Evaluating exploratory visualization systems: A user study on how clustering-based visualization systems support information seeking from large document collections

Category: Research

Abstract—Iterative, opportunistic and evolving visual sensemaking has been an important research topic as it assists users in overcoming ever-increasing information overload. Exploratory visualization systems (EVSs) maximize users' information gain through learning and have been widely used in scientific discovery and decision making contexts. Although many EVSs have been developed recently, there is a lack of general guidance on how to evaluate such systems. Researchers face challenges such as understanding the cognitive learning process supported by these systems. In this paper, we present a formal user study on Newdle, a clustering-based EVS for large news collections, shedding light on a general methodology for EVS evaluation. Our approach is built upon cognitive load theory that takes the users as well as the system as the foci of evaluation. The carefully designed procedures allow us to thoroughly examine the users' cognitive process as well as control the variability among human subjects. Through this study, we analyze how and why clustering-based EVSs benefit (or not) users in a variety of information seeking tasks. We also summarize leverage points for designing clustering-based EVSs.

Index Terms—Clustering, cognition, human-computer interaction, information visualization, information seeking.

1 INTRODUCTION

Large bodies of new knowledge resources appear every day. For example, The New York Times (http://www.nytimes.com/) and The Guardian (http://www.guardiannews.com/) recently opened their online archives, bringing nearly two centuries' worth of information to the public [18]. As information overload continues to grow, there is a dire need for information seeking systems to support iterative, opportunistic, and evolving information foraging and decision making. Such exploratory analysis stimulates significant cognitive changes through learning, and thus allows users to gain improved understanding to manage large amounts of information [23]. Targeting this need, many exploratory visualization systems (EVSs) have been developed to support users in conducting exploratory analysis with the aid of visualization techniques.

Despite the popularity of EVSs, researchers are still trying to understand the nature of users' exploratory process through EVSs. The design space of EVSs is still unclear. Therefore, evaluation becomes critical throughout the development process of EVSs. Effective evaluations can help EVS designers learn whether the interactions between EVSs and users promote the desired cognitive changes for a good exploration experience, which can lead to a better understanding of how the system works and its design space. However, existing visualization evaluation methods largely ignore the internal cognitive activities of the users. We argue that a systematic approach to measuring and analyzing the cognitive process of users during exploratory visual analysis is desired for EVS evaluations.

In this paper, we present a preliminary exploration of EVS evaluation, using clustering-based EVSs for large news collections as an example. Clustering-based EVSs, such as Google News Timeline (http://newstimeline.googlelabs.com) and NewsMap (http://newsmapper.jp), are important approaches for large document collections exploration. They tightly integrate document clustering algorithms and interactive visualization. The document clustering algorithms attempt to group documents together based on their semantic relevance; thus documents with similar topics can be placed in a single cluster. The visualization presents the clusters to users and allows them to quickly obtain a data overview and easily locate interesting documents through interactive visual exploration [26]. Although a number of user studies have been conducted on individual systems [24][14][17][26], few of them conducted detailed analysis on the subjects' underlying cognitive process.

We conducted a formal user study on clustering-based EVSs, both for a general methodology of EVS evaluation and for a deeper understanding of clustering-based EVSs. Towards a general methodology, we seek practical methods for measuring and analyzing the cognitive process in EVS evaluations. The study also explores the benefits and constraints of clustering-based EVSs in different information seeking tasks, as well as leveraging points in EVS design.

The user study was conducted with 36 subjects. Four typical information seeking tasks, 3,640 New York Times (http://www.nytimes.com/) news articles, and four test beds were used in this study. A user-centered approach was adopted to capture and analyze detailed cognitive load measures, qualitative measures, and quantitative measures. The major insights from this user study include:

- Guided by cognitive load theory [20], we developed a practical yet effective procedure to study the cognitive process of the subjects. According to the theory, there are three different kinds of cognitive load. Intrinsic cognitive load is affected by individual differences, such as prior knowledge and mental models, rather than the EVS. Extraneous cognitive load is caused by deficient system design and thus needs to be reduced. Germane cognitive load reflects enhanced learning by schema acquisition and automation and thus needs to be encouraged. To reduce the influence of prior knowledge differences on intrinsic
cognitive load, we tested the subjects’ prior knowledge and used the results to balance the groups in a between-subject design. Mis-matching mental models would increase intrinsic cognitive load. Conceptual description and system practice were used in the training session before the formal test to help the subjects build mental models matching the system to be used. During the formal test, we measured cognitive load factors, such as mental demand, physical demand, and frustration levels, and analyzed them together with user comments and other performance measures to distinguish extraneous cognitive load and germane cognitive load. This approach allowed us to gain a detailed picture of how clustering-based EVSs facilitated the subjects in different information seeking tasks, which was impossible otherwise.

- The study showed that the performance gain from a clustering-based EVS varied for different information seeking tasks. Four typical information seeking tasks, namely browsing, fact finding, information gathering, and revisit (e.g. go back to previously visited webpage), were used in the study. In complex tasks such as browsing and information gathering, the clustering-based EVS encouraged the subjects to explore unknown topics, enhanced their enthusiasm in the exploratory process, reduced their frustration level, and increased their confidence in these exploratory tasks. Meanwhile, the clustering-based EVS caused unnecessary cognitive load in simple tasks such as fact finding.

- The study revealed the importance of visual representation of cluster semantics to clustering-based EVSs. Three test beds were derived from a clustering-based EVS to tease apart the factors of clustering analysis and visual representation of cluster semantics. They were compared against the official New York Times (NYT) website http://www.nytimes.com, a normal news article website. The study showed that the clustering-based EVS outperformed the NYT website only when the clustering results were leveraged by pre-attentive visual representation of cluster semantics.

- A handful of leverage points of clustering-based EVSs have been collected from user comments, such as suggestions on how to meet users’ information needs at different stages of an information seeking process, and how to organize information in a clustering-based EVS.

2 RELATED WORK

User-centered evaluation uses both quantitative measures, such as task correctness and completion time, and qualitative measures, such as in-depth open-ended questionnaires and user comments. It is an important approach for evaluating EVSs [19]. For example, Hearst et al. [8] studied a search interface where open-ended questionnaires were used together with quantitative measures. They successfully learned the usefulness of each feature of the interface and identified leverage points where improvements could be made. Kules et al. [13] studied the change of user tactics when a categorized overview became available. They collected user comments for qualitative analysis in addition to recording quantitative measures. Consequently, they got a whole picture of how the categorized overview worked in shaping user tactics and a set of guidelines for exploratory search interfaces. We conducted a user-centered study with a distinct approach. By measuring and analyzing detailed cognitive load measures together with user comments and other measures, we examined the cognitive process of the subjects to obtain in-depth insights.

