Investigating Aspects of Visual Clustering in the Organization of Personal Digital Document Collections

by

Hoda Badesh

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
March 2013

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled 'Investigating Aspects of Visual Clustering in the Organization of Personal Digital Document Collections' by Hoda Badesh in partial fulfilment of the requirements for the degree of Master of Computer Science.

Dated: March 11, 2013

Research Co-Supervisors: _____

_____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: March 11, 2013

AUTHOR: Hoda Badesh

TITLE: Investigating Aspects of Visual Clustering in the Organization of Personal Digital Document Collections

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: M.C.Sc. CONVOCATION: May YEAR: 2013

_____

Signature of Author

DEDICATION PAGE

To my family, for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Organizing personal collections of digital documents can be frustrating for two main reasons. First, the effort required to work with the folder system on personal computers and the possible misplacement and loss of documents. Second, the lack of effective organization and management tools for personal collections of digital documents. The research in this thesis investigated specific visualization and clustering features intended for organizing collections of documents and built in a prototype interface that was compared to a baseline interface from previous research. The results showed that those features helped users with: 1) the initial classification of documents into clusters during the supervised stage; 2) the modification of clusters; 3) the cluster labeling process; 4) the presentation of the final set of organized documents; 5) the efficiency of the organization process, and 6) achieving better accuracy in the clusters created for organizing the documents.

LIST OF ABBREVIATIONS USED

DMSRPI	Data Mountain Search Results Presentation Interface
BI	Bubbles Interface
PI	Pie Interface

ACKNOWLEDGEMENTS

I am grateful to my supervisors, Dr. James Blustein and Dr. Evangelos Milios. Thanks to the IIE and British Petroleum who supported me. Special thanks to Mrs. Hanan Thabet from IIE. Thanks to everyone who participated in the studies conducted in this research. Thanks to the Faculty of Computer Science at Dalhousie University.

CHAPTER 1 INTRODUCTION

Personal documents grow in size and number rapidly. In the current state, desktop documents can be organized either manually in folder hierarchies or using special software such as: OpenText[1], IBM's Document Manager, and Google Desktop[2]. Manual organization can be very demanding since desktop computers may involve large collections of documents. Every type of software has its advantages and disadvantages. For instance, Google desktop presents its search results in a list provided from searching the index of keywords Google builds from the user documents. This type of presentation may require the user to go through large lists of result hits, formulate several queries, and eventually may or may not find the intended document.

When the pile of documents on a user's desktop grows extensively, organizing those documents into folders may become very time consuming. The use of software that presents lists of results may also be very ineffective. The use of clustering for organizing user desktop documents has had little consideration. User interfaces for assisting users with organizing their documents using aspects of clustering and visualization have not been thoroughly investigated.

Clustering is grouping together documents of the same type, genre, topic, etc. A categorization scheme has to be defined prior to applying clustering. Topical clustering and genre clustering have been investigated (Carpineto et al., 2009; Santini, 2006). The use of clustering in document presentation has been investigated for desktop retrieval as well as web retrieval (Knoll et al., 2009; Alhenshiri et al., 2010a). Clustering makes use of overviews of documents for conveying the different topics or genres covered in the document collection.

Visualization has been widely exploited in information retrieval. Several studies have investigated the effect of visualization on how users find relevant documents among search results (Suvanaphen and Roberts, 2004; Kules et al., 2008). Most search systems such as the Windows desktop search and Google Desktop offer a list-based presentation of search results. This kind of textual presentation does not convey enough features of

[1] www.opentext.com/

[2] Google discontinued the development of Google Desktop. http://desktop.google.com

either the individual documents or the entire collection to help the user find the intended documents especially in the case of broad queries such a single-term ones.

Visualization, however, can help the presentation of multiple features of search results (Badesh and Blustein, 2012; Alhenshiri and Blustein, 2011). Document features such as its size, last update, and type can be visualized. Features of the collection as a whole can also be visualized by showing documents of the same type connected or by showing documents with similar content under one category. Such visual clustering combines the benefits of visualization and clustering. Adding clustering and visualization to the presentation of search results can help users organize large collections of documents and find results more effectively and efficiently.

There are several tools/applications for desktop and online organization and sharing of documents such as Mendeley[3] and Zotero[4]. For example, Zotero collects all kinds of documents and builds a searchable index. This is basically similar to Google Desktop which no longer exists. Moreover, Mendeley is intended for organizing research papers in terms of generating bibliography, collaborating while working on articles, and accessing such content online easily. However, there is little done for using clustering and visualization for specifically organizing personal collections of digital documents on the desktop of a personal computer. The content of those documents may vary and the benefit of clustering in managing and organizing such collections has yet to be investigated.

There are several problems associated with managing and organizing personal documents on desktop computers. Those problems can be summarized in the following:

1- The size of the collection of documents on computers of personal nature grows very rapidly as users keep using their machines.
2- Manual organization of documents on desktop computers necessitates the use of folder structure which may result in:
    a. Excessive time consumption in the case of large collections.

[3] http://www.mendeley.com/
[4] http://www.zotero.org/

b. Losing documents due to the complex structures and the difficulties associated with manually searching those structures.

3- Organization tools may drive the user away for one or more of five reasons (Jones et al., 2008):
    a. Visibility, which is an issue when users cannot see or notice some or all of the important items that they would like to keep a track of in their peripheral vision.
    b. Integration happens when there is no interconnection between the user's old system and new system. Integration may influence visibility.
    c. Co-adoption. Suppose that a group of people work together as a team to develop or share a system for a certain purpose such as information sharing or scheduling. A part of the group fails because of unevenness of the distribution of the amount of work. As a result, they abandon their portion of the work.
    d. Scalability, which becomes an issue when a system cannot sufficiently handle greater complexity of project growth and progress, new circumstances, or increased collaboration.
    e. Return on investment, is an issue when users give up on an information management system because of its complexity and the effort required to learn about that system. Users consider such system as not worthwhile.
4- Search using desktop tools has problems associated with the presentation of the search results and the interaction with the user.

The research discussed in this thesis examines the use of different visualization and clustering features in an interface intended to help the user with:

1- Selecting documents to be used as cluster seeds in a supervised phase of a clustering process intended for organizing a personal collection of documents.
2- Interacting with the interface for editing the selection of cluster seeds.
3- Selecting and modifying the cluster labels based on the content of the documents.
4- Viewing the content of the documents while in the supervised phase of the clustering process.

5- Viewing the clusters created with document seeds at any time in a two-dimensional space.
6- Viewing the final set of clusters that includes the documents organized according to topics initially established by the user.

The research discussed in this thesis attempts to answer the following questions:

1- What is the effectiveness of using three options of document views (abstract, text cloud, full content) on how users classify their documents for organization?
2- What is the effectiveness of presenting the initial clusters during the classification process as bubbles containing glyphs of documents inside each corresponding cluster with different modification capabilities?
3- What is the effectiveness of having different views of clusters: as a list of cluster labels and as labeled bubbles?
4- What is the effectiveness of presenting the final set of documents clustered and organized in bubbles representing topics with their documents represented as glyphs?

The evaluation of interfaces built and investigated in this research used the ISO[5] usability measures of effectiveness, efficiency, and enjoyment (satisfaction). There are several definitions and frameworks of usability (Shneiderman, 1980). ISO defines usability as 'The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.' effectiveness is defined as 'the degree to which an interface facilitates a user in accomplishing the task for which it was intended' while efficiency is defined as 'the rate or speed at which an interface enable a user to accurately and successfully complete a task' according to UsabilityFirst[6]. Doll and Torkzadeh (1988) defined *user satisfaction* as 'the opinion of the user about a specific computer application, which they use.' Throughout this thesis, the engagement and user satisfaction are used changeably.

The research has gone through two stages. In the preliminary stage, a prototype using visual clustering with a data-mountains layout was investigated with a small collection of

[5] http://www.iso.org
[6] http://www.usabilityfirst.com/glossary/effectiveness/ and http://www.usabilityfirst.com/glossary/efficiency/

web search results. The study was intended to reveal the effect of using data mountains to group together web documents that belong a topic based on their content on how users perceive relevant topics and documents. The main stage of the research compared the use of visual clustering and visualized interaction during classifying documents to a more text-based interface intended for the same kind of clustering for organizing and managing personal collections.

The first study in this research used the factor of enjoyment only in evaluating the prototype interface. The enjoyment of the Data Mountain Search Results Presentation Interface (DMSRPI: see Chapter 3) was measured using different metrics such as the user perception of the helpfulness, usefulness, and effectiveness of the interface. In the second study, the efficiency of the interfaces was measured (to be compared) using the time needed to complete the tasks and the number of mouse clicks required. The effectiveness of the interfaces was measured using the accuracy of clustering identified by the users after completing the management and organization process.

The main study compared two interfaces, namely: the Bubbles Interface and the Pie Interface. The Pie Interface was used in a previous research to evaluate different clustering algorithms (Hu et al., 2012). The new interface (Bubbles Interface) was designed to overcome some of the issues users had with the Pie Interface. Both Interfaces used the same underlying clustering algorithm. The study was intended to evaluate the perceived effectiveness of the Bubbles Interface, its efficiency, and the user enjoyment of the interface compared to the Pie Interface.

The results showed that users favored the Bubbles Interface with respect to the following criteria:

1- The ease of selecting documents during the supervised classification stage.
2- The positioning of the document and cluster views on the display.
3- The feedback received from the interface during the organization process.
4- The future use of the interface for organizing the users' personal collections of documents.
5- The categorization of documents in the final results of the organization process.
6- The creation of cluster labels for organizing the document collection.

7- The interaction and visualization features embedded in the interface in general.

With similar collections of documents and similar groups of users, the Bubbles Interface may be effective for its purpose. It may serve users in categorizing personal document collections (digital documents). It could also be used for categorizing a small collection of web search results online. Future investigations may involve testing the interface with different groups of users with more types of documents. Using the interface for categorizing web documents can be considered in online clustering for web documents. The evaluation of the interface stopped at having the final sets of documents organized into clusters. The use of the interface may be investigated beyond this point. The interface can have keeping and re-finding capabilities, update of collection capabilities, and continuous clustering of the updated collections.

CHAPTER 2 RELATED WORK

The research discussed in this chapter is threefold. First, studies related to managing and organizing collections of digital documents are discusses. Second, the use of visualization in organizing collections of documents is illustrated. The last part discusses the use of clustering in managing collections of documents.

2.1 MANAGING COLLECTIONS OF DOCUMENTS

Managing and organizing information has been explored in different directions. Knoll et al. (2009) investigated how users view and manage desktop information in general. Jones et al. (2008) investigated important reasons behind giving up on certain personal information management tools. The strategies users follow to manage web information in order to be able to relocate and reuse information previously found are discussed in the work of Jones et al. (2003). Their work showed that users follow different keeping strategies to re-find and compare information later. The variety of managing and organizing strategies for personal information can be attributed to the fact that current tools lack important reminding, integration, and organization schemes (Cutrell et al., 2006).

Jones et al. (2008) found that users abandon the use of an information management tool for one or more of five closely related reasons: visibility, integration, co-adoption, scalability, and return on investment. Jones (2007) reviewed research in support of a more general preference for *way finding* methods that depend on a sense of digital location vs. direct search as the primary means for access to personal information (Civan et al., 2008). Bergman et al. (2008) indicated that direct search becomes the user choice for retrieving personal information after attempts for search by navigation fail.

Jaballah (2005) designed a desktop personal library manager to overcome the problems associated with the use of folder-based organization schemes. Users can browse and search their personal collections of documents by the document type, title, filename, date of modification, and so on. The interface was evaluated using a pilot study (two experts) followed by a learnability study and final diary study (6 participants). The results showed that even with the prototype's ability to harvest metadata about the files in the collection,

the users preferred the standard folder system. They reported that some actions on the prototype were difficult and that users spent most of the time trying to familiarize themselves with the interface.

To further emphasize the value of visual access to information for managing and organizing personal collections. Bauer (2005) built an interface intended to arrange piles of images or PDF documents in portraits. Each PDF file in the portrait is shown as one page containing images and parts of the text in the documents. Images are shown in their own piles. The closer the image to the user, the larger the size of the document is. The prototype allowed interactions with collections of documents to be logged over long periods. The prototype was not evaluated and it was expected to improve the user's experience with managing piles of personal documents and images.

Civan et al. (2008) compared the user behavior for organizing information using folders and using labels (tags). For the purpose of the comparison, Gmail[7] and Hotmail[8] were selected. Users organize their e-mail messages using different methods in the two systems. Gmail's users label or tag their messages; Hotmail's users put messages into folders. The two approaches were compared with respect to: 'retrieval performance, evolution in mappings between articles and folders/labels over time, and limitations to fully express one's internal conceptualization' (Civan et al., 2008). No clear winner was identified between tagging and placing. The study concluded that 'better support for information organization may need to go well beyond folders and tags or their artful combination' (Civan et al., 2008).

Managing information is concerned with how people store, organized, and re-find information (Elsweiler and Ruthavan, 2007). Information management systems are methods by which users find, categorize, and re-find information on daily basis. Research has considered personal information management. However, there is further need for investigating organizing and finding information in cases where the personal collection of documents grows extensively and when standard folder-based organization becomes demanding.

[7] http://mail.google.com
[8] http://www.live.com

2.2 VISUALIZATION

Visualization is a concept that has been in focus for research in information retrieval (Card et al., 1999; Risden et al., 2000; Roberts et al., 2002, Nguyen and Zhang, 2006; Friendly, 2009). Information visualization may improve users' performance by harnessing their innate abilities for perceiving, identifying, exploring, and understanding large volumes of data (Card et al., 1999; Friendly, 2009). There are several visualization-based prototypes that have been investigated for improving the effectiveness of search results (Kules et al., 2008).

Teevan et al. (2009) investigated the use of visual snippets in the presentation of web search results and compared the effectiveness of this approach to the conventional text snippets provided by search engines such as Google. The results showed that combining text with the most important images on a webpage may help users recognize the page more easily and be able to select documents of interest more effectively. The use of a *3D City Metaphor* in the work of Bonnel et al. (2006) also showed that users favored the visual presentation of clustered results. Visual thumbnails of search results that accompany textual presentations were also shown to be effective in searching for revisiting (Woodruff et al., 2001).

To further reveal the relationships (similarities) among documents to users for more effective exploration of search results, Zaina and Baranauskas (2005) designed a visual interface (called *ReVel*) for exploring search results. The interface uses a graphical representation of the documents presented. Result hits are connected via links representing similarities among documents presented on the display. ReVel allowed its users to integrate results of multiple sessions for further explorations; however, the visualization used content similarity as the only attribute conveyed to the user. Moreover, the display suffered from clutter due to visualizing similarities among all documents using graph vertices.

To provide topical overviews of search results, Paulovich et al., (2008) designed a search interface called *PEx-Web* that supported interpretation of collections of documents. PEx-Web permitted users to avoid excessive visiting to unwanted documents and to discover relevant documents based on visual topic representations through visual clustering. Both

approaches (Zaina,and Baranauskas, 2005; Paulovich et al., 2008) were shown to be effective when compared to raw presentations of search results. Providing a topical overview may result in very effective organization of personal collections of documents and help the user find their intended document efficiently.

Periscope, a prototype investigated by Wiza et al. (2004), showed that visualization was effective in presenting extensive numbers of results on a display and assisting users with finding information of interest. The use of visualization in results presentation was also investigated in the works of Bonnel et al. (2005, 2006). Furthermore, visualization plays an important role in how users explore collections of documents in techniques and prototypes that use visualized clusters (Kules et al., 2008; Carpineto, 2009).

Visualization of search results has been investigated in several layouts including the use of hyperbolic trees (Rivadeneira and Bederson, 2003), Scatterplots (Kules et al., 2008), Self-Organizing Maps (Au et al., 2000), and thematic maps such as in the visual search engine Kartoo[9]. Most of these approaches were intended to help users find relevant results. Exploring personal collections of documents has improved over the years using different layouts for presenting folder contents on personal computers. Some of these layouts used thumbnails with different sizes to visualize the collection of documents viewed.

Visualization of hierarchical data has been extensively researched in the past leading to numerous approaches. Early treemaps used the simplest treemap algorithm to implement which is called slice-and-dice. The main characteristics of slice-and-dice treemaps are long rectangles with a very small width. Shading indicates the order and the size of each rectangle reflects the size of the file. Some issues were found with the slice-and-dice treemaps that needed to be optimized to make treemaps more effective. For example, slice-and-dice treemaps often fell short in visualizing the structure of a tree. In addition, slice-and-dice treemaps presented files of the same size in vastly different shapes in the same area. These issues made comparing sizes problematic. The flat layout caused difficulties in understanding hierarchies and finding data (Bladh et al., 2004).

[9] http://www.kartoo.com

The Cushion algorithm is another extended version of the original treemaps algorithm that takes advantage of how a visual system is trained to interpret variations in shades as illuminated surface (Wijk, 1999). Various coloring options are available to show the size, the level, and other attributes of the documents in the leaf nodes. The cushion treemaps are very efficient in creating images in computers. Furthermore, the cushion treemaps visual structure is much more effective compared to the original treemaps. An implementation of this algorithm is the SequoiaView, for showing Windows PC files directories (Intranet, 2013).

The *iCluster* system that was introduced by Microsoft researchers aimed to help users manually classify a set of documents faster (Drucker et al, 2011). The process of cluster creation by users was monitored to assist the clustering algorithm and machine learning techniques to provide users with recommendations. These recommendations appear on the system's interface as links that connect unclassified documents with most related clusters after monitoring users' performance during creating clusters. iCluster does not give any recommendations until the user creates some clusters (semi-supervised clustering). Based on machine learning and active learning techniques, the system gets improved by learning each user's personal interests.

The results of the user study showed that the clusters were very different even with the same user and the overlap of having a document appear in multiple clusters was low. The clusters created were different from one user to the other, and almost all users were satisfied with their manual clustering and iCluster's recommendations. This explains how the iCluster prototype was more efficient than manual clustering systems and fully-automatic clustering systems. However, the random positioning of the clusters created was overwhelming. There should be a standard structure to organize clusters in order to help users remember where the cluster was. This characteristic would make the system more effective. In addition, the view of a cluster was largely cluttered.

The use of visualization might not be universally appropriate. The work of Alhenshiri et al. (2010a) showed that some users complained about issues of clutter in the visual search engine investigated. In addition, 3D visualizations can inhibit users and make interfaces more confusing (Kules et al., 2008). Visualization can also make the exploration of

search results more frustrating in cases where no meaningful axes are defined on the display (Turetken and Sharda, 2005). Moreover, some visualization layouts can be unproductive such as the use of *Data Mountains* for browsing tasks as demonstrated by Cockburn and McKenzie (2002). In several research works, visualization has been investigated along with clustering as discussed in the following section.

2.3 CLUSTERING

Clustering is a process intended for grouping together items that share similar characteristics and attributes. In web information retrieval, clustering is meant for grouping similar documents (Manning et al., 2008). The use of clustering has been widely investigated in web information retrieval (Ferragina and Gulli, 2005; Katifori et al., 2007). Clustering is usually intended to provide overviews of information categories (topics) in the result set. Hence, efficient subtopic retrieval is anticipated with the use of clustering in document presentations (Carpineto et al., 2009). Clustering can also decrease the need for scrolling over long lists of documents (Spink et al., 2001), resulting in more effective and efficient user performance. Moreover, clustering has other benefits including capturing meaningful themes in the search results, scalability, and domain independence (Efron et al., 2004).

In web information retrieval, clustering has been investigated in several prototypes such as in the work of Zamir and Etzioni (1999), Jing et al. (2006), and Alhenshiri et al. (2010b). Clustering has also been implemented in conventional search engines such as Clusty[10], Gceel[11], Northern Light[12], and Google (in their *see similar* feature and the former *Google Wonder Wheel*). Turetken and Sharda (2005) used a graph-based visualization that shows relationships between clusters. Their technique was shown to be more effective than ranked textual lists of results. Moreover, Cai et al. (2004) showed that clustering is very effective in image search. Although the performance of users with row presentations of web documents is comparable to their performance with clustering-based presentations, user preference usually comes in favor of clustering-based

[10] http://search.yippy.com/
[11] http://www.gceel.com/
[12] http://www.nlsearch.com/home.php

approaches (Carpineto et al., 2009). In addition, there are indications that clustering can even be more effective.

In addition, the concept of genre-based clustering is anticipated to improve the effectiveness factor in searching collections of documents especially collections of large nature. A genre is a class of documents that are similar with respect to content, structure (form), and functionality (Dong et al., 2008). Classifying documents by genre can be considerably accurate (Mason et al., 2009). Features based on which document genres are identified have also been studied and investigated (Ferizis and Bailey, 2006; Stubbe et al., 2007; Santini and Sharoff, 2009). However, the effect of genre-based clustering on the relevancy and accuracy of search results has yet to be investigated.

Berendt et al. (2010) introduced a tool that helps people in education (i.e. professors, students, etc.) while searching for scientific publications/literature. The system was meant to be a dialogue in which the user asks a question, and the system provides users with the answer by extracting the meaning using data mining techniques. The system retrieves the most related articles, and presents them in a text-based list. These lists (clusters) are separated by a bold line. Users can edit the clusters by deleting what they believe is not related to the cluster, assigning proper labels (the default label is 'Group#') to clusters, and reconstructing the clusters by moving results among different clusters. In addition, results can be saved for further processing or shared with other users. The usability evaluation of this interactive approach revealed that 82% of the users preferred the meaning extracting and clustering techniques used in this tool rather than their own way of searching. However, the quality of the clusters needed to be further considered by using larger clusters.

Clustering provides topical overviews of the information presented. This feature can be utilized in presenting large collections of documents. Clustering labels can play an important role in improving the effectiveness of users attempting to search collections of documents of different types and topics. However, there are some drawbacks associated with clustering. The labels selected for clusters may not reflect the content of documents in the cluster which may mislead the user when trying to find a particular document. In

addition, documents may belong to different topics and be associated with one cluster resulting in misleading the user.

In the case of larger collections of documents, speed is a concern in online clustering. Off-line clustering, on the other hand, may suffer in cases where personal collections change frequently and increase rapidly. Moreover, generating meaningful groups and effective labels is a problem usually recognized with the use of clustering (Shneiderman et al., 2000). Cluster labels are often generated from the titles and/or summaries of search results. In the case of using the entire document, creating a meaningful label can be very difficult due to having to deal with much more text than in the case of using summaries (Manning et al., 2008).

Clustering personal collections can be done in a semi-supervised fashion where the user selects the seeds around which the rest of the cluster documents are added by the clustering algorithm. The user interaction may help the process of organizing the documents in an effective way. Moreover, the user can select the labels representing the topics of the clusters created during the organizing-by-clustering process of their personal collections.

Prior to investigating classification and clustering of a large personal collection of documents using aspects of clustering and visualization, the next chapter discusses a preliminary study. In that study, clustering and visualization were applied to a small collection of documents from the web. The purpose of the study was to examine the effectiveness of the particular features embedded in a search interface on how users find information.

CHAPTER 3 INVESTIGATING ASPECTS OF VISUAL CLUSTERING WITH A SMALL COLLECTION OF DOCUMENTS

3.1 INTRODUCTION

This chapter presents a preliminary study in which a small collection of documents was used as a dataset. The study examined the use of clustering and visualization in organizing and presenting search results on the web. The study was motivated by the principle that presenting search results as a list of hits is often insufficient to assist users to find relevant documents and discover varied topics among search results. The use of visualization was intended to assist users to find relevant documents. A Data Mountain Search Results Presentation Interface (DMSRPI) (Badesh and Blustein, 2011a; Badesh and Blustein, 2011b) was designed and implemented to improve the effectiveness and efficiency of users searching the web through organizing the search results in a meaningful and easy to understand and perceive way.

In essence, the design of the DMSRPI interface was motivated by the need for a visualized clustering layout to improve how users find and re-find documents. The case of the web was investigated with a small collection of documents to develop recommendations for further studies that involve the use of aspects of visualization and clustering in organizing documents.

