Mining insights out of workflow data analysis

Feiyi Su

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 25.08.2020

Thesis supervisor:

Prof. Hämäläinen Perttu

Thesis advisor:

Valtteri Halla

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Feiyi Su

Title: Mining insights out of workflow data analysis

Date: 25.08.2020 Language: English Number of pages: 6+55

Department of Computer Science

Professorship: Human-Computer Interaction and Design

Supervisor: Prof. Hämäläinen Perttu

Advisor: Valtteri Halla

Customer support services such as IT helpdesk have to answer and fix a variety of issues for their customers. Implementing a highly efficient IT helpdesk is critical for an organization's daily operations. With an increasing number of issues, the IT helpdesk suffers from an overwhelming workload. However, non-transparency of workload and quality makes it hard to evaluate the performance of the IT helpdesk and the quality of services, affecting work productivity and service satisfaction. Thus, improving IT helpdesk operational management for higher issue resolution productivity is becoming necessary. This thesis is focused on mining insights from Futurice IT helpdesk workflow data and providing AI and data solutions, for making its workflow follow lean production principles. The thesis research evaluates the current state of Futurice IT helpdesk, provides AI and data solutions through a design prototype and discusses possible future recommendations.

Keywords: knowledge management, IT helpdesk workflow, software engineering, operational management, text mining, data mining, data analysis

Preface

I want to thank Professor Perttu Hämäläinen and my Futurice advisor Valtteri Halla for their good guidance.

Otaniemi, 05.08.2020

Feiyi Su

Contents

Abbreviations

SaaS	Software as a service
HC	Human Centre
FCFS	First-come, First-served
FIFO	First-in, First-out
LCFS	Last-come, First-served
LIFO	Last-in, First-out
VSM	Value Stream Map
EDA	Exploratory Data Analysis
KDT	Knowledge Discovery in Text
TSNE	t-distributed Stochastic Neighbor Embedding (is an algorithm for dimensionality reduction)
PCA	Principal Component Analysis
Infra	Infrastructure
AI	Artificial Intelligence

1 Introduction

1.1 Motivation

IT helpdesk serves a critical role in the information technology department in organizations by resolving information problems including hardware, software, or networks, and so on. According to research by Gartner Group, the average amount of information technology supported by IT help desks has increased in the past five years (Sandborn, 2001). An important reason for the growth is the spread and distribution of information technology, such as different personal computers, software applications, printers and servers in organizations. In addition, some studies (Marcella and Middleton, 1996; Sharer, 1998) have found that the more fragmented information technology is, the more support end-users need. As a result, the IT helpdesk has to support more and more information technology issues and meanwhile face an increasing workload.

Futurice, as one of the most successful software consulting companies in Finland, has a busy IT helpdesk, which provides strong support to ensure that Futurice providing timely, thoughtful, and upbeat service to its customers. However, overwhelming customer demands, rising operational costs, and increasing workload are pushing Futurice IT helpdesk to rethink saving IT operational costs meanwhile increasing productivity and service satisfaction.

Therefore, lean production is becoming the compass of Futurice's direction. Lean manufacturing evolves from the Toyota Production System over decades and has become a systematic production method of improving efficiency and performance through eliminating waste in the process of production, and creating economic values for the end consumers (Holweg, 2007; Shah and Ward, 2007; Čiarnienė and Vienažindienė, 2012). Taiichi Ohno, chief architect of Toyota's production systems, highlights the "seven wastes" that manufacturing operations need to strive to eliminate: overproduction, unnecessary transportation, inventory and worker movement, detection, overprocessing and waiting (Staats, et al., 2011).

Therefore, eliminating waste in IT helpdesk workflow and converting IT helpdesk workflow to lean production is an important goal that Futurice IT helpdesk pursues to achieve in its operational management.

1.2 Business problem

At present, the emerging business problem at Futurice IT helpdesk workflow is that there is no transparency in its workload and quality. The invisibility makes it hard to have a holistic picture of the situation with the workflows and the team finds it difficult to determine what to automate or outsource to keep its quality. Because of that, uncertain lead time of issue resolution and uncertain workload results in a problem that the IT helpdesk workers always get stuck during their work and have to struggle with the increasingly heavy workload.

1.3 Business objectives

For IT helpdesk workers, this thesis is beneficial for them to fully understand the insights of workflows. They can understand which tasks to be prioritized or decisions to be made. It will improve employees' health and satisfaction in the team and at the same time increase the satisfaction of end-users who receive assistance. Better customer service reflects positively on the business and company in the long term.

For each Futurice employee, this thesis is meaningful for having a faster, more predictable, more efficient, and happier IT helpdesk.

For Futurice, this thesis helps to set the optimal investment level for IT helpdesk operation in order not to lose more money through delays in IT support and infra than it costs to serve those needs. It might result in saving more money and resources and meanwhile, gaining more revenue from customer services.

For a wider market side, this thesis research hopes to turn IT helpdesk from a cost center to driving profits.

1.4 Research problems and questions

This thesis is a part of Futurice AI Exponential research, seeking a breakthrough in workflow automation and exploring how people can make wiser or more data-advised decisions by data analysis results.

With the help of data analysis and natural language processing and other AI techniques, the objective is to find automated ways to pursue an innovative transformation of IT helpdesk workflow to be a more lean production operation and also improve its knowledge management. Therefore, the research problem is stated as below:

Research problem: How to facilitate better operational management and knowledge work at IT helpdesk?

Accordingly, three research questions are defined as following:

RQ1: *How can workflow data analysis help to improve productivity (issue resolution) at the IT helpdesk?*

RQ2: *How can workflow data analysis help to improve operational management and knowledge work at the IT helpdesk?*

RQ3: *How can lean production be applied to improve operational management and knowledge work at the IT helpdesk?*

1.5 Theoretical and practical implications

For theoretical implications, this research is trying to convert theoretical knowledge into industrial-based practical use, especially the workflow practice. It is interesting to see that in the previous research, there have been some known good theories and philosophies for managing operations and knowledge management (Reinertsen, 1997). However, it is rare to see systematic research about dealing with a practical issue in an IT helpdesk workflow scenario.

For practical implications, this research is trying to produce a concrete tool to generate behavior changes in IT helpdesk workers. The tool could generate a general framework or application that can be *"transplanted"* from the IT helpdesk application into similar operations or scenarios such as client projects, Github issues, etc, thus optimizing similar workflows.

Overall, the meaning of this research is to seek new possibilities of practices by reaching the overlapping area (Fig. 1) of Futurice business goals, user's needs, and academic research (a combination of three fields).

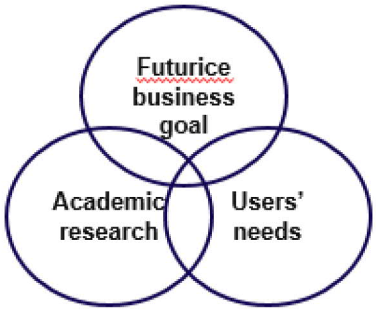


Figure 1: A combination of three fields

1.6 Scope and focus

This thesis will address problems for knowledge work, which are affecting operations at the Futurice IT helpdesk. The target user of this research are stakeholders of Futurice IT helpdesk. Stakeholders include IT helpdesk team members and customers who received assistance.

Essentially, the scope and focus are on how workflow data analysis (*how*) can be used to improve productivity and make knowledge flow faster (*why*), by (*scope and focus*) finding ways of achieving IT helpdesk lean production workflow, and then generating a data-driven tool for Futurice IT helpdesk (*what*).

The possibilities of ideas are implemented and validated partially at Futurice. However, a concrete implementation is not included in this thesis. The thesis mainly focuses on service design. This thesis mainly evaluates concepts and provides insights and recommendations for improvements.

1.7 Structure of the thesis

The rest of this thesis comprises the following main parts:

- **Case Background:** Case background analysis and identification of primary challenges (Section 2)
- **Theoretical background and literature review:** Review of relevant theoretical and methodological background (Section 3)

- **Data analysis and visualization:** Exploratory research using data analysis and visualization techniques to better understand the problems and gain insights into possible solutions, discussing and reflecting on the results (Section 4)
- **Prototype design and evaluation:** Prototyping and testing a candidate solution (Section 5)
- **Discussion and recommendation:** Discussing and proposing possible future recommendations for building a better IT helpdesk (Section 6)
- **Summary:** Summarizing the thesis research results and their impact and importance (Section 7)

2 Case Background

2.1 Organization

Futurice is a software consulting company, which started its business in 2000 in Helsinki. In the year 2020, Futurice has offices around Europe in Helsinki, Tampere, Berlin, Munich, Stockholm, Oslo, and London. The company comprises over 600 employees and its businesses are growing fast and spread over Europe. Futurice AI Exponential, as an internal initiative in Futurice, devotes itself to breaking traditional thoughts and implements AI transformation within the organization. It defines the future of knowledge work with the help of AI and data since artificial intelligence is now becoming the new wave of transformation to sweep across our lives and way of working. The initiative is focusing on helping the organization optimize knowledge management and achieving automation of the methods of doing work.

