2015

SenseCluster for exploring large data repositories

Yousra Harb
Dakota State University

Surendra Sarnikar
Dakota State University

Omar F. El-Gayar
Dakota State University

SenseCluster for Exploring Large Data Repositories

Yousra Harb
Dakota State University,
Madison SD
yaharb@pluto.dsu.edu

Surendra Sarnikar
Dakota State University,
Madison SD
surendra.sarnikar@dsu.edu

Omar El-Gayar
Dakota State University,
Madison SD
omar.el-gayar@dsu.edu

Abstract

Exploring and making sense of large data repositories has become a daunting task. This is especially the case for end users who often have limited access to the data due to the complexity of the retrieval process and limited availability of IT support for developing custom queries and reports based on the data. Consequently, traditional interfaces are no longer meeting these requirements. Instead, novel interfaces are required to fully support the sensemaking process. In this paper, we followed a design science approach and introduced a query clustering system (SenseCluster) that could serve as a quick exploration tool for making better sense of large data repositories. We also present an evaluation of the effectiveness of our artifact using cognitive walkthroughs.

1. Introduction

The term "sensemaking" refers to the process of gathering information and gaining understanding of the information to find meaning in a situation [1]. This process has been studied in various disciplines, for example in Human- computer interaction (HCI) [2], information systems [1], organizational studies [3], and communication [4]. In order to take advantage of emerging trends of open data and big data, knowledge workers have to survey and make sense of a large amount of data and understand its potential. Sensemaking research focuses on developing tools that support individuals to make sense of such complex information repositories. Some examples of such tools include tools designed to support information representation [6], information visualization [4], and organization of search results for users [5]. Thus sensemaking is a key issue in the information rich spaces, and designing user interfaces that enable users to make sense of large amounts of information in an easy and efficient manner is a major HCI research problem [6].

Understanding a users' space and tasks to be supported must be the first step to design any interface [5]. Conventional query-based interfaces are well suited for targeted information seeking tasks but are not designed for exploration of big data repositories or ill structured tasks such as sensemaking tasks. Oftentimes end users have a limited capability to write complex queries or procedures needed to retrieve or explore data repositories, and they are limited in data exploration due to the complexity of the retrieval process and limited availability of IT support and services to develop custom queries and reports based on the data. In data exploration tasks, users often do not have well defined or clear information about the data they want to explore. Therefore, a data exploration interface is needed to support viewing a dataset from multiple perspectives, levels of abstraction and summarization.

The tools currently available to access large datasets are designed for use by expert users such as database specialists and statisticians to transform and analyze the data. There is limited availability of tools that can help end users develop a broad understanding of the data, relationships among data elements, and sample datasets to explore potential trends. Since end users are the experts related to their problem domain and can best identify the potential for data to help with their information and decision support needs, providing them with easy access to data can greatly enhance productivity [7].

In this research, we propose a system called SenseCluster to address the above problems. The SenseCluster system is designed to support data exploration, visualization, and making sense of big data repositories. It builds upon a query clustering model and facilitates reuse of data queries that could serve as a quick exploration tool for large data repositories.

In the rest of the paper, we review relevant literature on sensemaking, exploration and

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.116

visualization, and identify the research gaps. We then present the design science research approach used in this paper for developing the SenseCluster artifact. Next, we present the SenseCluster Artifact including its overall architecture, TreeMap based visualization interface and details of query clustering algorithms. We then describe our implementation of the system and its evaluation using the cognitive walkthroughs process followed by conclusions and future work.

2. Related work

The development of more effective interfaces to support decision making is a key area for data warehouse and decision support research [8]. Enabling easy accessibility to data and data analysis tools is essential for organizations to derive the full value of the data warehouses [9]. A major limitation to develop effective interface to support easy exploration of data, data visualization, and analysis capabilities is lack of focus on designing data exploration systems targeted to end-users who are not familiar with query languages and do not have advanced retrieval skills. In order to understand the state of the research in developing data exploration systems for novice end users, we investigate literature in multiple research areas including sensemaking, human-computer interaction (HCI), visualization and exploration.