Detailed cognitive load analysis has not yet been widely adopted in information visualization system evaluations. Among the few existing studies, Kammerer et al. [11] conducted a user study to evaluate the effectiveness of a new design of exploratory systems. In the experiment, they evaluated the cognitive load of the subjects based on the NASA work load index [6]. Anderson et al. [3] measured brain activities using electroencephalography (EEG) to study cognitive load when comparing multiple visualization techniques. Their results showed that cognitive load measures extracted from EEG data can be used to quantitatively evaluate the effectiveness of visualizations. Kang et al. [2] used the NASA TLX survey to measure user cognitive load in a comparative study between single and multiple monitors. Our work goes beyond the existing work by distinguishing extraneous cognitive load and germane cognitive load through integrated analysis of cognitive load measures and rich qualitative and quantitative measures.

A lot of clustering-based EVSs have been developed. For example, INSPIRE [24] projects documents onto a 2D space so that clusters of documents form galaxies or mountains. The most significant keywords of the clusters are displayed as labels of the galaxies or mountains. PaperLens [14] groups papers by their topics. Each group is displayed in a time histogram and the common topic is displayed as a label. Newsmap (http://newsmap.jp) presents news clusters generated by Google News Aggregator in a treemap style visualization. Each cluster is represented by a rectangle in the treemap. The size of the rectangle indicates the number of documents in the cluster. The title of a representative news article in the cluster is displayed in the rectangle. Google News Timeline (http://newstimeline.googlelabs.com) allows users to view news clusters on a zoomable, graphical timeline. The time stamp, title, abstract, and sometimes a figure thumbnail of a representative news article are displayed for each cluster.

Evaluations have been conducted to assess the performance of clustering-based EVSs. For example, INSPIRE has been tested by a set of analysts [24]. Their reports showed that INSPIRE triggered creative thinking and justified the conviction that text visualizations have to make use of the cognitive and visual process. A formal user study was conducted on PaperLens [14]. It focused on the usability of the system using efficiency and accuracy measures. Sixteen tasks were used to test how PaperLens helped the subjects investigate research topics and their trends over time. Results showed that PaperLens assisted users in exploring the data with less user effort. Pirolli et al. [17] conducted a user study to compare a clustering-based EVS with a simple keyword-based search. The result suggested that the clustering-based EVS induced a more coherent conceptual image of the text collection and a richer vocabulary for con-
structing search queries. Zamir et al. [26] reported an empirical comparison between a standard ranked-list representation and a clustering-based representation for log files. Their study showed that the clustering-based representation influenced the number of documents the subjects read, the amount of time they spent, and their click distance. The purposes of the above user studies were mainly to evaluate the performance of a particular clustering-based EVS. Our evaluation had a different focus, which was to find general guidelines for the design and evaluation of clustering-based EVSs.

3 EVS evaluation: challenges and approaches

Toward a general methodology for EVS evaluation, we discuss several major challenges in evaluating EVSs and propose a set of approaches to address these challenges. The theoretical foundation underlying our approaches is introduced in Section 3.1. Our practice of these approaches in the user study of clustering-based EVSs is presented as examples to illustrate them.

3.1 Understanding the cognitive process

When interacting with EVSs, users are engaged in a sense-making process to bridge a knowledge gap that prevents them from accomplishing the tasks [3] [20]. EVS evaluation needs to assess how EVSs assist users in acquiring new knowledge through learning so that the gap in the cognitive process is bridged. To achieve this goal, Cognitive Load Theory [20], which consists of rich cognitive process models, procedures, and instructions, provides a solid theoretical foundation for EVS evaluation design.

According to cognitive load theory, two parts work together in a user’s cognition during an iterative exploration process: limited working memory and comparatively unlimited long term memory. The working memory is where important learning processes happen, while the long term memory is where users’ knowledge lies, including all the existing schemas. When new information is introduced, it is learned in the working memory to extract schemas for filling the knowledge gap. The schemas are then transmitted to the long term memory and saved. During this process, cognitive load is generated. Sweller [20] distinguished three types of cognitive load according to their sources:

- **Intrinsic cognitive load.** It is caused by the structure and complexity of the material being learned and cannot be influenced by system designers. Intrinsic cognitive load can only be reduced when needed schemas already exist in the long term memory.

- **Extraneous cognitive load.** This load is induced by system designs without sufficient consideration on the structure of information and the cognitive process. It is an overhead that interferes with the understanding of materials.

- **Germane cognitive load.** It represents users’ efforts to process and comprehend the materials. It is devoted to schema acquisition and automation and thus enhances learning.

Understanding the above three types of cognitive load is essential in EVS evaluations. First, users’ prior knowledge in the long term memory affects the intrinsic cognitive load and thus user variability needs to be carefully controlled in an EVS evaluation. Second, both extraneous cognitive load and germane cognitive load are imposed by the system design and vary from system to system. The indications of an effective EVS system are **low extraneous cognitive load and high germane cognitive load**, since the former hinders learning while the latter enhances learning. Therefore, it is critical to distinguish these two types of cognitive load during an EVS evaluation. We designed our evaluation following the above guidelines.

3.2 Conducting EVS evaluations in laboratory settings

The ideal evaluation approach for EVSs is longitudinal and in a naturalistic setting, because EVSs are often used in the context that is open-ended, progressive, and iterative. However, it is often costly and not efficient enough for system development, especially for the initial step of system design. In this paper, we focus on user-centered evaluations that combine controlled lab experiments with questionnaires and interviews. In particular, we discuss how to design tasks, control user variability, and motivate subjects in laboratory settings.

3.2.1 Task design

To reconstruct the multi-faceted exploration process in the laboratory setting, we suggest using existing task taxonomies to guide the task design, since they well summarize users’ activities in certain domains.

