.

- Drawing Steps

Here are the three drawing steps to complete our river trends:

Step 1: Suppose $r_i = f(t_i)$ is continuous between $[a, b]$, and there are n pairs of values between $[a, b]$, denoted as $(t_0, r_0), (t_1, r_1), ..., (t_n, r_n)$. Denote r_i as the total number of files created OR modified at time t_i, specifically, $a = t_0$, $b = t_n$.

Step 2: Denote each sub-coordinates by S_i, and r can be displayed by S_1,S_i,...,S_n where $i$represents the time t_i. We need to calculate $S_i(t_i) = r_{.i}$, where $i = 0, 1, ..., n-1$.

Step 3: Compute $S(x)$, $S(x)$ and $S"(x)$ (derivative, and second derivative), and Match first and second derivatives at left end with those at right end. Therefore, the flow curves are all continuous and smooth connected in the area $[a, b]$. Denote as: $S_i(t_{i+1}) = r_{i+1}$,$S_i^{'}(t_{i+1}) = S_{i+1}^{'}(t_{i+1})$, $S_i^{"}(t_{i+1}) = S_{i+1}^{"}(t_{i+1})$, where $i = 0, 1, ..., n-2$.

In drawing the river trends, we use a Coordinates Extension concept to display all the curves, which has not been proposed in the existed works. We choose the $x-axes$ positive side to explain how files are moved into different folders, while the $x-axes$ negative side represents how files are moved out of folders at the same time period. This method can simplify the exploration of data.

B. Visual Cues

Visual cues are popularly utilized in visual representation as it can improve the results qualities. For example, colour, size, and shape, these three cues are all proved to have stronger aesthetic influence on highlighting. What's more, visual cues normally does not work by itself but is often embedded in the other visualization techniques.

In our work, the design of iHDViewer, we choose colour cue to enhance and clarify the presentations. Because a proper colour matching can improve the reading speed and understanding degree, so we choose colors depends on the following three colour design principles [**?**].

Principle 1: If analogous values are used to define layers of attention, it can provide a good legibility when the data is in text format.

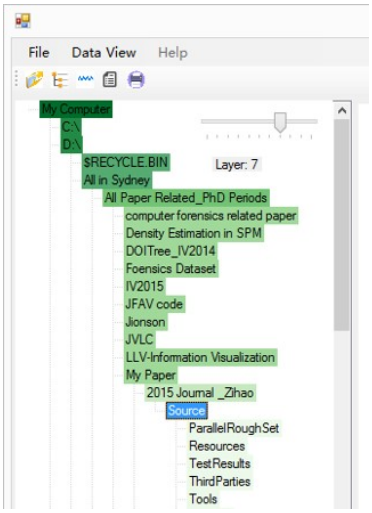

Fig. 3. An example of applying Color Design Principle 1

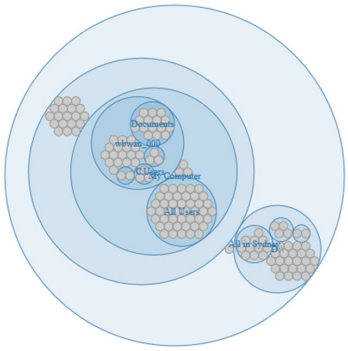

Fig. 4. An example of applying Color Design Principle 2

Principle 2: If the colors are grayer, or more pastel (closer to white), it will be easier to allow the use of saturated colours for highlighting

Principle 3: If the luminance contrast is higher, it will be easier to see the edge between one shape and another. If the contrast is too low, it can be difficult to distinguish between similar shapes, or even discern the shape at all.

We use cases to introduce the applications of the color principles in iHD Viewer. In the overview panel Fig. 3, according to Colour Design Principle 1, the darker the colour it is, the higher the layer of the file will be .

In the circled packed view Fig. 4, we choose light blue as the main colour to make it much clearer in the discovery of the chosen objects according to Colour Design Principle 2.