2.1 Sensemaking

Sensemaking is fundamentally a human activity. The process of sensemaking is often complex, dynamic, and involving data that is incomplete. Exploratory information seeking is a sensemaking activity in which there is a lack of knowledge or unclear information about the task, information space structure, and even the needed vocabulary or the right concepts [19]. In such exploratory searching, the user experience is a continuing series of knowledge acquisitions that bridge the gaps in understanding and form a chain of reasoning that helps to accomplish the task of sensemaking [10]. In [15], the authors point out that sensemaking is the process of creating understanding and awareness in ambiguous or ill-defined task. In [2], the authors present a theory of sensemaking as a process that is initiated when an individual recognizes the lack of understanding of events. Sensemaking is an active two loops of activities: a foraging loop and a sensemaking loop. Foraging loop involves seeking and extracting information. The sensemaking loop, on the other hand, involves iterative development of a conceptualization and includes activities such as skimming, examining details summarizing, and identifying patterns of concepts and relationship [11].

According to [2], sensemaking focuses on how users understand complex information spaces. . In their model of sensemaking, when interacting with large amount of information, the sensemaker creates representation to capture important features of the information in a way that support completing the task. Then the sensemaker identifies information of interest, encoding it in a proper representation. In the later stage of understanding of a sensemaking task, the sensemaker may find that the initial representation is inadequate to represent the sensemaking problem. In this case, the person is motivated to find better representation that fits with the sensemaking task.

2.2 Sensemaking task

Many of the tasks carried out online can be classified as known-tasks or sensemaking tasks. A known-task is concerned with finding "a single document, factoid, or snippet that satisfies the person's information need" [4]. Typically, searching to support such tasks is characterized by simple queries to retrieve the results, short-duration, and few search results retrieved. This kind of tasks is highly supported by major web search engines [4]. Sensemaking tasks, On the other hand, involve ambiguity, uncertainty, and discovery [12]. In addition, when a user engages in exploratory search, the searching for sensemaking tasks characteristics include: general rather than specific, open-ended, target multiple items, involve uncertainty, dynamic over time, and multi-faceted and complex [13]. Moreover, one of the key characteristic of sensemaking task is the number of queries required to find the needed information. According to [4], a large number of queries are needed for several reasons: to obtain better understanding of the task, to investigate independent aspects, and to react to newly-founded related items. In light with these characteristics, exploratory search systems should be designed and taken a step forward toward supporting exploratory search behavior. Given the complex nature of exploratory search, the design of such kind of exploratory search systems should focus on supporting interactive and dynamic exploration processes and not on search algorithms of classical interest to information retrieval. The interactive and dynamic exploration process play out between the user, the system, and the information sources in a task context [14]. In particular, such exploratory search challenges the interfaces of search engines, because "it requires support to all the stages of information acquisition, from the initial formulation

of the area of interest to the discovery of the most relevant and authoritative sources, to the establishment of relationships among the relevant information elements" [15]. According to [16], in order to engage people in exploratory search process, researcher should devise "highly interactive user interface."

2.3 Sensemaking and HCI

HCI uses sensemaking as "the cognitive act of understanding information" [6]. Designing user interfaces that enable them to make sense of a large amount of information in an easy and an efficient manner is a major challenge in HCI research [6]. Tools like "CoSense" support collaborative sensemaking have been proposed for use in different domains such as hospitals, classrooms, libraries [17]. Other systems proposed include "Entity Workspace" that supports making sense of large document collection [18]. In [19], the researchers propose a "SSIG" system which presents information as a tree structure and helps users to search, construct, reconstruct, and refine the tree presentation. However, there is limited literature on sensemaking systems for exploring large datasets consisting of many tables and data elements from many sources such as in the case of big data.

2.4 Visualization and exploration

In exploration and visualization research, the focus is on supporting people more engaged in exploratory search process to conduct lookup, learning, and investigation tasks through the development of highly interactive interfaces [16]. Examples in this area include "TaskSieve", which is a web exploration system with a task model to support information exploration and visualization [20]. Other systems designed for exploring web or document collections include "Jigsaw" a visual analytic system [21], "SenseMaker" [5] , "Scater/Gather" [22] and "Liquid Query" a querying system that supports multi-domain queries on the web [15]. In addition to web and document collections, sensemaking systems have also been proposed for network data such as "Apolo" which enables users to explore and making sense of large network data [23]. However, most literature is focused on sensemaking systems for document and web collections and is not suited for non-textual databases and database catalogs.

