2021

A Clustering and Treemap-based Approach for Query Reuse and Visualization in Large Data Repositories

Yousra Harb
Yarmouk University

Surendra Sarnikar
California State University, East Bay

Omar El-Gayar
Dakota State University

A clustering and TreeMap-based approach for query reuse and visualisation in large data repositories

Yousra Harb*

Yarmouk University,
Irbid 21163, Jordan
Email: Yousra.harb@yu.edu.jo
*Corresponding author

Surendra Sarnikar

California State University East Bay,
Hayward, CA 94542, USA
Email: surendra.sarnikar@csueastbay.edu

Omar El-Gayar

Dakota State University,
Madison, SD 57042, USA
Email: Omar.El-Gayar@dsu.edu

Abstract: The main objective of this paper is to develop a system to support data exploration tasks over large data repositories. In order to leverage the large amounts of data being generated, decision makers need to first explore the available data and understand its potential for helping with decision problems. In this paper, we present a query clustering and tree map approach that supports data exploration tasks through knowledge reuse, multiple data navigation paths and an easy to use point and click interface. We demonstrate the viability of the approach by building a prototype data exploration interface for health data from behavioural risk factor surveillance system (BRFSS). We evaluate the effectiveness of the artefact using cognitive walkthroughs and a user study. The results indicate that the proposed system is easier to use and reduces user effort for data exploration tasks when compared to a baseline faceted information system.

Keywords: query clustering; query reuse; query visualisation; query exploration; information retrieval; treemap.

Reference to this paper should be made as follows: Harb, Y., Sarnikar, S. and El-Gayar, O. (xxxx) 'A clustering and TreeMap-based approach for query reuse and visualisation in large data repositories', *Int. J. Business Intelligence and Data Mining*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Yousra Harb earned her PhD in Information Systems from Dakota State University (DSU), USA, MS in Information Systems, MIS from DSU, USA and Yarmouk University (YU), Jordan and Bachelor's in MIS from YU in Jordan. She is an Assistant Professor in MIS from the Department of

Management Information Systems at YU. Her research interests are in the areas of knowledge management, data analytics, healthcare analytics, e-commerce and e-government.

Surendra Sarnikar is an Associate Professor in the Department of Management at the College of Business and Economics, California State University East Bay. He holds a PhD in Management Information Systems from the University of Arizona. His research interests include business intelligence, data mining, healthcare analytics and socio-technical design of information systems. He has published several articles in multiple IS journals and has won best paper awards for his work in healthcare information systems at the Hawaii International Conference on System Sciences and the International Conference on Information Systems.

Omar El-Gayar is a Professor of Information Systems at Dakota State University. He has an extensive administrative experience at the college and university levels as the Dean for the College of Information Technology, United Arab Emirates University (UAEU) and the Founding Dean of Graduate Studies and Research, Dakota State University. His research interests include: analytics, business intelligence, and decision support with applications in problem domain areas such as healthcare, environmental management, and security planning and management. His inter-disciplinary educational background and training is in information technology, computer science, economics, and operations research. His industry experience includes working as an analyst, modeller and programmer. His numerous publications appear in various information technology related fields. He is a member of AIS, ACM, INFORMS and DSI.

This paper is a revised and expanded version of a paper entitled 'SenseCluster for exploring large data repositories' presented at 2015 48th Hawaii International Conference on System Sciences, IEEE, Kauai, HI, USA, 5–8 January 2015.

1 Introduction

According to IBM (2015), "2.5 quintillion bytes of data are generated every day". Approximately 90% of today's data has been created in the past few years alone. "The volume of business data worldwide is expected to double every 1.2 years" (Hagen et al., 2013). Wal-Mart, for example, handles more than a million customer transactions each hour that are stored in databases estimated to contain more than 2.5 petabytes of data (SAS, 2012). Large amounts of data are becoming available to decision makers due to the increasing number of people and enterprises that conduct business electronically (IDC, 2014), the increasing number of smart devices connected to the internet, and the growth in mobile data traffic (Gartner, 2015).

Despite the increasing amount of data that is available to decision makers, effectively utilising and making sense of such large volume of available data for decision making remains a major challenge (Keim et al., 2010; Elgendy and Elragal, 2014). In order to leverage data in decision making, it is important to provide decision makers with appropriate tools that inform them about the availability of data and its potential for use in decision making. Data catalogues, query-based tools, and faceted browsing interfaces

are the primary tools currently used to inform decision makers of available data and to provide a high-level overview of the data.

The capabilities of static data catalogues, static lists of datasets and even faceted browsing systems with categories, subcategories and filters, are limited to knowing high level metadata about available datasets but not in depth understanding of data which requires querying and retrieval. However, business users are often unfamiliar with querying languages and may lack query writing skills. Moreover, data exploration tasks often relate to a new domain or situation and are characterised by unclear information needs. Therefore, successful data exploration by business user exploration requires a tool that allows business users to easily interact with and query datasets from multiple perspectives, and can benefit from reusing previously developed retrieval models as a knowledge sharing mechanism.

In order to address the above problem and provide end users with a tool for deeper exploration of datasets, we present a query clustering and visualisation system that supports data users in the exploration of large data repositories. The goal of the system is to support business users who are often unfamiliar with query languages, with data exploration tasks by facilitating access to and reuse of pre-developed data retrieval models and analysis of the datasets from multiple different perspectives with an easy to use point and click interface.

The rest of this paper is structured as follows. We begin in the next section with a review of relevant literature on data exploration tools and highlight the research gaps. We then present the query clustering and visualisation system including its requirements, system architecture, TreeMap-based visualisation interface and details of the query clustering algorithm. We then demonstrate the viability of the approach by building a prototype data exploration interface for health data from the BRFSS (https://www.cdc.gov/brfss/data_tools.htm) dataset. We conduct cognitive walkthroughs and a user study for further evaluation of the effectiveness of the artefact followed by conclusions and future work.

2 Related work

Intuitive and easy to use interfaces are essential for data warehouses and repositories in order to support easy accessibility to data and data analysis tools, and be able to derive the full value of data warehouses for effective decision making (Watson and Wixom, 2007; March and Hevner, 2007). Users who are unfamiliar with the contents of the data or cannot write complex queries to retrieve data can encounter difficulty in exploring the data, understanding the underlying relationships latent in the data, and getting new insights.

In order to leverage large data repositories, users first need to explore and examine the data to help understand the potential knowledge, insight or patterns that can be extracted from the dataset. Such data exploration tasks often involve a new situation or problems, are complex and less structured, involve a new domain, and have an unclear information need. In addition, exploratory search is multifaceted and complex process and often involves multiple general and open-ended queries, and large number of items/documents retrieved (Wildemuth and Freund, 2012).

Information tasks can include both direct-search tasks and exploratory tasks. Direct-search task involves running few queries to retrieve a single document that satisfies the information need. On the other hand, exploratory tasks are those involving the examination of data without having an a priori understanding of what knowledge, information, or patterns it might contain (Baker et al., 2009).

Exploratory tasks differ from direct search tasks in terms of task structure, types of information required, and amount of information required (Al-Samarraie et al., 2017). One of the key characteristics of exploratory tasks is the number of queries that may be run to find the needed information. According to Golovchinsky et al. (2012), many queries are needed for several reasons: to obtain better understanding of the topic, investigate independent aspects, and to react to newly-founded related items. Therefore, it is important to develop tailored models to exploratory search that enable users to mitigate issues such as ambiguity and lack of focus (Hendahewa and Shah, 2017). Exploratory search systems should also help users manage their growing information needs during the sensemaking process (Qu and Furnas, 2008).