2.2 IT helpdesk at Futurice

Futurice IT helpdesk is responsible for resolving issues (usually called a ticket) of information technologies within the organization. Currently, there are 9 IT helpdesk workers, with all team members located at the Helsinki office. These 9 IT workers need to serve all of the over 600 Futurice employees. The average workload is on average more than 1000 tickets per month. The interaction between customers (requests) and IT workers (replies) are usually communicated by email or web form inside the ZenDesk system. The role of an operator on a ZenDesk is often called an IT agent. Besides, there are physical internal help desks that aim at offering help but only for the employees who are located in the Helsinki office.

2.3 IT helpdesk workflow application at Futurice

Specifically, Futurice IT helpdesk workflow depends on the ZenDesk platform where tickets are processed and manipulated. ZenDesk is a SaaS solution currently used by Futurice IT support and HC (Human Center) teams together. It can keep track of tickets where the state of tickets will change accordingly when manipulations have been done on them. Thus, IT helpdesk workflow can be traced by ticket states and also discussions between the customers (requesters) and IT agents (repliers).

An overview of the ticket lifecycle in ZenDesk is depicted in Figure 2. The entire workflow of Futurice IT helpdesk can be divided into sub-flows from timestamps and states. If treating workflow in ZenDesk as a ***"queue"***, one ticket experiences queuing time plus processing time, constituting its lead time.

Figure 2 also displays operations and corresponding states of customer experience and agent experience. The state is an important indicator of the workflow, which will be used frequently in the next data analysis part. There are six main states of tickets, separating unsolved and solved issues:

Unsolved issues:

New: a ticket which is coming into the system from various channels, but has not been assigned to any IT agent.

Open: when a new ticket is assigned to an IT agent, the state changes to open.

Pending: a ticket that is waiting for the response from customers.

On-hold: a ticket that is waiting for the support of third parties or external help.

Solved issues:

Solved: a ticket that is resolved and no more replies from customers.

Reopen: a ticket that is resolved and new replies from customers within 4 days.

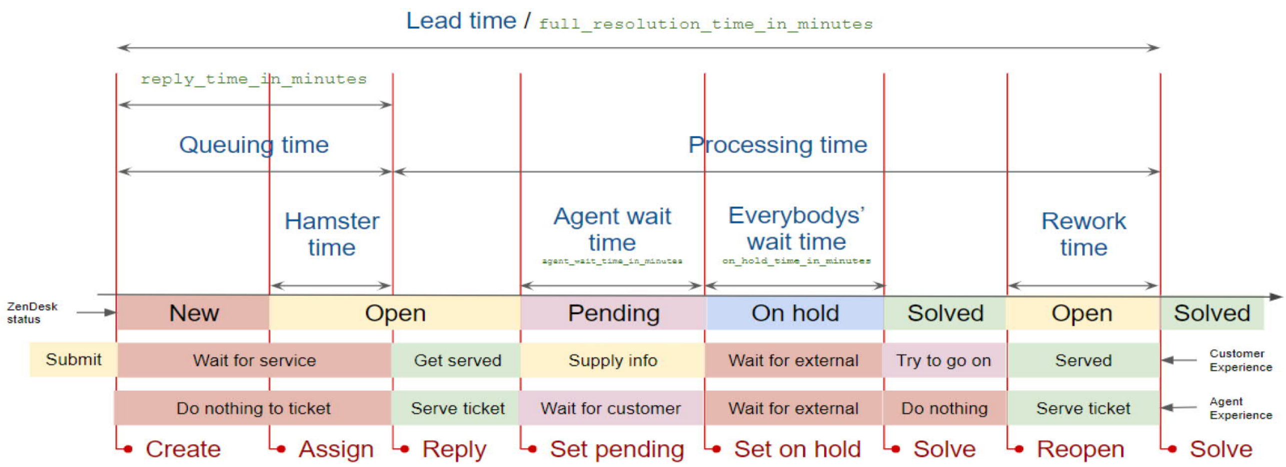


Figure 2: Complicated overview of ticket lifecycle

Simplifying the above lifecycle graph, Figure 3 highlights the waiting time from both customer experience and IT agent experience.

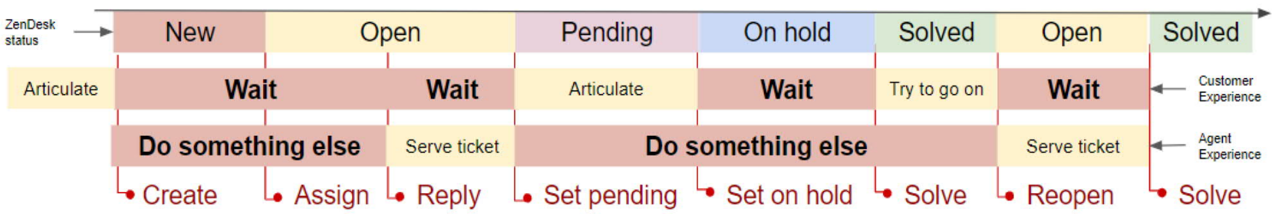


Figure 3: Waiting time highlights from customer experience and agent experience

From a customer's viewpoint, one experiences the entire lead time for having his/her problem solved. The customer mostly needs to wait for one's turn in queuing time and also needs to wait for issue completion or supplement information requests. Thus, a customer is to do other work to fill-in waiting times. From an IT agent's viewpoint, one begins his/her journey when assigned an issue. The agent mostly needs to wait for inputs from the customer's side or external support. Thus, the agent turns to serve other tickets to fill in waiting times. The waiting time from both sides produces time to waste.

The processing structure from an agent view can be further depicted as a macro-structure (Fig. 4).

For the whole of IT helpdesk workflow, IT agents are trying to shorten waiting time in the queue by reducing processing time as much as possible to improve customer experience.

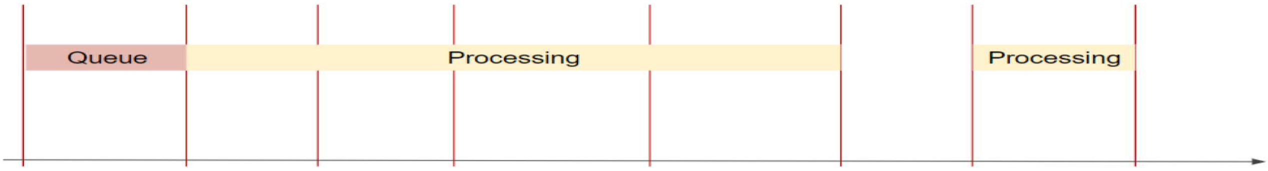


Figure 4: Macro-structure of processing from IT agent's view

2.4 Current knowledge flow at Futurice IT helpdesk

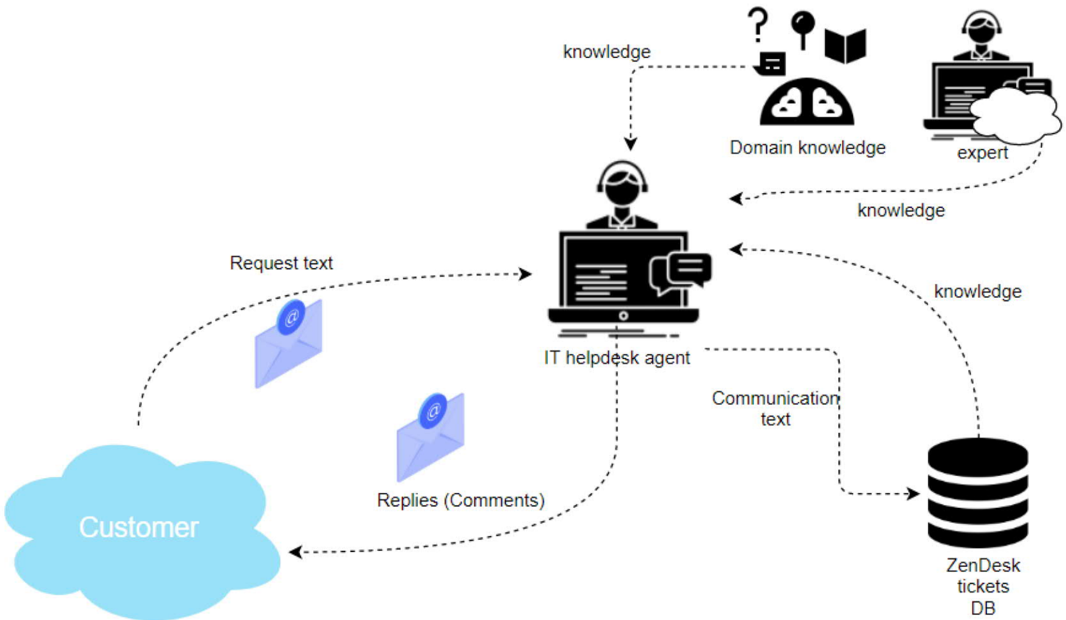


Figure 5: Information flow at Futurice IT helpdesk

The knowledge flow graph provides a holistic picture of how information transmits within the IT helpdesk workflow (Fig. 5). When an issue ticket from customers goes into the IT helpdesk, the agent is trying to resolve the ticket through accessing different knowledge sources. However, when facing repeated issues, the agent can not keep track of others experience of fixing similar issues, affecting their efficiency of solving issues.