**Practice:** The goal of our user study was to better understand how clustering-based EVSs assist users in an information seeking process. The tasks used in this study were designed based on a task taxonomy of high level web information seeking [12]. We believe that it can be easily extended to cover a large portion of information seeking tasks for other kinds of document collections. In this taxonomy [12], there are five categories of information seeking tasks, namely “browsing”, “fact finding”, “information gathering”, “transactions”, and “others”. We adapted the taxonomy and applied it to an online news collection information seeking scenario. In particular, we excluded “transaction”, which happens in E-commerce, since there is no transaction in news data. We chose “Revisit” from the “others” category, since it happens frequently when readers want to retrieve what they met before. In the user study, one task for each category was used, namely “browsing” (Task 1), “fact finding” (Task 2), “information gathering” (Task 3), and “revisit” (Task 4). Detailed descriptions of the tasks are provided in Table 1.

3.2.2 User variability control

Comparing multiple systems is a common practice in laboratory evaluations. To do so, either within-subject evaluation or between-subject evaluation can be used. Here we focus on between-subject evaluation. It avoids the learning effects since each subject only uses one system. However, it is influenced by the individual differences of the subjects, such as their prior knowledge, mental models, and demographic profiles [10]. Thus, user groups should be balanced. In the follow-
ing sections, we discuss user variability control on prior knowledge, mental model, and writing skill.

Prior knowledge Prior knowledge can affect subjects’ performance since schemas in long term memory can reduce intrinsic cognitive load. Unlike demographic profiles which can be collected using surveys, prior knowledge is not that easy to access. We design the prior knowledge test by bringing ideas from the Education domain. In Education practice, teachers gauge students’ prior knowledge when they enter a course or a program. There are several ways to evaluate student prior knowledge, such as prior knowledge self-assessments, concept maps, and concept tests [1]. We use the self-assessments since they are easy to conduct and score and have been proven effective in previous research [11].

In a self-assessment, a subject is asked to reflect and rate her/his level of knowledge and skill. A potential issue is that the subject may not be able to accurately assess their knowledge. However, accuracy can be improved when the questions clearly differentiate levels of knowledge. Identifying concepts and techniques that are needed in the exploration process can be of great help in generating effective questions. For concepts, we suggest five levels: never heard of $\rightarrow$ have heard of $\rightarrow$ have some idea $\rightarrow$ have a clear idea $\rightarrow$ can explain it. Also, we suggest five levels for techniques: never used it $\rightarrow$ have tried using it $\rightarrow$ can do simple interactions $\rightarrow$ can manipulate multiple functions $\rightarrow$ can easily use it and build results.

Practice: A between-subject design was used in our study. In particular, 36 subjects were assigned to four groups; one for each test bed. Each group had 9 subjects. The subjects had a prior knowledge test before the evaluation was conducted. The results were used to balance the groups. The prior knowledge test included 20 questions on concepts and 4 questions on techniques. The 20 concept questions were about the five most significant news stories in the news data (including the task-related ones), 4 for each. The questions included general ones, such as “How would you rate your knowledge regarding Haiti Earthquake?” and specific ones, such as “How would you rate your knowledge regarding humanitarian aid in Haiti Earthquake?”

The technique questions evaluated the subjects’ prior knowledge on computer usage, browsing and searching experience, database system usage, and information visualization system experience. For example, the question on “information visualization system” was “How familiar are you with information visualization systems?” All the questions were answered using the five level scales as described above. The test was scored 1- 5 (5 for highly knowledgeable). The subjects were sorted by their test scores from high to low, and assigned to four groups using a “Z” style. For example, the subjects ranked 1-12 were assigned to the groups as follows: Group 1: subjects ranked 1st, 8th, and 9th; Group 2: subjects ranked 2nd, 7th, and 10th; Group 3: subjects ranked 3rd, 6th, and 11th; Group 4: subjects ranked 4th, 5th, and 12th. After the initial assignment, the average scores for each group were calculated and minor exchanges were made to balance the groups.

The correlation between prior knowledge and user performance was analyzed to learn whether the clustering-based EVS compensated prior knowledge shortage of novice subjects.

Mental models Mental models are the “psychological representations that aid in understanding, explaining, or predicting how a system works” [9]. A matching mental model can greatly enhance a user’s experience with a system while a mismatching mental model can bring unnecessary mental barriers into a user’s exploratory process. Although it is interesting to study how the different mental models affect the effectiveness of an EVS, the effects of different mental models need to be controlled to reduce their influence on intrinsic cognitive load when they are not the focus of a study. To do so, the evaluation designers need to: (1) make sure that subjects understand how the system works; and (2) minimize the differences in the mental models that the subjects have already built. We suggest conducting mental model control in the training process before the formal testing, which is effective with low cost. Conceptual description and system practice [7], which have been proved effective in mental model building, can be employed.

Practice: In our study, each subject had a training session before the formal testing. In the training session, an instructor first introduced the system to the subject (conceptual description), including how the system processed the raw data, how the outcome should be interpreted, and how the interactions were processed by the system. Then, a training dataset was given to the subject. The subject was asked to freely interact with the system (system practice) and encouraged to talk to the instructor about each step she took, such as her thoughts about the interface, the aim of the interaction, and her prediction of the behavior of the system reacting to her interaction. In this way, the instructor observed how the subject learned the system and corrected her when her understanding of the system was inaccurate. Comments were recorded by the instructor for system improvement. When the subject felt ready for the formal testing, she was asked to describe the working process of
the system and how the system responds to different interactions. Again, misunderstandings were clarified when needed. The goal of such “exit description” was to make sure that the subject’s mental model matched the system to be tested.

Writing skills In EVS evaluations, the subjects are often required to summarize or describe information they collected. The writing skills of the subjects vary. The differences can affect the result assessment and thus need to be controlled since they are irrelevant to the system design. However, assessing the subjects’ writing skills is difficult and assigning groups based on the writing skills further increases the complexity of group construction. We propose a practical alternative: when assessing results whose qualities are affected by writing skills, ask the judges to explicitly rate the quality of writing besides other metrics. In this way, the influence of individual differences on writing skills can be separated from the influence of the system.

Practice: The information gathering task required the subjects to summarize the information they collected. We evaluated the results based on the information quality standards of Wikipedia (http://en.wikipedia.org/wiki/Information_quality). To simplify the evaluation process, three measures were generated from these standards. They are accuracy, completeness, and quality of writing. The accuracy measure indicated whether the information was true or not. The completeness measure indicated whether a summary covered all aspects of the whole story. The quality of writing measure captured all other metrics which depend on an individual’s writing skills.

