A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval

By

Feng Shi

Imperial College London, Dyson School of Design Engineering

A thesis submitted for the degree of Doctor of Philosophy

July 2018

Abstract

Data-Driven Design is an emerging area with the advent of big-data tools. Massive information stored in electronic and digital forms on the internet provides potential opportunities for knowledge discovery in the fields of design and engineering. The aim of the research reported in this thesis is to facilitate the design information retrieval process based on large-scale electronic data through the use of text mining and semantic network techniques.

We have proposed a data-driven pipeline for design information retrieval including four elements, from data acquisition, text mining, semantic network analysis, to data visualisation and user interaction. Web crawling techniques are applied to fetch massive online textual data in data acquisition process. The use of text mining enables the transformation of data from unstructured raw texts into a structured semantic network. A retrieval analysis framework is proposed based on the constructed semantic network to retrieve relevant design information and provoke design innovation. Finally, a web-based platform B-Link has been developed to enable user to visualise the semantic network and interact with it through the proposed retrieval analysis framework.

Seven case studies were conducted throughout the thesis to investigate the effectiveness and gain insights for each element of the pipeline. Thousands of design post news items and millions of engineering and design peer reviewed papers can be efficiently captured by web crawling techniques. Through the use of itemset mining and noun phrase chunking, a semantic network constructed based on these textual data is shown to capture more inherent design- and engineering-oriented concepts and relations, compared to the benchmarking approaches: WordNet, ConceptNet, NeLL and Wikipedia. A retrieval analysis framework has been developed with different retrieval behaviours to retrieve either common general or domain-specific concepts, explicit or implicit knowledge relations, which are found to satisfy various knowledge demands in our real design projects at the conceptual stage. Finally, the result of a user test is shown to be consistent with these findings.

Declaration

The coding for web-software B-Link was accomplished in collaboration with Liuqing Chen. Except where otherwise stated, the work in this thesis is my own. Parts of this thesis have been disseminated through conference or journal publications and are reused according to the respective copyright agreements. This is noted at the beginning of a chapter where it occurs.

Feng Shi

July 2017

List of Publications

This work has resulted in or contributed to the following publications at the time of submitting this thesis.

Full Journal Papers:

Shi, F., Chen, L., Han, J. and Childs, P., 2017. A data-driven text mining and semantic network analysis for design information retrieval. Journal of Mechanical Design, 139(11), p.111402. doi:10.1115/1.4037649.

Han J., Shi F., Chen, L., Childs P. R. N., 2018. The Combinator – A computer-based tool for creative idea generation based on a simulation approach. Design Science. 4, p. e11. doi: 10.1017/dsj.2018.7.

Han J., Shi F., Chen, L., Childs P. R. N., 2018. A computational tool for creative idea generation based on analogical reasoning and ontology. Artificial Intelligence for Engineering Design, Analysis and Manufacturing. doi:10.1017/ S0890060418000082.

Full Conference Papers:

Chen, L., Wang, P., Shi, F., Han, J. and Childs, P., 2018. A computational approach for combinational creativity in design. In DS92: Proceedings of the DESIGN 2018 15th International Design Conference (pp. 1815-1824).

Shi, F., Chen, L., Han, J. and Childs, P., 2017, August. Implicit Knowledge Discovery in Design Semantic Network by Applying Pythagorean Means on Shortest Path Searching. In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (pp. V001T02A053-V001T02A053). American Society of Mechanical Engineers.

Chen, L., Shi, F., Han, J. and Childs, P.R., 2017, August. A network-based computational model for creative knowledge discovery bridging human-computer interaction and data

mining. In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (pp. V007T06A001-V007T06A001). American Society of Mechanical Engineers.

Han, J., Shi, F., Chen, L. and Childs, P., 2017. The Analogy Retriever–an idea generation tool. In DS 87-4 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 4: Design Methods and Tools, Vancouver, Canada, 21-25.08. 2017.

Shi, F., Han, J. and Childs, P.R.N., 2016. A Data Mining Approach to assist design knowledge retrieval based on keyword associations. In DS 84: Proceedings of the DESIGN 2016 14th International Design Conference (pp. 1125-1134).

Han, J., Shi, F. and Childs, P.R.N., 2016. The Combinator: a computer-based tool for idea generation. In DS 84: Proceedings of the DESIGN 2016 14th International Design Conference (pp. 639-648).

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to first thank my supervisors Peter Childs and Marco Aurisicchio for providing the full support throughout this research. Peter, many thanks for your continuous enthusiasm, encouragements, promotion, and guidance on this work and also for providing freedom to explore and provoke my new ideas and thoughts during the research. Marco, thank you for showing the directions and feedbacks at every key stage of my PhD research.

Thank you very much to all my friends and colleagues in the Dyson school of Design engineering. Specially, I would like to thank Ji, Liuqing, Dongmyung, Min, and Ravi who have expanded my knowledge on design research methodology and product development, and collaborated with me on coding software and many other product design projects.

Thanks to China Scholarship Council (CSC) for providing the scholarship throughout the study, and thanks to Elsevier and Yanko design for providing the APIkey and data resources for web crawling.

Many thanks to my parents, who are always standing behind me.

Finally, I would like to thank Di who are always beside me, and Dati who is learning to walk in front of me.

Chapter 1 Introduction

In March 2016, something called 'AlphaGo' beat the human-world champion Lee Sedol in a five-game match in the game of Go. This become the biggest news in the whole world on that day. However, what we may not know is that, actually before the game with Lee Sedol, AlphaGo had already obtained the human knowledge by learning from millions of human moves and thousands of human-played games (Silver et al., 2016). The fundamental key for the success is the use or reuse of the huge amount of historical game data, even though the algorithm had a key role in the process. This example shows the power of huge amount of data in discovering new knowledge in the ancient game of GO.

Besides games, the power of data is now being used in various fields from science, engineering, medical care, to economy, social issue and our everyday lives. IBM's Watson can make much more accurate diagnosis on medical conditions than an expert (Chen et al., 2016). Tweets and News information are used to predict the trends of economic / stock market (Zhang et al., 2011). Amazon recommends relevant products based on our interests (Linden et al., 2003). Google Maps leads us to anywhere in a city. Security police officers use historical record data for crime detection (Nath, 2006). Biological data are used by researchers for discovering new drugs and protein interaction (Cohen and Hersh, 2005).

Big data approaches are now reshaping methodology, research, technology and development in many different fields. So How about its impacts on design? How does the huge amount of data affect the design engineering field? What is its impact in design? Can we utilize the power of data in facilitating and helping with the broad area of design involving design process, design knowledge/information, user

experience, conceptual ideation? The answer are positive and there are already many emerging research activities on this issue which is called Data-Driven Design (D^3).

1.1 Data-Driven Design

Data-driven design emerges with the rise of Big-Data economy, and it consists of three key components: **Data**, **Driven**, and **Design**. **Data** refers to the huge amount of data resources we can utilize such as user data, textual data, web data. **Driven** means the state-of-art data analytic tools and advanced technologies used to gather, analyse, interpret, and visualise the data including machine learning, data mining, natural language processing, internet of things and so on. **Design** points out the applications which involve different aspects of the broad design field such as user experience, design optimisation, and design information retrieval, on which we want to contribute and make improvement. In one sentence, data-driven design is to utilize the huge amount of data with advanced data analytic tools to help, improve and facilitate the diverse design aspects and activities.

Data

With the advent and rise of Cyber-Physical Systems or Internet of Things, there are various types of human-machine generated data resources that potentially contain valuable uncovered insights and knowledge that can create unprecedented challenges and opportunities to improve the theory, methods, tools and practices in the design field (Kim et al., 2017). Figure 1-1 shows some of the data resources available on social media and the Internet. One typical use case is to use the online customer reviews for improving the product features and services. Chaklader and Parkinson (2017) developed a novel approach which uses text mining on a large amount of consumer reviews to discover the most significant information variability for human-artifact interaction. This approach can quickly and economically help provide useful information to establish preliminary design specifications. Lim and Tucker (2017)

use the online product ratings to offer the customers more objective and accurate feedbacks by lowering the user rating bias caused by customers' optimism or pessimism. By applying opinion mining, product features can also be extracted from the online customers' reviews, then the future importance of product features can be predicted and determined through fuzzified frequency and sentiment scores (Jiang et al., 2017).

Figure 1-1 Data resources on social media and internet

Driven

Benefiting from the booming development of data science communities in recent years, many off-the-shelf data analytic tools are available to be readily applied on the large-scale data resources for design and engineering purposes. Figure 1-2 gives an interesting word cloud of the current most popular data analytic tools including text mining, natural language processing, data mining, and machine learning. As discussed above, one of the most commonly used tools for D^3 is text mining as well as natural language processing, since most of the electronic and digital data are texts in nature (Ur-Rahman and Harding, 2012). Lan et al. (2017) proposed an approach based on deep belief net (DBN) to automatically discover design tasks and quantify their interactions from design email archive, where the deep belief neural network is

used to learn a set of latent topic features from a simple word-frequency based input representation of the document. Dong and Agogino (1997) used natural language processing to induce a representation of design based on the analysis of syntactic pattern contained in the corpus of design documents. They first recognised the noun-phases from the documents and conducted clustering to investigate the inter-dependency between the terms. Then a Bayesian belief network was constructed that can describe a conceptual hierarchy for the specific domain of the design.

In addition to text mining and natural language processing, other techniques such as image processing and IoT are also becoming popular and have potential to be used in Data-Driven Design. A deep learning approach based on three-dimensional convolutional neural network has been applied to predict functional quantities of digital design concepts and also to discover the latent features of the products (Dering and Tucker, 2017). With the increasing amount of sensors applied on IoT, mobile and wearable devices, a huge amount of time-series measurements on the users' physical activities are generated. This user behaviour data can be extremely valuable to help improve the design of many beneficial applications such as health monitoring or activity recognition (Malekzadeh et al., 2018).

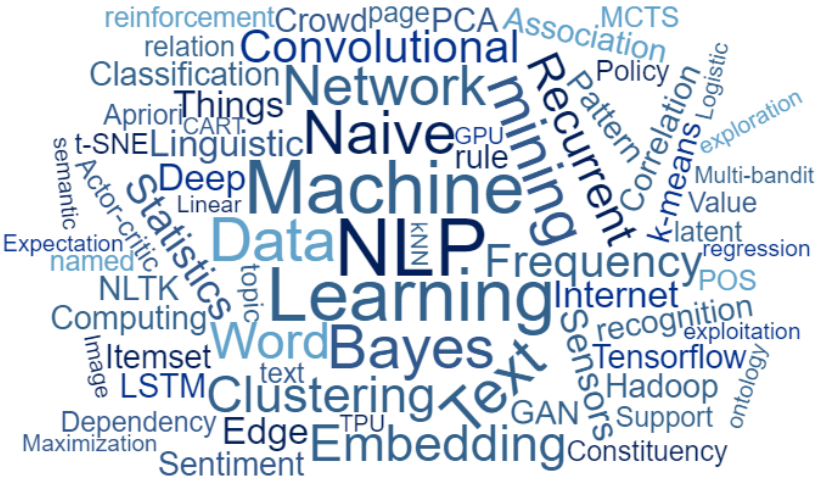


Figure 1-2 Popular data analytic tools

Design

Various and different aspects in the design field such as conceptual idea generation, design innovation, design information and knowledge, design optimisation, and design practice and assurance, are currently benefiting from the emergence of the large-scale data and advancements in data analytic tools, as mentioned above.

For conceptual design ideation, some studies, by leveraging the large amount of image data resources on the internet, have tried to generate random and relevant image stimuli relevant to the design query in order to provoke the designers' creativity for novel ideas (Han et al., 2018a, Han et al., 2017). Computer vision techniques such as generative adversarial networks have also been applied to generate combinational synthetic images of two separate design concepts to help with design innovation (Chen et al., 2018).

Design information retrieval and knowledge management, which are the material basis and driven force for conceptual ideation, are also a focusing application in Data-Driven Design. Luo et al. (2017) have tried to provide designers and engineers with valuable and useful design information by using patent mining techniques to identify technological neighbourhoods, where graph models are applied to analyse the proximity of patent domains. They hope to use this method to discover future design and technology opportunities and directions. Similarly, Song and Luo (2017) also focus on using data-driven patent mining approach to facilitate design information retrieval, where an iterative and heuristic methodology is proposed to exhaustively search for patents which are the precedents of the design of a specific technology or product in order to explore next generation design. In addition, data-driven approaches can also contribute to the automatic construction of design knowledge representation like design ontology. Cheong et al. (2017) developed a method to automatically extract functional knowledge relations from natural language text, which uses syntactic rules to parse subject-verb-object forms out of the text. The subject-verb-object forms can be formatted as semantic triples for the following ontology construction.

Other related design research areas where the communities are promoting and applying data-driven approaches, also include design optimization, design assurance and practice, as well as user experience. Experimental data are collected and analysed by machine learning algorithms to model the relationship between design parameter space and system response space, which can be further used by designers to perform the design optimisation (Chattopadhyay et al., 2017). Design assurance and practice is guaranteed by learning from the designers' every action data during the design process, where data mining approaches are applied to quantitatively study the processes from configuration of the design problem to finalisation of the design solution (McComb et al., 2017). User experience and Customers' satisfaction are identified and predicted by analysing the collected operating data from Smartphones of a large number of users (Zhang et al., 2017).

Therefore, we can see that "big" data approaches have already been actively utilised and had impact in the design field. However, data-driven design is still a very broad concept that we cannot fully cover here. Before focusing and narrowing down to a much more specific topic in data-driven design, we need to ask ourselves three questions corresponding to **Design**, **Data** and **Driven**:

- **Design**: what exact aspect or problem in design we want to contribute and solve?
- **Data**: what type of data resources are available and relevant, and can help for this aspect of design?
- **Driven**: which data analytic tools are appropriate to be applied to harness and exploit the power of this type of data to advance this aspect of design?

1.2 The Aim of this Thesis

The aim of this thesis can be more tightly defined by answering the above three key questions:

(1). **Design**: what exact or specific aspect in design we want to contribute to or solve?

Engineering design is a knowledge-intensive process. Various areas of knowledge and expertise are utilized in conducting each stage of the design activity including conceptual design, embodiment design, and detailed design (Bertola and Teixeira, 2003). However, during the design process, engineers usually spend a significant amount of time, about 60%, in searching for the right information among the highly diverse and unstructured knowledge resources (Ullman, 2002). Hence, effectively storing, managing, and retrieving design knowledge is one of the major issues for enterprises and industry to reduce the product development life-cycle time and costs, as well as to increase the quality and innovation elements of the product (Chandrasegaran et al., 2013). Therefore, this thesis aims to contribute to design information retrieval during design and engineering process by using data-driven approaches. Specifically, we will focus on the creative idea generation in conceptual design process, and the information retrieval is conducted by using concepts and relations as the basic elements.

(2). **Data**: what type of data resources are available and relevant, and can help for this aspect of design?

With the advent of the Big-Data economy, increasing amounts of large-scale data are generated, which provides an abundant resource base for knowledge mining and discovery in design field (Tuarob and Tucker, 2015, Ma and Kim, 2014). Since 80% of the available large-scale electronic and digital data are texts in nature (Ur-Rahman

and Harding, 2012), this thesis will focus on the textual data of electronic documents to facilitate the design information retrieval.

(3). **Driven**: which data analytic tools are appropriate to be applied to harness and exploit the power of this type of data to advance this aspect of design?

The capability to effectively harness the potential power of massive textual data provides opportunities to improve traditional design knowledge retrieval methods (Ishino and Jin, 2001). However, data are just the description of raw facts, and most of these human-and-machine-generated data are often highly unstructured and heterogeneous. It is necessary to process the raw textual data into information, which is the structured representation of data within specific context or usable format, and then people can acquire knowledge through understanding and learning of the information. Therefore, how to transform the large-scale unstructured data into structured information and then consequently discover meaningful knowledge is the main challenge for modern data-driven design information retrieval and knowledge discovery (Li and Ramani, 2007). Benefiting from the booming development of data science communities in the last two decades, many data analytic tools such as data mining, machine learning, association rule mining, natural language processing (NLP), network analysis and graph science have become available to be applied in design knowledge discovery for various purposes(Sangelkar and McAdams, 2012, Lan et al., 2017). Thus, investigating ways to adapt and apply these analytic tools on the massive textual data for design information retrieval is the key focus in this thesis. Text mining/NLP is currently the main method for information extraction and knowledge discovery used on raw text data (Allahyari et al., 2017). It can also automate the construction process of ontology and semantic network, which on the other hand, is one of the most efficient methods to represent and conceptualize the design knowledge (Chandrasegaran et al., 2013). Therefore, we can integrate text mining combined with ontology / semantic network to exploit the power of massive raw text resource for design information retrieval process.

The Aim

In conclusion, the aim of this thesis is to facilitate the design information retrieval process through the application of text mining (natural language processing) and ontology (semantic network) on "Big" textual data. As shown in Figure 1-3, the relations between data resource, text mining/NLP, Ontology/semantic network, and design information/ knowledge can be described as:

- Raw textual data act as the data resources for text mining /NLP.
- Text mining/NLP can realise the automatic construction of ontology and semantic network.
- Ontology / semantic network can effectively represent and conceptualise design knowledge and information.

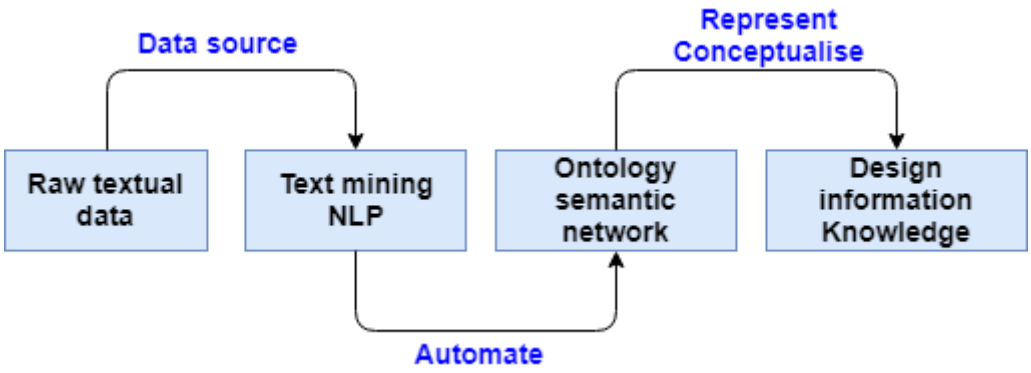


Figure 1-3 The relations between data, driven, and design in this thesis

1.3 Overview of the thesis

The motivation of this thesis corresponds with the recent emerging research activities in Data-Driven Design (D^3). With the advent of Big Data economy, the vision of D^3 is to investigate and explore the potential impacts and benefits that the large-scale human-machine generated data can have on the broad design field.

While data-driven design is a broad area, this thesis only focuses on a specific facet for each component of D^3. For data, this thesis solely relies on textual data. For analytic driven tools, we use text mining/NLP, and ontological/semantic network. Finally for design, we mainly focuses on the design information retrieval aspect of design. Therefore, the main contribution of this thesis is utilising text mining with ontological network techniques on huge amount of versatile textual data in order to facilitate the design information retrieval during design process and engineering activities.

Table 1-1 shows the overview of this thesis. A brief description of each chapter is provided.

Chapter 1 introduces this thesis under the background of Data-Driven Design. Here, the specific aim of this thesis is to harness the power of large-scale textual data to facilitate the design information retrieval. Text mining and ontological techniques are indicated to be the intermediate analytic driven tools to utilise the raw text for the benefits on the design information retrieval.

Chapter 2 explores the state-of-art applications of text mining and ontology techniques and their relations with the field of design information retrieval and knowledge representation.

Chapter 3 describes our practical data acquisition from Internet, which is the first element in our data-driven pipeline. A web crawling technique is implemented to harvest the raw textual data from Internet through two studies. Our first study (Study 1) focuses on the design news website where a thousand of design posts from the past five years are crawled. The second study (Study 2) uses academic literature in design and engineering fields as data resources where millions of papers are captured from Elsevier. The data captured in this chapter are used as our raw text sources for the following data analysis and information retrieval.

Chapter 4 transforms the unstructured textual data captured in chapter 3 into a structured semantic network. Specifically, noun phrase chunking techniques and academic keywords are used to recognise the knowledge concepts. Association rule and itemset mining on statistical occurrence frequency are used to build and represent the strength of the knowledge associations/relations between concepts. Disparity filter is implemented to remove the noise of the constructed semantic network. In order to evaluate whether the constructed semantic network can indeed capture more design-and-engineering-oriented knowledge concepts and relations, test is carried out to compare our semantic network with other three benchmark systems, namely, WordNet (Miller, 1995), ConceptNet (Speer et al., 2017) and NeLL (Carlson et al., 2010) (Study 3).

Chapter 5 proposes a design information retrieval framework based on the constructed semantic network constructed in chapter 4. In this retrieval framework, various novel criteria are developed to retrieve and rank both explicit and implicit associations under the unified standard by modelling probability, velocity layer and applying Pythagorean means. Study 4 evaluates the probability and velocity layer modelling, which are shown to be able to effectively quantify the correlation degree between concepts and yield results consistent with the human judgement on golden relations and none relations. Two detailed case studies (Study 5, Study6) are conducted to illustrate the practical use of the proposed framework and semantic network in real design information retrieval process, and also to investigate the retrieval behaviours of different criteria. Study 5 illustrates the process to explore information around a single concept of the design query, while Study 6 gives an example on retrieving knowledge relations between two concepts of the query.

Chapter 6 addresses on the final element of the data-driven pipeline: data visualisation and user interaction. A web platform B-Link is developed to enable the user to visualise our data and interact with the semantic network through the framework proposed in chapter 5. Besides, a user online test (Study 7) is designed

and conducted to investigate the users' perceptions on the retrieved knowledge associations from different aspects.

Chapter 7 gives an overall discussion of this research. This chapter summarises the key contributions of this research on what it means to the design information retrieval, and how this work explored new usage of the textual data to facilitate this area. It also shows that the future directions for this work.

Table 1-1 Overview of this thesis

Chapter	Objective	Key contribution(s)
One	Introduction	• Corresponds to emerging Data-Driven Design • Focus on textual data, text mining and ontological techniques, and design information for *Data*, *Driven* and *Design* respectively.
Two	Explore the use of text mining, ontology in design	• General application of text mining • The relation between text mining and ontological / semantic network • The use of text mining and ontology in design information retrieval and knowledge representation
Three	Identify data resources and acquisition methods	• Implement web crawler on texts resources from internet • Harvesting design news (Study 1) and academic literatures (Study 2)
Four	Transform unstructured raw text data into structured semantic network	• Build a semantic network specifically for engineering and design from scratch • Benchmarking with other public ontological databases (Study 3)
Five	Propose a design information retrieval framework based on the semantic network	• Develop various criteria to quantify the correlation degree of both explicit and implicit knowledge associations under a unified standard. • Evaluate the criteria by golden relations (Study 4) • Design information retrieval around a single concept (Study 5) and between two concepts (Study 6)
Six	Realise data visualisation and investigate user test	• B-Link platform to visualise and interact with the data • Investigate user perceptions on the retrieved information (Study 7)

Seven	Present conclusion and suggest future work	• Insights from the study and directions for future work

Chapter 2 Text, Ontology, and Design Information

Engineering design is a knowledge-intensive process. The acquisition, reuse and sharing of design knowledge are major facets that help to guarantee the smooth progress of every stage in design processes and activities. Design information retrieval, which is the most common approach to satisfy the intensive knowledge demands during design process, has already been studied over the last several decades with the arising of the science of information retrieval (IR) (Larson, 2010).

Design information retrieval is to search, gather and refine useful and relevant information from massive unstructured data to be used for help solving design queries and meeting the knowledge demands (Yang et al., 2005). The traditional design information retrieval approach is document-based retrieval by using text mining, where information is retrieved at fragment and document level (Salton and Buckley, 1988).

However, in recent years, many research teams and organisations have become more interested in retrieving and processing the information at semantic and concept level, where the ontological technique plays an important role in this progress (Gruber, 1995). Ontological structure provides a much more refined way to represent and conceptualise the specific knowledge concepts and relations between the concepts. This make it feasible for designers to retrieve the information at a semantic level based on the inherent meaning of design and engineering relations rather than the fragment or document level based on naïve keyword matching and full-text

indexing. This thesis is actually targeted towards developing an approach to support with the information retrieval based on semantic and concept level through the use of text mining and semantic network.

To construct the ontology, both manual and automatic approaches have been conducted. Manual approach uses hand-made pieces of information, and is usually suitable for building the ontology for a specific knowledge domain or project. However, the process is time-consuming and requires a huge amount of human efforts and expertise which is a non-trivial task and limited within a single knowledge domain (Li et al., 2008). In a digital information era, manually constructing ontology from the huge amount of data becomes impossible and not scalable. With the booming development of data science communities, techniques such as NLP and information extraction (IE) in the text mining area help to make it possible to automate the construction process of ontology based on the raw text from scratch.

As discussed above, we can see there exist interesting relationships between any two subjects, text mining, ontology and design information, as shown in Figure 2-1. In this chapter, we will explore the review of the relationship between text mining, ontology and design information retrieval, and find the gaps that suggest directions in our following research work.

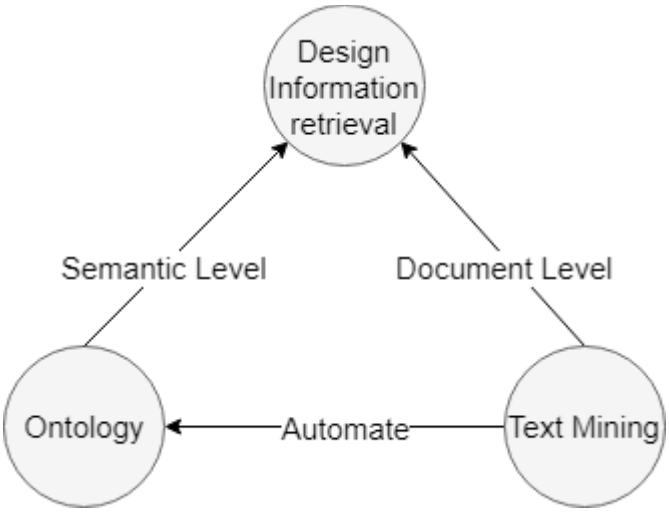


Figure 2-1 The relationship between design information retrieval, ontology and text mining

2.1 Text mining and the use in design and engineering

Text mining has received a great deal of attention in recent years, and there has been a substantial growth of its applications in various fields including both academic research and industrial as. This is mainly because that tremendous amount of textual data that are generated by human and machine everyday around the world, can potentially contain an invaluable source of hidden information and knowledge.

A report provided by EMC (Gantz and Reinsel, 2012), indicates that the amount of textual data will be over 40 zettabytes by 2020, which is fifty times of the amount in 2010. Imagine how much meaningful information, patterns, useful insights, and of course knowledge would be hidden within this huge amount of unstructured texts that could be beneficial to a variety of domains from technology, engineering and science to the humanities, social sciences, and education. Thus, there is an driver to develop methods and algorithms in order to effectively discover and extract the structured meaningful information and knowledge from the massive unstructured raw texts (Allahyari et al., 2017).

Text mining is a collection of such methods and algorithms designed to extract information and discover knowledge from texts based on different purposes. It involves a variety of specific tasks, such as, preprocessing, document retrieval, classification, topic modelling, clustering and information extraction (IE). Here, we will briefly discuss some of the most widely used text mining algorithms as well as their common applications in design information retrieval.

Pre-processing

Text preprocessing is the first step, but also a very important procedure in many text mining tasks. Since raw texts are highly unstructured and contain much noise data, pre-processing is to clean and reformat the text with better quality for the following feature engineering, which will have the significant influence on the performance of the algorithms. Text preprocessing usually consists of components such as string cleaning, tokenization, stop-words removal, lemmatisation and stemming:

- String cleaning: some text resources from social media (e.g. twitter) usually contain unwanted or useless characters that does not contribute any fundamental meanings to the text, such as hashtag #, email address, and web links which are desired to be removed. Regular expression is usually conducted to detect and filter these unwanted characters from the string.
- Tokenization: Given a sequence of text, tokenization is the task of chopping it up into pieces, called *tokens*. Tokenization can be conducted at different levels on the text, so a chopped token can either be a sentence, a word or a phrase. Certain characters can be discarded during this process, such as punctuations (Webster and Kit, 1992).
- Stop-word removal: the task is to remove the stop words from the texts, which are the words without having much meaningful information in the document. Three types of words are commonly considered as stop-words: (1). prepositions, conjunctions, pronouns, articles, etc. (2). Words occurring very often in all the documents, which means that they have little information to distinguish between different documents. (3). Words occurring very rarely in the document: this also indicates that they are of no significant relevance with the document (Saif et al., 2014, Silva and Ribeiro, 2003).
- Lemmatization: This is to get the basic dictionary form of the word. Due to morphology of words, a word may have various inflected forms, for example, “cat” and “cats”, “take”, “took”, and “taken”. It would be useful to use morphological analysis to transform the variants of a word into its basic form

as a single group in order to improve the accuracy and efficiency for following analysis. The specific process refers to getting the singular form of nouns, and the infinitive tense of verbs.

- Stemming: The purpose of stemming is the same as lemmatization, while the method it uses is much straight-forward. It's to get the base/root of the word which usually refers to a crude heuristic process that chops off the ends of words. Several stemming algorithms have been developed and are currently available in Natural Language Toolkit (NLTK) (Bird and Loper, 2004) , such as Porter stemmer, Lancaster stemmer, and Snowball stemmer (Porter, 1980).

Document retrieval

Document retrieval is one of the most common tasks in text mining for a wide variety of applications. It's mainly used for searching the relevant documents or text fragments to meet the requirements of users' queries.

In design and engineering fields, document retrieval is the major approach for traditional design information retrieval, that engineers depend mostly on keyword searches to identify textual documents for retrieving information to execute specific tasks in each design stage (McMahon et al., 2004). This leads to both the need for supporting the management of text documents and the need for unified system for the storage, classification, and retrieval of these documents, especially in initial stages of design. Traditional document retrieval systems are designed to follow full-text indexing and storage of information in a relevant manner (Chandrasegaran et al., 2013). Some of these systems provide additional engineering data organization and team collaboration information as well (Homer et al., 2002, Chen and Jan, 2000).

For an ordinary document management system, indexing the document is the key part to guarantee effective information retrieval. The vector space model (VSM)

(Salton et al., 1975) was the first and the most well-known way for document indexing. In a vector space model, structured representations of documents are generated by converting the unstructured text of each document into a numeric vector. Specifically, with a corpus of documents $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$, $\mathcal{V} = \{w_1, w_2, \dots, w_v\}$ is the set of vocabulary contained in this corpus. Let $f^d(w)$ be the weight of the term w for document d, for example, denoting whether term w occurred in document d with 1 being occurred and 0 being not occurred, then a document can be simply represented as a vector with Boolean values:

$$\boldsymbol{d} = [f^d(w_1), f^d(w_2), \dots, f^d(w_v)] \tag{2-1}$$

This is the naïve version of vector space model where the weight being 1 or 0 refers to whether the term occurs in a document. There are also other weight configurations to get more advanced VSM:

- Term frequency (TF) weighting: let the weight $f^d(w) = f_w^d$ just be the occurrence frequency of term w in document d.
- Term frequency inverse document frequency (TF-IDF) weighting (Salton and Buckley, 1988): the weight is calculated as:

$$f^d(w_i) = \frac{f_{w_i}^d}{\sum_{k=1}^{v} f_{w_k}^d} \cdot \log \frac{n}{F_{w_i}} \tag{2-2}$$

  where $f_{w_i}^d$ is the occurrence frequency of term w_i in document d, v is the size of vocabulary $\mathcal{V}$, n is the total number of the documents in the corpus, and F_{w_i} is the number of document which contain term w_i.

With TF or TF-IDF configurations, by setting a significance weight threshold, a set of significant terms/concepts beyond the threshold can be extracted and regarded as the most informative keywords/labels for each document. Therefore, the documents can be indexed by the keywords/labels with the highest weights. During design document retrieval process, relevant documents can be accordingly retrieved and

ranked based on the weights of the corresponding keywords matching with the user queries (Yu and Hsu, 2013).

The vector space model is also called "bag-of-words" model since it only considers the appearance of the words but ignore the order of the words in the text. However, despite its simplicity, VSM performs quite well in document retrieval and has also been broadly used in other text mining algorithms as shown below.

Text Classification

Text classification is another common task in text mining, which has a vast number of application in diverse fields. Instead of retrieving the relevant document for the user query in document retrieval, text classification aims to assign the predefined classes/labels to the documents (Mitchell, 1997). Text classification is a supervised learning task, where a training set $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ of documents with known class labels $\mathcal{L} = \{l_1, l_2, \dots, l_n\}$ are used to train a classification model $\mathcal{M}: \mathcal{D} \rightarrow \mathcal{L}$, which can be further used to predict the class/label of unknown new document.

There are many applications of text classification in industrial fields, such as email spam identification to determine whether an email could be spam or not spam (Blanzieri and Bryl, 2008), sentiment analysis on the movie review which aims to obtaining the feedback of the audience (Singh et al., 2013), and automated event/crime detection (Shen et al., 2006). In design and engineering domain, text classification is widely used for project document foldering and document categorisation in the information management system. In order to organize and improve the access to information stored in the system for the designers and engineers, engineering documents are usually needed to be classified and stored hierarchically into different categorisation according to the stages of the design process and engineering practice, types of knowledge requirements and demands, as well as the project divisions and components. Caldas and Soibelman (2003)

proposed a methodology to improve information organization and access in engineering construction management information systems by developing an automatic hierarchical text classification approach on engineering construction project documents according to project components.

In addition to document categorization in information management systems for large engineering projects, text classification has also been studied in product design for product feature analysis and innovative product development. One popular example is to use sentiment analysis on the customer's reviews on the product in order to investigate the importance of different product features and further predict the promising product feature in the future (Jiang et al., 2017). This is because sentiment analysis is fundamentally a text classification problem (Liu and Zhang, 2012). In a simple case, the problem is just to classify a piece of text into two classes/labels: Positive and Negative, while in a more advanced scenario, fine-grained sentiment levels are applied as class labels. Another use case is to apply text classification on the online product textual descriptions of e-commerce (Schulten et al., 2001) and the product technological patents (Liang and Tan, 2007). This helps discover the recent design trend and technology direction to show the guidance for innovative product development.

The traditional algorithms of text classification include: naïve Bayes, support vector machine (SVM), and k-nearest neighbour. All these algorithms are developed by using the features of the documents as input and labels of the documents as output. The feature of a document is literally a vector representation of the document that include all the useful information that can have potential impacts on the output label of the document. This vector representation, also called feature vector (Guyon and Elisseeff, 2003), is a concatenation of all these pieces of useful information where every piece of information is converted into a value and represented as a single element in the feature vector. We can integrate any features that we think may be useful into the feature vector, such as the words of the document, the sentiment of

the document, the author or even the time. The most commonly used feature for text classification is the vector produced by vector space mode (VSM) or bag-of-word (BOW) model, where each element indicates the TF, TFIDF or appearance of a particular term. It is usually directly used as the feature vector or part of the feature vector to represent the document.

K-nearest neighbour is the most simple algorithm (Han et al., 2001). It completely depends on the feature vector space to determine the class of the documents. For a new document d_x, the idea is to search its k nearest (most similar) neighbours within the documents in training set, where the similarity between the new document d_x and a document d_i in training set can be represented by the cosine similarity:

$$S(d_x, d_i) = \frac{\boldsymbol{f}_{d_x} \cdot \boldsymbol{f}_{d_i}}{\|\boldsymbol{f}_{d_x}\| \|\boldsymbol{f}_{d_i}\|} \tag{2-3}$$

where $\boldsymbol{f}_x$ and $\boldsymbol{f}_i$ is the feature vector of document d_x and d_i respectively. Then the most common label of these k nearest neighbours is regarded as the class of this new document d_x.

The naïve Bayes algorithm (McCallum and Nigam, 1998) uses the Bayes rule to calculate the probability of each label given the features of the document. With the assumption that every feature element is conditionally independent, the probability of a document belonging to class l_i can be shown as:

$$\begin{aligned} P(l_i|\boldsymbol{f_x}) &= \frac{P(\boldsymbol{f_x}|l_i)P(l_i)}{P(\boldsymbol{f_x})} \\ &= \frac{\prod_{k=1}^{m} P(f_k|l_i)\, P(l_i)}{P(\boldsymbol{f_x})} \end{aligned} \tag{2-4}$$

where $\boldsymbol{f_x}$ is the feature vector of the document, and f_k is the k^{th} element of the feature vector. By using equation (2-4), we can compute the probability of every label

for the document, and therefore the label with maximum probability can be selected as the predicted class of the document.

Support vector machine (SVM) (Cortes and Vapnik, 1995) is another powerful algorithm extensively used in text classification. It was originally developed as a binary linear classifier similar to logistic regression. Let y be the binary label of 0 or 1, the objective function in training process is to optimise:

$$\min_{\boldsymbol{\theta}} C \sum_{i=1}^{n} [y_i \max(0,1 - \boldsymbol{\theta}^T \boldsymbol{f}_i) + (1 - y_i) \max(0,1 + \boldsymbol{\theta}^T \boldsymbol{f}_i)] + \frac{1}{2} \sum_{j=1}^{n} \theta_j \qquad (2\text{-}5)$$

where C is a constant, $\boldsymbol{\theta}$ is the parameter of SVM model to be optimized, n is the size of the training set, y_i and $\boldsymbol{f}_i$ are the label and feature of document d_i respectively. In testing on an unknown document with feature $\boldsymbol{f}_x$, we will assign its class to be 1 if $\boldsymbol{\theta}^T \boldsymbol{f}_x > 0$, and 0 otherwise. For more advanced usage, we can use kernel methods to adapt SVM for nonlinear problem (Cristianini and Shawe-Taylor, 2000). Multiple SVM classifier can be created for multi-class classification where each SVM only distinguish one class from the rest (Weston and Watkins, 1998).

In addition to the above traditional algorithms, recently, many deep learning models including convolutional neural networks and recurrent neural networks (Lai et al., 2015) have also been applied and achieved great performance in text classification.

Clustering and Topic modelling

Differing from text classification, both clustering and topic modelling are unsupervised learning methods, which discover the insights from the raw data and input features without the need of any labels. The aim of clustering is to generate clusters/communities of the documents by grouping similar documents together. It is often called hard clustering problem where each document can belong to one cluster only. In topic modelling, we assume there exist a number of latent topics

within the corpus, and every document has a probability distribution over all the topics. Topic modelling is often referred as soft clustering problem as opposed to the hard clustering.

In design and engineering fields, text clustering and topic modelling techniques have been recently applied for the knowledge reuse and exploitation from a variety of past textual footprints of design and engineering activities such as emails, regular reports, change logs as well as different forms of social media. Grieco et al. (2017) applied clustering on the natural language text written on the Engineering Change Request (ECR), which is a type of engineering design log to record the frequently required change to redesign and alter the products and their components during the design process. The use of clustering on ECR can significantly help summarise the main features and changes added from the product development process of previous projects. Similarly, topic modelling has been implemented to analyse the past design emails archive in order to uncover design tasks and quantify their interaction (Lan et al., 2017). This helps the designers learn for the past, for example, what design tasks are actually carried out, and how they impact each other.

There are two major algorithms for hard clustering, namely, hierarchical clustering and k-mean clustering. In hierarchical algorithms, the documents are grouped into a hierarchical structure of clusters which can be constructed either through a top-down manner (divisive) or bottom-up manner (agglomerative) (Allahyari et al., 2017). A detailed introduction for hierarchical clustering can be found in (Murtagh, 1983). K-mean method (Bradley and Fayyad, 1998) is another more commonly used clustering algorithm that partitions all the n documents of the corpus into k clusters where the number of clusters k is predefined. The initial choice of k is very important for the final representations of the clusters and it is often determined through empirical and heuristic approaches (Alsabti et al., 1997, Kanungo et al., 2002). The procedures of k-mean algorithm can be shown as follows:

1) Randomly select k documents as the centroids of the k clusters
2) Assign every document to its nearest centroid based on the similarity measurements such as cosine similarity in Equation (2-3).
3) Calculate the new centroid for each cluster.
4) Repeat between step 2) to step 3) until convergence when the centroids of all the k clusters become stable.

However, in topic modelling, instead of grouping documents into separate groups, the aim is to extract the thematic information (topics) from the whole corpus. Topic modelling is a probabilistic unsupervised learning algorithm. Its probability assumption is that there exist K latent topics within the whole corpus where each document in the corpus is represented as a probability distribution over the K topics, and each topic is represented as a probability distribution over the vocabularies. The state-of-art algorithm for topic modelling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which assumes probability distributions of document-over-topics and topic-over-words follow Dirichlet distribution. Therefore, the process to generate the corpus $\mathcal{D}$ can be described as:

1) For each topic t_k where $k \in [1,2,\dots,K]$, we draw a probability distribution over the vocabulary: $\boldsymbol{\phi}^k \sim Dir(\boldsymbol{\beta})$. $\boldsymbol{\phi}^k$ is a v dimensional vector where v is the size of the vocabulary. $\boldsymbol{\beta}$ is the hyperparameter for topic-over-words Dirichlet distribution.
2) For each document d_i:
   - We firstly draw this document's probability distribution over the topics: $\boldsymbol{\theta_i} \sim Dir(\boldsymbol{\alpha})$, where $\boldsymbol{\theta_i}$ is a K-dimensional vector and $\boldsymbol{\alpha}$ is the document-over-topics Dirichlet distribution.
   - Then for each word position j in this document, we sample a topic assignment $z_j \sim multi(\boldsymbol{\theta_i})$, and then sample a word based on this topic $w_j \sim multi(\boldsymbol{\phi}^{z_j})$.

The key inferential problem in LDA is to estimate the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ where a variety of methods have been provided such as variational inference (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004).

Information Extraction

Last but importantly, information extraction is the key technique we will be discussed in the following section, since it is the critical method for the construction of Ontology and Semantic Network.

The aim of information extraction is to automatically extract structured information from unstructured texts, which has very broad applications in diverse domains including biomedicine, business (Aggarwal and Zhai, 2012) as well as engineering and design. Usually, information extraction refers to two fundamental tasks, namely, Named Entity Recognition (NER) and Relation Extraction (RE).

Named Entity Recognition refers to the identification of real-world entities in free-flow text. The entity could be as specific as the name of organization, person, location and gene (e.g. Facebook, Beijing, and Emily). They can also be as broad as noun phrases or any types of knowledge concepts. The task is to locate and identify the word sequence of the entity and classify it into predefined categories. The naïve methods for named entity recognition include matching the string against dictionary (for the name of person, location, etc.), and defining the regular expression patterns on part-of-speech (POS) tags (for noun phrase chunking) (Bird et al., 2009). A more advanced technique is to construct a simple supervised machine learning classifier to sequentially label each word through BIO tagging where the input features of the classifier can involve the current and contextual words as well as the POS tags of the current and contextual words. The state-of-art algorithms for named entity recognition include the probabilistic approaches such as hidden Markov models (Bikel et al., 1997) and conditional random fields (Settles, 2004), as well as the more

recent deep learning models, especially the recurrent neural networks (Miwa and Bansal, 2016).

Relation Extraction (RE) is a task carried out upon the named entity recognition. The RE task is to identify whether there exists relation between any two recognised entities and what relation type it is if given the predefined types of relations. For example, in the sentence "A **shaft** is used for **energy transmission**", we can identify that there is functional relation type between entities *Shaft* and *energy transmission*. One simple solution for this task is to use predefined linguistic patterns to match with string between two entities (Bird et al., 2009). The other solution is usually to consider it as classification problems: given predefined types of relation, how to categorize the relation between two entities into these predefined categories, therefore in which case, probabilistic methods and deep learning methods can be applied accordingly (Chan and Roth, 2010, Zheng et al., 2016).