In order to leverage open data, data users should be provided with novel interfaces to explore, analyze and identify the potential of data and associated queries [24], [25]. Most past approaches to this problem involve the development of sophisticated querying interfaces such as relational query processing system that uses microtask-based crowdsourcing [26], query formulation language [27] and SPARQL endpoint and RDF query language [28]. However, in order to provide easy access to data retrieval and analysis capabilities in the large data repositories, conventional query creation mechanisms are not very helpful [28].

Overall, our goal differs from previous research in that we aim to enable end users to quickly access and reuse pre-developed data retrieval and analysis models to analyze data and satisfy their information needs. In addition, our goal is to develop a system that facilitates the reuse of data queries and could serve as a quick exploration system for making better sense of large data repositories through an interactive visual interface.

3. Research approach

The paper followed a design science research approach [29]. Our main goal is to define and develop artifacts that support quick exploration and making better sense of big data repositories.

We first identified the problems of current querying systems: (1) Querying systems that exist for open data are targeted towards experts, and not end users who are not familiar with query languages and advanced information retrieval skills. (2) End-users have limited access to the data due to the complexity of the retrieval process and limited availability of IT support to develop custom queries and reports based on the data. (3) User interfaces limit end users' expectations to explore big data repositories or tasks especially sensemaking or ill-structured tasks.

We then defined specific objectives to infer the requirements of a possible solution to the aforementioned problems. The first objective is to find a solution that caters to end user or novice user who are not familiar with query languages and advanced information retrieval skills. A second objective is to introduce a system that provides end user with easy access to data, querying and analysis, and allow searchers to select a pre-existing query based on their preferences and without query writing requirements. Third, the system should support making sense of big data repositories and help user understand the available data sets, relationship among data, and the potential use of the data.

At the design and development stage, we inferred the requirements of the design features based on the theoretical foundations of the related field of human-computer interaction, exploration and visualization. Further, we combined knowledge and

techniques from the research fields in order to make design decisions that guide the directions of our approach.

Based on this theoretical foundation, we built artifacts to support our research objectives. Using the prototype, we evaluated the effectiveness of the system for data exploration and sensemaking purposes. We then compared the results with the interface design features specifications. The purpose of this step is to ensure that the user interface was correctly composed at the theoretical level.

4. Designing the artifact

4.1 Introducing "SenseCluster"- Initial design features

"SenseCluster" builds on a large body of research aimed at creating an understanding and awareness in ambiguous and ill-defined tasks. We inferred the requirements of the design features based on the theoretical foundations of the related field of human-computer interaction, exploration and visualization. Further, we combined knowledge and techniques from the research fields in order to make design decisions that guide the directions of our approach.

In this section, we look at how the design objectives satisfy the requirements for our proposed approach. The requirements and the design objectives are summarized in Table 1.

Table 1. Artifact design features

Requirements	Ref.	Design objectives
Target novice/ end users.	[30], [31], [32]	The system supports point and click functionality.
Eliminate query writing	[30] [33], [33]	The system supports query visualization. Users can easily select queries based on their preferences through clickable cluster, sub-clusters, and related queries.
Improve end-user accessibility for large data repositories	[23], [34], [35]	The system supports the functionality of re-using the data queries through query clustering and visualization
Support sensemaking processes where a user can: - Understand the data sets. - Understand the relationship among data. - Understand the potential of data.	[2] [33] [17]	User-friendly interface enables users to select/explore clusters, sub-clusters, and related queries. The system supports TreeMap interface for cluster visualization which provides a clear navigation path for exploring the queries by limiting navigation to a drill down/roll up actions. Such interface supports structured, open-ended, and exploratory tasks. Users can select data based on their preferences. The system supports report generation based on the data. The data are available in different views such as tables, charts, maps. The users can easily download the data in multiple formats

In contrast to many systems existing in the literature which focus on targeting experts, "SenseCluster" is built upon a query-clustering model as a potential solution that can enable end users to quickly access data retrieval and analysis models to analyze data and satisfy their information needs. Specifically, we proposed categories of feature sets that can be used to cluster a repository of queries, procedures and models into several clusters. Using a visualization scheme, the automatically developed clusters can then be further segmented to explore the available retrieval and analysis models for use. The underlying feature sets used for clustering can also be varied to cluster and explore the query and model repository from multiple perspectives.