2.5 Challenges

Futurice IT workers have to face over 1000 tickets per month and increasing issues (queues). The non-transparency of work queues and tasks causes challenges that urgently need to be addressed. To sum up, three top existing challenges at IT helpdesk from the employee's view, the organization's view, and the customer's view are:

1. From the IT helpdesk worker's view, a massive accumulation of unresolved issues and repeated issues in queues plus long processing time increases workload, resulting in low work efficiency and low satisfaction of employees.

2. From the company's view, delay in solving unresolved queues results in waste and loss of resources in the long term.

3. From the customer's view, long lead time leads to low satisfaction with customers.

The above challenges highlight the importance of this thesis research. Ultimately, employees are physically and mentally exhausted, resulting in being unable to ensure efficient work, further causing more accumulation in queues and this vicious cycle leads to both low employee and customer satisfaction. A large number of unsolved issues will also cause losses to the company and impact overall customer service.

Therefore, it is important to manage the increasing *"queues"*. Next, this thesis will explore the way of controlling *"queues"*, for increasing productivity meanwhile eliminating time to waste.

3 Theoretical background and literature review

Google Scholar was used primarily when searching for scientific literature. Most of the literature was found in IEEE Xplore Digital Library, Science Direct, and Wikipedia. Other articles were found from the recommendations of Google Scholar. The most frequent search terms were used separately or together with various combinations including lean manufacturing, knowledge management, queuing theory, natural language processing, machine learning, data mining, and text mining.

The search timeline of the literature review was limited from the beginning of the 1960s to the current year. Known theories and frameworks for managing operations and *"queues"* were selected such as lean movement, followed by the Theory of Constraints in the 1980s, and Reinertsen's queuing theory. Then, knowledge management and natural language processing related theoretical literature was also reviewed. These theories were used as theoretical support in the following sections.

3.1 Known theories and philosophies for managing operations

This part of the literature review is about different known theories and philosophies from 60's to 90's for managing operations that could be applied to IT helpdesk operations. Since the IT helpdesk service could be regarded as a ***"queue"*** service, relevant methods were also investigated. Literature of converting knowledge work such as IT helpdesk workflow to lean production was also reviewed.

3.1.1 Queuing theory (60's)

The queuing theory was proposed in about 1920 by a statistician named A. K. Erlang. He took up the problem of congestion of telephone traffic. The early work was extended to other general problems involving queues after World War II (Tiwari et al., 2016). David G. Kendall (1953) proposes describing queuing models using three factors written A/S/c. A queue can be served as a way of first-come, first-served (FCFS) or first-in, first-out (FIFO) or last-come, first-served (LCFS) or last-in, first-out (LIFO).

3.1.2 Lean movement and philosophy (70's)

The lean movement has been flourishing in almost all industries since the 1970s. Lean manufacturing started in the auto industry as the Toyota Production System in the 1970s and 1980s. To date, the implementation of lean beyond the manufacturing or knowledge work has received relatively little academic attention. Although there is some research about applying lean to the field of IT service, there are still rare related practices and case studies in a pure-service environment such as IT helpdesk.

Lean knowledge work

Staats et al. (2011) mention three differences between manufacturing and knowledge work that need to be focused on. First, knowledge work is generally more dynamic than traditional manufacturing work. For example, customers tend to be more likely to change their initial requirements during the production process. Second, the inputs, processes, and even outputs in knowledge work are intangible in nature, while manufacturing processes are visible. Third, manufacturing work only uses repetitive knowledge and processes, while knowledge work involves higher levels of knowledge and new processes.

The ultimate goal of lean manufacturing is to achieve the highest quality at the lowest cost, which refers to a basic absence of waste during the workflow. The method of elimination of these wastes is needed to be estimated carefully, however, the waste of knowledge is largely invisible. Furthermore, because knowledge work is much more complex, other than non-knowledge work, it requires multitasking, multi-principles, innovation, and cross-functional collaboration (May, 2005). It is challenging to define the structure of knowledge work in general. Meanwhile, it is also a challenge to improve knowledge of work performance.

Early research of lean in knowledge work is very scarce. Most of the studies are qualitative research. Value stream mapping (VSM) is one of popular methods for eliminating waste (Tyagi et al., 2014). In types of knowledge work, VSM helps to make intangible work processes more concrete. VSM emphasizes knowledge projection and information flow in the process of product or service generation (McManus, 2015). Rachman and Ratnayake (2016) describe the importance of standardization. McDermott and Venditti (2015) and Mann (2009) also applied work standardization in knowledge work.The method of eliminating waste in knowledge work is different from the traditional lean production method. Kruger (2014) and McDermott and Venditti (2015) both showed that it was easier to identify waste when work was more standardized. Rachman and Ratnayake (2016) point out the importance of identifying waste to meet engineering service-related requirements. Rahani and Muhammad (2012) carried out product flow analysis through value stream map in the manufacturing field.

May (2005) suggests two important ideas that firstly, attempting to go lean requires a much more human-centered method other than simply applying neat algorithms or toolboxes. Secondly, he also explained that information is the value base in knowledge work. Information needs to flow to the right people at the right time, in the right form, at the lowest cost and with the highest quality.

Visualization of knowledge control and tracking is also mentioned in some cases in the research of Toussaint and Berry (2013), McDermott and Venditti (2015), and Staats and Upton (2011). Sacks et al. (2009) implement visualization of a workflow to support lean construction. However, practical cases of visualization remained unclear and unspecified. Instead, visualizations were usually implied to be some key performance indicators on whiteboards.

Niall and Nick (2009) find that significant improvements in quality and cost are achieved with minimal investment by adopting lean tools in a pure service environment. They have proven empirically the effectiveness of applying lean in a pure service environment.

3.1.3 Reinertsen's work on queues (90's)

Since the 1990s, one of the important theories from Reinertsen's research is a method of managing queues. Based on the principle of product development flow proposed by Reinertsen, a queue is a hidden source of waste in the development flow. He realizes that the lateness is caused by uncertainty in the arrival times and the task duration. Since product development is a one-time activity, it causes largely uncertainty in task arrival times and task durations. Thus, the queue would be definitely generated in product development processes (Reinertsen, 1997). Queues increase variations, cycle time, operating expenses, and risk. In a pure-service environment, queues increase workload, slow down feedback, decrease motivation, reduced the quality of services.

Reinertsen (1997) also extends concepts of queuing theory and its application to product development. A queuing system indicates the relationship between a queue and a server. The time spent in line is called queue time. Service time refers to the time spent on being served. The total time from arrival to departure is called time in the system. The structure of a queuing system differs from the number of queues and the number of servers. One-queue-per-server structure (Fig. 6) and single-queue-multiple-servers (Fig. 7) structure are different. Single-queue-multiple-servers (or often referred to as M/M/s server discipline) occurred in queued settings where there are one or more servers, the customer should arrive at a random speed to be specified as a Poisson distribution for a given time period, and the service time as an exponential distribution.

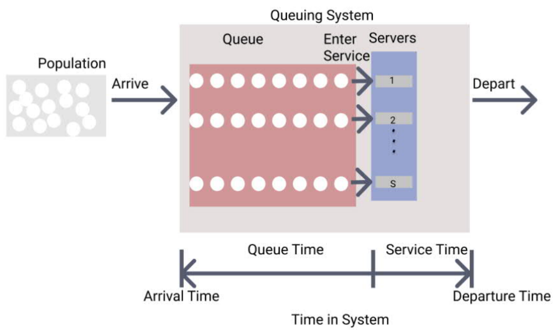


Figure 6: Queuing system: single-queue-per-servers structure

General feature of queuing theory

Taha (1987) also expresses three basic components of a queuing system. 1) the arrival (the entry of customers into the system), 2) the queue, and 3) the service (the customers get into the service). Tiwari et al. (2013) also mention that these three components that must be defined before queuing models can be mathematically

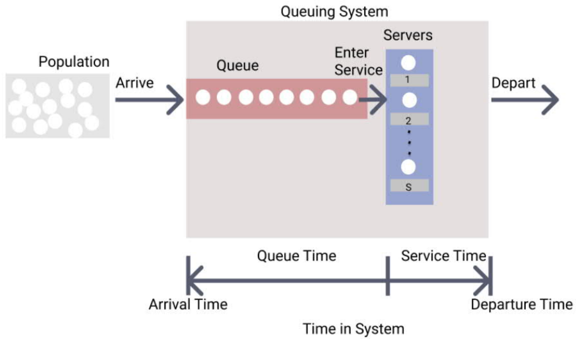


Figure 7: Queuing system: single-queue-multiple-servers structure

developed.

Queuing models assumption

Taha (1987) explains the most widely used and simplest queuing model: single-channel, single-phase model. It involves the following conditions.

1. Arrival service is on a basis of FIFO.
2. Each arrival customer has to wait for service, no matter how long the queue is.
3. Arrivals are independent and The average arrival rate does not change over time. Arrival is described in terms of Poisson probability. The sum of distributions comes from infinity or from a very large population.
4. Service time occurs in a negative exponential probability distribution. The average service rate is greater than the average arrival rate.
5. Service hours also vary by passengers. Their sum is independent of each other, but their average velocity is known.
6. The number of items in the queue at any given time and the number of queues experienced by a particular item are random variables.