3.3 Measurements

3.3.1 Time

Subjects’ completion time when conducting a task is often used to measure system performance. For exploratory tasks, completion time can be affected by factors other than the effectiveness of the EVS being evaluated. For example, curious subjects may spend a significant amount of time to read the raw data during the experiment. The evaluation designer should either limit such activities or exclude the time for such activities from result analysis.

Practice: A measure used in our study was immediate completion time. It recorded the time the subjects first found information that was useful for them. It was obtained either by self-reporting during the exploration or retroactively by examining relevant screen captures.

Sometimes capturing immediate completion time is not practical for open-ended exploratory tasks. Also, the subjects are often overly thorough in the test situation [10]. In these cases, posing time limitations to the tasks is a good alternative. To reduce the possible stress caused by time limitation, M. Kaki et al. [10] suggest using instructions such as “whatever can be found in the given time is acceptable”. In addition, a suitable time limitation which gets from the transaction log analysis or an average task time from pilot study can make the subjects’ behavior closer to the real behavior. Furthermore, we suggest allowing the subjects to exit open-ended tasks, such as browsing, at anytime they want. Thus the time the subjects spend on the tasks, named exploration time, can be analyzed together with the subjects’ comments to further learn about user engagement or frustration.

3.3.2 Measuring extraneous and germane cognitive load

As we discussed in Section 3.1, low extraneous cognitive load and high germane cognitive load indicate effective EVSs. Therefore, it is important to distinguish these two types of cognitive load. There are lots of direct or indirect cognitive load measurements in cognitive load theory, such as the NASA task load index questionnaire [6] and the direct measurement “dual-task approach” [4]. A good description of existing cognitive load measurements can be found in [16]. However, most measures, such as the cognitive load factors on mental demand and physical demand, do not differentiate extraneous and germane cognitive load. In other words, a high cognitive load measure can be either caused by high extraneous cognitive load or high germane cognitive load. Therefore, in EVS evaluations, it is important to analyze cognitive load measures together with the subjects’ comments and other qualitative and quantitative measures to distinguish extraneous and germane cognitive load.

The comments are obtained from open questions, post questionnaires, and interviews. To reduce the complexity of cognitive load analysis, we suggest classifying comments into three categories, namely “Engagement”, “Neutral” and “Frustration”. “Engagement” comments express excitement and encouragements when interacting with the system or describe how the subjects are motivated to explore more information and put more effort into the exploration process. Examples are comments such as “The keyword murder caught my eyes. Okay, let’s see what happened here” and “The information is well organized and already there. I would like to see more about what it will get.” “Neutral” comments describe how the subjects deal with the task, such as the comment “I put in the keywords and search for what I need.” “Frustration” comments report unexpected situations the subjects experience and difficulties they meet. Examples are comments such as “I need to go through each news article. I do not see a place to highlight what I want. Frustrating” and “Associating between different news articles costs me lots of time.”

The high values of certain cognitive load measures, such as the mental demand factor, indicate high germane cognitive load when they appear together with comments whose majority are “Engagement” comments. On the contrary, their high values indicate high extraneous cognitive load when they appear together with comments whose majority are “Frustration” comments.

Practice: In our study, we measured the subjects’ cognitive load after they conducted each task. We used a modified version of the NASA task load index questionnaire [6], which was designed to identify the cognitive variations in subjective workload within and between different types of tasks. Six workload-related factors, namely mental demand, physical demand, temporal demand, performance, effort, and frustration level, were measured using 5-point scale questionnaires to derive a sensitive and reliable estimation of workload. The detailed information is shown in Table 2.

User comments were collected in two ways. First, after each task, we asked the subjects to write down the challenges they met in the task and how they solved them. Second, we in-
4 THE USER STUDY

We conducted a formal user study on a clustering-based EVS, using the EVS evaluation methodology discussed in Section 3. Through this study, we explored the benefits and constraints of clustering-based EVSs when conducting different information seeking tasks. We also explored important design aspects that can affect the effectiveness of a clustering-based EVS, namely influence factors.

4.1 Test Beds

4.1.1 Influence Factors

To identify influence factors, we first examine existing clustering-based systems and identified several features that distinguish them from other document visualization approaches. First, a clustering-based EVS always groups a large collection of documents into clusters by some means. For example, PaperLens [14] groups papers by their topics and Newsmap (http://newsmap.jp) presents news clusters generated by Google News Aggregator. Second, clustering-based EVSs often provide visual representation for semantic information derived from the clusters. For example, labels in INSPIRE reveal the most significant keywords of the clusters [24]. Two influence factors are identified here: clustering analysis and visual representation of cluster semantics.

We posed the following questions in this study: do clustering-based EVSs really help users explore large document collections? Does the fact that relevant documents are organized into clusters alone lead to the advantages of clustering-based EVSs? How important is the visual representation of cluster semantics to clustering-based EVSs? The answers to the above questions can effectively guide further exploration of clustering-based EVSs' design space.

To answer these questions, we built three test beds teasing out clustering analysis and visual representation of cluster semantics. Since the most traditional way of news exploration is web searching, the New York Times (NYT) website (www.nytimes.com) was used as a baseline system in the evaluation. We compared the performance of the four test beds using a variety of information seeking tasks.

4.1.2 Customized Test Beds

Beside the baseline system, we created three test beds from Newdle (see Fig. 1) [25]. Newdle is a clustering-based EVS created recently for interactive exploration of large news archives. Systems such as NewsMap (http://newsmap.jp) and Google News Timeline (http://newstime.googlelabs.com) were not selected since they have a strong bias toward the most recent documents in the collections, which is beyond the scope of this study. Other reasons why we chose Newdle were that there were clear boundaries among the clusters in Newdle and that a semantic representation was explicitly provided for each cluster. These features made it easier to tease out the clustering analysis and semantic representation factors in Newdle than in other systems such as INSPIRE [24].

Newdle has the following components:

**Clustering Analysis.** Newdle clusters New York Times (NYT) (http://www.nytimes.com/) news articles based on their tags. The tags are manually generated by NYT editors and thus have a high quality. A document network is constructed where two documents with more than three shared tags are connected. Clustering is conducted using leading eigenvector community analysis [15] based on the network structure. We manually inspected the clusters visualized in the user study. The majority of them consisted of closely related documents. Noise (a small number of not so related documents) existed in some clusters. For example, the news cluster about the Toyota recall event might accidently contain one article about the spokesperson in its advertisement. However, such noise is small in most clustering algorithms.