The major use of information extraction for design information retrieval is to construct the design ontology and semantic network by identifying the design knowledge concepts and extracting the relations from the text. Design ontology and semantic networks are efficient ways to conceptualise and represent the design information and knowledge in structural graph forms (Rezgui et al., 2011). We will discuss this in detail in the next section.

2.2 Automatic Ontology Construction

The need for ontology in design information retrieval

In conclusion, many text mining techniques discussed above have been used in design and engineering domains for information retrieval purposes. However, most of these methods such as document retrieval, text classification, and document clustering are processing information at document or fragment level (Liang et al.,

2012). For example, in document retrieval, design documents are firstly indexed by using TF-IDF through vector space model, and then design queries based on keyword searches are used to match and retrieve relevant documents or textual fragments. In classification and clustering, each individual document is structurally represented by feature vectors, in which way, they can be either grouped into different clusters/communities of documents by unsupervised learning method (e.g., K-means, hierarchical models) based on the closeness between documents quantified through the metrics such as cosine similarity between vectors (Salton and Buckley, 1988), or categorized into different document classes by supervised learning methods (support vector machine (SVM), decision trees) (Murphy et al., 2014).

Therefore, all these techniques consider each document as a unit, and therefore can only process the information no more specific than document level or fragment level and retrieve/return a list of documents as the results for the designer's query.

In the age of exponentially increasing amounts of electronic documents and digitization of data, these traditional methods of document retrieval purely based on indexing and keyword matching becomes incapable of properly handling the highly contextual design knowledge in such a large data scale, and often retrieve "irrelevant" information with limited support (Iyer et al., 2005). Within a large-scale document set, a single keyword query will probably yield thousands of documents in the search results as long as any documents contain the keywords of the query, which will certainly lead to a lot of noise and irrelevant documents in the search results that are still difficult for the user to find the most useful information. Google has achieved a huge success on the webpage retrieval by developing the famous PageRank algorithm (Brin and Page, 2012). However, for a huge set of design documents without interlink or hyperlink between individual documents, PageRank algorithm is no longer applicable in this situation.

Thus, in the recent big data era, there have been increasing demands to process and extract information at more fine-grained semantic level, and this is where ontology techniques and information extraction (named entity recognition (NER), and relation extraction (RE)) come into play.

Unlike traditional document retrieval where processing information occurs at document or fragment level, the emergence of ontology-based techniques provides opportunities and effective mechanisms to process information at semantic level and extract and refine the relations between individual concepts from massive unstructured documents (Glier et al., 2014, Lim and Tucker, 2016, Lan et al., 2016), which can be subsequently structured and stored into design ontology-based systems (Rezgui et al., 2011, Chang et al., 2010, Liu et al., 2013) for knowledge representation and retrieval. This structured representation of design elements and associations can significantly help facilitate the information sharing among designers (Dong and Agogino, 1997). The extracted semantic relations can help designers obtain specific relevant knowledge concepts and capture a brief overview of the knowledge around their queries. More importantly, by incorporating the retrieved relevant concepts, designers can expand the original query to form a more informative query for further document retrieval, which will yield much more relevant document retrieval results (Dong and Agogino, 1997).

A design ontology system is actually a set of related fundamental design concepts with inherent associations such as design rules, constraints, contexts and rationale, which allows the designer to model one or more particular domains in terms of axiomatic definition or taxonomic structure (Mars, 1995). Due to its flexible and robust nature as well as formal protocol for representation (Gorti et al., 1998), the use of ontologies coupled with text mining enables designers to integrate and migrate valuable knowledge from originally unstructured-maintained documents into a richer structured conceptualization of the complex domain (Li et al., 2005). On the other hand, the development of text mining approaches prompts the rapid growth of

design ontology technology, and also enables ontology systems to be efficiently captured in the form of semantic network with vertices representing the individual concepts and objects, and edges representing the inherent relationships among concepts.

Current public ontology databases

There are several open-source public ontology databases available such as WordNet (Miller, 1995, Princeton, 2010), ConceptNet (Speer and Havasi, 2012, Luminoso, 2017), never end language learning (NeLL) (Carlson et al., 2010), and YAGO (Fabian et al., 2007). However, none of these establishes the semantic relations from a perspective related to design and engineering. WordNet is a hand-built large lexical ontology of English words, among which the semantic relations are mainly based on synonym, hypernym, and hyponym hierarchy. ConceptNet, NeLL, and YAGO all conduct unsupervised or semi-supervised learning approach to automatically extract entities and relations from resources on the internet (e.g., Wikipedia) to build huge semantic network. However, the extracted relation types in these systems are fixed and limited within a range of predefined rules such as *IsA*, *SubClassOf*, and *AtLocation*.

Design relations are highly diverse and contextual. For example, *fire* and *ceramic* have valuable association from the design engineering perspective because *ceramic* production may involve *firing* procedure and also the ceramic material is fireproof. However, this kind of useful relations cannot be captured by the limited predefined categories in above ontology systems. Therefore, in this big data era, there is an urgent need of applying the text mining approach to construct a design-oriented ontology database.

Construction of Ontology

Currently methods for ontology construction can be classified into three categories: hand-built methods, unsupervised methods, and supervised methods, as shown in Table 2-1.

Table 2-1 Previous works on information retrieval at document level and semantic level

Works	Document level	Semantic level			
		Hand-build	Unsupervised approaches		Supervised approaches
			Statistical	Linguistic	
(McMahon et al., 2004, Chen and Jan, 2000, Brin and Page, 2012)	√				
(Chang et al., 2010, Princeton, 2010, Ahmed et al., 2007)		√			
(Salton and Buckley, 1988, Liu et al., 2013, Ohsawa et al., 1998)			Co-occurrence frequency, clustering, vector space model		
(Li and Ramani, 2007, Speer and Havasi, 2012, Li et al., 2005)				POS tagger, phrase chunking, predefined-rule matching, Regular expression	
(GuoDong et al., 2005, Sun and Grishman, 2012, Socher et al., 2012)					SVM, maximum entropy, deep neural network

Hand-crafted approaches

Hand-built approach (e.g., WordNet) usually requires a large amount of human efforts and time, and therefore is often applied to construct domain-specific ontology representations (Chang et al., 2010, Princeton, 2010). Ahmed et al. (2007) described a six-stage methodology for developing ontologies for engineering design. They focus upon understanding a user's domain model through empirical research including interviews, literature reviews, and document analysis. The empirical research is integrated into the stages of their ontology development process, which involves identifying the root concepts of taxonomies and existing taxonomies, creating new taxonomies and testing for application. Chang et al. (2010) proposed a domain-specific process of ontology development method and implemented it in design for manufacturing. In their methodology, the structure, relations, and instances of the ontology are developed following a series of important phases including the concept categorization and class hierarchy development, slot categorization and development, identification and realization of relations among slots, and methods to support knowledge capture and reuse. Other representative methodologies for constructing ontology based on manual efforts include TOVE (Grüninger and Fox, 1995), ENTERPRISE (Uschold and King, 1995), and METHONTOLOGY (Fernández-López et al., 1997).

In practice, most of these hand-built ontology construction methods are often time consuming and human intensive. The domain-specific concepts and relations need to be manually input and edited by the experts in this field completely based on their own domain knowledge and expertise. Therefore, hand-crafted methods for ontology construction are only suitable for knowledge within specific domains and are not scalable to the massive data economy nowadays.

Unsupervised approaches

The above approaches are well developed for domain-specific ontology construction. However, in the big-data era with exponentially increasing amount of information, hand-edited/manual approaches require a huge amount of human efforts and time costs to develop large-scale ontology network within the context of massive unstructured data. The development of text mining and data analytic approaches provides opportunities for automatic ontology network construction. Unsupervised learning is currently a popular method to automatically extract semantic relations from the text. It commonly employs two kinds of approaches, namely, statistical approaches and linguistic approaches.

Statistical approaches

Statistical approaches are usually based on the descriptive statistics (e.g., occurrence frequency) and utilize various data analytic tools such as vector embedding and clustering to infer the relationship between two words or concepts (Salton and Buckley, 1988, Ohsawa et al., 1998, Munoz and Tucker, 2016, Bullinaria and Levy, 2007, Tous and Delgado, 2006, Juršič et al., 2012). Lim et al. (2010), Lim et al. (2011) and Liu et al. (2013) proposed an approach called document profile model in which single words or maximal frequent sequences are extracted as document profile elements, and average point wise mutual (avgPMI) is used to measure the strength of semantic association between terms and maximal frequent sequences. They perform facet modelling by clustering and aggregating similar entities to generate clusters of entity set. In their case, the labels assigned to the clusters are regarded as the concepts to represent the strongest semantic indicators of the cluster. Another kind of statistical method of building semantic relations proposed by Juršič et al. (2012) uses a vector model. Instead of representing a document by a feature vector of concepts with corresponding TF-IDFs, they inversely represent each concept by a vector of documents as: $\boldsymbol{w} = [f^{d_1}(w), f^{d_2}(w), \ldots, f^{d_n}(w)]$ corresponding to TF-IDFs of this word for n respective documents. Therefore, by embedding each concept in the semantic vector space, they are able to measure the strength of associations

among concepts via the cosine distance between vectors show in Equation (2-3), and conduct clustering algorithms to partition concepts into different levels of granularity as well.

These statistical approaches are not typical information extraction processes. Firstly, the extracted entities do not refer to the so called “named entity” which usually means the company name, person name, location name or even noun phrases, etc. Instead, every single word is regarded as an entity in this case. Secondly, the extracted relations do not have specific meaning. Rather than identifying the particular type of the relations from the predefined classes, the statistical approaches more focus on judging the existence and quantifying the relevancy of the relations purely based the statistics.

Linguistic approaches

Compared to statistical approaches, linguistic approaches (Speer and Havasi, 2012, Bateman et al., 2010, Marrero et al., 2013) utilizing a set of NLP tool kits are also widely used in unsupervised learning process to extract specific types of relations for semantic network construction. In-depth NLP for fully recognizing the syntactic structure and understanding the semantic meaning is still a nontrivial task. It becomes even more complicated when dealing with design ontologies due to the requirement of fulfilling both linguistic and domain-specific knowledge. To simplify the process, a shallow general NLP framework is often used for ontology construction (Li and Ramani, 2007, Li et al., 2005). In this shallow framework, domain-specific knowledge concepts and linguistic pattern rules should be first predefined. Then, analysis at syntactic level is conducted to extract the phrases by following the procedures including tokenizing, part of speech (POS) tagging, disambiguating, and phrase chunking. In the semantic recognition level, text string between the chunked phrases is matched with the predefined linguistic patterns to extract relations, where the matched patterns together with their two end noun phrases are structured into

many semantic triples, which are finally joined together as a semantic graph. A semantic triple is a kind of structure containing two end concepts and their relationship in between (Sintek and Decker, 2002). A dedicated web ontology language (OWL) (Bechhofer, 2009) as well as a popular ontology editor Protégé (Gennari et al., 2003) has been developed in the last decade to formally mark and organize semantic triples.

Overall, the linguistic approaches first parse out the syntactic structure, which is then evaluated to match with the predefined linguistic patterns to extract entities and relations. Even though the implementation of linguistic approaches explicitly upgraded the construction of ontology onto syntactic and semantic level, it is still a computationally intensive and expensive process (Collobert et al., 2011). Moreover, the main disadvantage of linguistic approaches is that it can only extract limited and predefined types of relations. From the perspective of creativity in design, an innovative event can often be observed to happen with implicit and 'surprising' links between design information and knowledge (Dorst and Cross, 2001). Thus, depending solely on predefined and 'default' linguistic patterns may possibly eliminate the innovative associations between knowledge concepts.

Linguistic approaches can be considered as the basic techniques for information retrieval. For named entity recognition, a dictionary of domain knowledge concepts is usually needed to recognise the entities from the text. Alternatively, noun phrase chunking is conducted to extract all the noun phrases as the entities. For relation extraction, each predefined type of relation corresponds with a set of linguistic patterns which are essentially regular expressions to match with the textual string between two entities. Although it uses unsupervised learning methods during text processing, a considerable amount of preliminary human work is required such as manually building the domain dictionary and compiling the linguistic patterns for each relation type. Also, the extracted relations can only be limited within the predefined types.

Supervised approaches

Recently, supervised learning methods have been applied for information extraction, specifically named entity recognition (NER) and relation extraction (RE) in machine learning communities, which is significant beneficial to ontology construction. Various techniques such as SVM, Naive Bayes, Maximum entropy, and conditional random fields (GuoDong et al., 2005, Sun and Grishman, 2012) are used to tackle the relation extraction task as classification problems. These methods usually require the practitioners to engineer handcrafted features or utilize features derived from existing ontologies (Rink et al., 2011). Some researchers have proposed end-to-end relation extraction without extra feature engineering by the use of deep neural network including recurrent neural network (Socher et al., 2012) and convolutional neural network.

However, there remain the same problems. The supervised models can only recognize limited types of relations, which have been predefined in the training set. This also means that, before automatically performing named entity recognition and RE, a large-scale text corpus with manually annotated relations of specific predefined type should be provided as the training set to train these supervised models. For engineering and design documents, it is a nontrivial task to create a manually annotated training set containing various types of relations from engineering and design perspectives regarding the functions, structures, mechanisms, methods, components, materials, and even power supplies.

2.3 The Use of Ontology in Engineering and Design

The above section has discussed various sophisticated ways developed for the construction of ontology and semantic network. Meanwhile, it's equally important to explore how to subsequently use the constructed ontology and semantic network in different domains. In the last decades, there is an increasing application of ontologies

across a broad variety of areas including medical, chemical, business, and engineering fields.

Medical field has widely applied the ontological platform for sharing and discovering the information about relationships between medical problems, treatments, symptoms, medicines, and tests extracted from the electronic medical records or documents (Rink et al., 2011). The ontology methodologies proposed in TOVE (Grüninger and Fox, 1995) and ENTERPRISE (Uschold and King, 1995) were originally used in enterprise modelling and project management process. The METHODOLOGY (Fernández-López et al., 1997) presented by Fernandez et al. for building ontology from scratch was applied in representing and storing the chemical specifications about chemical substances, elements, and attributes as well as the associations among them.

In the field of design engineering, Lim et al. (Lim et al., 2011, Lim et al., 2010) classified the ontological applications into three categories: (1) design information annotation, sharing and retrieval, (2) interoperability and interchange protocol between engineering systems and (3) product design configuration. For design information annotation, sharing and retrieval, they applied an automatically semantically annotated multifaceted ontology for product family modelling, and proposed a framework of information search and retrieval for product family design based on the multifaceted ontology. Following this, Liu et al. (2013) proposed a weighted multifaceted ranking schema to optimize the selection of the components that are semantically associated with the complex and heterogeneous design queries for product family design. For the purpose of information retrieval, Li and Ramani (2007) developed an ontology-based information search and retrieval framework, which was shown to outperform the traditional keyword-based search techniques.

Ontology also acts as a knowledge mapping structure between different engineering systems or criteria in order to facilitate and clarify the information exchange. Chang

et al. (2010) applied ontology in design for manufacturing strategies to facilitate the decision-making and optimization process in manufacturing with taking into consideration the complex interaction of both technical and economical criteria. Lin et al. (2004) also created a manufacturing system ontology (MSE) which is used as a mediator to coordinate collaboration and support the semantic interoperability across extend project teams. Another example is that ontology functions as an information exchange protocol for the interaction and integration among engineering knowledge where the ontology becomes the intermediate agent between systems. For instance, the use cases can be the information interchange between computer aided design application and computer aided process planning (Dartigues et al., 2007) as well as the machine element and engineering part libraries (Cho et al., 2006).

For a product design configuration, designers use ontology to represent the relationships between either physical design artefacts such as product components, models and specification among product families, or conceptual design rules such as design contexts, attributes and constraints (Felfernig et al., 2003, McGuinness and Wright, 1998, Soininen et al., 1998). Yang et al. (2009) presented an approach to represent product configuration knowledge through the use of semantic web technology (e.g. OWL, SWRL), and developed a product configuration engine based on the represented ontological knowledge, which can be the key factor for massive customization production. Nanda et al. (2006) proposed the Product Family Ontology Development Methodology (PFODM) where Formal Concept Analysis (FCA) is firstly used to identify the associations between design artefacts based on their properties and then OWL is implemented to refine and represent a product family ontology. This ontology-based methodology can effectively help provide a systematic and consistent design process for continuous development on a series of product family.

Therefore, based on the three categories of usage discussed above, we can see the essential feature of ontology is knowledge capture, sharing and reuse. Witherell et al. (2007) illustrated the potential value of ontology in representing application-specific knowledge while facilitating both the sharing and exchanging of this knowledge in engineering design. By using knowledge management (KM), ontological structures can also be integrated into design process management. One example is the H and J ontology (KM ontology), which was developed by Holsapple and Joshi (2004) to address the lack of a well-integrated framework that unify KM, and can be reused and further developed by KM practitioners.

In addition to the knowledge capture and reuse, some other works (O'Connor and Das, 2009, Jean et al., 2006, Mena et al., 2000) developed ontology-based query reasoning and processing for knowledge retrieval in semantic network. However, these query languages or methods were often developed case by case and highly depend on their respective systems of the established ontologies, which may be limited in terms of compatibility and universality.

2.4 Discussion and Motivation

This chapter firstly reviewed the state-of-art of text mining technique and its usage in design engineering. It has been shown that the traditional design information retrieval is largely based on the document retrieval, document classification and clustering which process the information at document or fragment level.

In the big data era, these traditional retrieval methods processing information at document level become incapable of handling such a large-scale of highly heterogeneous textual data. There is an increasing demand to process the information at semantic level where ontology comes into play which is the most suitable technology to process information at semantic level by extracting and representing the individual concepts and their inherent relations. However, manually

building the ontology or semantic network from massive text resources is impossible, and it is necessary to automate the construction process of ontology by using the information extraction technique in text mining. Therefore, the integration of text mining with ontological technology becomes the key solution for modern data-driven design information retrieval at fine-grained semantic level.

Although the review shows that many works has been conducted in the automatic construction of ontology and semantic network and their subsequent applications, there are still some issues needed to be tackled especially for the design information retrieval:

1) Although some large scale open-source public ontology databases are available nowadays, most of them are created for either common-sense or very specific domains (e.g. biomedical) and particular engineering projects. Still, there is no existing general ontology database for design and engineering knowledge. Therefore, with the massive available electronic technical documents, we are very keen to automatically construct a "WordNet" for the design and engineering knowledge.
2) In previous works, the entities are recognised by using either simple words (in statistical approach), hand-edited domain database (in linguistic approach), or annotated training set (in supervised approach). The relations are extracted mainly based on synonym, hypernym and hyponym, co-occurrence frequency, predefined linguistic patterns or training set with limited types of annotated relations. However, in our work, we aim to include much more broad knowledge concepts as entities and stand on a design and engineering perspective to establish more flexible and diverse relations between knowledge concepts.
3) Compared to the well-developed methods for constructing the ontology and semantic network, how to subsequently fully utilize and analyse the constructed semantic network is still an open question. The traditional applications of ontology mainly follow the naïve retrieval and reasoning methods based on the

individual entities and relations where a higher level of viewing it as a network structure is not achieved. It will be interesting for us to apply sophisticated network analysis techniques on the ontology in order to discover novel method for design information retrieval as well as creative idea generation.

4) Finally and most importantly, data visualisation can provide intuitive insight and knowledge. Human can leverage their visual system to intuitively discover patterns and identify new points. Therefore, graphic data visualisation can significant assist human with the knowledge discovery in an intuitive way. The data visualisation and user interaction with the ontology are rarely mentioned in previous work, except a limited number of old-fashioned software like Protégé. It would be good for us to use the state-of-art data-driven techniques to help user visualise the constructed ontology network and interact with it to retrieve information through a variety of network analysis techniques.

Keeping these four research gaps and motivations in mind, we will tackle each of these issues respectively throughout the following chapters from Chapter 3 to Chapter 6.

Chapter 3 Data acquisition: Web Crawler

Within the big-data economy, exponentially growing amounts of data are generated every second around the world, and becoming available in electronic and digital forms. The contents, types, forms and ways of dissemination of these data can be remarkably diverse. User and system generated data are being gathered through Cyber-Physical Systems or Internet of Things (IoT). Personal data and communications are exploding on the social media such as twitter, Facebook though bounded by privacy protocols. Massive news, blogs and publications are being posted and stored on the Internet. Such large-scale human and machine generated data, serving as valuable resources, create unprecedented challenges and at the same time unmatched opportunities for information retrieval and knowledge discovery (Kim et al., 2016).

However, these data are versatile and highly unstructured and contextualized. Understanding how to selectively and efficiently capture the relevant data is crucial for the following design information retrieval and knowledge discovery process. The aim of this chapter is to explore various types of *data forms*, *data resources* and *data acquisition techniques*, and identify suitable ones to build a raw database foundation for our data-driven design approach. Finally, we choose textual data as the main data form of our raw database, and apply the web crawling technique for efficient data acquisition. Two studies are conducted by using design posts (Study 1) and engineering and design publications (Study 2) respectively as the target data resource to build our raw database.

Some of the work described in this chapter has been previously published in (Shi et al., 2016):

1. Shi, F., Han, J. & Childs, P. 2016. A Data Mining Approach to assist design knowledge retrieval based on keyword associations. DS 84: Proceedings of the DESIGN 2016 14th International Design Conference. Copyright © Design Society 2016. Reprinted by permission of Design Society.

Some of the work described in this chapter has also contributed to the work published in (Han et al., 2018a, Han et al., 2016):

1. Han, J., Shi, F., Chen, L. and Childs, P.R., 2018. The Combinator–a computer-based tool for creative idea generation based on a simulation approach. Design Science. 4, p. e11.
2. Han, J., Shi, F. & Childs, P. 2016. The Combinator: A computer-based tool for idea generation. DS 84: Proceedings of the DESIGN 2016 14th International Design Conference.

3.1 Data Forms and Data Resources for design knowledge

Data, Information and Knowledge

Before we explore the very specific data forms and resources, a clear distinction should be made between the three basic terms: **Data**, **Information**, and **Knowledge**. Data is just the description of raw facts, and most of these human-and-machine-generated data are often highly unstructured and heterogeneous. It is necessary to process the data into information which is the structured representation of data within specific context or usable format. Knowledge can be then obtained through the understanding and learning of the information and their relationships which enables people to make action/decision (Liew, 2007, Bellinger et al., 2004). Thus, data and information are the sources for human to build and learn knowledge. It is true that one can be provided with an abundance of data but that does not automatically mean that one has acquired knowledge. Knowledge can only be

accumulated through human learning or experience by understanding and interacting with the data and information (Zins, 2007). Therefore, how to transform the large-scale unstructured data into structured information and then consequently discover meaningful knowledge is the main challenge for modern Data-Driven design information retrieval and knowledge discovery (Li and Ramani, 2007).

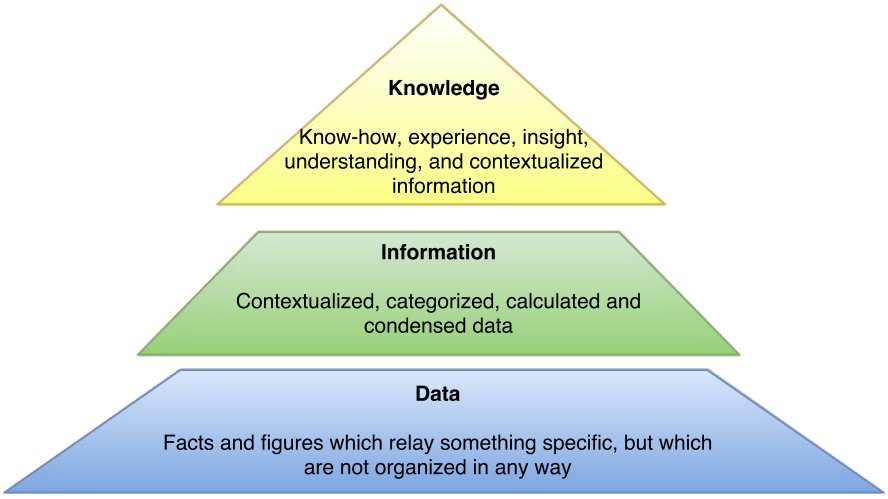


Figure 3-1 Hierarchical structure of data, information and knowledge, adapted from Bellinger et al. (2004)

In this chapter, by data acquisition we mean the capture of raw unstructured and heterogeneous data which are not organized in any way without going beyond to an information or knowledge level. The data obtained in this chapter will be used as our raw data which serves as the foundation and starting point for the subsequent information retrieval and knowledge discovery in following chapters.

Data Forms

Data can exist in many formats in the real world such as **numerical**, **graphic** and **textual** formats. For example, the temperature of a turbine blade, the price of an iPhone, the velocity of a spinning tube, the click volume of a video on YouTube, and the population growth in a developing country, are all **numerical** data, which are used to describe the degree, level and extent so that we can make comparison and distinguish between *fast* and *slow*, *expensive* and *cheap*, *big* and *small*. In design

and engineering fields, a large amount of information is available in the form of numerical data such as product dimensions, experiment parameters, user ratings. Design knowledge can be acquired by analysing these numerical data to gain insights, make predictions and improvement. Lim and Tucker (2017) offered a new method by analysing the histories and tendencies of the user rating scores to reduce the biases of reviewers' opinion. This numerical data analysis significantly helps the identification of true product quality in order to provide customers with more object and accurate feedback. The meter readings on energy usage, also characterized as numerical data, are collected through wireless transmission and network infrastructure across the country into a data centre to support the big-data driven smart energy management for the purpose of renewable and sustainable energy recycling (Zhou et al., 2016).

Graphic data such as images, CAD files, figures, videos, pictures, sketches, and drawings are also widely used in the fields of science, engineering, medical, art, architecture, design and also entertainment industry. Material scientists use scanning/transmission electron microscope to capture the micro images of atomic arrangement of carbon nanotubes. Aeronautical engineers establish the geometric and structural model of aircraft by elaborating on Computer-Aided Design (CAD) files. Medical doctors diagnose diseases based on the X-ray and CT photography. Artists paint pictures to express their emotions, feelings and thoughts, architects are making sketches to design buildings, and directors are producing videos to restore the history.

In industrial design and engineering design, graphs and diagrams are the key data format in product specification to show the product shape, illustrate the product function and behaviour, and instruct the assembly procedure, as well as represent the design idea. Therefore, many research works have been conducted by utilizing graphic data to assist design process or support design ideation and innovation. By fetching the online free image data, Han et al. (2016) developed a computer-based

tool, *Combinator*, to produce combined images of two or more design concepts to help designers generate creative ideas. Rapid advancement of sensor technologies enables the collection of massive population-based graphic shape data. Mining and analysis on these shape data help to obtain design knowledge of shape variability of the population and construct faithful 3D shape design models, which creates potential opportunities for mass customization, part-specific failure predication and just-in-time part maintenance (Wang and Qian, 2017). Instead of the population-based shape data, Dering and Tucker (2017) utilised product-based shape data paired with the product function to train a convolutional neural network, which is a able to recognise a product's function given its shape and therefore in turn create novel product's shape given a defined function.

Beyond numerical and graphic data, the ultimate format is **textual** data, since even information and knowledge acquired from numerical and graphic data can usually be transformed and narrated in textual data format. Thus, most information and knowledge are fundamentally grounded in texts, and the textual data format are useful for information and knowledge storage. In the current big-data economy, a large amounts of information are available in electronic and digital forms, approximately 80% of which are texts in nature (Yu et al., 2005). These textual sources are not only in the form of descriptive data formats such as technical papers, progress service reports about repair information, manufacturing quality documentation and customer help desk notes (Kornfein and Goldfarb, 2007), but also in the form of concise text formats including many industry specific terms and abbreviations (Ur-Rahman and Harding, 2012).

Recently in the emerging field of data-driven design, many research works rely on textual data as the main data sources for various design purposes. A large number of consumer text reviews are analysed by text mining to suggest human variability information that is essential for interaction, and weighted phrase rating on the text reviews can also help quickly and economically provide information useful to the

establishment of design specifications without any human intervention (Chaklader and Parkinson, 2017). Also based on the online customer reviews, Jiang et al. (2017) applied opinion mining and sentiment analysis on the review texts to extract information about customer preferences and expectation, which is then used to determine current importance and predict the future importance of different product features. In addition to user reviews, natural language text from Wikipedia pages has also been used in research work by Cheong et al. (2017) to automatically extract design function knowledge in the form of subject-verb-object triples through syntactic analysis. Song and Luo (2017) used textual data from patents, and applied patent mining to search for the precedents of the design of a specific technology or product for next generation design, which can be utilized by designers to explore of associations between retrieved patent data for new design opportunities.

According to the discussion above, textual data is currently the main data format, taking up 80% of the total available large-scale electronic and digital data. A large part of design information and knowledge is available in texts which become the ideal data format for design information retrieval and knowledge discovery, and are also widely used as the main data sources for data-driven design purposes. Therefore, the work in this thesis will also rely on **textual data**. In this chapter, we focus on the textual data acquisition and the obtained texts will then be used in the following elements of our data-driven pipeline.

Data Resources

The Internet has become one of the largest data resources in the world. In Nov 2016, Google has reported that their search index contained more than 130 trillion webpages and was well over 100,000,000 gigabytes in size (Google, 2017), and this does not include the “hidden” pages which was not indexed. The figure is certainly much bigger by now. Besides the vast amounts of data, the Internet also has several other beneficial properties for serving as an effective data resource for information

retrieval and knowledge discovery. Firstly, the Internet is updated in real time. Journalists are constantly updating news through online media, people are chatting through web apps, and website maintainers are keeping the page contents up to date. Secondly, Internet data are linked, structured and categorised (Pastor-Satorras et al., 2001). By conforming to the generic basic syntax of Uniform Resource Locator (URL) as shown in Eq. (3-1):

$$\textbf{schema}:\textbf{host}[/\textbf{path}][?\,\textbf{query}][\#\textbf{fragment}] \tag{3-1}$$

Every webpage can be linked through hyperlinks to other relevant pages, and groups of webpages in a website can be categorised under a common host address in a hierarchical structure through configuring the levels of path addresses. Such kinds of highly related and structured data could provide great efficiency for data acquisition process. Finally, the types of Internet data have great diversity including user and human generated data as well as machine and system generated data. Customer and user reviews are available on the websites of large electronic commercial companies such as Amazon, eBay, Alibaba. Social and personal data are exploding on social network media such as Facebook, twitter, LinkedIn. Various knowledge is edited on huge knowledge repositories (e.g. Wikipedia) and shared though online platform (e.g. Quora). Millions of literature resources and publications are stored in publishers' online databases such as those produced by Elsevier and Springer. Based on the above discussion, Internet data can be concluded to be vast in size, updated in time, related in structure, and diverse in type. Therefore, in our data-drive approach, the **Internet** is used as the abundant data resource for **textual** data acquisition.

3.2 Data acquisition by Web Crawler

To systematically and automatically extract the data from the Internet, a sophisticated robot often referred to as a **Web crawler** was developed in the late 1990s (Kausar et al., 2013), which has become the core technology of all the big

search engine companies in the world such as Google, Bing, and Baidu. By recursively following the hyperlinks contained in web pages, a Web crawler can start with only a few seed URLs and visit all the pages linked to the starting seeds, and then identify hyperlinks in these new pages to visit, and so forth. It behaves just like a spider crawling on the Internet based on the massive links between all the web pages. Therefore, instead of having all the target URLs in hand before conducting data acquisition, a Web crawler can be initialised with only a few URLs, and start to conduct data scraping process along with the collection of new target URLs simultaneously.

However, most of the time, a standard crawler which crawls through all the pages in an exhaustive search strategy is not necessary, since we only need specific data content for different purposes rather than all the data we have come across. This can be achieved by designing a focused crawler with specific policies or rules to configure its behaviour to fetch specific pages and extract relevant data for our interest only. The simplest policy or rule to guide a crawler is to make it only follow the URL addresses that satisfy a predefined regular expression[1]. Wang et al. (2013) developed their web crawler to only visit the news release websites of five well-known universities in UK and only parse specific sections to collect design-related information, which were then used to update their proposed *Effect Database* for design knowledge accumulation. In addition to regular expression, probabilistic and semantic models have been developed to evaluate the relevance of page contents for the crawler to decide whether or not to download a page or hit a link. Mukhopadhyay et al. (2007) designed a web crawler to crawl only through domain-specific webpages. They used ontological knowledge to identify relevant web pages for specific domain by assigning score to each term in the page content based on the term's specificity level in the domain ontology. The keywords in the link and the

[1] A regular expression is a sequence of characters that define a pattern used for string searching and matching Thompson, K. 1968. Programming techniques: Regular expression search algorithm. *Communications of the ACM,* 11, 419-422..

surrounding text of the link can also be utilized to speculate the relevancy of the document, which is reckoned measuring the semantic similarity between the keywords in the link and the taxonomy hierarchy of the specific domain (Yuvarani and Kannan, 2006).

Since the domain and path part of an URL address can somehow represent the page topic and the website structure (Kausar et al., 2013), in our following Study 1 and Study 2, we apply the simple policy (rule) by defining a set of regular expressions to configure the crawler's behaviour on only following specific types of URLs for our research purpose. In this case, *Scrapy* (Scrapinghub, 2017), an open source, collaborative, fast, light-weighted, and powerful crawling framework based on Python[2], can be directly adopted in our studies, and its architecture is shown as Figure 3-2.

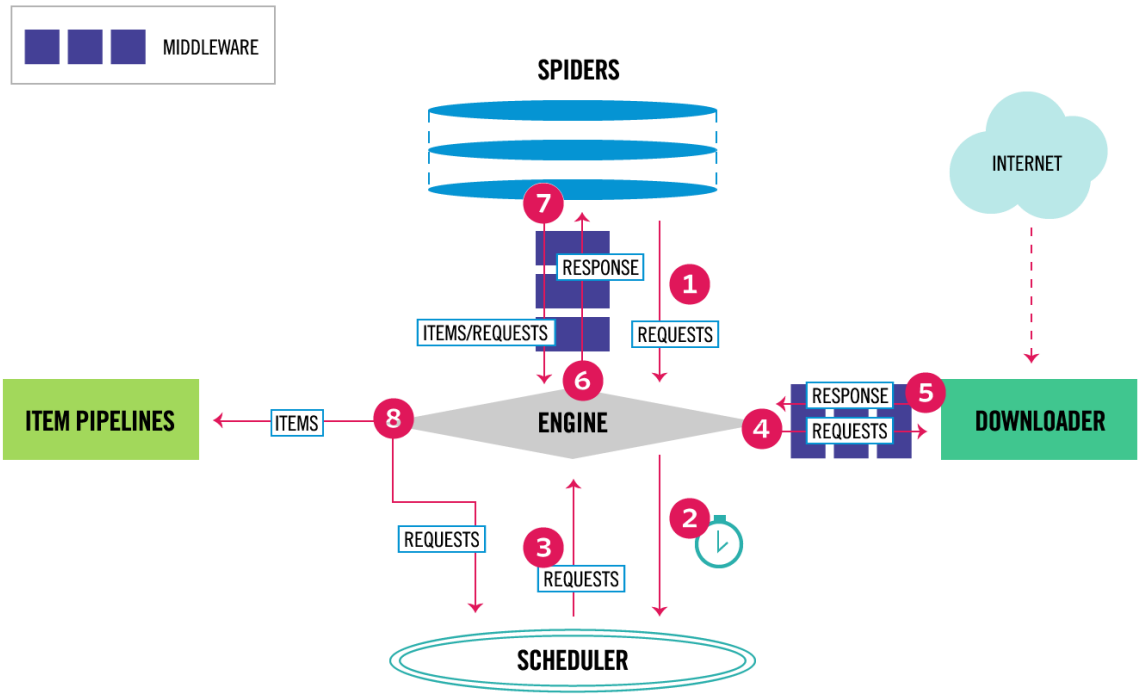


Figure 3-2 The common architecture of Web Crawler, adapted from (Aleksiūnas et al., 2017)

[2] A programming language just like C, Java, etc.

There are six main components in *Scrapy* (Aleksiūnas et al., 2017):

- The **Engine** controls the data flow between all components of the framework, and triggers events when certain actions occur.
- The **Downloader** directly accesses the web pages and transfers them to the engine which, in turn, feeds them to the spiders.
- **Spiders** are designed by users to parse the web pages and establish custom policies/rules to define what kinds of information to extract and what types of URL addresses to follow.
- The **Scheduler** gets input requests from the engine and arranges them for popping them later (also to the engine) when the engine requests them.
- The **Item Pipeline** is responsible for processing the items (information) once they have been extracted by the spiders. Typical tasks include validation, cleansing, and storage.
- **Middlewares** are specific user-definable hooks to process requests and responses passed between the Engine and the Spiders as well as the Downloader.

The data flow in Figure 3-2 can be described as (Aleksiūnas et al., 2017):

1) The Spider sends the initial Requests to the Engine.
2) The Engine puts the Requests in the Scheduler and asks for the next Requests to crawl.
3) The Scheduler sorts the Requests in a specifiable way, and sends back the next Requests to the Engine.
4) The Downloader gets the Requests from the Engine, which can be modified by the Middlewares.
5) The Downloader downloads the corresponding web page and passes it back to the Engine through the Middlewares.
6) The obtained Responses (web pages) are passed to the Spider for parsing.

7) The Spider parses the web pages to extract items (data) and generate additional Request, while the items are passed into the Item Pipeline, and the Requests are sent to the Engine and further to the Scheduler.
8) The extracted items are processed in the Item Pipeline.
9) The process repeats from step 2) and ends when there is no more Requests queued in the Scheduler.

The following studies illustrate our practical data acquisition process. We will conform to above **web crawling** architecture and data flow framework to extract **textual** data from **Internet** as the raw data foundation for our data-driven pipeline for design information retrieval.

3.3 Study 1: Capture Design Posts from *Yanko Design*

Background

Building on the above discussion, we will focus on capturing design-related information in the form of textual data from Internet resources. The goal of this study is to efficiently capture the raw texts from well-known design online magazines and design news websites as data sources such as YANKO DESIGN (Yamada, 2018), Red Dot Design Award (Reddot, 2018), and iF WORLD DESIGN GUIDE (iFDesign, 2018). These websites contain up-to-date design-related information including news about modern industrial design, technological innovation in product design, as well as the report on design trend of large design corporations. These text resources may potentially contain novel design concepts and relations between the design concepts, which could possibly be recognised and extracted by the following data text mining process. Table 3-1 shows several examples of design posts captured from YANKO websites involving various aspects including product design, technology, automotive and architecture. Each design post usually introduces one design topic, and provides several images and a paragraph of text description about the topic. We can see that

the text descriptions contain a wide range of rich design-related information including design concepts (e.g. biomimicry, three-wheeled design, double-rotor design), technical terms (e.g. 3D printing), mechanical components (e.g. saddle, pedal, handlebar), categorized products (e.g. bike, trike, scooter), various materials (thermoplastic resin, steel, timber) and shapes (curvy, pipe-esque, nested structure) as well as design functions and purposes (e.g. shielded from damage, hypnotic graphic, alert lighting, lock to prevent rolling, adjustable to grow with children, blends into serene setting).

Table 3-1 Examples of design posts from Yanko Design

Aspect & Name	Image	Textual description
Product design & Mist-ifying Shower head		Do you remember the old-timey Windows screensaver called **3D pipes**? Imagine turning that incredibly **hypnotic** graphic into an actual product. … Designed to look like nothing you've ever seen before, the shower head has a modern avatar and tends to leave one curiously observing and plotting the **flow of water** from its **curvy, pipe-esque** design. Nothing about the 3D Shower is traditional. Not even its manufacturing technique. Made using **3D printing**, the **shower head** uses **thermoplastic resin**, giving its quirky shape a different colour, material, and finish. …
Technology & Adventurer Drone		While a lot of **drones** follow the same four or six figure form, Alvix utilizes a specialized **double-rotor design** that makes it both incredibly agile and compact. … Its unique constructions consist of **four discs**: one for the **alert lighting mechanism**, two for the **rotors** and one for the **base**. When it's not being used, the rotors **swivel** inward where they're **shielded** from **damage**. … Like a flare that doesn't disappear, it uses **light** and **sound** to serve as a beacon for rescue personnel.

Automotive

&

Transformer of Trikes

Is it a **bike**? A **trike**? A **scooter**? The TF1 is actually all three! This transforming kid's trike/bike hybrid instantly transitions from a **three-wheeled** design that's perfect for training into a **two wheel bike**. Its **pedals** can also be removed so it becomes a **push scooter** little ones can ride. Not only does it adapt as children's skill evolve, it grows with them thanks to an **adjustable saddle and handlebars**. For added safety during training, its **rear wheels rotate** inward and outward to **lock** it in place and keep it from **rolling** uncontrollably.

Architecture

&

Bird watchers

You've probably heard of **biomimicry** as a **design trend** but have you ever heard of ***animal architecture mimicry***? … Glint is an architectural exploration inspired by the way **birds** build their own **nests** and other **structures**. It utilizes this unique form of **wildlife mimicry** to place **steel**, **timber** and **OSB** in a similar form as a **bird's nest**. The resulting organic shape blends in to the serene setting of the park. …

More importantly, there actually exist hundreds of this kind of design posts in these websites. This massive textual data, though highly unstructured and unorganized, can serve as a valuable raw textual source where useful design information can be extracted if proper data text mining approaches are applied.

Method

A focused web crawler is specially developed to crawl the Yanko design webpages. The structure of Yanko design website was investigated and can be illustrated as Figure 3-3.

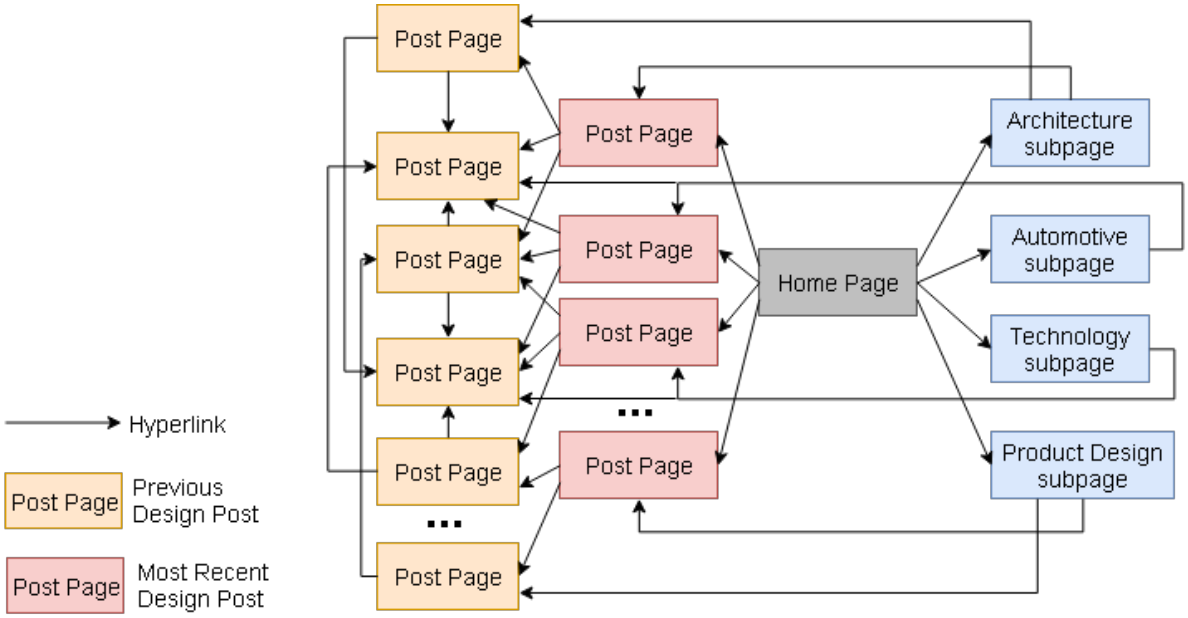


Figure 3-3 Structure of Yanko design website

The homepage divides the contents into four categories including product design, technology, automotive and architecture, each of which constitutes a subpage only containing hyperlinks to the most relevant design posts to its own topic. The homepage only shows links to several most recent posts. Each design post only contains one piece of design news, and is linked to its most related six design posts. The URL address of every design post is organized based on the chronological order, following the below syntax:

http://www.yankodesign.com/*year*/*month*/*date*/*topic*/

Therefore, the structure of Yanko design website is a huge unordered graph without providing a clear systematic way to crawl each piece of design post in sequence. In this case, our web crawler is programmed to utilise the hyperlinks between relevant design posts to gradually capture all the released design posts in Yanko Design website. The crawling framework is shown in Figure 3-4.