Above conditions produce equation for queuing systems (Taha, 1987):

λ = mean number of arrivals per time period (For example, per day).

μ = mean number of customers being served per time period.

When regarding to the arrival rate (λ) and the services rate (μ), the time period must be the same, which means If the λ is the average number of arrivals per day, then unit of μ must be in day.

For One-queue-per-server structure (Fig. 6), the queuing equations are:

Formulas	Definition
$Ls = \frac{\lambda}{\mu-\lambda}$	The average number of issues in the system being served.
$Lq = \frac{\lambda^2}{\mu-\lambda}$	Average number of issues in the queue waiting to be served.
$\rho = \frac{\lambda}{\mu}$	The probability that the service server is being utilized.
$Wq = \frac{\lambda}{\mu(\mu-\lambda)}$	The average time a issue spends waiting for service.
$Ws = \frac{1}{\mu-\lambda}$	The average time a issue spends in the system (queue time plus service time).
$P_0 = 1 - \left(\frac{\lambda}{\mu}\right)$	Probability of agents to be idle.

For single-queue-multiple-servers structure (Fig. 7), the queuing equations are (Hui and Tao, 2000): If suppose:

s = the number of service servers available,

λ = average arrival rate,

mu = average service rate at each service server.

Formulas	Definition
$Ls = \frac{\lambda}{\mu-\lambda}$	The average number of customers in the system being served.
$Lq = \frac{\lambda^2}{\mu-\lambda}$	Average number of customers in the queue waiting to be served.
$\rho = \frac{\lambda}{s\mu}$	The probability that the service server is being utilized.
$Wq = \frac{\lambda}{\mu(\mu-\lambda)}$	The average time of a customer spends waiting for service.
$Wt = Wq * inventories$	The average waiting time.
$P_0 = 1 - \left(\frac{\lambda}{\mu}\right)$	Probability of agents to be idle.

Moreover, Reinertsen (1997) illustrates that cumulative flow diagrams and other simple graphical techniques are powerful tools for depicting queues and analyzing queues.

3.2 Knowledge management

This part of the literature review is about research on knowledge management. For knowledge work, it is important to use full of knowledge and improve knowledge management.

Since the middle of the 90s, knowledge management has been on the rise globally. Knowledge includes personal knowledge and business processes or practices within an organization. Knowledge management attaches importance to the integration and continuous improvement of knowledge within the organization (Newell, 2015; McAdam, 2000). For consulting companies, knowledge management is key because of their product: Service is knowledge (Ambos and Schlegelmilch, 2009). Therefore, organizations need active knowledge management and knowledge utilization strategy in different states (Gattnar et al., 2014).

Knowledge is the core of doing innovation. Knowledge management (KM) refers to the coordination and utilization of an organization's knowledge assets through a systematic process, which is to acquire, organize, maintain, apply, share, and update different forms of knowledge of employees, thereby improving organizational performance, creating value, increasing efficiency and improving competitive advantage (Davenport and Prusak, 2000). Additionally, Omer et al. (2016) emphasized it was important to understand where, how, what forms to gain knowledge. Some researchers like Davenport and Prusak (2000) argue that if data was going to become information, it must be contextualized, categorized, calculated, and compressed. Bali et al. (2009) describe information as relevant and purposeful data. Gamble and Blackwell (2001) propose a theoretical framework in the form of a KM matrix which is divided into four stages. First of all, an organization must determine the source of knowledge. This must be followed by organizational work in this area. This step will help assess strengths and weaknesses and determine the relevance and re-usability of the knowledge. Then, knowledge must be socialized, where various technologies were used to help to share and spread as needed. After that, trying to internalize that knowledge, and deciding what to take action on within the organization. Botha et al. (2008) propose a three-pronged model: 1) knowledge creation and sharing, 2) knowledge organization, and acquisition, 3) knowledge sharing and dissemination, with overlapping and interacting between them. Later, Frost (2010) proposes a general knowledge management framework, which includes discovering and detecting, organizing and evaluating, sharing, reusing, creating, and acquiring knowledge.

3.3 Natural Language Processing (NLP)

This part of the literature review is about methods for data analysis. Since IT helpdesk workflow data mainly contained email text, natural language processing methods were investigated for reading, deciphering, understanding, and making sense of those human languages.

3.3.1 Text mining

Text mining is also called data mining or text analysis or Knowledge Discovery in Text (KDT). It is the process of acquiring high-quality information from text. Text mining (Hotho et al., 2005) is considered as three different perspectives: text mining as information extraction, text mining as text data mining and text mining as a KDT process in the database.

Marti (2003) explains that text mining is where a computer automatically extracts information from different written resources to discover new, previously unknown information. Written resources can be in the form of books, reviews, emails and so on. Haralampos and Babis (2001) state that the problem of text mining is to extract explicit and implicit concepts by using natural language processing techniques. Its purpose is getting insights into large amounts of textual data. Text mining uses

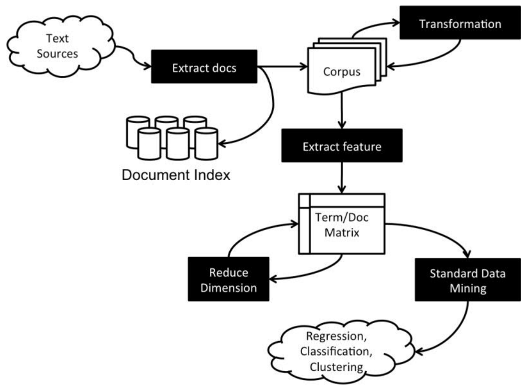


Figure 8: The workflow of text mining (Fan et al., 2005)

statistics, machine learning, information extraction, knowledge management, and others in its discovery process (Vishal and Gurpreet, 2009). A key factor is to link the extracted information together to form new facts or hypotheses that could then be explored further by more traditional experimental means. Figure 8 indicates details of the text mining process (Ho, 2014; Fan et al., 2005).

Essentially, the ultimate goal of text mining is to convert text into data for analysis through natural language processing and different types of analytical methods. Generally, after getting the results of text mining, the next step is doing steps of analysis, insights, and recommendations, finding out the relationship between key categories, identifying gaps and recommending key stakeholders valuable insights that could be used in business.

Typical text mining applications are text clustering, text categorization, entity extraction, sentiment analysis, document summarization, etc., which are now widely used for a variety of research and business needs (Ho, 2014; Fan et al., 2005).

3.3.2 Sentiment dictionary-based sentiment analysis

Sentiment analysis has become widely used to understand customers' emotions (Zhu et al., 2019). Sentiment analysis is a technology of text analysis technology to identify subjective opinions in text data (Loughran and McDonald, 2011). It usually involves dividing the text into categories such as *"positive"*, *"negative"* and in certain cases *"neutral"*. It can be seen a huge increase in the demand for emotional-analysis tools over the past years, organizations are willing to monitor what people say about their products and services (Loughran and McDonald, 2011).

Taboada et al. (2011) calculated the emotional scores of phrases, sentences, words and paragraphs through detailed rules, and used threshold method to judge emotional tendency. However, emotional analysis needs to consider context. Wu et al. (2017) constructed an emotional dictionary applicable to the shopping field. By sorting the TF-IDF of words and setting the threshold, the field emotion words are obtained to construct the emotion dictionary of shopping review emotion classification domain. Wu et al. (2017) extracted financial text features based on Apriori algorithm. They construct financial sentiment dictionaries and semantic rules to identify emotional units and forces.Then, the emotional orientation and emotional intensity of the text are deduced.

In this study, emotions can be analyzed from E-mail text, and emotion dictionary based analysis is an unsupervised classification method. The first step is to find out the emotional words, negative words, adverbs, conjunctions and so on in the sentence, and assign the weight to the emotional words. Then the emotional weight of the word is calculated according to the adverb of degree and the negative word. The emotional score of a sentence is obtained by combining the weight of conjunctions and emotional phrases in the sentence. Then the emotional tendency of the text is determined by combining the emotional results of the sentences (Xu et al., 2019).

3.3.3 Word embedding – Word2vec

Word2vec was created by Google in 2013. This predictive deep learning model is designed to compute and generate vector representations of words with high quality, continuous density, capture context and semantic similarity.In essence, these are unsupervised models that can absorb large text corpus, create possible word vocabulary, and generate dense word embedding for each word in the vector space representing that word (Mikolov et al., 2013). It usually specifies the size of the embedded vector of the word, and the total number of vectors is essentially the size of the vocabulary. This makes the dimension of the vector space much lower than that of the high-dimensional sparse vector space constructed by the traditional bag-of-words model.

3.3.4 K-means clustering algorithms

Clustering is one of the important data mining techniques to discover unstructured multidimensional data knowledge.

The method of identifying similar groups of data in a dataset is called clustering. Clustering is to divide population or data points into several groups so that the similarity between the same group of data points and other data points of the same group is greater than that of other groups. The purpose of clustering is to separate groups with similar characteristics and assign them to corresponding clusters (Kaushik, 2016; Kasim et al., 2013).