3.2 DESIGN AND IMPLEMENTATION

The DMSRPI was designed using the *prefuse*[13] visualization toolkit discussed in the work of Heer et al. (2007). Java Swing components were used in this design. The DMSRPI system was designed to satisfy the following criteria:

- Visualization of search results should make it feasible for users to recognize webpages of interest among the rendered results. Therefore, snapshots of webpages (thumbnails) are presented on the DMSRPI's display.
- Relationships among webpages (clustering) are shown without the clutter that results from using edges. The DMSRPI implementation uses the concept of a data

[13] http://prefuse.org/

mountain in which related webpages are shown close to each other and far from documents that belong to other topics.

- Focus+context are provided in the DMSRPI prototype as shown in Figure 1. The non-labeled DMSRP Interface figure is provided in Appendix F. This feature integrates focus and context into a single display where all parts are visible at the same time. The focus is displayed seamlessly within its surrounding context. Users overview the results in the form of clusters (mountains) with the ability to hover over each page to see details including the page's summary, URL, and title. In addition, hovering over a search hit results in magnifying the page thumbnail to provide more focus.
- The least amount of text is shown on the display while users explore search results. This helps users to focus on the similarities among search results especially for tasks that need more than one topic and sources of information such as information gathering (Kellar et al., 2007).
- The DMSRPI gives the user the ability to keep a subset of the search results under a label (the query keywords by default) and with the ability to add comments to the results kept. Re-finding the results saved can be either by searching past queries in a list or by keyword search.

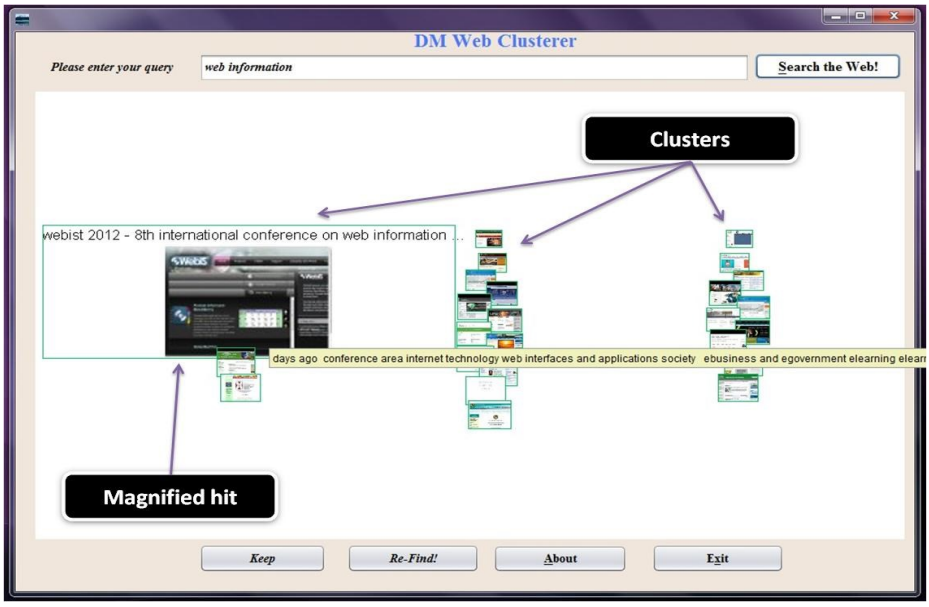


Figure 1. DMSRP Interface.

The interface gives the user the ability to keep a subset of the collection of search results presented during one search activity. The results are kept associated with the query and user comments. This feature is intended to improve re-searching by allowing keeping and re-finding for a group of webpages related to a query and most likely to a task the user is performing at the time of the query submission. In addition to the group of search results, the user is given the ability to add comments to the saved set of pages to make re-finding that set later more effective by searching the user comments, the titles of the pages, or a list of labels given to each group of pages kept for re-finding.

Organizing documents for re-finding is considered in the DMSRPI prototype. When the user attempts to search the web to re-find certain pages, sometimes, due to the evolutionary nature of the web and the continuous changes of the ranks given to webpages, users cannot re-find the same webpage even if they submit the same query. The feature added to DMSRPI makes re-searching easier by re-finding the results locally on the user's computer. The landmarks (search cues) kept for the user include the label given to the saved set of results (by default, the search query), the page title, and the date on which the results were kept.

3.3 DMSRPI SEARCH, PRESENTATION, AND RE-FINDING

When the user submits a search query, the system forwards the query to both Google and Yahoo search engines through their respective APIs (Google AJAX search API[14] and Yahoo Boss[15]). The results coming from both search engines are then filtered for elimination of repeated hits. The purpose of using search results from two different search engines is to cover as many topics as possible in the final set of results. Tyler and Teevan (2010) have shown that although the overlap between Google and Yahoo can be high, it is not significantly high. Therefore, having both search engines as search results providers may have some additional benefits.

The second step is computing the similarities among search results for presenting relationships. This is done in the DMSRPI using the cosine similarity (Manning et al., 2008), which is based on using content similarity. The similarity is computed between the

[14] http://code.google.com/more/
[15] http://developer.yahoo.com/search/boss/

summaries of webpages provided by the underlying search engines. Each page is compared to each and every other page in the result set. A threshold is used to decide on which documents belong together in one cluster. The threshold is automatically determined to produce a reasonable number of clusters that does not clutter the display.

Thirdly, the available directory of Google preview[16] is used for supplying webpage thumbnails to DMSRPI. Each page is assigned a thumbnail and a cluster to which it belongs as shown in Figure 1. Each page belongs to one cluster only (*hard clustering)* (Manning et al., 2008). Moreover, the presentation algorithm used in DMSRPI uses special characteristics of the display to present the results in the shapes of data mountains (clusters). That is, each cluster is presented taking the shape of a mountain. The closer the results to the user, the higher the rank. The ranks of the results depend on the order given by the underlying search engines.

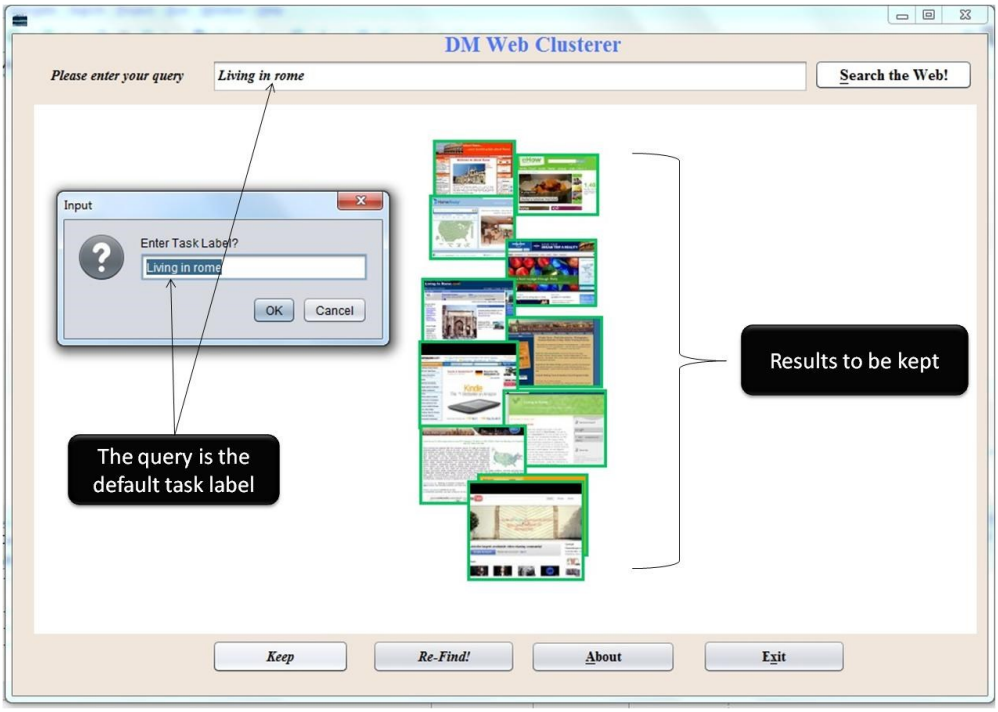


Figure 2. Keeping a set of search results.

The DMSRPI was built to support re-searching for Web documents by providing a form of organization by keeping selected sets of documents. Users can use the system to keep (save) complete search sessions by eliminating unwanted results from the display and

[16] Google Preview is no longer available.

leaving a group of documents to be kept for later re-finding under a task label (as Figure 2 shows). The user can later look up the query from a list of labels preserved earlier as shown in Figure 3. The results that were kept appear on the display taking the user back to the search context which he or she kept earlier. This feature was intended to support users trying to find the same results that were once rendered for a search query and that can no longer be found even by submitting the same exact keywords in the query.

Figure 3. Re-finding.

3.4 EVALUATION

To evaluate the DMSRPI, a pilot study was conducted to measure the user enjoyment of the interface and provide indications about its possible effectiveness. The study was intended to seek possible improvements needed in the interface and to provide indications for possibly conducting a much larger study to investigate the effectiveness of the DMSRPI. The study was also intended to reveal visualization and clustering aspects that could be used in organizing larger collections of different kinds of documents.

3.4.2. Study Procedure

Each participant was asked to fill out a pre-study questionnaire prior to conducting the evaluation. The DMSRPI interface was then explained to the participant. Afterwards, each participant was asked to use the interface to search the web and to also use the keep-and-re-find features. After the participants finished using the interface, they were asked to fill out an exit questionnaire. Figures 4 and 5 show the pre-study and post-study questionnaires.

Pre-study Questionnaire

(1) What search engines of the following do you usually use? (You can select more than one)
_____ Yahoo.
_____ Bing.
_____ Google.
Others:_____

(2) How satisfied are you with the current Web search engines that you use? *Select one please*

(very satisfied, satisfied, not sure, dissatisfied, very dissatisfied)

(3) Do you use bookmarks?
___ Yes.
___ No, why? _____

(4) How often do you use bookmarks to keep Web search results? *Select one please*
(very often, often, sometimes, rarely, never)

(5) Do you use any session keeper add-ons?
___ Yes.
___ No.
___ No idea!

Figure 4. Pre-study questionnaire.

Post-study Questionnaire

(1) Are you satisfied with the search interface?
___ Yes.
___ No.
___ Not sure.

(2) How effective do you think the presentation of search results was? *Select one please*
(very effective, effective, not sure, ineffective, very ineffective)

(3) How helpful do you think clustering the search results was? *Select one please*
(very helpful, helpful, not sure, helpless, very helpless)

(4) How interesting did you find the use of thumbnails to represent search results? *Select one please*
(very interesting, interesting, not sure, not interesting, not interesting at all)

(5) How helpful do you think keeping a subset of search results was?
(very helpful, helpful, not sure, helpless, very helpless)

(6) What are the most enjoyable features of the search interface?
_____
_____
_____
_____
_____

(7) What do you think should be improved about the interface?
_____
_____
_____
_____
_____

Figure 5. Post-study questionnaire.

3.4.2.1. Study Population

There were six experimental participants in the pilot study who were graduate students from the Faculty of Computer Science at Dalhousie University. The study was conducted in the Web Information Filtering Laboratory in the aforementioned Faculty. The study took approximately fifteen minutes with each participant and involved a short training session, a simple user task, and a pre-study and a post-study questionnaire.

3.4.3. Study Data

The data accumulated in the study involved only the answers to the survey's questions. No user interactions with the interface were recorded for the purpose of this study since the intention was to only measure some engagement factors with the DMSRPI visualization and clustering features for possible use in further studies. The data involved answers to close-ended questions in the questionnaires as well as comments stated by the participants.

3.4.4. Study Results

The study revealed some important indications about the benefits of the DMSRPI and its potential usability as an effective search interface for finding, managing, and re-finding search results.

Five of the six pilots indicated that they were satisfied with the DMSRPI with respect to the presentation of search results for finding relevant documents and also with respect to the keep-and-re-find feature intended for search results. Most participants (4/6) indicated that they believe the DMSRPI can be effective for presenting search results as clustered thumbnails. Five of the participants in the study thought that the clustering feature was helpful. All participants thought that using thumbnails in data mountains (clusters) for presenting search results was interesting. Five participants thought that the keep-and-re-find feature was helpful. The results of this part of the study are depicted in Figure 6.

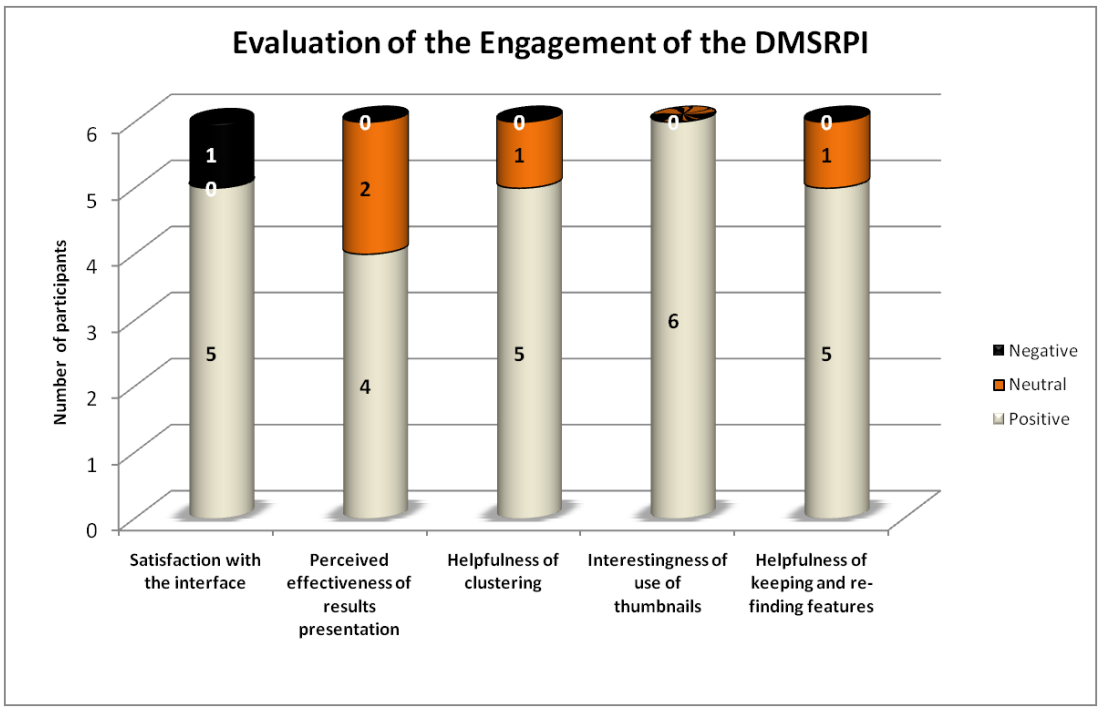


Figure 6. The DMSRPI engagement evaluation results.

Users in the study left 13 comments regarding some of the interesting features that they liked in the DMSRPI. The comments concerned the usefulness of the overview provided by the clustering scheme, the thumbnail view of webpages, the keep-and-re-find feature, and how the interface was easy to use. In addition, users left important comments regarding how the DMSRPI interface could be improved. Users complained about some display issues such as the size of the thumbnail when the user hovers over the result hits. Some other comments regarded the need for cluster labels, making keeping easier by selecting what should be kept in a drag-and-drop fashion, and some efficiency issues. The results of this part of the study are shown in Figure 7.

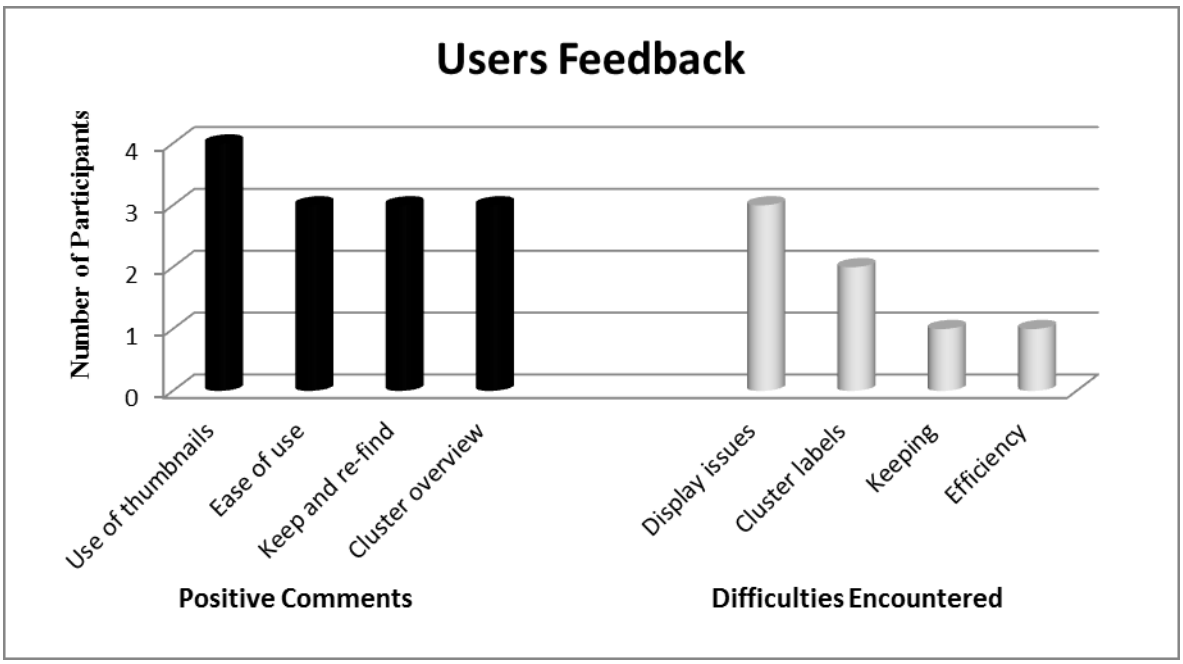


Figure 7. User comments regarding their preference and difficulties encountered with the DMSRPI.

3.5 Discussion

Research shows that there has been a great deal of focus on using visualization in documents representation for many reasons such as managing documents for re-finding and finding relevant documents. However, many of the visualization techniques discussed in the literature suffer from several drawbacks such as inefficiency and clutter due to inability to present a great deal of documents or due to expense associated with the technique in use.

The interface presented in this study was intended to improve how users find search results by exploiting visualization and visual clustering. The DMSRPI was also intended to enhance the user's ability to manage by keeping and later re-finding search results. The results of the pilot study indicate that the DMSRPI has a good potential as an effective tool that can be used for organizing documents for finding and re-finding purposes. Most users expressed great interest in using the interface and considered it as a helpful tool. The comments left by the participants in the study will be used to improve particular features in the current design of the interface and its functionality.

3.6 LIMITATIONS

Among the limitations of the study is the small number of participants who do not reflect the entire population of users searching the web. In addition, the study measured some engagement factors in the interface without considering the actual effectiveness and efficiency in finding and re-finding documents organized in clusters. The DMSRPI needs a larger scale study with different kinds of users. The evaluation should consider the effectiveness of the interface with regard to the relevant pages and topics covered in the cluster organization.

3.7 CONCLUSION

The DMSRPI prototype employed webpage snapshots for presenting each individual result while using a data mountain layout for presenting each cluster of documents that belong to a topic covered in the result set. The initial study conducted to evaluate the interface indicates that the DMSRPI can be a useful, helpful, and interesting tool for finding, managing, and re-finding web documents.

There are several recommendations which resulted from this study for future designs of interfaces intended for presenting clustered documents. First, the use of thumbnails and titles for reflecting the content of documents in clusters can be helpful in providing the focus into context (document details among other documents into a cluster among other clusters). Second, the organization of documents can be provided in more than one form of clustering. The data mountains' shape for clusters did not permit for having labels given how the documents were laid out into each cluster. Using an alternate approach for representing clusters could have been more effective.

CHAPTER 4 INVESTIGATING A USER INTERFACE FOR ORGANIZING PERSONAL COLLECTIONS OF DOCUMENTS

4.1 INTRODUCTION

The research discussed in this chapter extends the work of Hu et al. (2012), which involved building a clustering-based system with a simple interface to organize personal collections of desktop documents. This chapter discusses the design, implementation, and evaluation of a new interface (Bubbles Interface) which is compared to the Pie Interface discussed in the work of Hu et al. (2012). Both interfaces use the same underlying clustering algorithms and have the same purpose of assisting users with organizing their personal collections of digital documents. A user study was conducted to evaluate and compare the two interfaces.

The study discussed here was motivated by the problems associated with organizing personal collections of digital documents which are illustrated in Chapter 1. Those problems included:

1. The continuous growth of personal collections which makes organizing them manually on the desktop rather time consuming.
2. The difficulties associated with folder structures when used for organizing documents which include the loss of information and the consumption of a great deal of time.
3. The ineffectiveness of search and organization tools currently used which leads users to abandoning such tools.

The specific clustering and visualization features implemented in the Bubbles Interface were developed based on the results and observations noted in the study discussed in Chapter 3. The study in Chapter 3 resulted in developing the following recommendations:

1. Optimizing the shape of clusters to become more intuitive to the users.
2. Providing multiple options for viewing clusters of documents.
3. Providing more effective focus+context with less clutter in the case of zooming in and out.

The experience with the DMSRPI, which was intended to handle a small collection of documents, helped with choosing the clustering layout, the visualization features, and the overall design of the display in the Bubbles Interface. The data mountains layout was replaced with the bubbles layout in the new interface. However, the benefit of thumbnails was preserved in the new design. The Bubbles Interface provided better focus+context and a more effective labeling scheme.

4.2. Document Organization

In both interfaces, there are two stages to organize the personal collection of user documents. First, in the supervised clustering stage, the user choses from the collection of documents what is considered seeds of documents into each initial cluster. The user decides the number of clusters needed and can change that number by adding or eliminating clusters during the supervised stage. After finalizing the classification process and selecting a number of documents for each cluster as seeds, the user executes the clustering algorithm which uses the cluster seeds to gather the rest of the documents in the collection around the seeds selected by the user. This stage allows the user to move documents around before the non-supervised part of the clustering process (the clustering algorithm) is executed.

After the clustering is performed, the user can have the entire collection of documents organized into clusters. Content-based (topical) clustering is used in both systems.
How the supervised stage of the clustering process is performed and how the final results are presented to the user are what makes the difference between the Pie Interface designed earlier and Bubbles Interface presented in this research.

4.2.1. The Pie Interface

The Pie Interface has been used in the work of Hu et al. (2012) to evaluate different clustering algorithms. The enjoyment of the Pie Interface was measured yet with no comparisons in their study. The interface is divided into four sections as shown in Figure 8. The non-labeled Pie Interface figure is provided in Appendix G. A demonstration video showing the functionalities of the Pie Interface can be located at the following website: http://web.cs.dal.ca/~jamie/pubs/Movies/YemingHu%27sPhD/demo.htm.

4.2.1.1. The Supervision Panel Section

This section appears in the upper left corner in Figure 8. The pie chart in this section consists of a central circle that represents a document ID surrounded by two permanent sectors. These sectors are the *Trash* sector used by the user to get rid of any unwanted documents, and the *New Cluster* sector used by the user to drag any document to create a new cluster. Once a new cluster is created, a new sector with a different color will appear on the pie chart. Documents appear as yellow stripes on each sector (cluster). Documents are added to clusters by dragging and dropping the document in the corresponding sector that represents that cluster. Furthermore, users can merge any two clusters by dragging one cluster and dropping it into the other so that they become one sector on the chart.

Figure 8. The Pie Interface and the four subsections.

4.2.1.2. The To-be-Labeled Document View Section

This section is in the upper right corner of the interface. It shows the document ID, the document text cloud, and the entire content of the document. Labels can be selected from either the text cloud view or the entire content view of the document.

4.2.1.3. The Cluster View Section

It appears in the lower left corner of the interface. This section allows the user to view the cluster ID, the cluster label, and the text cloud of all the documents that comprise the cluster.

4.2.1.4. The Labeled Document View Section

This section lies in the lower right corner of the display. Its purpose is to allow the user to view information about a document in a cluster. The information includes the document ID, the ID of the cluster that contains the document, the keywords used to label that cluster, and the document content (i.e. document text cloud and document whole content). The user may want to make changes to particular documents that have already been assigned. This view allows them to view individual documents and make changes.