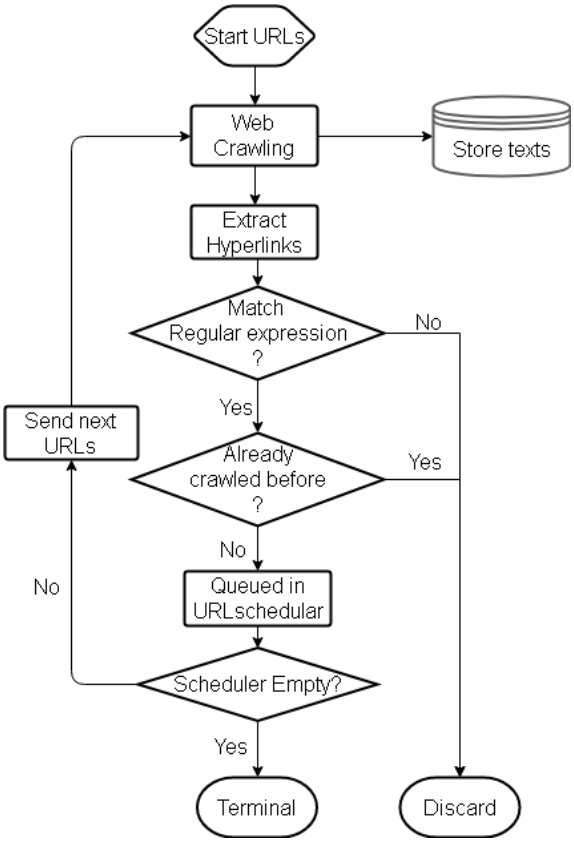


Figure 3-4 Web crawler framework for Yanko design

We begin by feeding the crawler with homepage address as the starting URL. The domain name "yankodesign.com" is used to restrict the crawler to only crawl within Yanko design web domain. Since all the URL addresses of design post pages conform to the above chronological syntax, a regular expression (regex) *r"/\d+/\d+/\d+/*"* is implemented to enable the crawler to recognise and only follow the hyperlinks to the design posts. Specific html elements of the post pages are selected and parsed to extract the textual description of each design post as shown in Table 3-1. A URL list is constructed to record all the URLs that have been encountered. All the extracted hyperlinks will be evaluated against the list and regex that only links which match with the regex and have not been processed before will be scheduled as new request URLs for the crawler, otherwise they will be discarded.

Results

Yanko design was firstly launched online in 2007, while this study was conducted in 2015. We initially aimed to fetch the textual description of all the design posts released during the period between 2007 and 2015 in one go, however, our crawler got blocked by the server after crawling 1000 posts. In order to avoid being blocked, our crawler should follow the robots.txt protocol of Yanko Design that it should wait 10~15 seconds after every crawl action. This will significantly slow down the crawling speed and is not considered in this study.

Three limitations of this study are worth noting. First, this study did not follow the robot.txt protocol and was blocked after crawling 1000 posts. Further study should follow the protocol to wait 15 s before the next crawling action, which however will result in a much slower crawling speed. Second, this study only uses the Yanko design website as the data source. Other websites related to design activities, news and design awards such as Red Dot Design and iF WORLD DESIGN GUIDE have not be explored in this study, although we found that there are a lot of information overlapping between Yanko Design and the other websites. Future studies will need to consider more web sources in order to increase the data variety. Third, the design news and posts mainly focus on the design-related information such as novel design concepts, design purposes and functions without containing enough engineering expertise and professional knowledge, which are also necessary and should be incorporated together in order to assist the general design information retrieval process. This will be addressed in Study 2 by focusing on textual resources where engineering concepts, technical terms and relations are available.

3.4 Study 2: Capture Engineering and Design Literatures from *Elsevier*

Background

The goal of this study is to crawl the resources about engineering expertise and knowledge which are complementary to industrial design posts and news captured in Study 1. Again, engineering design is a knowledge-intensive process where various areas of engineering knowledge and expertise are utilized to support the design process (Bertola and Teixeira, 2003). The design posts and news captured in Study 1 do provide raw texts to discover novel design concepts and relations, but it is still limited in amount and far from enough to satisfy the huge demands of design process on professional expertise and knowledge in engineering areas. In order to address this issue, this study use the academic papers in engineering and design fields as the data resource to further expand our raw database.

Academic literature such as research articles, review papers and technical briefs can provide abundant information in engineering and design fields. There are four advantages of using academic literatures as data resources. First, academic literature can cover a wide range of engineering and scientific domains, including design, mechanical, material, energy, civil, chemistry, physics, electronic, and computing. It is desirable since design activities need the support of a variety of knowledge in different fields (Pahl and Beitz, 2013). Second, the knowledge available in academic literature are detailed, specific and in-depth. Each research paper normally focuses on a specific topic in a particular engineering domain, and illustrates the domain knowledge in a professional way. Third, the engineering expertise provided in the academic literature is design-related, and can be interpreted from various design perspectives regarding the functions, structures, mechanisms, methods, components, materials, and even power supplies. Finally, the amount of academic literature is vast, and they are readily available online from large publishing companies such as Elsevier, Springer, Taylor & Francis, and SAGE Publications.

This study uses as the target resource literature from Elsevier, which is one of world's largest information and analytic company, and a major provider of scientific, technical, engineering, and medical publications. Elsevier is a global leader in

Science and Technical publishing. It currently publishes around 420,000 peer-reviewed research articles annually, and contains over 65 million document in total.

Method

To capture the metadata and texts of academic articles from Elsevier, data text mining API Keys (Elsevier, 2015) should be applied to obtain the permission. Elsevier assigned us a set of API keys and an agreement for non-commercial and research use only. The API Keys should be renewed periodically and always incorporated into header of each URL request, otherwise the request will be forbidden.

A web crawler coupled with the Elsevier API Key was programmed to fetch the metadata and extract the raw textual data of the academic papers in Elsevier corpus. A ScienceDirect sitemap was provided by Elsevier to structure the huge amount of documents in its corpus. The sitemap can be represented as a hierarchical tree as shown in Figure 3-5.

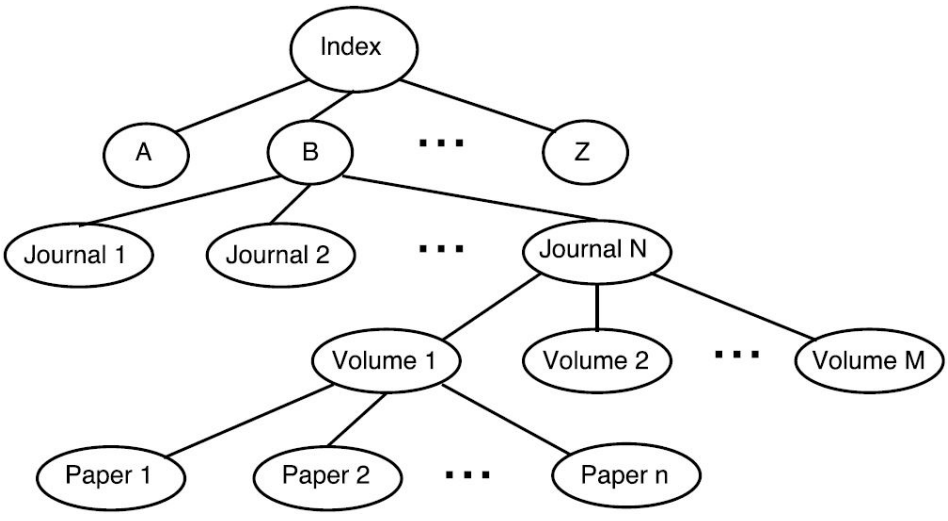


Figure 3-5 ScienceDirect Sitemap of Elsevier Corpus

In the first level, the sitemap indexes all the documents by the initial letter of the journal name. On the second level of the sitemap, each initial letter contains the

hyperlinks to a list of journals indexed by that Initial. Then, each journal page contains a set of volume URLs, and so forth. The volume page contains a set of paper URLs which eventually link to the metadata and raw text page of the corresponding paper. The structure of the sitemap can be utilised to facilitate our crawling process.

In this study, we only focus on the English papers published in recent decades, so the metadata of each article will be checked that journal name containing non-English word or papers published before 1995 will be discard. The metadata of each article, including authors, title, journal name, keywords, cover date, URL can be fetched by the crawler and saved into a local MySQL database. However, due to copyright issues and policy, each APIKey only has a limited quota (10,000) of papers to be parsed on full text in real-time, and the APIkey must be renewed when reaching its limited quota. Therefore in this study, in order to save the APIKey resource, we will fetch the metadata of every article the crawler comes across, but only parse full text in real time when the journal of this article is relevant to engineering, where the Elsevier's journal lists of Engineering, Energy, Chemical Engineering, Material Science and Computer Science are used (Elsevier, 2018).

Results

Due to the privacy policy, the Elsevier API key will expire after either reaching the limit quota or a specific period of time, and then the key should be renewed. This leads the whole crawling process to be extremely time-consuming because we should stop the crawler, renew the key and restart the crawler which make the crawling process discontinuous. This study was conducted between 2015 and 2016. The crawling process took up more than one year to fetch all the English journal papers available between 1995 and 2015 in the ScienceDirect corpus, indexed from initial letter A to Z by the journal name, as shown in Figure 3-5.

We have fetched 3,713,886 English papers in total published within the last 20 years. All the meta data of these articles, including URL addresses, titles, authors, journal names, publication dates and keywords, are stored in MySQL database. However, for the full content of the article, only 928,471 papers' full text are parsed due to the limit quota of the API.

3.5 Conclusion

In this chapter, differences between data, information and knowledge are defined and the chapter only focuses on the acquisition process of the raw data. Different data forms are investigated including numerical data, graphic data and textual data while a large part of information, approximately 80%, is available in textual data formats. Therefore, this research focuses on using textual data for our data-driven approach on design information retrieval. Internet is used as our data resource because of its several advantages that it's vast in size, updated in time, related in structure, and diverse in type. Finally, the web crawling technique is studied and applied in our two studies to extract the text data from Internet.

Study 1 extracts the data from YanKo design, a modern industrial design website. The design posts in this website provide up-to-date design-related information including news about modern industrial design, technological innovation in product design, as well as the report on design trend of large design corporations. Finally 1000 design posts are crawled within only 15 mins which is significantly faster than the manually browsing process. Study 2 focuses on academic literature, especially the papers in engineering and design domains. The hope is that the engineering research papers can strengthen the our raw data in terms of professional knowledge and expertise to support the knowledge demands in different stages of design activities, since engineering design is a knowledge-intensive process. The web crawling of Study 2 is a long process which took more than one year to finally fetch the metadata

of 3,713,886 papers, 928,471 of which are parsed along with the full texts in real-time.

The crawled design news, posts from design media websites, and metadata, full texts of academic papers described above will be used as the raw data for our following data analysis on design information retrieval.

Chapter 4 Mining and Mining: Network Construction

With the raw data captured in the previous chapter, the next step is to gain insights and information from these unstructured data. As shown in Figure 3-1, we need to move from the bottom unstructured raw data level to the upper structured information level. State-of-art data text mining techniques can be applied to transform the raw textual data into useful information which is the structured representation of data within specific contexts or usable formats, such as knowledge concepts, relations, semantic network and ontology.

This chapter aims to extract structured information from the raw textual data, in forms of knowledge concepts, relations, and semantic network to assist the design information retrieval. Data and text mining techniques including frequent itemset mining, association rule learning, and natural language processing are applied to recognise the knowledge concepts and evaluate the strength of the inherent relations, which are subsequently used to construct a huge semantic network. A disparity filter is implemented to filter the noise of the network to keep the relatively stronger knowledge relations. A study (Study 3) is conducted to evaluate the extracted knowledge concepts and relations by using a golden dataset, and results are compared with other benchmarking methods.

Some of the work described in this chapter has been previously published in (Shi et al., 2017a, Shi et al., 2016):

1. Shi, F., Chen, L., Han, J. & Childs, P. 2017. A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *Journal of Mechanical Design,* 139**,** 111402. Copyright © ASME 2017. Reprinted by permission of ASME.

2. Shi, F., Han, J. & Childs, P. 2016. A Data Mining Approach to assist design knowledge retrieval based on keyword associations. DS 84: Proceedings of the DESIGN 2016 14th International Design Conference. Copyright © Design Society 2016. Reprinted by permission of Design Society.

Some of the work described in this chapter has also contributed to the work published in (Han et al., 2017, Han et al., 2018b):

1. Han J., Shi F., Chen, L., Childs P. R. N., 2018. A computational tool for creative idea generation based on analogical reasoning and ontology. Artificial Intelligence for Engineering Design, Analysis and Manufacturing. doi:10.1017/ S0890060418000082.
2. Han, J., Shi, F., Chen, L. & Childs, P. 2017. The Analogy Retriever–an idea generation tool. DS 87-4 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 4: Design Methods and Tools, Vancouver, Canada, 21-25.08. 2017.

4.1 From Texts to Concepts, Relations, and Ontology

Although the captured raw texts can be directly used by traditional document retrieval approaches (McMahon et al., 2004) for engineering and design information retrieval, however, textual data are highly unstructured and the information can only be processed at fragment or document level. In traditional document retrieval, all the texts are indexed, and engineers depend mostly on keyword searches to retrieve the relevant textual fragments or documents based on the use of full text indexing and

matching (Salton and Harman, 2003, Salton, 1989). However, with increasing amounts of textual data, traditional document retrieval purely based on indexing and text-matching becomes incapable of properly handling the highly contextual design knowledge in such a large data scale, and often retrieves “irrelevant” texts with limited support (Iyer et al., 2005).

Unlike traditional document retrieval that processes information at document or fragment level, the emergence of ontology-based techniques provides opportunities and effective mechanisms to process unstructured textual data into structured information representation. Ontology-based technology process information at semantic level and extract and refine the relations between individual concepts from massive unstructured texts (Glier et al., 2014, Lim and Tucker, 2016, Lan et al., 2016), which can be subsequently structured and stored into design ontology-based systems (Rezgui et al., 2011, Chang et al., 2010, Liu et al., 2013) for knowledge representation and retrieval.

A design ontology system is actually a set of related fundamental design concepts with inherent associations such as design rules, constraints, contexts and rationale, which allows the designer to model one or more particular domains in terms of axiomatic definition or taxonomic structure (Mars, 1995). Due to its flexible and robust nature as well as formal protocol for representation (Gorti et al., 1998), the use of ontologies coupled with text mining enables designers to integrate and migrate valuable knowledge from originally unstructured-maintained documents into a richer structured conceptualization of the complex domain (Li et al., 2005). The development of text mining approaches has enabled rapid growth of design ontology technology, and also enables ontology systems to be efficiently captured in the form of semantic network with vertices representing the individual concepts and objects, and edges representing the inherent relationships among concepts.

Therefore, in this chapter, we firstly use data text mining techniques to extract the concepts and relations from our previously captured textual data. Then, an ontology network can be constructed based on the extracted concepts and relations. The raw data captured in Chapter 3 can be divided into two different data types for different data mining techniques. One data type is the metadata of academic papers. The other one is the full texts of design posts and academic papers. For the metadata, we directly recognise the keywords as the knowledge concepts and apply frequent itemset mining and association rule learning (Zaki et al., 2014) to evaluate the relations, while for the full text, we use natural language processing techniques to recognise the nouns or noun phrases as concepts, then syntactic analysis is implemented to parse the structure of sentences in order to assign the relations. A brief framework of the techniques used for the two different data types is shown in Figure 4-1. The specific details of this process are described in the following sections.

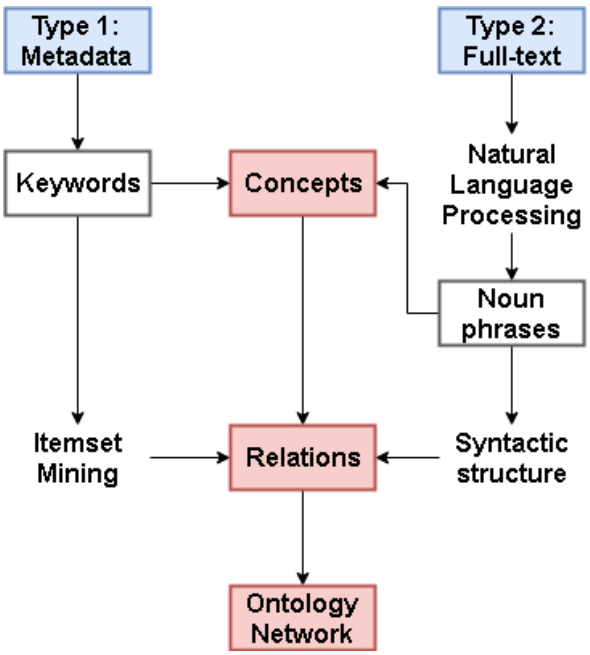


Figure 4-1 Brief summary of techniques used for the two different data types

4.2 Keyword Associations Mining

In this section, we use the structured keywords as one resource to construct our semantic network.

Advantages of Keywords

As discussed in Chapter 2, there exist significant limitations to apply current existing public ontology databases (e.g. WordNet, ConceptNet) for practical design information retrieval. It is due to two main reasons:

(1). Design relations are highly diverse and contextual. However, for WordNet, the relations are basically constructed on hyponym and hypernym. This cannot satisfy design information retrieval which involves high level association processes based on relations about methods, functions, materials, structures, etc. Thus, we need to explore the inherent associations between knowledge concepts from design and engineering perspectives.

(2). Design process needs professional knowledge and expertise. For ConceptNet, although it has well-established semantic relations between conceptual items, however, its knowledge domains are currently restricted in common sense, which are not suitable for responding to highly specialized engineering queries containing extensive expertise. Therefore, we need to focus on building associations between professional knowledge concepts instead of common sense.

The keywords of academic papers can overcome these limitations since they represent specialized knowledge concepts and have inherent high-level associations from design and engineering perspectives in concise format. Firstly, the keywords of academic journal papers in engineering fields concentrate on professional knowledge rather than common sense. They tend to be highly specialized and cover a broad range of concepts involving methods, principles, processes, materials, functions and

mechanisms in various engineering fields. Secondly, the associations between keywords within a single paper can be highly diverse and contextual. Keywords of one individual academic paper usually show the research objects, research methods and even principles and materials discussed in the paper. Thus, the inherent correlations between research objects, research methods and principles studied in a journal paper would result in the inherent associations between the keywords of this paper. Therefore, the keywords of a single academic paper are actually automatically related to each other within a high-level academic association scope.

Furthermore, keywords are concise. The form and source of fundamental knowledge concepts are important to build the professional knowledge network. Concise forms or patterns are effective for representing knowledge concepts in order to produce brief and clear knowledge network, which is easy and efficient for users to retrieve relevant information. Therefore, technical keywords are ideal to represent the knowledge concepts because of its simplicity and specificity.

In this section, we directly use the keywords of academic papers as the mining resource.

Linking Keywords

As discussed, the keywords within one single paper are often inherently related, which usually illustrate the topic, knowledge background, method, principle, problem or effect discussed within this paper. Since there exist correlations between the topic, knowledge background, method, principle and problem studied in this paper, the keywords representing these points are therefore also associated.

Here is an example of one academic paper (Coules et al., 2018) containing five sets of keywords: "Residual stress", "Fracture", "Neutron diffraction", "Digital image correlation" and "Finite element method". It can be seen that there exist high-level associations between these five sets of keywords from engineering perspective,

shown in Figure 4-2 as a relational network. Similarly, every paper has its own keywords network. The networks of different papers can be combined together by using their common keywords as the joints. Figure 4-3 shows the combined keywords network by linking Figure 4-2 with a second paper (Zhang et al., 2015). The keywords of the second paper are: "Laser shock wave", "Metal sheet", "Finite element method, "Deformation", "Residual stress", in which the common keywords "Finite element method" and "Residual stress" of both papers are used as the joints to link the two networks together. The keywords of the second paper extend the knowledge about the use and meaning of “finite element method” and “Residual stress” in metal sheet forming domain.

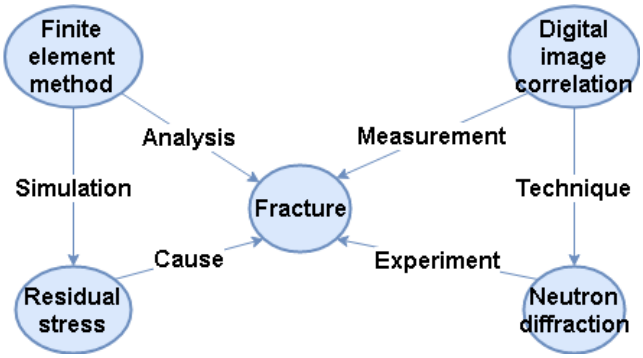


Figure 4-2 Keywords linking within a single paper

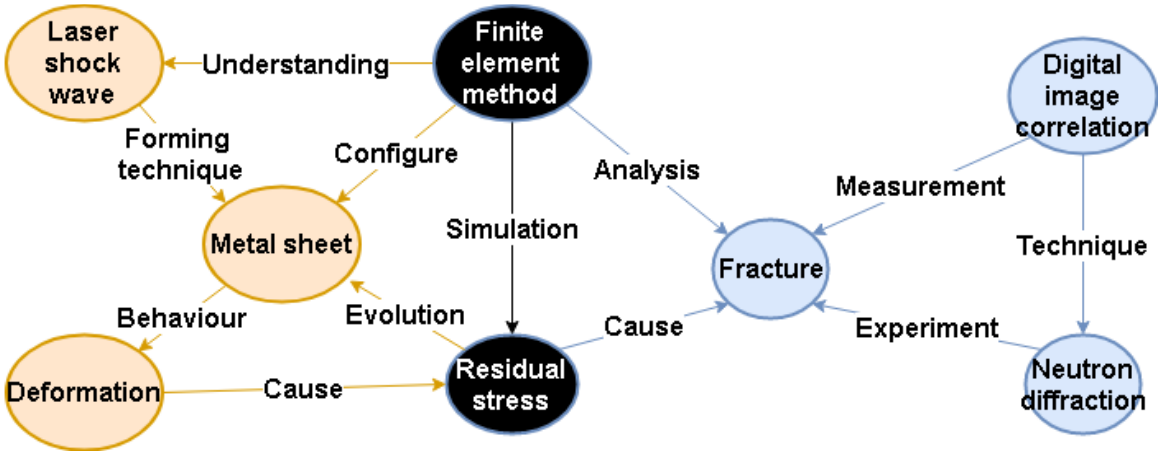


Figure 4-3 Combination of keywords networks of two papers

In this way, a knowledge network can be constructed by using the associations of keywords of massive academic papers. This huge network representing specialized

knowledge concepts with inherent high-level associations, is potentially more powerful than WordNet and ConceptNet for design information retrieval by overcoming the two limitations above.

Data Mining on Association Rules

However, for millions of papers in practice, it is impossible to manually construct the keywords network of each paper and combine millions of networks into a huge one by searching their common words as the join points, which would be extremely time-consuming. We can, however, use computational resources with data mining technology to automate this process. Data mining comprises powerful algorithms dedicated to massive data analysis, in which the itemset mining technique (Zaki and Hsiao, 2002, Pei et al., 2000) can help exploring the associations between keywords to construct the knowledge network.

As illustrated, keywords within one academic paper usually have high-level associations with each other, which is similar to the classic market basket analysis problem in itemset mining (Loraine Charlet and Kumar, 2012). In market basket analysis, items in one basket are considered to be associated between each other. Thus, we can view each paper as a supermarket basket and the keywords of the paper as the items in the basket. Furthermore, the strength of associations between keywords can be evaluated by the *support value* in itemset mining algorithm.

Let $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$ be a set of academic journal papers, where p_i denotes one single paper in the paper set $\mathcal{P}$. Let $\mathcal{K} = \{k_1, k_2, \dots, k_m\}$ be a total collection of the keywords of all papers in $\mathcal{P}$. A set $K \subseteq \mathcal{K}$ can be called a subset in $\mathcal{K}$. Thus, a basket transaction can be expressed as a tuple of the form $\langle p, K(p) \rangle$, where $p \in \mathcal{P}$ is an unique academic paper, and $K(p) \subseteq \mathcal{K}$ is the set of the keywords contained in p. A transaction database $\mathcal{D}$ can be built as a list of the basket transactions. Table 4-1

shows an example of basket transaction database $\mathcal{D}$. Here $\mathcal{P} = \{p_1, p_2, \dots, p_6\}$ is the set of papers and $\mathcal{K} = \{k_1, k_2, \dots, k_8\}$ is the set of all the keywords in $\mathcal{P}$.

Table 4-1 Basket transaction database

p	$K(p)$
p_1	$k_2\ k_4$
p_2	$k_1\ k_2\ k_5\ k_6$
p_3	$k_1\ k_5\ k_7$
p_4	$k_2\ k_8$
p_5	$k_1\ k_2\ k_5$
p_6	$k_3\ k_5\ k_7$

The transaction database $\mathcal{D}$ is actually a relation between the paper set $\mathcal{P}$ and keywords set $\mathcal{K}$. We say that paper $p \in \mathcal{P}$ contains keyword $k \in \mathcal{K}$ iff $(p, k) \in \mathcal{D}$ and $k \in K(p)$. A transaction database is essential in itemset mining as it contains all the fundamental information we need to evaluate the associations between keywords and construct the knowledge network.

An association rule is an expression $k_i \rightarrow k_j$, where k_i and k_j are two keywords that $k_i,\ k_j \in \mathcal{K}$. The *support value* of the rule in a transaction database $\mathcal{D}$ is the number of transactions in which both k_i and k_j co-occur:

$$sup(k_i \rightarrow k_j) = |\{p | \langle p, K(p)\rangle \in \mathcal{D} \text{ and } k_i, k_i \in K(p)\}| \quad (4\text{-}1)$$

For example, in Table 4-1, for association rule $k_1 \rightarrow k_5$, we found that papers p_2, p_3 and p_5 contain both k_1 and k_5. Thus, the *support value* of the association rule $sup\ (k_1 \rightarrow k_5) = 3$. In our case, association has no direction as we can find that $sup(k_5 \rightarrow k_1)$ is also 3. The matrix in Table 4-2 shows the *support values* of

associations between all keywords in transaction database $\mathcal{D}$. Each cell in the matrix represents the *support value* of the association rule between the keywords in row label and column label. This matrix is symmetric due to $sup(a \rightarrow b) = sup(b \rightarrow a)$. The diagonal line of the matrix, where row label and column label are the same keyword, illustrates the frequency of occurrence of this keyword in transaction database $\mathcal{D}$. Efficient algorithms have been developed to compute the *support values* of association rules in itemset mining such as Apriori Algorithm, Eclat Algorithm and FPGrowth Algorithm (Zaki et al., 2014), which are all applicable to existing databases.

Table 4-2 Support Values of all association rules in database $\mathcal{D}$

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
k_1	3	2	0	0	3	1	1	0
k_2	2	4	0	1	2	1	0	1
k_3	0	0	1	0	1	0	1	0
k_4	0	1	0	1	0	0	0	0
k_5	3	2	1	0	4	1	2	0
k_6	1	1	0	0	1	1	0	0
k_7	1	0	1	0	2	0	2	0
k_8	0	1	0	0	0	0	0	1

A keyword knowledge network could now be constructed by using the generated association rules with corresponding *support values*. In data mining analysis, the network can actually be represented as a graph data structure. A graph is a pair $G = (V, E)$, where V is a set of vertices, and $E \subseteq V \times V$ is a set of edges. In our case, keywords are the vertices in graph $V = \mathcal{K} = \{k_1, k_2, \dots, k_m\}$, and edges represent the association rules between keywords. An edge can be constructed between two vertices if the *support value* of association between this two keywords is nonzero. Thus an edge can be expressed as a tuple:

$$e_{ij} = \langle k_i, k_j, sup\ (k_i \rightarrow k_j) \rangle, \text{ subject to } sup\ (k_i \rightarrow k_j) > 0 \quad (4\text{-}2)$$

where $k_i,\ k_j$ are two vertices, and the nonzero *support value* $sup(k_i \rightarrow k_j)$ is the label of the edge. If there exists an edge between two vertices, we say the two vertices are adjacent. The edge set E can be formed as a list of such tuples $E = \{e_1, e_2, \dots, e_N\}$. Figure 4-4 shows the keywords network graph based on Table 4-2.

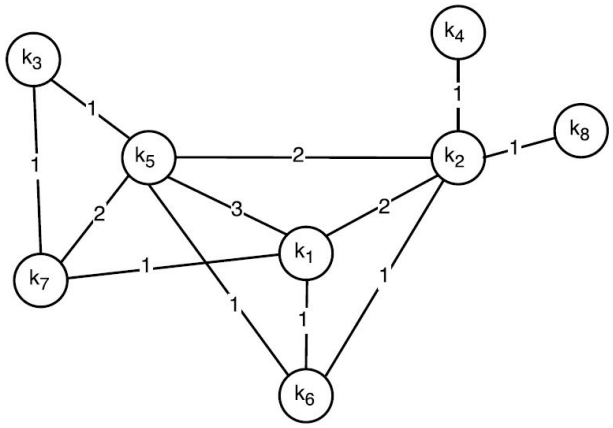


Figure 4-4 Graph representation of keywords network

Network Construction

Ontology-network construction was conducted by using the above itemset mining algorithm on the metadata of academic papers. The keywords and their associations are constructed and saved to a database which represented the keywords network. Since the network is a form of graph structure $G = (V, E)$, two relational tables are used to store the vertex set V and the edge set E respectively. The vertex table applies only one indexed column to store all the keywords $\mathcal{K}$, while the edge table contains three columns to respectively store two indexed vertices and one edge label (support value of the association between the two keywords) in the form of $\langle k_i | k_j | sup\ (k_i \rightarrow k_j) \rangle$.

The process of our algorithm to incorporate the keywords K of a single paper is described as follows: it firstly updates the vertex table through identifying and

adding keywords which are new to the network. Then, it iterates to construct associations between every two words in set K. If the association already exists in the edge table, its *support value* in the edge label will be increased by 1, otherwise, the association will be saved as a new edge in the network with initial *support value* to be 1.

4.3 Full Text Mining

In addition to the metadata of the academic articles, we also have the unstructured full text resources of design posts and academic papers where the full text descriptions or contents were parsed in real-time with crawling. In order to recognise knowledge concepts and relations from the full raw texts, text mining involving natural language processing techniques should be applied.

Statistical and Linguistic Approaches

As discussed in Chapter 2, there are currently two main unsupervised text mining methods for the automation process of concept recognition, relation extraction and ontology construction based on raw textual data, namely, statistical approach and linguistic approach. Statistical approach uses the descriptive statistics (mainly co-occurrence frequency or vector model) to infer the relationship between two words or concepts, while linguistic approach identifies the relationship by parsing the syntactic structure and semantic meaning of the sentences.

In statistical approaches, one common way is to utilize the co-occurrence frequency between words (Ohsawa et al., 1998, Munoz and Tucker, 2016, Bullinaria and Levy, 2007). Another method is the statistical vector space model (Tous and Delgado, 2006) based on the most well-known measurements term frequency-inverse document frequency (TF-IDF)(Salton and Harman, 2003):

$$\rho_i^x = \frac{f_i^x}{\sum_{k=1}^{n} f_k^x} \cdot \log \frac{N}{F_i} \tag{4-3}$$

Where ρ_i^x is the TF-IDF meaning the significance of term i in document x, f_k^x is the occurrence frequency of term k in document x, n is the total number of extracted terms contained in document x, N is the total number of documents in the particular repository to be processed, and F_i is the number of documents which contain term i. By setting a significance threshold, a set of significant terms/concepts beyond the threshold can be extracted from the document set. Each document can be represented as a vector of concepts with corresponding TF-IDFs. In this way, the document can be indexed by the concepts in the vector, and the similarity between documents can be compared through the closeness between vectors (Salton and Buckley, 1988). Alternatively, this model can also be transposed for ontology construction that each concept is inversely represented as a vector of documents with the same corresponding TF-IDFs (Juršič et al., 2012). Therefore, regarding concepts as the vertices of semantic network, the association between concepts can be evaluated by the similarity measure between the vectors of concepts. However, this statistical vector model only works well to build a semantic web within small and homogeneous collections under a common design topic (Brin and Page, 2012). Practically, a large number of design documents usually describe complex heterogeneous engineering expertise as well as various design processes and specifications.

To improve the scalability, NLP with the development of pattern recognition in machine learning has recently been implemented in constructing semantic web (Bateman et al., 2010). In-depth NLP for fully recognizing the syntactic structure and understanding the semantic meaning is still a non-trivial task. It becomes even more complicated when dealing with design ontologies due to the requirement of fulfilling both linguistic and domain-specific knowledge. To simplify the process, a shallow general NLP framework is often used for ontology construction (Li and Ramani,

2007, Li et al., 2005, Marrero et al., 2013). In this shallow framework, domain-specific knowledge concepts and pattern rules should be firstly predefined. Then analysis at syntactic level is conducted to extract the phrases by following the procedures including tokenizing, part of speech (POS) tagging, disambiguating and phrase chunking. In the semantic recognition level, relations are extracted by matching the text between two phrases with the predefined linguistic rules. The matched rules together with their two end phrases are structured into many semantic triples which are finally joined together as a semantic graph. A semantic triple is a kind of structure containing two end concepts and their relationship in between (Sintek and Decker, 2002).

Even though the implementation of NLP upgraded design information retrieval onto syntactic and semantic level, it is still a computationally intensive and expensive process (Collobert et al., 2011). From the perspective of creativity in design, an innovative event can often be observed to happen with implicit and 'surprising' links between design information and knowledge (Dorst and Cross, 2001). Thus, depending solely on predefined and 'default' linguistic patterns may possibly eliminate the innovative associations between knowledge concepts.

A combination in our full text mining approach

As discussed above, solely depending on either pure statistical or linguistic approach will have their own advantages and limitations:

(1). For statistical approaches, the advantage of co-occurrence frequency based methods is that it can recognise any kind of relations between any words and concepts. However, in the meantime, it will also lead to a considerable amount of noises in the extracted results. For example, the extracted concepts might be meaningless and not professional, and the relations can be "irrelevant". Vector models can build well-related semantic networks within small and homogeneous

collections under a common design topic, but have limitations in terms of data scale and knowledge heterogeneity when provided with a huge amount of texts which describe complex heterogeneous engineering expertise as well as various design processes and specifications.

(2). For linguistic approaches, the advantage is the data scalability and specificity that it can be applied to a large scale of textual documents to extract specific types of relations for semantic network construction. The linguistic analysis can first parse out the syntactic structure, which is then evaluated to match with the predefined linguistic patterns to extract entities and relations. However, this is also the limitation that specific types of linguistic patters should be predefined and therefore, it can only extract limited and predefined types of relations. Design and engineering knowledge are highly diverse and contextual containing various types of relations from different perspectives, which can hardly be satisfied by solely matching with the limited types of predefined linguistic patterns and rules.

Therefore, in order to capture various kinds of relations and also to adapt to large-scale data environment, we combine the advantages of both statistical and linguistic approaches in our full text mining process. Specifically, both statistical and linguistic techniques are integrated into a hybrid unsupervised learning algorithm by combining simplified NLP tools and frequent itemset and association rule mining, where the concepts are extracted at linguistic level by using syntactic analysis, while the strengths of the associations are evaluated at statistical level by using co-occurrence frequency. Hence, our full text mining process consists of two steps: simplified NLP extraction and itemset association mining.

Simplified NLP extraction

The procedures of our simplified NLP extraction are conducted as shown in Figure 4-5, by using spaCy[3] and the natural language toolkits[4] (NLTK) (Bird et al., 2009).

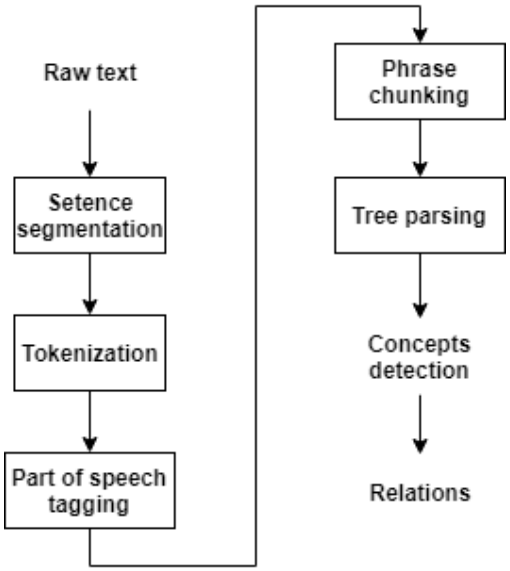


Figure 4-5 Procedure of simplified NLP extraction

Whenever a piece of text is fed into the unsupervised learning algorithm, the text is segmented into separate and complete sentences. Then, a word tokenizer and part-of-speech tagger are used to separate every single word and to tag the part of speech of each word respectively. The part-of-speech tag is one of the most useful pieces of information for phrase chunking. By defining chunk grammar based on the part-of-speech tags, we can conduct noun phrase chunking to search noun phrases within a sentence. Chunk grammar consists of rules indicating how a noun phrase can be formed by matching with the pattern of the part-of-speech tags of a sequence of words. The rules can simply be regular expression, for example <*DT*>*?*<*JJ*>***<*NN*>, which means that an noun phrase chunk can be formed when we finds a word sequence starting with an optional article, followed by any number of adjectives and finally followed by a noun (Bird et al., 2009).

[3] spaCy is an open source industrial-level natural language processing software. https://spacy.io/
[4] Another software initially developed for education purpose. https://www.nltk.org/

A constituency tree can be parsed to represent sentence structure through NLTK Stanford parser (Manning et al., 2014). Figure 4-6 shows an example of a sentence tree parsed at clause level and phrase level.

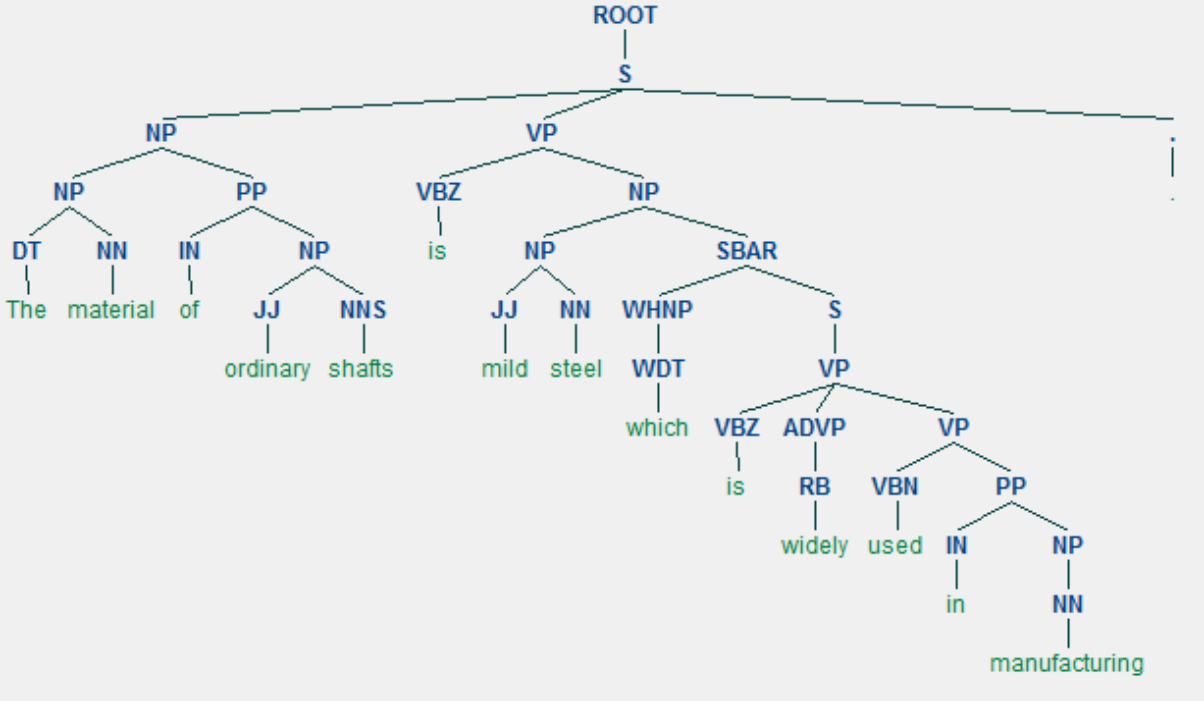


Figure 4-6 Tree representation of a sentence at phrase and clause level

Therefore, the **noun chunks** (noun phrases that do not contain any other noun phrases. E.g. *the material*, *ordinary shafts*, *mild steel*, *manufacturing*, in Figure 4-6) which we usually just simply refer as "noun phrases" in this thesis, can be easily extracted from both the subject and object parts at different clause levels for each sentence. In practice for efficiency, instead of firstly parsing out the constituency tree through NLTK Stanford parser and then searching the noun phrases which do not contain any other noun phrases programatically, we rather prefer to use spaCy to extract noun chunks straightforward. In this case, the constituency tree will be solely used to check if two noun chunks/phrases are within the same clause as required below.

Stop-word *(e.g. prepositions, conjunctions, pronouns, articles, and also word occurring very rare)* removal is used to refine the extracted noun phrases (chunks)

to become essential concepts in different forms such as *adj+noun, noun+noun, gerund, gerund phrase,* etc. Afterwards, these essential concepts are used as the nodes in our ontology network, and the associations are constructed to link the nodes of concepts which appear within the same sentence or clause. Synonyms and polysemy, while worth exploring, are not considered at the moment, since it is difficult to identify synonyms of professional and technical terms in such broad engineering fields, and recognizing polysemy is a not trivial NLP task, which is out of scope for our research.

Itemset Association Mining

In order to evaluate the strength of the constructed association, frequent itemset mining, which has been implemented in Section 4.2 for keyword association mining, is also similarly used in here. Instead of assuming items are related in the same supermarket basket or keywords are related to one another within the same academic article, here we assume that noun phrases of the subject and object parts within the same sentence are related. Based on another assumption that concepts in the same clause would be more relevant to each other than concepts in the different clauses of the same sentence, therefore when we get a sentence, we increase the weight value by 1 for the association between concepts in the same clause while increasing the weight by reduced value ($0 < \tau < 1$) between concepts in the same sentence but different clauses. In our case, we let τ be the median value 0.5.

The increase of the weight will happen in real time with the expansion of the network, as illustrated by an example in Figure 4-7, where *ST* means the sentence, *CL*, *SB*, and *OB* are the clause, subject, and object, respectively, of the sentence, and C_n is the concept (noun phrase) parsed out from *SB* or *OB*.