A query clustering and visualization system can provide end users with easy access to such models and enable discovery and retrieval of relevant queries and analysis models. An overview of the proposed query clustering system is given in Figure 1. A brief description of the key components of the system is given below.

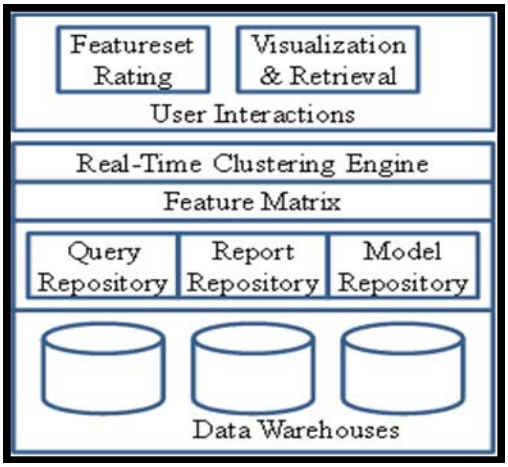

Figure 1. System Overview

4.2 System architecture

Query, Report and Model Repository: The query report and model repositories are used to store queries, reports and statistical models developed to satisfy various user information needs. In addition to the queries and models, the repositories may also contain user annotations describing the query or model.

Feature Matrix: The feature matrix is an index structure for representing the queries, reports and model in the form of features. The feature model consists of a representation of the SQL Query characteristics based on the relational algebra model (Projection, selection, union, difference, product, intersection, joins), a representation of the statistical models as characterized by the statistical modeling techniques used and model variables, and text annotations of queries and statistical models. In addition other key features captured include database tables, views and fields used in a query or a model.

Clustering System: The clustering system is used to automatically cluster the queries and models in the repository to enable visualization and selection of appropriate queries and model by end User. We propose to use hierarchical clustering method such as Hierarchical Agglomerative Clustering (HAC) to automatically cluster the queries and models. Further discussion about the process of identifying the queries cluster is given in section 4.4.

Interactive User Interface: The user interface enables user ratings of different feature sets, dynamic feature weighting in response to user ratings and real-time clustering of queries and models based on the feature weights. Such real feature manipulation and clustering can enable the end user to explore the query repository from multiple perspectives.

Searching the literature for appropriate approaches for information visualization, different approaches are suggested by the literature such as: TreeMap (Hierarchical data) and Graph data (Network data) [36]. Basically, different data visualization approaches serve different purpose and choosing the appropriate data display should fit with the design purpose. For instance, Graph visualization aims to develop summary views of graphs to help users who know nothing or little about the data make sense and explore graphs [23]. In designing "SenseCluster" we chose a TreeMap interface for cluster visualization. TreeMap has already been accepted as a powerful technique for visualizing hierarchical data [37]. In our study, we use a TreeMap method to provide a clear navigation path for exploring the queries by limiting navigation to a drill down/roll up actions. In addition, once the select cluster is identified; it greatly reduces user effort by displaying all relevant queries grouped together within a cluster. One of the most important aspects of "SenseCluster" is that query cluster assignments are not mutually exclusive and multiple hierarchies can be generated for navigating the queries.

4.3 Feature selection for query clustering

A key component of the query clustering system is the feature matrix and the set of features that are used to cluster queries and models into clusters. We identify six categories of features that can be used for clustering queries and models and describe the rationale for using the feature sets in Table 2.