One of the effective ways used to aid sensemaking of large data repositories is information visualisation. Information visualisation field involves representing data/information in a way that enable users to interactively explore it, to gain insight, to enhance human cognition of data, to draw conclusions, and make better decision. Sensemaking can inform the design of visualisation (Haider et al., 2019). In this context, several tools have been proposed for supporting the sensemaking process in the information visualisation research. Examples in this area include 'MAMView', a visualisation and data exploration tool that helps users in understanding the data indexed by metric access methods (Vieira et al., 2010) and 'TaskSieve', which is a web exploration tool with task model to support information exploration and visualisation (Kang and Stasko, 2012). Other systems designed for exploring unstructured data such as web or document collections include 'INVISQUE', an intelligence analysis tool helps in searching, clustering, and sorting documents (Rooney et al., 2014), 'TexTonic' which supports interactive visualisation for exploration of large unstructured text collection (Paul et al., 2019), and 'Jigsaw', a visual analytic system helps visualise connections among entities extracted from document collections for sensemaking tasks (Kang and Stasko, 2012).

More recent techniques for supporting exploratory searches in document repositories include unique query suggestion techniques for exploratory session for preventing null results (Li et al., 2017) leveraging search trails of other users in similar contexts (Hendahewa and Shah, 2017), interactive information retrieval (Ruotsalo et al., 2018), and ontology and fuzzy clustering-based approaches (Alam and Baulkani, 2017; Bhavani et al., 2019). 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 make sense of large network data (Chau et al., 2011). While several systems have been proposed to support sensemaking and exploration of documents and web collections, there is limited literature on approaches for exploring large datasets and data catalogues through visualisation of related queries.

With the proliferation of open data initiatives and in order to leverage open data, novel interfaces are needed to help data users easy access to data retrieval, analysis capabilities, and identify the potential of data and associated queries in the large data repositories (Zuiderwijk et al., 2012; Eberius et al., 2012). The approaches proposed to address this problem involve the development of sophisticated querying interfaces such

as relational query processing system that uses microtask-based crowdsourcing (Franklin et al., 2011), query formulation language (Jarrar and Dikaiakos, 2012) and SPARQL endpoint and RDF query language (Auer et al., 2007). More recent development in this area includes the development of a natural language and visualisation tool for querying dimensional data such as OLAP cubes (Djiroun et al., 2019). However, these systems offer limited support for knowledge (query) reuse to explore and retrieve data and more importantly, they do not directly support data exploration task but are rather designed for focused data retrieval or data integration across datasets.

Overall, our research extends prior research in that we aim to facilitate the access, visualisation and reuse pre-developed data retrieval models (queries) by data users to analyse data and satisfy their information needs. The proposed approach is designed to serve as a data catalogue exploration tool that allows data users to gain insights about available data, their potential use in decision making, and to quickly identify potential areas for detailed exploration and analysis.

3 Design of a query clustering and visualisation system

3.1 Design requirements

Table 1 highlights key design requirements while the following paragraphs provide the underlying rationale and related literature.

Table 1 Artefact design features

Objectives	*Theory base*	*Design features*
Support knowledge reuse	Taxonomies are often used to index knowledge and enable reuse of knowledge (Ahmed, 2005).	The system supports the functionality of re-using the data queries through query clustering and visualisation approach.
	"Clustering is helpful for clarifying and sharpening vague queries" (Hearst, 2006).	
Facilitate data query exploration	Interactive query tool for detailed viewing of data from variety of perspectives (Kules et al., 2009).	Query cluster assignments are not mutually exclusive and multiple hierarchies can be generated for navigating the queries, and select/explore clusters, sub-clusters, and related queries.
	TreeMap has already been accepted as a powerful technique for visualising hierarchical data (Tu and Shen, 2007).	
	Tree structure helps users search, construct, reconstruct, and refine the selected information (Qu, 2003; Kules et al., 2009; Paul and Morris, 2009).	The system supports TreeMap interface for cluster visualisation 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 task.
Support data users	Exploratory search interfaces should be user-centered aiming at helping users explore, learn, and use information (Pearce et al., 2011).	The system supports point and click functionality.
		The system supports query visualisation.
	'Direct data manipulation with a schema later approach' improves usability (Jagadish et al., 2007).	Direct data retrieval with a schema later approach; users can easily select queries based on their preferences through clickable cluster, sub-clusters, and related queries.

3.1.1 Support knowledge reuse

Taxonomies are often used to index knowledge and enable reuse of knowledge (Ahmed, 2005). In web search, query clustering has been used to organise users' query terms into a hierarchical structure and automatic development of query taxonomies to provide deeper analysis of domain specific terminology and discovery of term relationships (Chuang and Chien, 2003).

In collaborative web search, knowledge reuse has been enabled by mining query logs for providing query recommendations (Balfe and Smyth, 2005). Query recommendations help non-expert users, who do not have sufficient domain knowledge or specific and well-formed information requests, easily reuse and learn from past queries. Query reuse can also be helpful in exploratory search tasks (Shah and Marchionini, 2009). Generally, users formulate a large number of queries in exploratory search and often reuse the same or similar queries for the same information need. Consequently, query reuse can facilitate and accelerate the task of formulating new queries in exploratory search.

To best facilitate this reuse, query clustering approach has been used to identify and group similar queries in search engine query logs to effectively reuse related queries identified based on previously issued queries (Baeza-Yates et al., 2005). Clustering approaches are specifically helpful in data exploration as "clustering is helpful for clarifying and sharpening vague queries" (Hearst, 2006), and thus helps users understand the data sets, the relationships among the data, and the potential use of the data. However, most previous work on automatic query clustering is in the area of web search and unstructured textual queries in the form of keywords or phrases. In this paper, we develop an approach to cluster structured queries (SQL) and generate taxonomies of query clusters to help explore data repositories.

3.1.2 Facilitate data query exploration

Data visualisation techniques can help in the quick exploration of large datasets. Moreover, data visualisation enables users to capture insights and analytical thoughts which are intermediate products of sensemaking (Keim, 2001; Zhang and Soergel, 2014). Data visualisation is relevant in situations where the individual is exploring the data, analysing and discovering patterns and potential relationships (Stasko, 2007). Research studies have shown that interactive data visualisation can support data exploration processes by providing detailed viewing of data from different perspectives as well as presenting data at different levels of granularity (Livny et al., 1996; Keim, 2001).

Furthermore, hierarchical faceted categories are better for exploration process as they allow exploration from multiple perspectives (Hearst, 2006). Using faceted hierarchical scheme, users can easily navigate content along different dimensions in exploratory search where their information needs is unclear (Niu et al., 2011). In this context, TreeMap has already been accepted as a powerful technique for visualising hierarchical data (Tu and Shen, 2007). It can effectively display the overall hierarchy as well as the detailed data attributes from individual data items in a dataset.

The proposed approach uses a TreeMap interface for query cluster visualisation. The approach supports Shneiderman's (1996) 'overview first, zoom and filter, then details-on-demand' guideline for visual query and data exploration. The TreeMap method is used 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 query cluster is identified; it greatly reduces user effort by displaying all relevant queries grouped together within a cluster.

