THE DESIGN AND EVALUATION OF INFORMATION VISUALIZATION FOR SCHOLARLY DATABASES IN REDUCING COGNITIVE OVERLOAD AND IMPROVING COMPREHENSION

AZAM MAJOONI GOLKHATMI

UNIVERSITI SAINS MALAYSIA

2021

THE DESIGN AND EVALUATION OF INFORMATION VISUALIZATION FOR SCHOLARLY DATABASES IN REDUCING COGNITIVE OVERLOAD AND IMPROVING COMPREHENSION

by

AZAM MAJOONI GOLKHATMI

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

September 2021

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my dear supervisor Prof. Dr. Mona Masood for the continuous support of my study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I am short in word to express her contribution to this thesis, with criticism, suggestions and discussions. I could not have imagined having a better supervisor and mentor for my Ph.D. study. I would like to make grateful acknowledgement to the Dean and all staff members of the Centre for Instructional Technology and Multimedia, USM. Indeed, I am fortunate to be surrounded by kind and helpful staff. There is no word to express my deep feeling to my lovely parents, sisters, brother and husband for their strong cooperation and inspirations. I am greatly thankful to God, that I am so much fortunate to all good things together.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AOI	Area of Interest
APPV	Author Publication Progress Visualization
CAMD	College of Arts, Media and Design
CLS	Cognitive Load Score
CLT	Cognitive Load Theory
CTML	Cognitive Theory of Multimedia Learning
ETG	Eye Tracking Glasses
ITPC	Cognitive Theory of Multimedia Learning, Integrated Model of Picture & Text Comprehension
JS	JavaScript
LFR	Long Fixation Rate
mm	Millimeters
ms	Milliseconds
OPNV	Overall Publication Network Visualization
PD	Pupil Dilation
RED	Remote Eye Tracking Glasses
SLVi	Saccade Long Vertical
SPDmin	normalized minimum sum of pupil diameters for each participant
TF-IDF	Term Frequency–Inverse Document Frequency
USM	Universiti Sains Malaysia

LIST OF APPENDICES

REKA BENTUK DAN PENILAIAN VISUALISASI MAKLUMAT PANGKALAN DATA ILMIAH UNTUK MENGURANGKAN BEBAN KOGNITIF DAN MENINGKATKAN PEMAHAMAN

ABSTRAK

Penyelidikan ini adalah kajian multidisiplin dalam bidang komunikasi visual. Objektif utama kajian adalah untuk mencadangkan kaedah mengatasi kesan beban maklumat bagi pelajar pascasiswazah dan penyelidik semasa bekerja dengan pangkalan data ilmiah. Kajian ini dimulakan dengan kajian awal mengenai masalah dan idea pengkaji mengenai pangkalan data ilmiah. Kaedah yang paling berjaya untuk menangani konsep sumber maklumat yang besar, yang merupakan sumber utama beban maklumat, adalah pengumpulan dan visualisasi maklumat dan mengubahnya menjadi kelompok dan kumpulan. Reka bentuk visualisasi dalam penyelidikan ini berdasarkan aturan Gestalt untuk mencapai reka bentuk yang mudah dan pelaksanaan yang mantap serta bersifat informatif dan berkesan dalam mengurangkan kesan beban maklumat. Oleh sebab reka bentuk dan pelaksanaan visualisasi memerlukan penilaian dan pengemaskinian, model yang digunakan dalam kajian ini adalah Model Penghampiran Berturut (*Successive Approximation Model*). Kesan langsung dari maklumat yang berlebihan pada penyelidik adalah peningkatan beban kognitif. Oleh itu, untuk menguruskan maklumat berlebihan melalui visualisasi, kaedah yang dicadangkan harus dapat mengurangkan beban kognitif. Dalam kajian ini, teknologi penjejakan mata digunakan untuk menilai beban kognitif pengguna reka bentuk visualisasi maklumat tersebut. Untuk mengautomatikkan dan menormalkan perkiraan beban kognitif, kaedah baharu berdasarkan bilangan dan jangka masa lekapan (titik

berhenti mata), bilangan, kelajuan, dan percepatan sakad (pergerakan mata yang cepat), jumlah kelipan mata per minit dan dinamik diameter pupil digunakan. Visualisasi akhir mampu menstabilkan min diameter pupil dengan menangani masalah seperti memusatkan perhatian dan kegelisahan dari maklumat besar. Hasil akhir kajian ini merangkumi mekanisme anggaran beban kognitif dan kerangka reka bentuk visualisasi baharu. Kerangka visualisasi yang dikembangkan mampu membantu penyelidik dan pelajar pendidikan tinggi untuk melakukan tinjauan literatur dengan cepat antara puluhan ribu penerbitan yang berkaitan dan mencari hubungan antara trend penyelidikan.

THE DESIGN AND EVALUATION OF INFORMATION VISUALIZATION FOR SCHOLARLY DATABASES TO REDUCE COGNITIVE OVERLOAD AND IMPROVING COMPREHENSION

ABSTRACT

This research is a multidisciplinary study in the field of visual communication. The main objective of this study is to propose methods for tackling information overload effects for postgraduate students and researchers while working with scholarly databases. The study begins with a preliminary study on the problems and ideas of the researchers regarding the scholarly databases. The most successful method for dealing with the concept of large information resources, which is the main source of information overload, is aggregation and visualization of the information and transforming it into clusters and groups. The visualization design in this research is based on Gestalt laws in order to achieve a simple design and robust implementation which is both informative and effective in reducing the information overload effect. As the design and implementation of visualizations require evaluation and update, the model used in this study is the Successive Approximation Model. The direct effect of the information overload on the researchers is increased cognitive load. Therefore, to undertake information overload using visualizations, the proposed method should be able to reduce cognitive load. In this study, eye-tracking technology is used to evaluate the cognitive load of the users of the information visualizations designs. In order to automate and normalize the cognitive load estimation, a new method based on the number and duration of fixations (eye stop points), number, speed, and acceleration of saccades (rapid eye movements), number of blinks per minute and dynamics of the pupil diameter is used. The final visualizations were capable of stabilizing the mean pupil diameter by addressing the problems such as split attention and anxiety from huge information. The final outcome of this study includes a novel cognitive load estimation mechanism, a novel visualization design framework and set of visualizations designed and improved for scholarly databases. The developed visualization is capable of assisting researchers and higher education students to quickly perform literature review between tens of thousands of related publications and find relation between research trends.