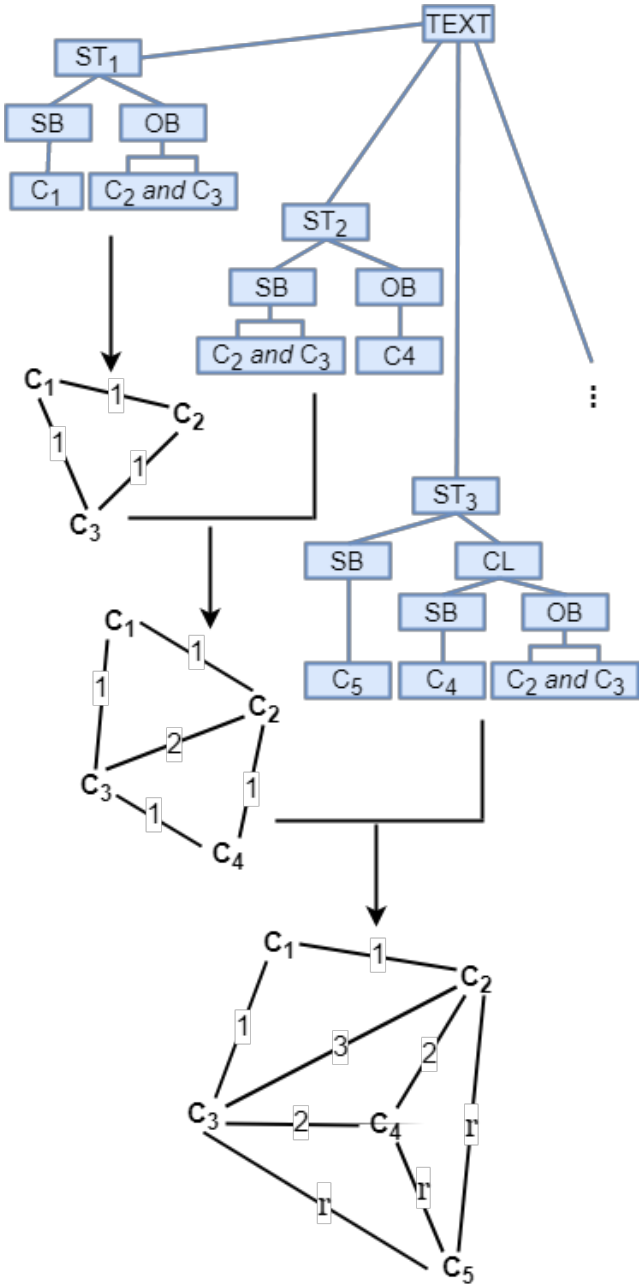


Figure 4-7 Constructing an ontology network and evaluating relation strength simultaneously

4.4 Unifying and Disparity Filtering

Unifying the two types of associations

So far, we have two types of associations, namely, the keyword associations built in Section 4.2, and the associations between noun phrases extracted from full text in

Section 4.3. The comparison of the two kinds of associations can be shown in Table 4-3. The first type of associations directly uses keywords as concepts and builds relations between any two keywords of the same paper, and the weight is the number of papers containing both keywords. The second type of associations directly uses noun phrases as the concepts, and builds relations between any two noun phrases within the same sentence, where the weight depends on the number of sentences containing both concepts.

Table 4-3 Comparison of the two types of associations

	Nodes/Concepts	Edges/Relations	Weights
First (Section 4.2)	Keywords	Built within the same paper	Number of papers containing both connected concepts of the relation
Second (Section 4.3)	Noun phrases	Built within the same sentence	Number of sentences containing both connected concepts of the relation

We can see that these two types of associations essentially follow the same idea and can be unified into one semantic network. The nodes that represent exactly the same knowledge concept can be directly merged together. The edges linking the same pair of concepts can be merged together, while the statistical weights of the merged edges should be summed up to be the final merged weight. In this way, a unified semantic network can be constructed.

Disparity Filtering

As the concepts and relations are extracted from the massive quantity of academic papers in the Elsevier site and design posts of various types of design topics, the quantity of the associations around a concept can be very high, and valuable knowledge associations may be concealed and mixed with other less meaningful noisy relations. In order to reduce the graph size and improve the quality of the

associations, a disparity filter (Serrano et al., 2009) is conducted to remove the noisy relationships with relatively lower weights around the concepts. The raw weights should be normalized by $\bar{w} = w/s$ to be used for disparity filter, where w is the raw weight of the relation, and s is the sum of raw weights around the node. Based on a null hypothesis that the normalized weights $\bar{w}$ of the associations around a certain node of degree k are produced by a random assignment from a uniform distribution, the probability density of one of these weights taking a particular value x is

$$\rho(x) = (k-1)(1-x)^{k-2} \tag{4-4}$$

The disparity filter proceeds by identifying which links for each node should be preserved in the network. A link is considered to be statistically heterogeneous if its normalized weight $\bar{w}$ rejects the null hypothesis. Therefore, by imposing a significant level α, the significantly relevant edges of a node will be those whose normalized weights $\bar{w}$ with respect to that node and the degree of that node k satisfy

$$\int_0^{\bar{w}} \rho\,(x)dx = (k-1)\int_0^{\bar{w}} (1-x)^{k-2}dx > \alpha \tag{4-5}$$

Finally, an edge will be preserved if it is significantly relevant to both of its connected nodes. In this way, the semantic network can be filtered robustly to retain statistically relevant relations between the design knowledge concepts.

Results

By unifying the two types of associations extracted from keywords and full-text, a huge semantic network is obtained containing millions of knowledge concepts and relations. A disparity filtering is conducted on this huge semantic network to remove the statistically irrelevant relations and preserve the significantly relevant relations. The details and parameters of the obtained huge semantic network are shown in Table 4-4.

Table 4-4 Details and parameters of the obtained huge semantic network

Parameters	Explanation	Value
N_{papers}	Total number of engineering academic papers used	928,471
N_{posts}	Number of design posts crawled from Yanko design	1,000
τ	Reduced weight for association between concepts in same sentence but different clauses	0.5
N_{nd}	Number of nodes in the merged network before noise filtering	3,157,377
N_{eg}	Number of edges in the merged network before noise filtering	29,991,894
α	Significant level of disparity filter	0.65
N_{nd}^{α}	Number of nodes in the merged network after disparity filtering	536,507
N_{eg}^{α}	Number of edges in the merged network after disparity filtering	3,726,904
w_{max}	Maximum value of the raw weight	2914.0
w_{min}	Minimum value of the raw weight	0.5

For the significant value α, as it increases from 0 to 1, the graph size will become smaller and the threshold of relevant degree for relations to be preserved will be higher. For $\alpha < 0.5$, we found that it still yields a very large graph containing more than 10 million edges, which leads to a very low retrieval speed of the subsequent network analysis. The graph size can be significantly reduced to below 4 million edges for an acceptable query speed (within 2 s) when α is higher than 0.6. The retrieved top ranked results will almost keep the same until α is increased above 0.8 where even some very relevant relations are removed and valuable information begin to loss. In order to have an acceptable query speed and also keep the relevant associations as complete as possible, we chose the value 0.65 from the ideal range $0.6 < \alpha < 0.8$.

In conclusion, the work of above sections achieves the transformation of the unstructured data into structured information through the process from raw text, to concepts, relations and finally the semantic network. This filtered huge semantic network will be used throughout our following design information retrieval process.

4.5 Study 3: Concepts and Relations, Precision and Recall

Golden dataset

To evaluate our constructed semantic network, we use a standard list of common machine elements and mechanical engineering concepts, which is shown in *Table 1.3 of Mechanical Design Engineering Handbook* (Childs, 2013), as a technology-based dataset to validate whether our constructed ontology network indeed contains more design and engineering-oriented knowledge concepts and relations. The dataset readily lists 205 inherent-related machine elements and engineering concepts, which are already divided into nine categories in the dataset. Table 4-5 shows some examples of the machine elements and engineering concepts for each category. Some concepts may belong to more than one category such as *Roller* included in both *Energy transmission* and *Friction reduction*. For convenience, we call these listed 205 machine elements and engineering concepts as "golden concepts" in our following experiments.

Table 4-5 Some examples of machine elements and engineering concepts in each category

Categories	Related machine elements and concepts
Energy conversion	Turbomachinery, Internal combustion engines, …
Energy transmission	Gear, belts, chains, couplings, cranks, …
Energy storage	Flywheel, spring, fluid accumulator, pressure vessel, …
Locating	Threaded fasteners, washer, nails, pins, clamps, …

Friction reduction	Rolling elements, bearings, recirculating ball, …
Switching	Clutches, ratchet, pawl, valve, bimetallic strip…
Sealing	Seals, lip ring, ferrofluidic seals, gasket, O rings, …
Sensor	Temperature, thermocouples, pressure, manometer, piezoelectric, laser Doppler, …
Miscellaneous mechanism	Hinges, pivot, drills, pulleys, centrifuges, filters, linkage, …

Among the 205 golden concepts, human efforts and expertise have been conducted to judge whether any two of the 205 concepts are related to each other. We regard a pair of concepts as being related if the two concepts satisfy at least one of the following three criteria:

(1). Similar function: the two concepts have the similar engineering function or belong to the same class. For example, both thermocouple and thermometer are used for temperature measurement purpose, so we identify the existence of relationship between thermocouple and thermometer (thermocouple ⇔ thermometer) where "⇔" means the two concepts are associated. For another example, both spur gear and helical gear are different types of gears while in the same class; thus, we identify spur gear and helical gear to be relevant (spur gear ⇔ helical gear).

(2). Attachable structure: one of the two concepts can be combined or adapted with the other one. For example, pump system can often be adapted with centrifuge; so there also exists association between pump and centrifuge (pump ⇔ centrifuge). For another example, roller, bearing, and rope are often combined as a partial structure for transmission system; thus, we can have relations among all the three concepts (roller ⇔ bearing, bearing ⇔ rope, rope ⇔ roller).

(3). Transferable knowledge: there exists knowledge intersection and dependency or cause-effect between the two concepts. For example, the design of

turbomachinery involves the use of knowledge about fluid dynamics, so we regard turbomachinery to be related with fluid dynamics (turbomachinery ⇔ fluid dynamics). Another example is that heat exchanger and combustion engine both share the knowledge in terms of the heat transfer; thus, heat exchanger and combustion engine are identified to be related to one another (heat exchanger ⇔ combustion engine). Besides, for an example from cause-effect perspective, the trigger of a bimetallic strip depends on the changes of temperature; therefore, we can also say that there exists association between bimetallic strip and temperature (bimetallic strip ⇔ temperature).

Finally, a minimum of 565 golden relations were identified between the 205 golden concepts by human judgment independent of the test in order to circumvent coding bias. Table 4-6 shows some examples of these human-judged relations among the 205 concepts for each criterion. Each relation actually represents a pair of concepts selected from the 205 concepts (e.g., cam ⇔ lever), which satisfy at least one of the above three criteria. Some relations may potentially satisfy more than one criterion.

Table 4-6 Some examples of the 565 human-judged relations for each criteria

Similar function	Attachable structure	Transferable knowledge
o ring ⇔ gasket	Bolt ⇔ nut	Generator ⇔ centrifuge
Spur gear ⇔ worm gear	Cam ⇔ lever	Spring ⇔ energy storage
Cone clutch ⇔ disk clutch	Piston ring ⇔ internal combustion engine	Manometer ⇔ pressure
Fan ⇔ heat exchanger	Fuel cell ⇔ electric motor	Thermopile ⇔ energy conversion
Pyrometer ⇔ thermocouple	Thermometer ⇔ pitot tube	Gardon gauge ⇔ heat flux
Reciprocating engine ⇔ rotary engine	Combustor ⇔ turbine	Piezoelectric ⇔ stress
Hose ⇔ pipe	Bearing ⇔ rivet	Flywheel ⇔ energy storage
Latch ⇔ fastener	Compressor ⇔ propeller	Laser Doppler ⇔ velocity measurement
Belt ⇔ chain	Pulley ⇔ rope	Brake ⇔ friction

Rubber ⇔ sealant	Pivot ⇔ wheel	Damper ⇔ shock absorber
…	…	…

Three public available huge ontology-based databases (WordNet (PrincetonUniversity, 2010), ConceptNet(Speer and Havasi, 2012), and NeLL (Carlson et al., 2010)) can be used as baselines to compare with our constructed network. As stated in Chapter 2, they are three representative ontology-based approaches, which establish the semantic relations through the hand-built method, unsupervised and semi-supervised learning, respectively from the whole internet, and all of them aim to capture all the words / phrases as well as any possible relations between the words / phrases.

Retrieving Concepts

First, we check whether the 205 golden concepts are contained in our ontology network and the three benchmark systems in order to test whether our constructed ontology network has indeed captured the engineering-specific knowledge concepts. For this purpose, we use the concept retrieval rate C_R as the metric of concept retrieval:

$$C_R = \frac{n_C}{N_C} \tag{4-6}$$

where N_C is the number of golden concepts, which is 205 in this case, and n_C means how many of these N_C concepts are contained in the system. For example, we find that WordNet only contains 133 of these 205 concepts and therefore its C_R rate is 0.65. Table 4-7 compares the retrieval rate C_R of each approach for different categories. We can see that our approach outperforms other databases for total retrieval rate C_R, which indicates that our constructed ontology network covers more engineering-specific knowledge concepts than the three benchmark systems.

Specifically, our approach involves more concepts in the categories of *energy conversion*, *friction reduction*, *sealing*, and *sensor*, while ConceptNet shows better performance regarding the categories of *locating*, *switching*, and *miscellaneous mechanism*.

Table 4-7 Concept retrieval results

Categories	WordNet	ConceptNet	NeLL	Our approach
Total rate C_R	0.65	0.75	0.19	**0.82**
Energy conversion	0.71	0.86	0.25	**1.0**
Energy transmission	0.66	**0.76**	0.21	**0.76**
Energy storage	0.4	**0.8**	0.13	**0.8**
Locating	0.62	**0.76**	0.15	0.71
Friction reduction	0.3	0.3	0.1	**0.7**
Switching	**0.68**	**0.68**	0.11	0.63
Sealing	0.54	0.62	0.15	**0.77**
Sensor	0.81	0.87	0.29	**1.0**
Miscellaneous mechanism	**1.0**	**1.0**	0.27	0.95

Note: Boldface values are the maximum value in each row

Node Strength of Retrieved Concepts

As shown in Table 4-7, our constructed ontology network retrieves 168 (82%) of the original 205 concepts. Therefore, we conduct further analysis on each of the retrieved 168 concepts by evaluating their node strength in our constructed network, where the node strength of a concept means the sum of raw weights of all edges incident to that node in the network. Table 4-8 ranks the retrieved 168 concepts by their node strength and shows the top ten and last ten concepts. We can see that all of the top ten concepts (except *fuel cell*) are very general concepts such as *stress*, *temperature*, and *sensor* while the last ten concepts with the lowest node strength are all very specific concepts in particular domain like *worm gear*, *pitot tube*. This indicates that general concepts usually have high node strength due to their diverse associations

with other concepts in wide engineering areas while specific concepts are shown to have lower node strength and only associate with limited knowledge particular within its own domain.

Table 4-8 Node strength of the retrieved 168 concepts in our constructed network

Highest Ten nodes	Concept	Node strength	Lowest Ten nodes	Concept	Node strength
1	Stress	34,453	159	Gardon gauge	2
2	Temperature	21,101	160	Pitot tube	2
3	Fuel cell	12,739	161	Drum clutch	2
4	Resistance	9299	162	Worm gear	2
5	Sensor	8490	163	Hydrodynamic bearing	2
6	Friction	7393	164	Solid mass	2
7	Strain	4916	165	Torsion bar	2
8	Heat exchanger	4295	166	Gas spring	2
9	Pressure	3175	167	Mechanical face seal	2
10	Energy storage	2458	168	Hydrostatic bearing	2

Therefore, our constructed network can utilize the node strength property to roughly arrange a top-down concept structure from broad general level to specific detailed level within in a particular domain. For example, we select the concepts regarding “temperature” and “heat” from the retrieved 168 concepts, and then arrange these concepts based on their node strength and relations in our constructed network. Figure 4-8 plots the arranged concepts along with their node strength and relations, where node strength is used to position a concept in vertical direction, and relations are used to aggregate associated concepts in horizontal direction.

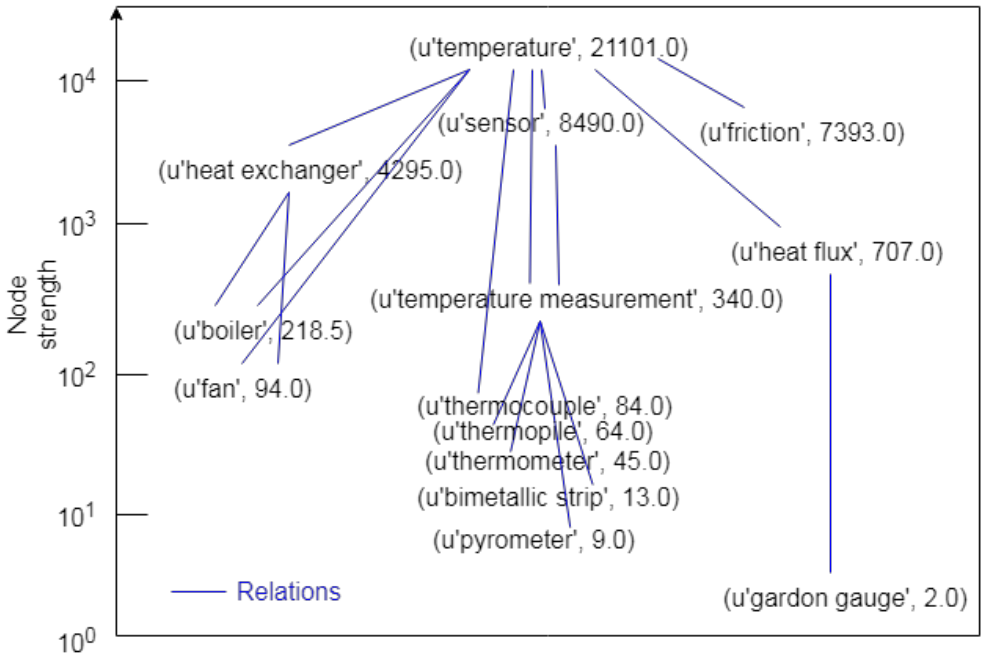


Figure 4-8 A top-down concept structure arranged by the node strength property

For example, (u'thermocouple', 84.0), meaning the node strength of concept *thermocouple* is 84.0, is located at 84.0 of the vertical coordinate, while in horizontal direction, it is just positioned closer with its related concepts (e.g., *sensor*, *temperature measurement*, *thermopile*) and separated from unrelated concepts (e.g., *heat exchanger*, *fan*).

We can see that it automatically forms a rough concept structure, which starts from a very general concept (*temperature*) at the top and extends downward to three specific branches (except *friction*). The left branch associates from *heat exchanger* to two specific heat transfer instances (*fan* and *boiler*). The middle branch goes down through *sensor* and then *temperature measurement*, and finally reaches a variety of specific methods for temperature measuring such as *thermocouple* and *bimetallic strip*. The right branch connects *temperature* with *heat flux*, which further extends to *gardon gauge*, a specific instrument for the measurement of high intensity radiation. Overall, it is shown in Figure 4-8 that the use of the node strength can reasonably divide the concepts into different levels of specific degree to help structurally illustrate the captured design and engineering relations.

Retrieving Relations

Here, by using the 565 human-judged golden relations, we aim to compare the performance of our constructed network and benchmark systems in retrieving the design and engineering-oriented relations. For a system, we first retrieve all its relations existing between the 205 golden concepts, and denote these retrieved relations as set $\boldsymbol{V}$. Then, let $\boldsymbol{H}$ denote the set of the 565 human-judged relations among the 205 concepts. The precision P, recall R, and $F1$ score of this system for relation retrieval can be accordingly represented as:

$$P = \frac{|\boldsymbol{V} \cap \boldsymbol{H}|}{|\boldsymbol{V}|} \tag{4-7}$$

$$R = \frac{|\boldsymbol{V} \cap \boldsymbol{H}|}{|\boldsymbol{H}|} \tag{4-8}$$

$$F1 = \frac{2PR}{P + R} \tag{4-9}$$

Table 4-9 Compares the relation retrieval results of each approach. Through natural language toolkits, WordNet is found to have totally 59 relations between the 205 concepts, in which 55 relations belong to the 565 golden relations while the other four relations fall out of the scope of the 565 golden relations. Therefore, we have $|\boldsymbol{V}| = 59$ and $|\boldsymbol{V} \cap \boldsymbol{H}| = 55$, resulting in a very low recall 0.097, but an extremely high precision 0.932. The low recall is due to the tree data structures of WordNet where relations are limited to either between synonyms, sister terms, or between hypernym and hyponym. For example, most of retrieved relations from WordNet are like (spur gear ⇔ worm gear), (spur gear ⇔ gear), (cone clutch ⇔ disk clutch), and (spiral spring ⇔ leaf spring), which hardly contain associations between distant concepts such as (spring ⇔ energy storage). The high precision makes sense because the relations in WordNet are also hand-crafted, which should be consistent with the similar-function criteria of our human judgment for relations.

Table 4-9 Relation retrieval results

Categories	Recall	Precision	F1
WordNet	0.097	**0.932**	0.176
ConceptNet	0.159	0.882	0.270
NeLL	--	--	--
Our approach	**0.409**	0.802	**0.542**

Note: Boldface values are the maximum value in each column

We use the official web API to retrieve the relations from ConceptNet and NeLL, respectively. ConceptNet is found to contain 102 relations between the 205 concepts, in which 90 relations are true positive belonging to the 565 relations and the other 12 relations are false positive, providing an improved recall 0.159 (compared to WordNet) and still a very high precision 0.882. ConceptNet essentially follows several common predefined linguistic patterns (e.g., *IsA*, *HasProperty*, *FormOf*, *PartOf*, *AtLocation*) to extract relations from text, which, to some extent, enriches the types of extracted relations. For example, the retrieved relations of ConceptNet are not restricted to simple relations of synonyms or hypernym, but extend to some sophisticated relations satisfying the attachable structure or transferable-knowledge criteria such as (o ring ⇔ rubber), (pulley ⇔ chain), (friction ⇔ brake), and (fan ⇔ propeller). Also, another advantage is that linguistic patterns inherently communicate the human expressions and judgements, which indirectly contributes to the high precision of ConceptNet. However, it is still far from sufficient for the predefined linguistic rules to fully recall the inherent design and engineering relations due to the highly diverse and complex nature of the design knowledge.

For NeLL system, which has a very low concept retrieval rate only containing 39 of the 205 golden concepts at first time, it can hardly retrieve any relations between these concepts. This means the ontology database of NeLL system is not specialized in design and engineering knowledge concepts and relations.

Compared to the above benchmark systems (WordNet and ConceptNet), our constructed ontology network actually sacrifices a little bit of precision for much improvement of recall, as shown in Table 4-9. In our constructed ontology network, 288 relations are retrieved between the 205 golden concepts, in which 231 relations are true positive belonging to the golden relations while the other 57 are false positive, resulting in a much improved recall 0.409 with a little degraded precision 0.802. Our approach actually unfreezes the restriction on the predicates of predefined linguistic patterns, and uses either the keyword associations or a simplified NLP instead to detect the noun phrases of subjects and objects among which the relations are established using itemset mining at sentence level. This will certainly augment our approach with more diverse types of associations between any possible concepts, and therefore sophisticated relations can be captured between design and engineering concepts such as "energy conversion ⇔ piezoelectric," "rim seal ⇔ turbomachinery," "compressor ⇔ turbine," "friction ⇔ temperature," "clutch ⇔ brake," "electric motor ⇔ fuel cell," and "pitot tube ⇔ velocity measurement." However, the drawback of our approach is that we unavoidably take a risk of bringing in more noisy relations established between two unrelated concepts (e.g., "heat exchanger ⇔ sound," "fuel cell ⇔ sensor") that just happen to be the keywords in the same paper or the objects of the same sentence, which therefore leads to a poor precision.

For this reason, the disparity filter was incorporated in our approach to statistically remove the noise and preserve the significantly relevant relations. Although the final produced precision 0.802 is not as perfect as WordNet and ConceptNet, it is still acceptable. Overall, with a small amount of cost in precision, our constructed ontology network largely boosts the recall of design engineering relations and significantly improves the whole *F1* score. We may further argue that increasing the significant level α of disparity filter may help eliminate more noise and irrelevant relations to further improve the precision, but meanwhile this will also potentially

remove the golden relations having an adverse impact on recall. In our case, we just persist with the optimal α value of 0.65 for fast query and complete information as discussed in Section 4.4, and leave the adjustment of the α value for future work to explore a balance between precision and recall.

4.6 Discussion and Conclusion

This chapter has presented the process of constructing the structured ontology network based upon the unstructured textual data. The necessity of this process is because the traditional document retrieval methods solely depending on unstructured textual documents can only process and retrieve the information at fragment level, while structured ontology-based approaches focus on extracting relevant information at the semantic level. Overall, this ontology network construction can be concluded as a process from raw texts, to concepts, relations and finally the network.

Two novel mining approaches are developed to extract the concepts and relations from the previously captured raw textual data. One is to creatively exploit the inherent relations between keywords by using association rule learning based on the metadata of the academic papers. The other approach combines a simplified natural language processing procedure with itemset mining to extract the concepts and relations from full texts of the design posts and academic papers. In this approach, phrase chunking is used to recognise the noun phrases as the knowledge concepts, and itemset mining is used to build relations between noun phrases within the same sentence. The extracted concepts and relations of this two approaches are unified together to construct a huge semantic network which is then cleaned by a disparity filter to remove the noisy, irrelevant relations with relatively lower weights.

At the end of this chapter, we used a golden dataset containing professional engineering design concepts and relations to evaluate our constructed ontology

network. The intention is to investigate whether our approaches can indeed capture more knowledge concepts and relations from the design and engineering perspectives and contexts. We have compared the precision and recall of our ontology network with three other public ontological databases including WordNet, ConceptNet and NeLL. We found our approach outperforms the other three benchmarking methods in two aspects:

- Firstly, our ontology network contains more specific and professional concepts and technical terms compared to the other three benchmarking systems. It is because our data resource involved a huge amount of academic research articles and we captured both the keywords and the noun phrases, which can achieve a high coverage on the technical terms and phrases. However, in WordNet, most entities are single separate words, and do not include a lot of phrases and terms. While for ConceptNet, the data resources are mainly focusing on common sense, which results in a constructed ontology network representing general knowledge concepts. Therefore, both WordNet and ConceptNet has a low retrieval rate on design and engineering-specific knowledge concepts.
- The other aspect is that our constructed ontology network contains more design- and engineering-oriented relations, and the types of the relations are much more diverse. WordNet and ConceptNet either rely on the synonym, hypernym, and hyponym or use the predefined linguistic patterns to build the relations, which potentially limits the types of relation they can extract. For our approach, we exploit the inherently diverse associations between keywords to increase the chances for discovering novel relations. Also, instead of following predefined linguistic rules, we skip the restriction of predicates in the sentence and link the noun phrases between subject and object. This increases the possibility to establish more sophisticated relations, for example,

satisfying the attachable structure or transferable knowledge criteria, but the drawback is that noisy relations may be involved in.

Therefore, an interesting point that may be explored further in our ontology network is to investigate the balance between precision and recall by adjusting the significant level α of the disparity filter. Since it falls out of the scope of our following information retrieval process, we will leave this for future work.

Chapter 5 The use of Explicit vs Implicit Networks for Data Insights

The previous chapter explored extraction of structured information in the form of a huge semantic network from the captured raw unstructured textual data. Now, having this constructed semantic network, our key question is how to use this network? How can we utilise this semantic network for supporting design activities such as design information retrieval and idea generation. When provided with a design query, how do we explore the relevant knowledge concepts, and how to rank and retrieve the most plausible, relevant and useful ones? Do we only consider the neighbour nodes around the query in the network as relevant knowledge concepts? If not, how to evaluate the correlations between distant concepts? In addition to relevant knowledge concepts, is there way to find novel links in the network to provoke design creativity and innovative insights?

This chapter aims at tackling above questions by proposing a retrieval framework based on semantic network analysis to support the design information retrieval and improve design innovation and idea generation. Specifically, this framework explores both explicit knowledge associations and implicit associations for the design queries. Novel criteria are established to retrieve and rank both explicit and implicit associations under a unified standard by using various approaches such as applying Pythagorean means, and modelling dedicated probability layer and velocity layer in the semantic network. Dijkstra's shortest path searching is used to discover the knowledge associations either around a single query concept or between two query concepts. Three studies (Study 4, Study 5, Study 6) are conducted to evaluate the

effectiveness of the established novel criteria, and demonstrate the use of our proposed retrieval framework in real practice to support the design relations retrieval and provoke creative idea generation.

Some of the work described in this chapter has been previously published in (Shi et al., 2017b, Shi et al., 2017a):

1. Shi, F., Chen, L., Han, J. & Childs, P. 2017b. Implicit Knowledge Discovery in Design Semantic Network by Applying Pythagorean Means on Shortest Path Searching. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, V001T02A053-V001T02A053. Copyright © ASME 2017. Reprinted by permission of ASME.

2. Shi, F., Chen, L., Han, J. & Childs, P. 2017a. A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *Journal of Mechanical Design,* 139**,** 111402. Copyright © ASME 2017. Reprinted by permission of ASME.

5.1 Explicit and Implicit Knowledge Associations

Relevant Concepts and Relations

As shown in previous chapter, the combination of text mining and ontology technologies has already enable us to extract structured knowledge associations from unstructured textual data and subsequently represent and store the knowledge associations in form of structured semantic network. Obviously, the next step would be how to utilise this constructed semantic network to assist design and engineering –related activities such as information retrieval and idea generation.

The major use of design ontology and semantic network for supporting design information retrieval is to retrieve useful design relations and concepts related to the original design query and need (Zhang and Yin, 2008, Darlington and Culley, 2008, Chandrasegaran et al., 2013). This can help designers and engineers obtain specific and relevant knowledge concepts and relations about their queries, which can be useful in four ways:

(1). The specific and professional knowledge concepts related to the queries may possibly directly provide feasible solutions to their problem from the perspectives regarding the functions, structures, mechanisms, methods, components, materials, and even power supplies.

(2). The retrieved knowledge concepts may represents relevant knowledge areas of the query. This enables the designers to be aware of the related domains of their query, and encourages them to progress farther into related domains for an idea or solution.

(3). Alternatively, the retrieved concepts can be further used by the designer to expand the original query to form a more informative query for further document retrieval (Dong and Agogino, 1997).

(4). The extracted relations can help designers build a hierarchical structure of the knowledge around their queries (as illustrated in Figure 4-8), and finally capture a brief overview of the retrieval process.

This can significantly facilitate the information retrieval process at different stages. However, to retrieve relevant concepts and relations, the first question would be how should we define “related or relevant”? Should we only consider as relevant concepts the neighbour nodes of the query in the semantic network? Even if we have defined things to be “related”, how should we rank the relevance? Is the relation with the maximum weight to be the most relevant?

Explicit vs Implicit

Most of the knowledge associations extracted by the well-established text mining methods are explicit in nature, which means the associated two knowledge concepts are directly related following a co-occurrence manner in the texts and are extracted by either statistical significance or linguistic approaches (Bullinaria and Levy, 2007). However, design creativity and innovative events can often be observed to happen in implicit, loose and surprising links between design information and knowledge (Dorst and Cross, 2001). In order to improve design innovation and idea generation, it may be also worthwhile to explore implicit associations where two concepts are not directly relevant, but they can be indirectly linked through a series of other bridging concepts in between to help provoke innovative insights. Many current research activities mainly focus on how to construct the ontologies and semantic networks by extracting explicit associations between concepts through text mining methods. However, few studies are found to focus on how to better analyse and fully utilize the already existing and well-established semantic networks to explore implicit associations.

Although statistical approaches and linguistic approaches have been extensively used as the two prevailing unsupervised text mining techniques to extract and recognize the structured knowledge associations and information patterns from massive unstructured textual documents, however, both of them are essentially limited to the recognition of co-occurrence, and are therefore only able to discover explicit relations between two individual concepts that appear together within the document level, paragraph level or even sentence level (Feldman and Sanger, 2007).

Not only within design domain, actually, the knowledge associations extracted by many studies in various domains are essentially explicit in nature by using the text mining and information extraction. For example, this includes the associations between symptoms and medical effects in healthcare industry (Spasic et al., 2005),

the relations between criminal records and legal statements in law practice (Wyner et al., 2010), and even the relationships between teaching materials and recipients preference in education area (Munoz and Tucker, 2016). In the domain of product design, associations between lead users and product features have been discovered by textual analysis on the large scale social media networks (Tuarob and Tucker, 2015). Patent documents are often used to explore the associations among the professional expertise for technique trend analysis and design innovation inspiration (Liang et al., 2012). Similarly, research papers and design progress reports for extracting the associations between design configurations and products can help designers with the knowledge reuse for ideation and planning in manufacturing system (Efthymiou et al., 2015).

Therefore, many research teams have already studied extracting the explicit knowledge associations from raw texts and directly linking the knowledge concepts to construct various kinds of ontologies and semantic networks. In order to make full use of our already well-established semantic networks for idea generation and design innovation, it may be also worthwhile to explore not only the explicit knowledge associations of directly related concepts based on the co-occurrence manner, but also the implicit associations between concepts that are not directly linked but can be indirectly connected through a series of other bridging concepts.

Formal Representation of Explicit and Implicit Associations

In a design semantic network, explicit knowledge associations directly linking the concepts can be regarded as the basic edges connecting the nodes. Thousands of explicit associations (edges) join together to constitute the backbone of the semantic network. While, implicit knowledge associations indirectly linking the concepts can be regarded as the paths connecting through a series of intermediate bridging nodes. For a simple example in Figure 5-1, if node A represents the design query, direct edges such as (A, B), (A, C) and (A, D) can be regarded as explicit knowledge

associations linked to explicit knowledge B, C and D. On the other hand, concepts X and Y can be discovered as the implicit knowledge, which are indirectly linked to A via different paths as implicit knowledge associations, e.g. $(A-B-C-X)$, $(A-D-Y)$. From the design innovation perspective, creative events are believed to often happen in the implicit associations which link design query A with implicit knowledge concepts X, Y into a coherent chunk (Dorst and Cross, 2001).

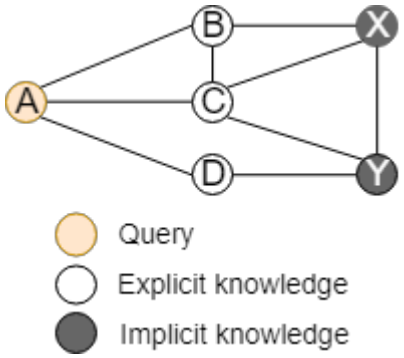


Figure 5-1 Illustration of explicit and implicit knowledge associations in semantic network.

5.2 Retrieval framework

As discussed above, the major use of the design ontology and semantic network is to retrieve relevant knowledge concepts and relations for the design queries to support the knowledge demands of different stages of the design process, where both the explicit and implicit knowledge associations will be focused in our research. Therefore with this regards, we propose a retrieval framework to retrieve both explicit and implicit knowledge associations, where unified standards / criteria are set up to quantitatively evaluate and rank their correlation with the query concepts. The proposed retrieval framework is shown in Figure 5-2. Firstly, the design query is reinterpreted or represented as one or more design concepts as the key query concepts. If the query can be represented as only one concept, knowledge associations are explored around this single concept. Otherwise if the query is

represented by more than one design concepts, in addition to exploring around each concepts, we also construct concept pairs and search the association paths between the two concepts for each pair. We can explore both explicit and implicit knowledge association around a concept, while paths searched between two concepts are usually implicit associations. Then unified standards are set up to quantitatively evaluate and rank the correlation degree of explicit associations (between directly linked nodes) and implicit associations (between indirectly linked nodes) under the same criteria, which can be achieved by the modelling of probability and velocity layer as well as the use of Pythagorean means in the semantic network. Finally, shortest path search is conducted based on the above unified quantitative criteria to retrieve the explicit and implicit knowledge associations as the retrieval results including relevant concepts and relations.

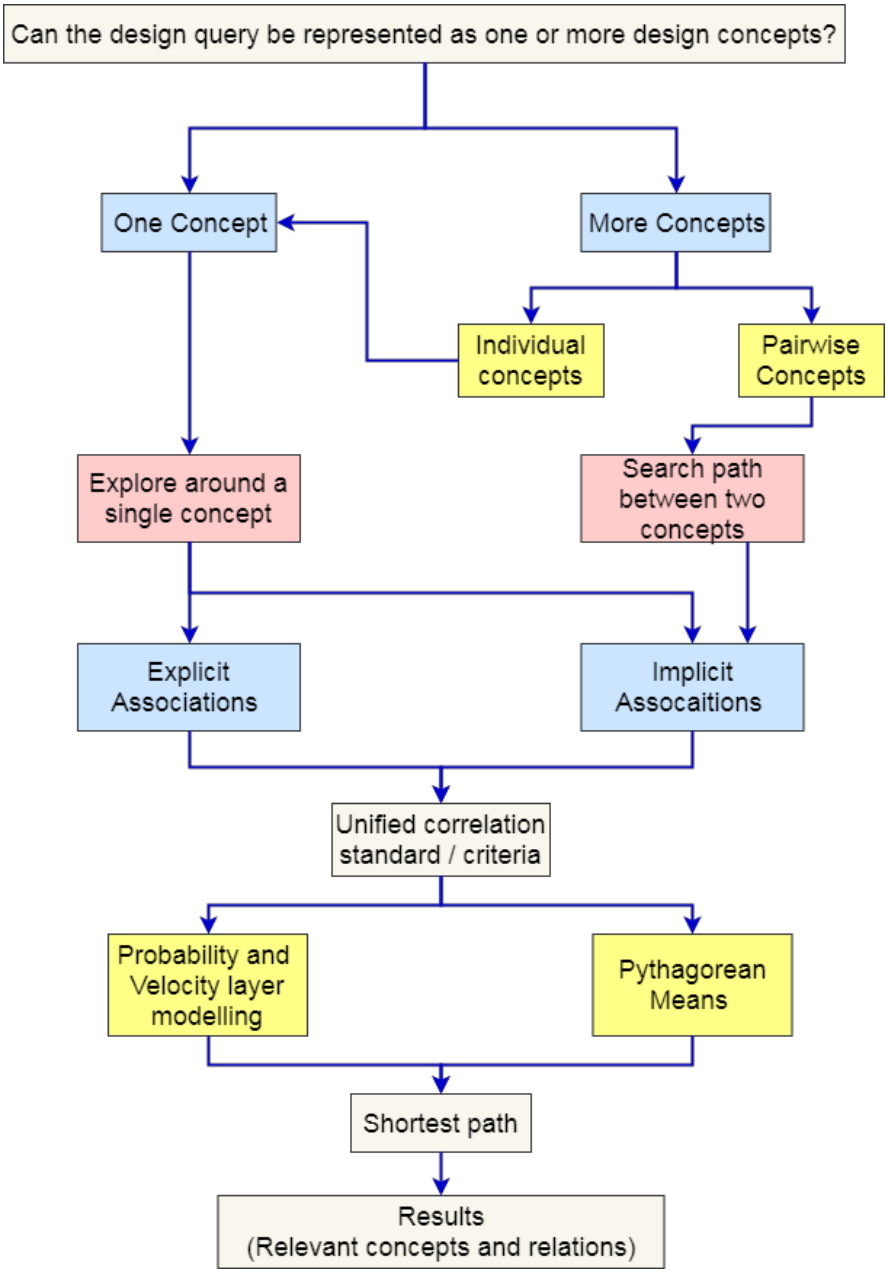


Figure 5-2 Proposed retrieval framework

5.3 Probability and Velocity Layer Modelling

Since we need to retrieve both explicit and implicit knowledge associations for a query, a unified standard should be set up to retrieve and quantitatively rank both of them under the same criteria. For this purpose, probability layer and velocity layer are innovatively modelled in the semantic network as a unified quantitative standard

to quantify the correlation degree between any two concepts regardless of whether it's explicit or implicit associations.

Strength of Explicit Associations – Weight Normalisation

In a semantic network, explicit knowledge associations are the basic edges, and implicit knowledge associations are the paths consisting of multiple edges, which means implicit knowledge association is essentially a concatenation of a series of interconnected explicit knowledge associations. Therefore, in order to evaluate the correlation degree of implicit knowledge associations, the strength of explicit knowledge associations should be firstly quantified. Weight values of the edges provide valuable information for characterizing the strength of explicit knowledge associations where higher weight indicates stronger and closer relevance of the explicit association between the two directly linked concepts. However, the raw weights usually need to be normalized with practical statistical meanings for comparison.

Since different normalization methods are based on different perspectives and scales, and may potentially have distinct effects on the subsequent analysis, in our research, two normalized weights are used to quantitatively represent the strength of explicit knowledge association in the interval [0,1]:

$$\overline{w}_{ij}^{g} = {(w_{ij} - w_{min})}\Big/{(w_{max} - w_{min})} \tag{5-1}$$

$$\overline{w}_{ij}^{l} = {w_{ij}}\Big/{s_i} \tag{5-2}$$

where w_{min} and w_{max} are the minimum and maximum values of the raw weights in the whole network, w_{ij} is the raw weight of the association between node i and j, and s_i is the sum of raw weights incident to the node i. Specifically, which one of the two

nodes i, j to be counted for s depends on the retrieval direction, which is $i \rightarrow j$ in this case.

Equation (5-1) performs feature scaling normalization from a global perspective, in which $\overline{w}_{ij}^{g}$ expresses the significance of the strength of an explicit association compared to the whole semantic network (Antoniou and Tsompa, 2008). On the other hand in Eq. (5-2) from local perspective, the normalization through local fluctuation $\overline{w}_{ij}^{l}$, which is already applied in previous noise filter in Chapter 4.4, focuses on the relative importance of an explicit association to its own connected node i (Serrano et al., 2009).

Probability and Velocity Layer Modelling and Analysis

In order to retrieve the most relevant knowledge concepts for a design query, the traditional and common way is just to retrieve the neighbours of the query (explicit associations) as related concepts, which are simply ranked in order of the raw weight values of the associations between the neighbours and query, and this does not consider the non-neighbour concepts (implicit associations), which can be indirectly connected to the query through other bridging concepts. As discussed, for design information retrieval, innovative and valuable insights probably happen between loose and implicit linked information (Dorst and Cross, 2001). Hence, in order to retrieve and rank both directly and indirectly linked nodes (explicit and implicit) under the unified standard, probability and velocity analysis of the network simulating the information flow are modelled to quantify the correlation degree between any two concepts.

Probability analysis is established on a directed graph, where $\overline{w}_{ij}^{l}$ normalized by local fluctuation is the weight value of the directed edge $i \rightarrow j$. Assuming the probability that the information flows from a node i to its neighbors is 1, $\overline{w}_{ij}^{l} = {}^{w_{ij}}/_{s_i}$ represents the probability of the information flowing through the association $i \rightarrow j$ from i to j.