In the river flow panel and ring Vis panel Fig. 5, colour is mainly used for labelling. As for labelling element, contrast and analogy are two worthy references. Therefore, according to Colour Design Principle 3, to better differentiate the objects, contrasting colours are better choices, while for grouping objects, analogous colours normally take more effective results.

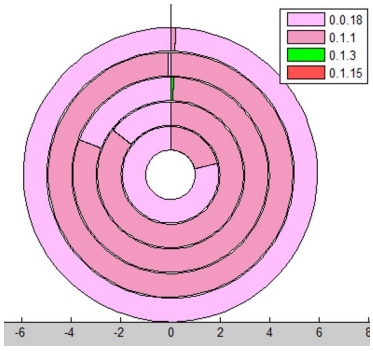

Fig. 5. An example of applying Color Design Principle 3

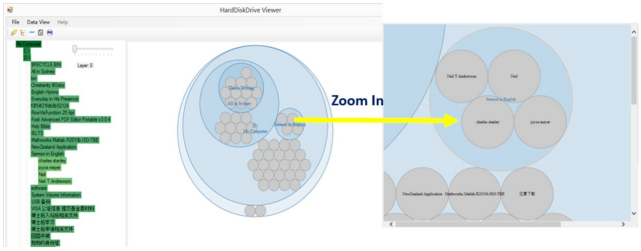

Fig. 6. The Interaction Mechanism of iHD Viewer - Zoom in and Zoom out

C. Interaction Mechanism

One of the most important capabilities of a visualization tool is to maximize human capabilities to perceive complicated data. A navigation mechanism can enable users to interactively adjust views to reach a clearer understanding of a complex graph. Therefore,we apply an interaction mechanism zoom in and zoom out. See Fig. 6.

Another usefulness of interaction mechanism is helping users to choose their preference information. This allows users to maintain the perception of what they would like to discover from the large information spaces. We take Fig.7 as an example: We use a decimal form to display the path of each folder because it is easier to be located technically, every decimal number is determined by the storing path of computer files. For example, In folder $0.1.5$, the first number "0" represents the computer root directory; the second number "1" represents the secondary path of the folder is the first folder of the secondary level of computer storage system; the third number "5" represents the third layer path of the folder is the fifth folder of the third level in this computer storage,Suppose a computer is under-examined and the analysts requires details about four folders, $0.1.5 -> 0.1.11 -> 0.1.19 -> 0.1.22$ in a time period of $[2016-01-01, 10:23:15; 2017-05-12, 10:23:15]$.

After starting iHD Viewer, the users are required to set a number to segment the long time period into small parts from $[2016-01-01, 10:23:15]$ to $[2017-05-12, 10:23:15]$. In our test, we set "5" as the time interval, which means the time period are equally divided into five segments, see below:

- $[2005/12/02/00/00/00, 2016/03/11/00/00/00]$

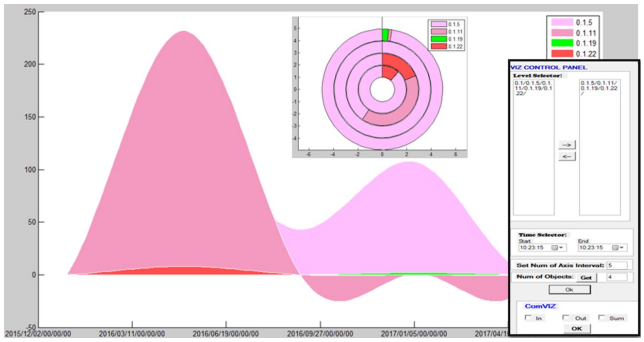

Fig. 7. An example for Interaction Mechanism

- $[2016/03/11/00/00/00, 2016/06/19/00/00/00]$
- $[2016/06/19/00/00/00, 2016/09/27/00/00/00]$
- $[2016/09/27/00/00/00, 2017/01/05/00/00/00]$
- $[2017/01/05/00/00/00, 2017/04/15/00/00/00]$

Then users then can choose their desired visualization methods to display the status of this hard disk storage. The visualization methods contain flow trend view and ring view. We use Fig. 7 as an example to explain these two views.