CHAPTER 1
INTRODUCTION

1.1 Introduction

This research is focused on evaluation of the design efficiency of scholarly databases and providing guidelines for improvement and design of visualizations for them in the form of a prototype platform. This platform aims to provide an environment that takes advantage of visual exploration to reduce the information overload effect in scholarly databases and augment users' cognition. The scholarly databases have a significant role in expanding academicians and researchers' knowledge. There are dozens of open and limited access scholarly databases available such as Scopus, Google Scholar, Web of Science, ScienceDirect and IEEE. Most of these databases use the traditional querying method for searching and browsing through the publications, which eventually could result in "information overload" of the users. The issue of information overload in the scholarly databases has been usually discussed in relation with rapid increase of the online articles and ease of access to the publications. For example, Rowlands and Nicholas (2005) mentioned that the growth in the size of the literature is a key problem for both librarians and scholars.

In this study, the Gestalt principles (Ware, 2020) along with multimedia learning theories are used to design the prototype platform. The influence of the Gestalt grouping principles of proximity, similarity, continuity, connectedness, closure, common region/common fate on comprehension and reducing cognitive load in scholarly databases is studied using an eye tracking experiment. In this experiment, the Gestalt principles are manipulated by parameters of size and color. It is expected that this study will be able to provide a beneficial direction to higher education students and researchers.

In this chapter, initially, a brief background of the study is presented to provide a big picture of the current research and the existing problems in this area. In the problem statement section, the basic problems, preliminary investigations and the motivation to the research are discussed shortly and research problems are stated. In the objectives section, the research objectives and the main goals of the current research are described. The research questions section includes the questions about the methods and variables involved in the main objective, which addresses the issue of information overload in scholarly databases. The rest of the sections provide detailed information about the research hypotheses, the significance of study, limitations, and scope of the research contributions, theoretical framework, research flow and finally operational definitions.

Nowadays with the increase of the importance of information in academic databases for higher education students and researchers, visualizing information can provide a concise and essential device that assists researchers to comprehend and analyze and find correlation in large amounts of information effectively in a much shorter time (Shapiro, 2018). Libraries and academic databases now have huge resources, but according to the recent findings the users are not capable of using these resources efficiently (Wang & Loftis, 2020)

In the past two hundred years, the scientific literature has had exponential growth, although the acceleration the growth rate has reduced slightly in the past few years, but still both number of disciplines and the amount of the literature on each discipline is still steadily increasing (Larsen & von Ins, 2010). Accordingly, the process of literature review for the researchers, with this increase of the academic documents in each discipline has become a complicated task. The emergence of the online scholarly databases has provided a great tool for the researchers in different research areas.

However, the variety and distribution of these databases along with large number of the records in these databases has caused some difficulties (De Bellis & Bellis, 2009; Mikki, 2009) and has raised the following questions: Which database has more coverage over the contents?, What is the trend of the research and what are the related fields?, What metrics should be used to analyze the development of a topic in a discipline?, and finally how to refine the enormous datasets in a meaningful pattern (Cobo & Herrera, 2011).

The mentioned questions are the outcome of huge information resources and limited processing ability of the researchers. Researchers as the main clients of the academic databases have to go through hundreds of numerous search results of the academic databases. With the exponential growth of scientific literature, search results on each topic provide a big dataset of the related papers. In this case, the large number of the search results accompanied by difficulty and expenses of the access to the majority of paid sources makes the task of literature review a challenging and time-consuming process (Calhoun, 2014). In this procedure, the researchers receive too much of information and have to deal with the overload of information. Typically, information overload leads to inaccurate perception from the information (Zhuang et al., 2011).

According to Kirsh (2000), the inaccurate perception of information is due to two major factors of interruption and multitasking. Both interruption and multitasking lead to the split attention effect, which is the result of moving from one information source to another. The other major factor is the limitation of the cognitive capacity, namely germane cognitive capacity (Tabbers et al., 2004). The split attention results in the increase in the extraneous cognitive load, occupying most of the available cognitive capacity of the users. As a result, the remaining capacity for the germane cognitive load shrinks. The germane cognitive load is dedicated to the processing and construction of permanent store of knowledge or the schema. Therefore, with most of the cognitive capacity occupied by extraneous cognitive load, the schema building process does not happen flawlessly, accordingly the user's comprehension is reduced (Sweller et al., 1998; Tabbers et al., 2004).

In the modern world, which is filled with different various sources of information (Al-Kumaim et al., 2021; Bawden & Robinson, 2020; de Koning et al., 2020), learning is usually challenged by split attention effect. The term "Information Overload" is commonly used for the condition when the learners have to deal with several sources of complex information (Zhuang et al., 2011). Information overload generally occurs when the available information resources are too much for the consumers, and therefore consumers find it extremely difficult to use the information (Grether et al., 1985). The concept of information overload has been studied in different research fields with diverse terms, such as information fatigue syndrome (Wurman, 2001), communication overload (Meier, 1963), knowledge overload (Hunt & Newman, 1997) and cognitive overload (Vollmann, 1991). According to Bawden et al. (1999), information overload can happen when access to the potentially useful information becomes a major difficulty.