Thus, the probability of the information flowing through a path of n concepts ($C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_n$) is

$$P(C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_n) = \frac{w_{12}w_{23} \ldots w_{(n-1)n}}{s_1 s_2 \ldots s_{n-1}} = \prod_{k=1}^{n-1} \frac{w_{k(k+1)}}{s_k} = \prod_{k=1}^{n-1} \overline{w}_{k(k+1)}^{l} \tag{5-3}$$

Therefore, we can use the probability value of the most probable path between C_1 and C_n to evaluate the degree of correlation between the two concepts as shown in Eq. (5-4), where $R_p(C_1, C_n)$ is the correlation degree of the two concepts in probability analysis, $\mathbb{C}$ means the whole concept space of the network. In order to search the maximum probability path, dijkstra's shortest path algorithm (Dijkstra, 1959) can be applied by transforming the weight to $-\log \overline{w}_{ij}^{l}$, as shown in Eq. (5-5), where $\wp$ is the set of all paths between the two concepts C_1 and C_n.

$$R_p(C_1, C_n) = \max_{\substack{C_i \in \mathbb{C} \\ k \geq 0}} \{P(C_1 \rightarrow \overbrace{C_i \rightarrow \cdots}^{k} \rightarrow C_n)\} \tag{5-4}$$

$$\underset{\wp}{\operatorname{argmax}} \left\{ \prod_{k=1}^{n-1} \overline{w}_{k(k+1)}^{l} \right\} = \underset{\wp}{\operatorname{argmin}} \left\{ \sum_{k=1}^{n-1} -\log \overline{w}_{k(k+1)}^{l} \right\} \tag{5-5}$$

Alternatively in the velocity analysis, weight can be simulated as the velocity of information flow between concepts, since higher information flow velocity meaning less transmission time implies the closer association between the two concepts. In order to scale the flow rate in the whole network within the range [0,1], $\overline{w}_{ij}^{g}$ by feature scaling can be regarded as the flow velocity of the corresponding link because it is normalized globally and compared to the overall network. If the distance between two adjacent concepts is one, the time cost of information flowing through the path $(C_1 \rightarrow C_2 \rightarrow \cdots \rightarrow C_n)$ is

$$T(C_1 \to C_2 \to \cdots \to C_n) = \frac{1}{\overline{w}_{12}^{g}} + \frac{1}{\overline{w}_{23}^{g}} + \cdots + \frac{1}{\overline{w}_{(n-1)n}^{g}} = \sum_{k=1}^{n-1} \frac{1}{\overline{w}_{k(k+1)}^{g}} \tag{5-6}$$

The degree of correlation between C_1 and C_n in velocity analysis, $R_v(C_1, C_n)$, can be evaluated by the fastest path with minimum time cost between the two concepts as shown in Eq. (5-7). Shortest path algorithm can also be implemented to search the fastest path by transforming the weight to $1/\overline{w}_{ij}^{g}$.

$$R_v(C_1, C_n) = \min_{\substack{C_i \in \mathbb{C} \\ k \geq 0}} \{T(C_1 \to \overbrace{C_i \to \cdots}^{k} \to C_n)\} \tag{5-7}$$

Thus, for a same design query, the retrieved results of the two different analyses can possibly be completely different. In the analysis on probability layer, the most related knowledge concepts are ranked by the correlation degree R_p in Eq. (5-4) from a local perspective using normalization Eq. (5-2). The information will always flow to the most probable surrounding node with regard to the start node in each association step of the path. This enables designers to find the domain-specific concepts within the original problem domain space of the query source node. While unlike probability analysis, the analysis on velocity layer uses R_v in Eq. (5-7) to rank and retrieve the concepts from a global perspective using weight normalization Eq. (5-1), in which the information would prefer to flow on the most significant edges with regard to the whole network and therefore may cross different domains. Through velocity metric, designers can retrieve the general and comprehensive concepts located at the backbone of the network to investigate the intersection with other domains.

5.4 Study 4: Probability and Velocity analysis on the golden relations

Correlation Degree

In this study, the human-judged 565 golden relations in Study 3 is reused to investigate the effectiveness of probability and velocity analysis. The correlation degree for each of these golden relations can be quantified by the probability and velocity analysis, respectively, by using the distance of the shortest path between the two concepts in the probability and velocity layer based on Equation (5-5) and Equation (5-7).

In the prior study, Study 3, Table 4-9 shows our network totally retrieves 231 (41%) from the 565 human-judged golden relations, which means that each of the 231 recalled concept pairs is an explicit knowledge association directly linked by an edge in our constructed network, while the concept pairs of the other 334 undetected relations can be either disconnected without any paths in between, or implicit associations if they are indirectly linked through other bridging concepts in our semantic network. With the assistance of our developed probability and velocity analysis, we are able to evaluate the correlation degree between any two concepts no matter whether they are directly linked by an edge or not.

Therefore, based on Equation (5-4) and (5-7), we search in our network the most probable path and fastest path between the two concepts for each of the 334 undetected relations as well as the 231 recalled relations. Only nine relations (out of the 334 undetected relations) are found to have no paths between their two concepts. Therefore, for the 556 relations (231 recalled and 325 undetected) that have the paths between their two concepts, we compute the path distances (Equation (5-5) and Equation (5-6)) for their most probable path and fastest path, respectively. In order to make comparison, we also do the same for all other possible combinations of the

168 recalled concepts (as shown in Table 4-7), where 11,189 pairs of concepts are found to have paths in between.

Figures 5-3 and 5-4 show the histograms of the shortest path distances between all these pairs of concepts in probability analysis and velocity analysis, respectively. It can be seen that the 556 human-judged golden pairs have significantly smaller path distances, in which the 231 recalled pairs take the smallest distance values. The other 325 undetected golden relations, though not directly linked and become as implicit associations in our network, are also shown to have relatively smaller distances compared to the other possible pairs. Hence, by assessing the path distance of the most probable path and fastest path, the probability and velocity analysis can potentially provide useful and effective information about the correlation degree between any two directly or indirectly linked concepts (explicit or implicit associations), and the evaluated correlation degree is consistent with the human judgements.

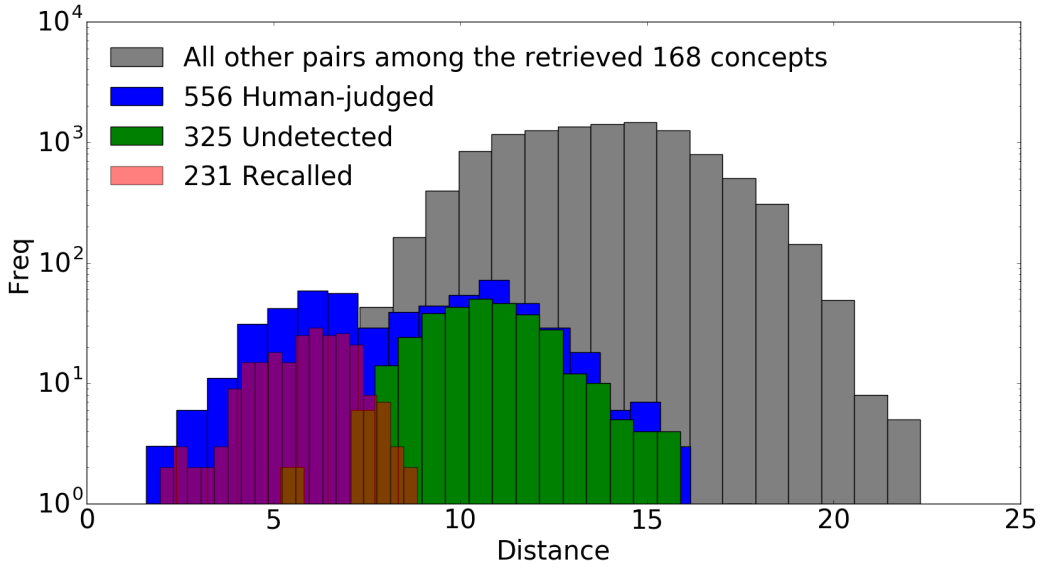


Figure 5-3 Histogram of the shortest path distances in probability analysis

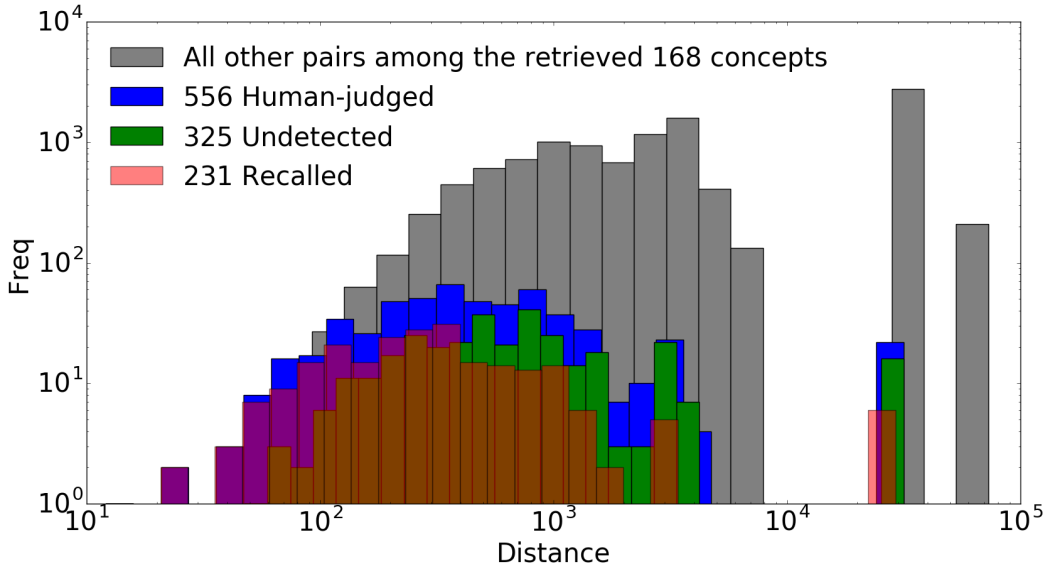


Figure 5-4 Histogram of the shortest path distances in velocity analysis

Retrieval Behaviour

In order to investigate the retrieval behaviours of the probability analysis and velocity analysis, we evaluate the node strength on both the most probable path and fastest path between any two of the 168 recalled concepts. We found the average node strength on fastest paths is more than ten times higher than the most probable paths. Figure 5-5 randomly samples some pairs and shows the node strength of both the most probable path and fastest path for each pair. As illustrated in the Study 3 in Chapter 4, the node strength corresponds with the specific level of the concept in knowledge hierarchical structure, therefore we can accordingly conclude that probability analysis tends to retrieve domain-specific concepts with lower node strength while the velocity analysis prefers to search general and broad concepts with higher node strength. This approves the statements in Section 5.3.

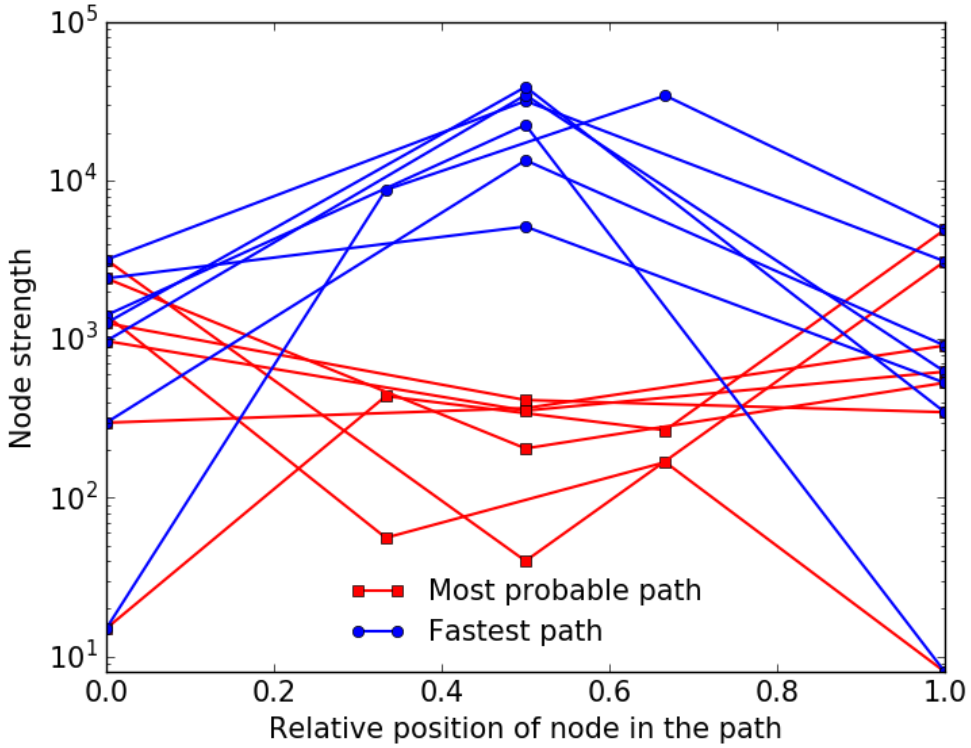


Figure 5-5 Node strength on the most probable paths and fastest paths

5.5 More Criteria by Pythagorean Means

In Section 5.3, the use of probability and velocity modelling has achieved to quantify the correlation degree of both explicit and implicit associations (between directly and indirectly linked concepts) under a unified standard. Besides, the retrieval behaviours of the analysis on probability and velocity layers are very different. Are there any other approaches that can uniformly quantify the correlation degree of explicit and implicit association, and also have distinct retrieval behaviours? Since engineering design activities incorporate different and various kinds of knowledge associations, it would be very beneficial if we can explore more possible quantitative criteria with different retrieval behaviours which can retrieve different types of knowledge concepts or relations to satisfy the different and various kinds of knowledge demands during engineering design activities.

In addition to probability and velocity modelling, this section applies the Pythagorean means to set up unified standards for quantifying the correlation degree between any concepts either directly linked (explicit associations) or indirectly linked (implicit associations).

With the defined strength of explicit associations in Section 5.3, the correlation degree of an implicit knowledge association can be evaluated and quantified based on its contained explicit associations. If we let direct edge (C_i, C_j) represent the explicit association between two concepts C_i and C_j, an implicit association can be then represented as a path $(C_1 - C_2 - \cdots - C_{n+1})$ containing n explicit associations (C_k, C_{k+1}) that $k \in \{1,2, \ldots, n\}$. Section 5.3 actually has discussed about using normalized weight $\bar{w}_{ij}$ to quantify the strength of explicit association(C_i, C_j), and proposed the use of probability and velocity assumptions to quantify the correlation degree of a path $(C_1 - C_2 - \cdots - C_{n+1})$. Similarly in this section, we also propose another two novel assumptions to quantify the correlation degree of a path from a different perspective in terms of Pythagorean means.

Assumption 1

Since an implicit knowledge association is essentially a concatenation of a series of explicit associations, the accumulation of the strength of the contained explicit associations (edges) can potentially indicate the correlation degree of the implicit association (path), that stronger explicit associations (edges) should produce a higher correlation degree of the corresponding assembled implicit association (path). Therefore, in order to reflect the overall strength of all the explicit associations contained in an implicit association $(C_1 - C_2 - \cdots - C_{n+1})$, three classical Pythagorean means (PMs) including the arithmetic mean (AM), the geometric mean (GM), and the harmonic mean (HM) can be respectively applied on the normalized weights of all edges contained in the path:

$$PMs(C_1 - C_2 - \cdots - C_{n+1}) = \begin{cases} AM: & \frac{1}{n} \times \sum_{k=1}^{n} \overline{w}_{k,k+1} \\ GM: & \sqrt[n]{\prod_{k=1}^{n} \overline{w}_{k,k+1}} \\ HM: & \frac{n}{\sum_{k=1}^{n} \frac{1}{\overline{w}_{k,k+1}}} \end{cases} \quad (5\text{-}8)$$

Assumption 2

The correlation degree of an implicit association should decrease with the increase of the quantity of the edges contained in the path n. It is reasonable that a long path with multi-step explicit associations could certainly cross different knowledge domains and link irrelevant information at the two ends. On the other hand, without the constraint of the quantity of included edges n, it may lead to very long path, and therefore tremendous computing workload to search the implicit association with the highest Pythagorean means (PMs) in the huge semantic network.

Therefore, based on both of the **Assumption 1** and **Assumption 2**, the correlation degree of an implicit association (path) can be represented as:

$$R(C_1 - C_2 - \cdots - C_{n+1}) = f_R(PMs, n) \quad (5\text{-}9)$$

According to **Assumption 1**, we have

$$\frac{\partial f_R}{\partial PMs} > 0 \quad (5\text{-}10)$$

meaning that the correlation degree gets higher with the increasing of the Pythagorean means (PMs) of the normalized weights of all edges contained in the path.

According to **Assumption 2**, we can have

$$\frac{\partial f_R}{\partial n} < 0 \tag{5-11}$$

meaning that the correlation degree gets lower with the increasing of the number of the included edges n in the path.

In order to apply shortest path algorithms in searching the knowledge associations in sequence from highest correlation degree to lowest, the correlation degree of the implicit association should be inversely mapped into the distance of the corresponding path, which means that shorter distance of the path corresponds to higher correlation degree, and therefore higher PMs and smaller n. Thus, the partial derivatives of path distance with respect to PMs and n are opposite to the partial derivatives of correlation degree with respect to PMs and n, so that the distance of path can be defined as:

$$D(C_1 - C_2 - \cdots - C_{n+1}) = f_D(PMs, n) \tag{5-12}$$

$$\frac{\partial f_D}{\partial PMs} < 0 \tag{5-13}$$

$$\frac{\partial f_D}{\partial n} > 0 \tag{5-14}$$

To satisfy the conditions of partial derivatives in Equation (5-13) and (5-14), and enable the total distance to be allocated to each individual edge, three specific constructor functions for path distance $f_D(PMs, n)$ can be established for each of the three Pythagorean means respectively:

$$f_D(AM, n) = (1 - AM) \times n = \sum_{k=1}^{n} (1 - \overline{w}_{k,k+1}) \tag{5-15}$$

$$f_D(GM,n) = (-\log GM) \times n = \sum_{k=1}^{n} (-\log \overline{w}_{k,k+1}) \tag{5-16}$$

$$f_D(HM,n) = \frac{1}{HM} \times n = \sum_{k=1}^{n} \frac{1}{\overline{w}_{k,k+1}} \tag{5-17}$$

Thus, for the path distances represented in Equation (5-15), (5-16) and (5-17), we can respectively assign $1 - \overline{w}_{k,k+1}$, $-\log \overline{w}_{k,k+1}$ and $1/\overline{w}_{k,k+1}$ as the distance of individual edge. By substituting $\overline{w}_{k,k+1}$ with the two normalized weights as shown in Equation (5-1) and (5-2) from global and local perspective in Section 5.3, six specific criteria can be established to represent the distance of the path $(C_1 - C_2 - \cdots - C_{n+1})$, as shown in Table 5-1, where the directed weight $\overline{w}_{ij}^{l}$ normalized by local fluctuation is converted to the undirected values, $(\overline{w}_{ij}^{l} + \overline{w}_{ji}^{l})/2$, $\sqrt{\overline{w}_{ij}^{l} \times \overline{w}_{ji}^{l}}$, and $2\overline{w}_{ij}^{l}\overline{w}_{ji}^{l}/(\overline{w}_{ji}^{l} + \overline{w}_{ji}^{l})$, for AM, GM and HM respectively.

Table 5-1 Criteria for quantifying the correlation degree of any paths

Criteria	Pythagorean mean	Normalization	Path distance
AM_g	AM	Feature scaling	$\sum_{k=1}^{n} (1 - \overline{w}_{k,k+1}^{g})$
AM_l	AM	Local fluctuation	$\sum_{k=1}^{n} (1 - \frac{\overline{w}_{k,k+1}^{l} + \overline{w}_{k+1,k}^{l}}{2})$
GM_g	GM	Feature scaling	$\sum_{k=1}^{n} (-\log \overline{w}_{k,k+1}^{g})$
GM_l	GM	Local fluctuation	$\sum_{k=1}^{n} (-\log \sqrt{\overline{w}_{k,k+1}^{l}\overline{w}_{k+1,k}^{l}})$
HM_g	HM	Feature scaling	$\sum_{k=1}^{n} \frac{1}{\overline{w}_{k,k+1}^{g}}$

HM_l	HM	Local fluctuation	$\sum_{k=1}^{n} \frac{\overline{w}_{k,k+1}^{l} + \overline{w}_{k+1,k}^{l}}{2\overline{w}_{k,k+1}^{l}\overline{w}_{k+1,k}^{l}}$

Since the correlation degree is inversely mapped into the path distance, paths with the highest correlation degree can be discovered by searching the shortest paths based on the six criteria of path distance in Table 5-1.

In summary, by using Pythagorean means, we create six novel criteria to quantitatively evaluate the correlation degree of knowledge associations, and these criteria are featured in three aspects:

(1). Path distance: the six novel criteria are embodied in six different types of path distance where shorter distance corresponds to higher correlation degree of the path. Therefore, shortest path searching algorithm can be applied to search the shortest paths (knowledge associations) orderly starting from highest correlation degree.

(2). Unified standards: since an edge is essentially a simple path, these six criteria can be also used to evaluate the correlation degree of edges exactly in the same way as paths. Therefore, the six criteria are also unified standards to evaluate both the explicit associations (edges) and implicit associations (paths) under the same ways.

(3). Diversity: the six criteria actually evaluate the correlation of knowledge associations from very broad and diverse perspectives. Some of these criteria inherently involve the probability and velocity analysis. For example in Table 5-1, *GM_l* is actually similar to the analysis on probability layer that can be transformed to represent the probability of the information flowing through the path, while *HM_g* is actually equivalent to the analysis of velocity layer which represents the time cost of information flowing through the path.

5.6 Study 5: Exploring from a Design Query of a Single Concept

In this section, we will practically use the above established criteria to assist the information retrieval and knowledge discovery in a real design case study. The criteria based on probability and velocity analysis as well as the use of Pythagorean means will be applied to quantify the correlation degree of both explicit and implicit knowledge associations under the unified standards. The retrieval behaviour of each criteria will be investigated and compared. We will follow the previously proposed retrieval framework in Figure 5-2 at Section 5.2 in order to retrieve the relevant concepts and relations about the design query. In this study, the design query can be represented as a single concept and therefore the routine for exploring around a single concept will be illustrated.

Background

The case study is performed to demonstrate the use of our constructed huge semantic network combined with our established criteria in a real product design project conducted among doctoral students in the design engineering school at Imperial College. The aim of this project is to develop a household water-purifier product, which can convert seawater into direct drinking water for people living near the coast. Home-based, sustainability and suitableness for the seaside conditions are the desired requirements of the product. The design team responsible for this project consists of five Ph.D. students who have general engineering design background, but none of them has relevant experience and expertise in chemical engineering or water purification technology.

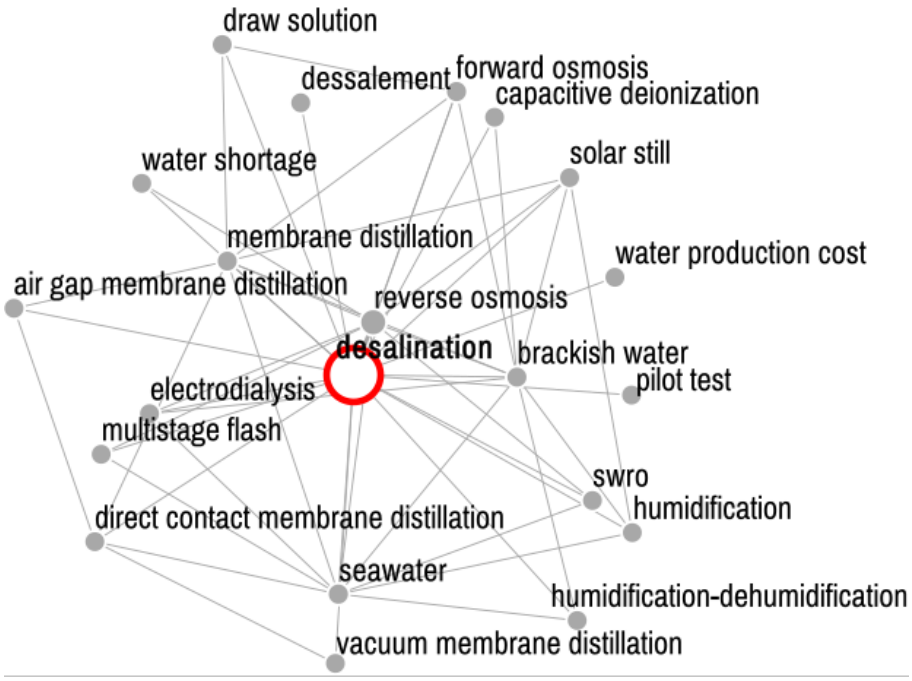
Information Retrieval Process

First Retrieval Phase

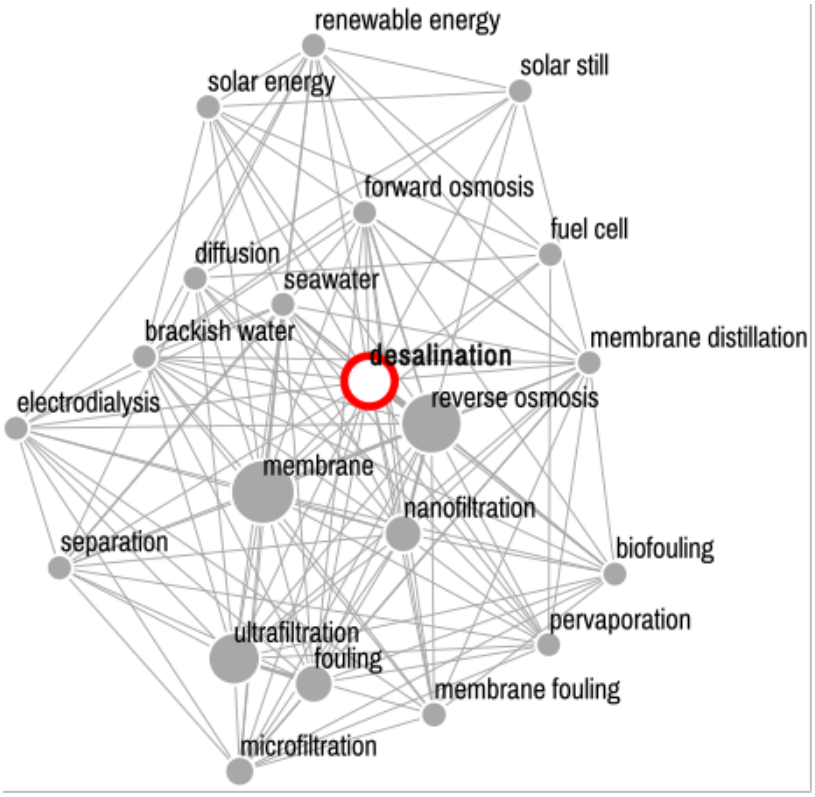
The team decided to retrieve related knowledge concepts from the semantic network to help facilitating the initial idea generation in conceptual design. They concluded and refined the design problem as one word: “desalination,” which was used as the design query input to start the first retrieval phase.

Results from the Analysis on Probability and Velocity Layers

In this first phase, the most relevant knowledge concepts to desalination were retrieved from both probability and velocity analysis of the network, respectively, as shown in Fig. 5-6, which illustrates the top 20 relevant concepts in each kind of analysis ranked by the corresponding correlation degree.



(a).



(b)

Figure 5-6 Top 20 relevant concepts from probability and velocity analysis: (a) Probability analysis ranked by correlation degree R_p and (b) velocity analysis ranked by correlation degree R_v

The corresponding most probable paths of the retrieved top 20 concepts of probability analysis as well as the fastest paths of the retrieved top 20 concepts of velocity analysis are, respectively, shown in Table 5-2. It can be seen that the results of velocity analysis show some general concepts related to desalination, like nanofiltration, solar energy, seawater, renewable energy, membrane, which are the cross-domain concepts that also relate with other areas while most knowledge concepts retrieved from probability analysis are about various and specific methods and mechanisms for use in desalination, such as reverse/forward osmosis, multistage flash, membrane distillation (MD), and Capacitive deionization.

Table 5-2 The corresponding paths starting from “desalination” to the top 20 relevant knowledge concepts in each analysis

ID	Most probable paths in probability analysis starting from “desalination”	Fastest paths in velocity analysis starting from “desalination”
1	→Reverse osmosis	→Reverse osmosis
2	→Brackish water	→Reverse osmosis → nanofiltration
3	→Forward osmosis	→Solar energy
4	→Humidification-dehumidification	→Reverse osmosis→membrane
5	→Humidification	→Forward osmosis
6	→Solar still	→Seawater
7	→Membrane distillation	→Reverse osmosis→ultrafiltration
8	→Air gap membrane distillation	→Membrane distillation
9	→Direct contact membrane distillation	→Electrodialysis
10	→Draw solution	→Reverse osmosis→ultrafiltration→microfiltration
…	…	…
18	→Multistage flash	→Reverse osmosis→membrane →pervaporation
19	→Capacitive deionization	→Solar energy→renewable energy
20	→Vacuum membrane distillation	→Reverse osmosis→ultrafiltration →membrane fouling

Results from the criteria of Pythagorean means

By using the six criteria in Table 5-1 as path distance respectively, dijkstra’s shortest path algorithm is conducted starting from “desalination” to retrieve the implicit knowledge associations in sequence from highest correlation degree to lowest. Here we mainly want to investigate the effectiveness of implicit associations, and therefore the results of explicit associations (edges) just linked to the adjacent neighbours of the query “desalination” are filtered out. The top three implicit knowledge associations with the highest correlation degrees obtained under each criteria are

presented in Table 5-3, and the corresponding graph of these knowledge associations in the semantic network is shown in Figure 5-7.

Table 5-3 The top three implicit knowledge associations start from desalination under different criteria

Criteria	1st	2nd	3rd
AM_g	→ morphology → surface structure	→ morphology → roughness	→ morphology → topography
AM_l	→ capacitive deionization → electrosorption capacity	→ ammonium bicarbonate → thermolytic solution	→ direct contact membrane distillation → pure water productivity
GM_g	→ optimization → genetic algorithm	→ adsorption → kinetics	→ membrane → fuel cell
GM_l	→ multi-stage flash → condenser tube	→ mechanical vapor compression → heat pump	→ air gap membrane distillation → pilot md plant
HM_g	→ reverse osmosis → ultrafiltration	→ solar energy → renewable energy	→ reverse osmosis → ultrafiltration → microfiltration
HM_l	→ forward osmosis → draw solution	→ electrodialysis → ion-exchange membrane	→ membrane distillation → hydrophobic membrane

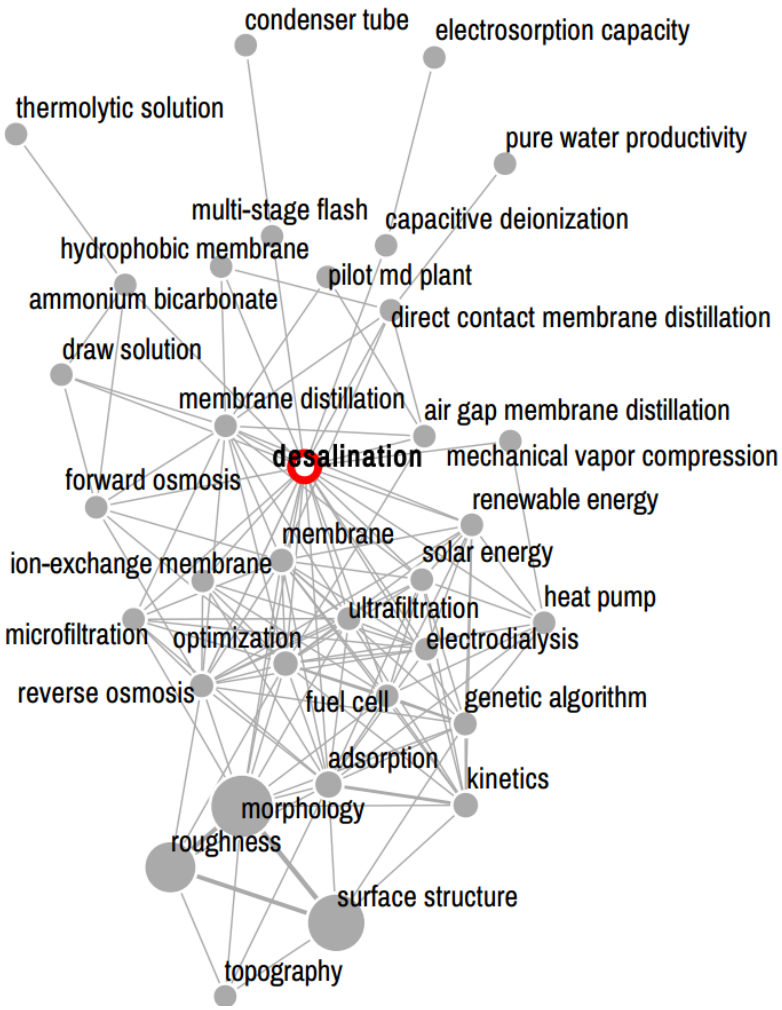


Figure 5-7 The top three implicit associations around desalination for the six criteria.

It can be found that the retrieved implicit associations facilitate the information flowing from one concept to another through multiple association steps. This process comes across different knowledge concepts in various areas, and can potentially help designers discover unexpected but useful information to provoke idea generation in conceptual design stage. For example, the implicit associations through multi-stage flash to condenser tube, through electrodialysis to ion-exchange membrane, through membrane distillation to hydrophobic membrane, and through solar energy to renewable energy, all provide feasible and rich knowledge concepts involving different design aspects from desalination mechanisms, methods, to components, materials and even power supplies.

In order to satisfy the design requirements for sustainability and household conditions of usage, the team chose “solar energy” from the velocity analysis as the potential power source for the product due to the abundant sunshine by the sea and its renewability while methods such as multistage flash and electro dialysis, which mainly depend on electricity or consume too much energy for industrial application, were excluded. After investigating all the retrieved knowledge concepts, membrane distillation was chosen and considered to be one proper solution suitable for pure solar energy-powered home appliance. Meanwhile, the results in the probability analysis happen to provide information about several specific technologies of MD including direct contact MD, air gap MD and vacuum MD, as shown in Table 5-2.

Second Retrieval Phase

In order to further explore relevant knowledge about membrane distillation, the team conducted a second retrieval phase on our semantic network using “membrane distillation” as the design query. In a similarly way, relevant knowledge concepts around membrane distillation were retrieved from both probability and velocity analysis. Results of velocity analysis this time show some general concepts similar to the previously retrieved results in the first phase like desalination, solar energy, etc. While the results of probability analysis in this second phase contain some new useful knowledge concepts including temperature polarization, hydrophobic membrane, nanofiber membrane, etc. The team found that temperature polarization is the core working principle of membrane distillation process, and hydrophobic and nanofiber membranes are commonly applied as the main purifying component in the membrane distillation device.

Third Retrieval Phase

In the same way, a third retrieval phase was conducted focusing on “hydrophobic membrane” as the design query. Interestingly again, in this third retrieval phase, all

the concepts extracted from velocity analysis are the old existing concepts such as membrane distillation, membrane, and desalination, which have already been retrieved in the previous retrieval phases. New information is obtained only from probability analysis, which provides various specific materials used for the hydrophobic membrane, including polypropylene, polyethersulfone, polytetrafluoroethylene, polyvinylidene fluoride, etc.

Figure 5-8 briefly illustrates the whole process of the three retrieval phases with respective design queries from *desalination*, *membrane distillation* to *hydrophobic membrane*, and concludes some key results from both probability and velocity analysis in each phase. Taking into consideration the costs, design requirements, and use conditions, the design team selected the potentially feasible knowledge concepts as highlighted in Fig. 5-8, and then incorporated and associated them into one design conceptual draft: membrane distillation based on temperature polarization principle is applied as the fundamental mechanism for the water purifier; solar energy is utilized as the power supply to generate temperature difference; the core functional part for distillation can implement air-gap structure with polytetrafluoroethylene hydrophobic membrane as the replaceable filtering component.

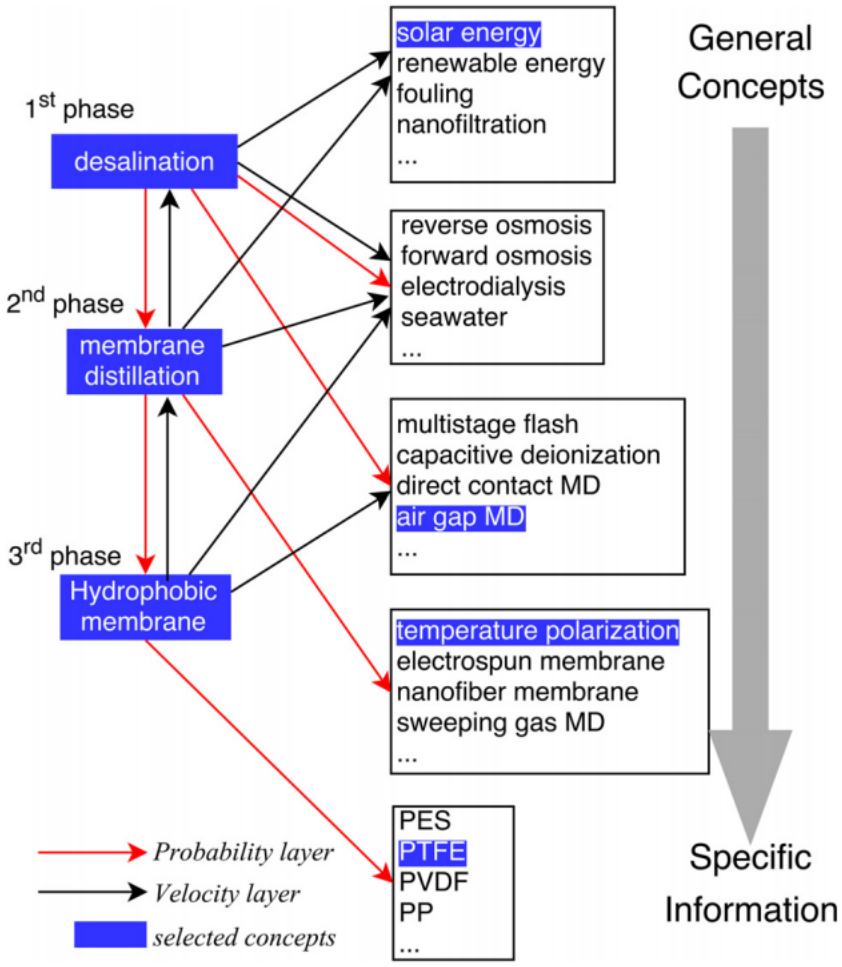


Figure 5-8 The whole three-phase retrieval process

Discussion

In the above case study, none of the team members had a background and specific expertise relating to chemical or water cleaning field. However, the use of our constructed huge semantic network enabled them to investigate and explore relevant knowledge concepts around design queries in unfamiliar engineering fields. Although our raw data are highly heterogeneous collected from various engineering fields including design, mechanical, material, energy, civil, chemistry, physics, electronic, and computing, the retrieved relations in the results show high correlation between the knowledge concepts, which indicates the suitability of the proposed method for knowledge heterogeneity. In another aspect, unlike the predefined linguistic patterns that can only extract limited and specified types of relationship,

the retrieved result of our itemset mining-based approach contains more flexible and diverse knowledge relationships between two concepts in perspectives from, for example, mechanisms, methods, components, materials, and even power supplies, which can hardly be extracted by simply matching with predefined linguistic patterns. Benefiting from the network structure, the retrieval process can be conducted recursively that the result explored from previous phase can be used as the design query in the next phase, which enables the multiphase retrieval process. The retrieved knowledge concepts are shown to be useful and have high correlation to help the team achieve a design solution.

Design process is knowledge-intensive, but it's usually impossible for the design team to have expertise in every knowledge domain. The use of our constructed semantic network coupled with the proposed retrieval framework and the unified criteria for quantifying correlation degree of knowledge associations can effectively help the design team to conduct design information retrieval in other domains which are possibly beyond their expertise. The aim is to satisfy the variety of knowledge needs during real-world design activities.

Retrieval behaviour

The three-phase retrieval process shown in Fig. 5-8 automatically develops a concept structure, in which comprehensive and general concepts such as solar energy and desalination are at the top, while subsequent detailed and specific information of this field such as distillation technology and hydrophobic materials are located at the bottom of the structure. It can be found that in each retrieval phase, results from probability analysis tend to get knowledge in the lower specific level of the concept structure while results of velocity analysis tend to be in the upper general level.

To further investigate the difference between probability and velocity analysis, we evaluate the node strength of the retrieved paths corresponding to the top 20

concepts of each analysis. Figure 5-9 shows the results for each of the three retrieval phases. The vertical axis is the strength of the node. Horizontal axis is the position of the node in the path, where zeroth position means the start point of the path. We can see that the result is consistent with the conclusion in Sec. 4.5 that the broad concept (desalination) has the higher node strength while the specific element (hydrophobic membrane) has lower node strength. Besides, the retrieval behaviours of probability and velocity analysis are also shown to be consistent with Figure 5-5. The probability analysis tends to retrieve low strength nodes of specific concepts through broad and short paths, which is similar to breadth first search, while the velocity analysis usually searches for general concepts with high node strength through deep and long paths like depth first search.

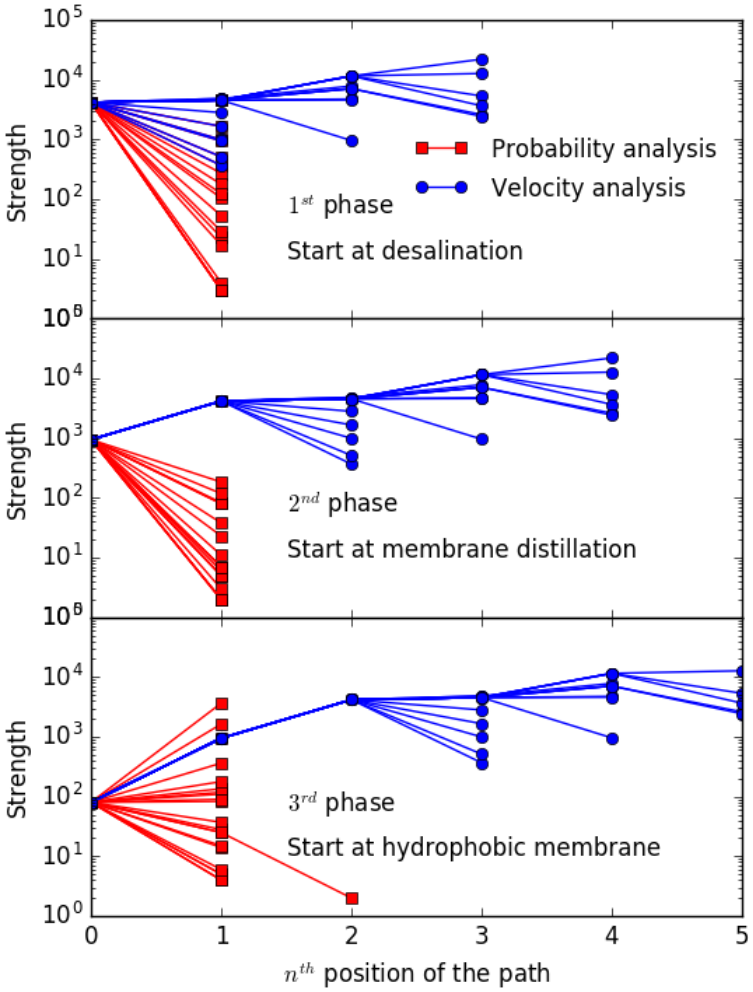


Figure 5-9 Node strength of retrieved paths in the three retrieval phases

Therefore, the different retrieval behaviours of the two distinct analyses concluded above can yield different types of knowledge concepts. For the example of Figure 4-8 and concept navigation process in Figure 5-8, probability analysis usually explores more specific level below the design query in the top-down concept structure to retrieve details within the domain of the query concept. On the contrary, velocity analysis, starting from the query concept, searches upward through the concept structure for more general knowledge concepts that may cross to other different but relevant domains. In comparison, given *desalination* as the design query, the probability analysis will output specific methods such as *multistage fla*sh and

membrane distillation while the velocity analysis leads to some broad concepts such as *solar energy*, *renewable energy*, *microfiltration*, which may be used in a variety of engineering processes other than water purifying. Similarly, taking *temperature measurement* as the query, probability analysis retrieves the two specific instruments *thermocouple* and *pyrometer* for measuring temperature while the results from velocity analysis may come across to other different but relevant domains such as *sensor*, *heat transfer*, and *friction*.

Hence, probability analysis can explore rich domain-specific knowledge particular in the focusing area, which provides a solid background in professional knowledge and helps the designers to develop domain know-how solutions. The velocity analysis, by retrieving general and comprehensive concepts, can guide the designers to the upper level of the concept structure with a wider scope on other related areas, which enables the designers to utilize and bridge other knowledge domains with current problem domain for further improving the idea generation process. In the practical sense, the features of the two methods are complementary, and a combination of both could potentially help satisfy the diverse knowledge demands during engineering design activities.