<table>
<thead>
<tr>
<th>Factors</th>
<th>End Points of Scale (1/5)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Demand (MD)</td>
<td>Low / High</td>
<td>How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?</td>
</tr>
<tr>
<td>Physical Demand (PD)</td>
<td>Low / High</td>
<td>How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)?</td>
</tr>
<tr>
<td>Temporal Demand (TD)</td>
<td>Low / High</td>
<td>How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?</td>
</tr>
<tr>
<td>Performance (OP)</td>
<td>Good / Poor</td>
<td>How successful do you think you were in accomplishing the goals of the task set by experimenters? How satisfied were you with your performance in accomplishing these goals?</td>
</tr>
<tr>
<td>Effort (EF)</td>
<td>Low / High</td>
<td>How hard did you have to work (mentally and physically) to accomplish your level of performance?</td>
</tr>
<tr>
<td>Frustration Level (FR)</td>
<td>Low / High</td>
<td>How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?</td>
</tr>
</tbody>
</table>
Fig. 1. (a) Topic overview of Newdle. (b) Search result display of Newdle. (c) Clustering-only system. Each cluster is displayed in a document list. (d) New York Times website. (e) Plain system. All documents are displayed in one list ordered by time stamps.
**Topic Canvases.** Newdle pre-attentively represents the most shared tags in a cluster in a rectangular area named “topic canvas”, as shown in Fig. 1 (a). These tags provide a high level overview of the semantics of the news articles in this cluster. Thus, the topic canvas provides visual representation of the cluster semantics. In a topic canvas, the most shared tags are displayed using Wordle, a tag cloud like visualization that packs a large number of tags with varying font sizes and colors into a small screen space [22]. The colors of the tags represent their categories assigned by NYT editors, such as people, organizations, locations, and topic descriptors. For example, all the location tags are in yellow and all the person tags are in white in Fig. 1 (a). The size of a tag is proportional to the number of articles with it within the cluster. The overlaid line graph on a topic canvas is a time graph revealing the daily number of articles in this cluster.

**Document Lists.** Besides topic canvases, Newdle also presents snippets of documents within a cluster using a list, where the documents are ordered by their time stamps in descending order (see Fig. 1 (c)). The snippets can be displayed in a detail mode or a compact mode. In the detail mode, the titles, tags, summaries, time stamps, and authors of the documents are displayed. In the compact mode, only titles are displayed. Note that the titles are good indicators of the content of NYT news articles since they are carefully chosen. Users can switch between the two modes. Clicking on a title leads the users to the full text. The colors of the titles range from blue to white, indicating the age of the documents.

**Multiple Views.** Newdle provides an overview and a detail view. The overview allows users to browse the major clusters in the collection. The detail view allows users to examine clusters of interest. To generate the overview, the clusters are sorted by the number of documents. Topic canvases of clusters are displayed in a grid, as shown in Fig. 1 (a). Users can learn the semantics of the clusters at a glance and quickly drill down to clusters of interest. After users select clusters of interest from the overview, a detail view of the selected clusters are generated. In the detail view, the topic canvases and document lists of the selected clusters are displayed side by side, as shown in Fig. 1 (b). Users can open a document (a news article) by clicking its title in the document list.

**Interactions.** Newdle allows users to conduct in-depth analyses on clusters, tags, and documents through a rich set of interactions. From the overview, users can select clusters containing a set of tags of interest or clusters related to a focus cluster. The search results are visually presented in the detail view. From the detail view, users can open a document and examine its details from the document list.

Three test beds were derived from Newdle to tease out the visual representation and the clustering analysis. The first test bed kept the basic features of Newdle, namely the clustering analysis, the document lists, and the topic canvases. We refer to it as Newdle. The second test bed was derived from the first bed by removing the topic canvases. The most recent news articles are highlighted in the document list as labels for each cluster (see Fig. 1 (c)). We call it the clustering-only system. The third test bed was a plain system without clustering analysis. All documents were displayed in one document list, where they were ordered by the time stamps in descending order. Fig. 1 (e) shows the plain system. In addition, the NYT website was used as a baseline system. Table 3 summarizes the features and interactions of all four test beds used in the study.

By comparing Newdle against the plain system and the NYT website, we evaluated the performance gain of clustering-based EVSs. By comparing Newdle and the clustering-only system, we examined the performance gain from the visual representation of cluster semantics. Comparing the clustering only system against the plain system and the NYT website allowed us to assess the performance gain by grouping relevant documents into clusters.

### 4.2 Data

The data used in the user study were 3,640 news articles fetched from the New York Times (NYT) online RSS feeds in January 2010. We set constraints in the NYT website using the advanced search function to make sure that the subjects using the NYT website accessed the same set of data as the other subjects.

Since online news is a typical text source in huge data volumes, we believe insights from this study can be extended to text exploration in many other domains, such as digital libraries and archived reports.

### 4.3 Subjects

Thirty-six UNCC students (17 male, 19 female) participated in the user study. The ages of the subjects ranged from 20 to 28 years old. Eighteen of them were Computer Science majors. Twelve students were Communication majors. Four were Electronic Engineering students. The other two subjects were a Mathematics major and a Chemistry major. All subjects reported at least one year of computer experience. One week before the study, invitation emails were sent out with an informed...
consent form and a prior knowledge test. Subjects who wanted to join the study replied to the email with the signed consent form and completed prior knowledge test. The test had a maximum score of 120. There were 5 subjects whose scores were within the range 62~58, 24 subjects within the range 55~50, and 5 subjects within the range 50~46. The subjects were assigned to four groups according to their prior knowledge test scores using the method described in Section 3.2.2. Each group worked on one of the four test beds in the user study.

The average prior knowledge test scores for Newdle, Clustering-only, NYT, and plain group were 53.875 (Highest:62, Lowest:46), 53.875 (Highest:61, Lowest:48), 53.5 (Highest:58, Lowest:48), and 53.625 (Highest:58, Lowest:48). The standard deviations of the four groups in the same order were 4.55, 3.91, 3.07 and 3.50.

Twenty-six subjects were native speakers of English; the other ten subjects also spoke and wrote fluent English. The numbers of native speakers in Newdle, clustering-only, NYT, and plain systems were 6, 7, 6, and 7 respectively.