In Fig. 7, a main river flow view represents the changing frequencies of folders in a specific time period. The wider the river trend, the more files are moved into this folder. In addition, the $x-axes$ positive side explains how many files are moved into the system at the same time period, while the $x-axes$ negative side represents how many files are moved out of the folders.

The ring view is the other graph to directly compare the changing histories of each folder in the same time periods. In this view, we discover that folder named "$0.1.5$" is changed more frequently than the other folders, and according to this discovery, we look back the file system and find that this folder contains many songs. So we roughly predict the user might be a music fans; To compare, folders named "$0.1.11$" and "$0.1.19$" are only changed in the fourth time period, and file "$0.1.22$" is modified more times during the first and second time period, samely, acoording to these two findings, we look back the file system again and observe that this folder contains all the stuff related with study abroad, which means the owner of the computer is working overseas during that time; Further, in the fifth time period $[2017/01/05/00/00/00, 2017/04/15/00/00/00]$, none file was generated or modified, so if the analyst would like to find some clues, they can ignore the fifth time period, and pay more attention on other time periods, which will save searching or investigation time to some extent.

V. CASE STUDY

To put this tool in context, let us consider the examples of analyzing a private personally-owned computer. The aim is to discover and predict the activities of computer user/owner through analyzing computer files moving states.

The test computer is an ASUS laptop, iHD 1TB, Memory 8G, Intel Core i5-5200U, 2.7GHZ. OS: Win 8.1.

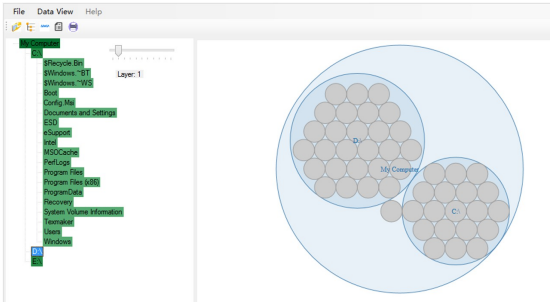

Fig. 8. The hierarchical view of a hard disk drive

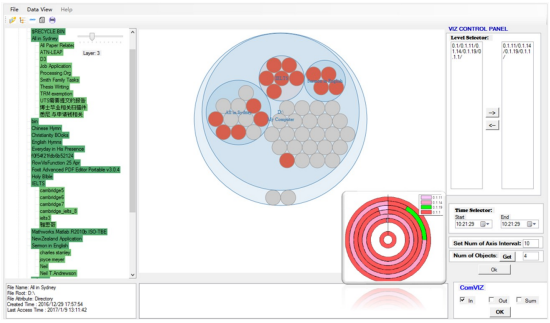

Fig. 9. Data Exploration - A representation of files' moving in activities by circle view and ring view

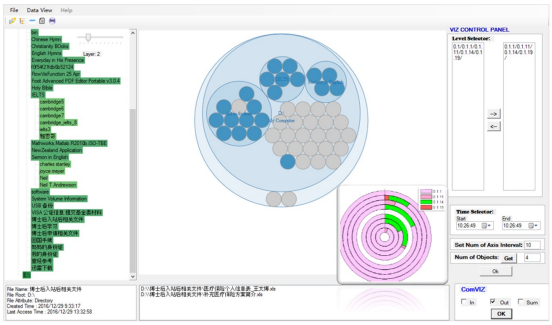

Fig. 10. Data Exploration - A representation of files' moving out activities by circle view and ring view

Firstly, loading computer files, and the data would be read by file parser. And a hierarchical view would be displayed on the left panel of the tool. See Fig.8.

Then the analyzer can set the constraints at the VIZ CONTROL PANEL to further understand the details of the files movements and then predict the user activities.

To make it more clear, here we use two examples to explain how our tool is useful for data analysts and forensic investigators.

- Data Exploration

This example aims to display how the four view functions are worked together to explore data.