In some cases, information overload is compared to information poverty. Sweetland (1993) mentions that too much information delays its usage and reduces its understanding. One might argue the researchers can ignore large search results and documents, just go through limited number of more cited documents. However, the new researchers and junior higher education students, which are less familiar with the distribution of the scientific literature and the main sources are prone to start a previously completed task and have their time spend on it (May, 2000; Zellers et al., 2008).

Several methods are proposed to deal with the problems caused by the information overload in corresponding fields such as defensive documentation in engineering (Anette et al., 2000; Jabbari, 2018), web page ranking and clustering in web search results (Irfan & Ghosh, 2018; Turetken & Sharda, 2004), training intelligent systems for emails (Soucek & Moser, 2010), and application of graphics in financial data (Zhuang et al., 2011). In some cases, the researchers not only have to deal with the information overload but also, they have the problem of knowledge overload. The difference between the knowledge and information can be expressed as: information provides answers to questions such as “what”, “where”, “how many”, “how long” but knowledge aims to answer “why” and “how” (Burkhard, 2004).

One of the ways of refining the search results and finding the patterns in the large datasets is by the means of information visualization (Börner et al., 2005). Information Visualization is the visual representations of abstract data, information, and knowledge to amplify human cognition. Information visualization has become a very active research field in the past 30 years, especially with the emergence and widespread use of personal computers (Johnson, 2004). The field of large datasets visualization has improved a lot from both technical and artistic aspects and many of the problems had been studied in this period (Rhyne et al., 2004) but there are still a lot of open problems in the process of generating visualization for the transfer of the knowledge.

The perspectives towards the preparation of a method for knowledge visualization as visual representations to transfer knowledge between users can be classified into three groups: 1) Knowledge, 2) Type of Visualization, and 3) Recipient (Figure 1.1) (Burkhard, 2004).

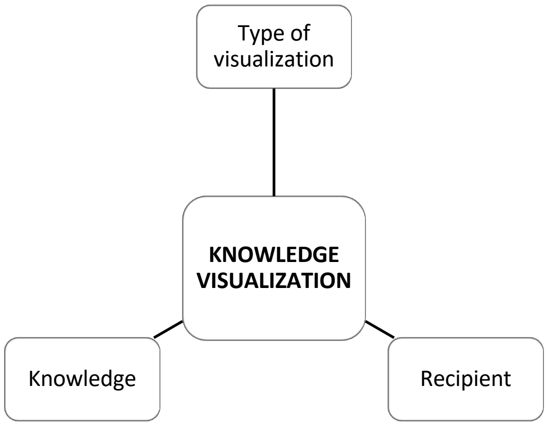

Figure 1.1 Knowledge visualization (Burkhard, 2004)

During the past decade, there has been some efforts in the process of the visualization of the Scientometrics (Broilo et al., 2016; Eppler & Mengis, 2004). Most of the scholarly databases, such as “Web of Science” (WoS), “Scopus” and “Microsoft Academic” have started providing diagrams and mappings to display statistics on bibliometric information in the search results. The provided figures usually present statistics such as a number of the articles published in each year and number of the citations, citation per year and related literature (De Bellis & Bellis, 2009; Martín et al., 2018). However, one of the most important factors in the visualization, which is the graphics of the visualizations and mappings (Garfield, 2009), has received much less attention (Jan & Ludo, 2010).

1.2 Background of the Study

1.2.1 Overview

The age of information and the emergence of the World Wide Web together with ease of access to the sources of information have had numerous effects on the way academicians and researchers access to the scientific documents. The issues such as information retrieval, big data, and the information overload are some of the problems that scientists have to deal with (Lehman & Miller, 2020). According to Ding et al. (2000), the traditional methods of information retrieval have become less effective and huge growth in the amount of information has led to major difficulties in effective information retrieval. They also mention that the three main problems that have raised with the rapid development of the World Wide Web and ease of information availability are “information overload, query expansion, and uncertainty principle”.

Particularly, in the past decade, with the rapid growth in the amount of academic literature on the web and improvements in the information retrieval technology, several multidisciplinary digital online databases such as Scopus, WoS, and Google Scholar were established (Garfield, 2009; Larsen & von Ins, 2010).

The digital academic databases usually provide numerous services for the researchers and academicians. Some of these services are searching, categorizing, ranking and sorting. Before emergence of these databases citation/reference tracking between huge number of documents (Bakkalbasi et al., 2006) was a very difficult task to perform. The quality of the service of the academic digital libraries, rely on the relevance and accuracy of the results in responses to the queries.

Like any other search engine, concepts of coverage, clustering, sorting, ranking, importance and relevance and finally representation of the search results comes across with this difference that unlike conventional databases, users of the academic databases usually follow a specific trend of work and have a certain discipline of research (Kulkarni et al., 2009).

However, the enormous amount and diversity of the related content available on the scholarly databases has made it difficult for the clients to access the refined and fully related information (Lehman & Miller, 2020; Zhuang et al., 2011). Browsing the titles and reading abstracts among the gigantic query results of the scholarly databases is an extremely time-consuming task.

This is not a new phenomenon and has been studied in the past few decades under the title of information overload (Bawden et al., 1999; Eppler & Mengis, 2004; Grether et al., 1985; Hargittai, et al. 2012). The clients usually try to refine these large datasets by the aid of options provided in the side bars of the online databases (e.g. Google scholar, Scopus, WoS and etc.) (Bakkalbasi et al., 2006; Martín et al., 2018).