4.2.1.5. The Clustering Process

The IDs of unclassified documents appear automatically in the supervision panel. The information of the document appears in the 'To-be-Labeled Document' view to allow the user to review the document content. The user can also select keywords—if wanted—to label the cluster to which the document will be assigned. To create a new cluster labeled with keywords that were selected from the document, the user can drag the document ID to the *New Cluster* sector on the pie chart. Documents are viewed in the same way and can either be used to create a new cluster or to be assigned to a cluster that has already been created.

4.2.2. The Bubbles Interface

This interface was designed and implemented to allow users to organize their personal collection of documents based on clustering and using aspects of visualization in both the classification stages, which is the supervised portion of the process, and the final presentation stage of the organization process. The Bubbles Interface (shown in Figure 9 and Appendix H) is designed to overcome several disadvantages in the Pie Interface. To

describe the difference between the two interfaces, the following is a discussion of each part of the Bubbles Interface followed by a description of the organization process.

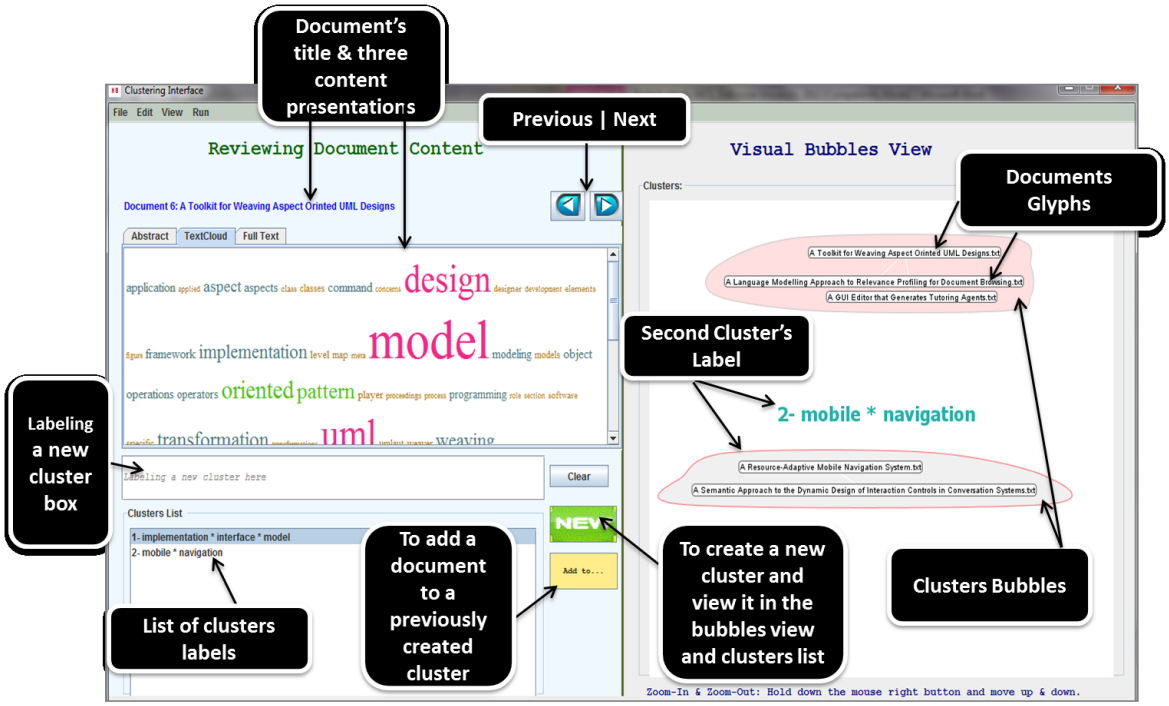


Figure 9. The Bubbles Interface and the most important features and views.

4.2.2.1. The Document View

This part of the interface concerns showing users the content of the document during the supervised stage of the clustering process. There are three options for viewing the document content in the Bubbles Interface. The document title is available with all of these views.

- The *Abstract View*: this view is shown in Figure 10. The text of the abstract of the document is shown to the user and the user can flip back and forth among the documents in the collection to see each abstract. Users can double-click any term in the abstract to add it to the label of the current cluster that is being created.

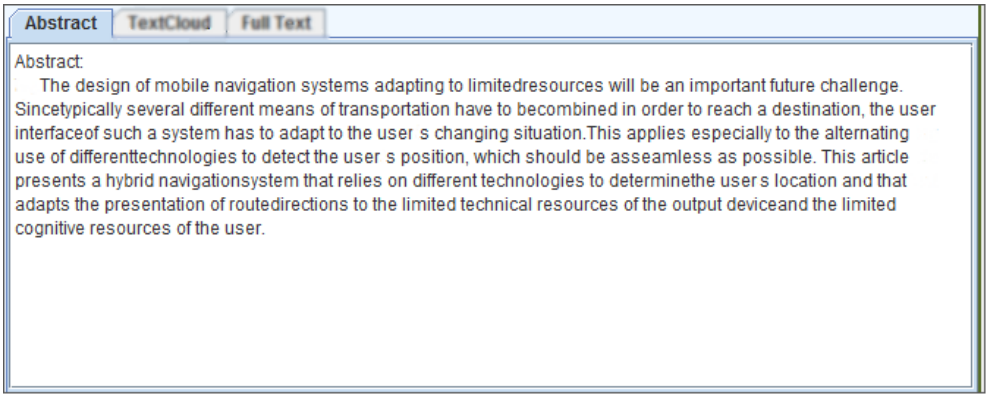


Figure 10. The abstract view of the document content.

- The *Text Cloud View*: this is shown in Figure 11. In this view, the most frequent terms in the document are shown as a text cloud. The frequency of the terms is reflected using text size. Colors are used to help the users to see different terms.

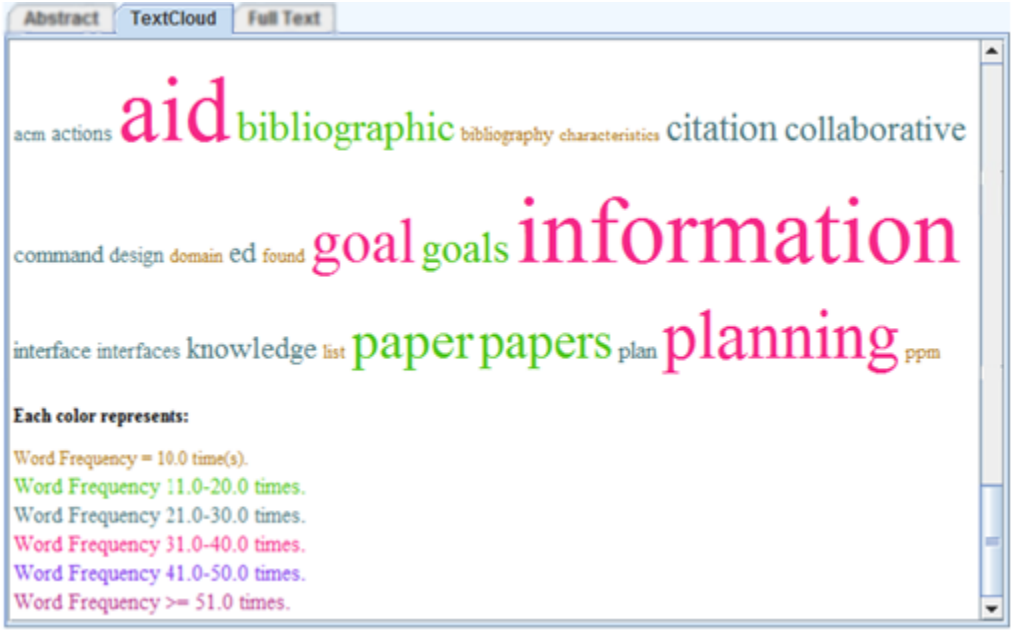


Figure 11. The text cloud view of the document content.

- The *Full Text View*: In this view, the entire text of the document is shown to the user as depicted in Figure 12. In this view, the article's title is colored in blue while the rest of the text is colored in black with a well-organized structure.

Figure 12. The full-text view of the document content.

- The *PDF View*: This view is used when the user selects a document which has already been assigned to a cluster, from the visual clusters by left-clicking its corresponding glyph. The PDF view will appear in a separate window. The content of the document will appear in a PDF format.

4.2.2.2. The Clusters View

This view provides assistance to the user in two different ways. First, as the user adds documents to clusters, modifies clusters by eliminating documents, changes cluster labels, or removes clusters, the user can notice the change immediately on this view. The bubbles (clusters) and the contained glyphs (documents) will be affected by the changes to help the user keep up with the supervised stage and their created topics.

The second purpose of the view is to provide the final organization (clustering) of the documents. Documents belonging to the same topic (i.e. having similar content) are visualized as glyphs within a bubble that simulate a cluster (see Figure 9). The user can hover over any cluster to see its label and identify the underlying topic. The user can see the document titles on the glyphs representing the documents within each cluster. Zooming is provided for better views of the clusters and the documents in the clusters.

4.2.2.3. The Cluster Labeling and the Clusters List

The cluster labeling box is a text box in which the user can type the label of the cluster. In addition, the user can double-click any terms in the document view (abstract, text cloud, or full text) to be added to the cluster label in the labeling box. The clusters list has all the

labels of the clusters that the user has already created. The user can click on any label in the list to make it active and modify the corresponding cluster or show its content in the cluster view. The user can delete clusters from the list.

4.2.2.4. The Dropdown Menu

This menu has additional functionalities that users can perform. For example, the user can merge any two clusters already created into one cluster by selecting *Merge Clusters* from the *Edit Menu*. The merge window pops up to allow the user to select the clusters to be merged. The outcome of the merging process is a new cluster that consists of all the documents in all of the merged clusters. The title of the new cluster will be a combination of the titles of the merged clusters separated by slashes '/'.

4.2.2.5. The Organization Process on the Bubbles Interface

The organization process on the Bubbles Interface goes through the following steps:

- The interface presents the user with the title of the document in addition to one of three viewing choices defaulting to the document abstract. The other options, which the user can use, are the view of the entire article of the document or the text cloud view of the document. The user can go back and forth through the collection of documents to see these elements and decide whether to create a new cluster that starts with the current document or to add the current document to an already established cluster.
- The user continues viewing documents and adding those documents to clusters. The document could either be used to start a new cluster or to be added to the cluster being formulated or even to an already finished cluster. Documents can be removed from clusters created and each of those clusters can be modified at any time to conform to the user needs.
- Not only can the user see a list of documents added to the cluster and a list of clusters being created during the supervised stage, but they can also see bubbles of clusters containing the glyphs representing the documents inside each cluster. This is done to make it more feasible to the user to see and understand what they have created so far, to eliminate unnecessary documents from clusters, to modify

clusters as needed, and to see the changes applied to clusters visually. The user sees each cluster with its label.

- In the case of using a document to establish a new cluster, the cluster is given a label by the user. The label can be entered manually or terms from the currently displayed document content by double-left clicking the corresponding term to instantly appear in the cluster labeling box. The label can be modified at any time before the end of the supervised stage of the clustering process.
- After creating all clusters based on how many topics users decides to have in their organized collection, the user submits the clusters (with their seed documents) to the underlying clustering algorithm and gets the results back. The results are viewed as:
    a. Bubbles representing clusters labeled according to the user choice as shown in Figure 13.
    b. Glyphs inside the cluster bubbles representing the documents (with document titles).
- The user then can see the entire collection of documents clustered in the bubbles view and organized by topic (cluster label). The current labeling scheme is what the user selects for each cluster. The user can (temporarily or permanently) eliminate any number of documents within a cluster and also any number of clusters for reducing clutter. Table 1 compares the two interfaces.

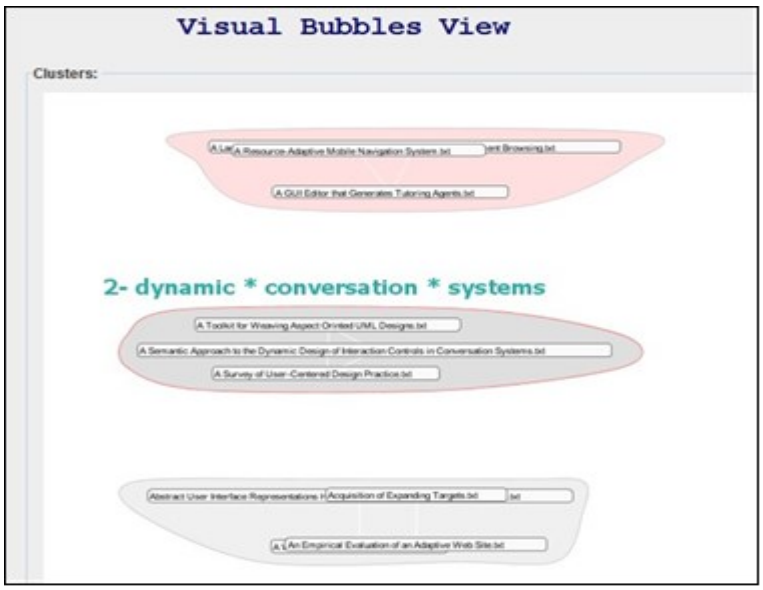


Figure 13. Cluster bubbles and document glyphs.

Table 1. A comparison between the Bubbles and the Pie Interface Features

No.	Features	PI	BI
1.	Document representation	In a circle with a document ID	As a document title with a document index
2.	Mechanism of showing documents	Automatically after classifying the previous document	Using "*Previous* | *Next"* buttons
3.	Permanent document content view(s)	Plain text cloud + whole content	Abstract only + colorful text cloud + full text
4.	Other document content view	None	PDF format in a new window
5.	Creating clusters mechanism	Drag & drop a document into the "*New Cluster*" sector to create a new sector (cluster) containing a yellow stripe (document)	Click the "*New"* button. The new cluster label will be added to the Clusters List. A new bubble (cluster) with a glyph (document) will appear in the Visual Bubbles View
6.	The case of creating a cluster without a label	Although it is incorrect, the interface allows users to do so.	An error massage will pop up asking the user to create a label first.
7.	Visual view of clusters	a) Pie chart presentation b) Stripes (documents) within a sector (cluster) c) Not zoom-able	d) Visual bubbles presentation e) Glyphs (documents) within a bubble (cluster) f) Zooming in and out, and moving the bubbles around
8.	Viewing one cluster at a time	Not applicable	Allowed
9.	Skipping document(s)	Users are allowed to do one of two things to a document. 1) Assign it to a cluster. 2) Send it to the "*Trash"* sector (will not be considered in the clustering phase).	Allowed by hitting the "*Next"* button

4.3. EVALUATION

A user study was conducted to evaluate the Bubbles Interface by comparing it to the Pie Interface. The study was meant to measure the effectiveness, efficiency, and enjoyment of each interface and compare particular findings in the results.

4.3.1. Study Population and Demographics

Ten participants took part in this study. Even though the advertisement was to students from both the computer science and the marine biology departments, only computer science students responded. Of the ten participants in the study, eight were males and only two were females. The ages of two participants were between 18 and 22. The ages of the remaining eight participants were between 23 and 30. All participants were graduate students.

4.3.2. Study Design, Location, and Procedure

The design of the study was complete factorial and within-subjects. The design involved two processes and one task (2x1). A complete factorial design means that the two independent variables (process and task) and their levels (Bubbles Interface and Pie Interface for the process, and the one task used for the task variable) are involved. In the within-subjects design, all participants evaluated the two interfaces using the one task given in the study. The study used a counterbalanced design to ensure controlling for order effect. In other words, both interfaces had the same chance of being used first.

The study was conducted over the duration of two days in October of 2012. The study was conducted in groups and took two sittings. The study took place in the usability laboratory at the Faculty of Computer Science, Dalhousie University. The consent form signed by every participant in the study is shown in Appendix A. The study procedure is shown in Table 2. A collection of 30 documents was used in the study for organization by the participants on each interface. These 30 documents were randomly selected from the data set used in the work of Hu et al. (2012).

The number of documents in this dataset was determined based on: 1) the documents will be given to the participants two days ahead of the actual study to explore; 2) the evaluation of the final results of clusters was human-performed. Each participant had to

check each cluster by ensuring whether or not each documents was in the correct cluster from the user's perspective. Using a larger collection would have been very time consuming for a study that had a limited fund. The limited number of documents does not allow for generalizing the results of the research to all kinds of documents and all sizes of personal collections of documents. The research selected academic articles because it was the choice in the work of Hu et al. (2012) to which the Bubbles Interface is compared in this research. In addition, the participants are familiar with this type of documents. The titles of the documents used during the study are listed in Appendix D. The actual documents can be accessed at: http://web.cs.dal.ca/~badesh/Study/StudyPDFs.zip.

Table 2 The procedure of the user study that was conducted to evaluate the Bubbles and Pie interfaces started two days before the Actual evaluation. This was done to give the participants time to review the study dataset.

Time (min.)	**Procedure**	
5	Study Consent Form	
60	Study 30-Document Collection	
Two days later		
10	Background Questionnaire	
Stage 1		
	Group 1 (Bubbles Interface)	Group 2 (Pie Interface)
5	Training session on how to use the BI	Training session on how to use the PI
15	Study Task	Study Task
10	Clusters Evaluation	Clusters Evaluation
10	BI Post-testing Questionnaire	PI Post-testing Questionnaire
Stage 2		
	Group 1 (Pie Interface)	Group 2 (Bubbles Interface)
5	Short tutorial on how to use the PI	Short tutorial on how to use the BI
15	Study Task	Study Task
10	Clusters Evaluation	Clusters Evaluation
10	PI Post-testing Questionnaire	BI Post-testing Questionnaire
30	Focus group discussion (All participants)	

4.3.3. Study Methodology

The experimental design accounted for the possible effects of *order* using two conditions in a *within-subject* design. The possible main effect of the independent variable (the interfaces) was controlled by randomly selecting with which interface the participant started. A background questionnaire was used to gather demographic data about the participants. It was also used to collect information about the size of the participants' personal collections of documents and any tools they use to organize their documents. The questionnaire is provided in Appendix C.

Every participant was given 30 documents from which they could select 12 documents as seeds to clusters (1–12 clusters). They were given 15 minutes to classify the 12 documents into initial clusters. This was the supervision stage. The ten participants were split into two teams (Team1 consisted of four participants while Team 2 consisted of six participants). The two teams met on two different days. On the first day of the study, the team met and the evaluation was completed as follows:

1- The team was divided into two groups (Group 1 and Group 2).
2- Each group was given a training session (approximately 5 minutes) on how to work on each interface.
3- Group 1 started working on the Bubbles Interface while Group 2 started working on the Pie Interface.
4- The participants were given the 30 documents used in the study two days ahead to familiarize themselves with the collection.
5- Every participant was asked to classify 12 documents from the collection of 30 documents into any number of clusters (1-12 clusters). The description of the study task can be seen in Appendix B.
6- After completing the classification process, the interface called the underlying clustering algorithm used in the work of Hu et al. (2012), and the remaining 18 documents were assigned by the algorithm to finish the clustering stage.
7- Every participant was asked to evaluate the clustering process by deciding whether or not each of the documents was assigned to the correct cluster from the participant's point of view.

8- Every participant was asked to complete a post-testing questionnaire about the interface they used. The questionnaires are shown in Appendix C.

9- The groups were then switched to follow the same steps 5 through 8 as described above.

10- A focus group discussion took place after complete the task on both interfaces. The focus group discussion guide can be found in Appendix E.

The study was meant to evaluate the effectiveness, efficiency, and enjoyment of each interface and compare the two interfaces for measuring possible improvements in the Bubbles Interface over the Pie Interface. The efficiency is measured by recording the time spent by every participant during the study and the number of mouse clicks needed to complete the study. The perceived effectiveness and the engagement of the interface are measured through the data accumulated in the questionnaires and the accuracy of the clustering process.

The design of the experiment in the work of Hu et al. (2012) influenced the design of the current experiment in many ways. First, the document collection used in both studies is the same. Second, the current study gave the users the documents in advance since they complained a lot about the time they took to familiarize themselves with the documents in the study of Hu et al. (2012). The design of the Bubbles Interface attempted to change the visualization used in the Pie Interface and provide more interaction and display of content features as seen in Table 1.

4.3.4. Study Data

The data collected in the study came from: the generic pre-study questionnaire, the post-task questionnaire for each interface, and the focus group discussion. Logs were also used to count the number of mouse clicks needed by each participant to complete the study. All questionnaires used Likert-scale questions with room for additional written comments.

1. *The Background Questionnaire (Appendix C.1.)*: An online demographic questionnaire was administered at the beginning of the study. General demographic questions that concerned: the age of the user, their gender, the time users spend on organizing their own document collections, and the user

experience with document organization tools were asked. Participants had the right to provide no answers for the questions asked in the questionnaire.

2. *The Post-task Questionnaires (Appendices C.2. & C.3.):* Two identical online post-task questionnaires were administered. Questions pertaining to usability, visualization features, and interface layout were used in both questionnaires.
3. *The Focus Group Discussion (Appendix E):* After evaluating the two interfaces, all participants met for a group discussion. The discussion was led by the researcher who illustrated the purpose of the meeting. Then, the researcher re-asked the questions from the study questionnaire for obtaining more details from the participants and to perform qualitative analysis.

4.3.5. Study Results

This section reports the results of the user study. The results are of three kinds. First, answers to the questions on the background questionnaire are discussed. Second, the study results coming from the user responses to the post-task questionnaires are analyzed. The last part of the results concerns the efficiency results while comparing the two interfaces evaluated in the study.

4.3.3.1. Background Questionnaire

The questionnaires asked about the sizes of the collections of documents with which users deal in usual and how long it would take them to organize their personal collections. Of the participants, 20% (2/10) indicated that the size of their personal collections was under 100 documents. Sixty percent (6/10) of the participants have collections of sizes between 100 and 500 documents. The remaining 20% (2/10) of the participants have document collections of sizes over 500 documents. Eighty percent of the participants indicated that organizing and classifying their collections of documents would take between one and two hours. The remaining 20% (2/10) of the participants indicated taking three to four hours to organize their collections of documents.

All ten participants reported using their own computers to organize their document collections. The majority of participants (7/10) indicated that they are usually very frustrated with organization activities of huge collections of documents on their machines. Some additional applications and tools they use to store their documents

included: Google Docs[17] (which is now Google Drive[18]), Skydrive[19], and Dropbox[20]. Most of the participants spend between one and two hours to organize their data on their local machines. The participants reported the following issues and difficulties they have with organizing their documents:

- Time consuming. Individually classifying large collections of electronic documents is overwhelming and takes too long with unsatisfactory results.
- Getting lost with subfolders. The hierarchical structure of folders hides documents and confuses the searcher. Users in some cases, re-search the web for documents they know that they have previously downloaded on their machines as they indicated.
- Other reasons include forgetting the name of the document and the ineffectiveness of desktop search.

Of the participants, 80% indicated that they would like to have effective organization tools for their personal collections of digital documents. Only two participants stated that they have used tools such as Google Desktop and Mendeley[21]. They had to get rid of these tools eventually because they were ineffective and time consuming as those two participants indicated.

4.3.3.2. Post-Task Questionnaires

The study used a post-task questionnaire for each interface after the user completed the task. Each questionnaire had 16 Likert-scale questions that measured engagement factors considered in the study. The questions used involved the option of 'other' in most cases so that the user could provide different answers from the choices given. Both questionnaires had the same questions except for two cases (questions 11 and 12) that depended on the interface being evaluated. In all of the questions that used Likert-scales, the neutral case (i.e. the answer of 'not sure') was ignored from the analysis. The first and second choices of the 5-point Likert-scale were merged and considered as one choice. The same procedure was followed with the fourth and fifth choices.

[17] https://docs.google.com
[18] Drive.google.com
[19] http://windows.microsoft.com/skydrive/download
[20] https://www.dropbox.com/
[21] http://www.mendeley.com/

The data was evaluated using the *z-test* (Downy et al., 2004) because: 1) one variation of the z-test can be used for comparing two means (equivalent to *Student t-test*). 2) Another variation of the z-test can be used for comparing two proportions (equivalent to *Chi Square*). The study questionnaires are shown in Appendix C. Additional comments added by the users to the answers of the questions are listed categorized along with their corresponding questions in Appendix I. The following discussion goes through the results in each individual case measuring the engagement of the interfaces.