3.1.3 Support data-users

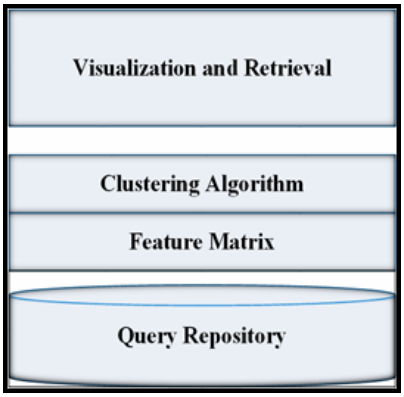
End user performance during query writing, especially in exploration tasks, is affected by the ambiguity in the information request and the complexity of the target solution (Casterella and Vijayasarathy, 2013), user understanding of the query domain and their ability to translate their understanding of the query domain correctly into query language (Ashkanasy et al., 2007; Bowen et al., 2009). Query formulation can be a difficult task for users who are unfamiliar with database schemas and knowledge of query language. Inadequate knowledge of database structure or the querying language often leads to erroneous results (Borthick et al., 2001).

In order to support such users who are not familiar with query languages and/or do not have the time to formulate complex queries, the interfaces should be user-centred and aimed at helping users explore, learn, and use information (Pearce et al., 2011). In order to allow more intuitive user interaction with the system, the data visualisation approach should support 'direct data manipulation with a schema later approach' (Jagadish et al., 2007) such that users are not required to comprehend the schema of the database and formulate queries in terms of that particular schema. Moreover, users may find point-and-click functionality much easier to use than to type queries (Jagadish et al., 2007). The proposed approach adopts the above principles by enabling a point and click interface that allows for reuse and execution of queries leading to direct data retrieval such that users can explore and navigate datasets without requiring knowledge of the database schemas.

3.2 Components

The proposed approach includes several components (Figure 1). The *query repository* stores queries developed to satisfy various user information needs. In addition to the queries, the repository may also contain user annotations describing the query. The queries can be manually generated over time, or automatically generated (Hansen et al., 2013). The *feature matrix* is an index structure for representing the queries in the form of features. The feature model consists of a representation of the SQL Query characteristics based on relational algebra model (projection, selection, union, difference, product, intersection, joins), a representation of the statistical models as characterised by the statistical modelling techniques used and model variables, and text annotations of queries. In addition, other key features captured include database tables, views and fields used in a query.

Figure 1 System overview (see online version for colours)



The *clustering algorithm* is used to automatically cluster the queries in the repository to enable visualisation and selection of appropriate queries by end user. We propose to use hierarchical clustering method such as hierarchical agglomerative clustering (HAC) techniques to cluster the queries. Different criteria were used to generate multiple clustering hierarchies. We set different weight for features set (table, fields, fields values). Moreover, we specified the maximum distance needed to connect parts of the clusters at four levels. As a result, multiple hierarchical clusters can be generated to enable the user to navigate the query repository from more than one perspective. The generated clusters can be evaluated according to the quality and performance measures (Steinbach et al., 2000) such as cluster homogeneity and completeness (Amigó et al., 2009). Detailed description about this process is provided in 'hierarchical clustering process' section.

In designing the query cluster 'QC' we chose a TreeMap interface for cluster visualisation in order 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 'QC' is that query cluster assignments are not mutually exclusive and multiple hierarchies can be generated for navigating the queries. Such feature can enable the end user to explore the query repository from multiple perspectives.

3.2.1 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. We identify six categories of features that can be used for clustering queries and describe the rationale for using the feature sets in Table 2.

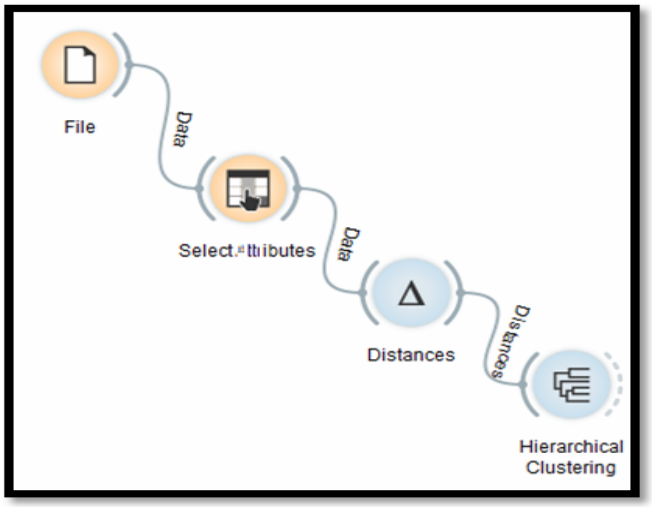
Table 2 Features for query clustering

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 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.
Text annotations and comments	The features extracted from text annotations, comments and any meta-data associated with a query.

3.2.2 Hierarchical query clustering

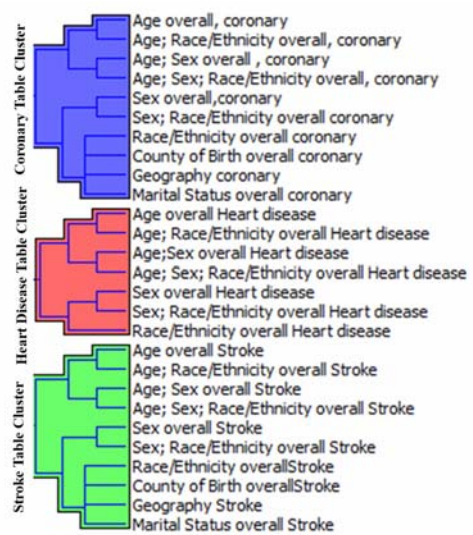
The hierarchical clustering method is used to automatically cluster the queries and models. The process of identifying the hierarchical clusters is conducted through a series of steps. First, we uploaded the input queries document to the software, filtered the feature sets, and defined distance metrics (see Figure 2). Moreover, several agglomerative techniques were used to produce a series of clusters: single linkage, complete linkage, average linkage, and ward's linkage.

Figure 2 Hierarchical clustering process overview (see online version for colours)

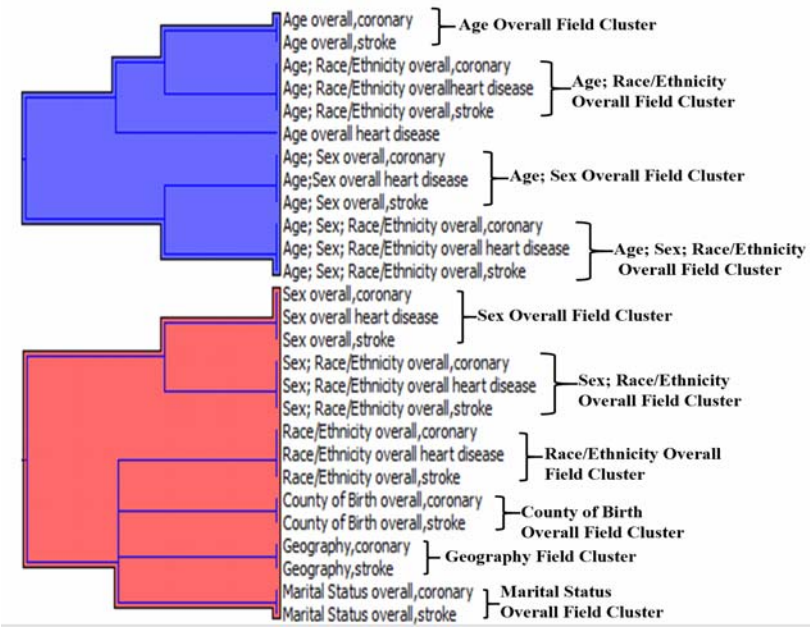


Second, to generate multiple cluster hierarchies, we assigned different weights for tables, fields, and fields values. Figure 3 shows one example of the generated clusters. TableClustering [Figure 3(a)] is clustered according to the tables (e.g., coronary, heart disease, stroke), while FieldClustering [Figure 3(b)] is clustered according to the fields (e.g., age, sex, race/ethnicity, etc.).