The information classification options available in scholarly databases are meant to assist the researchers, sort the dataset in different orders (such as, author name, number of citation, date of publication) or limit the search results to specific fields of studies, years, affiliations, journals, availability of full text and so on. Even some of the scholarly databases allow the clients to either store their search results in their personal online account or download (limited number of) the results in standard formats such as BibTex, Endnote or RefWorks format (Goodman, 2005; Reed, 2018).

In general, scholarly databases are units for providing refined information for the clients and have little to deal with the construction of the knowledge on the client side. Although the refining tools aim to reduce and localize the search results their efforts in the abstraction of the information are limited to just presenting statistical information about them (Burkhard, 2004; Valente et al., 2019).

Besides refining the information, some of the well-known scholarly databases (such as WoS, Scopus and Microsoft Academic) have provided diagrams and graphs demonstrating bibliometric statistics like the number of publications per year, per journal, per author or per affiliation or number of citations per document, etc. According to the preliminary investigations, the existing visualizations were found to be useful but in their current form are not sufficient and are very difficult to comprehend (Majooni et al., 2014).

One way to address the information overloads effect is the application of information mapping techniques or in other words, information visualization (Ware, 2020). The term information visualization refers to efforts made to prepare graphics for abstraction of information to increase cognition of the viewers from the information. In the recent few years a tremendous amount of surveys, reviews and researches were conducted in the information visualization as a subcategory of visual communication field (Figure 1.2). The main goal of visualization can be summarized as revealing patterns and the trends of the information, and bringing up insight about the information (Geroimenko & Chen, 2006 pp:18-19; Munzner, 2014). In many cases, the visualized information is actually statistical data about phenomena, population, changes of the parameter, time series, characteristics of a system, mathematical predictions and so on (Kmetova, 2010; Ware, 2020).

An efficient information visualization can be described as a visual communication channel, which lets the viewer see the patterns in data easily. Therefore, the science of pattern perception can be used as a method for providing a basis for a design decision.

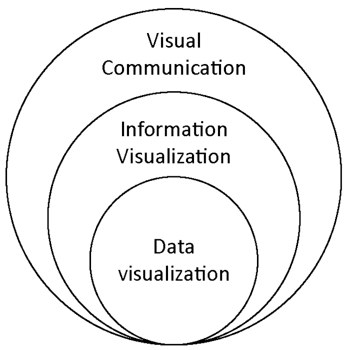

Figure 1.2 Relationship between visual communication, information visualization, and data visualization

On the other hand, it is very difficult to design visualizations without any theoretical framework. A visualization is a construction built by the meaningful organization of graphical elements in order to convey the intended message to the viewers. Bertin (1983) has tried to provide a comprehensive classification for all graphic marks in terms of how they could express data. Visual semiotics as a sub-domain of semiotics can be used to analyze the way visual images communicate a message.

It is often assumed that any visual communication is based on the language of visuals and providing a clear and understandable communication can only take place between two parties if they both speak the same language. Using the language of visualization, or in the other words, grammar and vocabulary of the visualization, namely visual literacy are necessary for the design of a good visualization (Ware, 2020). Semiotics provides a theoretical framework in order to study quality and the effectiveness of visualization. Graphs and diagrams as constructive units of the visualizations are the outcomes of a social interaction and communication. Therefore, having a comprehensive knowledge of design principles, design elements, signs, and symbols, allows the designer to provide easily perceivable visualizations.

Several researchers in the past few decades have tried to come up with new models and applications to visualize the bibliographic information for the purpose of science mapping (Cobo & Herrera, 2011). Visual Science Mapping is also a method of information visualization for bibliometric statistics. In the visual science mapping the techniques of the information visualization, including the transformation of the information into visual elements and providing a special connection between the elements is operated using various visualization methods. For instance, in the science mapping methods described by Cobo and Herrera (2011), the temporal placing of the elements (symbols of literature and authors) in the environment is the mere attempt of providing a visual structure of the information that is more understandable to the user in a glance.

Small (1999, p. 799) defines science map as a “spatial representation of how disciplines, fields, specialties, and individual papers or authors are related to one another”. He mentions that this relation can be either physical proximity such as the geographical position of the affiliation or other physical features. According to Small (1999), the main objective of science mapping is to create a global structure and overview of large datasets and enable users to explore the detailed structure under it. This structure could be represented using visualization. The goals and objectives of science mapping are more abstract (Small, 1994) than overcoming information overload issue. However, some of the strategies used in the process of building science maps can be helpful in reducing the information overload effect by producing a conceptual map of science (Ding et al., 2000; Leung et al., 2017). Chen (2006), indicates the importance of the multidisciplinary integration of information visualization with Scientometrics, philosophy of science and science mapping is. He also mentions that information visualization has potentials similar to the telescope and microscope. Stating that a “telescope enabled Galileo to see the moons of Jupiter, and a microscope made it possible for Pasteur to see bacteria that enabled him to understand disease processes”. Chen (2006, p. 6), explains that in information visualization effective presentation of information is critical to influence decision-making, in science mapping, therefore, designers must understand how users collaborate.

Cobo and Herrera (2011) in their article "science mapping software tools" have reviewed nine of the most well-known science mapping software tools and tried to find out the advantages of each software by comparing them using five different points of view (the preprocessing methods, the bibliographic network available, the normalization and measures used, type of analysis and other extra aspects). The results of the analysis and the comparison indicate that no science mapping software tool is perfect and includes all the capabilities of preprocessing, clustering and representation of the information and therefore, there is a demand for generating more comprehensive science mapping tools (Cobo & Herrera, 2011). Besides the discussed issues, Cobo and Herrera argued that the nine science mapping tools were based on simple geometric concepts of shape and line, and the design principles were not taken care of. In order to achieve a robust design, it is necessary that the design principles and educational technology theories be taken into consideration. Thus, the basic concepts of visual communication and visual literacy techniques should be considered in the design of visualizations.