1. How easy was the selection of documents for each cluster?

Nine participants chose 'easy' and 'very easy' for the Bubbles Interface, while only three participants found the Pie Interface to be 'easy' with regard to selecting documents for each cluster. The difference between the two proportions of participants (9/10 and 3/10) was significant ($z = 2.739$, $p < 0.007$). Figure 14 shows the comparison of the two interfaces with respect to how easy it was to select documents for clusters.

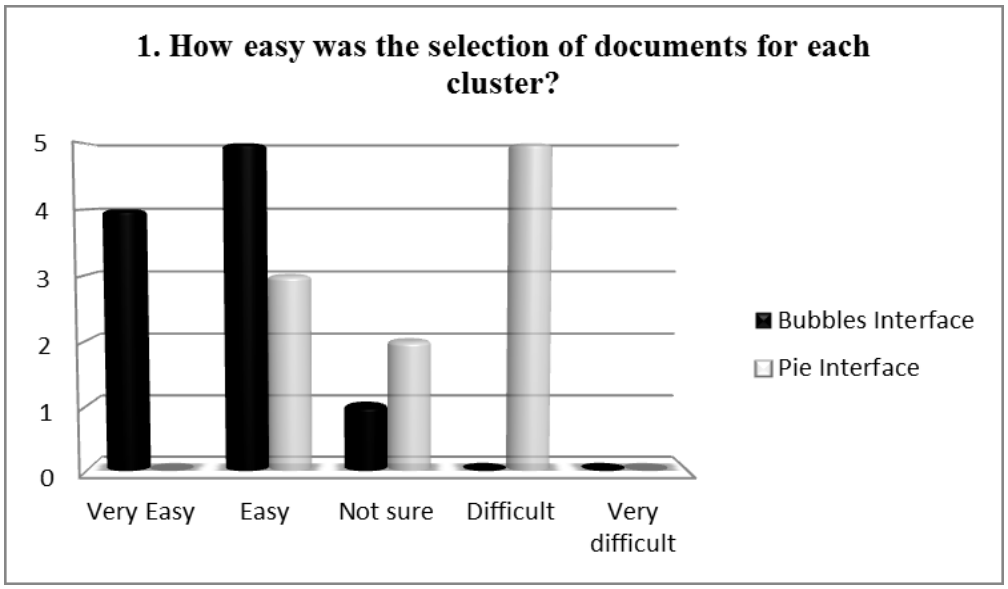


Figure 14. The documents selection process was considered to be easier with the Bubbles interface than with the Pie interface.

2. How effective (helpful & useful) did you find creating labels for a new cluster?

On the Bubbles Interface, eight participants (8/10) indicated that creating cluster labels was 'effective'. The remaining two participants selected the neutral choice 'not sure' on the Likert-scale. On the Pie Interface, five participants chose 'effective' while three participants selected the 'not effective' choice on the scale. The difference between the

two proportions of participants who considered the labeling feature as effective (8/10 and 5/10) was not significant ($z = 1.41$, $p = 0.16$). Figure 15 shows the participants' responses on the effectiveness of assigning labels to new clusters.

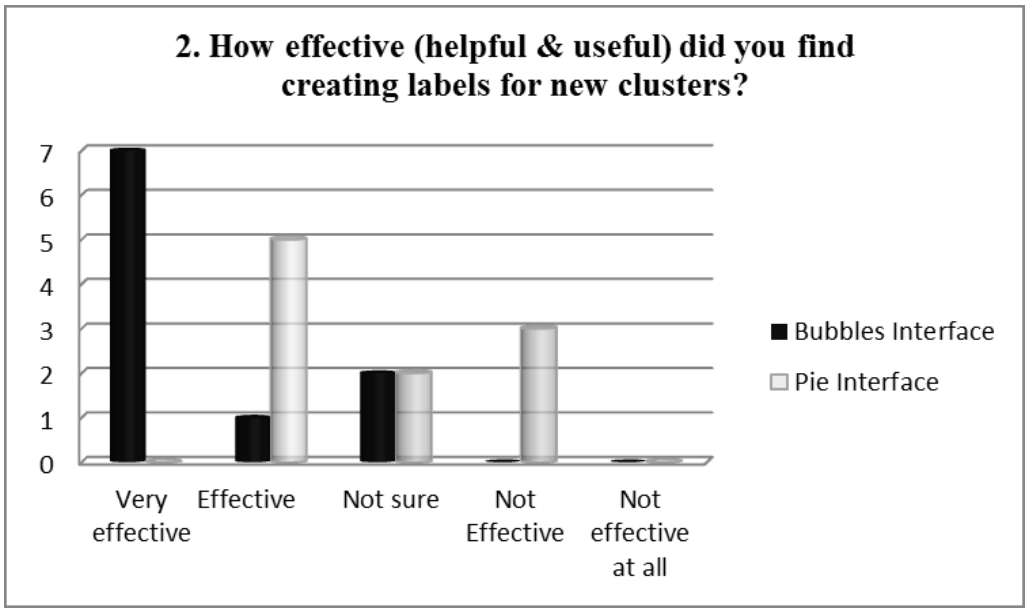


Figure 15. The labeling feature was considered equally helpful on both interfaces.

3. How easy was modifying a cluster to add or remove documents?

On the Bubbles Interface, 70% of the participants (7/10) found it to be easy to modify clusters created during the supervision stage. Two participants indicated that it was difficult to modify the clusters while the remaining one selected the neutral choice 'not sure'. On the Pie Interface, eight participants (8/10) found modifying clusters to be easy. One participant found it to be difficult while the remaining one was 'not sure'. The difference between the proportions of participants who found it easy to modify clusters on the Bubbles Interface and those who found easy to modify clusters on the Pie Interface was not significant ($z = -0.52$, $p = 0.60$). The ease of using the feature of modifying clusters results are shown in Figure 16.

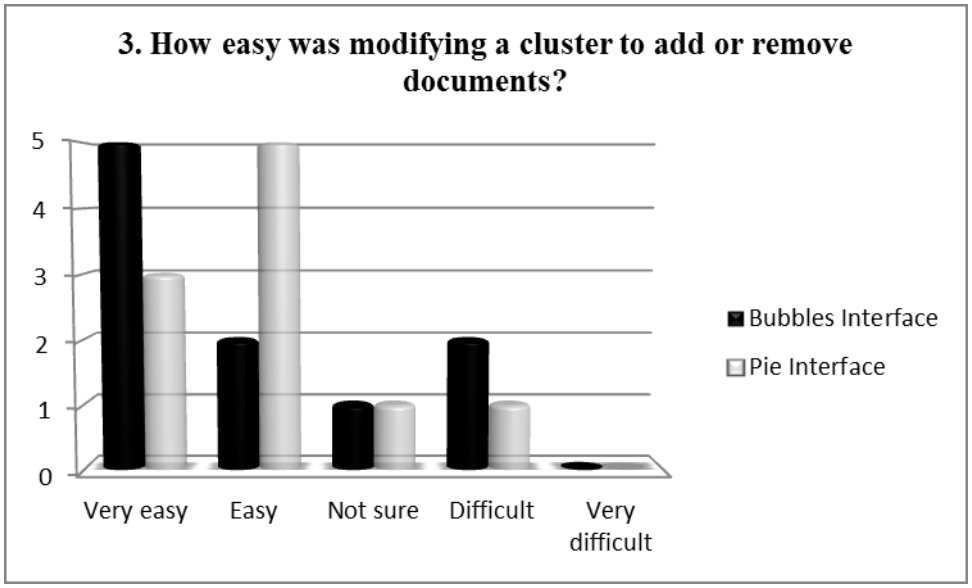


Figure 16. The modify-clusters feature was considered equally easy on both interfaces.

4. How clear did you find the view of your selected documents into the initial clusters?

On the Bubbles Interface and during the supervision stage, six participants (6/10) liked the clear presentation of their initial clusters. Two participants indicated that it was not clear while the rest selected the neutral choice 'not sure'. During the supervision stage on the Pie Interface, five participants liked the clear presentation of their initial clusters. Three participants found it unclear while two participants selected the 'not sure' choice. The difference between the proportion of participants who found the presentation of the initial clusters clear on the Bubbles Interface and those who found the presentation of the initial clusters clear on the Pie Interface was not significant ($z = 0.45$, $p = 0.56$). Figure 17 shows the results regarding the clarity of the participants of the initial clusters.

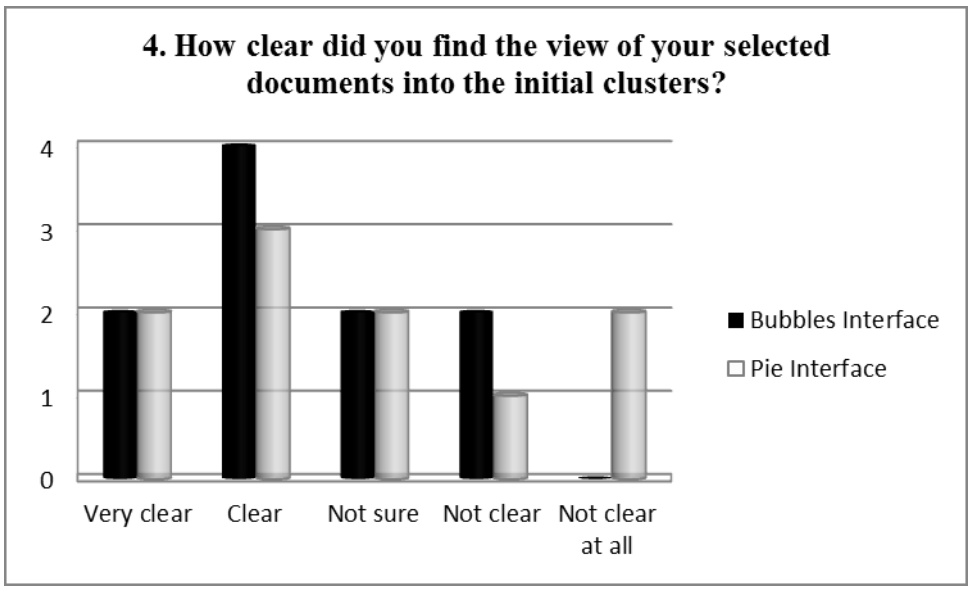


Figure 17. The presentation of the initial clusters was equally clear on both interfaces.

5. How helpful and effective did you find the final view of the clusters created by the system?

On the Bubbles Interface, four participants (4/10) found the final presentation of the clusters to be helpful and effective. Three participants (3/10) indicated that it was not helpful or effective because of the overlapping of the documents names while the three remaining participants (3/10) selected the neutral choice 'not sure'. On the Pie Interface, four participants (4/10) found the final presentation of the clusters to be helpful and effective. Four participants (4/10) considered it not helpful or effective while two participants (2/10) were 'not sure' (see Figure 18). The difference between the proportions of participants who found the final presentation of the clusters helpful and effective on the Bubbles Interface and those who found the final presentation of the clusters helpful and effective on the Pie Interface was not significant ($\mathbf{z} = 0$, $p = 0.99$).

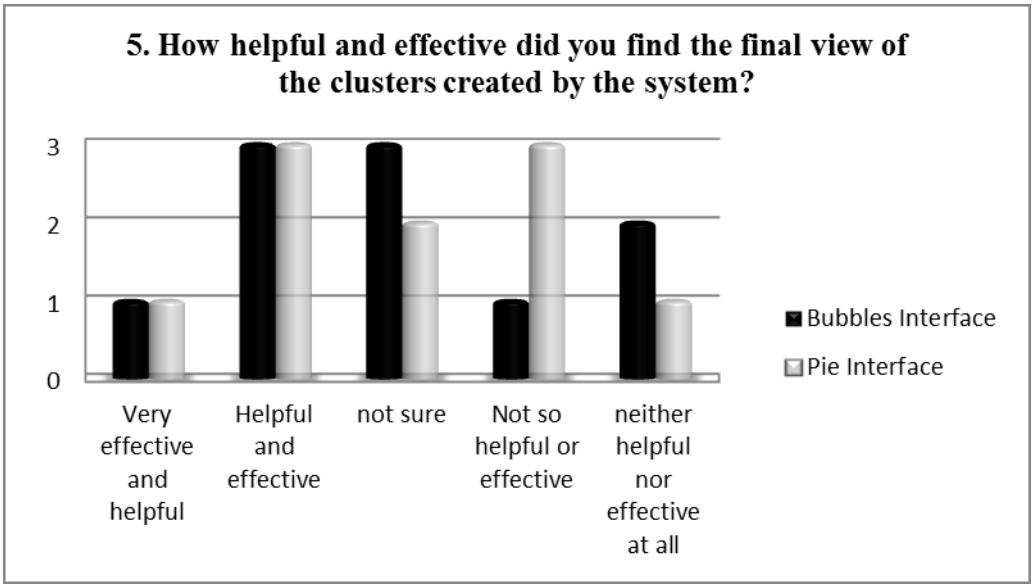


Figure 18. The presentation of the final clusters that were created by the system was helpful and effective on both interfaces.

6. How do you rate the presentation of elements on the interface?

All participants (10/10) rated the elements on the Bubbles Interface as effective and commented that the layout was intuitive and easy to understand. Four participants (4/10) rated the elements on the Pie Interface to be effective while four participants rated those elements as not effective. The remaining participant (1/10) selected the neutral choice 'not sure' (see Figure 19). There was a significant difference between the proportions of participants who found the presentation of the elements on the Bubbles Interface to be effective and those who found the presentation of the elements on the Pie Interface to be effective ($z = 2.93$, $p < 0.003$).

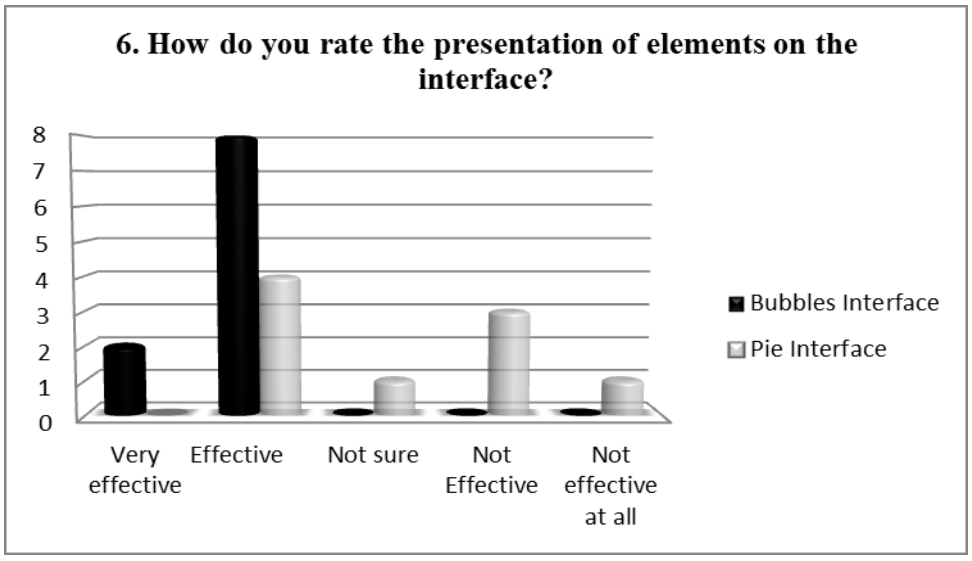


Figure 19. The presentation of elements on the Bubbles interface was highly considered to be easy and intuitive.

7. How do you rate the positioning of the document view and cluster view on the screen?

The positioning of the document view and cluster view on the Bubbles Interface were considered effective by 70% of the participants (7/10). Two participants rated the views as not effective while only one participant selected the 'not sure' choice. On the Pie Interface, the positioning of the document view and cluster view were considered as effective by three participants (3/10). Four participants (4/10) rated the view as not effective and the remaining three participants (3/10) selected the 'not sure' choice. There was a significant difference between the proportions of participants who rated the positioning of the document view and cluster view on the Bubbles Interface as effective and those who rated the positioning of the document view and cluster view on the Pie Interface as effective. Figure 20 depicts the comparison ($z = 2.25$, $p < 0.02$).

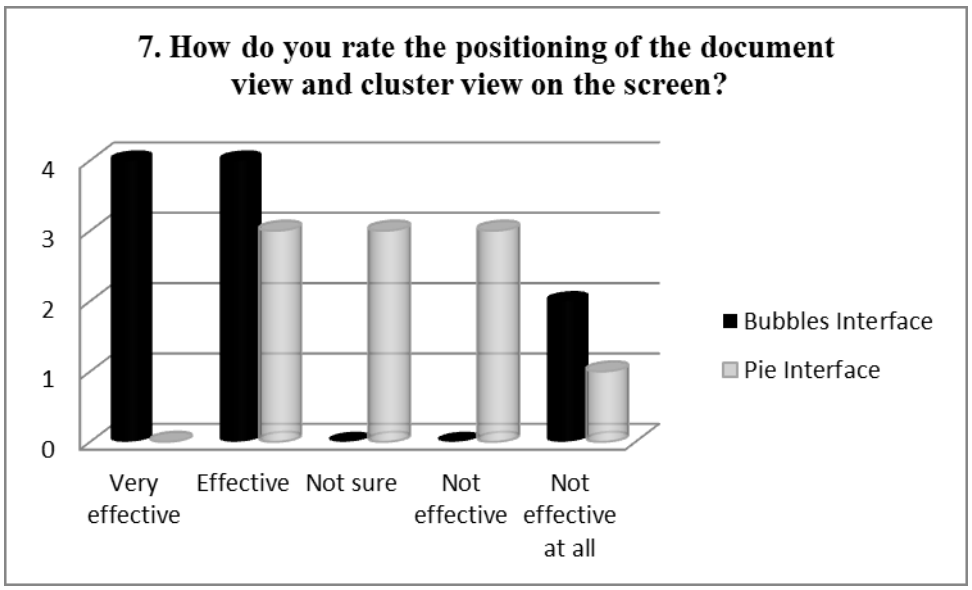


Figure 20. The positioning of the document and cluster views was rated as effective with regard to helpfulness in reviewing the documents' content and the labels of the clusters.

8. How easy was it to undo actions on the interface?

On the Bubbles Interface, eight participants (8/10) rated the ability to reverse actions as easy. One of the remaining two participants rated it as difficult and the other one selected the 'not sure' choice. On the Pie Interface, three participants (3/10) rated the ability to reverse actions as easy while three other participants (3/10) rated it as difficult. The remaining four participants (4/10) selected the neutral choice of 'not sure' (see Figure 21). The difference between the two proportions of participants who found reversing actions to be easy on both interfaces was significant ($z = 2.25$, $p < 0.02$).

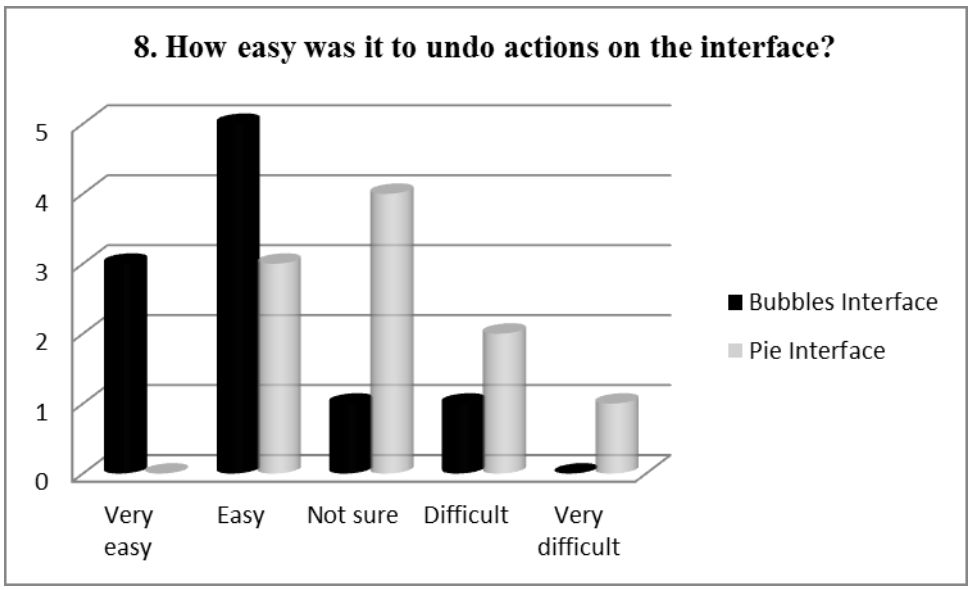


Figure 21. Reversing actions on the Bubbles interface was reported as easy and effortless.

9. Was the feedback from the interface helpful to you?

The feedback from the Bubbles Interface was considered as clear and helpful by eight participants (8/10), not clear or helpful by one participant (1/10), and not applicable by one participant (1/10). The feedback from the Pie Interface was considered as clear and helpful by three participants (3/10), not clear or helpful by two participants (2/10), and not applicable by one participant (5/10) (see Figure 22).There was a significant difference between the proportions of participants who found the feedback from the Bubbles Interface as clear and helpful and those who found the feedback from the Pie Interface as clear and helpful ($z = 2.25$, $p < 0.02$).

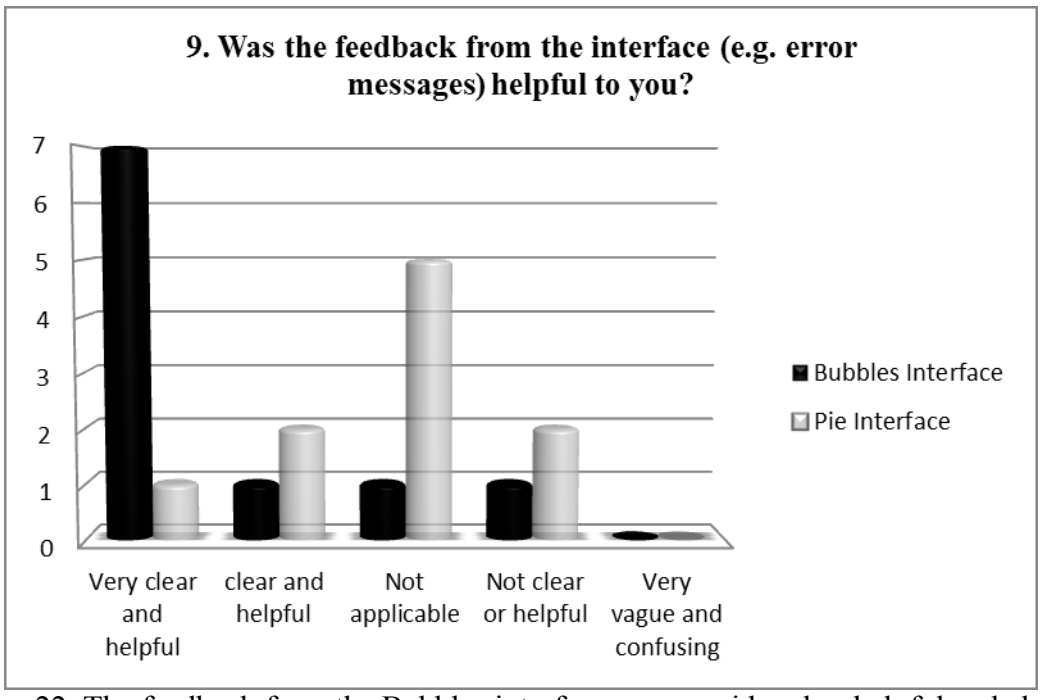


Figure 22. The feedback from the Bubbles interface was considered as helpful and also guiding compared to the Pie interface which provided almost no feedback whatsoever.

10. How helpful and effective do you think the interface will be with organizing your collection of documents?

Of the participants, seven users (7/10) considered the Bubbles Interface to be helpful and effective with organizing their own collections of documents. Two participants (2/10) considered it as neither helpful nor effective. Two participants considered the Pie Interface as helpful and effective with organizing their own collections of documents. Four participants (4/10) considered it as neither helpful nor effective (see Figure 23). There was a significant difference between the proportions of participants who considered the Bubbles Interface to be helpful and effective for organizing their own collections of documents and those who considered the Pie Interface to be helpful and effective for organizing their own collections of documents ($z = 2.24$, $p < 0.02$).