The visualization constraints are shown in Fig.9. We choose four folders as the targeted objects, "0.1.11","0.1.14","0.1.19" and "0.1.1". We set a time period from "$2015/12/1, 10:31:37$" to "$2017/5/15, 10:31:37$" and divide this time period into 10 parts. Now we take a look at the visualization results.

In Fig. 9, the orange bubbles represent the files which were created at the same time period. In the same folder layer, we can observe that all the files in folders named IELTS $(0.1.11)$ and Sermons in English $(0.1.14)$ were created, while folders named All in Sydney $(0.1.1)$ and Postdoc related Files $(0.1.19)$ had some files being created in the same period. and when we look at the ring view, the comparisons of changing frequencies among these four folders are more obvious. In the ring view, each ring represents a time interval. In our case, because we had set 10 segments for the time period $[2015/12/1, 2017/5/15]$, so in the second and the third time intervals, none file is created, while most of files in folder Sermos in English are generated in the fifth and the ninth time interval, it is also very easy to see that files generated in folder Postdoc related files are mainly at the seventh time period. From the above information, the analysts can predict that computer owner might stay in Sydney during $[2015/12/110:31:37 - 2017/5/1510:31:37]$, and in the later of this time period, the computer owner might have activitis related with postdocs.

Furthermore, iHD Viewer also supported flow trend analysis, see Fig.11, and Fig.10 is a supplemental analysis result for Fig. 11. These two views cooperatively display how the files are moved out from folders All in Sydney $(0.1.1)$, IELTS $(0.1.11)$,Sermons in English $(0.1.14)$ and Postdoc related Files $(0.1.19)$ during $[2015/12/110:31:37 - 2017/5/1510:31:37]$ under Disk D. For example, in Fig.11, $[28 - Mar - 201610:31:3724 - Jul - 201610:31:37]$ is the most fluctuating part, which means most activities of these folders are happened in this time period, and according to this clue, we can reload data and visualize this time period through ring view, see Fig.10, to find more detailed events happened during this time period; At the same time, from ring view presented in Fig.10, we found that nearly all these folders had been operated, except folder All in Sydney $(0.1.1)$, so we can guess that this folder was less interested by users than the other folders during this time period.

In general, although the fesults displayed through ring view and river view contain similar information, we believe that different views can enlarge investigators analyzing perspectives.

- Events Tracking

The second case is to explain the effectiveness of the events tracking approach of iHDViewer(the corporation of flow trend view and circle view), which is specifically convenient for forensic investigators.

We take Fig.7 as an example. It is easy to observe that ring view is more directly to compare the changing histories of each folder in the same 5 time periods. And the ring view can directly track the changing frequencies of each folder. For example: In Fig.7, a folder named "0.1.5" is changing with a higher frequencies compared with the other

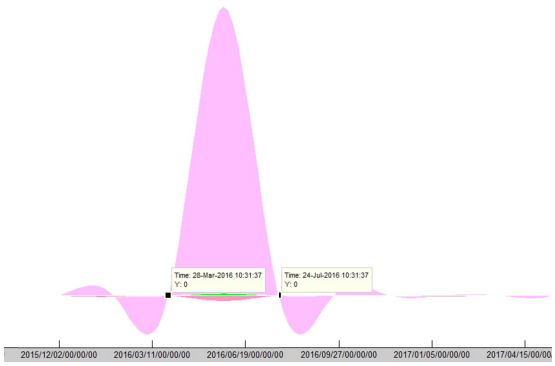

Fig. 11. Data Exploration - A Flow Trend View

three folders within the same time period. After observing the visualization result, users can recheck the file system to find the related clues. In this case, the analysts can discover that this folder contains many songs, which helps to predict the user's activities or hobbies in the corresponding stage. Accroding to the folders' name and contents, we guess that the user might have interests in music and the music style is relatively active; In addition, from the ring view, we can also discovered that folders named "$0.1.11$" and "$0.1.19$" are all only changed in the fourth time period; and files in folder "$0.1.22$" are modified more times than the other folders, so based on these new findings from ring view panel, we check the file system again and discover that this folder contains all the stuff related with study abroad, which could help the analysts to predict that the computer owner might be working overseas during that time; The last obvious finding in this ring view is in the fifth time period $[2017/01/05/00/00/00, 2017/04/15/00/00/00]$, none file was generated or modified, therefore, according to the phenomenon, computer investigator will focus more on the activities or events happened during the other time segments, which would save their investigation time and improve the working efficiency.