K-means is an iterative clustering algorithm that tries to find the local maximum value in each iteration.The algorithm works as follows (Kaushik, 2016):

1. Specify the number of clusters K required.
2. Each data point is randomly assigned to a cluster.
3. Compute the cluster center.
4. Each point is redistributed to the nearest cluster center..
5. Re-calculate cluster centroids.

Repeat steps 4 and 5 until we reach global optimization.

3.3.5 Silhouette coefficient

Clustering can divide data patterns into meaningful groups, but the number of groups is difficult to determine (Dinh et al., 2019). Profile coefficient is an evaluation method of clustering effect, which was first proposed by Peter J. Rousseeuw in 1986. It combines cohesion and separation. Based on the same raw data, it can be used to evaluate the influence of different algorithms or different operation modes of algorithms on the clustering results. Kaufman et al. (1990) introduced the silhouette coefficient to represent the maximum value of the mean s(i) of all data in the entire data set. Many researchers implement a solution of using the silhouette coefficient to find an optimal number of clustering (Aranganayagi and Kuttiyannan, 2008; Zhou and Gao, 2014; Dinh et al, 2019). Zhou and Gao (2014) indicate a method to automatically determine the cluster number by using the silhouette coefficient and the sum of square error. Experimental results show that this method can find the optimal number of clusters and effectively cluster the data patterns. Aranganayagi and Kuttiyannan (2008) also express that objects are grouped into a cluster on the minimum dissimilarity value. In the merging process, the objects are relocated using the silhouette coefficient. Dinh et al (2019) also use a method based on silhouette analysis to evaluate the quality of the different clusters obtained in the previous step, and select the optimal K in the K-means clustering algorithm.

For K-means, the data it will classify is divided into K clusters. For each vector in the cluster, its contour coefficient is calculated separately (Kaufman et al., 1990).

For one of the points i:

Calculate a (i) = average (distance between i vector and all other points in the cluster to which it belongs)

Calculate b (i) = min (the average distance between the i vector and all points in the cluster adjacent to it) Then the silhouette coefficient of the i vector is:

$$S(i) = \frac{b(i) - a(i)}{max\left\{a(i), b(i)\right\}} \tag{1}$$

a (i) : mean of the dissimilarity degree of i vector to other points in the same cluster

b (i) : the minimum value of the mean dissimilarity of the i vector to other clusters

It can be seen that the silhouette coefficient is between [-1,1], and the closer to 1, the better the cohesion and separation effect will be.

The average of the silhouette coefficients of all points is the sum of the silhouette coefficients of the clustering results (Kaufman et al., 1990).

Overall, literature reviews provide methods for managing "queues" and knowledge work and also provide technologies for data mining.

4 Data analysis and visualization

In this thesis, the aim was to firstly, understand how workflow data analysis could improve productivity (***RQ1***), secondly, understand how workflow data analysis could help to improve IT helpdesk operational management and knowledge work (***RQ2***), thirdly, understand how lean production could be applied to improve IT helpdesk operational management and knowledge work (***RQ3***). To find the answers to these research questions, this section presents a series of experiments aimed at mining insights about Futurice's IT helpdesk problems, to inform the prototype described in Section 5.

4.1 Overview of methods

Both quantitative and qualitative were used to address the research questions. Scientific research and literature review were used to find theories aligned with research problems. Qualitative research included observations, interviews, and focus group discussions, aiming to conduct a series of user validation activities. Quantitative research included data visualization, queuing model, exploratory data analysis, sentiment analysis, text mining, K-means clustering, aiming to mine insights from workflow data.

Firstly, it identified challenges and detected waste in workflow by conducting observation and interviews. Graphic visualization and a queuing model were implemented to visualize current workflow states. Meanwhile, it was focused on data analysis with multiple techniques to mine insights and try addressing existing problems. The aim was to help eliminate waste while increasing productivity. The process was conducted in several sub-sections: exploratory data analysis (EDA), and then further data and text mining, K-means clustering, and evaluation. The purpose of clustering was to derive issue categories from workflow data to optimize services for different categories of issues.

Next was to apply insights found in the previous phases into prototype design, which could improve IT helpdesk workflow become more lean production by eliminating waste during its workflow. It was also focused on testing and evaluating a candidate solution based on the prototype and collecting feedback from target users.

4.2 Data collection

Workflow data was extracted from the ZenDesk system containing ticket information in accordance with Futurice data privacy policies. The filtering rules of ticket data from ZenDesk were as follows:

- The time was from 01. January.2020 - 13. March.2020, a total of 72 days.
- The data only contained tickets belonging to the IT support team, no HC team data was included.

After filtering, there were 1881 original ticket data in total. ZenDesk workflow data fields mainly used for analyzing included: *Ticket handler team (recipient), Identification (Ticket ID, Ticket URL), Assignee (Agent ID, Agent name), Content (Subject, Body, Comments), Metadata (Status), Time Stamps (Created, Assigned, First-reply, First-solved, Solved), Metrics Calculated from Time Stamps (Full resolution time).*

4.3 User research and data visualization

This phase was to understand more about IT helpdesk problems of existing or potential waste during its workflow that impacted the work efficiency.

4.3.1 Current state estimation

In order to help users understand the current *"queue"* and state more intuitively, a streamgraph was used to visualize the workflow. A streamgraph is stacked area graph that moves around a central axis to form an organic shape of flows. The visualization of streamgraph was generated by D3JS, which provided an interactive visual representation of data with rich features and animations. Moreover, a queuing model was used to estimate the IT helpdesk dynamic process and evaluate its current workload.

4.3.2 Observation and Interviews

For observation, an IT agent performed the whole process by operating the ZenDesk system and explaining it as he operated. After that, I interviewed 4 Futurice IT helpdesk agents and showed them current state estimation graphs, and they elaborated insights of the following questions from a system point of view:

1. Whether they noticed that there were many unsolved tickets in queues.
2. How they behaved when facing different types of unsolved and unsatisfactory tickets.
3. How they decided priorities of tickets.

Moreover, I interviewed Tiina (IT helpdesk operation manager) and she elaborated insights of the following questions from operational and knowledge management point of view:

1. Whether she noticed that there were many unsolved tickets in queues.
2. How she took actions when IT agents facing large workloads.
3. How she decided resource allocation and adjusted the plan.
4. How she implemented knowledge management.

4.3.3 Current visualization of "queue"

Graphic visualization

The workflow streamgraphs (Fig. 9) qualitatively displayed the flow of tickets of a month. The duration was scaled to February. Each area represented one ticket. The user could update the streamgraph by filtering status and agent name, and fuzzy query with key words. The color represented the number of emails of one ticket. The darker the color, the more emails in this ticket.

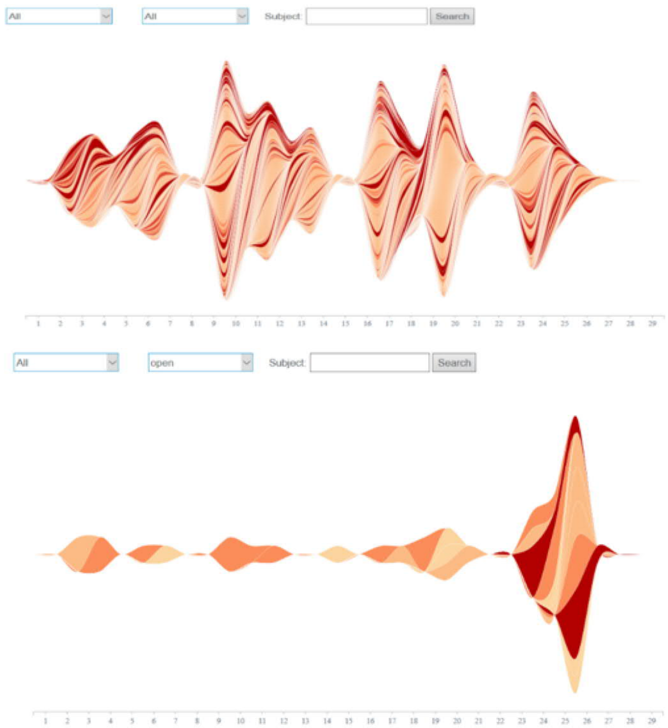


Figure 9: Workflow streamgraphs of all tickets (*above)* and open tickets (*below*) (Fan et al., 2005)

From workflow streamgraph, IT helpdesk agents had to handle with a long queue, roughly speaking, when each start of a new week, the number of tickets had an obvious increase. There were also still open tickets kept cumulative day by day. By the end of the month, there were a large number of pending tickets accumulated.