Figure 3 (a) Query sample clustering – 'TableClustering' (b) Query sample clustering – 'FieldsClustering' (see online version for colours)



(a)



(b)

We then manually examined the generated cluster hierarchies and assigned different scores based on a cluster homogeneity metric to evaluate the quality of the clusters (Amigó et al., 2009). Cluster homogeneity assesses the query cluster generated and its ability to cluster similar queries together. 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 considered the tables, fields, and fields values which 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 $C \notin$ to the same T, F or V.

The homogeneity score was then computed for the generated clusters. Figures 3(a) and Figure 3(b) show some similar queries from one hierarchal cluster that are clustered together. The clusters with higher homogeneity score are combined in one cluster. The query cluster assignments are not mutually exclusive and multiple hierarchies can be generated for navigating the queries. For example, Figure 3(a) depicts the queries clustered according to the tables (e.g., coronary, heart disease, and stroke tables). Figure 3(b) depicts the queries according to the fields in condition (e.g., age overall field, age; race/ethnicity overall field, etc.). The user can navigate the same query through different hierarchies.

3.2.3 Visualisation of query clusters

The basic idea of our approach is that the developed queries are grouped into clusters according to several criteria as explained in the previous section. The structure of a hierarchical cluster is essentially a TreeMap view with each node representing a cluster or sub-cluster. Each node has its name, branch name, leaf node, and description. The results windows open in a secondary window, so the user can delve deeper and analyse the data tables whereas the main visualisation remains in the original window. The size of each cluster or sub-cluster in the tree map is determined by the number of the sub-clusters or/and the queries in that cluster. The bigger the cluster, the more the number of sub-clusters or/and queries.

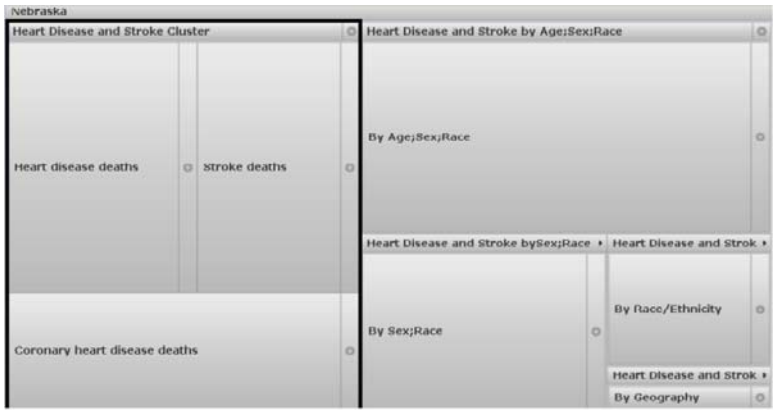
4 Prototype demonstration

We demonstrate the viability of the approach by building a prototype data exploration interface for health data from the BRFSS dataset which is a large surveillance dataset on topics including health-related risk behaviours, chronic health conditions, and use of preventive services. We implement the system for a set of diabetes and heart stroke disease-related queries on the dataset. Specifically, we created a set of 200 queries related to diabetes and heart stroke diseases indicators in South Dakota and Nebraska. Indicators included diabetes education, glucose monitoring, deaths due to diabetes complications,

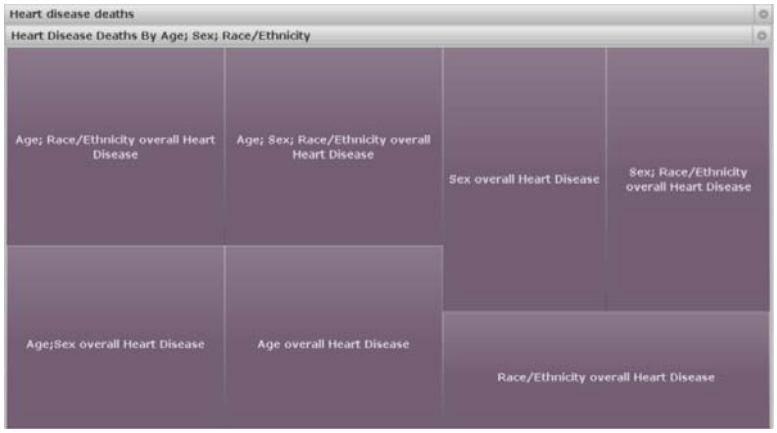
coronary, heart diseases, stroke diseases and variations in the indicators by age, sex, race/ethnicity, economic status, education, insurance status, etc. We then looked through the extracted files and preprocessed the input queries. In particular, we organised every query into table, field in condition, and fields values in condition statements. For example, the query 'glucose monitoring by (4-year college degree or more)' is defined as follows:

- table: glucose monitoring
- fields in condition: educational attainment
- field values in condition statements: 4-years degree or more.

Figure 4 (a) Sample screenshot of 'heart disease and stroke' cluster (b) A secondary window of 'heart diseases deaths' sub-cluster queries (see online version for colours)



(a)



(b)

Using TreeMap we can visualise and navigate the resultant query clusters as shown in Figure 4. The size of each cluster or sub-cluster in the tree map is determined by the number of the sub-clusters or/and the queries in that cluster. The bigger the cluster, the more the number of sub-clusters or/and queries [see Figure 4(a)]. In Figure 4(a), the

'heart disease and stroke' is the biggest cluster since it has the largest number of sub-clusters and queries. Figure 4(b) shows a secondary window of one 'heart disease and stroke' sub-cluster's queries.

4.1 Cognitive walkthroughs

In order to assess the interface design, we developed data exploration tasks related to diabetes in South Dakota and heart diseases and stroke in Nebraska. The evaluation process session lasted about 2 hours in which we evaluated ideal and alternative paths to achieving the task using the proposed approach. An ideal sequence of steps or user interactions to accomplishing the task was identified for the cognitive walkthrough. Each step was then analysed 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 task, design recommendations for addressing potential issues were recorded.

In order to compare the interface of both the proposed approach and a baseline faceted interface system (BFIS), we designed two exploratory tasks. The common aspects of exploratory task are ambiguity, uncertainty, and discovery. The searcher lacks the knowledge to formulate the query and even the required vocabulary or the right concepts. The two sensemaking tasks designed for walkthrough evaluation pertain to healthcare topics. We define the ambiguity in the tasks as scenarios that might require the participants to clarify and redefine the topic themselves. Each task was split into a series of steps that the user has to perform for completing the task. The number of individual steps required to perform the task indicates the complexity level of a user-interface system (Richards and Egenhofer, 1995).