Table 2. Features for Query Clustering

Feature Category	Description
SQL Features	This set of features includes SQL language elements and operators such as Select, Join, Where, etc. The SQL features provide an indication of the type and complexity of SQL queries used to retrieve data and generate reports.
Tables	The tables used in a model or a query are an important feature that can be used to differentiate between queries. The tables represent the source data of the queries and could potentially indicate similarity between queries.
Fields retrieved	The fields retrieved are the fields specified following the select keyword of an SQL query. The fields retrieved are among the most important indicators of the purpose and information retrieved by a query.
Fields in filter conditions	The fields specified in conditional statements include those specified

	under 'where' and 'having' conditions of an SQL query and influence the type of records retrieved by a query.
Statistical Functions	This set of features includes statistical functions used in a query or model such as AVG, COUNT etc. and any advanced statistical functions used in analytical models.
Text Annotations and comments	The features extracted from text annotations, comments and any meta-data associated with a query.

4.4 Identifying the query cluster

In order to define more efficient cluster of the given queries, we propose to use hierarchical clustering method to automatically cluster the queries and models. For our experiment, we identified the input queries based on Health Indicator Warehouse dataset (http://www.healthindicators.gov/), which is a large open data warehouse consisting of a database of community health data from around the USA.

We then preprocessed the input queries and organized them into tables (e.g. diabetes education), fields in condition (e.g. educational attainment), and field values in condition statements (e.g. high school). After that, we uploaded the input queries document to the software, filtered the feature sets, and defined distance and distance metric (see figure 2).

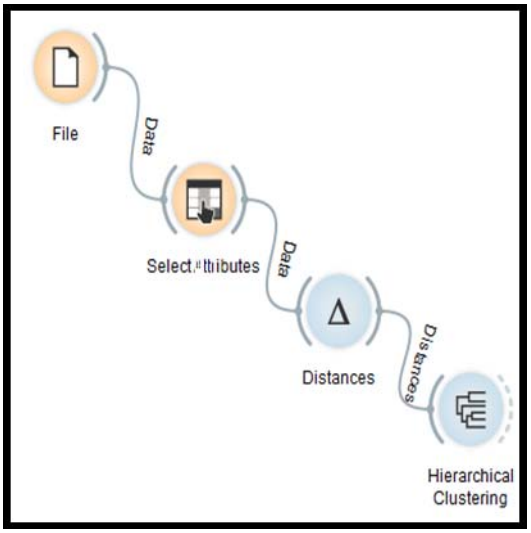

Figure 2. Hierarchical clustering process

Several agglomerative techniques were used to produce a series of clusters: single linkage, complete linkage, average linkage, and ward's linkage.

We developed a cluster homogeneity metric to evaluate the quality of the clusters. Cluster homogeneity assesses the query cluster generated and its ability to cluster similar queries together. In order to calculate cluster homogeneity, we manually checked each query in all generated clusters. We used human evaluators to judge the similarity between queries within a cluster. They took into account the tables, fields, and fields values that the queries belong to. In particular, a query cluster was given high homogeneity score if it clusters similar queries together that belong to the table, field, and field in values that should belong to. On the other hand, low homogeneity score was given to a cluster that separates similar queries across tables, fields, and field in values. More specifically, let C be a cluster and $C = \{q1, q2, q3, \ldots, qn\}$, q a query, T a table, F a field, and V a field value.

- C is given a score of 1 if queries in C $\in \{T \cap F \cap V\}$.
- C is given a score of -1 if queries in C are $\notin$ to the same T, F, or V.

The homogeneity score was then computed for the generated clusters. Figure 3 shows some similar queries from one hierarchal cluster that are clustered together.

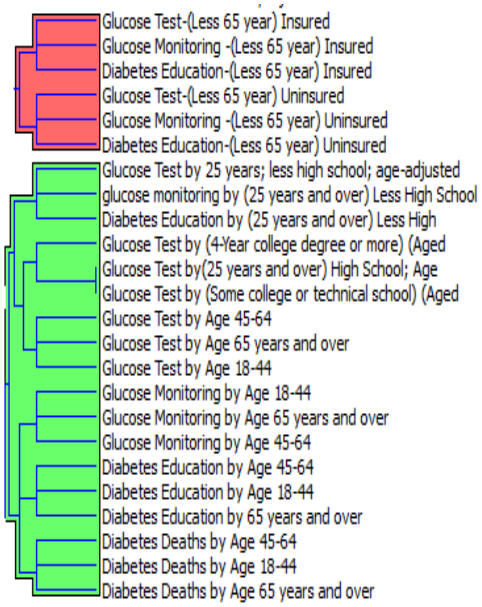

Figure 3. Query clustering sample

5. Implementation and Evaluation

5.1 Query clustering system for Diabetes related Health Indicators

In order to test the our key assumption that our proposed SenseCluster is more suited for the sensemaking and exploration of large datasets for end users who are not familiar with query languages and advanced information retrieval skills than query based approaches, we implemented the system for a