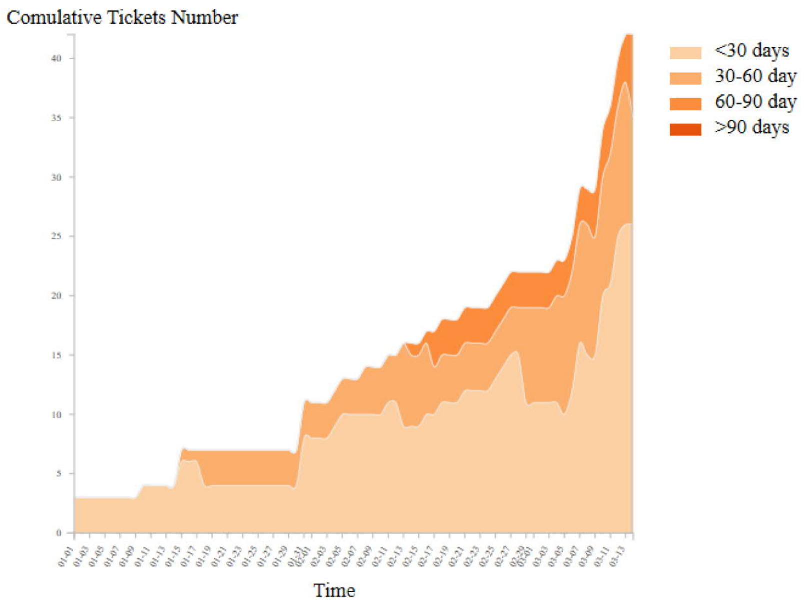


Figure 10: Cumulative tickets graph

To be more quantitative, the cumulative tickets graph (Fig. 10) represented long-term unsolved queues, which could affect customer experience and overall IT helpdesk performance. There were 26 cumulative unsolved tickets within 30 days, 12 cumulative unsolved tickets in 30-60 days, 4 cumulative unsolved tickets 60-90 days, and no tickets over 90 days. It could be seen that the queue of unsolved tickets increased day by day.

Queuing Model

Furthermore, the M/M/s queuing model (Reinertsen, 1997) was applied to evaluate the current system performance and provide evidence of the workload. The below tables showed the metrics and formulas used in the queuing model. Based on the current system, there were the following metrics calculated from timestamps:

- Full resolution time = full resolution time (data field)
- Processing time = solved time – assigned time
- Requester wait time (queue time) = assigned time – created time

After calculation:

items	**Processing time / hour**	**Full resolution time / hour**	**Wait time / hour**
mean	82.12	121.09	38.97
Std.	368.98	412.53	163.28

Mean data was as input data for this queuing model, for estimating states of system. Therefore, inputs were following four variables:

- s was the number of service agents available, currently ***9 IT agents***,
- λ was average tickets arrival rate,
- inventories was ***42*** unsolved tickets, which was calculated from the above calculation,
- μ is average service rate at each channel. Service rate = 1 /processing time.

Formulas	**Metrics**	**Definition**
$Ls = \frac{\lambda}{\mu-\lambda}$	Length in system	Average number of tickets in the system being served.
$Lq = \frac{\lambda^2}{\mu-\lambda}$	Queue length	Average number of tickets in the queue waiting to be served.
$\rho = \frac{\lambda}{s\mu}$	Utilization rate	Percentage of time of an IT agent is being utilized by a ticket.
$Wq = \frac{\lambda}{\mu(\mu-\lambda)}$	Queuing time	The average time of a ticket waiting for service.
$Wt = Wq \cdot inventories$	Requester wait time	The average time of a ticket spent in the queue waiting for the service.
$P_0 = 1 - \left(\frac{\lambda}{\mu}\right)$	Probability of idle agents	Probability of agents to be idle.

Currently the arrival rate was ***25 tickets per day***, and the average processing time was ***3.42 days***, so the average service rate was ***2.8 tickets per day***. Then, after calculation in queuing model, Ls = 130.52, Lq = 121.59, Wt = 204.12, Wq = 4.86, ρ= 99.21%.

Therefore, it could be seen that the utilization rate was nearly 100%. When the incoming tickets increased over 25 tickets, the system was in a 100% overloaded situation.

In summary, the current performance of IT helpdesk service was that average processing time: 82.12h, average lead time (resolution time): 121.09h, average requester wait time: 39.97h, and the system utilization rate was nearly 100% almost every day.

"Waste" in IT helpdesk workflow

Based on Taiichi Ohno's "seven wastes" in a manufacturing operation that needed to strive to be eliminated, for Futurice IT helpdesk, wastes mainly derived from three "wastes":

1. Inventories of unsolved tickets.
2. Long requester wait time, affecting lead time.
3. Long processing time, affecting lead time.
4. Unreasonable resource allocation (resource waste when issues were wrongly assigned to the person who could not solve the issue and it would be re-assigned).

4.3.4 Problems that may cause "waste" (affecting lead time)

After observation and interviews in the problem validation process, I identified several problems from both system and operational management perspectives.

From a system's view, problems derived from:

- All incoming tickets went into ZenDesk staying in the same long queue, and IT helpdesk agents picked up tickets randomly from the queue without understanding the prioritization of issues. They labeled tickets with an *"Urgent"* tag by their own experience.
- They could not know the issue type and the difficulty of this issue before they open and read email content.
- IT helpdesk agents could not judge customers' emotions timely, only when they got complaint calls.

From the operational management(IT helpdesk operator)'s views, problems derived from:

- She had no way of knowing which types of issues had the most problems and low satisfaction rate and which resources were most needed. She could not make adjustments timely.
- She could not evaluate the overloaded work of employees.
- She could not centralize and distribute knowledge and experience of resolving tickets for future reference. If the problem arises again but is assigned to another agent, the second agent must learn and solve the problem without any prior knowledge.

4.4 Data analysis experiments

Next, data analysis focused on sentiment analysis and clustering analysis for understanding the complexity of tickets and insights from different types of issues.

4.4.1 Sentiment analysis

For understanding attitudes from customers to IT helpdesk services, sentiment analysis was implemented. The goal was to extract sentiments from email content from customers. Email content included subject and description text.

1. **Dictionary preparation**. The sentiment analysis method was based on the sentiment dictionary. It defined sentiments into positive, neutral, and negative emotions. Sentiment dictionaries used were downloaded from WordStat Sentiment Dictionary 2.0 (01/26/2018), which was found on website: https://provalisresearch.com/.
   - **Sentiment dictionary:** The weight of each word was set to 1.
   - **Dictionary of degree adverbs:** Different degrees of adverbs corresponded to different weights.
2. **Text pre-processing.** These steps were used for converting text features including all email text (description and subject) from human language to machine language for further processing. The pre-processing included:
   - **Tokenization:** Parsed email text data into smaller units called tokens.
   - **Cleaning:** removed all the undesirable content, including punctuation removal, stop words removal.
   - **Normalization:** converted any non-text information into textual equivalent.
   - **Lemmatization and Stemming:** transformed text into a canonical (root) form, for deriving the root word.
3. **Compute sentence scores.** It was iterated through the words in the sentence. If it was negative, added 1 to negative score to determine whether there was adverb in front of negative word. If there was, multiplied the corresponding weight. The sentiment value of *positive_score - negative_score* was the final score; if greater than 0, it was *"Positive"*; if smaller than 0, it was *"Negative"*, otherwise, it was *"Neutral"*.

4.4.2 Customers' emotions

From sentiment analysis, it (Fig. 11) presents three common sentiments: negative, positive, neutral from customer's sentiments. Around 74.4% tickets from customers indicate a positive attitude, while still about 11% customers express dissatisfaction or a negative attitude during their communication with Futurice IT helpdesk. After reviewing dissatisfied cases, it shows that most of reasons of bad feedback include long processing time and providing incorrect solutions. It reveals that IT helpdesk services still need to continue improvements.

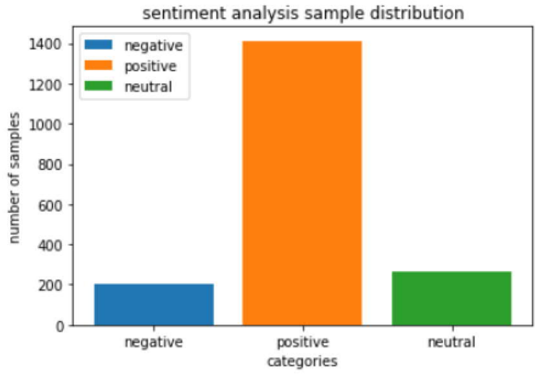


Figure 11: Sentiment analysis from customer email content

4.4.3 Clustering analysis

If problem categories can be categorized by topic, IT agents that resolve tickets could be able to understand the problem more easily and focus on solving tickets that have the same category. Clustering was a way of splitting a long queue and achieving tickets categorization. In this section, there were two experiments for clustering: text clustering (documents clustering) and clustering analysis with multiple features.

Experiment 1 - Text clustering

In this experiment, the goal was to classify tickets (IT issues) according to email content from customers. Text clustering was an unsupervised learning technique for grouping similar documents. Documents were gathered dynamically, not through using predefined topics. The basic clustering algorithm created a topic vector for each document and measures the fit of the document in each cluster. In this analysis, text clustering belonged to document level clustering which aimed to regroup documents about the same topic (Feldman et al., 2007). K-means clustering algorithm was used and TSNE plot was used for dimension reduction. TSNE diagrams are dimensionality reduction techniques that graphically simplify large data sets.

1. **Text pre-processing.** This step was the same as in the above sentiment analysis.

2. **Feature extraction.** In this step, it converted documents into vectors using the word embedding model (Doc2vec). It computed a feature vector for each document in the corpus.