The first task used in the cognitive walkthrough for comparison was to prepare a report that includes information about 'the status of diabetes indicators for 65 years old and above, for South Dakota residents' in a given system. For each task, the user thoughts, actions, and potential errors are recorded. The second task was to identify SD Diabetes trends (incidence of diabetes onset by Age group over last 4 years.). Table 3 displays the procedures for the first task. We gave a number for each category as (1 = user thoughts, 2 = actions, 3 = issues) and (1, 2, 3, …) refers to the number of individual step in each category. QC stands for the proposed query clustering with TreeMap visualisation approach and BFIS stands for the base line system. BFIS is a system that represents an instantiation of a faceted catalogue browsing systems that we are using as a benchmark.

The comparison of the conceptual complexity of the two user-interfaces design is based on:

1 prerequisites knowledge necessary to complete the task

2 the number of steps required to complete the task

3 the potential issues that might arise during implementing the actions.

We presented the previous knowledge needed for the completion of each step as user thoughts. User actions present the basic steps to complete the task. Finally, problems will disclose the potential issues that could occur during task implementation.

We used the number of user thoughts and user actions needed to complete the actions to assess the user-interface design. The less user thoughts required, the less time a user will need to learn the system and understand the dataset, and the less actions necessary to complete the tasks, the faster a user will perform the tasks, and the fewer number of potential errors means the higher likelihood a user will successfully completing the task. If one system meets all the previous three measures, then it would be considered easier to perform the task than the other one.

In task 1 – step 1, we described user thoughts, in both systems, as we expect to find the status of diabetes indicators for 65 age and older South Dakota residents. The main argument made in user actions in QC system was the interface has many boxes, one for 'variation by cluster' and another by 'diabetes indicators'. At this point, the user would not be sure what to choose. The same argument was made for BFIS where the user would see many diabetes-related indicators and is not sure what to select. Potential issues include: the user would experience many clusters and sub-clusters in QC system and many health indicators and pop-screens in BFIS. In this step, both systems have the same number of user thoughts, actions, and potential issues.

In step 2, in QC system, the user goal was defined as to find the diabetes indicators for 65 age and older group. The main interface actions were to click on '+' sign for age cluster which resulted in all age groups in one screen. At this point, the user would locate 65 age and older group. To enhance the interface design, we commented that the placement of clusters and '+' sign for each cluster is unclear. In BFIS, the user goal was also defined as to find the diabetes indicators for 65 age and older group. The user also would understand that he needs to drill down to get the specific details. User actions would start with finding diabetes-related health indicator ('glucose test'). The user would then click on some tabs to select the age dimension and then find the specific age group. The potential issues that would encounter the user include search complexity to find the right information. Overall, using the QC approach, less user thoughts and less number of actions are required to complete this step.

In step 3, we defined the user goal, in QC system, as to retrieve different diabetes indicators for 65 age and older group. The arguments made on interface actions were the screen clearly lists two indicators for that age group, the user can easily find and click on 'glucose monitoring' and 'glucose test' to explore the results. Since the user would easily accomplish this step, the researchers didn't record any issues on the interface design.

In BFIS, the researchers defined the user goal as to retrieve different diabetes indicators for 65 age and older group. The user actions summarised as: first, the user needs to backtrack to the main indicators window. Second, the user should select 'glucose monitoring'. The user then should click on 'data tab' and then 'advanced tab' to get the age dimension list and then the right age group. Finally, the user can explore the results. No issues were identified further as the user would learn from the previous steps that he or she needs to do extensive search to locate the right specific detail. The researchers observe that, in QC system, less action are necessary to complete this step as compared to BFIS. Moreover, the user can perform the same task using two different pathways. Therefore, it supports the argument that the querying pathways in QC system are not mutually exclusive and the user can have multiple ways to conduct the same task.

Table 3 Cognitive walkthrough procedures

Task 1 – step 1

QC

BFIS

QC	QC1	User Thoughts	
		QC1.1	Thinking of finding the status of diabetes indicators for 65 and older SD resident by exploring TreeMap screen.
	QC2	Action	
		QC2.1	See many boxes, one is 'diabetes indicators', another by 'educational attainment', etc.
		QC2.2	Not sure what to click
	QC3	Issues	
		QC3.1	Users would encounter many clusters, subclusters.
BFIS	BFIS1	User Thoughts	
		BFIS1.1	Thinking of finding the status of diabetes indicators for 65 and older SD resident from the come up indicators list.
	BFIS2	Action	
		BFIS2.1	See many health diabetes-related indicators
		BFIS2.2	Not sure what the starting point
	BFIS3	Issues	
		BFIS3.1	User would encounter many indicators and come up screens.

Table 3 Cognitive walkthrough procedures (continued)

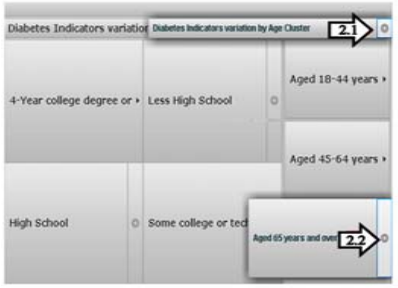
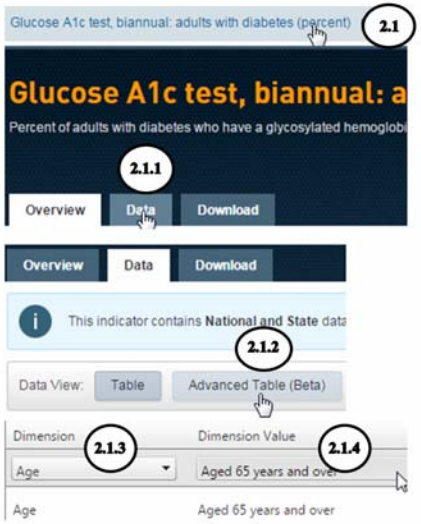
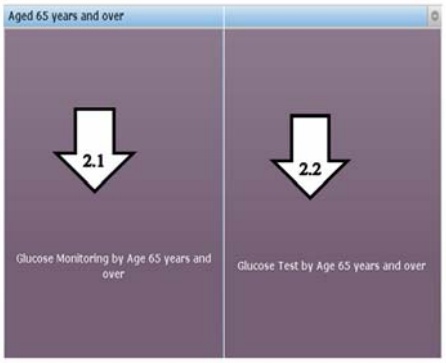
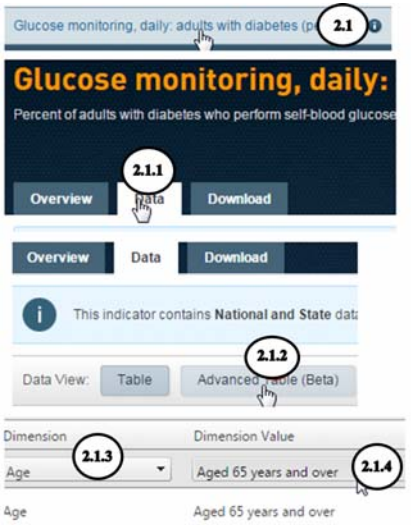
		Task 1 – step 2	
QC		QC1	User thoughts
			QC1.1 Thinking of getting diabetes indicators for 65+ age group.
		QC2	Action
			QC2.1 Click on '+' sign of diabetes indicators variation by age cluster which resulted in all age groups.
			QC2.2 Click on aged 65 years and over.
		QC3	Issues
			QC3.1 Unclear placement of the '+' sign for each cluster.
BFIS		BFIS1	User thoughts
			BFIS1.1 Thinking of getting diabetes indicators for 65+ age group.
			BFIS1.2 The user understand that he needs to drill down in order to excel the information and find the answer.
		BFIS2	Action
			BFIS2.1 Click on clucose test indicator
			BFIS2.1.1 Click on data tab
			BFIS2.1.2 Click on advanced table tab
			BFIS2.1.3 Select Age dimension from the list.
			BFIS2.1.4 Select Aged 65 years and more value form the list.
		BFIS3	Issues
			BFIS3.1 It is quite complex-: extensive search to get the specific details.