In addition to the retrieval behaviours of probability and velocity analysis, we can also see that the six criteria established by Pythagorean means are also showing diverse and distinct retrieval behaviours with different kinds of implicit knowledge associations. Table 5-3 illustrates the different information retrieval behaviours of each of the criteria. The retrieved results of *HM_l, HM_g* and *GM_l* show more relevance to the original query desalination than *GM_g, AM_g and AM_l*. For example, the concepts (e.g. reverse osmosis, membrane distillation, multi-stage flash) obtained in *HM* and *GM_l* are more helpful and related to the design problem than the concepts (e.g. morphology, optimization, ammonium bicarbonate) extracted by *AM* and *GM_g*. One reason is that to yield a large Harmonic Mean (*HM*) essentially has a higher requirement on the closeness between the addends than to

yield a large Arithmetic Mean (*AM*) or Geometric Mean (*GM*). Another reason is that local fluctuation normalisation tends to extract more domain-specific concepts than feature scaling.

From the perspective of a knowledge specific level, we can find in Table 5-3 that the criteria using feature scaling normalization, namely *AM_g*, *GM_g* and *HM_g*, tend to retrieve general knowledge concepts such as morphology, optimization, and reverse osmosis, compared to the criteria normalized by local fluctuation (*AM_l*, *GM_l*, *HM_l*) which tend to get domain-specific knowledge concepts such as capacitive deionization, air gap membrane distillation, and hydrophobic membrane. This can also be illustrated in Figure 5-10 which evaluates the node strength of the top 20 retrieved implicit knowledge associations for each criteria, where the strength of a node means the sum of raw weights of all the edges incident to that node. The vertical axis of Figure 5-10 shows the strength of the node, and horizontal axis is the position of the node in the path. The node at 0^{th} position represents the design query *desalination* which is the start point of the knowledge association. A very general concept usually has extensive connections with many other concepts in various engineering areas, and therefore yields a higher strength of the corresponding node. While, a specific concept is only linked to limited knowledge within its own domain, resulting in a lower node strength. Thus, we can see from Figure 5-10 that results retrieved by criteria of feature scaling normalization (*AM_g*, *GM_g*, *HM_g*) contain general concepts with higher node strength, while the criteria using local fluctuation normalization (*AM_l, GM_l*, *HM_l*) tend to retrieve specific concepts with lower node strength. The Harmonic Mean (*HM*) can help extracting longer paths with more association steps than *AM* and *GM*.

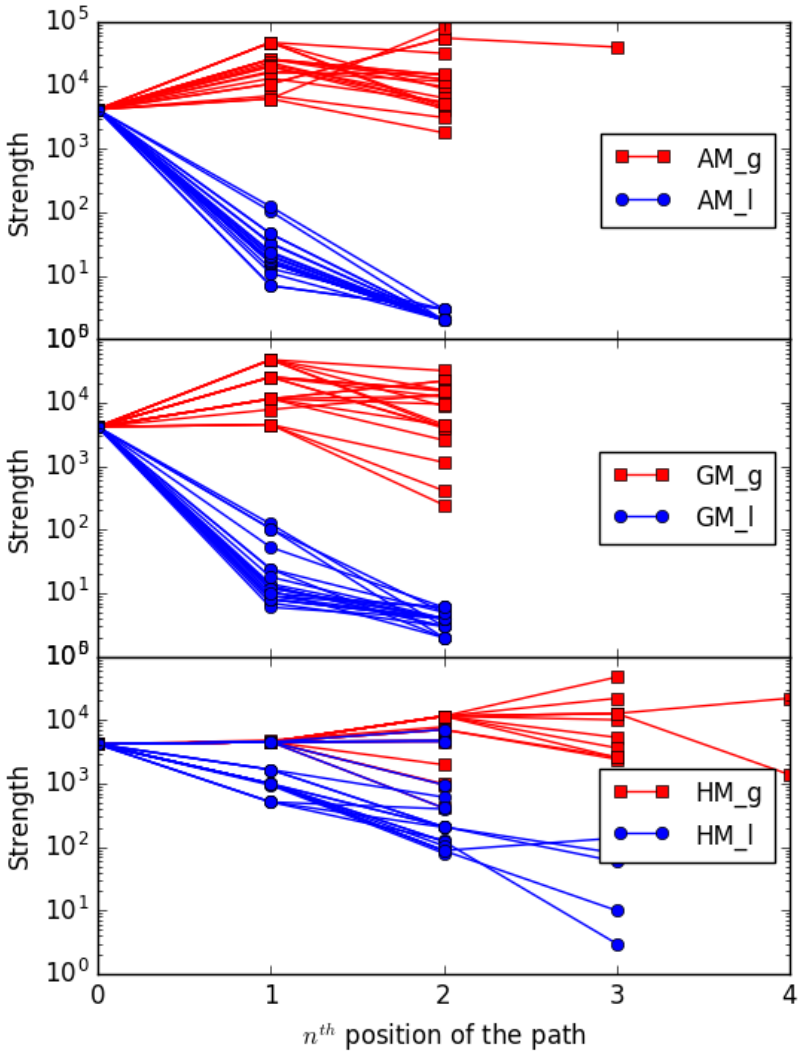


Figure 5-10 Node strength of the top 20 implicit associations around desalination for each criteria

5.7 Study 6: Search Paths between Two Concepts of a Design Query

In contrast with Study 5 looking for knowledge associations around the design query of a single concept, the Study of this section aims to explore the knowledge associations between two distant knowledge concepts since the design query need to be represented by more than one single concept. It is believed that the bridging

concepts and associations between two knowledge domains can usually inspire innovative insights and provoke creativity exploration (Koestler, 1964).

The initial purpose of this Study is to find the potential collaboration opportunities between design engineering and civil engineering departments to promote interdisciplinary cooperation for new projects. Therefore, the retrieval framework was used to search the possible knowledge associations between design engineering and civil engineering in order to discover the potential opportunities for innovation and generate creative ideas by bridging two engineering domains.

Since *design engineering* is a very broad area, we can firstly explore relevant sub-areas of design engineering in the semantic network, and then select one of the promising sub-areas as a start point to discover implicit associations with civil engineering. After a brief exploration, relevant knowledge concepts such as *creativity*, *ergonomics*, *product development*, and *robotics* are retrieved around *design* as shown in Figure 5-11.

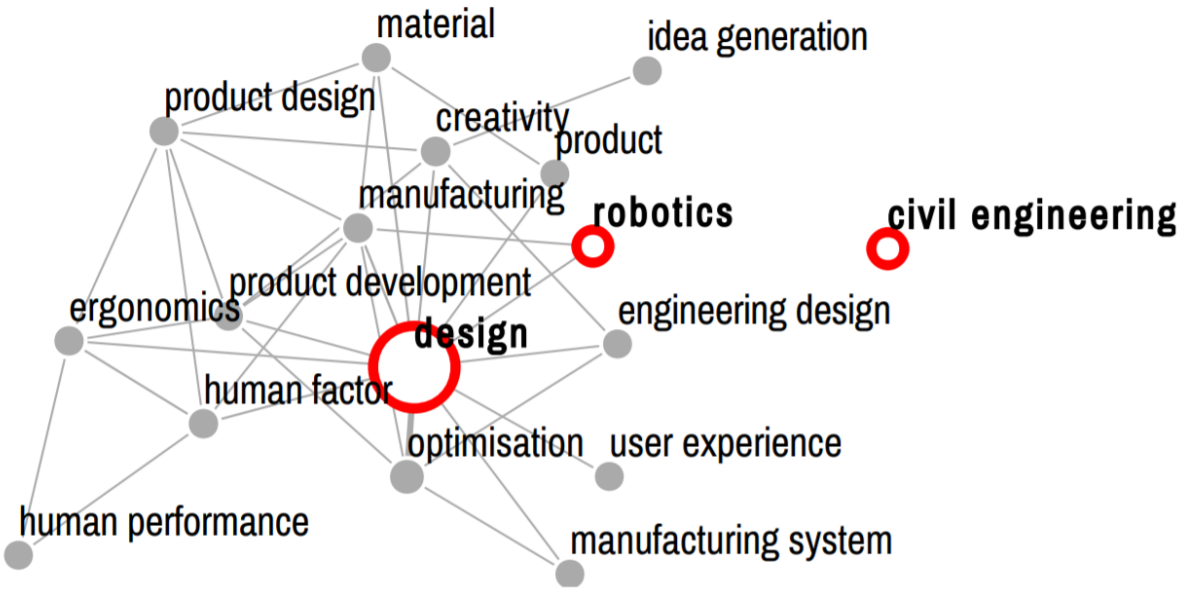


Figure 5-11 Sub-areas of design engineering to be considered

We select, for example *robotics*, from Figure 5-11 as one interesting sub-area to provoke creative ideas with *civil engineering*. Thus, the implicit knowledge

associations between *robotics* and *civil engineering* can be discovered in order from highest correlation degree to lowest, by conducting shortest paths searching between the two nodes, where the six criteria established by Pythagorean means in Table 5-1 are used as the path distance. Each discovered implicit knowledge association, after being evaluated based on common sense and expertise, is used to inspire innovative insights about the potential applications of *robotics* in *civil engineering* area. For each criteria, Table 5-4 shows some examples of the discovered high-correlated implicit associations, and the corresponding inspired ideas and comments. The graph representing these association paths in the semantic network is shown in Figure 5-12.

Table 5-4 Examples of the discovered high-correlated implicit knowledge associations and corresponding comments and ideas under each criteria

Criteria	Implicit associations	Comments & Creative ideas about "How To Use *Robotics* in *Civil Engineering*?"
AM_g	**robotics** → construction → **civil engineering**.	Automatic construction, e.g. modular building.
	robotics → machine learning → classification → **civil engineering**.	Machine learning in classification of types of construction.
AM_l	**robotics** → arthroplasty → cement → **civil engineering**.	Cement building components like jointing bones.
	robotics → automation → pavement → **civil engineering**.	Automatically construct pavement.
GM_g	**robotics** → surgery → fracture → concrete → **civil engineering**.	Detection and repair of concrete cracks.
	robotics → neural network → compressive strength → concrete → **civil engineering**.	Prediction of concrete strength by neural networks.
GM_l	**robotics** → automation → crack sealing → pavement → **civil engineering**.	Automatic crack sealing for pavement.
	robotics → remote handling → metrology → **civil engineering**.	Measurement conducted by robots at high-risk locations.

HM_g	**robotics** → laparoscopy → kidney → liver → oxidative stress → hydrogen peroxide → oxidation → coating → corrosion → concrete → non-destructive testing → **civil engineering**.	An interesting path across multiple domains connecting robotics, medicine area, chemical phenomenon, and civil engineering sequentially into a coherent chunk.
HM_l	**robotics** → path planning → construction automation → earthwork → road construction → pavement → **civil engineering**.	Automatic road planning.

From Table 5-4, it can be seen that the discovered implicit associations link *robotics* and *civil engineering* through diverse intermediate bridging concepts in various domains. The bridging nodes may be general knowledge concepts between the two domains such as construction, machine learning and classification, which indicate the potential use of robotics technology to construct a building, and the use of machine learning to classify the types of construction. The bridging concepts can also be specific subjects linking the two knowledge domains. For example, intermediate concepts like automation, pavement, crack sealing and road construction can provoke ideas that robotics techniques could be applied for automatic pavement construction and crack maintenance. Moreover, the implicit associations may possibly link through surprising and irrelevant fields such as the area of medicine (e.g. surgery, arthroplasty, laparoscopy, fracture), in which case, the application of robotics in medical surgery such as fracture and joint operation can inspire the idea of using microrobot techniques for concrete fracture cementing and reinforcing. From the above examples, we can see that the implicit knowledge associations, which are actually based on a quasi-relevance and serendipity manner, can discover innovative bridging concepts connecting the two domains to significantly provoke conceptual idea generation.

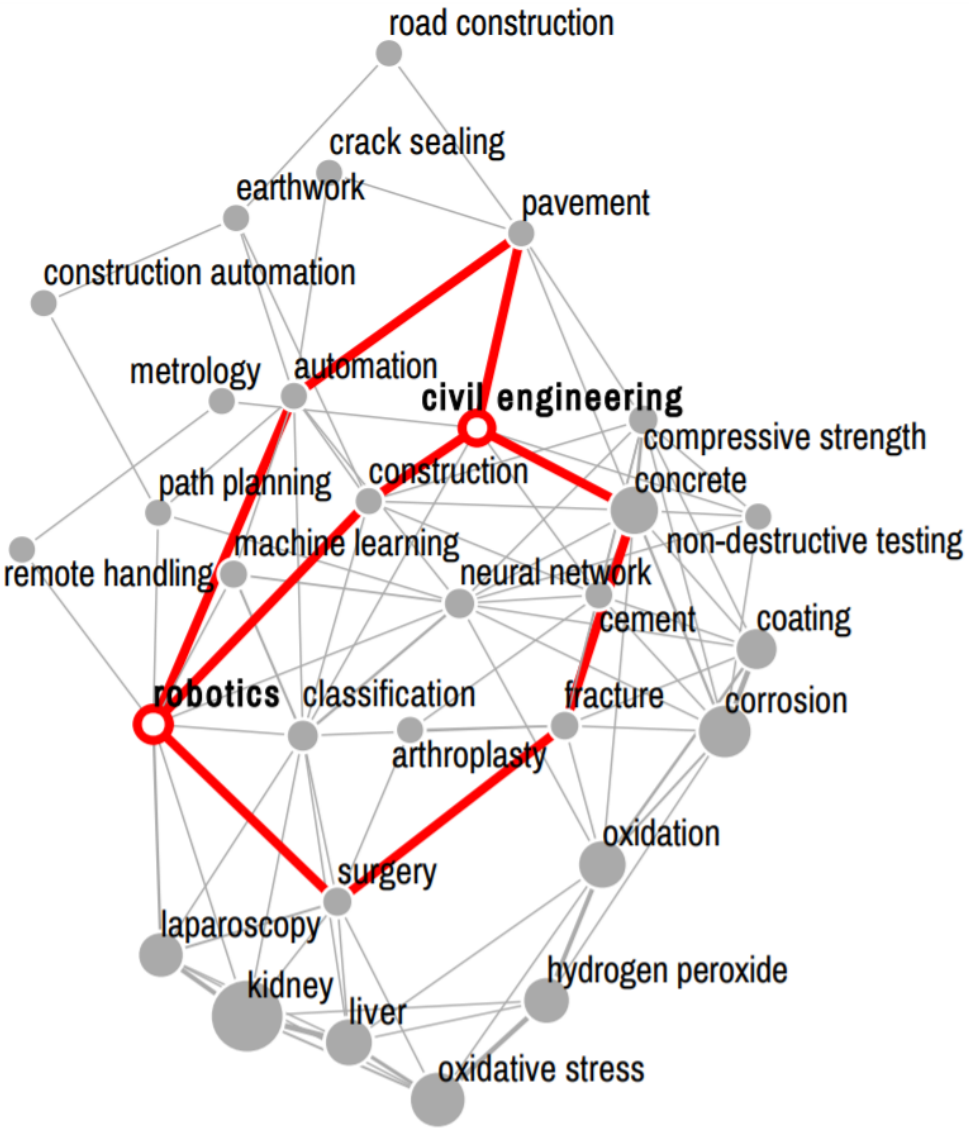


Figure 5-12 Graph representation of the examples of implicit associations with high correlation degree between robotics and civil engineering

In order to investigate the retrieval behaviours of different criteria, the node strength of the top 20 discovered implicit associations between *robotics* and *civil engineering* are evaluated for each criteria, as shown in Figure 5-13. Similarly, the vertical axis means the strength of the node, and the horizontal axis shows the relative position of a node in the path where 0.0 and 1.0 mean the start and end point of the path representing *robotics* and *civil engineering* respectively. The result of Fig. 5-13 shows consistent retrieval behaviours with Figure 5-10. AM_g, GM_g and HM_g tend to find paths containing general concepts with high node strength, while AM_l,

GM_l and HM_l discover paths of specific knowledge with low node strength. This is mainly due to the different perspectives of the two weight normalization techniques where feature scaling and local fluctuation are respectively based on global view and local view. Besides, the quantity of edges contained in a path n is shown to increase in the results from AM to HM. Also as illustrated in Table 5-4, AM tends to retrieve short implicit association across less edges focusing on relevant knowledge within the same domain, while HM tends to retrieve long implicit association with more edges across multiple distant domains (e.g. medicine, chemical area), and GM shows a balance between them. This kind of behaviours is achieved by the different statistical natures of the three Pythagorean means.

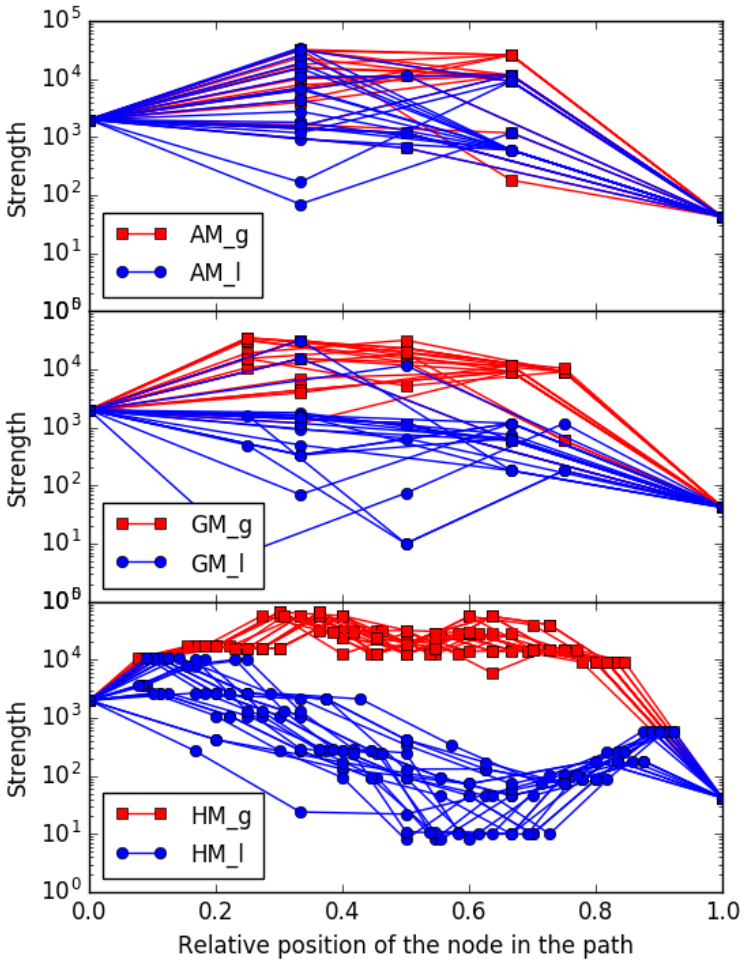


Figure 5-13 Node strength of the top 20 implicit associations between robotics and civil engineering for each criteria.

5.8 Conclusion

This chapter focuses on the question of how to use the constructed semantic network to support design information retrieval.

We proposed a design information retrieval framework based on our constructed semantic network, in which the major use of the semantic network is to retrieve useful knowledge concepts and relations relevant to the design query and needs. However, how to define "relevant" is not a binary problem of Yes or No, 0 or 1. The correlation between concepts may be better to be quantitatively evaluated rather than using an absolute Yes-relevant or No-irrelevant. Traditional ontology analysis methods only consider the adjacent neighbours of a concept as the relevant concepts, and ignore all other concepts out of the neighbour scope as well as the essential network structure of ontology. While in our research, we innovatively define the association between directly linked concepts as explicit knowledge associations, and the association between indirectly linked concepts through intermediate nodes as implicit knowledge associations. We consider both explicit and implicit knowledge associations as potentially relevant concepts and relations, and take advantages of the network structure of our semantic network.

Therefore, instead of judging the relevance by only looking at the directly linked concepts (explicit associations), we quantitatively evaluate the correlation between any two concepts regardless of whether they are directly or indirectly linked. This means that we actually consider both explicit and implicit knowledge associations under the same standards to be retrieved and ranked. We developed two sets of quantitative criteria in order to evaluate the correlation degree of explicit and implicit knowledge associations under the unified standards. One set of criteria is to model the probability and velocity layers of the semantic network, and uses the probability of the most probable path and the time cost of the fastest path to evaluate the correlation degree between any two concepts. The other set of criteria, which

extends the probability and velocity analysis, establishes six novel criteria by applying Pythagorean means on assessing the path distance between any nodes no matter they are directly or indirectly linked.

The criteria were applied in real design case studies to retrieve both explicit and implicit knowledge associations either around a single concept or between two concepts of the design queries. The results show that the knowledge concepts and relations, which are retrieved from our constructed semantic network through the proposed framework and quantitative criteria, can significantly support the design activities and assist with the idea generation at conceptual stage. More importantly, different criteria show significantly different retrieval behaviours. Some criteria tend to retrieve domain-specific knowledge concepts while some tend to explore general knowledge concepts. Some tend to search short and broad associations while some tend to search for deep and longer associations. These different retrieval behaviours of the different criteria indicate the great potential of our constructed semantic network and proposed retrieval framework in satisfying the different and various knowledge demands during engineering and design activity.

Chapter 6 Data Visualisation and User Interaction

The previous chapters have described the construction of the elements for a data-driven pipeline from data acquisition (web crawling), text mining (unstructured to structured), and to a semantic network analysis for supporting design information retrieval by providing useful design relations and concepts related to the design query and need. The final missing block of this data pipeline is the ability to visualise our constructed huge semantic network and more importantly to let a user interact with this semantic network.

In chapter 5, we have already proposed an information retrieval framework based on unified quantitative criteria of correlation degree to provide both explicit and implicit knowledge associations either around a single knowledge concept or between two concepts of the design query. This framework is proved by Study4, Study5 and Study6 to satisfy the different and various kinds of knowledge demands during engineering design activities. However, all these analyses conducted on the semantic network and the information retrieval process are performed on the back-end. It is necessary to provide some mechanisms to bring the data into the front-end so that the users can intuitively visualise the data, interact with the data, and more importantly evaluate the data and retrieved information from a more diverse perspectives and comparative ways.

Therefore, in this chapter, a data visualisation and user interaction web platform called 'B-Link' is developed and released on the Internet. This platform inherently

integrates the constructed huge semantic network (in Chapter4) with the developed semantic network analysis methods and unified correlation degree criteria established in Chapter5. This means that it allows the users to practically manipulate our constructed semantic network through various sophisticated analysis algorithms to retrieve and readily visualise the related information in an interactive way. Based on B-Link, we have also conducted an online test of user study to evaluate and compare our retrieved knowledge associations with other existing public ontology databases from a variety of aspects.

Some of the work described in this chapter has contributed to the work previously published in (Chen et al., 2017, Chen et al., 2018):

1. Chen, L., Shi, F., Han, J. & Childs, P. R. 2017. A network-based computational model for creative knowledge discovery bridging human-computer interaction and data mining. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, V007T06A001-V007T06A001. Copyright © ASME 2017. Reprinted by permission of ASME.
2. Chen, L., Wang, P., Shi, F., Han, J. & Childs, P. 2018. A COMPUTATIONAL APPROACH FOR COMBINATIONAL CREATIVITY IN DESIGN. DS92: Proceedings of the DESIGN 2018 15th International Design Conference. 1815-1824. Copyright © Design Society 2018. Reprinted by permission of Design Society.

6.1 Data Visualisation Techniques and User Interaction

Data visualisation

Data visualisation is to generate visual representation of the data, where the data are usually pre-processed and structured in a certain way. It can abstract information in some schematic form, including attributes, relations or variables for the units of information so that the information can be directly captured and conveyed throughout the visual representation (Friendly, 2008). From another perspective, data visualisation inherently involves the process to convert data into information where data are organized, processed and represented in a structured way to communicate meaningful information (Shneiderman, 1996). This literally means that the raw data cannot just be directly visualised since the result will be meaningless. Procedures such as categorizing, classifications, pre-analysis are necessary to be conducted before the actual visualisation. That's why data visualisation is positioned at the end of our data pipeline.

With the growing amount of data and the relations between data, suitable data visualisation techniques are becoming increasingly important. An abundant amount of opportunities exist to discover new knowledge and gain insights from the massive quantity of accessible data. One well-known approach is through data mining which is based on statistical approaches to discover the patterns, trends and insights by following a quantitative analysis manner. However, another useful approach is actually through the data visualisation to directly obtain insights. Instead of using the quantitative analysis to do inference in data mining, data visualisation on the other hand utilise and leverage the human intuition and instinct visual system to directly detect the patterns, spot trends and identify new knowledge (Heer et al., 2010).

Data visualisation has several advantages for supporting the representation of the information, because it's **concise**, **structural**, and **intuitive** (Keim, 2002).

- The first significant feature of data visualisation is its conciseness. A single visual graph can often capture the richest information of the facts, which otherwise may need to be described by a large amount of textual paragraphs.

For example, a histogram illustrates the numerical distribution of a variable over a continuous range divided by consecutive and non-overlapping intervals (Sereda et al., 2006). However, such task is difficult and even impossible to be completed by text descriptions. Especially when there are hundreds of bins (intervals), we cannot naively describe the value of each intervals by using texts.

- The second advantage is that data visualisation can represent information in a structural and relational manner. The histogram groups data into different bins/intervals, and the bins are organized and positioned in sequence from the lowest intervals on the leftmost to the highest intervals on the rightmost.
- Finally and most importantly, data visualisation can provide intuitive insight and knowledge. Human can leverage their visual system to intuitively discover patterns and identify new points. Still taking a histogram as an example, we can instantly understand the overall distribution of the variables at the very first glance, and directly find out the areas of the highest and lowest density distribution. As an example Fermat's last theorem was solved by Andrew Wiles (Wiles, 1995) by transforming the problem from numerical space into geometric space to push up the knowledge boundary. Therefore, graphic data visualisation can significant assist human with the knowledge discovery in an intuitive way.

Based on the above advantages features, we consider data visualisation to be the last and very necessary component in the pipeline of our data-driven approach.

Visualisation Techniques

With the arrival of the big data economy, a lot of novel visualisation techniques have been developed for sophisticated related data in multi-dimensions and large scale rather than the traditional 2D, or 3D dimensions in a small scales (Card et al., 1999, Ware, 2012, Spence, 2001, Schumann and Müller, 2013). All these techniques can be

classified from two perspectives, namely, the type of data structure to be visualised, and the visualisation techniques (Keim, 2002).

The types of data structure can be diverse and may involve:

- 2-dimensional data, such as cartogram, Choropleth, geographical maps (Stolte et al., 2002, Abello and Korn, 2002),
- Temporal data, which is similar to linear visualisation but it depends on time series usually with a start and finish point,
- Multidimensional data, such as relational tables (Stolte et al., 2002, Kreuseler et al., 2000), Pie charts, Histograms, Scatter Plots
- Text and hypertext including news articles and web documents (Havre et al., 2002). This is a rare data type to be directly visualised where the texts often need to be abstracted and compressed.
- Hierarchical data, such as Dendrogram, Ring Chart, Tree Diagram, which categories data into different groups from a top-down structure.
- Network and Graph Data, e.g. hyperlinked web documents, telephone calls (Abello and Korn, 2002, Kreuseler et al., 2000)
- Algorithms and software, as such flowcharts and operation diagrams (Stolte et al., 2002)

The traditional visualisation techniques used for the data type described above can be categorised as follows (Keim, 2002):

- Standard two or three dimensional display, for example bar charts and x-y plots
- Display based on stacks, such as treemaps and multi-dimensional stacking (Shneiderman, 1992, Johnson and Shneiderman, 1991, Ward, 1994)
- Icon-based displays including star icons, circle icons, square icons, needle icons (Abello and Korn, 2002)

- Dense pixel displays. This involves the use of the graph sketches, segmenting techniques as well as the recursive patterns (Keim, 2000, Abello and Korn, 2002)
- Geometrically transformed displays. This refers to different parallel and vertical coordinates in multi-dimensional space used for landscapes (Kreuseler et al., 2000)

The above mentioned techniques were developed in the last two decades to satisfy the ever-increasing demands for data visualisation tasks due to the exponentially growing amount of human-machine generated data. However, these techniques still have some limitations. Firstly, they don't have flexibility to change their dimensionality. For a particular visualisation technique, they usually only fit with a specific data type with a fixed dimensionality, and are not adaptable for data types with a different dimensionality. Secondly, most of these techniques only generate static visual display, and are not able to create animation or dynamic features which are also important in modern data visualisation and can potentially convey richer information. Finally, these techniques are usually not interactive that users are not able to operate and manipulate with the data.

In the recent years, there is an emerging usage of an approach called data-driven documents (D3), which designs, manipulates and represents the document elements by exclusively using data driven approach. It visualises the data using HTML, SVG and CSS in universal standards without tying the user to a stationary framework (Bostock, 2018). Therefore in D3 approach, document elements are exclusively driven and manipulated by the underlying data, while in turn, data are visualised through the representation of document elements at the meantime. This is actually a complementary process. Figure 6-1 shows some sample techniques in D3.js.

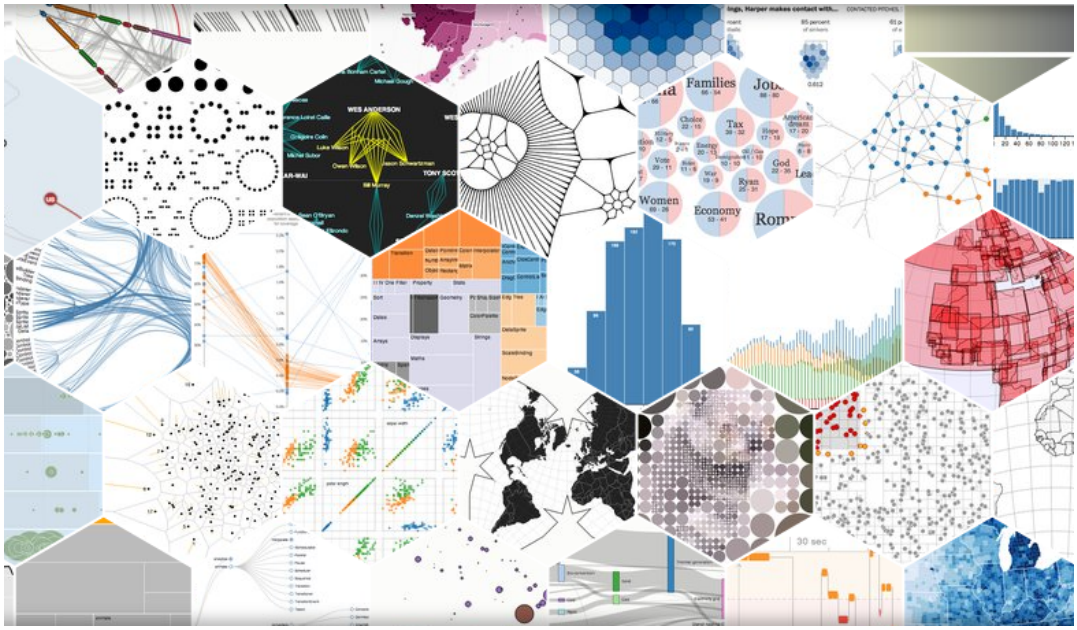


Figure 6-1 Samples gathered from D3 approaches adapted from (Bostock, 2018)

D3 approach is the state-of-art cutting edge technique for both data visualisation and user interaction. It overcomes the limitation of the previous techniques developed in the last decades. Since D3 is completely data-driven document that all the elements and styles, literately everything displayed on the document are controlled and defined by the underlying data, therefore the dimensionality of the data to be visualised can be highly flexible and can be increased or reduced according to the user's needs. Second, D3 approaches can use animation and dynamic features to provide extra information during the visualisation process. Finally and most importantly, D3 approach enables the user interaction with the visualised data so that user events such as click, scroll, drag and query can trigger back-end data analysis to generate new data layout for satisfying the user needs.

6.2 B-Link

As discussed above, D3 approach is one of the recent most advanced data visualisation and interaction techniques, that has flexibility for data dimensionality, incorporates animation and dynamic features, and provides interaction mechanism.

Therefore, in this chapter, we will mainly rely on D3.js technique to build our front-end platform to visualise and interact with our constructed huge semantic network through the proposed retrieval framework. The platform we developed is called 'B-Link', which mainly integrates two components: data visualisation and user interaction. The data visualisation component is to extract and visualise the relevant parts or portions of our constructed semantic network based on the user's query. The user interaction component provides functionalities to allow the users to query and manipulate with our semantic network through the network analysis methods and retrieval framework proposed in Chapter 5.

Data Visualisation

Since the fundamental data structure of our semantic network is essentially a graph, we apply the Force-Directed Graph in D3.js to layout our network data. In Force-Directed graph, each knowledge concept is a node and the relation between two concepts is represented by an edge between the nodes. Physical simulation of multiple forces is applied to characterize the dynamic placement and movement of these nodes and edges. Specifically, three kinds of forces are simulated on the graph. Firstly, many-body force applies mutually amongst all the nodes. It can simulate attraction (gravity) among the nodes if the strength is set to be positive, or electrostatic charge (repulsion) if the strength is negative. Since we want the knowledge concepts to be clearly displayed rather than messed up, we apply a small amount of electrostatic charge to simulate repulsion effect between nodes. Moreover, spring force is applied to simulate the links between nodes that places related characters in closer proximity, while unrelated characters are farther apart. This spring force pushes linked nodes together or apart according to the desired link distance. The strength of the force is proportional to the difference between the linked nodes' distance and the desired distance, similar to a spring. Besides, we also applied a centre force translating nodes uniformly so that the centre of mass of all nodes (if all nodes have equal weight) is at the centre of the viewport. Figure 6-2

shows an example of the force-directed layout on a portion of our semantic network data.

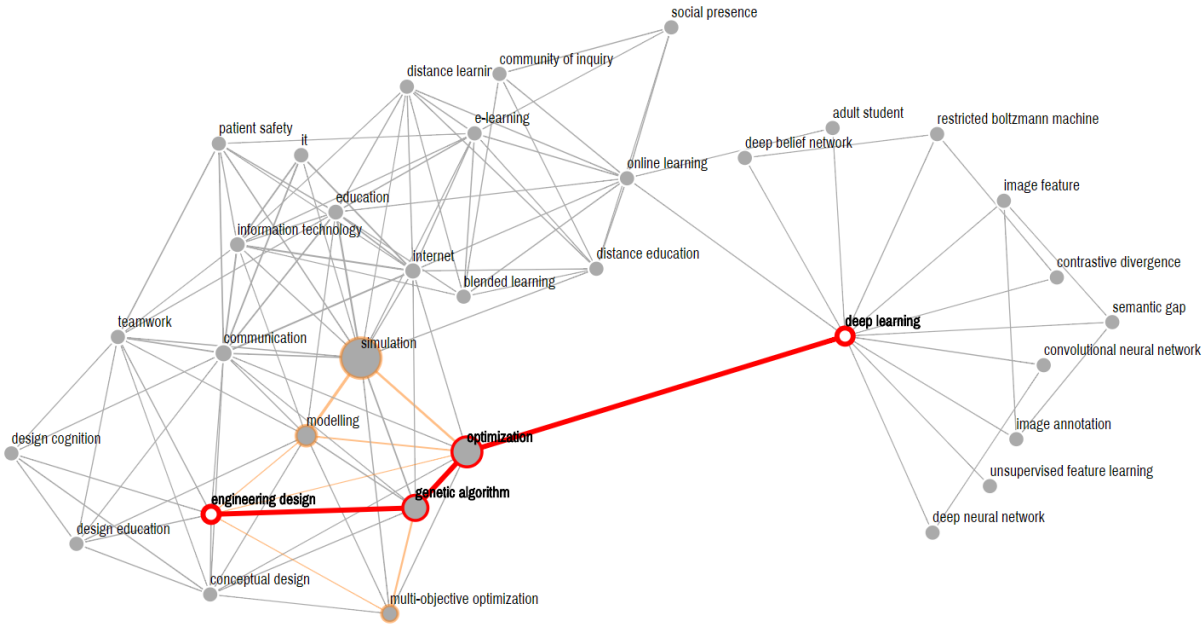


Figure 6-2 An example of force-directed graph

As shown in Figure 6-2, there are many graphic features that we can utilise to represent the rich dimensionalities of the data. In our case, we can have totally six dimensionalities/features to communicate different types of information to the user:

- **Text label**: a text label can be associated with each node representing the knowledge concept
- **Node size**: the size/radius of the node can indicate the node strength which means the sum of raw weights of all edges incident to that node in the network. As discussed in Chapter 4.5 (Study 3), the strength of the node corresponds to the specific level of the knowledge concept. General concepts usually have higher node strength and therefore obtain large node size in the visualisation, such as 'simulation' in Figure 6-2, while on the other hand, specific concepts with lower node strength are visualised to be node with smaller size/radius in the graph such as 'convolutional neural network' in Figure 6-2.

- **Link strength**: the link strength is based on the Hooke's law:

$$F = -kx \tag{6-1}$$

where x is the length of extension/compression, and k is the Hooke's constant. Hence, we can set different k for different edges according to the relevance of the edge. We assign higher k to edge with higher weights, so that more relevant nodes will have a stronger link and more stable distance between them than less related nodes.

- **Link distance**: Similarly, we can use the target link distance (the original length of spring) to correspond with the relevancy degree of the edge. We set shorter target link distance for more relevant edges, and longer distance for less related edges. Specifically, the six criteria in Table 5-1 can be used directly to represent the link distance.
- **Link width:** The width of link can also be used as an additional feature to represent the correlation degree of the link between two nodes. We use thicker link to represent high weighted associations while thinner link for low weighted associations.
- **Colour:** The colour of nodes and edges can be used to highlight the selected or focused concepts and relations under the user queries during interaction.

Therefore, the above six features in visualisation can provide great flexibility to represent different dimensionalities of our data. Rich information can be captured and contained within a single graph, which can be intuitively discovered by the user.

In addition to the flexibility of data dimensionality discussed above, ours B-Link platform also incorporates the animation and dynamic feature in the visualisation process. All the dynamic movements including position, velocity and accelerated velocity are computed based on the physical simulation of the inherent applied forces. This dynamic physical simulation makes the visualisation more realistic and much easier to get insights. For example, the closely related nodes/concepts will be pulled together and irrelevant nodes are pushed apart automatically according to the

link strength and distance defined in the simulation, therefore relevant concepts can be automatically grouped together to form various clusters/communities being representations of different knowledge domains. All of these can be simply accomplished through the pure physical simulation without applying any sophisticated clustering algorithms.

User Interaction

B-Link platform also supports user interaction functionality and behaviours. It enables users to practically retrieve, query and browse a semantic network by following the retrieval framework and quantitative correlation degree criteria proposed in Chapter 5. The user interaction involves two parts: browsing interaction which does not trigger the back-end analysis, and retrieval interaction which does trigger the back-end data analysis.

The **browsing interaction** is mainly used to facilitate user with the information visualising and browsing process. This includes operations such as translating, zooming, dragging, clicking and highlighting on the graph. User can move and zoom in and out of the viewport to either look into specific part or have an overview on the whole structure of the graph. They can freely drag the node in the graph to observe the full details of the links and relations around the node, and can also simply click on the node to highlight and select the corresponding concepts for subsequent retrieval interaction and analysis. Besides these simple operations, we also provide an information panel to view the retrieval results in form of list as shown in Figure 6-3, in addition to the force-directed network graph as shown in Figure 6-2. This panel lists and ranks the retrieved knowledge associations in order from the highest correlation degree to lowest, which just cannot be illustrated by the network graph structure. However, on the other hand as previously discussed, network graph structure can intuitively show groups of closely related concepts as clusters, and also demonstrate the specific level (node strength) of concepts by using the node

radius/size, which in turn cannot be achieved in information panel. Therefore, the information panel and force-directed graph are complementary, and a combination of them can help provide the user with the richest possible information.

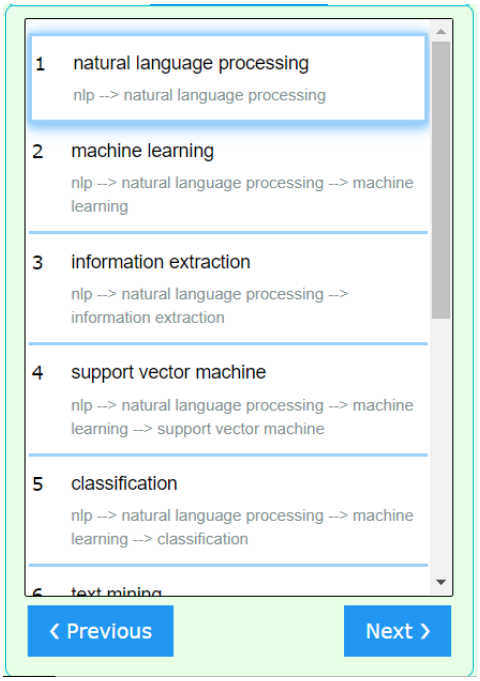


Figure 6-3 Information panel to view results in form of list

For **retrieval interaction**, we incorporate the network analysis techniques and retrieval framework proposed in Chapter 5 in order to enable user to query information from our back-end semantic network. We create a function panel which provides three query methods corresponding with the retrieval framework in Chapter 5, namely, *Explore*, *Search Path*, and *an additional Cluster*.

Explore function is used to explore and retrieve the relations around a single knowledge concept similar to Study 5. It uses the single source Dijkstra's shortest path algorithm which starts from the source query to retrieve all the reachable nodes in order from the shortest distance. The unified criteria for correlation degree developed in Chapter 5 are used to configure the distance of individual edge. Specifically, since we want to retrieve both general and specific knowledge concepts related to the query, we apply two different criteria with distinct retrieval behaviours

in this explore panel. One is *HM_n* for general concepts, and the other is *GM_l* for specific concepts. The equations and retrieval behaviours of this two quantitative criteria can be referred in Chapter 5 and Table 5-1. In addition, a minimum step function is provided on the Explore panel, where knowledge associations with edges less than the number of the defined minimum step can be removed out. Therefore, both the explicit and implicit knowledge associations are retrieved and ranked under the unified correlation degree standard with the minimum step being 1. However, if we increase the minimum step to be 2, explicit associations (two concepts directly linked by a single edge) will be removed so that 'implicit' associations will be retained only.

Search Path function aims to find the implicit associations/relation paths between two knowledge concepts which is similar to Study 6. Dijkstra's shortest path algorithm is applied to discover and rank the knowledge association paths between two concepts in order from the highest correlation degree to lowest based on our unified standards. We also configure two different criteria of path distance with different retrieval behaviours. Specifically, we use *HM_l* to search for long paths with professional/specific concepts, and *GM_n* for short and basic/general concepts. The details of these two criteria can be found in Chapter 5 and Table 5-1.

In addition to Explore and Search path, we also provide clustering analysis in the function panel to group closely related concepts into different communities. Two clustering algorithms are implemented. The first one is k-mean clustering where the user can define the number of clusters to be generated within the current graph. The other one is called Markov clustering that the user can define the granularity of the clusters by using an inflation factor in which case the number of generated clusters is dependent on the defined granularity.

Figure 6-4 shows the hierarchical structure of the function panels for retrieval interaction. Figure 6-4 (a) is the function menu of the three query analysis. Figure 6-

4 (b) is the Explore panel with adjustable minimum steps, where 'General' option corresponds to the *HM_n* criteria, and 'Specific' corresponds to the *GM_l* criteria. Figure 6-4 (c) shows the Search Path panel where 'Professional' option uses the *HM_l* criteria and 'Basic' option uses *GM_n* criteria for the path distance. Figure 6-4 (d) is the clustering panel containing the options of k-mean algorithm and Markov algorithm.