3. **K-means clustering.** After feature extraction, the text was ready for clustering. K-Means clustering algorithm was chosen in this step. Overall, it computed similarities of vectors and clustered different text documents based on the features generated.

4. **Evaluation.** Measuring the quality of a clustering algorithm was to be important as the algorithm itself. Evaluated it in two ways: 1) Silhouette coefficient 2) External quality measure: conducted focus group discussion to evaluate text clustering result.

Experiment 2 - Clustering with additional features

The features could be added more or be deleted to adjust and find better groups for applying on different business scenarios. Therefore, in order to further fractionize ticket issues, in the following experiment, I classified tickets by inputting additional features in clustering. Timestamps features were used to distinguish tickets with long or short wait time, long or short processing time, long or short full resolution time.

Before the clustering process, it was necessary to select appropriate features to "feed" the clustering algorithm.

1. **Feature correlation analysis.** Correlation analysis was used to explore whether there was a strong correlation between these features. Here I analyzed the correlation between processing time, requester wait time, and full resolution time. An intuitive way to see the correlation between features were using scatter matrix. For a more quantitative analysis, Python Seaborn library was used. According to the calculation formula of correlation coefficient, the correlation coefficient on the diagonal was 1. The closer the color of the sub-graph was to light, the closer the correlation coefficient was to 1, and the stronger the feature correlation, and it was positive correlation. When the sub-graph color is close to the other extreme (-1), the feature correlation was also strong and negative correlation.

2. **Feature scaling: Normalization processing.** The numbers needed to be normalized accordingly, that is, all index values were at the same quantitative level. Normalization was to make the features of different dimensions have a certain degree of numerical comparison, which could greatly improve the accuracy of the classifier. It was a linear transformation of the original data, so that the result could fall into the interval of [0, 1]. Linear normalization formula: $x' = \frac{x - min(x)}{max(x) - min(x)}$

3. **Principal Component Analysis.** In order to understand which features contribute the most to sample classification, PCA, also known as Principal Component Analysis, was used for dimension reduction. To put it simply, a certain number of components could be calculated, and each component could explain a certain proportion of data.

4. **Feature extraction.** Merged features and retrieved feature vector matrix.

5. **K-means clustering.** The same as it was in above text clustering experiment.

6. **Evaluation.** The same as it was in above text clustering experiment.

4.4.4 Clustering and evaluation

The two clustering outcomes are visualized as below (Fig. 12), all ticket data were clustered by categories and each color represented one category. It clustered the most relevant type of tickets into one category.

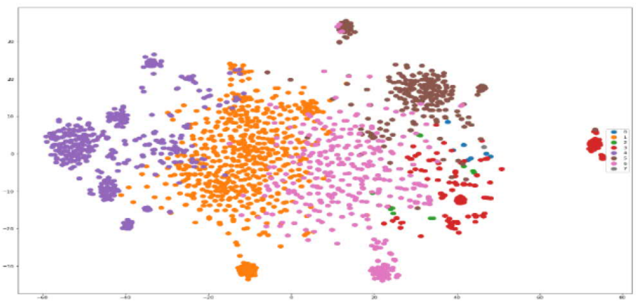


Figure 12: K-means clustering result

To determine the number of clusters, the method was to calculate the silhouette coefficient of each data, which is a measure of whether the clustering is reasonable and effective. The silhouette coefficient, combined with the cohesion and separation of the clustering, was used to assess the effect of clustering. The value range is [-1, 1], with -1 representing complete dissimilarity and 1 representing complete similarity. Therefore, the closer to 1, the better the clustering effect of the data.

The specific calculation method was as follows: For each sample point i, the average distance between point i and all other elements in the same cluster was calculated, denoted as a (i), to quantify the cohesion within the cluster. Selected a cluster b outside i, calculated the average distance between all points in i and b, traversed all other clusters, and found the nearest average distance, which was denoted as b (i), which was the neighbor class of i, to quantify the degree of separation between clusters. For sample point i, silhouette coefficient:

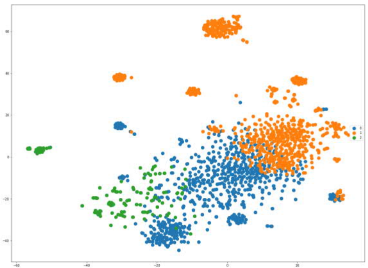
$$s\left(i\right)=\frac{b\left(i\right)-a\left(i\right)}{max\left(a\left(i\right),b\left(i\right)\right)}$$

The silhouette coefficients of all points were calculated, and the average value was the overall silhouette coefficient of the current clustering, which measured the tightness of the data clustering The average silhouettes factor was calculated in the sklearn function *silhouette_score()*, while *silhouette_samples()* returned the silhouettes factor for each point.

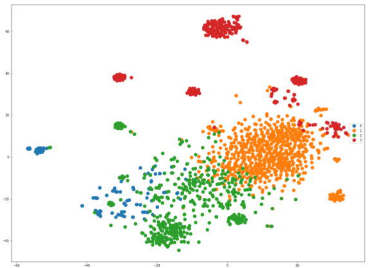
I set the number of clusters from 3 to 8 in K-means clustering, then calculated the silhouette coefficient, and selected the number of clusters with the best silhouette coefficient as the choice of the K-means algorithm.

4.4.5 Experiment 1 - Text clustering outcomes

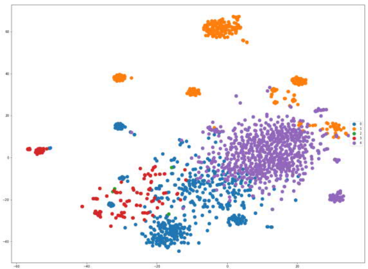
The below graphs (Fig. 13 (a)-(f)) represent the results when defining cluster numbers from 3 to 8, clustering silhouette coefficient scores.



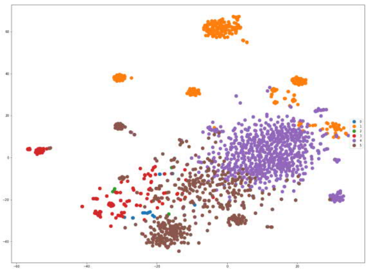
(a) When cluster= 3,silhouette_score= 0.11428619355781511



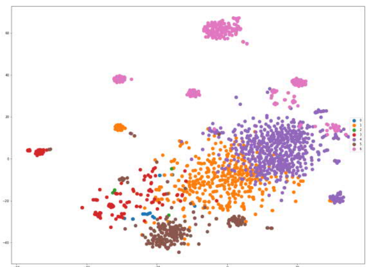
(b) When cluster= 4,silhouette_score= 0.07464751817492797



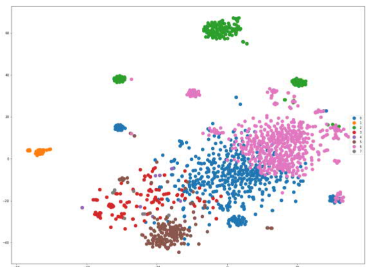
(c) When cluster= 5,silhouette_score= 0.07870270263204197



(d) When cluster= 6,silhouette_score= 0.08199680552462897



(e) When cluster= 7,silhouette_score= 0.07646695712035526



(f) When cluster= 8,silhouette_score= 0.08108441310181946

Figure 13: (a) – (f). Text clustering results

It can be seen that when the cluster number is 3, the silhouette score equals about 0.114, which is the closest to 1. Therefore, when cluster= 3, it generates a better outcome of clustering.

4.4.6 Experiment 2 - Clustering with multiple features outcomes

Scatter matrix and heatmap of feature correlation

The scatter matrix is symmetrical. Except for the density function graph on the diagonal, the other subgraphs represent the relation between different features, if the relationship between features is approximately linear, those features are strongly correlated. On the contrary, the scatters of other feature columns are disordered, which indicate that the correlation between features was not strong. From the scatter matrix (Fig. 14), processing time and full resolution time are strongly correlated.