set of diabetes related queries on the Health Indicator Warehouse (HIW) dataset[1]. HIW provides "a single source of national, state, and community health information" on various health indicators in the USA. Specifically, we developed a large set of queries related to diabetes indicators such as diabetes education, glucose monitoring, deaths due to diabetes complications and variations in the indicators by age, economic status, education, insurance status etc.

5.2 Query clustering system evaluation using cognitive walkthrough

In order to conduct a preliminary evaluation of the SenseCluster system prior to user studies, we conducted a cognitive walkthrough evaluation. Specifically, the development team evaluated the interface and played the role of would be users to reveal any possible problems and deficiencies of the interface and the mismatches between system capabilities and user goals. To assess the interface design, we designed an information exploration and sensemaking task related to diabetes in South Dakota. The task used for the cognitive walkthrough was to understand Diabetes trends in South Dakota. The evaluation process session lasted about 2 hours in which the development team evaluated ideal and alternative paths to achieving the task using SenseCluster.

In order to initiate the cognitive walkthrough, an ideal sequence of steps or user interactions were identified for accomplishing the task. Each step was then analyzed in detail from a user perspective. For each step the development team outlined user thoughts and interface actions that could be executed and tried to identify possible problems that users would possibly encounter in executing the step and alternative actions that could be taken by a user. Following the evaluation of each step, design recommendations for addressing potential issues were recorded as illustrated in Figure 4.

Overall, the TreeMap interface for cluster visualization provided a clear navigation path for exploring the queries by limiting navigation to a drill down/roll up actions. Moreover, once the selected cluster is identified, it greatly reduced user effort by displaying all relevant queries grouped together within a cluster. In addition, query cluster assignments are not mutually exclusive and multiple hierarchies are generated for navigating the queries. With respect to limitations of the proposed system, we observed that the interface design in terms of clusters sizes, placement and color should be given more attention. In addition, we also identified that cluster names and query names need to be descriptive in order to guide the users through the most optimal path of the cluster hierarchy for finding relevant queries.

6. Conclusions and Future Work

The goal this research is to support end-user exploration and sensemaking of data in the context of large data repositories. We propose a solution that could support end-user sensemaking, exploration and visualization activities of big data repositories and facilitate the reuse of data queries for better decision making. We have implemented a prototype of the system based on the health indicators warehouse dataset and performed a cognitive walkthrough as a preliminary evaluation of the effectiveness of the artifact for data exploration and sensemaking purposes.

Future research will focus on refinement of the prototype to address the usability problems identified through the cognitive walkthrough. In addition, we plan to use focus groups [38] to further evaluate the design artifact. Specifically, focus groups will allow us to promote the use of the proposed SenseCluster, investigate its performance relative to query-based system, e.g., the HIW system, and qualitatively assess users' attitude, feelings, and beliefs.