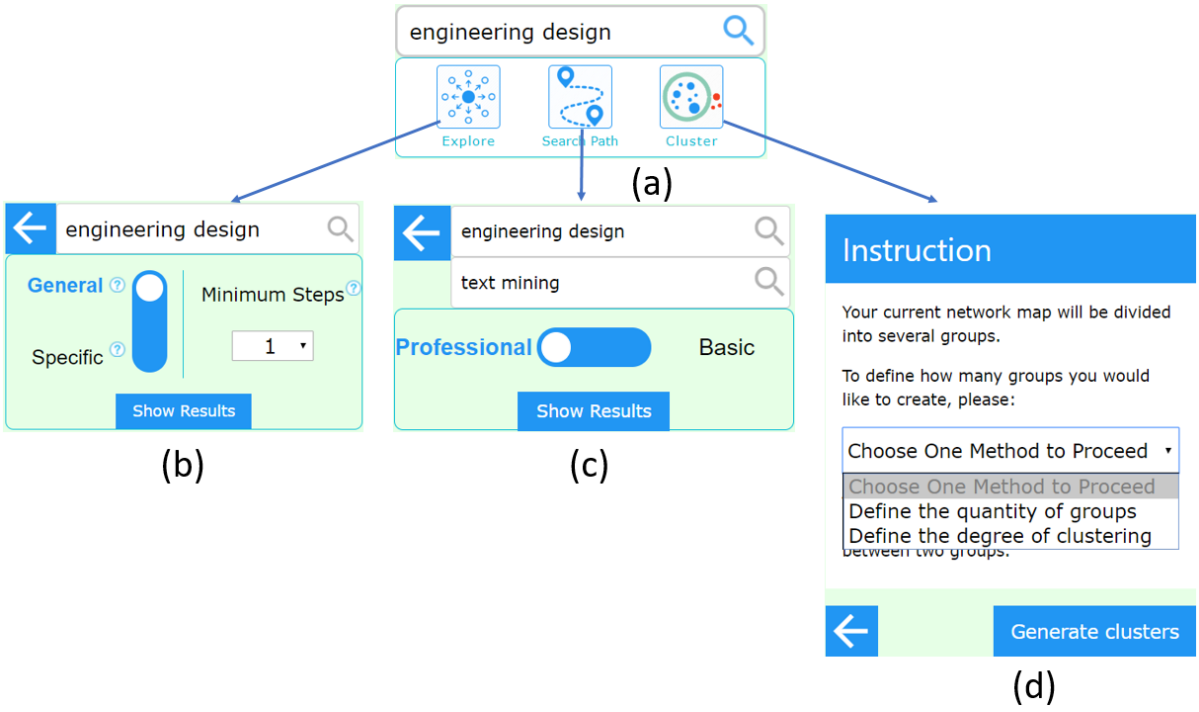


Figure 6-4 Hierarchical structure of the function panels

6.3 Study 7: User Test

Background

In this study, we conduct an online user test to evaluate the knowledge concepts and relations extracted from our huge constructed semantic network. This online user test is created based on the B-Link platform and incorporates the *Explore* and *Search path* functions of the B-Link into the test. It enables user to retrieve related concepts for a single arbitrary concept and also search implicit knowledge

associations between any two concepts. In this test, the results are retrieved from both our semantic network as well as other public benchmarking knowledge resources. Users are required to compare the results between our approach and benchmarking approaches, and give their ratings on different aspects based on their own expertise. In order to prevent bias, the users are not able to know the corresponding approaches of the retrieved results when performing evaluation. Finally, significant statistics test can be conducted on the user ratings to make comparison between our approach and benchmarking approaches and also between different quantitative criteria of our own approach.

We invite 24 participants from different domains in the broad design and engineering fields. Their expertise involve mechanical design/engineering, physics, tribology, chemical engineering, energy, computer science, information technology, interaction design, user experience, contamination design, civil engineering, artificial intelligence, business, start-up, creativity, psychology, etc. The 24 participants consist of PhD students and research associates from academic field as well as experts from industry, aged from early 20s to late 30s. All the online tests were conducted in London, UK although the cultural background of the participants involves various countries from Asia, Mideast, Europe, and North America to South America. Every participant conducted the online test individually under a 30 min time framework.

Evaluation tasks

The online test includes three evaluation tasks in total:

(1) First evaluation task:

In first task, users are required to input an arbitrary knowledge concept of their own expertise. Then, six approaches will be used to retrieve six different groups of concepts related to their input query concept. For each user, the six approaches are

randomly selected from our total 10 available approaches as shown in Table 6-1. Approaches 1-6 are based on the resource of our constructed semantic network but use the six different correlation degree criteria of Table 5-1, since different correlation degree criteria can have different retrieval behaviours to retrieve different types of concepts as discussed in Chapter 5. Four other approaches rely on the public benchmarking knowledge resources. Approach 7 uses WordNet to find the relevant concepts which are usually the hyponyms, hypernyms and synonyms of the input concept. The WordNet is readily accessible from NLTK (Bird et al., 2009). Approach 8 and 9 use respective online API to retrieve the relevant concepts from ConceptNet database (Speer et al., 2017) and NeLL database (Carlson et al., 2010). Approach 10 utilises the 'See also' section in Wikipedia (Wikipedia, 2017), which usually shows other related entries of current entry.

Table 6-1 Ten available approaches to retrieve the relevant concepts

Approaches	Resource	Criteria/method
1	Our semantic network	*AM_g*
2	Our semantic network	*AM_l*
3	Our semantic network	*GM_g*
4	Our semantic network	*GM_l*
5	Our semantic network	*HM_g*
6	Our semantic network	*HM_l*
7	WordNet	NLTK WordNet corpus
8	Concept	Online API
9	NeLL	Online API
10	Wikipedia	MediaWiki API

In order to relieve the task intensity, the platform will randomly and only select 6 approaches from the 10 approaches for each user. For each approach, the top 10

concepts related to the query concept will be retrieved and shown to the user as a group. In other speaking, six different groups of results will be retrieved simultaneously by using six different approaches, and each group contains the 10 top relevant concepts retrieved from the corresponding approach. In order to avoid bias, the order of the six groups is shuffled and the corresponding approach for each group is not disclosed to the user, where these groups are just labelled as group 'A', 'B', ..., and 'F'.

The user needs to evaluate each group in terms of three aspects: *Relevancy*, *Informativity*, and *Specificity*, using a rating scale of -5 ~ 5:

- For *relevancy*, -5 means "the associations generated in this group are NOT at all relevant and NOT related to the user's input", and 5 means "very relevant to the user's input".

- For informativity, -5 means "the associations generated in this group are NOT informative and NOT insightful at all", and 5 means "very informative and insightful".

- For specificity, -5 means "the generated concepts in this group are very general and common concepts/knowledge", and 5 means "very specific, detailed, and professional concepts/knowledge".

As an example, the web page layout of this first task is shown in Figure 6-5.

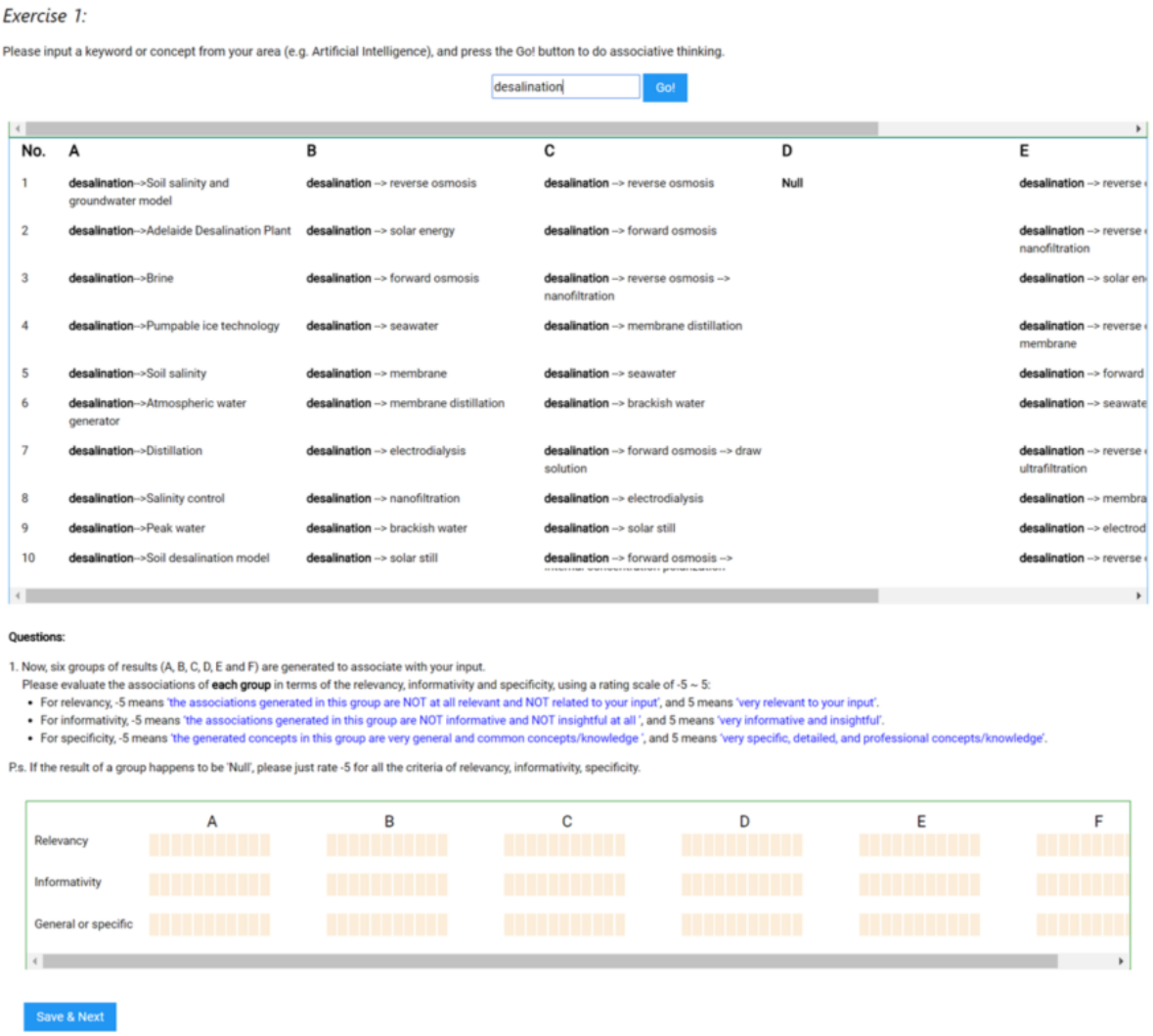


Figure 6-5 Layout of the online test first task

(2). Second evaluation task:

In this second task, the same knowledge concept in the first task will be reused. The difference is that this task will retrieve implicit associations around the knowledge concept. Instead of retrieving both explicit and implicit knowledge associations under the unified correlation degree standards from our semantic network in the first task, this second task sets the minimum association step to be 2 so that explicit associations (two concepts directly linked by a single edge) will be discarded and 'implicit' associations will be retrieved only.

Since other public benchmarking databases WordNet, ConceptNet, NeLL and Wikipedia (approaches 7, 8, 9, 10) are unable to retrieve implicit associations or multi-step associations, this task will just use approaches 1~6 in Table 6-1 to retrieve the implicit knowledge associations. Similarly, each of the six approaches will retrieve top 10 implicit associations with the highest correlation degree as a result group, and six groups of results respectively based on the six approaches will be provided for the user to evaluate and compare. The order of the six approaches is shuffled and the corresponding approaches will not be disclosed.

Also, the user will need to evaluate the result of each group in terms of the same three aspects as the first task.

(3). Third evaluation task:

In the third evaluation task, the users are required to input two arbitrary knowledge concepts, and the knowledge association paths between the two concepts will be retrieved. Again, since the other benchmarking approaches are unable to find a relation path between two concepts, we will only use the six approaches 1~6 based on our semantic network in this task, as shown in Table 6-1.

As discussed in Chapter 5, each approach based on a different criteria of correlation degree has different retrieval behaviour and will retrieve different type of knowledge association paths. In this task, the top five paths with the highest correlation degree will be retrieved for each approach. The results of the six approaches will be evaluated and compared by the users. Similarly, the order of the six approaches is shuffled and the corresponding approach for each group of results will not be disclosed to the user.

In this task, we evaluate on a different aspect *interestingness* instead of *informativity*, since the initial intention to search relation paths between two unrelated/arbitrary concepts is to help provoking the novelty and creative idea

generation in design and engineering process. Therefore, each group is evaluated based on three aspects of *Relevancy*, *Interestingness*, and *Specificity*, using a rating scale of -5 ~ 5:

- For relevancy, -5 means 'In this group, the relation paths are NOT relevant and not reasonable at all for connecting the two inputs', and 5 means 'very relevant and reasonable'.

- For interestingness, -5 means 'In this group, the relation paths are NOT interesting at all', and 5 means 'very interesting'.

- For specificity, -5 means 'In this group, the bridging concepts generated between the two inputs are very general and common concepts/knowledge ', and 5 means 'the bridging concepts are very specific, detailed, and professional concepts/knowledge'.

Results

All the input queries and corresponding ratings of the 24 participants are recorded into the database for following data analysis. Due to participants' diverse background, their input queries involve various knowledge concepts in a broad engineering and design areas. For example, the inputs in Task 1 and Task 2 include "deep learning", "behaviour change", "3D printing", "thermodynamics", "start-up", etc. The search path queries in Task 3 have some more interesting concept pairs like ("human-centred design", "circular economy"), ("synthetic biology", "religion"), ("3D printing", "energy") and even ("natural language processing", "stock market").

We finally collected 1296 ratings (data points) in total, which consist of 24 participants' ratings on three tasks where each task needs the ratings on six approaches in terms of three aspects (so $1296 = 24 \times 3 \times 6 \times 3$). These rating data

can be used for the statistical significant test to compare the difference of performance among our semantic approaches and benchmarking approaches.

Since each participant has evaluated multiple groups of result, we conducted paired dependent significant test to evaluate the difference between every pair of groups. Namely, *dependent t-test* is used for two groups with normal distributed difference, and *Wilcoxon signed rank test* will be used otherwise. *Kurtosis test* is applied to check the normality of the sample difference.

(1) Results of task 1:

Task 1 was to explore the relevant concepts around a single query concept, where the participants evaluated the six groups of results retrieved by six different approaches. Since the six approaches are randomly selected from 10 available approaches for each participant, it is not guaranteed that given a pair of two approaches, all the participants have evaluated both of the two approaches simultaneously. Therefore, when doing a dependent significant test between two approaches, we only select the data points of the participants who have evaluated both of the approaches, so the samples size would be smaller than 24. For example, the sample size of dependent statistical test to compare 'HM_l' and 'GM_g' is 12 in this case.

Figure 6-6 shows the **p-value** results of the statistical significant test to compare every pair of the 10 approaches in Table 6-1. We use the significant level of 0.1 to distinguish the significant difference, where result of p-value below 0.1 are considered to be significantly different and shown in darker background with white-colour text, otherwise it will be considered to be not significantly different and shown in lighter background with black-colour text.

Figure 6-6(a), (b) and (c) show the test results for *relevancy*, *informativity* and *Specificity* respectively. For relevancy, the overall mean value for each approach is: (HM_g 0.72), (HM_l 1.87), (GM_g 3.0), (GM_l 2.69), (AM_g 3.57), (AM_l 1.33),

(WordNet -2.33), (ConceptNet 0.4), (NeLL -4.93) and (Wikipedia -1.0). We can obviously see that *WordNet* and *NeLL* are significantly different from other approaches, and after looking into the scores, we found these two approaches have much lower ratings for relevancy compared to other approaches. It is simply because most of the users' queries such as *behaviour change*, just cannot be found in the WordNet and NeLL databases. On the other hand, *GM_g*, *GM_l*, and *AM_g* are shown to have significantly higher score of relevancy compared to the rests of the approaches: *HM_g*, *AM_l*, *ConceptNet,* and *Wikipedia*. Interestingly, *HM_l* has no significant difference from any other approaches except *WordNet* and *NeLL*. Therefore, based on the user ratings, the performance of the ten approaches, regarding the relevancy of their retrieved concepts to the single query, can be ranked as follows:

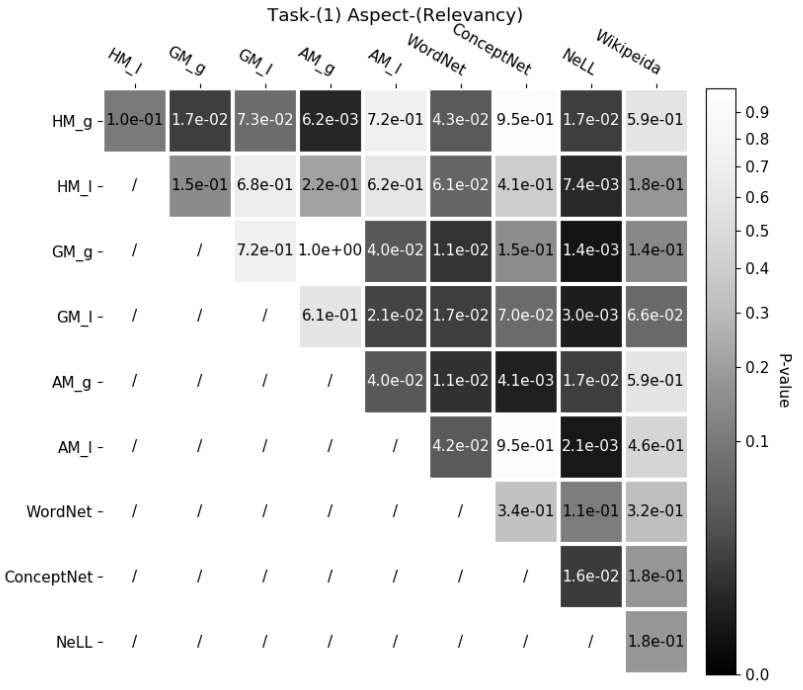
- High relevancy approaches: GM_g, GM_l, AM_g, (HM_l)
- Median relevancy approaches: HM_g, AM_l, ConceptNet, Wikipedia (HM_l)
- Low relevancy approaches (due to the limited concept coverage): WordNet, NeLL.

This indicates that overall, our semantic network can retrieve significantly higher relevant concepts compared to the other public benchmarking databases though ConceptNet and Wikipedia can compete equally with some of our criteria.

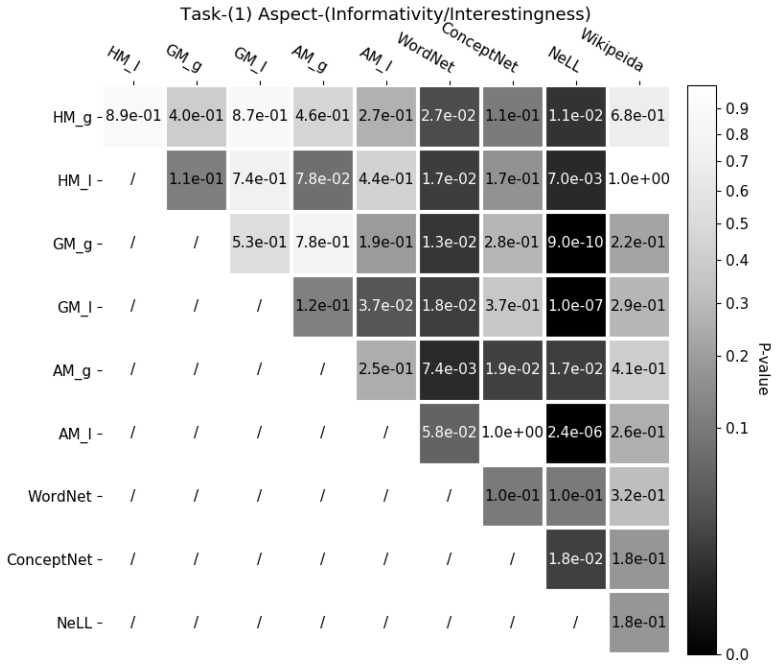
For informativity shown in Figure 6-6(b), similarly, the WordNet and NeLL are significantly less informative than other approaches simply because they are just unable to retrieve anything for the various user queries. There are not much significant differences either between our approaches and ConceptNet and Wikipedia or among our own six approaches. This is probably because different approaches and criteria have different retrieval behaviour targeting on different types of relevant knowledge concepts and it's hard to say which is more informative, since all the retrieved information can be useful in different cases. Just one point to notice is that

AM_g is significantly more informative than HM_l, and GM_l is more informative than AM_l. It's probably because that AM_g conducts a very BFS compared to the very DFS of HM_l, and the high relevancy of GM_l contributes to its informativity when compared with AM_l with less relevancy.

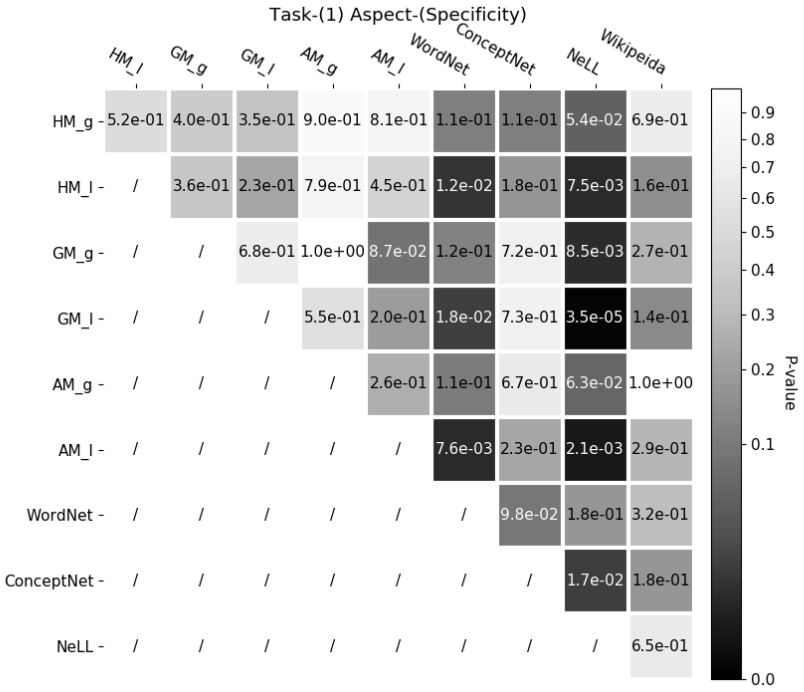
We also cannot tell much difference among the approaches on *specificity* in this task as shown in Figure 6-6(c), since the minimum knowledge association step in this task is 1 and we only show the top 10 concepts, which is difficult for the user to distinguish between general and specific concepts in such short association step and small amount of concepts. Except WordNet and NeLL, the only significant difference is shown between GM_g and AM_l, which is reasonable that GM_g using global normalisation tends to retrieve general concepts while AM_l based on local normalisation tends to yield specific concepts.



(a) Relevancy



(b) Informativity



(c) Specificity

Figure 6-6 Paired significant test results in Task 1

(2) Results of task 2:

This second task focuses on evaluating the implicit knowledge associations where the explicit knowledge associations directly linking the query to a concept within only one association step are filtered out/removed from the results. In other words, only multi-step knowledge associations from each approach will be retrieved for the users. Since the other benchmarking approaches are not able to retrieve implicit knowledge associations (multi-step associations), we can only make comparison between the six criteria of our own approaches.

Since there are totally only six available approaches in this task instead of 10 available in previous task, the user will evaluate the results retrieved by each of the six approaches. Therefore, the sample size for each approach on each aspect is exactly 24. Similarly, we conduct *paired t-test* or *Wilcoxon signed rank* to compare the significant difference between every pair of the six approaches in terms of *Relevancy*, *Informativity*, and *Specificity*. The P-value of the test results are shown in Figure 6-7 (a), (b) and (c) for the three aspects respectively.

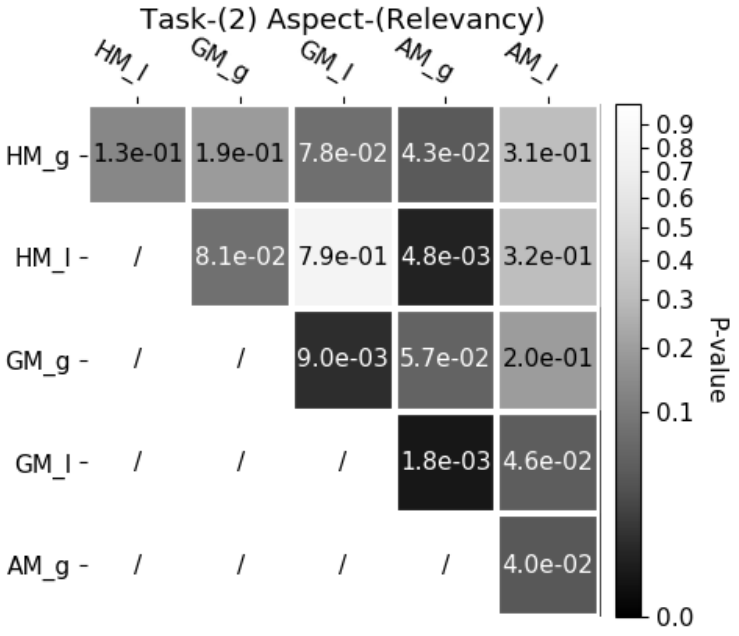
For relevancy, the mean score for each approach is: (HM_g, 1.34), (HM_l, 2.17), (GM_g, 0.58), (GM_l, 2.46), (AM_g, -0.46), and (AM_l, 1.5). In general, it can be seen that the local fluctuation weighting usually has higher score than the feature scaling weighting from global perspective. From Figure 6-7 (a), the relevancy score of AM_g is significantly lower than all the other approaches, while GM_l has significantly much higher relevancy score than other approaches, which is followed by HM_l. Based on the significant test, we can rank the performance of the six approaches regarding the relevancy of the retrieved implicit knowledge associations as follows:

$$GM\_l > HM\_l > HM\_g, AM\_l, GM\_g > AM\_g$$

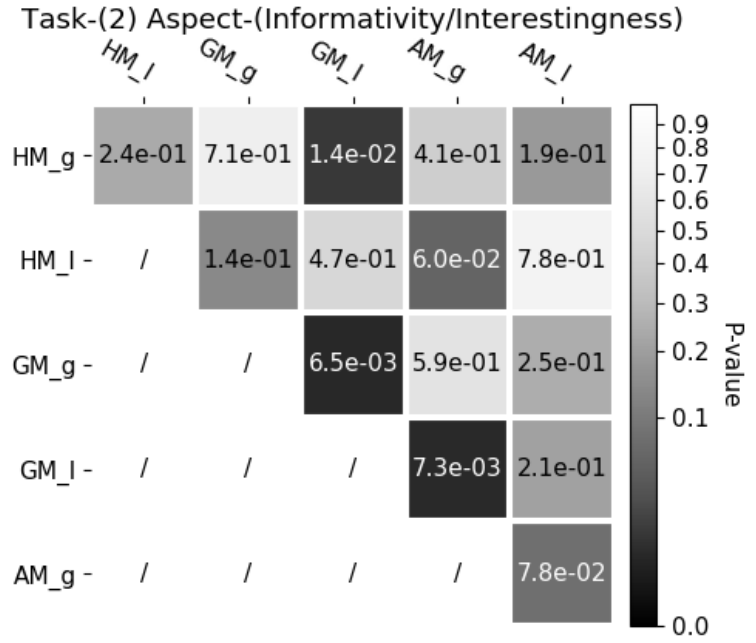
For informativity aspect, the mean score of these approaches are: (HM_g, 0.54), (HM_l, 1.5), (GM_g, 0.375), (GM_l, 2.17), (AM_g, -0.08), and (AM_l, 1.29).

Interestingly, the test result for informativity is very similar with the results of relevancy. This is probably because that the users tend to think that relevant knowledge relations are also informative. Based on Figure 6-7 (b), it shows that GM_l is significantly more informative than HM_g, GM_g, and AM_g. Also, AM_g is significantly less informative than HM_l, GM_l, and AM_l.

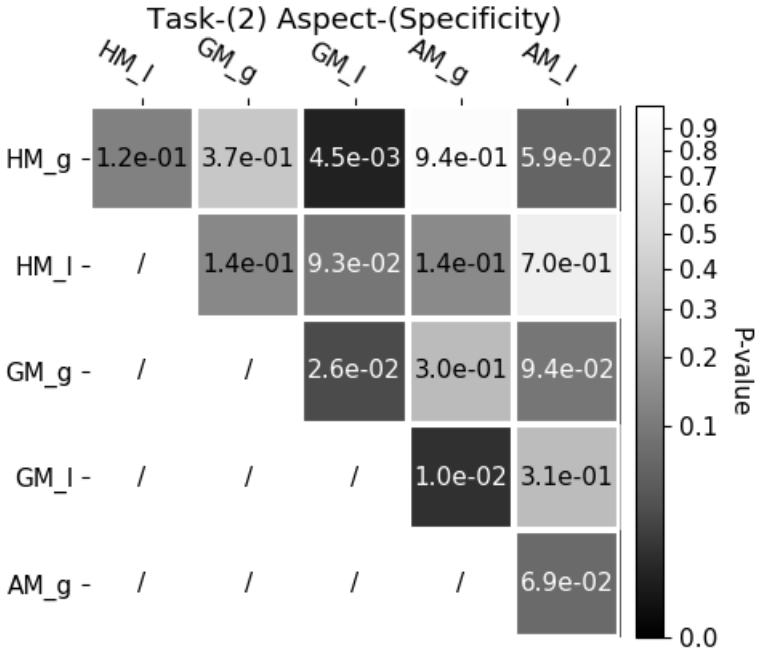
For specificity, the mean score is: (HM_g, 0.17), (HM_l, 1.54), (GM_g, 0.46), (GM_l, 2.71), (AM_g, 0.08), and (AM_l, 2.00). The significant test of Figure 6-7 (c) just shows a beautiful result which is consistent with our discussion in Chapter 5. The statistical test together with the mean values clearly illustrate that the HM_l, GM_l and AM_l using local normalization will retrieve significantly more specific concepts than HM_g, GM_g and AM_g using global normalization. Also, from the above mean values and the paired test between GM_l and HM_l, we can see that GM (Geometric mean) algorithm tends to get more specific concepts than AM (Arithmetic Mean) and HM(Harmonic Mean). Differing from task 1, this second task focuses on retrieving multi-step knowledge associations and therefore the specificity degree of concepts can be more intuitive for the users to distinguish between different approaches.



(a). Relevancy



(b) Informativity



(c) Specificity

Figure 6-7 Paired significant test results in Task 2

(3) Result of task 3:

This third task aims to search the knowledge association paths between two query concepts in order to find the bridging concepts between the two domains and also get possible novel insights to provoke innovation. Again, the other benchmarking approaches are unable to find path between concepts, therefore, we just keep only using the six criteria based on our constructed semantic network. The sample size of each approach on each of the three aspects is 24. The p-values of the paired significant test are shown in Figure 6-8 (a), (b) and (c) for relevancy, interestingness and specificity respectively.

For relevancy, the mean score of each approach is: (HM_g, 0.67), (HM_l, -0.21), (GM_g, 1.13), (GM_l, 2.13), (AM_g, 1.04), and (AM_l, 1.54). Based on the statistical significant test in Figure 6-8 (a), it can be seen that GM_l has significantly higher relevancy degree than the other approaches except AM_l. On the other hand, it also clearly shows that HM_l has the significantly lowest relevancy compared to other approaches except HM_g. There are no significant differences shown between HM_g, GM_g, and AM_g. The performance of the six approaches in terms of the relevancy of the retrieved relation paths between two query concepts can be ranked as:

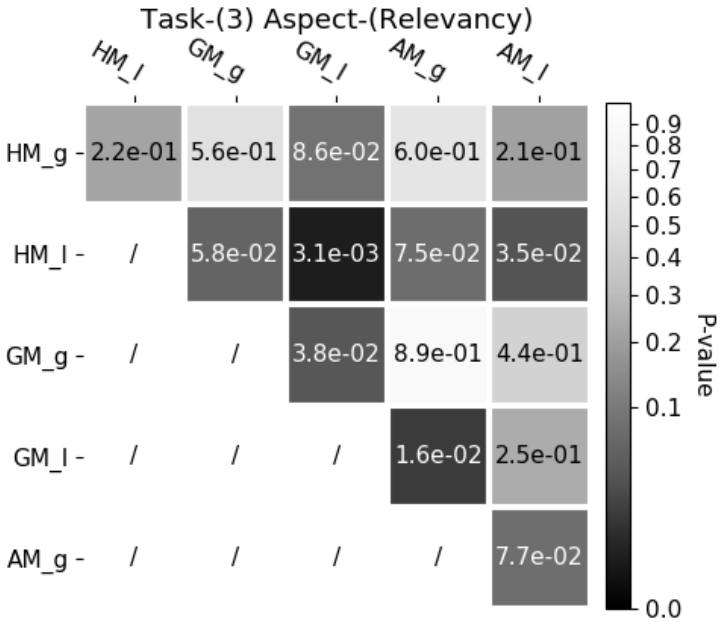
$$GM\_l, (AM\_l) > (AM\_l)\ AM\_g, GM\_g, (HM\_g) > (HM\_g), HM\_l$$

Also, the mean values show that in task, HM (harmonic mean) algorithms has lower relevancy degree than AM (Arithmetic mean) and GM (Geometric Mean). It is because HM tends to retrieve longer paths between two concepts which usually cross multiple distant/irrelevant domains. This leads to an adverse impact on its relevancy degree when evaluated by the users.

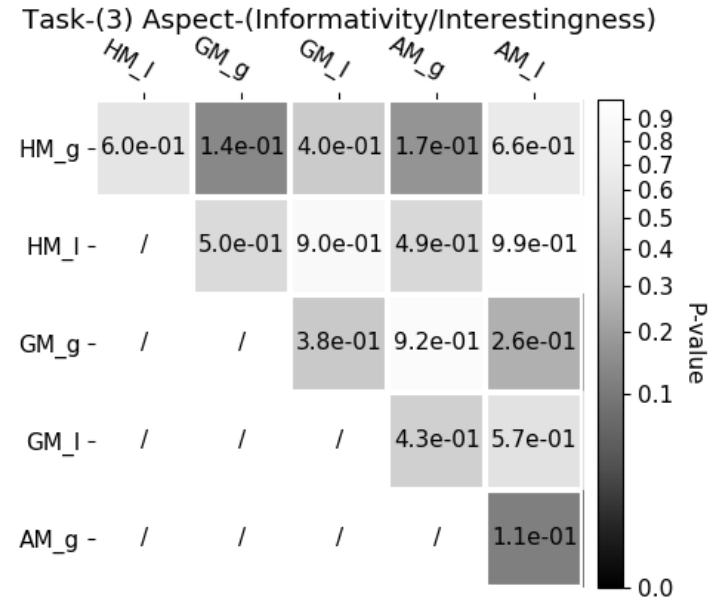
We initially expected to see quickly which approach/ correlation criteria can search and identify significantly more interesting knowledge relations paths that can potentially provoke generation of creativity idea. However, the statistical test results

in Figure 6-8 (b) show that there are no significant differences between the six approaches. It is probably because people usually perceive “interestingness” in different ways. Nevertheless, if we look at the mean score of the six approaches which are: (HM_g, 0.63), (HM_l, 0.21), (GM_g, -0.29), (GM_l, 0.04), (AM_g, -0.25), and (AM_l, 0.29), we can have an intuition that HM algorithm, to some extent, seems to have higher score on interestingness than AM and GM in this case. This is also because the longer paths retrieved by HM will link through multiple distant and other surprising knowledge domains which makes the knowledge associations perceived to be interesting.

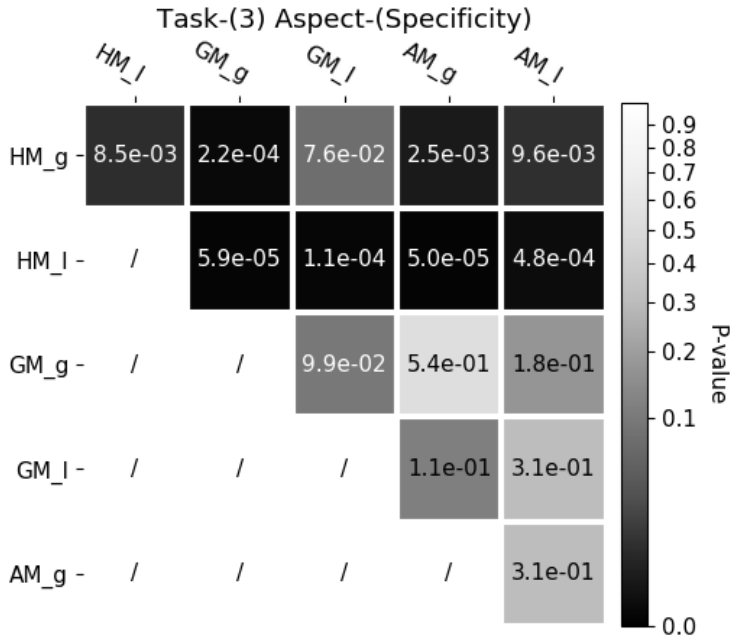
For the specificity, by looking at the mean score for each approach as (HM_g, 1.21), (HM_l, 3.29), (GM_g, -1.71), (GM_l, -0.375), (AM_g, -1.458), and (AM_l, -0.917), we obviously see that local fluctuation weighting normalization tends to search for specific concepts while global feature scaling weighting normalisation usually retrieves general concepts, which is consistent with the discussion in Chapter 5. Besides, we can also see from Figure 6-8 (c) that HM algorithm is significantly different from GM and AM algorithms. This indicates that when searching relation paths between two concepts, HM tends to explore significantly more specific bridging concepts than GM and AM. Another significant difference is found between GM_g approach and GM_l approach, which is reasonable due to the two different normalisation techniques.



(a) Relevancy



(b) Interestingness



(c) Specificity

Figure 6-8 Paired significant test on the six approaches in task 3

6.4 Conclusion

This Chapter has described the methods used for the final data visualisation and user interaction of our data-driven design pipeline. Specifically, D3 (Data-driven document) with a forced-directed graph is applied to realise the multi-dimensional and dynamic visualisation of our previously constructed huge semantic network in Chapter 4. A web platform called "B-Link" has been developed. It implements the retrieval framework, unified correlation degree criteria/standards for both explicit and implicit associations, and semantic analysis with shortest path searching proposed in Chapter 5 to enable the user to interact, retrieve and query on our constructed semantic network.

Finally, an online-user test was conducted to evaluate the effectiveness of our approaches from the user's perspective. Based on the domain expert ratings,

significant statistical test results show that our constructed semantic network coupled with the proposed unified criteria of correlation degree can retrieve significantly more relevant knowledge associations compared to other four public benchmarking databases. This indicates our approaches are effective and can indeed capture and retrieve the knowledge associations from the perspective of engineering and design expertise.

In addition, the test results also demonstrate the distinct differences between the retrieval behaviours of our different criteria. For example, GM algorithm tends to retrieve more relevant and informative knowledge associations than AM and HM criteria, while HM, in case of searching relation paths between two concepts, can retrieve slightly more interesting paths with significantly more specific concepts which link through multiple distant knowledge domains. Furthermore, there shows significant differences on knowledge specificity between the techniques of local fluctuation normalisation and feature scaling normalisation, where the former tends to retrieve specific concepts and the later usually search for general concepts which are consistent with the discussion in Chapter 5.

Chapter 7 Conclusions

7.1 Research summary

A summary of the works carried out in this thesis is illustrated schematically in Figure 7-1.

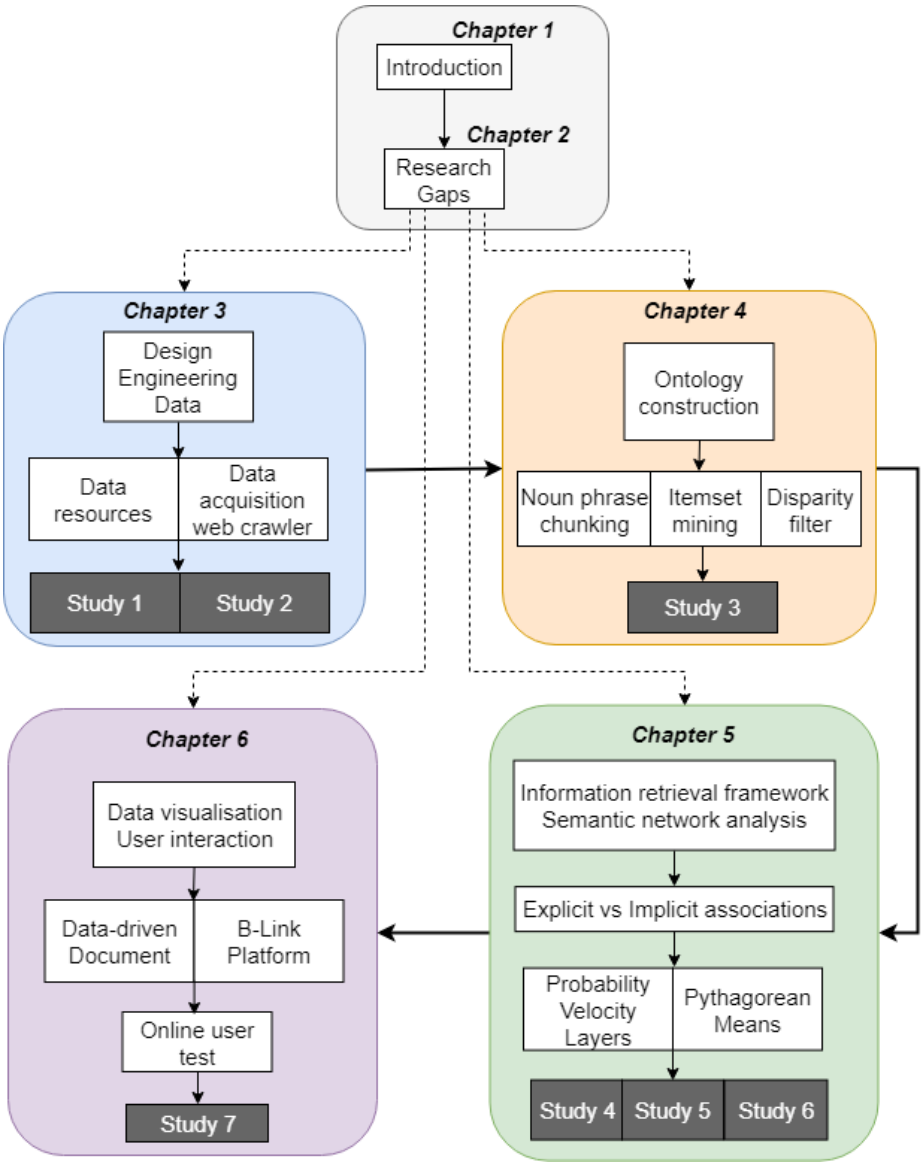


Figure 7-1 Schematic summary of the thesis

As shown in Figure 7-1, this thesis contains two research structures:

- One structure can be interpreted as follows. A review of existing work has been undertaken (Chapter 2) and four current research gaps/issues identified, namely, the domain of ontology, the construction techniques, the use and analysis of the semantic network, and finally the visualisation and interaction. Then Chapter Three, Four, Five and Six are conducted to tackle each of these four issues respectively.
- Another structure is as follows. Chapter three to six actually form a data processing pipeline for data-driven design information retrieval. The pipeline starts from data acquisition (Chapter three), then data transformation from unstructured to structured (Chapter four), to data analysis and retrieval (Chapter five) and finally data visualisation and interaction (Chapter six).

Totally seven studies have been conducted throughout the thesis to execute and evaluate the specific elements proposed in each chapter.

7.2 Key Findings and Contributions

This thesis corresponds with the emerging field of data-driven design in the recent big-data economy. For 'design', we specifically focus on the design information retrieval aspect. For 'data', we utilise the massive electronic and digital textual data available online. For 'driven', we based on the well-developed text mining and ontological techniques. In one word, the aim of this thesis is to facilitate the design information retrieval process (design) through the use of text mining and ontological technology (driven) on a large-scale of online textual data (data).