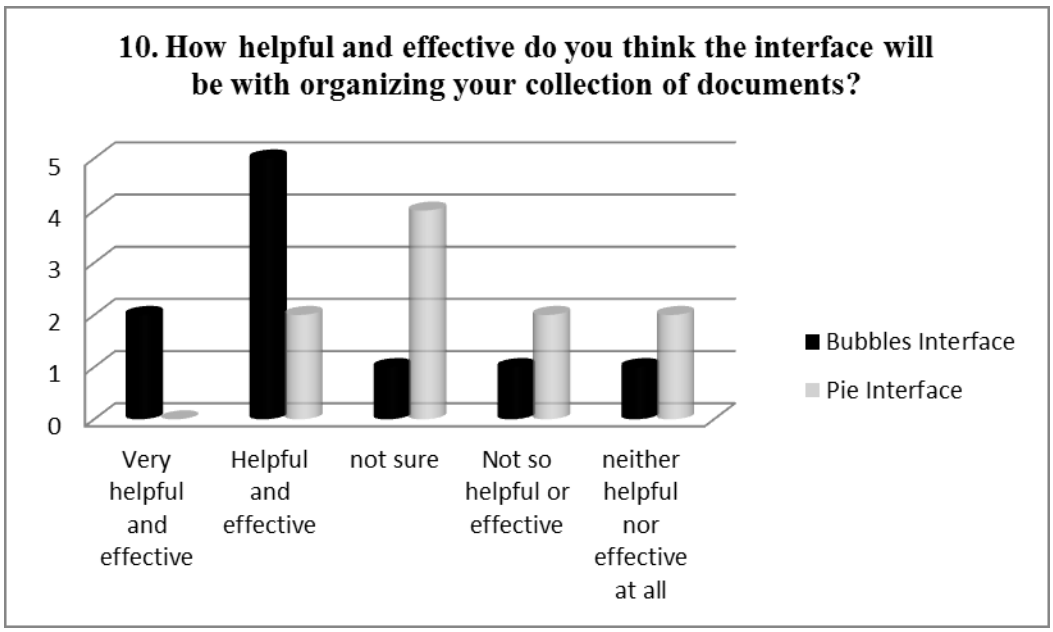


Figure 23. The future helpfulness and effectiveness of the interfaces as perceived by the participants who highly preferred to use Bubbles interface to organize their own document collections.

11. How effective (useful & helpful) was each of the following interaction features:

A. Bubbles Interface

Participants were asked about their satisfaction with the following interaction features in the Bubbles Interface. The results of the user responses are shown in Figure 24.

1) Immediately showing any added clusters in the clusters list.
2) Instantly showing new clusters in the visualized (bubbles) view of clusters.
3) Going back and forth among documents by showing the abstracts and titles immediately.
4) Selecting words (terms) from documents so they get added immediately to the label of the cluster.
5) Choosing a cluster from the clusters list to add the current document to that cluster.
6) Continuously modifying clusters you created initially.
7) Seeing your cluster label when you hover over the bubble that represents the cluster.
8) Deleting any document or cluster from the bubbles view.
9) Zooming in and out in the bubbles view.

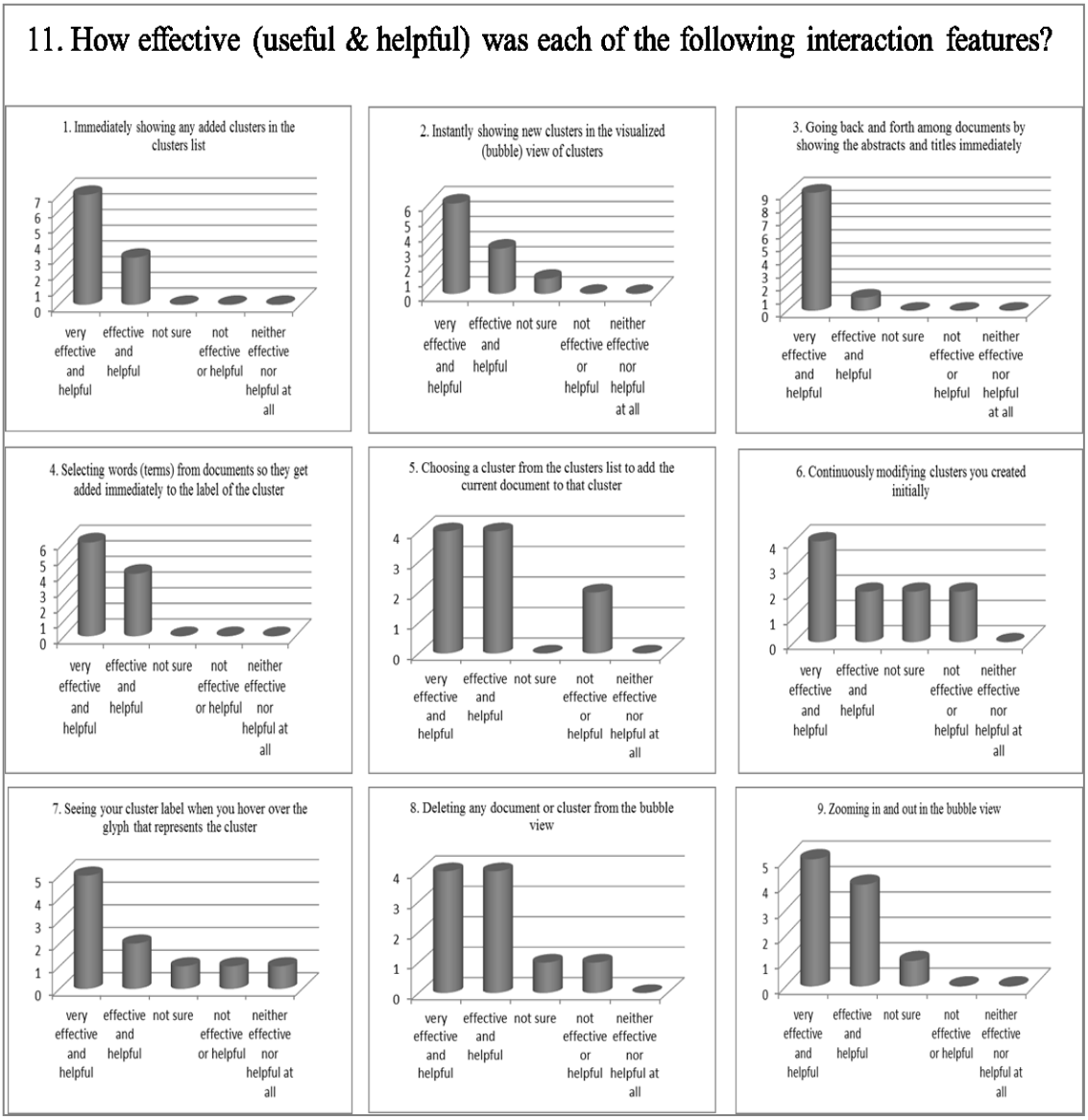


Figure 24. Users' evaluation of interaction features in the Bubbles interface.

All participants (10/10) considered the immediate showing of any added clusters in the clusters list as an effective feature. Nine participants (9/10) considered instantly showing new clusters in the visualized view of clusters (bubbles) as an effective feature. Only one participant selected the neutral choice 'not sure'. All participants (10/10) considered going back and forth among documents and immediately showing the abstracts and titles of these documents as an effective and helpful interaction feature. Selecting words (terms) from documents so they get added immediately to the cluster's label was rated as effective and helpful by all the participants (10/10). Eight participants (8/10) considered

the feature of adding a document to a previously created cluster effective while two participants (2/10) considered this feature ineffective.

The ability to modify clusters that have already been created was rated as an effective feature by six participants (6/10). Two participants (2/10) rated this feature as not effective. The ability to see a cluster label while hovering over the bubble that represents the cluster was rated as effective and helpful by seven participants (7/10). Only two participants (2/10) rated this feature as not effective. Eight participants (8/10) rated the ability to delete any document or cluster from the bubbles view as effective while only one participant (1/10) rated it as ineffective. The zooming in and out in the bubbles view was preferred and rated as effective and helpful by most of the participants (9/10) whereas one participant (1/10) found it to be ineffective.

B. Pie Interface

Participants were asked about their satisfaction with the following interaction features in the Pie Interface. The results of the user responses are shown in Figure 25.

1) Dragging a document/cluster for either creating a new cluster or modifying it.
2) Coloring any added clusters in the pie chart view.
3) Showing new clusters with a certain color and ID in the pie chart view.
4) Showing a new document with an ID on the pie chart view, and the document information (e.g. title and content) in the labeled-document view.
5) The ability to select words (terms) from documents so they get added immediately to the label of the cluster
6) Continuously modifying clusters you create initially.
7) Viewing a text cloud for all the documents in a cluster on the content-of-cluster panel.
8) Deleting any document or cluster by dragging them to the trash.

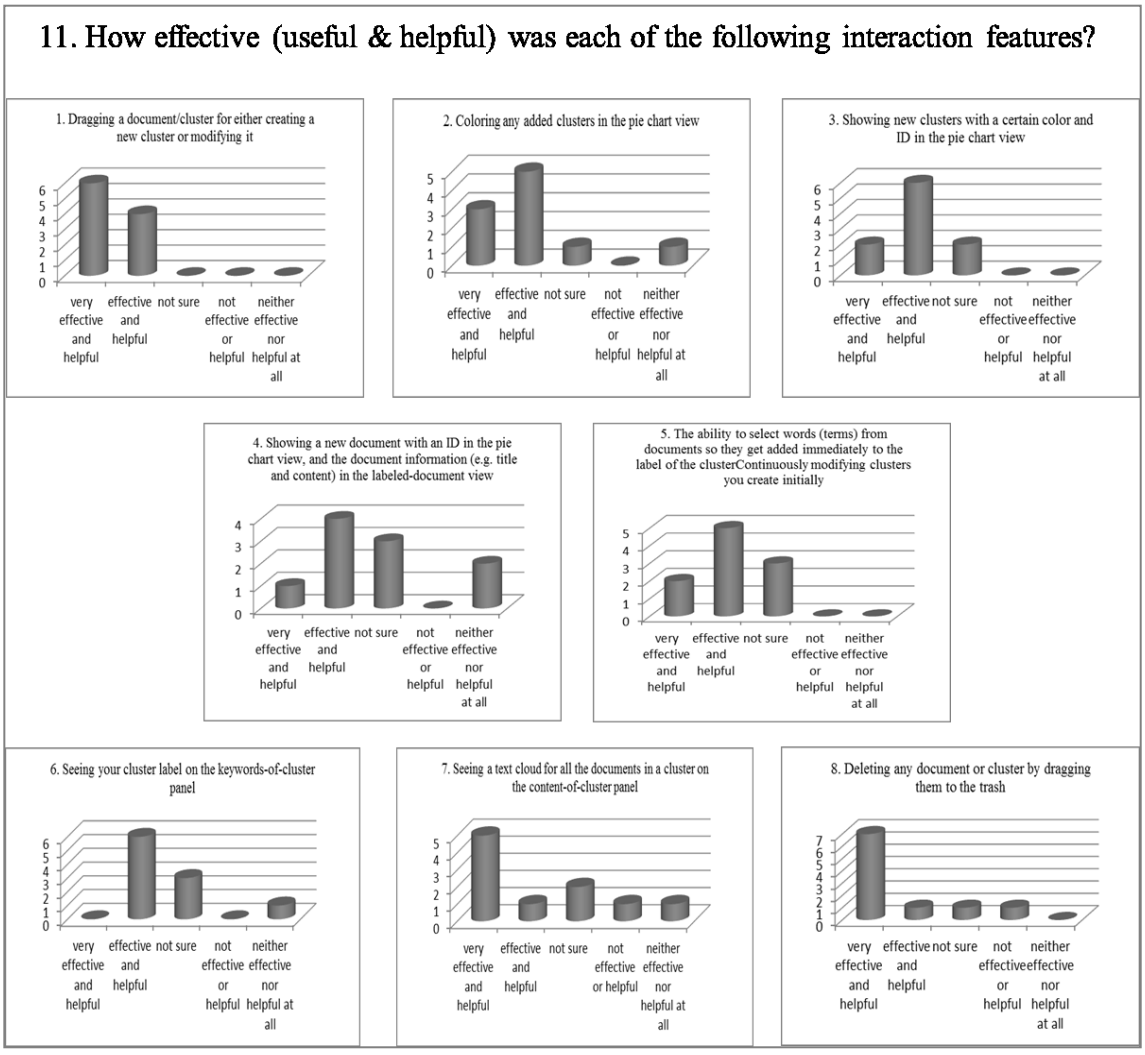


Figure 25. Users' evaluation of interaction features in the Pie interface.

12. How effective (useful & helpful) was each of the following visualization features?

A. Bubbles Interface

The participants were asked about the usefulness and helpfulness of the following visualization features in the Bubbles Interface. Their responses are depicted in Figure 26.

1) The display of the text cloud of the document.
2) The display of the title of the document.
3) The list of clusters displayed to you under the labeling box.
4) The display of the abstract of the document.
5) The visualization of the clusters in the bubbles view while creating the initial set of clusters.
6) The visualization of the final set of clusters.

7) The layout of the items on the screen.
8) The display of the full text of the document.
9) The use of different colors with the text cloud presentation of the document content.

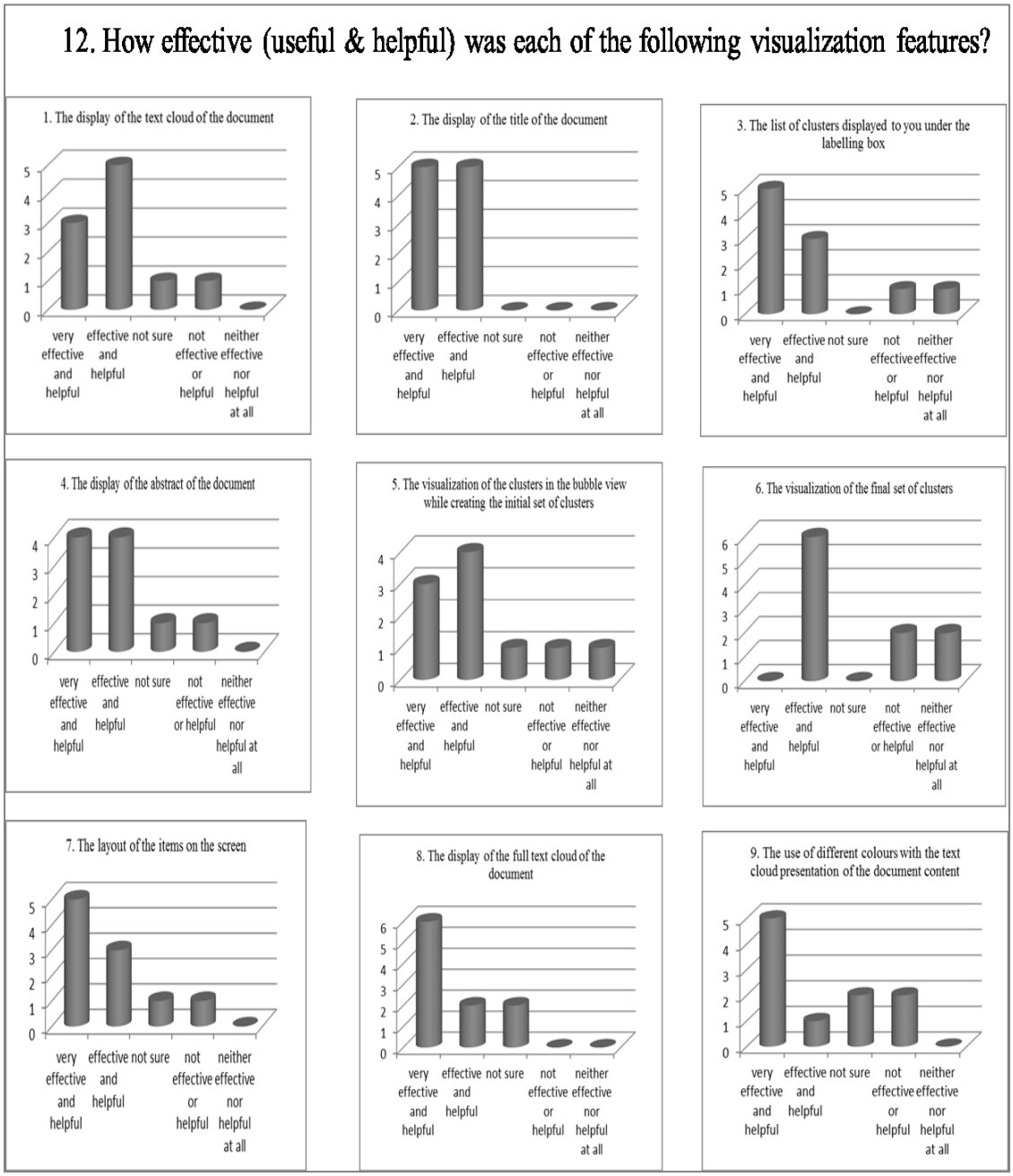


Figure 26. Users' evaluation of visualization features in the Bubbles interface.

Eight participants (8/10) considered the display of the text cloud of the document an effective feature while one participant found it ineffective. All participants (10/10)

considered the display of the title of the document as effective and helpful. Eight participants (8/10) found the list of clusters (displayed under the labeling box) as effective and it provided an overview of all clusters created. Eight participants (8/10) preferred the abstract view of the documents while one participant found it least effective.

Seven participants (7/10) found the visualization of the clusters in the bubbles view while creating the initial set of clusters to be an effective feature. The remaining two participants found it to be ineffective. Six Participants (6/10) found the visualization of the final set of clusters as effective while four participants found it ineffective. The layout of the items on the interface was rated as effective by most participants (8/10) and only one participant found it ineffective. Eight participants (8/10) liked the full text view of document content and found it effective. Six participants (6/10) found the use of different colors with the text cloud presentation of the document content as effective while two participants found it ineffective.

During the group discussion, the participants were asked about the effectiveness and helpfulness of the three views of the document content (abstract, text cloud, and full text). Six participants indicated that the abstract view was effective. Seven participants stated that the text cloud view on the Bubbles Interface was effective while only one user liked the full text view. However, the results are not reflective of the entire population of users because of the small number of participants involved in the study. In addition, the users were only reading the articles to classify them which may have implied time restrictions that resulted in favoring the shorter views.

B. Pie Interface

The participants in the study were also asked to evaluate the usefulness and helpfulness of the following visualization-related features on the Pie interface. Their responses are shown in Figure 27.

1) The display of the text cloud of the document content.
2) The presentation of documents within a cluster as stripes.
3) The creation of the label of the cluster.
4) The display of the full text of the document content.

5) The visualization of the clusters in the pie chart view when you create the initial set of clusters.
6) The display of the final set of clusters.
7) The layout of the items on the screen.

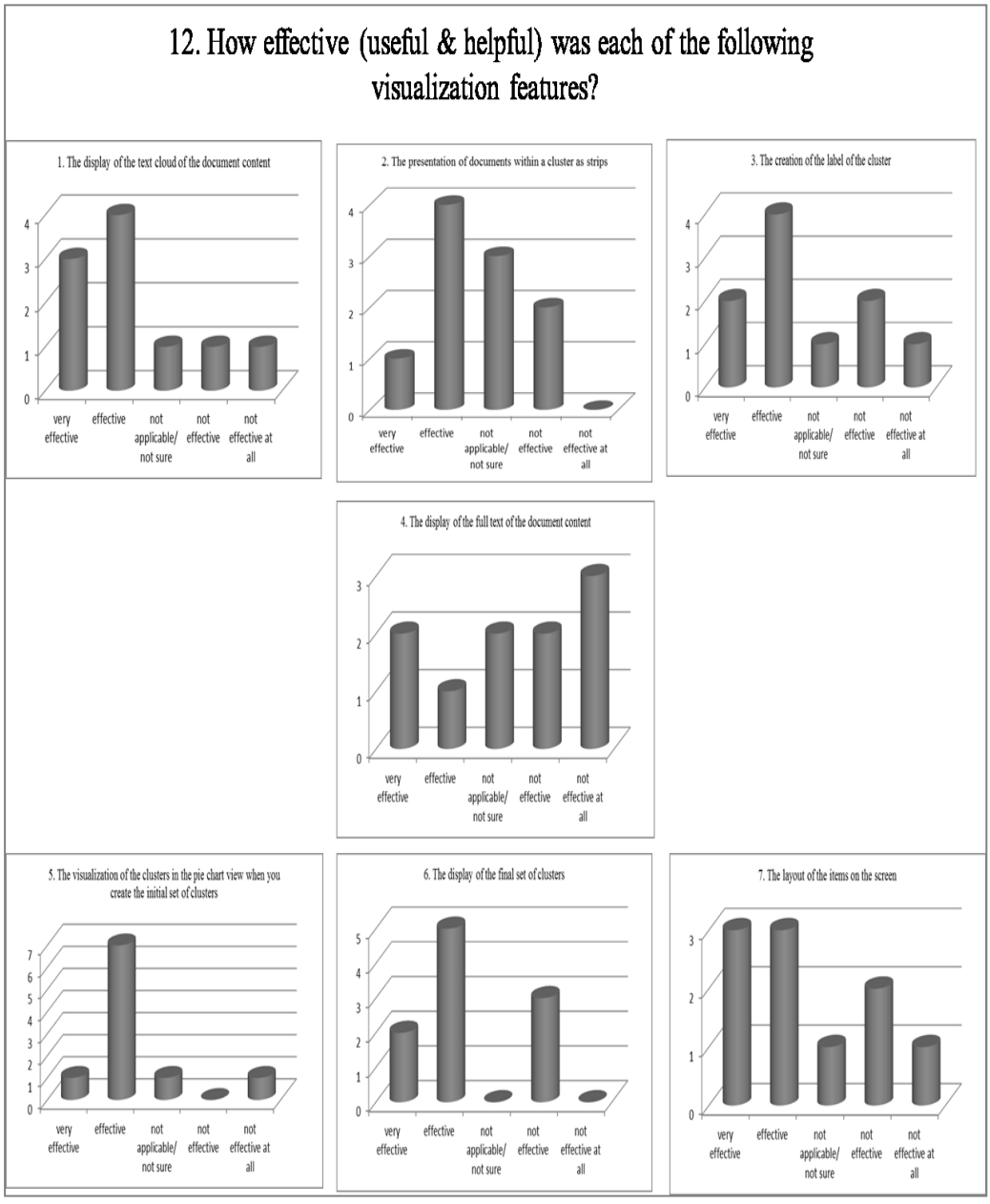


Figure 27. Users' evaluations of visualization features in the Pie interface.

Seven participants (7/10) found the display of the text cloud of the document content as effective while two participants (2/10) found it ineffective. Only half of the participants (5/10) found the presentation of the documents assigned to clusters as stripes to be

effective. The creation of cluster labels was considered as easy and helpful by six participants (6/10) while three participants considered this process to be difficult. Half of the participants (5/10) considered the display of the full text of the document content as neither helpful not useful. Only three participants (3/10) considered it to be helpful and useful.

Most participants (8/10) considered the visualization of the clusters in the pie chart view while creating the initial set of clusters to be effective, and only one participant found it ineffective. Seven participants (7/10) considered the display of the final set of clusters as effective while three participants (3/10) found it ineffective. Only three participants (3/10) considered the layout of the items on the interface as effective and six participants considered it ineffective. When asked, participants indicated that the Pie Interface layout was very confusing.

13. Did the interface help with the categorization of the documents in the final results?

In the case of the Pie Interface, 40% of the participants (4/10) answered the question with 'Yes'. Two participants favored the easiness of moving documents from one cluster to another. Another participant indicated that the documents categorization of the final results was very helpful. The remaining 60% of the participants (6/10) answered the question with 'No'. Most of the participants stated that having documents numbers (stripe numbers) and clusters' numbers (sector number) were not helpful at all. Users indicated that those numbers were confusing and frustrating most of the time. The documents and clusters information should appear clearly while moussing over the document or the cluster. Moreover, the interface layout was a mess in terms of having many sections and subsections that caused confusion and delay of the supervision stage as indicated by participants.