Table 3 Cognitive walkthrough procedures (continued)

Task 1 – step 3			
QC		QC1	User Thoughts
		QC1.1	Thinking to retrieve different diabetes indicators for 65+ group.
		QC2	Action
		QC2.1	Click on glucose monitoring for 65+
		QC2.2	Click on glucose test for 65+
		QC3	Issues
		QC3.1	No issues.
BFIS		BFIS1	User thoughts
		BFIS1.1	Thinking to retrieve different diabetes indicators for 65+ group.
		BFIS2	Action
		BFIS2.1	Backtrack to the main indicators list.
		BFIS2.1	Click on glocuse monitoring indicator
		BFIS2.1.1	Click on data tab
		BFIS2.1.2	Click on advanced table tab
		BFIS2.1.3	Select age dimension from the list.
		BFIS2.1.4	Select aged 65 years and more value form the list.
		BFIS3	Issues
		BFIS3.1	No issues.

Overall, the cognitive walkthroughs indicate that the proposed approach requires fewer steps in user actions to complete the tasks and user thoughts than the baseline system. The QC user interaction interface is less complex, because its appropriateness in providing a clear navigation path for exploring the queries, and the system ability to reduce user effort and time by displaying all relevant queries grouped together within a cluster. In addition, it provides multiple ways to perform the task along with lesser number of steps to complete it. However, the walkthrough revealed usability issues related to the interface design in terms of clusters sizes, placement, colour, and inadequate description for clusters and queries. As a consequence, we refined the prototype to address the usability problems identified through the cognitive walkthrough evaluation.

4.2 Evaluation

According to Karoulis (2006), evaluation by experts is considered a rigorous approach as it can help discover problems in actual practice. However, real users will also encounter problems which expert evaluators tend to under-estimate or to not perceive. More accurate and reliable results can be achieved by using a 'combinatory evaluation' which always provides better results. In this study, in order to achieve further validity to the results, a user study (via focus groups) is used to test the results obtained from cognitive walkthrough methodology.

Using focus groups for evaluating design artefact is relatively new in the information system field (Smolander et al., 2008). Focus groups can be effectively applied to evaluate the utility of the design artefacts (Tremblay et al., 2010). Several reasons make focus groups an appropriate evaluation technique for design artefact projects including the flexibility to handle several design topics and domains, the direct contact with respondents, the large amounts of rich data, and the emergence of ideas or opinions based on respondent's comments (Tremblay et al., 2010). Several steps for the conduct of focus groups are outlined in the literature (Stewart et al., 2007; Bloor et al., 2001).

As the focus group technique is a powerful technique in obtaining users' attitudes, feelings, and beliefs, the study objectives aim at investigating the performance of our proposed approach. Specifically, we compare the performance of the proposed approach with the baseline system. The open-ended questions are designed to elicit user comments and opinions about the systems. It is always challenging to design representative sensemaking tasks, data exploration, and visualisation studies (Kules et al., 2009). Designing such tasks requires inducing exploratory style search rather than simple or direct style search. In developing the tasks, we first chose the candidate topics based on the available dataset and then refined the tasks to ensure that they were not too easy to qualify for use in an exploratory search. We used a topic that involved understanding and learning certain type of diseases in US states, given datasets from healthcare domain.

Potential participants were identified via university emails resulting in focus groups consisting of four graduate students. The focus group were given a brief description of the study, and they were asked to sign the consent form and fill out pre-questions to assess their knowledge about sensemaking and large open data databases. The participants were seated in a U-shape arrangement to encourage collaboration. In the first moments of the focus group discussion, the moderator welcomed the participants and demonstrated both approaches and used slides presentation to explain the aim of the study and describe the focus group process.

During the focus group session, the subjects were randomly divided into two groups. We followed a crossover design approach. In this design approach, the first group completed one sensemaking task on the experimental system first and then another sensemaking task on baseline system. On the other hand, the second group started with baseline system to complete one sensemaking task and then used the experimental system (query clustering) to complete the other sensemaking task. Instructions on completing the tasks were given to the participants. After completing the tasks, comparative assessment questions were asked to compare and contrast the performance of both systems. A tape recorder was used to tape the discussion. Everything participants say was strictly confidential – real names were not used in any report.

Baker et al. (2009) propose various ways to measure the quality of sensemaking experience including, objective measures such as time-related measures and the quality of hypotheses generated by exploring the data, or subjective measures such as satisfaction or confidence with using the sensemaking assistance tool. In this research, we focused on the participants' overall satisfaction with using the proposed approach during the data exploration tasks according to the following dimensions: ease of use, user friendliness, accessibility, time to complete the task, and successfully completing the task.

5 Results and discussion

The findings from the focus group sessions show that overall, for both tasks, there was a satisfaction with using the proposed approach for sensemaking during data exploration tasks compared to the query-based approach. For example, the following quotes from some participants showed evidence of the simplicity and ease of use QC compared to the baseline system:

- "QC is quite simple, quite user interactive. The clustering is very easy to use."
- "After using the QC system, I think it is easier for end users to find queries more quickly especially for decision makers."

Participants indicated that the approach also enabled them to easily explore the queries in less time, e.g., "I feel that the QC is very helpful because it is easy to drill down and up". "With QC, it takes less time for users to get the data they are looking for". These findings are consistent with our initial findings from the cognitive walkthrough session.

In addition, the participants also discussed some of the advantages of each search method more generally. In particular, the 'user friendliness', 'accessibility', 'successfulness in completing the task' and 'the timeliness to complete the task' characteristics of the QC approach appealed to many participants. According to them, the system was user friendly, interactive, and enables users to easily access to the queries and quickly find the desired information:

- The clustering is very simple and the user can easily find what he needs.
- "The QC system can provide the user the accessibility to the data that he needed more than BFIS."
- "After using the clustering system, I think its easier for end users to find queries more quickly especially for decision makers."

- "I think the main advantage of QC system is that it enables users to find what they are looking for in less time."

There was a clear consensus that QC is more appropriate for end users who are not familiar with compiling queries than the baseline system. For example:

- "For basic users, the QC gives easier access to the queries."
- "Hitting queries to BFIS is quite complex and users need to navigate and drill down in order to get the results."

The participants also indicated that QC has a clear methodology for enabling the users to find what they are looking for, e.g., "the clustering system can provide the user the accessibility to the data that they needed more than BFIS, especially with specific queries and specific information".

The proposed approach was also recommended for users with limited health domain knowledge:

- "Clustering is better for those who are not familiar with healthcare."
- "For general user, QC is quite easy for them to understand and interpret the information."
- "The QC is quite simple and clear enough."

Overall, participants found the proposed approach much easier to use to understand and interpret the data. Participants suggested enhancing the QC to empower users to create queries in satisfying their information needs: "the clustering system should be provided with more flexible tool to enable end user to make his own queries". These results are also consistent with recent findings in clinical settings that indicate domain experts prefer longer and more technical queries (Tamine and Chouquet, 2017).