[1] http://www.healthindicators.gov

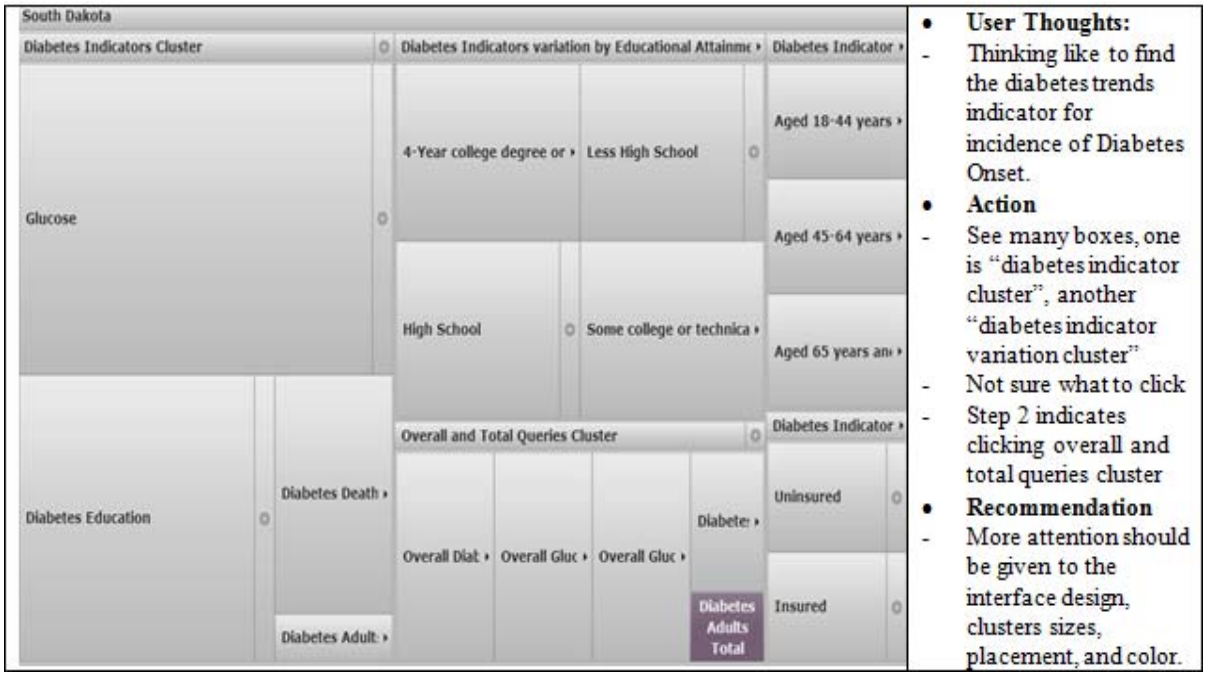

Figure 4. Evaluation of each step using cognitive walkthroughs

7. References

[1] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking II: A macrocognitive model.," *IEEE Intell. Syst.*, vol. 21, no. 5, pp. 88–92, 2006.

[2] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of Sensemaking," in *Proceedings of ACM INTERCHI'93*, 1993, pp. 269–276.

[3] K. E. Weick, *Sensemaking in Organizations*. CA, USA: Sage Publications Inc, Thousand Oaks, 1995.

[4] G. Golovchinsky, A. Diriye, and T. Dunnigan, "The future is in the past: designing for exploratory search," in *Proceedings of the 4th Information Interaction in Context Symposium*, 2012, pp. 52–61.

[5] M. Q. W. Baldonado and T. Winograd., "Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests," in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 1997, pp. 11–18.

[6] S. Whittaker, "Making sense of sensemaking," in *HCI remixed: Reflections on works that have influenced the HCI community*, 2008.

[7] M. Spahn and V. Wulf, "End-User Development for Individualized Information Management: Analysis of Problem Domains and Solution Approaches," *Enterp. Inf. Syst.*, pp. 843–857, 2009.

[8] S. March and A. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decis. Support Syst.*, vol. 43, pp. 1031–1043, 2007.

[9] H. Watson and B. Wixom, "The current state of business intelligence," *Computer (Long. Beach. Calif).*, vol. 40, pp. 96–99, 2007.

[10] B. Dervin, "Sense-making theory and practice: an overview of user interests in knowledge seeking and use," *Inf. Knowl. Manag.*, vol. 2, no. 2, pp. 36 – 46, 1998.

[11] P. Pirolli and S. K. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of International Conference on Intelligence*, 2005.

[12] B. Kules and R. Capra, "Designing exploratory search tasks for user studies of information

seeking support systems," in *Proceedings of JCDL '09*, 2009.

[13] B. M. Wildemuth and L. Freund, "Assigning search tasks designed to elicit exploratory search behaviors," in *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, 2012, p. 14.