Most participants (9/10) answered the question with 'Yes' for the Bubbles Interface indicating that the prototype helped them with the categorization of the documents in the final results. Only 10% of the participants (1/10) selected 'No' with no mentioning of any reasons.

Participants who found the Bubbles Interface helpful with categorizing documents provided valuable comments. One participant indicated that the feature of adding

documents to previously created clusters was very helpful. Another participant reported that the feature of merging clusters was very useful. In addition to the merging feature, the abstract view of documents was very effective with respect to selecting labels for clusters. Three participants indicated that the visual presentation of clusters was very helpful and intuitive. One participant commented that the ability to see a PDF view of any document in any visual cluster 'bubble' was very useful. Most participants reported that the visual presentation of clusters provide an effective overview of the documents and their titles within clusters.

There was a significant difference between the proportions of participants who responded with 'Yes' to the question in the case of the Bubbles Interface and those who responded with 'Yes' in the case of the Pie Interface ($z = 2.34$, $p < 0.01$). Figure 28 shows comparisons of the two interfaces.

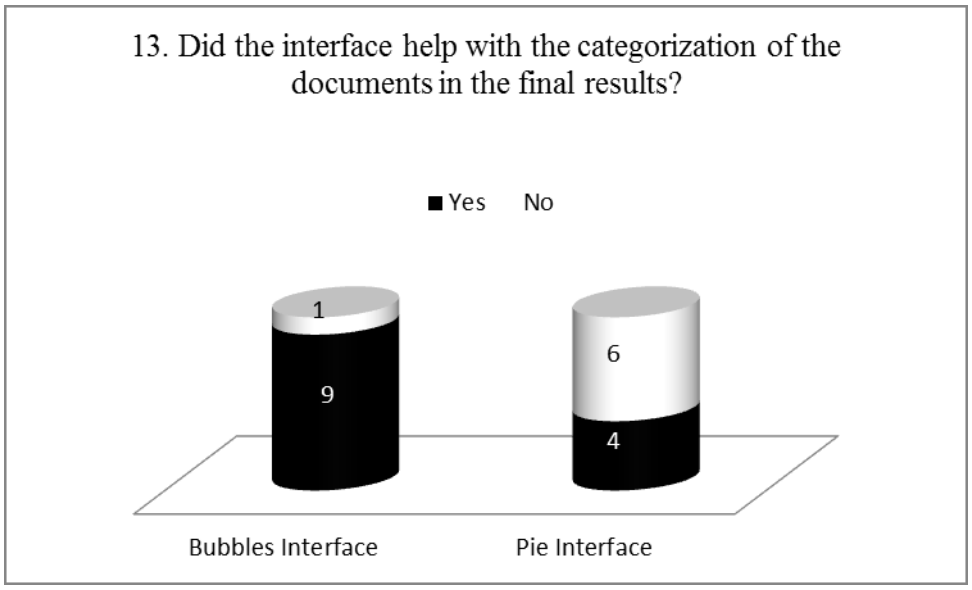


Figure 28. Both interfaces were helpful with the categorization of the documents in the final results, but the Bubbles interface was highly preferred by the participants.

14. Are there any features that should be added to the interface to make it more effective?

Participants provided valuable feedback regarding future optimization for both interfaces. For the Bubbles Interface, participants suggested that the cluster label should appear closer to the cluster while hovering over the bubble representing the cluster. Participants

also recommended that the color of the document's title should change to red as soon as the classifying of that document is completed. A search feature should be added to allow the user to find a document already classified into a cluster. The glyphs that represent the documents and convey documents' titles within a cluster should not be cluttered and should have an organization scheme according to some criteria. For instant, documents should be sorted alphabetically according to their titles within the cluster bubble. When the user selects a cluster from the list, the corresponding bubble should be highlighted in a way that makes it easy for the user to see that cluster bubble. The size of the different views on the interface should be customizable. For example, if the user wants to re-size of the bubbles view area, they should be able to do so.

15. Did you encounter any challenges while working on the interface?

On the Pie Interface, four participants answered with 'Yes' to challenges encountered. Two participants indicated that they could not figure out how to re-label clusters. Another issue was the need to scroll up and down to see the content of the document due to the small size of the viewing area. One participant indicated the inability to see the cluster label and title when he or she hovers over it. The rest of the participants reported that they did not encounter any challenges. In the case of the Bubbles Interface, three participants answered with 'Yes'. They indicated that they encountered one issue which was the inability to resize the visual clusters view. The rest of the participants answered with 'No' to the question indicating that no challenges were encountered.

4.3.3.3. Interface Efficiency

The number of mouse clicks (left, right, and middle) during the study sessions was logged. The number of mouse clicks while using the Pie Interface was 301.3 on average (SD=66.6). In the case of using the Bubbles Interface, the average number of mouse clicks was 208.5 (SD=142.41). A two-sample-for-means z-test showed that no significant difference existed between the number of mouse clicks on the Bubbles Interface and the number of mouse clicks on the Pie Interface (z = -1.86, p = 0.06). However, by removing the outliers in the case of the Bubbles Interface, the difference became significant (z = -6.22, p<0.0001).

4.3.3.4. Interface Effectiveness

To measure the effectiveness of the interfaces and compare the Bubbles Interface to the Pie Interface, every participant was asked to evaluate the accuracy of the final clustering of the 30 documents used. Every participant was asked to determine which documents were assigned to the correct clusters and which documents were assigned to the incorrect cluster based on the cluster topic built by the participant. The results are shown in Table 3. The two-samples-for-means z-test results ($z = -2.93$, $p < 0.003$) indicated that there was a significant difference between the two interfaces with respect to the number of documents accurately clustered as perceived by the participants.

Table 3. The numbers of documents considered by the participants to be accurately clustered.

Participants	Bubbles Interface	Pie Interface
P1	23	28
P2	25	18
P3	21	12
P4	28	15
P5	28	15
P6	25	18
P7	21	18
P8	23	15
P9	27	13
P10	19	28
Mean	**24**	**18**
SD	**2.9**	**5.4**

4.4. DISCUSSION

There are strong indications that the Bubbles Interface is better. Users achieved higher clustering accuracy with the Bubbles Interface than they did with the Pie Interface. One participant indicated that "*navigation among the document content views was much easier with the Bubbles Interface*". The Bubbles Interface may have helped users with assigning the appropriate documents together to represent a topic (cluster). It may have also helped users with identifying the documents in each cluster in the final results. The labeling process on the Bubbles Interface may have also helped with identifying the accurate topic of both the documents during the supervised classification stage and the

clusters during the final presentation stage. One participant stated that he "*did very well in assigning documents into correct clusters with the Bubbles Interface.*"

The data logged during the study showed that users worked more efficiently on the Bubbles Interface than they did on the Pie Interface. This improvement was demonstrated by the number of mouse clicks needed to complete the task on the interfaces. The Bubbles Interface required significantly fewer mouse clicks by the user than the Pie Interface to complete the same task. However, there was no significant difference between the times needed to complete the task on the Bubbles Interface and the times needed by users to complete the same task on the Pie Interface.

Performing more clicks on the Pie Interface can be attributed to the user's need for very frequent scrolling in order to see the document content. This kind of scrolling was not needed as frequently on the Bubbles Interface. The reason for completing the tasks on both interfaces with no significant difference in the time needed can be attributed either to the nature of the task itself or to other factors that were not measured in the study.

There have also been significant differences between the interfaces with respect to factors related to how users organized collections of digital documents. The difference between the number of participants who found the process of selecting documents for clusters to be easy on the Bubbles Interface and those who found it easy on the Pie Interface was significant. This may indicate that the approach that was used to show the document content to the user was more effective on the Bubbles Interface. It may also indicate that users found it easier to perceive the cluster content and see where the new document belongs during the supervised initial classification. The use of visualization to allow the user to view the cluster content at any time may have played a role in the user preference of the interface.

Users also found the presentation of elements on the Bubbles Interface to be more effective than the presentation of elements on the Pie Interface. Users indicated during the post-study discussion that the layout was more effective on the Bubbles Interface. They commented that the layout was intuitive and easy to understand and that no confusion or frustration was caused with the organization of the interface elements. During the group discussion, one participant stated "*I had some difficulties viewing the*

document content both with the text cloud and the whole content view. The area customized for displaying the content was not sufficient. It should be larger on the Pie Interface." Another participant indicated "*I got really lost with the Pie Interface because I always forget how to review the document and cluster content.*"

The positioning of the document view and cluster view on the display was considered effective and helpful by significantly more users of the Bubbles Interface than users of the Pie Interface. The participants indicated during the discussion that the layout of the Bubbles Interface was helpful and provided more focus on the intended view. However, the layout of the Pie Interface was confusing and frustrating as stated by the participants. Reversing actions on the Bubbles Interface was significantly easier than it was on the Pie Interface. The participants reported they "*found the interaction with the Bubbles Interface easier because of the nice layout that was easy to understand.*"

The design of the Bubbles Interface allowed the user to more effectively progress with the supervision stage with the ability to go back and reverse any action performed. Those actions included creating a cluster, adding a document, skipping a document for later classification, and so on. On the Pie Interface, however, users were not allowed to undo some important actions. For example, users were not allowed to skip a document during the supervision stage and go back to choose a cluster for that document later or ignore classifying the document. Users had to delete and eventually exclude the document from the organization process.

The feedback given by the interfaces was different. Significantly more users favored the feedback given by the Bubbles Interface. For example, all the messages given by the Bubbles Interface were clear and given to serve many purposes. However, the only feedback message given to the user while using the Pie Interface was the delete conformation message when the user attempts to delete a cluster or a document. The users stated that the feedback of the Bubbles Interface was more helpful and reduced the need for asking the researcher questions to clarify the reactions of the interface. Two participants stated that they liked "*to have something on the interface indicating how many documents we have already classified and how many documents remain.*"

The Bubbles Interface was favored by users for future use with organizing their collections. The difference was significant between the number of users who considered the Bubbles Interface and those who considered the Pie Interface for future use. The participants thought that the supervision process was easier to perform on the Bubbles Interface, the layout of the Bubbles Interface was less confusing, the interaction with the Bubbles Interface was much easier, and that the Bubbles Interface had more comprehensible representations of the clusters.

Participants preferred the categorization of documents in the final results provided by the Bubbles Interface over the categorization provided by the Pie Interface. They indicated that in the case of the Pie Interface, there was little information about the documents in each cluster. The interaction with the interface to obtain more information about the clusters and the documents was hard. The area dedicated for the clusters and the documents on the Pie Interface was very limited and caused frustration to the users.

With respect to the process of creating labels for clusters, there was a substantial difference between the number of users who considered the Bubbles Interface to be effective and those who considered the Pie Interface to be effective. Even though the difference was not significant, most users (8/10) acknowledged the effectiveness of the Bubbles Interface compared to only half of the users (5/10) in the case of the Pie Interface. Users indicated that in the case of the Bubbles Interface, the creation of labels was in one known place to the user and was less confusing than in the case of the Pie Interface.

The study asked participants about particular interaction features in both interfaces. The two-samples-for-means z-test indicated that there is a substantial difference between the number of participants who considered the interaction features on the Bubbles Interface and the Pie Interface ($z = 1.78$, $p<0.07$). As shown in Figures 16 and 17, users favored the interaction features provided on the Bubbles Interface although the difference was not significant. In the case of the Pie Interface, users indicated that it was hard to perform the classification of the initial clusters and it was hard to interact with the process of choosing documents for clusters. Some participants indicated liking the drag-and-drop feature for selecting documents for clusters. This feature can be readily added to the

Bubbles Interface in future design. Actually, adding documents to the bubbles representing clusters can make it even more enjoyable for users to create the initial clustering because of the visual nature of the bubbles. Users can see the change reflected on the display immediately by having the document assigned as a glyph to the cluster.

Users were also asked to evaluate visualization features in both interfaces. The difference between the number of users who rated the visualization features on the Bubbles Interface as effective and those who rated the visualization features on the Pie Interface as effective was significant ($z = 2.45$, $p<0.01$). For example, all users (10/10) of the Bubbles Interface indicated that the display of the title of the document was effective. In addition, most users (8/10) preferred the text cloud on the Bubbles Interface. In the case of the Pie Interface, users did not like the stripes on the Pie chart and considered it very uninformative.

Even though the Bubbles Interface has promising features in the organization of documents, it may have issues of clutter with very large collections. Different design parameters may need to be adjusted such as the glyph size for the document and the size of the bubble representing the cluster. The quality of clustering of a large collection of documents can be evaluated in the case of using the Bubbles Interface by evaluating the seeds selected for the clusters. It will be almost impossible (very time consuming) to ask users in a laboratory experiment to measure the accuracy of the final results in the case of very large collections of documents. However, the seeds chosen by the users to be given to the underlying clustering algorithms can be evaluated by comparison to a ground truth.

4.5 LIMITATIONS

The study had a very limited number of participants due to fund restrictions. The participants were computer science students who may not reflect all kinds of users, yet they represent early adopters. The number of documents organized in the experiment was also limited. Having larger collections may present other difficulties and issues that were not revealed in the current study. The type of documents was limited to academic articles which may not reflect all kinds of documents in personal collections.

4.6 CONCLUSION

The investigation compared specific features in a prototype interface against a baseline interface from previous research (Hu et al, 2012). The results of the investigation showed that the new prototype interface had a better layout and helped users with: 1) the initial classification of documents into clusters during the supervised stage; 2) the modification of clusters; 3) the cluster labeling process; 4) the presentation of the final set of organized documents; 5) the efficiency of the organization process, and 6) the actual accuracy of the cluster for organization process.

CHAPTER 5 CONCLUSIONS

The research discussed in this dissertation started by exploring the concepts of managing and organizing personal collections of documents, visualization, and clustering. The research built a prototype to explore the use of aspects of visualization and clustering in finding relevant documents in the case of a small collection. The prototype used the layout of data mountains to represent clusters of documents. The evaluation showed that users benefited from the use of the layout in finding relevant documents. Moreover, the design showed possible uses of clustering and visualization features in managing and organizing personal collections of documents, which was investigated in the later study.

The research followed by building an interface that was compared to the work of Hu et al. (2012). That work was intended to evaluate different clustering algorithms. The interface was used to assist the user with the selection of the initial clusters during the supervised phase of the clustering. The final results of organizing the documents into clusters was shown to the user too. The interface built for the current research was built based on recommendations from the study in Badesh and Blustein (2012) and was compared to the interface used in the work of Hu et al. (2012).

The evaluation was intended to compare the interfaces with regard to the effectiveness in organizing collections of personal desktop documents into clusters. Moreover, the efficiency of performing the organization in terms of the selection of cluster seeds during the supervision stage of the clustering process was measured and the interfaces were compared. Finally, different engagement factors were considered in the comparison taking into account the users' perspective.

The results of the final study in this research showed that the Bubbles Interface had several advantages over the Pie Interface used in the work of Hu et al. (2012). The Bubbles Interface incorporated several features in the design and implementation that helped the user achieve the task of organizing and managing the given documents more effectively, efficiently, and with more engagement. The Bubbles Interface was more intuitive, a feature that helped with several aspects of the organization process and resulted in more efficiency and effectiveness in the process of managing and organizing the documents.

The Bubbles Interface helped its users organize the documents faster and perform quicker when selecting the initial set of documents for clusters in which the entire collection of documents will be organized. Moreover, the users ended with more accurate organization of the documents into clusters. Finally, users indicated that the Bubbles Interface had: 1) better presentation of the elements on the display; 2) better positioning of the document and cluster views; 3) more helpful reverse action capabilities; 4) easier to understand and more helpful feedback; 5) and better organization of the final document collection than the Pie Interface.

Reflecting on the research questions listed in Chapter 1, users of the Bubbles Interface indicated equivalent preferences of the use of the abstract and text cloud views. The effectiveness of the text cloud was similar to that of the abstract of the document during the classification process. The users considered presenting clusters as bubbles containing glyphs (documents) to be effective. The modifications (interactions) capabilities were also considered effective by the users in the case of the Bubbles Interface. Even though 50% of the users considered that the list of labels for clusters was not a necessary view, the rest of the participants wanted keeping both the bubbles and the list views of clusters. The list would give an idea about the clusters without the content of each cluster which would help in many cases as the users indicated. Finally, most users (9/10) stated the effectiveness of presenting the final set of clusters in bubbles containing glyphs that represent the documents. Only 60% of the users indicated that the Pie Interface had effective presentation of the final set of clusters.

Not only did the visualized clustering features implemented in the Bubbles Interface improve usability, but they achieved more effective categorization of the results as discovered by the users. The results of the study showed that the use of visual clustering helped users with identifying the documents initially chosen as seeds for clusters and also resulted in more effective clustering compared to the Pie Chart used in the Pie Interface (as shown in Section 4.3.5). Moreover, the visualization of the titles and summaries of the documents may have also helped users with the choices of documents for each cluster which may also have led to the improved effectiveness in the final organization stage.

The second and main study of this research was motivated by the need for more effective organization of personal collections of documents (as discussed in Chapter 1) and by the results developed the smaller-scale study illustrated in Chapter 3. Both studies used different features of clustering involving aspects of visualization to organize documents for similar purposes. The visualization concepts were developed based on research in the literature and were integrated with aspects of clustering to build the interface investigated in Chapter 4 which was also guided by the recommendations from the first study. Future work under the topic of organization of personal collections of digital documents may continue with investigating issues beyond the clustering stage. What the user can do with the documents after the organization is completed should be considered for investigation involving re-clustering and the decrease and increase in the number of documents in the collection.

There are particular features of the Bubbles Interface that were essentially built based on recommendations from the study discussed in Chapter 3 and the results of the work of Hu et al. (2012). Those features may have contributed to the users' perceived effectiveness of the interface. First, the cluster visualization was intuitive so that documents appear as glyphs in a bubble container showing what looks like a cluster of grapes for instance. This visualization makes it easier to see at every time what content the cluster has. Second, the documents' visualization inside each cluster with the use of focus is very intuitive. The use of focus+context shows the glyph of the document with the title (focus) within the cluster (context) in an easy to understand and compare manner. The user can hover over the content of the cluster readily and effectively.

Third, the multiple views of the clusters (list + intuitive bubbles) is another factor that may have contributed to improving the interface by satisfying multiple preferences which may suite users with different expertise. Fourth, the instant reflection of the content (documents) inside each cluster as the user adds seeds to clusters and the easiness of removing documents (glyphs from bubbles) is another factor that may have contributed to the effectiveness perceived in the Bubbles Interface. Finally, the fixed view of the document content helped users find what they were looking for easily and quickly. This factor had a lot of complaints on the Pie Interface. There are other interaction (editing and

modifying) features that have also helped with the improvement achieved with the Bubbles Interface.

With regard to specific aspects of future work, further investigations may start by considering adding some missing features to the Bubbles Interface such as a search feature that would allow users to search for documents before and after they are assigned to clusters. The interface may also be considered for evaluation in field settings with personal document collections. The Bubbles Interface could also be improved by adding the ability to save clusters for future viewing and searching for documents. It can also be enhanced by adding the ability to modify clusters already created after accomplishing the organization process. This can be done by allowing the users to move documents around and add new documents with and without the need for re-clustering.

If the study is to be repeated in a controlled laboratory environment, and the two interfaces were to be compared, a slightly different approach will be followed. The emphasis will be more on the process of choosing the initial seeds for clusters with respect to the data logged in the study. Manual clustering of the collection of documents will have to be performed first by experts. The collection will be split into two sets of documents of different types. Each set will be given to a focus group which will be asked to classify the documents into clusters (the ground truth). The clusters will be used for later comparisons.

The participants will work on one interface using one set of the documents and on the other interface using the other set of documents. The time needed to select the initial set of clusters will be recorded and not the entire time of the study. The same underlying clustering algorithm will be used and the final set of results will not be considered for any measures. The study will measure for the effect of document to see if one set of documents was different from the other set. An algorithm will be built and implemented to compare the initial set of clusters (seeds only) to the ground truth created manually by users prior to the study. The results of the comparison will show to which extent was the interface effective with helping the user accurately select documents for their clusters. The user opinions will still be accumulated in questionnaires.

The research can be extended in two directions. First, the interface can be utilized in web search. The user submits a query and gets a chance to classify a subset of the results into initial clusters (topics) the user needs to cover. The interface can then re-present a larger subset of the results organized into clusters. This may lead to discovery of more topics in the results and finding relevant results more effectively. Second, the interface can be investigated on small-screen devices. Those devices are becoming very popular and the sizes of the personal collections of documents accumulated on those devices are on the rise according to Vertic[22]. The presentation of the final organized clusters may help users with locating documents on those devices.

[22] http://www.vertic.com

BIBLIOGRAPHY

Alhenshiri, A., & Blustein, J. (2011, July). Exploring visualization in web information retrieval. *International Journal for Internet Technology and Secured Transactions, 3*(3), 320-330.

Alhenshiri, A., Brooks, S., Watters, C., & Shepherd, M. (July, 2010a). Augmenting the visual presentation of web search results. *The 5th International Conference on Digital Information Management*, (pp. 101-107). Thunder Bay, ON, Canada.

Alhenshiri, A., Shepherd, M., Watters, C., & Duffy, J. (September, 2010b). Web information gathering tasks: a framework and research agenda. *the International Conference on Knowledge Discovery and Information Retrieval*, (pp. 131-140). Valencia, Spain.

Au, P., Carey, M., Sewraz, S., Guo, Y., & Rugers, S. (July, 2000). New Paradigms in Information Visualization. *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 307-309). Athens, Greece.

Badesh, H., & Blustein, J. (June, 2011a). Visual Clustering in Web Search: An Effective Approach. *In Proceeding of the 2011 Information Society Conference* (i-society), London, UK, , (pp. 34-38).

Badesh, H., & Blustein, J. (December, 2011b). Improving Finding and Re-finding Web search Results Using Clustering and Visualisation. *In International Journal of Intelligent Computing Research* (IJICR) , *2*(1-4), 228-234.

Badesh, H., & Blustein, J. (February, 2012). VDMs for Finding and Re-finding Web Search Results. *2012 iConference,* (pp. 419-420). Toronto, ON, Canada: ACM.

Bauer, D. F. (2005). Spatial Tools for Managing Personal Information Collections. *HICSS2005* (pp. 104-106). Hawaii, USA: IEEE Computer Society.

Berendt, B., Krause, B., & Kolbe-Nusser, S. (2010, January). Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing and Management: an International Journal, 46*(1), 1-10.

Bergman, O., R., B.-M., & Nachmias, R. (April, 2006). The project fragmentation problem in personal information. *SIGCHI Conference on Human Factors in Computing Systems*, (pp. 271-274). Montréal, QC, Canada.

Bladh, T., Carr, D., & Scholl, J. (June, 2004). Extending Tree-Maps to Three Dimensions: a Comparative Study. *the 6th Asia-Pacific Conference on Computer-Human Interaction (APCHI2004)*, (pp. 50-59). Rotorua, New Zealand.

Bonnel, A., Lemaire, V., Cotarmanc'h, A., & Morin, A. (2006). Effective organization and visualization of web search. *the 24th IASTED International Conference on Internet and multimedia systems and applications*, (pp. 209-216). Anaheim, CA, USA.

Bonnel, N., Cotarmanac'h, A., & Morin, A. (2005). Meaning Metaphor for Visualizing Search Results. *9th International Conference on Information Visualization*, (pp. 467-472). London, England.

Cai, D., He, X., Li, Z., Ma, W., & Wen, J. (2004). Hierarchical Clustering of WWW Image Search Results Using Visual, Textual, and Link Information. *12th Annual ACM International Conference on Multimedia*, (pp. 951-959). New York, NY, USA.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think.* San Francisco, CA, USA: Morgan Kaufman Publishers.

Carpineto, C., Osiński, S., Romano, G., & Weiss, D. (2009, July). A Survey of Web Clustering Engines. *ACM Computing Surveys, 41*(3).