In summary, the results from the cognitive walkthrough and user study indicate that the query clustering and visualisation systems are better suited than faceted browsing for data exploration tasks. The results of the cognitive walkthrough show that both systems have an expressive power as they allow users to successfully perform the task. However, the query clustering and visualisation approach performed better than the baseline faceted system as it enables the users to perform the task in fewer steps and provided multiple ways to perform the same task. Further, the TreeMap interface for cluster visualisation provides a clear navigation path for exploring the queries by limiting navigation to a drill down/roll up actions. Once the selected cluster is identified, it greatly reduces user effort by displaying all relevant queries grouped together within a cluster. Hence, the user can get a quick overview of the dataset, explore it rapidly, and facilitate the reasoning process. Therefore, users can easily select queries based on their preferences through clickable clusters and sub-clusters. Moreover, users can easily explore and navigate datasets without requiring knowledge of how to formulate queries in terms of a particular database schema.

The participants from the focus group confirmed the simplicity and ease of use of the proposed approach. The participants also indicated that the query clustering system was a better fit for developing an understanding of the available dataset given the ease of use of the system, the ability to quickly find and reuse relevant queries and the cluster/sub-cluster organisation of the queries.

6 Conclusions and future work

In this paper, we have proposed a query clustering and tree map approach to support data exploration tasks and facilitate easy exploration of large datasets through reuse of pre-developed data retrieval models. Specifically, we make three major contributions in this paper. Our first contribution is the identification of a useful and effective feature set for representing and clustering queries. Our paper is among the first to identify features of structured queries such as SQL for the purposes of clustering. Second, we developed a process for generating multiple cluster hierarchies using different feature weighting schemes that allows for navigating the query dataset from multiple perspectives and facilitates query reuse. Third, we demonstrate the utility of the query clustering and TreeMap-based visualisation approach for data exploration tasks by building a prototype and evaluating it using cognitive walkthroughs and focus groups. Our results indicate that the query clustering and visualisation approach greatly reduces the effort needed to navigate and understand new datasets. The proposed approach is especially beneficial for users who are unfamiliar with querying languages and the data domain.

The results in this paper have implications for both practice and design. In terms of implications for practice, the prototype system design can be used as a blueprint to readily implement such a system for large multi domain data repositories and improve the accessibility of open datasets. In terms of design implications, our paper is among the first to formulate a clustering problem for structured query datasets and identify and extract useful features of structured queries. This paves the way for a new stream of research in many directions including the development of more effective and automated mechanisms for extracting query features. Alternative clustering algorithms for structured query datasets, new visualisation techniques for visualising query datasets and new mechanisms for query reuse. In future research, we intend to further explore further development of the prototype through the use of fuzzy clustering algorithms and large-scale experimental evaluations of the system.

References

Ahmed, S. (2005) 'Encouraging reuse of design knowledge: a method to index knowledge', *Design Studies*, Vol. 26, No. 6, pp.565–592.

Alam, M.G. and Baulkani, S. (2017) 'Reformulated query-based document retrieval using optimised kernel fuzzy clustering algorithm', *International Journal of Business Intelligence and Data Mining*, Vol. 12, No. 3, pp.299–318.

Al-Samarraie, H., Eldenfria, A. and Dawoud, H. (2017) 'The impact of personality traits on users' information-seeking behavior', *Information Processing & Management*, Vol. 53, No. 1, pp.237–247.

Amigó, E., Gonzalo, J., Artiles, J. and Verdejo, F. (2009) 'A comparison of extrinsic clustering evaluation metrics based on formal constraints', *Information Retrieval*, Vol. 12, No. 4, pp.461–86.

Ashkanasy, N., Bowen, P.L., Rohde, F.H. and Wu, C.Y.A. (2007) 'The effects of user characteristics on query performance in the presence of information request ambiguity', *Journal of Information Systems*, Vol. 21, No. 1, pp.53–82.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007) 'Dbpedia: a nucleus for a web of open data', *Proceeding of Semantic Web*, pp.722–735, Springer Berlin Heidelberg.

Baeza-Yates, R., Hurtado, C., and Mendoza, M. (2005) 'Query recommendation using query logs in search engines', *Current Trends in Database Technology-EDBT 2004 Workshops*, January, pp.588–596, Springer Berlin Heidelberg.

Baker, J., Jones, D. and Burkman, J. (2009) 'Using visual representations of data to enhance sensemaking in data exploration tasks', *Journal of the Association for Information Systems*, Vol. 10, No. 7, pp.533–559.

Balfe, E. and Smyth, B. (2005) 'An analysis of query similarity in collaborative web search', *Advances in Information Retrieval*, pp.330–344, Springer Berlin Heidelberg.

Bhavani, R., Prakash, V. and Chitra, K. (2019) 'An efficient clustering approach for fair semantic web content retrieval via tri-level ontology construction model with hybrid dragonfly algorithm', *International Journal of Business Intelligence and Data Mining*, Vol. 14, Nos. 1–2, pp.62–88.

Bloor, M., Frankland, J., Thomas, M. and Robson, K. (2001) *Focus Groups in Social Research*, Sage, London.

Borthick, A.F., Bowen, P.L., Jones, D.R. and Tse, M.H.K. (2001) 'The effects of information request ambiguity and construct incongruence on query development', *Decision Support Systems*, Vol. 32, No. 1, pp.3–25.

Bowen, P.L., O'Farrell, R.A. and Rohde, F.H. (2009) 'An empirical investigation of end-user query development: the effects of improved model expressiveness vs. complexity', *Information Systems Research*, Vol. 20, No. 4, pp.565–584.

Casterella, G.I. and Vijayasarathy, L. (2013) 'An experimental investigation of complexity in database query formulation tasks', *Journal of Information Systems Education*, Vol. 24, No. 3, p.211.

Chau, D.H., Kittur, A., Hong, J.I. and Faloutsos, C. (2011) 'Apolo: making sense of large network data by combining rich user interaction and machine learning', *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp.167–76, ACM.

Chuang, S.L. and Chien, L.F. (2003) 'Automatic query taxonomy generation for information retrieval applications', *Online Information Review*, Vol. 27, No. 4, pp.243–255.

Djiroun, R., Boukhalfa, K. and Alimazighi, Z. (2019) 'Data cubes retrieval and design in OLAP systems: from query analysis to visualisation tool', *International Journal of Business Intelligence and Data Mining*, Vol. 14, Nos. 1–2, pp.267–298.

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

Elgendy, N. and Elragal, A. (2014) 'Big data analytics: a literature review paper', *Proceedings of Advances in Data Mining. Applications and Theoretical Aspects*, Springer International Publishing, pp.214–227.

Franklin, M.J., Kossmann, D., Kraska, T., Ramesh, S. and Xin, R. (2011) 'CrowdDB: answering queries with crowdsourcing', *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp.61–72.

Gartner (2015) *What's Driving Mobile Data Growth?* [online] http://www.gartner.com/newsroom/id/2977917 (accessed August 2016).

Golovchinsky, G., Diriye, A. and Dunnigan, T. (2012) 'The future is in the past: designing for exploratory search', *Proceedings of the 4th Information Interaction in Context Symposium*, pp.52–61, ACM.

Hagen, C., Ciobo, M., Wall, D., Yaday, A., Khan, K., Miller, J. and Evans, H. (2013) *Big Data and the Creative Destruction of Today's Business Models* [online] http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192 (accessed August 2016).