The main goal of the current research is to develop prototype of a visualization system, named as Scientific Information Visualization, (SCiNFOVIS) with the goal of assisting researcher by reducing the information overload and boosting their comprehension of the acquired datasets from academic databases (e.g. Scopus or Google Scholar).

1.2.2 Theoretical background

Deploying multimedia in the design of a visualization system enables the designer to convey the information required by the user through both the verbal and visual channels to the learners.

Each of the existing technological mediums possess some attributes that provide especial capabilities for designing interactive or static multimedia-based materials. For instance, e-Learning environments, informative slides, and interactive interfaces each one takes advantage of one method to transfer the knowledge. The symbolic coding and decoding system embedded in each of these technologies rely on the way information is processed by the learners (Salomon, 1994).

However, research suggests that inappropriate application of multimedia in the design of the instructional materials could lead to unexpected effects such as the split attention and cognitive overload. One example is the use of representational slides, which has been debated so far by many of the researchers. Tufte, for example, argued that power point slides are similar to a drug that makes us stupid by “degrading the quality and credibility” of communication (Tufte, 2003).

According to the Dual coding theory, the information processing of human consists of two independent, yet interconnected, systems: a verbal system and a visual system. The effect of processing in both channels (combining verbal and visual information) leads to significantly better results and lower cognitive load. Two major multimedia learning theories, namely Mayer’s Cognitive Theory of Multimedia learning (CTML) and Schnotz’s Integrated Theory of Text and Picture Comprehension. There has been a great amount of research on the effects of combination of visual and verbal information in both theories and several principle are proposed by them (Mayer & Moreno, 2003; Schnotz & Kulhavy, 1994).

When subjects are provided with different mutually referring information sources, such as written text and a graphic; the separate presentation could adversely affect the process of learning. This phenomenon is usually referred as the split attention effect (Ayres & Sweller, 2005). There are several hypotheses regarding the causes of split attention. One hypothesis states that the source of this effect lies in the inability of the subjects in integrating two sources of information, whereas the other believes that the subject finds it difficult to switch between two sources and loses the last position on the verbal or graphical representation (van Gog & Scheiter, 2010). The clear issue here is that split attention in many cases leads to rapid increase in the cognitive load and accordingly slows down the process of learning. Several studies are conducted in the past few years to identify the main causes of problems such as split attention effect and also proposed strategies to overcome these problems (Mayer & Moreno, 2003). Most of the methods applied in these studies use the changes of the cognitive load as a dependent variable and measure of quality. There are several methods to estimate the changes in the cognitive load, which are either subjective or objective. The subjective measures depend on the evaluation of the participants from the experiment, whereas the objective methods monitor the physical changes of the participants such as the heart rate (Paas & Van Merriënboer, 1994) and eye tracking data (Paas et al., 2003).

Cognitive Load Theory (CLT) is based on information processing assumptions such as Paivio's dual coding and Mayer's dual channel assumption (Mayer & Massa, 2003). The CLT proposes that each learner has only a limited amount of working memory. This statement is of great importance in the design of instructional materials. The main requirement of learning is optimizing the process of information management in working memory, and when the cognitive load in the working memory exceeds the limits of working memory, the learning process either slows down or completely stops. Therefore, optimization of information processing in the working memory has the highest rank in the CLT. According to CLT, it is believed that there are three different types of cognitive loads: Intrinsic cognitive load, Extraneous cognitive load and Germane cognitive load (Chandler & Sweller, 1991). The intrinsic cognitive load is related to the information presented and, in some cases, can be reduced by the aid of scaffolding and chunking techniques. The extraneous cognitive load is usually imposed to the learner by the environmental variables and distractions in the presentation of the information. Finally, the germane cognitive load refers to the mental workload engaged in processing the information.

Figure 1.3 provides a conceptual image of different types of cognitive load. In this figure, each of the cognitive loads is separately demonstrated on a conceptual rectangle, which represents the total cognitive capacity.

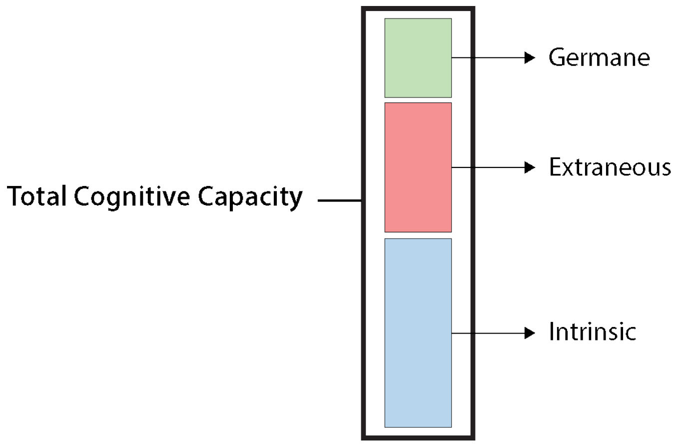

Figure 1.3 Total Cognitive Capacity (Young et al. 2014)

It should be considered that the germane and intrinsic cognitive load is different for advanced and beginner learners. In the case of an early learner, intrinsic cognitive load can be extremely high that leads to overload in the working memory and consequently no information processing can take place (Young et al., 2014). However, in the case of an advanced learner performing a task with no intention to learn, there is no information-processing going on, and therefore no germane load is present. When the expert is performing an active learning task, the germane cognitive load is active and extraneous cognitive load is constant (Figure 1.4).