A list of key novelty points is as below:

- Instead of processing and retrieving information at document and fragment level as the traditional document retrieval approaches, our approach processes and retrieves information at the semantic and concept level.

- Differing from existing ontology network focusing on common-sense and domain-specific knowledge, our approach tries to construct a semantic network from a design and engineering perspective to build design- and engineering- oriented relations.
- Instead of manually handcrafting the ontology in traditional method, we automate the semantic network construction through text mining.
- Differing from the existing automatic ontology construction method extracting either fixed types of concepts or relations, our automatic approach removes the constraints by extracting flexible concepts and relations.
- Instead of only considering the explicit knowledge association in traditional ontology-based information retrieval, our approach considers and quantifies the implicit knowledge associations from a network analysis perspective.
- In order to leverage the human intuition and instinct visual system to directly detect the patterns, spot trends and identify new knowledge, our approach developed a data-visualisation system to support the information retrieval process.

The key findings resulting can be described as follows.

Chapter 2 studies previous works on the application of text mining and ontology in design and engineering fields. It was found that most text mining techniques process information at document level and are usually used for document retrieval, document classification and clustering in design information retrieval. Ontological methodology provides the opportunities to process information at semantic and concept level. However, current public ontology databases are either created for common-sense knowledge or edited for specific domains (biomedical) or particular engineering projects. There is no ontology database particularly focusing on design and engineering relations from a broad perspective, which raises the needs to build a "WordNet" for design and engineering knowledge. The construction methods for ontology can be divided into three categories: manual, unsupervised and supervised

approaches. Manual approaches are time-consuming in the case of big data scenario with large-scale textual data. Supervised approaches require large amount of annotated training set where various types of design and engineering relations (e.g. mechanism, function, structure, materials) should be manually labelled. However, such kind of training data does not exist in design and engineering field at this moment. For unsupervised approaches, statistical methods usually directly use single word as the entities and haven't consider using the phrase and knowledge concept consisting of multiple words as a whole unit to be the targeted entity, while linguistic methods use predefined language patterns / regular expressions for string matching where only limited and fixed predefined types of relations can be extracted. Therefore, regarding the highly complex and diverse nature of design and engineering relations, a new automatic approach for constructing design and engineering semantic network is needed. With the essential target on knowledge reuse and representation, the specific applications of ontology and semantic network in design engineering can be classified into three categories: design information annotation, sharing and retrieval, interoperability between engineering systems, and product design configuration. However, in these applications, ontology is used only based on its individual related concepts and relations where the network structure of the ontology was not fully utilised. It would be valuable to apply sophisticated network analysis techniques on the semantic network to facilitate the design information retrieval processing. Finally, data visualisation and user interaction were rarely focused in the previous works on design information retrieval.

The fundamental building block for data-driven design is data. Chapter 3 identifies the available data resources and data acquisition methods used in this research. Fundamental differences exist between data, information and knowledge. Data is the description of raw facts which is highly unstructured and heterogeneous, while information is the structured representation of data within specific context or usable format and knowledge can be then obtained through the understanding and learning of the information and their relationships which enables people to make

action/decision. Raw data can be divided into different forms including numerical data, graphic data and textual data while a large part of information, approximately 80%, is available in textual data format, which is the most commonly used data form for design and engineering information retrieval. Internet is the ideal resource for textual data because it's vast in size, updated in time, related in structure, and diverse in type. Web crawling is the powerful data acquisition technique for internet resource. Two studies are carried out by using web crawling techniques on different internet resource in order to gather the raw data to prepare for following design information retrieval. Study 1 extracts the texts from YanKo design, a modern industrial design website, where 1000 design posts are crawled within only 15 mins which is significantly faster than the manually browsing process. Study 2 focuses on the academic literatures / papers in engineering and design domains. APIKeys were applied to the Elsevier service and the crawling process took more than one year to finally fetch the metadata of 3,713,886 papers, 928,471 of which are parsed along with the full texts. The amount of publications has increased exponentially in the last decade from around only 7,182 papers at 1995 to about 309,790 papers in 2015, and journals indexed by the initial letter *I, J* in the name have the largest volumes of papers.

The raw texts captured in Chapter 3 is unstructured data, and is necessary to be processed into structured information for following retrieval purpose. Chapter 4 present our novel ways to transform the data from raw texts to concepts, relations and final structured ontology / semantic network. Briefly, frequent itemset and association rule mining are applied to build relation between any keywords within the same paper and any noun phrases within the same sentence, and then a huge semantic network is created based on these millions of relations and concepts. Disparity filter is conducted to remove the noisy, irrelevant relations and retain the significantly relevant relations. Study 3 uses a golden dataset containing professional engineering design concepts and relations in order to evaluate whether our approach can indeed capture more knowledge concepts and relations from the design and

engineering perspectives and contexts. We compare the precision and recall of our ontology network with three other public ontological databases including WordNet, ConceptNet and NeLL. The results show that our ontology network has a higher retrieval rate on more specific and professional concepts and technical terms compared to the other three benchmarking systems. Node strength is shown to be related with the specific level of the concepts that general concepts usually have higher node strength while specific concepts often have lower node strength. This useful feature can be utilized to form a top-down concept structure for a particular domain based on the specificity level of each concept. For the performance on retrieving design- and engineering-oriented relations, our constructed ontology network has a higher recall but lower precision which yield a better total F1 score compared to the benchmarking methods.

After the construction of the ontology network, the next question is how to use this network to benefit the actual design information retrieval process. Chapter 5 presents an information retrieval framework based on the use of the constructed semantic network in Chapter 4. In this framework, probability and velocity layer modelling, and Pythagorean means are implemented to develop novel criteria to retrieve and quantify the correlation degree of both explicit and implicit knowledge associations under the same unified standard. Study 4 shows that the golden relations are indeed closer and have small distance in the probability and velocity layer. This indicates that probability and velocity analysis can help provide useful information about the correlation degree between any two concepts, which is consistent with the human judgement. Two real design case studies were conducted to demonstrate the practical use of the framework in retrieving information either around a single concept (Study 5) or between two concepts (Study 6) of the design queries. The results show that the knowledge concepts and relations, which are retrieved from our constructed semantic network through the proposed retrieval framework, can significantly support the design activities and assist with the idea generation at conceptual stage. More importantly, using different criteria of

correlation degree shows significantly different retrieval behaviours. Probability analysis and local fluctuation (AM_l, GM_l and HM_l) tend to retrieve domain-specific knowledge concepts while velocity analysis and feature scaling (AM_g, GM_g and HM_g) tend to explore general knowledge concepts. Besides, AM tends to retrieve short implicit association across less edges focusing on relevant knowledge within the same domain, while HM tends to retrieve long implicit association with more edges across multiple distant domains, and GM shows a balance between them. These different retrieval behaviours of the different criteria indicate the great potential of our constructed semantic network and retrieval framework in satisfying the different and various knowledge demands during engineering design activity.

Chapter 6 presents the final block of data visualisation and user interaction in our data-driven design pipeline. The Data-Driven Document technique is used to visualise our constructed semantic network as a forced-directed graph in a multi-dimensional and dynamic manner. A web platform called "B-Link" is developed to enable the user to interact and query with our constructed semantic network through the proposed retrieval framework and unified quantitative criteria of correlation degree. Study 7 carried out an online-user test to investigate the difference between our approach and benchmarking approaches (including Wordnet, Conceptnet, NeLL, and Wikipedia) and also among the six criteria of our own approach. Statistical tests (dependent t test and *Wilcoxon signed rank test*) are conducted to investigate the significant differences. Overall, our approach can retrieve significantly more relevant knowledge associations compared to other four public benchmarking databases from the users' perspective. GM_g, GM_l and AM_g have the best performance on the relevancy with minimum step being 1, where ConceptNet and Wikipedia also show relatively good performance and can compete with criteria HM_g, AM_l and HM_l of our approach, while WordNet and NeLL have the worst performance in this case. Regarding the different behaviours of the six criteria of our own approach, GM algorithm tends to retrieve more relevant and informative knowledge associations than AM and HM criteria, while HM, in case of searching relation paths between two

concepts, can retrieve slightly more interesting paths with significantly more specific concepts which link through multiple distant knowledge domains. A significant difference exists between the knowledge specificity between the fluctuation normalisation and feature scaling normalisation, where the former tends to retrieve specific concepts and the later usually search for general concepts.

Overall, the key aim of this thesis, i.e. to support design information retrieval through the use of text mining and semantic network has been achieved. This can be reflected specifically by each individual study:

- Study 3 and study 7 shows benchmarking result that our semantic network can indeed capture more design-oriented concepts and relations compared to the existing ontology databases.
- Study 3 and study 5 shows that by using the node strength, our approach can build top-down hierarchical concept structure to support by guiding the direction of information retrieval process.
- Study 4 shows that our approach can help detect and justify the golden design relations judged by human experts.
- Study 5 demonstrates the practical use of our approach can indeed help with design information retrieval activities in real-world design projects.
- Study 6 proves that our approach can provoke creative ideas and design innovation in practical problem solving.
- Study 4, study 5, study 6 and study 7 show the different kinds of retrieval behaviours that our approach is able to achieve, which can actually satisfy the variety of knowledge demands during the engineering design process.

7.3 Directions for future work

This thesis discusses the design information retrieval under the context of data-driven design (D^3) in this big-data era. Instead of retrieving information from particular data resources for specific design domains and activities in traditional design information retrieval, this thesis has an emphasis on harnessing the power of broad, noisy, real-world, large-scale data to facilitate the design information retrieval from a general perspective. An interesting point raised by this is how to utilise the data resources across domains under this big-data economy. In traditional methods, the data generated in a particular project or domain will be only used in this particular project and this particular domain. However, it would be valuable to explore ways of reusing the data from other seemly "irrelevant" domains or activities.

One example is that the user behaviour data can be applied into many domains such as business, design, and healthcare. This raises the issue of data sharing and reuse across domains. The two studies in Chapter 3 consider the use of design news posts and academic literatures, which is limited. It's worth considering in future work what kind of other data resources even in distant domains can be utilised for design purposes.

The second point needed to be further addressed is the highly-noisy nature in the variety of data resources of this big-data era. If we want to reuse and share data across domains or projects, we need to know how to remove and more importantly control the noise to retain the most relevant data. Chapter 4 controls the noise in the constructed semantic network through a naïve statistical disparity filter with a fixed significant level α. Further studies might investigate the effects of controlling the noise through different significant levels on the performance of the retrieved results (e.g. precision, recall).

Chapter 5 proposed an information retrieval framework based on the constructed semantic network. The relations between two concepts are retrieved purely based on their statistical correlation degree quantified by our criteria which unify both directly linked (explicit) associations and indirectly linked (implicit) associations. However, no concrete meaning is identified for the relation between two concepts. As discussed in Chapter 2, it's due to three main reasons: 1). The manual approach, which is time-consuming and human-effort intensive, would just be impossible in the case of large-scale data. 2). The predefined linguistic patterns in unsupervised approaches are far from satisfying the highly contextual and diverse types of design and engineering relations. 3). For supervised learning methods, the problem is the lack of an annotated training dataset specially labelled with a variety of design- and engineering- oriented relations. With the development of machine learning techniques, many other domains, especially biomedical domain have created their own annotated training dataset for domain-specific relation extraction. Therefore, another promising future work for data-driven design information retrieval would be to create a training set where relations are exclusively annotated based on the design and engineering perspective, and subsequently train a model based on this dataset. This will also greatly advance the automatic construction of design and engineering ontology for many other applications in design engineering field.

References

Abello, J. & Korn, J. 2002. MGV: A system for visualizing massive multidigraphs. *IEEE Transactions on Visualization and Computer Graphics,* 8**,** 21-38.

Aggarwal, C. C. & Zhai, C. 2012. *Mining text data*, Springer Science & Business Media.

Ahmed, S., Kim, S. & Wallace, K. M. 2007. A methodology for creating ontologies for engineering design. *Journal of computing and information science in engineering,* 7**,** 132-140.

Aleksiūnas, P., Hoffman, P., Korobov, M., Dorneles, E. & Ricardo, J. 2017. *Architecture overview - Scrapy 1.5.0 documentation* [Online]. Available: https://docs.scrapy.org/en/latest/topics/architecture.html [Accessed March 12, 2018].

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Alsabti, K., Ranka, S. & Singh, V. 1997. An efficient k-means clustering algorithm.

Antoniou, I. & Tsompa, E. 2008. Statistical analysis of weighted networks. *Discrete dynamics in Nature and Society,* 2008.

Bateman, J. A., Hois, J., Ross, R. & Tenbrink, T. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence,* 174**,** 1027-1071.

Bechhofer, S. 2009. OWL: Web ontology language. *Encyclopedia of database systems*. Springer.

Bellinger, G., Castro, D. & Mills, A. 2004. Data, information, knowledge, and wisdom.

Bertola, P. & Teixeira, J. C. 2003. Design as a knowledge agent: How design as a knowledge process is embedded into organizations to foster innovation. *Design Studies,* 24**,** 181-194.

Bikel, D. M., Miller, S., Schwartz, R. & Weischedel, R. 1997. Nymble: a high-performance learning name-finder. Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 194-201.

Bird, S., Klein, E. & Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.".

Bird, S. & Loper, E. 2004. NLTK: the natural language toolkit. Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 31.

Blanzieri, E. & Bryl, A. 2008. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review,* 29**,** 63-92.

Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.

Bostock, M. 2018. *D3.js - Data-Driven Documents* [Online]. Available: https://d3js.org/ [Accessed May 01, 2018].

Bradley, P. S. & Fayyad, U. M. 1998. Refining Initial Points for K-Means Clustering. ICML. Citeseer, 91-99.
Brin, S. & Page, L. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks,* 56**,** 3825-3833.
Bullinaria, J. A. & Levy, J. P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods,* 39**,** 510-526.
Caldas, C. H. & Soibelman, L. 2003. Automating hierarchical document classification for construction management information systems. *Automation in Construction,* 12**,** 395-406.
Card, S. K., Mackinlay, J. D. & Shneiderman, B. 1999. *Readings in information visualization: using vision to think*, Morgan Kaufmann.
Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R. & Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. AAAI. Atlanta, 3.
Chaklader, R. & Parkinson, M. B. 2017. Data-Driven Sizing Specification Utilizing Consumer Text Reviews. *Journal of Mechanical Design,* 139**,** 111406.
Chan, Y. S. & Roth, D. 2010. Exploiting background knowledge for relation extraction. Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 152-160.
Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F. & Gao, W. 2013. The evolution, challenges, and future of knowledge representation in product design systems. *Computer-aided design,* 45**,** 204-228.
Chang, X., Rai, R. & Terpenny, J. 2010. Development and utilization of ontologies in design for manufacturing. *Journal of Mechanical Design,* 132**,** 021009.
Chattopadhyay, P., Mondal, S., Bhattacharya, C., Mukhopadhyay, A. & Ray, A. 2017. Dynamic Data-Driven Design of Lean Premixed Combustors for Thermoacoustically Stable Operations. *Journal of Mechanical Design,* 139**,** 111419.
Chen, L., Shi, F., Han, J. & Childs, P. R. 2017. A network-based computational model for creative knowledge discovery bridging human-computer interaction and data mining. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, V007T06A001-V007T06A001.
Chen, L., Wang, P., Shi, F., Han, J. & Childs, P. 2018. A computational approach for combinational creativity in design. DS92: Proceedings of the DESIGN 2018 15th International Design Conference. 1815-1824.
Chen, Y.-M. & Jan, Y.-D. 2000. Enabling allied concurrent engineering through distributed engineering information management. *Robotics and Computer-Integrated Manufacturing,* 16**,** 9-27.
Chen, Y., Argentinis, J. E. & Weber, G. 2016. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics,* 38**,** 688-701.
Cheong, H., Li, W., Cheung, A., Nogueira, A. & Iorio, F. 2017. Automated Extraction of Function Knowledge From Text. *Journal of Mechanical Design,* 139**,** 111407.
Childs, P. R. 2013. *Mechanical design engineering handbook*, Butterworth-Heinemann.

Cho, J., Han, S. & Kim, H. 2006. Meta-ontology for automated information integration of parts libraries. *Computer-Aided Design,* 38**,** 713-725.
Cohen, A. M. & Hersh, W. R. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics,* 6**,** 57-71.
Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research,* 12**,** 2493-2537.
Cortes, C. & Vapnik, V. 1995. Support-vector networks. *Machine learning,* 20**,** 273-297.
Coules, H., Horne, G., Venkata, K. A. & Pirling, T. 2018. The effects of residual stress on elastic-plastic fracture propagation and stability. *Materials & Design,* 143**,** 131-140.
Cristianini, N. & Shawe-Taylor, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press.
Darlington, M. J. & Culley, S. J. 2008. Investigating ontology development for engineering design support. *Advanced Engineering Informatics,* 22**,** 112-134.
Dartigues, C., Ghodous, P., Gruninger, M., Pallez, D. & Sriram, R. 2007. CAD/CAPP integration using feature ontology. *Concurrent Engineering,* 15**,** 237-249.
Dering, M. L. & Tucker, C. S. 2017. A Convolutional Neural Network Model for Predicting a Product's Function, Given Its Form. *Journal of Mechanical Design,* 139**,** 111408.
Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *Numerische mathematik,* 1**,** 269-271.
Dong, A. & Agogino, A. M. 1997. Text analysis for constructing design representations. *Artificial Intelligence in Engineering,* 11**,** 65-75.
Dorst, K. & Cross, N. 2001. Creativity in the design process: co-evolution of problem–solution. *Design studies,* 22**,** 425-437.
Efthymiou, K., Sipsas, K., Mourtzis, D. & Chryssolouris, G. 2015. On knowledge reuse for manufacturing systems design and planning: A semantic technology approach. *CIRP Journal of Manufacturing Science and Technology,* 8**,** 1-11.
Elsevier. 2015. *Elsevier Developer Portal* [Online]. Available: https://dev.elsevier.com/index.html [Accessed June 01, 2015].
Elsevier. 2018. *Shop and Discover Books, Journals, Articles and more* [Online]. Available: https://www.elsevier.com/books-and-journals [Accessed Jul 18, 2018].
Fabian, M., Gjergji, K. & Gerhard, W. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. 16th International World Wide Web Conference, WWW. 697-706.
Feldman, R. & Sanger, J. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge university press.
Felfernig, A., Friedrich, G., Jannach, D., Stumptner, M. & Zanker, M. 2003. Configuration knowledge representations for Semantic Web applications. *Ai Edam,* 17**,** 31-50.
Fernández-López, M., Gómez-Pérez, A. & Juristo, N. 1997. Methontology: from ontological art towards ontological engineering.
Friendly, M. 2008. A brief history of data visualization. *Handbook of data visualization*. Springer.
Gantz, J. & Reinsel, D. 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future,* 2007**,** 1-16.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F. & Tu, S. W. 2003. The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies,* 58**,** 89-123.

Glier, M. W., Mcadams, D. A. & Linsey, J. S. 2014. Exploring automated text classification to improve keyword corpus search results for bioinspired design. *Journal of Mechanical Design,* 136**,** 111103.

Google. 2017. *How Google Search Works* [Online]. Available: https://www.google.com/intl/ALL/search/howsearchworks/crawling-indexing/ [Accessed April 03, 2018].

Gorti, S. R., Gupta, A., Kim, G. J., Sriram, R. D. & Wong, A. 1998. An object-oriented representation for product and design processes. *Computer-aided design,* 30**,** 489-501.

Grieco, A., Pacella, M. & Blaco, M. 2017. On the application of text clustering in Engineering Change process. *Procedia CIRP,* 62**,** 187-192.

Griffiths, T. L. & Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences,* 101**,** 5228-5235.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies,* 43**,** 907-928.

Grüninger, M. & Fox, M. S. 1995. Methodology for the design and evaluation of ontologies.

Guodong, Z., Jian, S., Jie, Z. & Min, Z. 2005. Exploring various knowledge in relation extraction. Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 427-434.

Guyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of machine learning research,* 3**,** 1157-1182.

Han, E.-H. S., Karypis, G. & Kumar, V. 2001. Text categorization using weight adjusted k-nearest neighbor classification. Pacific-asia conference on knowledge discovery and data mining. Springer, 53-65.

Han, J., Shi, F., Chen, L. & Childs, P. 2017. The Analogy Retriever–an idea generation tool. DS 87-4 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 4: Design Methods and Tools, Vancouver, Canada, 21-25.08. 2017.

Han, J., Shi, F., Chen, L. & Childs, P. R. 2018a. The Combinator–a computer-based tool for creative idea generation based on a simulation approach. *Design Science,* 4.

Han, J., Shi, F., Chen, L. & Childs, P. R. N. 2018b. A computational tool for creative idea generation based on analogical reasoning and ontology. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*.

Han, J., Shi, F. & Childs, P. 2016. The Combinator: A computer-based tool for idea generation. DS 84: Proceedings of the DESIGN 2016 14th International Design Conference.

Havre, S., Hetzler, E., Whitney, P. & Nowell, L. 2002. Themeriver: Visualizing thematic changes in large document collections. *IEEE transactions on visualization and computer graphics,* 8**,** 9-20.

Heer, J., Bostock, M. & Ogievetsky, V. 2010. A tour through the visualization zoo. *Queue,* 8**,** 20.

Holsapple, C. W. & Joshi, K. D. 2004. A formal knowledge management ontology: Conduct, activities, resources, and influences. *Journal of the Association for Information Science and Technology,* 55**,** 593-612.
Homer, G. R., Thompson, D. M. & Deacon, M. 2002. A distributed document management system. *Computing & Control Engineering Journal,* 13**,** 315-318.
Ifdesign. 2018. *iF WORLD DESIGN GUIDE* [Online]. Available: https://ifworlddesignguide.com/ [Accessed April 03, 2018].
Ishino, Y. & Jin, Y. 2001. Data mining for knowledge acquisition in engineering design. *Data mining for design and manufacturing*. Springer.
Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y. & Ramani, K. 2005. Shape-based searching for product lifecycle applications. *Computer-Aided Design,* 37**,** 1435-1446.
Jean, S., Aït-Ameur, Y. & Pierra, G. 2006. Querying ontology based database using ontoql (an ontology query language). OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 704-721.
Jiang, H., Kwong, C. & Yung, K. 2017. Predicting future importance of product features based on online customer reviews. *Journal of Mechanical Design,* 139**,** 111413.
Johnson, B. & Shneiderman, B. 1991. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. Proceedings of the 2nd conference on Visualization'91. IEEE Computer Society Press, 284-291.
Juršič, M., Sluban, B., Cestnik, B., Grčar, M. & Lavrač, N. 2012. Bridging concept identification for constructing information networks from text documents. *Bisociative Knowledge Discovery*. Springer.
Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. & Wu, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence,* 24**,** 881-892.
Kausar, M. A., Dhaka, V. & Singh, S. K. 2013. Web crawler: a review. *International Journal of Computer Applications,* 63.
Keim, D. A. 2000. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on visualization and computer graphics,* 6**,** 59-78.
Keim, D. A. 2002. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics,* 8**,** 1-8.
Kim, H., Liu, Y., Wang, Y. & Wang, C. 2016. Special issue: Data-driven design (d3). *Journal of Mechanical Design,* 138**,** 128002.
Kim, H. H. M., Liu, Y., Wang, C. C. L. & Wang, Y. 2017. Special Issue: Data-Driven Design (D3). *Journal of Mechanical Design,* 139**,** 110301-110301-3.
Koestler, A. 1964. The act of creation.
Kornfein, M. M. & Goldfarb, H. 2007. A Comparison of Classification Techniques for Technical Text Passages. World congress on engineering. 1072-1075.
Kreuseler, M., Lopez, N. & Schumann, H. 2000. A scalable framework for information visualization. Information Visualization, 2000. InfoVis 2000. IEEE Symposium on. IEEE, 27-36.
Lai, S., Xu, L., Liu, K. & Zhao, J. 2015. Recurrent Convolutional Neural Networks for Text Classification. AAAI. 2267-2273.
Lan, L., Liu, Y. & Lu, W. F. 2016. Discovering a hierarchical design process model using text mining. ASME 2016 International Design Engineering Technical

Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, V01BT02A024-V01BT02A024.
Lan, L., Liu, Y. & Lu, W. F. 2017. Automatic discovery of design task structure using deep belief nets. *Journal of Computing and Information Science in Engineering,* 17**,** 041001.
Larson, R. R. 2010. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology,* 61**,** 852-853.
Li, Z., Liu, M., Anderson, D. C. & Ramani, K. 2005. Semantics-based design knowledge annotation and retrieval. ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, 799-808.
Li, Z. & Ramani, K. 2007. Ontology-based design information extraction and retrieval. *Ai Edam,* 21**,** 137-154.
Li, Z., Raskin, V. & Ramani, K. 2008. Developing engineering ontology for information retrieval. *Journal of Computing and Information Science in Engineering,* 8**,** 011003.
Liang, Y., Liu, Y., Kwong, C. K. & Lee, W. B. 2012. Learning the "Whys": Discovering design rationale using text mining—An algorithm perspective. *Computer-Aided Design,* 44**,** 916-930.
Liang, Y. & Tan, R. 2007. A Text-Mining-based Patent Analysis in Product Innovative Process. Boston, MA. Springer US, 89-96.
Liew, A. 2007. Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice,* 8**,** 1-16.
Lim, S. & Tucker, C. S. 2016. A bayesian sampling method for product feature extraction from large-scale textual data. *Journal of Mechanical Design,* 138**,** 061403.
Lim, S. & Tucker, C. S. 2017. Mitigating Online Product Rating Biases Through the Discovery of Optimistic, Pessimistic, and Realistic Reviewers. *Journal of Mechanical Design,* 139**,** 111409.
Lim, S. C. J., Liu, Y. & Lee, W. B. 2010. Multi-facet product information search and retrieval using semantically annotated product family ontology. *Information Processing & Management,* 46**,** 479-493.
Lim, S. C. J., Liu, Y. & Lee, W. B. 2011. A methodology for building a semantically annotated multi-faceted ontology for product family modelling. *Advanced Engineering Informatics,* 25**,** 147-161.
Lin, H.-K., Harding*, J. A. & Shahbaz, M. 2004. Manufacturing system engineering ontology for semantic interoperability across extended project teams. *International journal of production research,* 42**,** 5099-5118.
Linden, G., Smith, B. & York, J. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing,* 7**,** 76-80.
Liu, B. & Zhang, L. 2012. A survey of opinion mining and sentiment analysis. *Mining text data*. Springer.
Liu, Y., Lim, S. C. J. & Lee, W. B. 2013. Product family design through ontology-based faceted component analysis, selection, and optimization. *Journal of Mechanical Design,* 135**,** 081007.
Loraine Charlet, A. & Kumar, A. 2012. Market basket analysis for a supermarket based on frequent itemset mining.
Luminoso. 2017. *"ConceptNet"* [Online]. Luminoso, Cambridge, MA. Available: http://conceptnet.io/ [Accessed Apr 21, 2017].

Luo, J., Yan, B. & Wood, K. 2017. InnoGPS for Data-Driven Exploration of Design Opportunities and Directions: The Case of Google Driverless Car Project. *Journal of Mechanical Design,* 139**,** 111416.
Ma, J. & Kim, H. M. 2014. Continuous preference trend mining for optimal product design with multiple profit cycles. *Journal of Mechanical Design,* 136**,** 061002.
Malekzadeh, M., Clegg, R. G. & Haddadi, H. 2018. Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis. Internet-of-Things Design and Implementation (IoTDI), 2018 IEEE/ACM Third International Conference on. IEEE, 165-176.
Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & Mcclosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 55-60.
Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. & Gómez-Berbís, J. M. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces,* 35**,** 482-489.
Mars, N. J. 1995. *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing 1995*, Ios Press.
Mccallum, A. & Nigam, K. 1998. A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization. Citeseer, 41-48.
Mccomb, C., Cagan, J. & Kotovsky, K. 2017. Mining process heuristics from designer action data via hidden markov models. *Journal of Mechanical Design,* 139**,** 111412.
Mcguinness, D. L. & Wright, J. R. 1998. Conceptual modelling for configuration: A description logic-based approach. *AI EDAM,* 12**,** 333-344.
Mcmahon, C., Lowe, A., Culley, S., Corderoy, M., Crossland, R., Shah, T. & Stewart, D. 2004. Waypoint: an integrated search and retrieval system for engineering documents. *Journal of Computing and Information Science in Engineering,* 4**,** 329-338.
Mena, E., Illarramendi, A., Kashyap, V. & Sheth, A. P. 2000. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and parallel Databases,* 8**,** 223-271.
Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM,* 38**,** 39-41.
Mitchell, T. M. 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill,* 45**,** 870-877.
Miwa, M. & Bansal, M. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.
Mukhopadhyay, D., Biswas, A. & Sinha, S. 2007. A new approach to design domain specific ontology based web crawler. Information Technology,(ICIT 2007). 10th International Conference on. IEEE, 289-291.
Munoz, D. & Tucker, C. S. 2016. Modeling the Semantic Structure of Textually Derived Learning Content and its Impact on Recipients' Response States. *Journal of Mechanical Design,* 138**,** 042001.
Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D. & Wood, K. 2014. Function based design-by-analogy: a functional vector approach to analogical search. *Journal of Mechanical Design,* 136**,** 101102.

Murtagh, F. 1983. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal,* 26**,** 354-359.
Nanda, J., Simpson, T. W., Kumara, S. R. & Shooter, S. B. 2006. A methodology for product family ontology development using formal concept analysis and web ontology language. *Journal of computing and information science in engineering,* 6**,** 103-113.
Nath, S. V. 2006. Crime pattern detection using data mining. Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on. IEEE, 41-44.
O'connor, M. & Das, A. 2009. SQWRL: a query language for OWL. Proceedings of the 6th International Conference on OWL: Experiences and Directions-Volume 529. CEUR-WS. org, 208-215.
Ohsawa, Y., Benson, N. E. & Yachida, M. 1998. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 12-18.
Pahl, G. & Beitz, W. 2013. *Engineering design: a systematic approach*, Springer Science & Business Media.
Pastor-Satorras, R., Vázquez, A. & Vespignani, A. 2001. Dynamical and correlation properties of the Internet. *Physical review letters,* 87**,** 258701.
Pei, J., Han, J. & Mao, R. 2000. Closet: An efficient algorithm for mining frequent closed itemsets. ACM SIGMOD workshop on research issues in data mining and knowledge discovery. 21-30.
Porter, M. F. 1980. An algorithm for suffix stripping. *Program,* 14**,** 130-137.
Princeton. 2010. *"About WordNet." WordNet* [Online]. Princeton University. Available: http://wordnet.princeton.edu [Accessed Apr 21, 2017].
Princetonuniversity. 2010. *"About WordNet." WordNet. Princeton University*. [Online]. Available: https://wordnet.princeton.edu/ [Accessed Apr 21, 2017].
Reddot. 2018. *Red Dot Design Award: Home* [Online]. Available: https://en.red-dot.org/ [Accessed April 03, 2018].
Rezgui, Y., Boddy, S., Wetherill, M. & Cooper, G. 2011. Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction? *Computer-Aided Design,* 43**,** 502-515.
Rink, B., Harabagiu, S. & Roberts, K. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association,* 18**,** 594-600.
Saif, H., Fernández, M., He, Y. & Alani, H. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter.
Salton, G. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*.
Salton, G. & Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management,* 24**,** 513-523.
Salton, G. & Harman, D. 2003. *Information retrieval*, John Wiley and Sons Ltd.
Salton, G., Wong, A. & Yang, C.-S. 1975. A vector space model for automatic indexing. *Communications of the ACM,* 18**,** 613-620.
Sangelkar, S. & Mcadams, D. A. 2012. Adapting ada architectural design knowledge for universal product design using association rule mining: A function based approach. *Journal of Mechanical Design,* 134**,** 071003.
Schulten, E., Akkermans, H., Botquin, G., Dörr, M., Guarino, N., Lopes, N. & Sadeh, N. 2001. The e-commerce product classification challenge. *IEEE Intelligent systems,* 16**,** 86-89.

Schumann, H. & Müller, W. 2013. *Visualisierung: Grundlagen und allgemeine Methoden*, Springer-Verlag.
Scrapinghub. 2017. *Scrapy | A Fast and Powerful Scraping and Web Crawling Framework* [Online]. Available: https://scrapy.org/ [Accessed April 03, 2018].
Sereda, P., Bartroli, A. V., Serlie, I. W. & Gerritsen, F. A. 2006. Visualization of boundaries in volumetric data sets using LH histograms. *IEEE Transactions on Visualization and Computer Graphics,* 12**,** 208-218.
Serrano, M. Á., Boguná, M. & Vespignani, A. 2009. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences,* 106**,** 6483-6488.
Settles, B. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, 104-107.
Shen, D., Yang, Q., Sun, J.-T. & Chen, Z. 2006. Thread detection in dynamic text message streams. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 35-42.
Shi, F., Chen, L., Han, J. & Childs, P. 2017a. A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *Journal of Mechanical Design,* 139**,** 111402.
Shi, F., Chen, L., Han, J. & Childs, P. 2017b. Implicit Knowledge Discovery in Design Semantic Network by Applying Pythagorean Means on Shortest Path Searching. ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, V001T02A053-V001T02A053.
Shi, F., Han, J. & Childs, P. 2016. A Data Mining Approach to assist design knowledge retrieval based on keyword associations. DS 84: Proceedings of the DESIGN 2016 14th International Design Conference.
Shneiderman, B. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG),* 11**,** 92-99.
Shneiderman, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations. Visual Languages, 1996. Proceedings., IEEE Symposium on. IEEE, 336-343.
Silva, C. & Ribeiro, B. 2003. The importance of stop word removal on recall values in text categorization. Neural Networks, 2003. Proceedings of the International Joint Conference on. IEEE, 1661-1666.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V. & Lanctot, M. 2016. Mastering the game of Go with deep neural networks and tree search. *nature,* 529**,** 484-489.
Singh, V. K., Piryani, R., Uddin, A. & Waila, P. 2013. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on. IEEE, 712-717.
Sintek, M. & Decker, S. 2002. TRIPLE—A query, inference, and transformation language for the semantic web. International Semantic Web Conference. Springer, 364-378.

Socher, R., Huval, B., Manning, C. D. & Ng, A. Y. 2012. Semantic compositionality through recursive matrix-vector spaces. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 1201-1211.
Soininen, T., Tiihonen, J., Männistö, T. & Sulonen, R. 1998. Towards a general ontology of configuration. *Ai Edam,* 12**,** 357-372.
Song, B. & Luo, J. 2017. Mining patent precedents for data-driven design: the case of spherical rolling robots. *Journal of Mechanical Design,* 139**,** 111420.
Spasic, I., Ananiadou, S., Mcnaught, J. & Kumar, A. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics,* 6**,** 239-251.
Speer, R., Chin, J. & Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. AAAI. 4444-4451.
Speer, R. & Havasi, C. 2012. Representing General Relational Knowledge in ConceptNet 5. LREC. 3679-3686.
Spence, R. 2001. *Information visualization*, Springer.
Stolte, C., Tang, D. & Hanrahan, P. 2002. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics,* 8**,** 52-65.
Sun, A. & Grishman, R. 2012. Active learning for relation type extension with local and global data views. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 1105-1112.
Thompson, K. 1968. Programming techniques: Regular expression search algorithm. *Communications of the ACM,* 11**,** 419-422.
Tous, R. & Delgado, J. 2006. A vector space model for semantic similarity calculation and OWL ontology alignment. International Conference on Database and Expert Systems Applications. Springer, 307-316.
Tuarob, S. & Tucker, C. S. 2015. Automated discovery of lead users and latent product features by mining large scale social media networks. *Journal of Mechanical Design,* 137**,** 071402.
Ullman, D. G. 2002. *The mechanical design process*, McGraw-Hill Science/Engineering/Math.
Ur-Rahman, N. & Harding, J. A. 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications,* 39**,** 4729-4739.
Uschold, M. & King, M. 1995. Towards a methodology for building ontologies.
Wang, X. & Qian, X. 2017. A Taylor expansion approach for computing structural performance variation from population-based shape data. *Journal of Mechanical Design,* 139**,** 111411.
Wang, Z., Childs, P. & Jiang, P. 2013. Using web crawler technology to support design-related web information collection in idea generation. DS 75-6: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 6: Design Information and Knowledge, Seoul, Korea, 19-22.08. 2013.
Ward, M. O. 1994. Xmdvtool: Integrating multiple methods for visualizing multivariate data. Proceedings of the Conference on Visualization'94. IEEE Computer Society Press, 326-333.
Ware, C. 2012. *Information visualization: perception for design*, Elsevier.

Webster, J. J. & Kit, C. 1992. Tokenization as the initial phase in NLP. Proceedings of the 14th conference on Computational linguistics-Volume 4. Association for Computational Linguistics, 1106-1110.
Weinreich, H., Obendorf, H., Herder, E. & Mayer, M. 2008. Not quite the average: An empirical study of Web use. *ACM Transactions on the Web (TWEB),* 2**,** 5.
Weston, J. & Watkins, C. 1998. Multi-class support vector machines. Citeseer.
Wikipedia. 2017. *Wikipedia, the free encyclopedia* [Online]. Available: https://en.wikipedia.org/wiki/Main_Page [Accessed Apr 23, 2017].
Wiles, A. 1995. Modular elliptic curves and Fermat's last theorem. *Annals of mathematics,* 141**,** 443-551.
Witherell, P., Krishnamurty, S. & Grosse, I. R. 2007. Ontologies for supporting engineering design optimization. *Journal of Computing and Information Science in Engineering,* 7**,** 141-150.
Wyner, A., Mochales-Palau, R., Moens, M.-F. & Milward, D. 2010. Approaches to text mining arguments from legal cases. *Semantic processing of legal texts.* Springer.
Yamada, T. 2018. *Yanko Design | Modern Industrial Design News* [Online]. Available: http://www.yankodesign.com/ [Accessed April 03, 2018].
Yang, D., Miao, R., Wu, H. & Zhou, Y. 2009. Product configuration knowledge modeling using ontology web language. *Expert Systems with Applications,* 36**,** 4399-4411.
Yang, M. C., Wood, W. H. & Cutkosky, M. R. 2005. Design information retrieval: a thesauri-based approach for reuse of informal design information. *Engineering with computers,* 21**,** 177-192.
Yu, L., Wang, S. & Lai, K. 2005. A rough-set-refined text mining approach for crude oil market tendency forecasting. *International Journal of Knowledge and Systems Sciences,* 2**,** 33-46.
Yu, W.-D. & Hsu, J.-Y. 2013. Content-based text mining technique for retrieval of CAD documents. *Automation in Construction,* 31**,** 65-74.
Yuvarani, M. & Kannan, A. 2006. LSCrawler: a framework for an enhanced focused web crawler based on link semantics. Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on. IEEE, 794-800.
Zaki, M. J. & Hsiao, C.-J. 2002. CHARM: An efficient algorithm for closed itemset mining. Proceedings of the 2002 SIAM international conference on data mining. SIAM, 457-473.
Zaki, M. J., Meira Jr, W. & Meira, W. 2014. *Data mining and analysis: fundamental concepts and algorithms*, Cambridge University Press.
Zhang, L., Chu, X., Chen, H. & Xue, D. 2017. Identification of performance requirements for design of smartphones based on analysis of the collected operating data. *Journal of Mechanical Design,* 139**,** 111418.
Zhang, W. & Yin, J. 2008. Exploring Semantic Web technologies for ontology-based modeling in collaborative engineering design. *The International Journal of Advanced Manufacturing Technology,* 36**,** 833-843.
Zhang, X., Fuehres, H. & Gloor, P. A. 2011. Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences,* 26**,** 55-62.
Zhang, X., She, J., Li, S., Duan, S., Zhou, Y., Yu, X., Zheng, R. & Zhang, B. 2015. Simulation on deforming progress and stress evolution during laser shock forming with finite element method. *Journal of Materials Processing Technology,* 220**,** 27-35.

Zheng, S., Xu, J., Zhou, P., Bao, H., Qi, Z. & Xu, B. 2016. A neural network framework for relation extraction: Learning entity semantic and relation pattern. *Knowledge-Based Systems,* 114**,** 12-23.
Zhou, K., Fu, C. & Yang, S. 2016. Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews,* 56**,** 215-225.
Zins, C. 2007. Conceptual approaches for defining data, information, and knowledge. *Journal of the Association for Information Science and Technology,* 58**,** 479-493.

Appendices

Permission from Design Society

July 16, 2018

Dear Design Society,

I am completing my PhD thesis at Imperial College London entitled 'Data-Driven Design by Text Mining and Semantic Network for Design Information Retrieval'.

I seek your permission to reprint, in my thesis, the published version of three papers I published in DESIGN 2016 14th and DESIGN 2018 15th International Design Conference:

Shi, F., Han, J. and Childs, P.R.N., 2016. A Data Mining Approach to assist design knowledge retrieval based on keyword associations. In *DS 84: Proceedings of the DESIGN 2016 14th International Design Conference*.

Han, J., **Shi, F.** and Childs, P.R.N., 2016. The Combinator: A computer-based tool for idea generation. In *DS 84: Proceedings of the DESIGN 2016 14th International Design Conference*.

Chen, L., Wang, P., **Shi, F.**, Han, J. and Childs, P., 2018. A COMPUTATIONAL APPROACH FOR COMBINATIONAL CREATIVITY IN DESIGN. In *DS92: Proceedings of the DESIGN 2018 15th International Design Conference* (pp. 1815-1824).

I would like to include the paper in my thesis which will be added to Spiral, Imperial's institutional repository http://spiral.imperial.ac.uk/ and made available to the public under a Creative Commons Attribution-NonCommercial-NoDerivs licence.

If you are happy to grant me all the permissions requested, please return a signed copy of this letter. If you wish to grant only some of the permissions requested, please list these and then sign.

Yours sincerely,

Feng Shi

Permission granted for the use requested above:

I confirm that I am the copyright holder of the paper above and hereby give permission to include it in your thesis to be made available, via the internet, for non-commercial purposes under the terms of the user licence.

Signed:

Name: Ross Brisco

Organisation: The Design Society

Job title: Administrator

Permission from ASME

FW: Reprint published paper in my dissertation

Beth Darchi <DarchiB@asme.org>

Wed 18/07/2018 15:25

To: Shi, Feng <f.shi14@imperial.ac.uk>;

Dear Prof. Shi,
This permission has been revised to reflect all request. It is our pleasure to grant you permission to use **all or any part of** the following ASME papers:

- A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval, by Feng Shi; Liuqing Chen; Ji Han; Peter Childs, J. Mech. Des. 2017; 139(11)
- Implicit Knowledge Discovery in Design Semantic Network by Applying Pythagorean Means on Shortest Path Searching, by Feng Shi; Liuqing Chen; Ji Han; Peter Childs, Paper Number DETC2017-67230
- A Network-Based Computational Model for Creative Knowledge Discovery Bridging Human-Computer Interaction and Data Mining, by Liuqing Chen, Feng Shi, Ji Han and Peter R. N. Childs, Paper No. DETC2017-67228

cited in your letter for inclusion in a PhD dissertation to be published by Imperial College London.

Permission is granted for the specific use as stated herein and does not permit further use of the materials without proper authorization. Proper attribution must be made to the author(s) of the materials. **Please note:** if any or all of the figures and/or Tables are of another source, permission should be granted from that outside source or include the reference of the original source. ASME does not grant permission for outside source material that may be referenced in the ASME works.

As is customary, we request that you ensure full acknowledgment of this material, the author(s), source and ASME as original publisher. Acknowledgment must be retained on all pages where figure is printed and distributed.

As a note, ASME does not sign any forms for permission. Please accept this letter as proof.

Many thanks for your interest in ASME publications.

Sincerely,

Beth Darchi
Publishing Administrator
ASME
2 Park Avenue
New York, NY 10016-5990
darchib@asme.org