Haider, J.D., Gastecker, B., Pohl, M., Seidler, P., Kodagoda, N. and Wong, B.W. (2019) 'Sense-making strategies in explorative intelligence analysis of network evolutions', *Behaviour & Information Technology*, Vol. 38, No. 2, pp.198–215.

Hansen, P., Järvelin, A. and Järvelin, A. (2013) 'Exploring manual and automatic query formulation in patent IR: initial query construction and query generation process', *Journal of Documentation*, Vol. 69, No. 6, pp.873–898.

Hearst, M.A. (2006) 'Clustering versus faceted categories for information exploration', *Communications of the ACM*, Vol. 49, No. 4, pp.59–61.

Hendahewa, C. and Shah, C. (2017) 'Evaluating user search trails in exploratory search tasks', *Information Processing & Management*, Vol. 53, No. 4, pp.905–922.

IBM (2015) *Big Data at the Speed of Business* [online] http://www.ibm.com/software/data/bigdata/what-is-big-data.html (accessed August 2016).

IDC (2014) *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things* [online] http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm (accessed August 2016).

Jagadish, H.V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A. and Yu, C. (2007) 'Making database systems usable', *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data – SIGMOD '07*, pp.13–24, New York, New York, USA, ACM Press.

Jarrar, M. and Dikaiakos, M.D. (2012) 'A query formulation language for the data web', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp.783–798.

Kang, Y. and Stasko, J. (2012) 'Examining the use of a visual analytics system for sensemaking tasks: case studies with domain experts', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 18, No. 12, pp.2869–2878.

Karoulis, A. (2006), 'Evaluating the LEGO-RoboLab interface with experts', *Computers in Entertainment (CIE)*, Vol. 4, No. 2, p.6.

Keim, D.A. (2001) 'Visual exploration of large data sets', *Communications of the ACM*, Vol. 44, No. 8, pp.38–44.

Keim, D.A., Kohlhammer, J., Ellis, G. and Mansmann, F. (Eds.) (2010) *Mastering the Information Age-Solving Problems with Visual Analytics*, Eurographics Association, Goslar, Germany.

Kules, B., Capra, R., Banta, M. and Sierra, T. (2009) 'What do exploratory searchers look at in a faceted search interface?', *Proceedings of the 2009 Joint International Conference on Digital Libraries – JCDL '09*, p.313, ACM Press, New York, New York, USA.

Li, X., Schijvenaars, B. and de Rijke, M. (2017) 'Investigating queries and search failures in academic search', *Information Processing & Management*, Vol. 53, No. 3, pp.666–683.

Livny, M., Ramakrishnan, R. and Myllymak, J. (1996) 'Visual exploration of large data sets', *Proceedings of SPIE – The International Society for Optical Engineering*, San Jose, CA, pp.263–274.

March, S. and Hevner, A. (2007) 'Integrated decision support systems: a data warehousing perspective', *Decision Support Systems*, Vol. 43, No. 3, pp.1031–1043.

Niu, N., Mahmoud, A. and Yang, X. (2011) 'Faceted navigation for software exploration', *Proceedings of 2011 IEEE 19th International Conference on Program Comprehension (ICPC)*, pp.193–196, IEEE.

Paul, C.L., Chang, J., Endert, A., Cramer, N., Gillen, D., Hampton, S., Burtner, R., Perko, R. and Cook, K.A. (2019) 'TexTonic: interactive visualization for exploration and discovery of very large text collections', *Information Visualization*, Vol. 18, No. 3, pp.339–356.

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

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

Qu, Y. (2003) 'A sensemaking-supporting information gathering system', *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pp.906–907, ACM.

Qu, Y. and Furnas, G.W. (2008) 'Model-driven formative evaluation of exploratory search: a study under a sensemaking framework', *Information Processing and Management*, Vol. 44, No. 2, pp.534–555.

Richards, J. and Egenhofer, M. (1995) 'A comparison of two direct-manipulation GIS user interfaces for map overlay', *Geographical Systems*, Vol. 2, No. 4, pp.267–290.

Rooney, C., Attfield, S., Wong, B.L.W. and Choudhury, S. (2014) 'INVISQUE as a tool for intelligence analysis: the construction of explanatory narratives', *International Journal of Human-Computer Interaction*, Vol. 30, No. 9, pp.703–717.

Ruotsalo, T., Peltonen, J., Eugster, M.J., Głowacka, D., Floréen, P., Myllymäki, P., Jacucci, G. and Kaski, S. (2018) 'Interactive intent modeling for exploratory search', *ACM Transactions on Information Systems (TOIS)*, Vol. 36, No. 4, p.44.

SAS (2012) *Big Data Meets Big Data Analytics* [online] http://www.sas.com/resources/whitepaper/wp_46345.pdf (accessed August 216).

Shah, C. and Marchionini, G. (2009) 'Query reuse in exploratory search tasks', *Poster at Workshop on Human Computer Interaction and Information Retrieval (HCIR) 2009*, 23 October, Washington, DC.

Shneiderman, B. (1996) 'The eyes have it: a task by data type taxonomy for information visualization', *Proceedings of IEEE Workshop on Visual Languages '96*, pp.336–343.

Smolander, K., Rossi, M. and Purao, S. (2008) 'Software architectures: blueprint, literature, language or decision?', *European Journal of Information Systems*, Vol. 17, No. 6, pp.575–588.

Stasko, J. (2007) 'Visualization for information exploration and analysis', *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*, Herrsching Am Ammersee, Germany, pp.7–8.

Steinbach, M., Karypis, G. and Kumar, V. (2000) 'A comparison of document clustering techniques', *KDD Workshop on Text Mining*, Vol. 400, No. 1, pp.525–526.

Stewart, D.W, Shamdasani, P.N. and Rook, D.W. (2007) *Focus Groups: Theory and Practice*, 2nd ed., Sage Publications, Newbury Park, CA.

Tamine, L. and Chouquet, C. (2017) 'On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings', *Information Processing & Management*, Vol. 53, No. 2, pp.332–350.

Tremblay, M.C., Hevner, A.R. and Berndt, D.J. (2010) 'Focus groups for artifact refinement and evaluation in design research', *Communications of the Association for Information Systems*, Vol. 26, No. 1, pp.599–618.

Tu, Y. and Shen, H.W. (2007) 'Visualizing changes of hierarchical data using treemaps', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, No. 6, pp.1286–1293.

Vieira, M.R., Chino, F.J., Traina Jr., C. and Traina, A.J. (2010), 'A visual framework to understand similarity queries and explore data in metric access methods', *International Journal of Business Intelligence and Data Mining*, Vol. 5, No. 4, pp.370–397.

Watson, H. and Wixom, B. (2007) 'The current state of business intelligence', *Computer*, Vol. 40, No. 9, pp.96–99.

Wildemuth, B.M. and Freund, L. (2012) 'Assigning search tasks designed to elicit exploratory search behaviors', *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, ACM, Cambridge, California, USA.

Zhang, P. and Soergel, D. (2014), 'Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking', *Journal of the Association for Information Science and Technology*, Vol. 65, No. 9, pp.1733–1756.

Zuiderwijk, A., Janssen, M. and Choenni, S. (2012) 'Open data policies: impediments and challenges', *Proceeding of 12th European Conference on eGovernment – ECEG 2012*, pp.794–802.