Similar to the second case, we can apply this to speculate more events of the first case. The following are the analyzing results based on the first case:

- In the late 2015 and the beginning of 2016, the computer owner was focusing more on "Enlgish learning" (Fig.9).
- From Dec, 2015 to Mar, 2017, the owner was probably staying in Sydney, or might communicate more with people who are in Sydney (Fig.10-11).
- In the late 2016, the owner was more likely to start a postdoc job (Fig.11).

We had also verified our hypothesises with the computer owner, and found that our conclusions are almost close to the facts. For this reason, the application provides to the analyzer a useful tool which is responsible for showing the files moving history and roughly explaining the owners' activities.

VI. RESULTS AND EVALUATION

In general, iHDViewer acted without significant problems at runtime. It has been realized that the use of Breadth-First-Search (BFS) would cause the running time relatively slow, but this is now put on our enhancement list of things to do in the future, and we plan to use parallel computing to improve it. Some important extensions that might be considered for further optimizing the functionality of this tool as follows.

- Support for platforms
  It is necessary for the visualization application to support all the available platforms because it increases its portability and extensibility to all the file systems. Such as, suppose a visualization program that can display files from any system, without worries about data formats. But before such optimizations can be utilized to the tool, there are two important prerequisites: A sound understanding on the files systems, and how are digital data stored in hard disk drives.
- Technical optimization on visualizing anonymous file
  This aim is especially designed for forensic investigators. In computer forensic investigation, the deleting operation is always a suspicious event, and the files which are deleted are always critical clues for the investigators. Therefore, a major issue for analysts is to find these files quickly and accurately, including their attributes: file extension, file size, date time, etc. In our tool, we will highlight these files in the circle view panel and point them out in the flow trend view to help the user intuitively find their storing paths. The details of the suspicious files will be also shown on the list view panel. A possible solution is to get the detail of deleted files by other investigation tools first, such as Disk Investigator, and then we save the data in a readable format for iHDViewer; lastly, use our tool to further analyze it visually.
- Improvement in the flow trend view
  We use cubic spline interpolation to display the files moving trend. Although it has a good stability, it is difficult to keep a balance between information visibility and system running consumption. The higher of the number of segments, the more of running time it costs. Another problem is about setting a proper Constraint to ensure the information clarity. As shown in Fig. 12, with the number of segments increase, the more data changes can be explored (from (a) to (c)), however, when the number sets to a boundary, the data lost its value, see (d), where the user nearly could not see any information. Lastly, we also hope more information can be observed on the flow trend.
- Optimization on circle view
  In our current circle view, when the name of the documents are long and the number of the documents are big, the bubbles would be blurred and it is hard for users to find the target files. Therefore, we plan to display the top 20 files in a circle or sub-circles and find a method to show a suitable length of the file name without losing

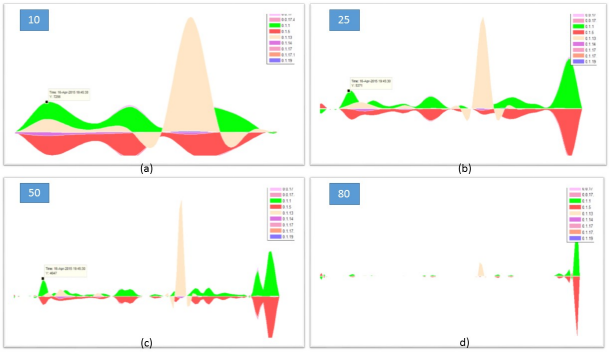

Fig. 12. A comparison between different number of segments in Flow Trend View. (a) The number of segments is 10; (b) The number of segments is 25;(c) The number of segments is 50; (d)The number of segments is 80.

its keywords. In addition, to help the analyzer better understands the movements of files, we plan to use a new time slide bar for users to control the variations of time, and make the circle view more user-friendly.