4.4 Procedure

The subjects took the study in a laboratory setting one by one. The study lasted about two hours for each subject. Before the study, the instructor explained the goals of the user study to the subject. The background information about extraneous cognitive load and germane cognitive load was explained. The subject was informed to pay attention to the cognitive load she/he experienced during the experiment.

A twenty minute training session was first conducted. The capabilities of the test bed and the upcoming tasks were explained to the subject by an instructor who supervised the experiments of all the subjects. Details of the training session were introduced in Section 3.2.2.

The test session followed the training session. The subject was asked to conduct the four information seeking tasks presented in Table 1, one by one using the test bed. 15 minutes, 5 minutes, and 5 minutes were given to the browsing, fact finding, and revisit tasks, respectively. There was no time limit on the information gathering task and the completion time and the immediate completion time were recorded. All the questions were given in a text editor. Subjects’ screen activities were recorded. After each task, the subject answered the cognitive load questionnaire (see Table 2) and open-ended questions, such as how they felt about the system and how they rated the cognitive load ratings. This session was concluded with an interview about the user experience with the system and a subjective questionnaire (see Table 8).

5 RESULTS

The collected data include (1) prior knowledge test scores; (2) effectiveness measures; (3) cognitive load measures; (4) classified comments; and (5) subjective ratings. They were analyzed together as follows:

1. The correlation between the prior knowledge test scores and the effectiveness measures was calculated to learn whether a test bed can compensate knowledge shortage in novice users.
2. The cognitive load measures were analyzed together with the classified comments and effectiveness measures to reveal the strengths and types of the cognitive load experienced by the subjects.

In the following, we report the results for each of the four tasks: browsing, fact finding, information gathering, and revisit.

5.1 Task 1: Browsing

Task: Find as many distinct topics from the dataset as possible. Describe each of them using a few sentences.

<table>
<thead>
<tr>
<th></th>
<th>Newdle</th>
<th>Clustering-only</th>
<th>NYT</th>
<th>Plain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness</td>
<td>♯ of Topics</td>
<td>5(0.67)</td>
<td>4(0.71)</td>
<td>4(0.60)</td>
</tr>
<tr>
<td>Exploration Time (min)</td>
<td>11.39(3.51)</td>
<td>8.13(1.06)</td>
<td>10.86(2.72)</td>
<td>7.62(1.14)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comments</th>
<th>Category</th>
<th>Subjects</th>
<th>Mentions</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newdle</td>
<td>Engagement</td>
<td>9</td>
<td>20</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>9</td>
<td>15</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>6</td>
<td>10</td>
<td>1.66</td>
</tr>
<tr>
<td>Clustering-only</td>
<td>Engagement</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>9</td>
<td>26</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>NYT</td>
<td>Engagement</td>
<td>4</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>9</td>
<td>20</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plain</td>
<td>Engagement</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>9</td>
<td>17</td>
<td>1.88</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>7</td>
<td>8</td>
<td>1.14</td>
</tr>
</tbody>
</table>

From the cognitive load analysis (as shown in Fig. 2 and Table 4), we noticed the order of mental demand (MD) is: Newdle > Clustering-only > Plain > NYT. Newdle users had the highest MD. NYT users had the lowest MD. Clustering-only system users had slightly higher MD than base line system users. In contrast, NYT users had the highest physical demand (PD). Newdle and the clustering-only system have the same lowest PD. The time demand (TD) exhibited a different trend. NYT
and Newdle users (NYT > Newdle) had higher TD than the other two systems (Plain>Clustering-only). Newdle users had more confidence than users of the other three systems. The plain system users had the lowest confidence in their answers. NYT users gave the highest rating of effort (EF). Newdle users had the second highest EF rating. Surprisingly, the plain system users had the lowest EF rating. The order of frustration level was opposite to EF: NYT < Newdle < Clustering-only < Plain.

Table 4 shows Newdle users had the largest number of “Engagement” comments among the four test beds. For example, there were comments such as “The keyword cloud is interesting. I realized some news that I did not know before and explore them. Funny.” The comments suggested that Newdle users had high MD since they were intrigued to explore more unknown news topics and became more engaged in the exploratory process. Clustering-only and NYT users had more comments related to “Neutral”. For NYT, four users commented on the “Engagement” while clustering-only had only one user. The comments showed that NYT users liked the website in their exploration. They enjoyed navigating through hyper links on the webpages. This explained the high physical demand for NYT users. Users commented that “There is no challenge. I am familiar with this kind of news website. Most of my time was moving from one page to another and got what I need.” The plain system had almost the same number of comments in the “Neutral” and “Frustration” categories. According to the comments, it was hard to associate different news articles using the plain system.

The browsing task’s effectiveness was measured by “% of topics” and “exploration time”. As shown in Table 4, on average, Newdle users found 5 news topics; clustering-only and NYT users found 4 news topics, respectively; and the plain system users only found 3. “Exploration time” grew as the “% of topics” increased. It indicated how the subjects were engaged in this free exploration task. Four Newdle users and three NYT users reached the 15 min time limit. All the clustering-only and plain system users exited before 15 mins. Together with the cognitive load data and comments, these two effectiveness measures suggested that Newdle and NYT users were more engaged in their task than clustering-only and plain system users.

The correlations between prior knowledge and “% of topics” were analyzed using t-tests. The correlations were positive for NYT users \((r = .51, p < 0.05)\) and the plain system users \((r = .42, p < 0.05)\). In contrast, there was no positive correlation for Newdle users and the clustering-only system users (Newdle: \(r = -.55\); Clustering only: \(r = -.32\)). The correlation analysis between the effectiveness measure and the prior knowledge of subjects suggested that the performance difference between novice users and experienced users in the browsing task can be reduced by clustering-based EVSs.

### 5.2 Task 2: Fact Finding

**Task:** Find as many articles as possible about humanitarian aid in the Haiti earthquake. Save the links of each article.

Overall the cognitive load of task 2 was low (as shown in Fig. 3). All the ratings were below 3. There were no big differences among the four test beds in mental demand (MD). NYT and the plain system users’ MD ratings were slightly higher. On the other hand, Newdle and clustering-only system users had slightly higher physical demand (PD). Clustering-only and plain system users experienced higher time demand (TD) than Newdle and NYT users. Newdle users were the most confident in their performance while the plain system users were the least confident. The performance ratings were almost the same between clustering-only and NYT users. NYT users experienced the lowest frustration level in task 2.