Civan, A., Jones, W., Klasnja, P., & Bruce, H. (2008). Better to organize personal information by folders or by tags?: The devil is in the details. *Journal of the American Society for Information Science and Technology, 45*(1), 1-13.

Cockburn, A., & McKenzie, B. (2002). Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. *SIGCHI Conference on Human Factors in Computing Systems*, (pp. 203-210). Minneapolis, Minnesota, USA.

Cutrell, E., Dumais, S., & Teevan, J. (January, 2006). Searching to Eliminate Personal Information Management. *In Communications of the ACM (Special Issue: Personal Information Management)*, (pp. 58-64).

Doll, W. J., & Torkzadeh, G. (1988). The Measurement of End User Computing Satisfaction. *MIS Quarterly , 12*(2), 259-274.

Dong, L., Watters, C., & Shepherd, M. (2008). An Examination of Genre Attributes for Web Page Classification. *41st Annual Hawaii International Conference on System Sciences*, (pp. 133-143). Waikoloa, Hawaii, USA.

Dowdy, S., Wearden, S., & Chilko, D. (February, 2004). *Statistics for Research.* (3, Ed.)

Drucker, S. M., Fisher, D., & Basu, S. (2011). Helping users sort faster with adaptive machine learning recommendations. *INTERACT'11 Proceedings of the 13th IFIP TC 13 International Conference on Human-computer interaction*, (pp. 187-203). Springer-Verlag Berlin, Heidelberg.

Efron, M., Elsas, J., Marchionini, G., & Zhang, J. (2004). Machine Learning for Information Architecture in a Large Governmental Website. *4th ACM/IEEE-CS Joint Conference on Digital libraries*, (pp. 151-159). Tucson, AZ, USA.

Elsweiler, D., & Ruthavan, I. (2007). Towards Task-based Personal Information Management Evaluations. *the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 23-30). Amsterdam, The Netherlands: ACM.

Ferizis, G., & Bailey, B. (2006). Towards Practical Genre Classification of Web Documents. *15th International Conference on World Wide Web*, (pp. 1013-1014). Edinburgh, Scotland.

Ferragina, P., & Gulli, A. (2005). A Personalized Search Engine Based on Web-snippet Hierarchical Clustering. *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, (pp. 810-810). Chiba, Japan.

Friendly, M. (n.d.). *Milestones in the History of Thematic Cartography, Statistical, Graphics, and Data Visualization.* Retrieved 03 2009, 2009, from York University Archives: http://www.math.yorku.ca/scs/gallery/milestone/

Grewal, R. S., Jackson, M., Burden, P., & Wallis, J. (2000). Visual Representation of Search-engine Queries and their Results. *1st International Conference on Web Information Systems Engineering*, (pp. 352-356). Hong Kong, China.

Havre, S., Hetzler, E., Perrine, K., Jurrus, E., & Miller, N. (2001). Interactive visualization of multiple query results. *the IEEE symposium on information visualization 2001 (INFOVIS'01)*, (pp. 105-112). San Diego, California, USA.

Heer, J., Card, S. K., & Landay, J. A. (2005). Prefuse: A toolkit for Inter- active information visualization. *the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 421-430). Portland, Oregon, USA.

Hoeder, O. (2008). Web information retrieval support systems: the future of web search. *the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 29-32). Washington DC, USA,: IEEE.

Hu, Y., Milios, E., Blustein, J., & Liu, S. (2012). Personalized Document Clustering with Dual Supervision. *the 12th ACM Symposium on Document Engineering* (pp. 161-170). Paris, France: ACM.

Intranet: Eindhoven, university of technology. (n.d.). Retrieved January 05, 2012, from University of Technology Eindhoven: http://w3.win.tue.nl/nl/onderzoek/onderzoek_informatica/visualization/sequoiaview/about_sequoiaview/

Jaballah, I. C. (2005). Managing Personal Documents with a Digital Library. *Research and Advanced Technology for Digital Libraries, 9th European Conference, Research and Advanced Technology for Digital Libraries, 9th European Conference*, (pp. 18-23). Vienna, Austria.

Jing, F., Wang, C., Yao, Y., Deng, K., Zhang, L., & Ma, W. (2006). IGroup: Web Image Search Results Clustering. *ACM Multimedia Information Retrieval (MIR)* (pp. 377–384). 2006: ACM.

Joho, H., & Jose, J. M. (2006). A comparative study of the effectiveness of search results presentation on the Web. *Lecture Notes in Computer Science, SpringerLink, 3936*, 302-313.

Jones, E., Bruce , H., Klasnja, P., & Jones, W. (2008). I Give Up! Five Factors that Contribute to the Abandonment of Information Management Strategies. *68th Annual Meeting of the American Society for Information Science and Technology (ASIST 2008).* Columbus, OH.

Jones, W. (2007). *Keeping Found Things Found: The Study and Practice of Personal Information Management.* San Francisco, CA, USA: Morgan Kaufmann Publishers.

Jones, W., Bruce, H., & Dumais, S. (2003). How do People Get Back to Information on the Web? How Can They Do It Better? *9th IFIP TC13 International Conference on Human-Computer Interaction.* Zurich, Switzerland.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology Visualization Methods: a Survey. *ACM Computing Surveys, 39*(4), Article 10.

Kawano, H. (2000). Overview of Mondou Web Search Engine Using Text Mining and Information Visualizing Technologies. *International Conference on Digital Libraries*, (pp. 234-244). Kyoto, Japan.

Kim, K. (2008). Effects of emotion control and task on Web searching behavior. *Information processing and management: An International Journal, 44*(1), 373-385.

Knoll, S. H., Hoff, A., Fisher, D., Dumais, S., & Cutrel, E. (2009). Viewing Personal Data Over Time. *CHI 2009 Workshop on Interacting with Temporal Data*, (pp. 1-4). Boston, MA, USA.

Kobayashi, T., Misue, K., Shizuki, B., & Tanaka, J. (2006). Information gathering support interface by the overview presentation of web search results. *the 2006 Asia-Pacific Symposium on Information Visualisation*, (pp. 103-108). Darlinghurst, Australia.

Kules, W., Wikson, M. C., Schrafel, C., & Shneiderman, B. (2008). *From keyword search to exploration: how result visualization aids discovery on the Web.* Southampton, UK: School of Electronics and Computer Science, University of Southampton.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press.

Mason, J. E., Shepherd, M., & Duffy, J. (2009). An N-Gram Based Approach to Automatically Identifying Web Page Genre. *HICSS 2009*, (pp. 1-10). HI, USA.

Nguyen, T., & Zhang, J. (2006). A Novel Visualization Model for Web Search Results. *IEEE Transactions on Visualization and Computer Graphics, 12*(5), 981-988.

Nortel, C. J., & Kazman, R. (1997). WebQuery: searching and visualizing the web through connectivity. *The 6th International Conference on World Wide Web*, (pp. 1257-1267).

Paulovich, F., Pinho, R., Botha, C. B., Heijs, A., & Minghim, R. (2008). PEx-WEB: content-based visualization of web search results. *12th International Conference on Information Visualization*, (pp. 208-214). London, UK.

Risden, K., Czerwinski, M. P., Munzner, T., & Cook, D. B. (2000). Initial examination of ease of use for 2 D and 3 D information visualizations of web. *International Journal of Human-Computers Studies, 53*(5), 695-714.

Rivadeneira, W., & Bederson, B. B. (2003). *A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable User Interfaces.* Baltimore, MD, USA: University of Maryland.

Roberts, J. C., Boukhelifa, N., & Rodgers, P. (2002). Multiform Glyph Based Web Search Result Visualization. *Proc. 6th International Conference on Information Visualisation*, (pp. 549-554). London, England.

Santini, M., & Sharoff, S. (2009). Web Genre Benchmark under Construction. *Language Technology and Computational Linguistics (JLCL)*, 25(1).

Shneiderman, B. (1980). *Software Psychology: Human factors in computer and information systems* . Winthrop Publishers.

Shneiderman, B., Feldman, D., Rose, A., & Grau, X. F. (2000). Visualizing Digital Library Search Results with Categorical and Hierarchical Axes. *5th ACM International Conference on Digital Libraries*, (pp. 57-66). San Antonio, TX, USA.

Spink, A., Wolfram, D., Jansen, M., & Saracevic, T. (2001). Searching the Web: the Public and Their Queries. *Journal of the American Society for Information Science and Technology*, 226-234.

Stubbe, A., Ringlstetter, C., Zheng, T., & Goebel, R. (2007). Incremental Genre Classification. *Colloquim Hel in Conjunction with Corpus Linguistics.* Birmingham, UK.

Sugiyama, K., Hatano, K., & Yoshikawa, M. (2004). Adaptive web search based on user profile constructed without any effort from users. *the 13th International Conference on World Wide Web*, (pp. 675-684). New York, NY, USA.

Sutcliffe, A. G., & Patel, U. (1996). 3D or not 3D: is it nobler in the mind? *the BCS Human Computer Interaction Conference on Peopleand Computers XI*, (pp. 79-94). London, UK.

Suvanaphen, E., & Roberts, J. C. (2004). Textual difference visualization of multiple search results utilizing detail context. *the Theory and Practice of Computer Graphics 2004 (TPCG'04)*, (pp. 2-8). Washington, DC, USA.

Teevan, J. (2008). How People Recall, Recognize, and Reuse Search Results. *ACM Transactions on Information Systems, 26*(4), Article No. 19.

Teevan, J., Cutrell, E., Fished, D., Drucker, S. M., Ramos, G., & Andre, P. (2009). Visual snippets: summerizing web pages for search and revisitation. *the 27th International Conference on human factors in computing systems*, (pp. 2023-2032). Boston, MA, USA.

Turetken, O., & Sharda, R. (2005). Clustering-based Visual Interfaces for Presentation of Web Search Results: An Empirical Investigation. *Information Systems Frontier, 7*(3), 273–297.

Tyler, S. K., & Teevan, J. (2010). Large scale query log analysis of re-finding. *the 3rd ACM International Conference on Web Search and Data Mining.* Santa Cruz, CA, USA.

Wijk, J. J. (1999). Cushion treemaps: Visualization of Hierarchical Information. *IEEE Computer Science* (pp. 73-79). IEEE.

Wiza, W., Walczak, K., & Cellary, W. (2004). Periscope: a System for Adaptive 3D Visualization of Search Results. *9th International Conference on 3D Web Technology*, (pp. 29–40). Monterey, CA, USA.

Woodruff, A., Faulring, A., Rosenholtz, R., Morssion, J., & Pirolli, P. (2001). Using Thumbnails to Search the Web. *SIGCHI Conference on Human Factors in Computing Systems*, (pp. 198-205). Seattle, WA, USA.

Zaina, C. M., & Baranauskas, M. C. (2005). Revealing relationships in search engine results. *the 2005 Latin American conference on Human-computer interaction*, (pp. 120-127). New York, NY, USA.

Zamir, O., & Etzioni, O. (1999). Grouper: a Dynamic Clustering Interface to Web Search Results. *8th International Conference on World Wide Web* (pp. 1361-1374). Toronto, ON, Canada: ACM.

APPENDIX A. Consent Form

DALHOUSIE UNIVERSITY
Inspiring Minds

Informed Consent Form

Project Title
Interactive Document Collection Visualization (IDCV) for Organizing Personal Information

Project Contact
Hoda Badesh, Faculty of Computer Science, 6050 University Avenue, Halifax, NS, B3H 1W5, Canada, phone: 902-494-2093, Fax: 902-492-1517, e-mail: badesh@cs.dal.ca

Who Will Be Conducting the Research
Hoda Badesh, the principal investigator and Jamie Blustein and E. Milios, the research supervisors (jamie@cs.dal.ca and eem@cs.dal.ca).

Introduction
You are invited to participate in a research study being conducted by Hoda Badesh who is a Master's student at the Faculty of Computer Science, Dalhousie University. This study is being done as part of a research project.

In order to participate in this study, your major should be computer science or marine biology. The information in this consent form will outline any possible risks, inconvenience and discomfort that you might experience.

Participation is voluntary, and participants are free to withdraw at any time without repercussions. If you have any concern and question about the study, please do not hesitate to ask the principal investigator, *Hoda Badesh*.

Purpose
This study is an exploration to determine what operations users employ often to organize their document collection and whether user input can help machine-learning algorithm to organize the collection better based on users' point of view.

Study Design
This is a focus group in which you will be asked to classify 12 documents from 30 documents into clusters, fill in pre-study and post-study questionnaires, and join in a post-testing discussion. The results of the study can be seen in possible publications.

What you will be asked to do

The following table indicates the study activities and time associated with the tasks in which you will be asked to participate:

Page 1 of 4

Activities	*Time*	*Comments*
Participants will be given the consent form to read and sign two days before conducting the study.	5 minutes	
Participants will be provided with the data set (given in a flash memory or attached with an e-mail).	60 minutes	It may take one hour to skim through the whole data set.
Next is the study day Agenda:		
Participants will complete an online demographic questionnaire.	10 minutes	
Participants will start the two testing stages*. Stage 1, participants will do the following: Each group will be given a short tutorial on how to use the user interface and become familiar with it.	5 minutes	*Groups will be working on different interfaces at the same time. After finishing the first stage, interfaces will be switched between the groups.
Participants will be asked to perform the study task.	15 minutes	
Participants will be asked to evaluate the final clustering and decide which documents were clustered correctly based on the user's point of view.	10 minutes	
Each participant will then fill out an online post-study questionnaire about their satisfaction with the user interface.	10 minutes	
Stage 2, interfaces will be switched so the each group has a chance to use the interface they did not use in the first stage. Each group will be given a short tutorial on how to use the user interface and become familiar with it.	5 minutes	
Participants will be asked to perform the study task.	15 minutes	
Participants will be asked to evaluate the final clustering and decide which documents were clustered correctly based on the user's point of view.	10 minutes	
Each participant will then fill out an online post-study questionnaire about their satisfaction with the user interface.	10 minutes	
The principal investigator, Hoda Badesh, will lead the	30 minutes	

group discussion. She will discuss the effectiveness of the presentations on each interface, the ease of interaction, and the layout architecture with the focus group members.		

Who can Participate in the Study

This study needs participants who are computer science or marine biology students and read scientific papers at least monthly. You should be able to summarize its topic after reading a paper. The study will be conducted at the usability room in the Faculty of Computer Science at Dalhousie University.

Risks of Participation

Participation as a user does not involve risks beyond those encountered in daily life. Note that withdrawal from the study is optional in case of discomfort.

Benefits of Participation

Your participation in the study will help us investigate the effectiveness of our user interface. Your participation will be greatly appreciated, and we expect that it will help us to learn more how users can interact with document clustering and develop new tools to aid researchers to better organize their personal library of academic papers.

Compensation

As a way to thank you for your participation in the study, you will be given $20 in cash upon the completion of the entire study.

Time Commitment

The study tasks will need about 180 minutes to complete.

Confidentiality & Anonymity

All participants will be asked not to disclose anything said within the context of the discussion. By agreeing to participate, you agree to not disclose to others outside this event anything said within the context of the discussion. All identifying information will be removed from the collected materials, and all materials will be stored securely in the researchers' password-protected computer.

Permission to Quote:

I may wish to quote your words directly in reports and publications resulting from this. With regards to being quoted, please check yes or no for each of the following statements:

Researcher may publish documents that contain quotations by me under the following conditions: (Answer ONLY one of the following)		
1. I agree to be quoted directly (my name is used).	Yes	No
2. I agree to be quoted directly if my name is not published (I remain anonymous).	Yes	No
3. I agree to be quoted directly if a made-up name (pseudonym) is used.	Yes	No

By signing this consent form, you are indicating that you fully understand the above information and agree to participate in this study.

Participant's signature _____

Questions

If you have any questions or concerns about this study, please contact the principal investigator, Hoda Badesh, via the email specified above.

If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, you may contact Catherine Connors, Director, Research Ethics, Dalhousie University, for assistance at (902) 494-1462, ethics@dal.ca

I have read the explanation about this study. I have been given the opportunity to discuss it and my questions have been answered to my satisfaction. I hereby consent to take part in this study. However I realize that my participation is voluntary and that I am free to withdraw from the study at any time.

Participant Signature:

Would you like to share the results of the study? (Yes/No)?

If yes, please leave your email address: .....

_____ _____
Name Signature

Principal Investigator Signature:

___Hoda Badesh____ _____
Name Signature

APPENDIX B. Study Task

Study Task[1]

You are asked to classify 12 documents from a 30-document collection into clusters. Each new cluster you create has to have at least one labeling word. After getting the final results, please judge the clusters in terms of whether documents in the right cluster or not. A list of all documents names are provided to check in front any document that is incorrect cluster based on your perspective.

1- This task will be applied on both interfaces.

APPENDIX C. Questionnaires

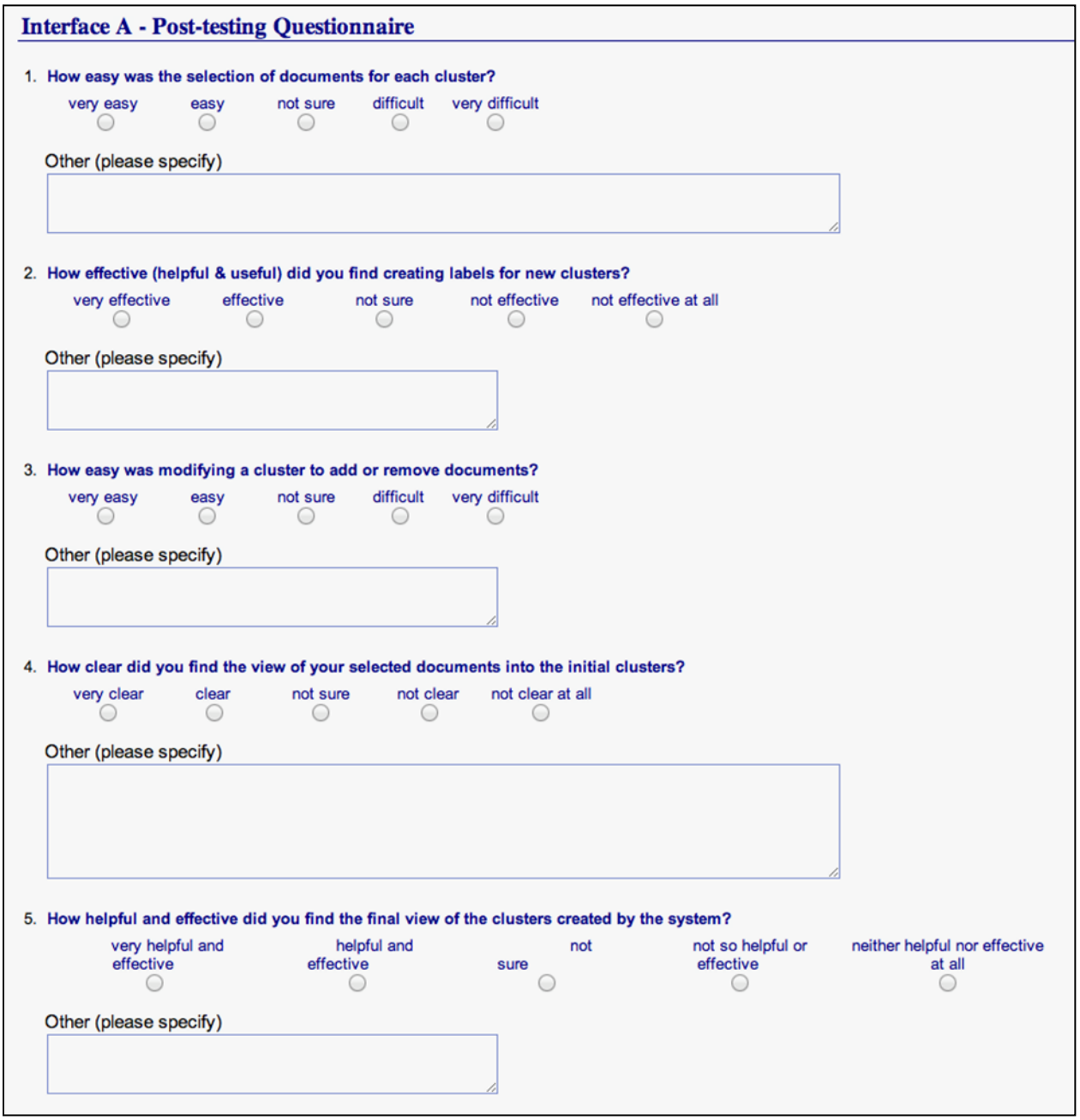
C.1. Background Questionnaire

Background Questionnaire

Please answer the questions below:

1. Age:
   18-22 23-30 >30
2. Gender:
   Male. Female. Other.
3. Are you?
   Undergraduate. Graduate.
4. How big is your own document collection? (approximately)
   <=100 documents 100-500 documents >500 documents
5. How many hours (per week) do you spend on organizing your document collection:
   1-2 hour(s) 3-4 hours >=5 hours
6. Where do you usually prefer to keep your electronic document collection:
   Your own machine (desktop/laptop computer) External memory (USB driver, external hard driver)
   Other:_____.
7. Do you have any difficulties finding a good strategy to organize your document collection?
   Yes. No.
   If yes, please tell us what difficulties do you face?
   _____.
8. Have you used any application for organizing your document collection?
   Yes. No.
   If yes, was it useful? (please answer with yes or no)
   _____.