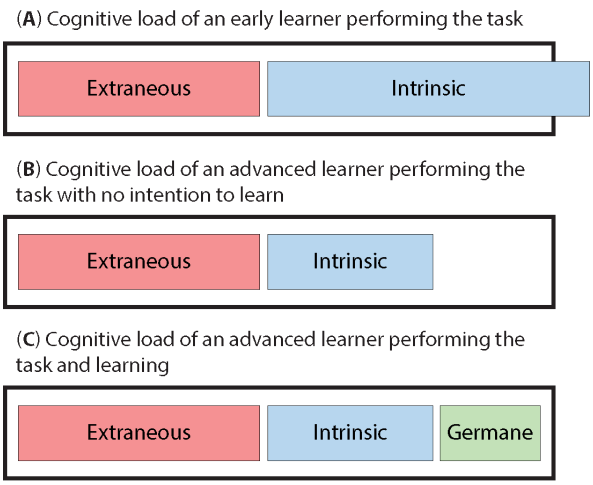

Figure 1.4 Cognitive load in early and advanced learner (Young et al., 2014)

Literature review reveals that eye tracking experiments conducted to estimate cognitive load are performed with passive record & analysis method. These effects reduce the efficiency of the analyses by interfering with the effect of independent variables and limit the capabilities of the designer of the instructional materials in testing different strategies. Numerous studies indicate existence of a strong relationship between amount of the mental workload and the pupil dilation (Ahern & Beatty, 1979; Bernhardt et al., 1996; Kahneman & Beatty, 1966; Klingner et al., 2008; van der Wel & van Steenbergen, 2018; Xu et al., 2011). Bernhardt et al. demonstrated that the response to the increase in the mental workload has an instant effect on the pupil dilation.

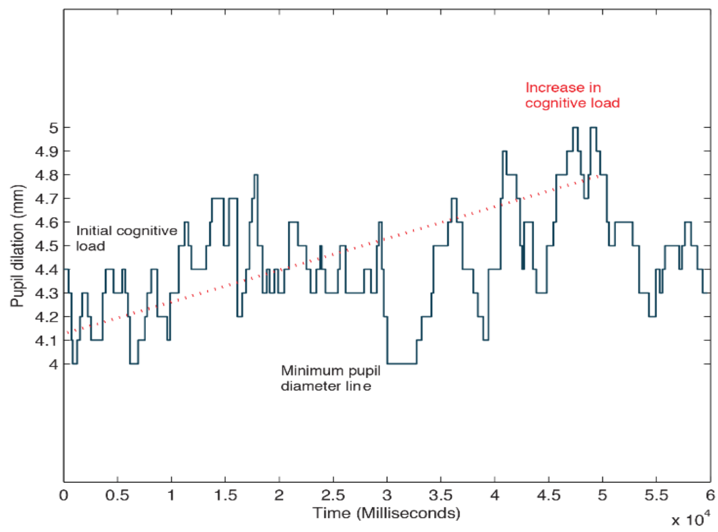

Figure 1.5 Pupil dilation and the Cognitive load (results of eye tracking experiment)

For instance, Figure 1.5 demonstrates the changes in the pupil dilation of the subject while reading a context containing both visual and textual information. The pupil diameter of the subject fluctuates during the experiment, in regards the elements he is going through. The red dashed line in the figure shows the trend of the increase in the pupil dilation, which is correlated with the cognitive load of the subject. In the last 10 seconds of the experiment, the subjects have finished reading the content and has comprehended the context. The pupil size and accordingly cognitive load has suddenly dropped roughly to initial start point. This example provides a conceptual representation of how the pupil dilation and cognitive load are connected to each other.

1.3 Problem Statement

In the past few decades, the online academic databases (Scholarly databases) have grown to be an important part of the process of the research and innovation in the modern world. Scholarly databases play an essential role in the process of literature review, research and innovation (Björk et al., 2020; Callahan, 2014; Demir, 2020;

Denney & Tewksbury, 2013; Sastry & Sastry, 2013; Steward, 2004; Torraco, 2005). Meanwhile the extremely rapid growth of the literature on every subject has resulted in information overload for researchers. The difficulty in information retrieval and inability of absorbing too much information yields to confusion and bias in the evolution of literature review (Ding et al., 2000).

The rapid progress of internet technology has expedited and supercharged the production of information in our modern information society (Al-Kumaim et al., 2021). Information overload has been described as a big problem in the past two decades. It's been mentioned in science, medicine, education, politics, government, business and marketing, smart city planning, news access, personal data tracking, family life, social media use, and online shopping, among other places, and it's even inspired literature (Bawden & Robinson, 2020). In order to identify the general aspects of information overload, numerous research have been conducted in different fields in the past few decades (Bawden & Robinson, 2008, 2020; Hunt & Newman, 1997; Vollmann, 1991; Wurman, 2001). Information overload is a state in which the reader receives too much relevant information about a subject to process (Zhuang et al., 2011). For instance, in the case of researchers and scholarly databases, there usually exist thousands of relevant potentially useful documents for a keyword on a scholarly database like Scopus. For some of the keywords, this number is too many and the researcher can hardly go through browsing the title words only. Bawden and Robinson (2008) mention that there exists no single definition for information overload but in summary of the provided definitions in the existing literature, information overload occurs when the information received leads to the hindrance rather that improve of cognition from the potentially useful information. This increase in difficulty not only leads reduction in the quality of the research but also imposes extra mental workload on the users. Kirsh (2000) introduced the relation between the task difficulty (problem hardness) and the consumption of working memory. He mentions, "The increase in cognitive overload seems an inevitable consequence of the complexity of our information intense environments." (p.22). One of the most obvious consequences of Cognitive overload is reduction of comprehension from the context, this is due to increase in the extraneous cognitive load that occupies bigger portion of the working memory and leaves little or no more space for the germane cognitive load which is in charge of processing information and creating schema.