Most comments were in the “Neutral” category (see Table 5). For NYT, one user expressed engagement in this task. Other systems did not have comments in “Engagement”. According to the comments, most NYT and the plain system users worked on identifying relevant news articles from the returned search results and refining their search strategy. Most Newdle and the clustering-only system users benefited from the clustered results and thought “It is clear and organized well”.

“Interactive recall” and “Interactive precision” [21] were used to measure effectiveness of task 2. Interactive recall measures the fraction of the documents that are relevant to the query...
and successfully found. Interactive precision states the fraction of found documents that are relevant to the users information need. As shown in Table 5, all the four test beds had 1 on “interactive recall”, namely that all the subjects were able to find relevant articles. Newdle and the clustering-only system users had higher “interactive precision” than NYT and the plain system users. This suggested that NYT and the plain system users included more unrelated articles in the answer.

Correlation analysis was conducted between “interactive precision” and prior knowledge. It showed a positive correlation for all the four test beds in this task ($r > 0.5, p < 0.05$).

### 5.3 Task 3: Information Gathering

**Task:** Summarize the activities of President Obama on Human Health Insurance.

Fig. 4 showed that the trends of mental demand, physical demand, performance, effort and frustration among the four test beds were similar: Newdle < Clustering-only < NYT < Plain. The task did not have a time limit, so the time demand ratings of all four test beds were low. The plain system users experienced the highest frustration level. Newdle users had the lowest frustration level.

Table 6 showed that for NYT and Newdle, there are 2 users who provided “Engagement” comments for either system. Newdle users described their excitement about quickly locating the information they needed. NYT users liked the highlighting of searched keywords in the result list. For NYT, 9 users gave “Frustration” comments as opposed to 3 for Newdle. NYT users provided comments such as “Lots articles are returned to me. I can see why they are returned, but how are they related? It is hard to form a story.”. “It is terrible to see so many articles that I need to summarize. Organizing them and finding association among them cost me most of time”, and “sometimes I am lost when jumping through links in the article. Too much information”. This together with the cognitive load measures suggested that NYT users had higher extraneous cognitive load than Newdle users.

It seemed that the semantic representation in Newdle reduced extraneous cognitive load. For the clustering-only system, no subjects gave “Engagement” comments, 5 subjects expressed “Frustration”, and the clustering-only system’s cognitive load measures were higher than those of Newdle. Some clustering-only system users complained that it was not easy to understand the relationship between the searched keywords and the returned results and extra reading was needed.

The plain system received the largest number of “Frustration” comments and the highest cognitive load measures. Obviously it caused the highest extraneous cognitive load on the subjects. According to the comments, the reason might be the lack of cues on the associations among the news articles.

The effectiveness measures (see Table 6) were mostly consistent with the cognitive load analysis. Newdle users had much shorter immediate completion time than other users, namely that they were able to locate desired information much faster than other users. The task completion time’s order was: Plain < Newdle < NYT < Clustering-only. However, the plain system users got the lowest overall score on result quality. From the cognitive load analysis above, we knew that plain system users had the highest frustration. These together might suggest that plain users were so frustrated that they exited the task much quicker. The detailed quality measures (see Fig. 5) indicated that the results from Newdle users were more accurate and complete than that from other subjects. According to Fig. 5 and the immediate completion time, the clustering-only system
performed better than NYT and the plain system, but there was a significant performance boost from the clustering-only system to Newdle, which indicated the importance of the visual representation of cluster semantics. Overall, Newdle outperformed the other three test beds in this task.

Correlation analysis was conducted between result quality and prior knowledge. It showed a positive correlation for the plain system ($r = .65$, $p < .05$) and no strong correlation for other three test beds ($r < 0$, $p < .05$).

### 5.4 Task 4: Revisit

**Task:** List as many keywords as possible that can be used to retrieve articles in task 3.

In this task, the plain system and Newdle had slightly higher mental demand and frustration levels than the other test beds (see Fig. 6). According to Table 7, no user expressed “Engagement” in this task. Five Newdle users expressed their feelings of “Frustration”. The comments showed that they tried to remember the keywords in the “topic canvas”, which resulted in unnecessary cognitive load.

The effectiveness of task 4 was measured by the number of keywords listed. As shown in Table 7, NYT users performed better than the other users. Correlation analysis showed that effectiveness and prior knowledge were positively related for all the test beds ($r > 0$, $p < .05$).

### 5.5 User Preferences

Fig. 7 showed the subjects’ answers to the post study subjective questions, which are listed in Table 8. Newdle users were more enthusiastic on using the system to conduct “Information gathering” and “Browsing” tasks than other users (Q9 and Q10). They also liked the presentation and structure of the visualization (Q1), and thought it was easy to navigate in the news collection using the system (Q2). It seemed that Newdle and the clustering-only system required more training than the other systems (Q4). According to the subjects, NYT was the best system among the four test beds for querying and fact finding (Q3 and Q8). NYT also received a high preference rating for using the system to conduct browsing tasks. Interestingly, NYT had better ratings than the clustering-only system on all the questions, while Newdle had better ratings than NYT on most questions. These results revealed the importance of visual representation in clustering-based EVSs.

### 6 Insights and Leverage Points

#### 6.1 Insights for EVS evaluation

The most important contribution of this paper is applying cognitive load theory, especially Sweller’s three types of cognitive load theory [20], to guide the design and result analyses of EVS
evaluation. We analyzed detailed cognitive load measures together with classified user comments. This approach measured and distinguished germane cognitive load and extraneous cognitive load, the former being beneficial while the latter not. For example, the high mental demand experienced by Newdle users in the browsing task, together with lots of comments about engagement, indicated that Newdle encouraged the subjects to explore more information, which was positive. On the contrary, the high mental demand experienced by the plain system users in the information gathering task was negative, and triggered many comments of frustration. Subjects indicated that the plain system hindered their information gathering task by increasing their extraneous cognitive load. Without linking cognitive load measures and qualitative comments, such in-depth knowledge about the exploration process would have been hard to obtain. We recommend using this systematic yet practical approach to evaluate subjects’ cognitive process in EVS evaluations.