VII. SUMMARY DISCUSSION

During the initial development, the program looked like a common visualization tool to visualize hard disk drive files. Subsequently, it became apparent that building an integrated visualization world to track user activities through displaying the files movements. This made the program heavier in terms of idea novelty, development effort, information accuracy and diversification. In addition, it is helpful for data analyzer to do further information exploration, including inference and prediction.

Although such an integrated views are not important for an investigator to work on real forensic cases, the idea of revealing file trends' to track computer owner's activities is unique compared to other open source forensic tools. It is useful in quickly understanding the file systems using statements and also generally grasping the computer owners activities in a user-friendly and visual manner. This will help the investigators and data analysts to go through the large volume data in a computer related forensic analysis in an easy-to-use way.

VIII. CONCLUSION

Current file system visualizing techniques provide capability to present the original hierarchical data structure and some with fundamental methods of file classification on file attributes, file size, file extension, or file date time, etc.

However, none filesystem visualization tool has been proposed to track file movements and then predict user activities. Furthermore, most of the existed techniques focus on the analysis of files, lack the ability of analyzing relationships among folders, which is also very important for analyzers or investigators to save investigation time.

iHDViewer is a new visualization tool for analyzing computer files, which can be divided into three parts from a visualization point of view, including: structure overview, tracking file movements view and a content + detail view. Many strength and weaknesses identified in developing of the tool have been highlighted throughout the paper. The currently implemented visualization forms presented in this paper are only the beginning of more extended features to come in future versions of the tool.

ACKNOWLEDGMENT

We thank all reviewers for their valuable comments. This work was supported in part by the National Natural Science Foundation of China NO. 61502306, the China Young 1000 Talents Program.

REFERENCES

[1] A. Beveridge. *Forensic investigation of explosions*. CRC press, 2011.
[2] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, 17(12):2412–2421, 2011.
[3] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *IEEE transactions on visualization and computer graphics*, 20(12):2281–2290, 2014.
[4] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan. Themedelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics*, 21(5):672–685, 2015.
[5] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky. Newslab: Exploratory broadcast news video analysis. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 123–130. IEEE, 2007.
[6] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE transactions on visualization and computer graphics*, 8(1):9–20, 2002.
[7] F. Heimerl, Q. Han, S. Koch, and T. Ertl. Citerivers: Visual analytics of citation patterns. *IEEE transactions on visualization and computer graphics*, 22(1):190–199, 2016.
[8] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE transactions on visualization and computer graphics*, 18(1):93–105, 2012.
[9] S. McKinley and M. Levine. Cubic spline interpolation. *College of the Redwoods*, 45(1):1049–1060, 1998.
[10] Q. V. Nguyen, S. Simoff, and M. L. Huang. Using visual cues on doitree for visualizing large hierarchical data. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 1–6. IEEE, 2014.
[11] G. G. Robertson, J. D. Mackinlay, and S. K. Card. Cone trees: animated 3d visualizations of hierarchical information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 189–194. ACM, 1991.
[12] C. Shi, Y. Wu, S. Liu, H. Zhou, and H. Qu. Loyaltracker: Visualizing loyalty dynamics in search engines. *IEEE transactions on visualization and computer graphics*, 20(12):1733–1742, 2014.
[13] B. Shneiderman and M. Wattenberg. Ordered treemap layouts. In *Proceedings of the IEEE Symposium on Information Visualization 2001*, volume 73078, 2001.
[14] M. Stone. Choosing colors for data visualization. *Business Intelligence Network*, 2, 2006.
[15] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. Zhu, and R. Liang. Evoriver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1753–1762, 2014.
[16] W. Wang and M. L. Huang. Parallel coordinates visualization of large data investigation on hdds. In *Computer Graphics, Imaging and Visualization (CGIV), 2013 10th International Conference*, pages 93–99. IEEE, 2013.
[17] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1763–1772, 2014.