In this study, initially a preliminary survey was conducted in 2014/2015 about the application of academic databases, information overload, and different facilities available on the scholarly databases (Majooni et al., 2014). The participants were the higher education students at Universiti Sains Malaysia. Literature review and the results of the preliminary survey support the claim that there exist information overload and information retrieval difficulties in the scholarly databases. The majority of the participants were not satisfied with the existing facilities (e.g. sorting and visualization options) in the scholarly databases. In addition, they had difficulties finding the relation between different research documents and finding the trend of the research. Most of these difficulties had roots in the large number of search results leading to information overload and lack of efficient methods for grouping information (representation and clustering) (Rachfall et al., 2014). The problem of information overload and difficulty in information retrieval in the scholarly databases (Roetzel, 2019), could be categorized into three major problems:

1) Difficulty in summarizing, finding, understanding the relation and trajectory of scientific documents.
2) Inadequacy of existing visualization systems in the academic databases.

3) Lack of an interactive multivariate visualization for scholarly and academic databases.

Although emergence of the academic databases and the number, quality, accessibility and retrieval of the information have improved, yet there are many unresolved issues. For instance, in the field of information visualization for the academic databases, the principles of the design are rarely considered, therefore the visualizations are insufficient from different perspectives. The results of the survey and the eye tracking experiments on the existing information visualization in the academic databases, demonstrate that the researchers have difficulty in digesting information embedded in the existing graphs and visualizations. Thus, this study aims to design alternative visualizations and evaluate the proposed visualizations using cognitive load of the users as a dependent variable. The main problem with huge number of the search results is the difficulty of going through mage pages which eventually leads to split attention. Therefore, a main problem is representation of huge data and achieving meaningful connection between related pieces of the data.

1.4 Preliminary Investigation

In this study in order to identify the habits, usage intensity and limitations of the scholarly databases, initially a survey initially a survey was conducted in 2013 and again in 2015 whereby 200 participants from different fields of research participated in this survey. Besides exploring the clients experience using the academic databases and scholarly databases (such as frequency of usage, popularity of databases and usability of the facilities), the survey had four major objectives:

1) To identify importance of visualizations compared versus the filtering/clustering options provided by the academic digital libraries.
2) To investigate the popularity and frequency of multiple and single disciplinary academic database, between the researchers of different fields.

3) To achieve an initial understanding about the most used classification tools on academic databases.
4) To find out about the difficulties and problems the researchers face while using the academic databases.

Therefore, the survey consisted of four sections: in the first section the academic and personal information was collected, in the second section, the access frequency was investigated and participants were asked about their most used databases. The third and final section consisted of questions about the existing facilities and possible updates in the academic databases. The survey was conducted on the internet by the aid of online questionnaire designed or the identical paper-based copy of it. The hard copy replica of the survey has been distributed in several academic events, workshops and colloquiums between the higher education students at Universiti Sains Malaysia. In order to reach the participants, the link to the soft copy of the survey has been emailed to the academic staff and the higher education students.

The outcomes of the statistical analysis of the questionnaires provided that the existence of visualization for dealing with large queries is significantly important and the participants had difficulties processing too much information related for each query.

In several studies (De Moya-Anegón et al., 2007; Etxebarria & Gomez-Uranga, 2009; Groote & Raszewski, 2012; Zhu & Liu, 2020) the three major academic databases (WoS, Scopus and Google Scholar) are investigated and their cons and pros are compared with each other. With the prior knowledge that the participants have full access to these libraries inside campus network, the popularity of the databases is presented in Figure 1.6.

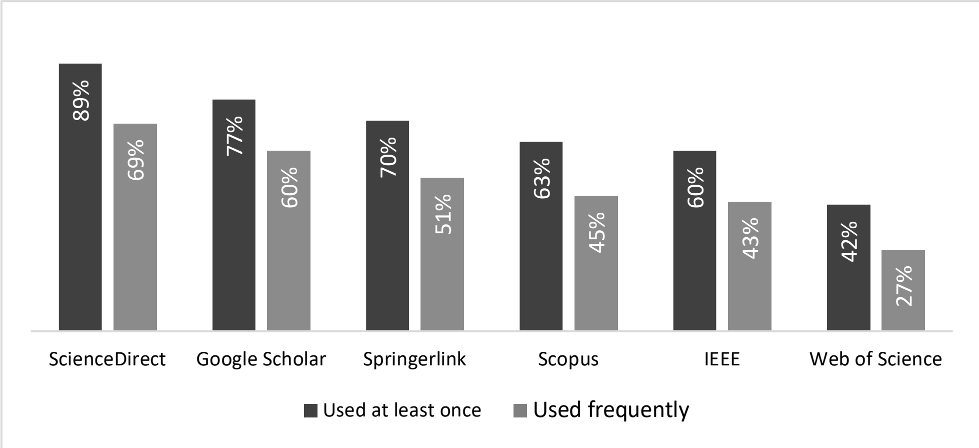

Figure 1.6 Preference of the participants in employing the databases

Although both Scopus and WoS are accessible inside the campus, Google Scholar and search engine of the ScienceDirect were the most preferred databases. The coverage of the Google Scholar is 94% of the citing references, which might be one of the reasons why the researchers in USM employ it more than its major competitors (Scopus and WoS). The other reasons can be the simplicity of the user interface, faster query speed and ease of access (Etxebarria & Gomez-Uranga, 2009).