[14] Y. Qu and G. W. Furnas, "Model-driven formative evaluation of exploratory search: A study under a sensemaking framework," *Inf. Process. Manag.*, vol. 44, no. 2008, pp. 534–555, 2008.

[15] A. Bozzon, M. Brambilla, S. Ceri, and P. Fraternali, "Liquid query: multi-domain exploratory search on the web," in *Proceedings of the ACM 19th international conference on World wide web*, 2010, pp. 161–170.

[16] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[17] S. A. Paul and M. R. Morris, "CoSense: enhancing sensemaking for collaborative web search," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1771–1780.

[18] D. Billman and E. A. Bier, "Medical Sensemaking with Entity Workspace," in *Proceddings of CHI 2007*, 2007, pp. 229–232.

[19] Y. Qu, "A sensemaking-supporting information gathering system," in *In CHI'03 extended abstracts on Human factors in computing systems2003*, 2003, pp. 906–907.

[20] J. W. Ahn, P. Brusilovsky, J. He, D., Grady, and Q. Li, "Personalized web exploration with task models," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1–10.

[21] Y. Kang and J. Stasko, "Examining the Use of a Visual Analytics System for Sensemaking Tasks : Case Studies with Domain Experts," *Vis. Comput. Graph. IEEE Trans.*, vol. 18, no. 12, pp. 2869–2878, 2012.

[22] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections," in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, 1992, pp. 318–329.

[23] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos, "Apolo: making sense of large network data by combining rich user interaction and machine learning.," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 167–176.

[24] A. Zuiderwijk, M. Janssen, and S. Choenni, "Open Data policies: impediments and challenges," in *Proceeding of 12th European Conference on eGovernment – ECEG 2012.*, 2012.

[25] J. Eberius, M. Thiele, K. Braunschweig, and W. Lehner, "DrillBeyond: enabling business analysts to explore the web of open data.," in *Proceedings of the VLDB Endowment, 5(12).*, 2012, pp. 1978–1981.

[26] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB: answering queries with crowdsourcing," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 2011, pp. 61–72.

[27] M. Jarrar and M. D. Dikaiakos, "A query formulation language for the data web," *Knowl. Data Eng. IEEE Trans.*, vol. 24, no. 5, pp. 783–798, 2012.

[28] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *Proceeding of semantic web . Springer Berlin Heidelberg.*, 2007, pp. 722–735.

[29] K. Peffers, T. Tuunanen, and S. Rothenberger, M. A. Chatterjee, "A design science research methodology for information systems research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[30] R. W. White, S. M. Drucker, G. Marchionini, M. Hearst, and M. c. Scheaefel, "Exploratory Search and HCI : Designing and Evaluating Interfaces to Support Exploratory Search Interaction," in *Proceedings of Human Factors in Computing Systems, CHI '07 extended abstracts on Human factors in computing systems . ACM*, 2007, vol. 49, no. 4, pp. 2877–2880.

[31] J. Pearce, S. Chang, B. Alzougool, G. Kennedy, M. Ainley, and S. Rodrigues, "Search or explore: do you know what you're looking for?," in *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, 2011, pp. 253–256.

[32] M. L. R. R. J. Myllymaki, "Visual exploration of large data sets," *Hum. Vis. Electron. Imaging*, vol. 2657, p. 263, 1996.

[33] B. Kules, R. Capra, M. Banta, and T. Sierra, "What do exploratory searchers look at in a faceted search interface?," in *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09*, 2009, p. 313.

[34] D. A. Keim, "Visual exploration of large data sets," *Commun. ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[35] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 103–114, 1996.

[36] M. Alsaleh, A. Alqahtani, A. Al-Salman, and A. Alarifi, "Visualizing PHPIDS log files for better understanding of web server attacks," in *Proceedings of the Tenth Workshop on Visualization for Cyber Security*, 2013, pp. 1–8.

[37] Y. Tu and H. W. Shen, "Visualizing changes of hierarchical data using treemaps," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1286–1293, 2007.

[38] M. C. Tremblay, A. R. Hevner, and D. J. Berndt, "Focus groups for artifact refinement and evaluation in design research.," *Commun. Assoc. Inf. Syst.*, vol. 26, no. 27, pp. 599–618, 2010.