C.2. Bubbles Interface – Post-testing Questionnaire



Interface A - Post-testing Questionnaire

1. **How easy was the selection of documents for each cluster?**

   very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

   Other (please specify)

2. **How effective (helpful & useful) did you find creating labels for new clusters?**

   very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

   Other (please specify)

3. **How easy was modifying a cluster to add or remove documents?**

   very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

   Other (please specify)

4. **How clear did you find the view of your selected documents into the initial clusters?**

   very clear ○ clear ○ not sure ○ not clear ○ not clear at all ○

   Other (please specify)

5. **How helpful and effective did you find the final view of the clusters created by the system?**

   very helpful and effective ○ helpful and effective ○ not sure ○ not so helpful or effective ○ neither helpful nor effective at all ○

   Other (please specify)

6. **How do you rate the presentation of elements on the interface?**

very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

Other (please specify)

7. **How do you rate the positioning of the document view and cluster view on the screen?**

very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

Other (please specify)

8. **How easy was it to undo actions on the interface?**

very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

Other (please specify)

9. **Was the feedback from the interface (e.g. error messages) helpful to you?**

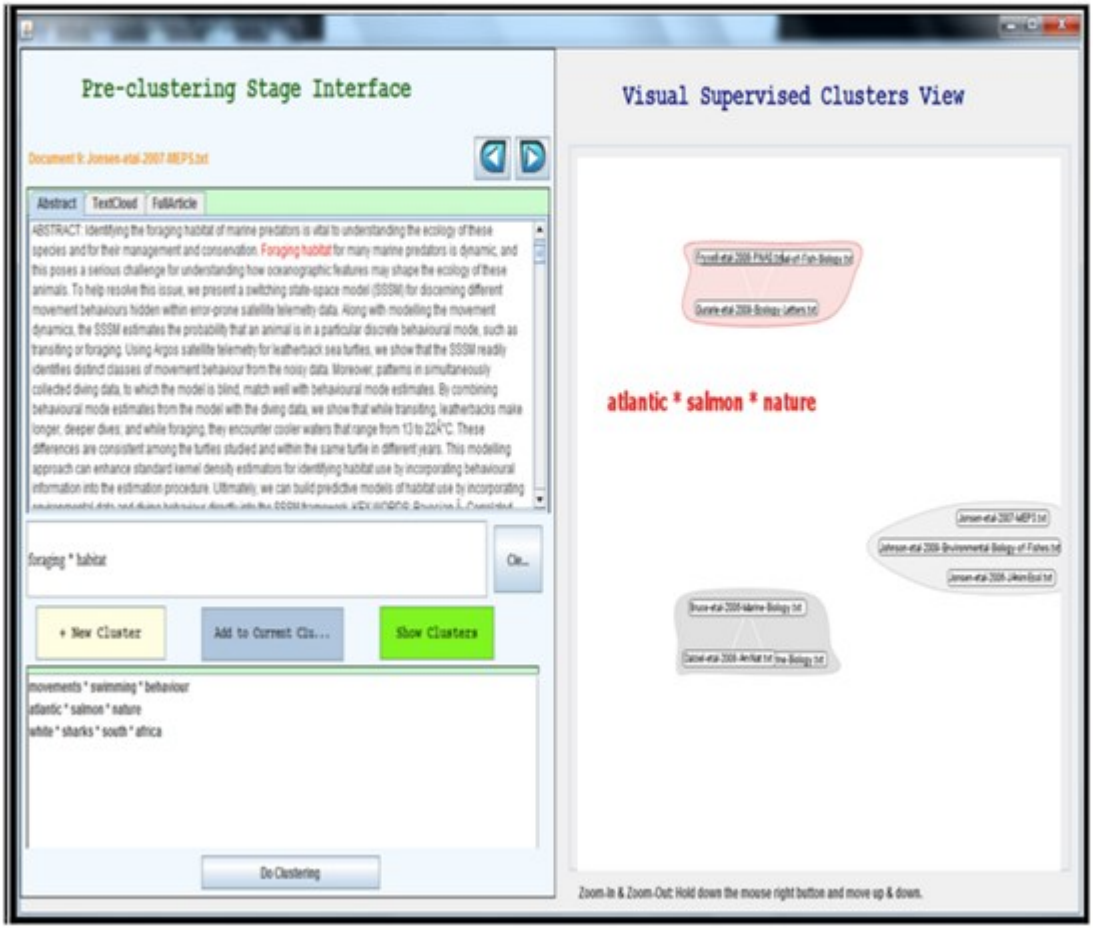
very clear and helpful ○ clear and helpful ○ not applicable ○ not clear or helpful ○ very vague and confusing ○

Other (please specify)

10. **How helpful and effective do you think the interface will be with organizing your collection of documents?**

very helpful and effective ○ helpful and effective ○ not sure ○ not so helpful or effective ○ neither helpful nor effective at all ○

11. **How effective (useful & helpful) was each of the following interaction features?**



	very effective and helpful	effective and helpful	not sure	not effective or helpful	neither effective not helpful at all
Immediately showing any added clusters in the clusters list	○	○	○	○	○
Instantly showing new clusters in the visualized (bubble) view of clusters	○	○	○	○	○
Going back and forth among documents by showing the abstracts and titles immediately	○	○	○	○	○
Selecting words (terms) from documents so they get added immediately to the label of the cluster	○	○	○	○	○
Choosing a cluster from the clusters list to add the current document to that cluster	○	○	○	○	○
Continuously modifying clusters you created initially	○	○	○	○	○
Seeing your cluster label when you hover over the glyph that represents the cluster	○	○	○	○	○
Deleting any document or cluster from the bubble view	○	○	○	○	○
Zooming in and out in the bubble view	○	○	○	○	○

Add your comments here

12. **How effective (useful & helpful) was each of the following visualization features?**

	very effective	effective	not applicable/ not sure	not effective	not effective at all
The display of the text cloud of the document	○	○	○	○	○
The display of the title of the document	○	○	○	○	○
The list of clusters displayed to you under the labelling box	○	○	○	○	○
The display of the abstract of the document	○	○	○	○	○
The visualization of the clusters in the bubble view while creating the initial set of clusters	○	○	○	○	○
The visualization of the final set of clusters	○	○	○	○	○
The layout of the items on the screen	○	○	○	○	○
The display of the full text cloud of the document	○	○	○	○	○
The use of different colours with the text cloud presentation of the document content	○	○	○	○	○

Other (please specify)

13. **Did the interface help with the categorization of the documents in the final results?**

Yes ○ No ○

How?

14. **Are there any features that should be added to the interface to make it more effective?**

Yes ○ No ○

If yes, please list them:

15. **Did you encountered with any challenges while working on the interface?**

Yes ○ No ○

If yes, please tell us what challenges did you have;

16. **Any other issues you had with the interface?**

Yes. ○ No. ○

If yes, please list them:

C.3. Pie Interface – Post-testing Questionnaire

1. **How easy was the selection of documents for each cluster?**

   very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

   Other (please specify)

2. **How effective (helpful & useful) did you find creating labels for new clusters?**

   very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

   Other (please specify)

3. **How easy was modifying a cluster to add or remove documents?**

   very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

   Other (please specify)

4. **How clear did you find the view of your selected documents into the initial clusters?**

   very clear ○ clear ○ not sure ○ not clear ○ not clear at all ○

   Other (please specify)

5. **How helpful and effective did you find the final view of the clusters created by the system?**

   very helpful and effective ○ helpful and effective ○ not sure ○ not so helpful or effective ○ neither helpful nor effective at all ○

   Other (please specify)

6. **How do you rate the presentation of elements on the interface?**

   very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

   Other (please specify)

7. **How do you rate the positioning of the document view and cluster view on the screen?**

very effective ○ effective ○ not sure ○ not effective ○ not effective at all ○

Other (please specify)

8. **How easy was it to undo actions on the interface?**

very easy ○ easy ○ not sure ○ difficult ○ very difficult ○

Other (please specify)

9. **Was the feedback from the interface (e.g. error messages) helpful to you?**

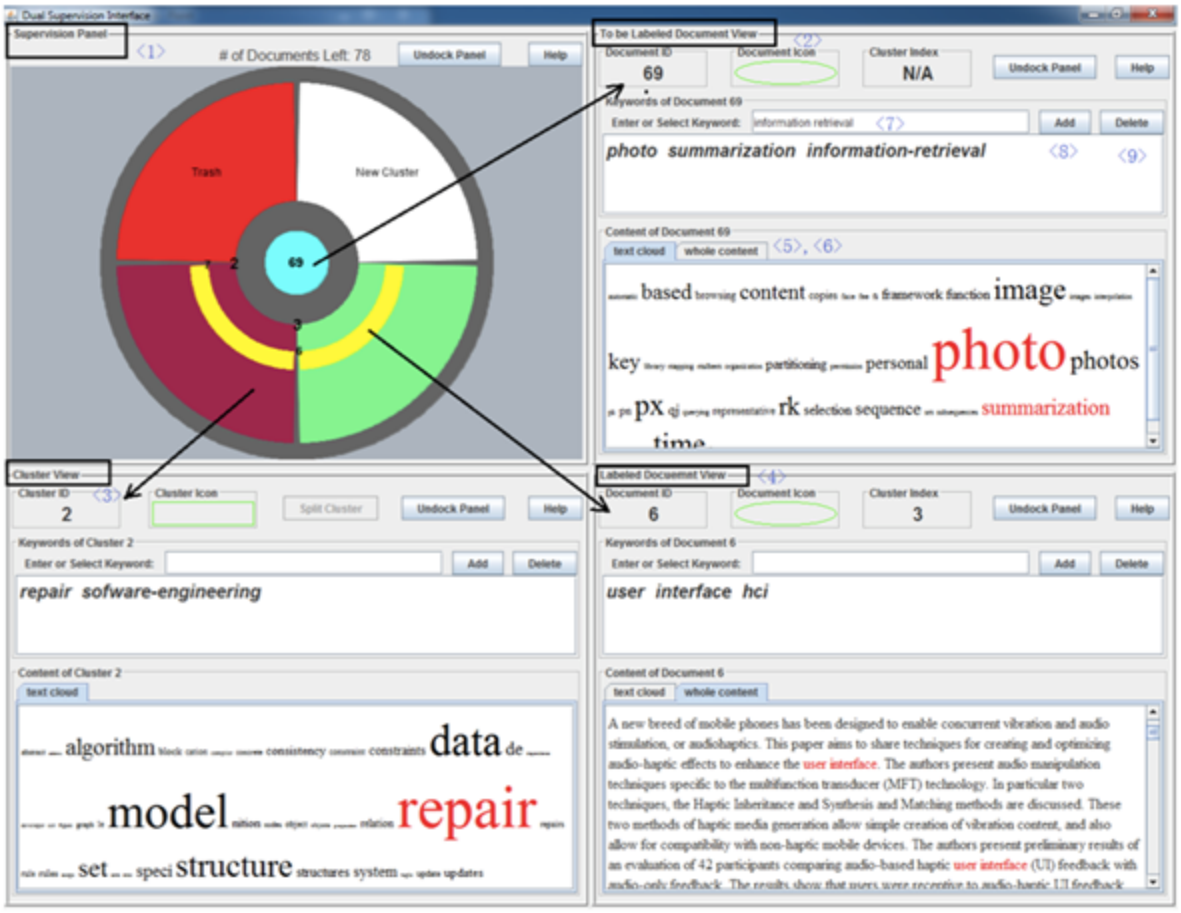
very clear and helpful ○ clear and helpful ○ not applicable ○ not clear or helpful ○ very vague and confusing ○

Other (please specify)

10. **How helpful and effective do you think the interface will be with organizing your collection of documents?**

very helpful and effective ○ helpful and effective ○ not sure ○ not so helpful or effective ○ neither helpful nor effective at all ○

11. **How effective was each of the following interaction features?**



	very effective and helpful	effective and helpful	not sure	not effective or helpful	neither effective not helpful at all
Dragging a document/cluster for either creating a new cluster or modifying it	○	○	○	○	○
Coloring any added clusters in the pie chart view	○	○	○	○	○
Showing new clusters with a certain color and ID in the pie chart view	○	○	○	○	○
Showing a new document with an ID in the pie chart view, and the document information (e.g. title and content) in the labeled-document view	○	○	○	○	○
The ability to select words (terms) from documents so they get added immediately to the label of the clusterContinuously modifying clusters you create initially	○	○	○	○	○
Seeing your cluster label on the keywords-of-cluster panel	○	○	○	○	○
Seeing a text cloud for all the documents in a cluster on the content-of-cluster panel	○	○	○	○	○
Deleting any document or cluster by dragging them to the trash	○	○	○	○	○

12. **How effective (useful & helpful) was each of the following visualization features?**

	very effective	effective	not applicable/ not sure	not effective	not effective at all
The display of the text cloud of the document content	○	○	○	○	○
The presentation of documents within a cluster as strips	○	○	○	○	○
The creation of the label of the cluster	○	○	○	○	○
The creation of the label of the cluster	○	○	○	○	○
The display of the full text of the document content	○	○	○	○	○
The visualization of the clusters in the pie chart view when you create the initial set of clusters	○	○	○	○	○
The display of the final set of clusters	○	○	○	○	○
The layout of the items on the screen	○	○	○	○	○

Other (please specify)

13. **Did the interface help with the categorization of the documents in the final results?**

Yes ○ No ○

How?

14. **Are there any features that should be added to the interface to make it more effective?**

Yes ○ No ○

If yes, please list them:

15. **Did you encountered with any challenges while working on the interface?**

Yes No

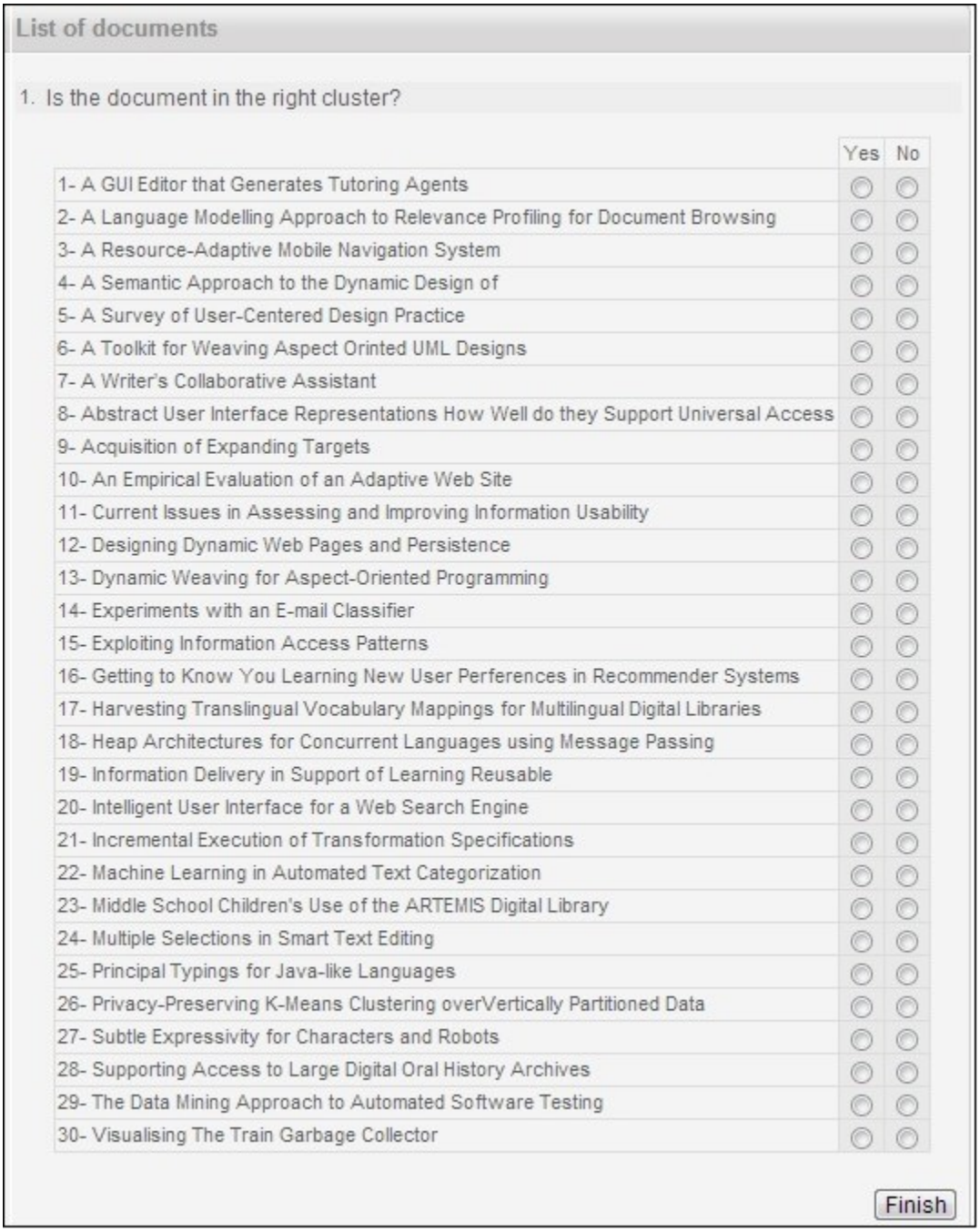
If yes, please tell us what challenges did you have;

16. **Any other issues you had with the interface?**

Yes No

If yes, please specify:

APPENDIX D. List of Documents



List of documents

1. Is the document in the right cluster?

	Yes	No
1- A GUI Editor that Generates Tutoring Agents	○	○
2- A Language Modelling Approach to Relevance Profiling for Document Browsing	○	○
3- A Resource-Adaptive Mobile Navigation System	○	○
4- A Semantic Approach to the Dynamic Design of	○	○
5- A Survey of User-Centered Design Practice	○	○
6- A Toolkit for Weaving Aspect Orinted UML Designs	○	○
7- A Writer's Collaborative Assistant	○	○
8- Abstract User Interface Representations How Well do they Support Universal Access	○	○
9- Acquisition of Expanding Targets	○	○
10- An Empirical Evaluation of an Adaptive Web Site	○	○
11- Current Issues in Assessing and Improving Information Usability	○	○
12- Designing Dynamic Web Pages and Persistence	○	○
13- Dynamic Weaving for Aspect-Oriented Programming	○	○
14- Experiments with an E-mail Classifier	○	○
15- Exploiting Information Access Patterns	○	○
16- Getting to Know You Learning New User Perferences in Recommender Systems	○	○
17- Harvesting Translingual Vocabulary Mappings for Multilingual Digital Libraries	○	○
18- Heap Architectures for Concurrent Languages using Message Passing	○	○
19- Information Delivery in Support of Learning Reusable	○	○
20- Intelligent User Interface for a Web Search Engine	○	○
21- Incremental Execution of Transformation Specifications	○	○
22- Machine Learning in Automated Text Categorization	○	○
23- Middle School Children's Use of the ARTEMIS Digital Library	○	○
24- Multiple Selections in Smart Text Editing	○	○
25- Principal Typings for Java-like Languages	○	○
26- Privacy-Preserving K-Means Clustering overVertically Partitioned Data	○	○
27- Subtle Expressivity for Characters and Robots	○	○
28- Supporting Access to Large Digital Oral History Archives	○	○
29- The Data Mining Approach to Automated Software Testing	○	○
30- Visualising The Train Garbage Collector	○	○

Finish

APPENDIX E. Focus Group Discussion Guide

Focus group discussion guide

Introduction – Welcome and introduce myself.

1- Explain the general purpose of the discussion.
2- Address the issue of confidentiality. All information collected will be confidential and Participants names will not be disclosed neither will any attributions for quotes be made in my final report (unless the participant give us a permission to use his name in the consent form). I hope this encourages you to speak openly.
3- Ground rules: these rules can help set the boundaries for decorum (e.g. mobile phones on silent) and for interaction and exchange (e.g. listening to others, no interrupting, speaking up).
4- Getting involved:

- Did the work on the first user interface make it easier to use the second?
- How difficult was the work on the interfaces?
- Which interface layout do you like most?

5- Description

- How would you describe interaction with each interface?
- Did you enjoy working on the interfaces? Which one did you enjoy most? Why?
- Did you find the visualization of clusters helpful and useful? Which presentation was more effective than the other (Pie Char View/Bubbles)? Was it as expected?

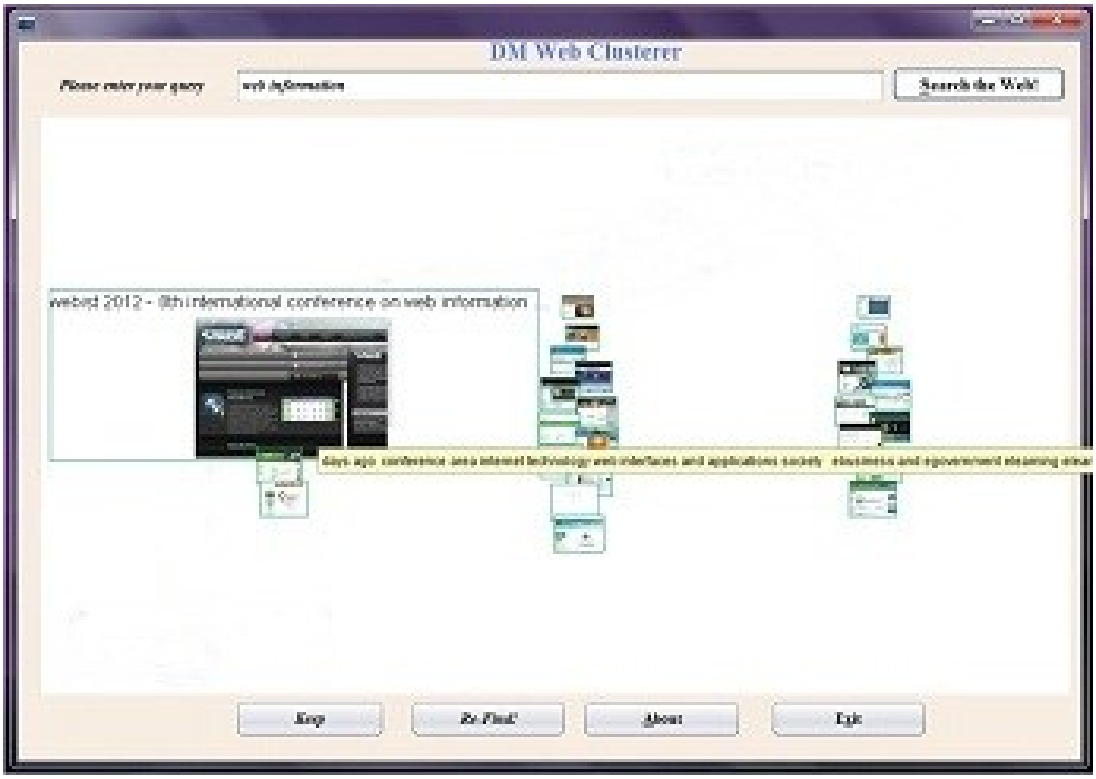
6- Improvement

- How would you like us to improve the layout of the interface?
- Are there any suggestions regarding the visual presentations?
- How would you like us to improve the color scale for both interfaces?
- Do you think that the interaction on both interfaces should be improved? How?

7- Conclusion

How could the interfaces be enhanced further?
What features/issues do you feel are important?
Any other comments?

APPENDIX F. DMSRP Interface without Labels



APPENDIX G. Pie Interface without Labels

APPENDIX H. Bubbles Interface without Labels

APPENDIX I. Qualitative Data from Post-testing Questionnaires

Question		Tool	Comments
1	How easy was the selection of documents for each cluster?	PI	- Easy for some and difficult for others. - It was difficult to see the documents labels, the cluster labels and the document content.
		BI	- It was amazing. - It was easy for some documents such as those related to HCI or interfaces but difficult for some others.
4	How clear did you find the view of your selected documents into the initial clusters?	PI	- Color of the cluster text should not be the same as the document text. - I could not see the name of documents and clusters. If I want to see them I need to click on each color to see which color belongs to which cluster. - Sometimes it is hard to remember which color stands for the target cluster.
		BI	- Confusing. - Visual representation of the cluster view needs to be improved.
5	How helpful and effective did you find the final view of the clusters created by the system?	PI	- Colors meant nothing to me on the pie chart when the important information is missing and they were confusing.
		BI	- It's a bit hard to read the documents' titles because of the overlapping.
6	How do you rate the presentation of elements on the interface?	PI	- The view area for the document content is very small. I had to scroll every time to read a line. - The color of the cluster should not be same as the color of the text. - Updating the labels was not available. - It is hard to remember what the numbers on the chart stand for (e.g. the content of the document or at least the title of document). - It gave me a headache.
		BI	- The colors in the text cloud were confusing.

Question		Tool	Comments
7	How do you rate the positioning of the document view and cluster view on the screen?	PI	- Zooming would be helpful. - The screen should maximize which would allow us to navigate more easily
		BI	- A counter should be there to indicate how many documents have been clustered and how many documents have been left.
11	How effective (useful & helpful) was each of the following interaction features?	PI	- Dragging the documents to trash was easy. - The color of the cluster should not be same as the color of the text. Maximize button should be there. - There is no way of going to the previous document. - Hovering arrow over the cluster or document should give the name of that cluster.
		BI	- Bubbles are moving away from the screen automatically. It is annoying for me because every time I need to see the cluster, I need to drag it to the center of the screen. - Moving the documents from one cluster to another in the cloud would have been very helpful. - Visual representation of the cluster view needs to be improved.
12	12. How effective (useful & helpful) was each of the following visualization features?	PI	- The presentation of documents within a cluster could be improved to something else than strips.

Question		Tool	Comments
13	Did the interface help with the categorization of the documents in the final results?	PI	- Moving the documents from one cluster to another was amazing. - Document strips are not helpful. It should show the document title, abstract, and content in larger view so I can see it clearly. - It should show the document and cluster name when I hover my cursor over it instead of clicking on it. - Too many boxes to choose from to add labels for the clusters. - It is confusing to go back and forth among documents. - It is hard to know the documents that exist on the pie chart. - It is hard to view the document's name in the content view.
		BI	- Very easy to add documents to previously created clusters. - Document cluster helps to select the potential labels for clusters. - Merging clusters was a very useful feature for me. - It was very helpful. I like the idea of cloud and just by clicking one can access the PDF files. - I can get a clear idea of how many clusters I created and get an overview of what kind of documents I put in them.
14	Are there any features that should be added to the interface to make it more effective?	PI	- No document title was visible anywhere. - When you look closely, it is merged with the full text along with other names in the same font size and color. - There is less possibility that the user will be able to identify the title of document. - Come up with a different design.

Question		Tool	Comments
16	Any other issues you had with the interface?	PI	- Difficult to view the presentation of the clusters. - The interface has no good layout, interaction, and clarity. - The information was hidden in this interface. - This interface always requires more time to work on.
		BI	- The interface can be divided to two permanent-vertical parts: 1) Visual bubbles view. 2) The two views of the document content (abstract and text cloud).