6.2 The role of clustering-based EVSs in information seeking

The study revealed that Newdle, the clustering-based EVS, performed better than the other test beds in the browsing and information gathering tasks. These two tasks were typical exploratory tasks and shared the following features: (1) there was a large amount of information to be accessed; (2) subjects had vague targets before starting the visual exploration; and (3) subjects had drifting information need during the exploratory process.

In the browsing task, Newdle triggered the subjects to explore more information, which enhanced their learning. NYT website offered lots of multimedia resources like pictures and videos. Such visualizations, together with the hyper links provided, also triggered the subjects to explore information. The performance of the clustering-only system was superior to the plain system while inferior to Newdle. It indicated that clustering helped information seeking, while its results were much more useful when they were leveraged by visualization, such as the pre-attentive word cloud provided by Newdle.

Information gathering was the most complex task among the four tasks. Subjects not only needed to find the information, but also had to associate, digest, and represent the information. The results showed that the performance gain from using Newdle was the most significant in this task. It encouraged the subjects to explore unknown topics, increased their enthusiasm in the exploratory process, compensated their shortage of prior knowledge, reduced their frustration level, and increased their confidence in the complex task. We thus recommend using clustering-based EVSs in applications where complex tasks are frequently conducted.

Meanwhile, clustering-based EVS designers need to carefully consider how to assist simple tasks such as fact finding. Although there was no big difference among the four test beds in the fact finding task, Newdle users did have a higher frustration level than NYT users. It needs to be studied how to give users appropriate information for the task at hand without overwhelming them with unnecessary information. This is consistent with a previous study in [5], which showed that visualization is a superior interface for complex, spatial, and inferential learning, and not so much the case for hunt-and-find simpler tasks.

6.3 What users need at different stages of information seeking

The detailed examination of the comments revealed that users had different information needs at different stages of an information seeking process:

- At the beginning of an exploration, an overview is of great help. Users appreciated the fact that information is well organized and a global picture of an article collection is provided in Newdle. According to user comments, “organization” was the most essential requirement at this stage. Users would like to see unrelated information separated, core information highlighted, and relationships between different information units easy to access. Designed in such a way, clustering-based EVSs can help users get an overview of a large document collection with reduced efforts.

Newdle offered two types of overviews to users (as shown in Fig. 1). At the end of the training section, we asked the subjects which one they would like to use as the starting point. All the subjects selected the second one (Fig. 1 (b)). They thought the first one (Fig. 1 (a)) was “kind of overwhelming”, “too much information”, and “busy for eyes”. In contrast, the list view in the second one made the visualization more “organized” and “understandable” to them. However, after they used the system for a while and become familiar with the system, most of them began to make use of the first view. In the interview, they explained that “At first, I like the one with list because it seems more organized to me. However, after playing with the system, the other view is not that overwhelming and I can see more information when I want.”. “Because I would like to see more news events at one time and I know what are under those tags. It is compact.”, and “Ever since the last task, using the visualization became easier, I grasped it quicker.” These comments suggested that the ability of users to digest information is evolving in the learning process. EVS design can take it into consideration. A visualization with raw data presented in the way users are most familiar with is easier to understand without overwhelming novice users. During the learning process, options can be given to the users so that they can switch to a compact view with more information.

- During the exploration, users interacted with the system. At this stage, the most desired feature of the system was to emphasize the “association” between the subjects’ actions and the results returned to them. For example, the clustering-only system frustrated users because they needed to read the article titles to evaluate the relationship between the returned results and the query in Tasks 2 and 3, whereas the search keywords were not highlighted in the results. On the other hand, NYT users in our experiment appreciated the highlighting of query keywords in the returned results.
The subjects provided lots of useful feedback on the information organization in clustering-based EVSs. They are summarized in the following sections.

### 6.4.1 Revealing relationships among clusters
The study showed that the relationships among the clusters, if any, should be explicitly presented to users. Our experiment subjects called for better organization of the clusters in Newdle. A subject commented "It is difficult to see the relationships between different topics. For example, I saw several clusters about Toyota recall. But it is hard to know what the relations are between these Toyota news." User feedback indicates the need for better layout of the clusters in the design space to explicitly convey the relationships among them. We also suggest providing interactions to allow users to compare and associate clusters.

### 6.4.2 Using document lists
Interestingly, in this study, our subjects liked the view in Newdle where the topic canvases and the document lists were displayed side by side. It seemed that although the cluster semantic representation helped users locate topics of interest, there still was ambiguity in the representation. The document list helped diminish this ambiguity by providing detailed information. In the document list, we provided document snippets such as titles, tags, and summaries. Note that clustering-based visualization systems often provide other snippet information such as the logo of the information source, images, or even a few sentences from the documents. Most users commented that the snippets were useful. All the users agreed that the titles were very useful. In addition, the document list allowed the subjects to quickly access individual documents, which was preferred by the subjects.

### 6.4.3 Organizing documents within a cluster
Our experiment subjects expressed interests for the hottest, oldest, and latest news in a cluster. This suggested clustering-based visualization systems could provide flexible ways to organize the documents within a cluster to support different tasks. For example, they can be grouped by key persons or locations and ordered by time stamps, hotness, and similarities. Their effectiveness for different exploration tasks needs to be further explored and evaluated.

# 7 Conclusion and future work
In this paper, we propose a systematic approach to evaluating EVSs and present our utilization of this approach in evaluating clustering-based EVSs. EVSs go beyond truth finding and present our utilization of this approach in evaluation. In the document list, we provided document snippets such as titles, tags, and summaries. Note that clustering-based visualization systems often provide other snippet information such as the logo of the information source, images, or even a few sentences from the documents. Most users commented that the snippets were useful. All the users agreed that the titles were very useful. In addition, the document list allowed the subjects to quickly access individual documents, which was preferred by the subjects.

### 7.1 Revealing relationships among clusters
Our user study provided in-depth insights about how clustering-based EVSs worked. Our results indicated the benefits, limitations, and circumstances of clustering-based EVSs in supporting different information seeking tasks. The results showed that clustering-based EVSs can benefit complex information seeking tasks such as browsing and information gathering. Insights and leverage points from our study provided promising direction on clustering-based EVS design.

Moving forward, we will further explore the design space of clustering-based EVSs. For example, we plan to evaluate how the performance of clustering-based EVSs is affected by differences in the methods and quality of the clustering analysis as well as different visual representation of the cluster semantics. We will also apply this methodology in the evaluation of other exploratory visualization systems, such as graph visualization systems and multivariate visualization systems.

### References