According to the results of the survey, researchers from all the fields use ScienceDirect and Google Scholar more regularly than the other databases, whereas IEEE and WoS belongs to particular research fields. For example: most of the participants in favour of IEEE were from Computer Science and Engineering fields, which is normally expected to be like that.

The immense amplitude of the literature and Scientometric records on a topic or a keyword in the academic databases leads to information overload effect. Several strategies are suggested to overcome the information overload effect (Hargittai et al., 2012). Field of information visualization has received considerable attention as a strategy to address this issue (Tate, 2008). In the survey conducted for the preliminary investigation, the importance and effectiveness of existing graphical elements and visualizations in the academic databases has been questioned using two separate questions. The results of these two questions are separately investigated to study the validity of the results. The first question requires the participants to specify significance of visualizations in the presentation of the Scientometric information in the queries. In the second question, the graphical elements, and visualization in search results are compared with other methods such as grouping, filtering, and sorting options in advanced search options. According to the results, overall 58% of the participants believed that visualizations are very important in abstraction of the queries in the academic databases and 31% have selected that it is nice to have visualization alongside textual representation of the queries. However, overall 11% have considered visualizations are not important at all.

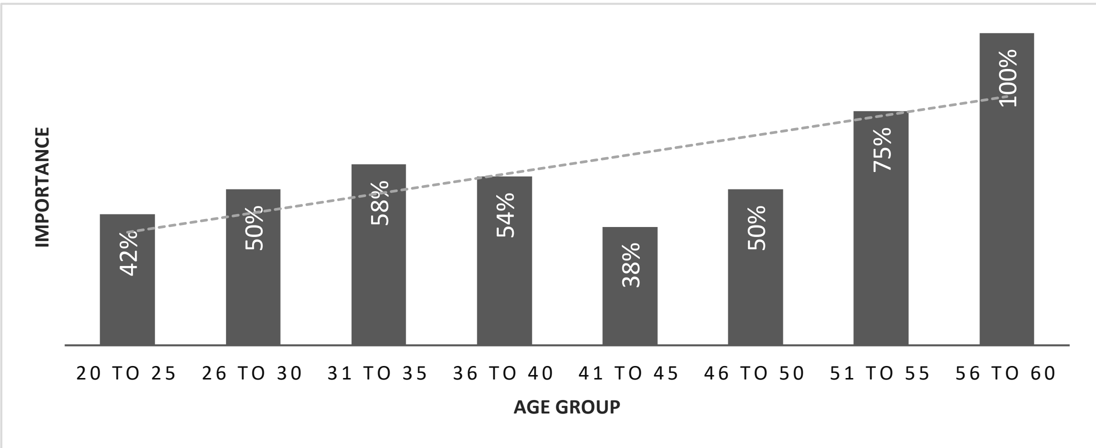

Figure 1.7 Importance of the visualizations versus age of the participants

Figure 1.7 demonstrates statistics related to the importance of the visualizations versus the age group of the participants. The graph suggests that the age group (41 to 45) has the least interest in the visualizations whereas the three age groups of (56 to 60), (51 to 55) and (31 to 35) are more in favour of the visualizations and believe in their importance in the presentation of the queries. The interesting results represented in Figure 1.7, indicate that advanced users are more in favour of visualizations. This could be due to higher amount of the insight that can be acquired from visualizations in comparison to textual representation of the data.

Understanding and learning through visual contents is much faster than the textual and other types of receiving information. The same concept can be true for the researchers trying to review the literature or understand connection concerning different concepts and disciplines (Huang et al., 2009).

In the preference section of the survey, the participants are asked to grade the importance of different options on the academic databases. These options can be very helpful in limiting the search results, excluding the unwanted results and finding the trend of research.

In addition, the importance of the visualizations for citations and references is compared with these options. The five major options are:

1) Sorting by number of citations
2) Grouping by author name
3) Grouping by publication/journal
4) Sorting by impact factor
5) Sorting by date of publication

The answers to this question of the survey are presented in Figure 1.8.

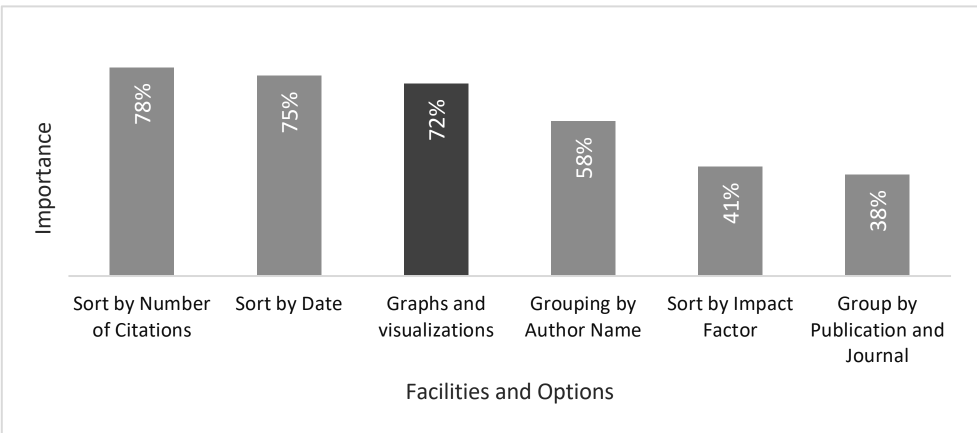

Figure 1.8 Importance of graphs vs other facilities in academic databases