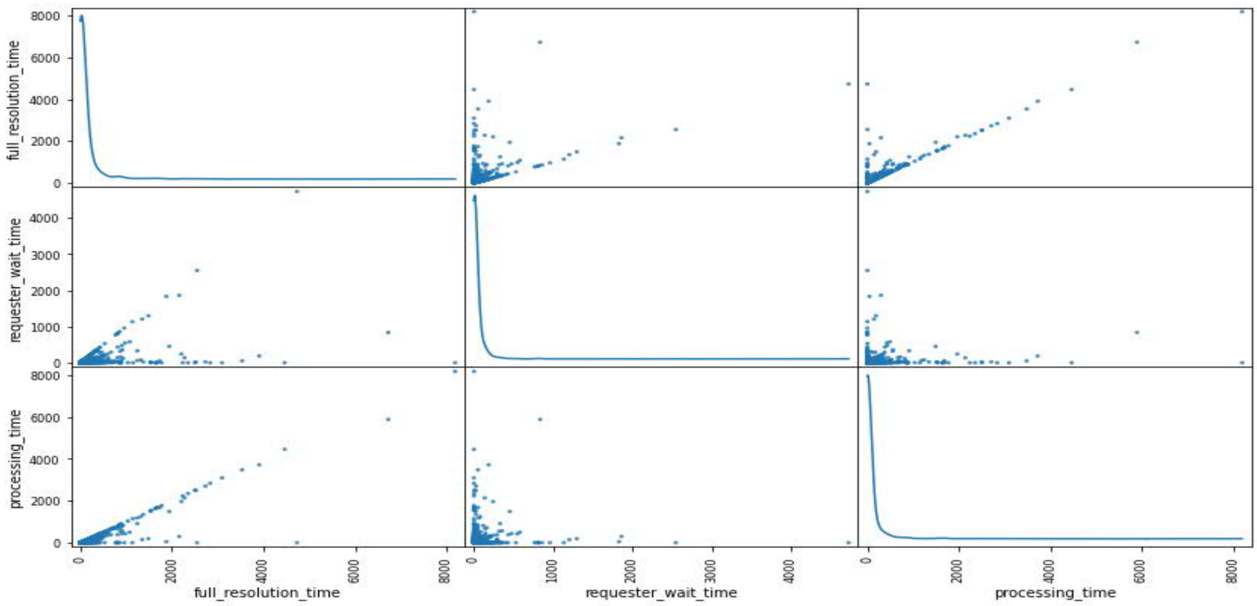


Figure 14: Scatter matrix

More specifically, as can be seen from the heatmap (Fig. 15), there is a strong correlation between full resolution time and processing time, and there is also a correlation between request wait time and full resolution time, while the other features are relatively weak or even irrelevant, which is consistent with the conclusion of the dispersion matrix diagram.

Graph of Principal Component Analysis

The corresponding characteristic weight value and variance interpretation of three components were obtained in the principal component analysis (Fig. 16).

The variance interpretation of the first component is 0.7838, which was that this combination of components could explain 89.01% of the data. As it could be seen, the full request time and processing time had a very large positive weight, while requester wait time had only a small weight.

From above exploratory data analysis, the features selected for clustering were those that could distinguish a ticket from different perspectives. Since full resolution time and processing time had close correlation, I removed full resolution time. Input features for clustering in this experiment 2 were:

1. Processing time

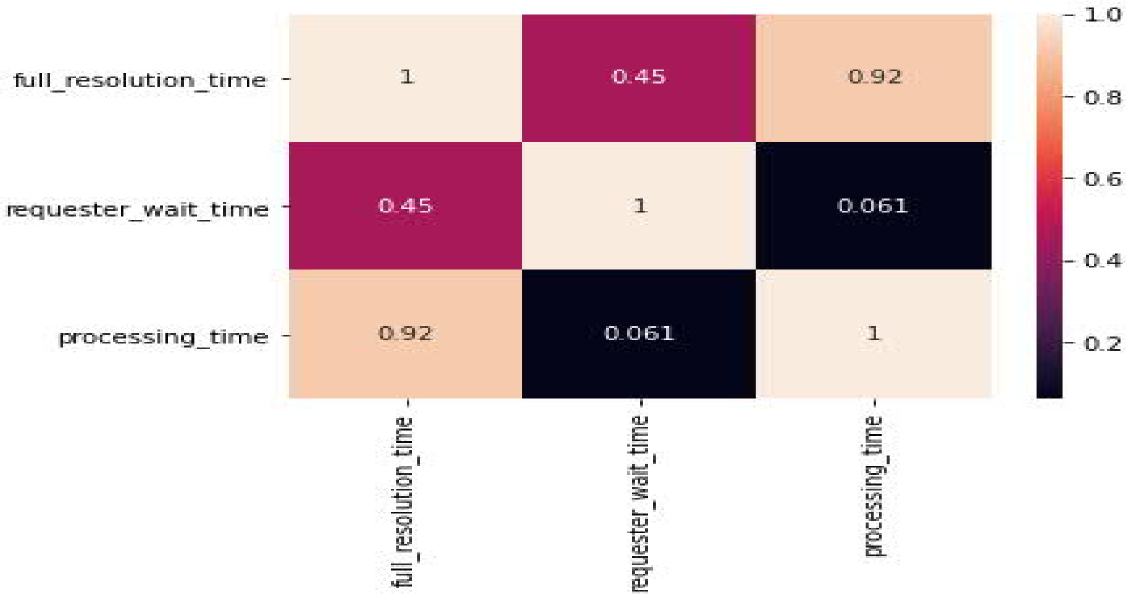


Figure 15: Heat map

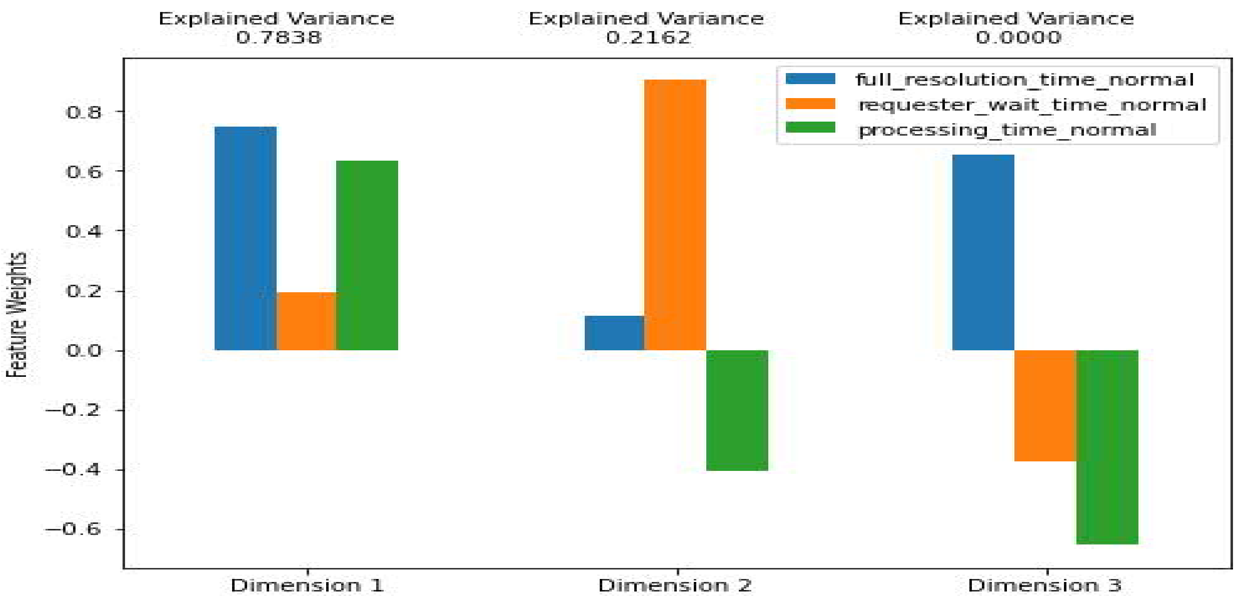
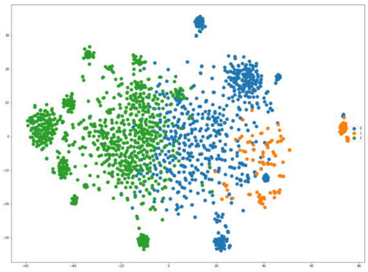


Figure 16: Principal Component Analysis

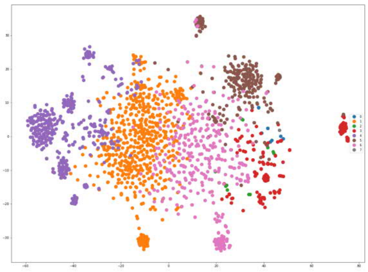
2. Requester wait time
3. All email text

After the clustering, the below graphs (Fig. 17 (a)-(f)) represent the results when defining the cluster numbers from 3 to 8 and their silhouette coefficient scores.

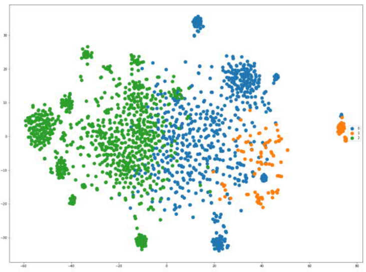
It can be seen that when the cluster number is 3, the silhouette score equals about 0.183, which is the closest to 1. Therefore, 3 is the cluster number to produce a better outcome of data.



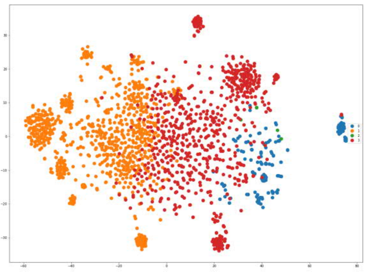
(a) When cluster= 3,silhouette_score= 0.1830219077507251



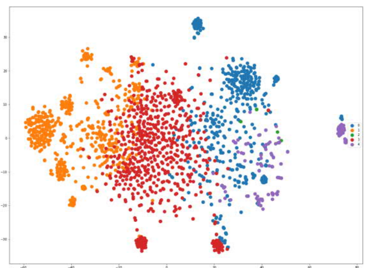
(b) When cluster= 4,silhouette_score= 0.16763715372587937



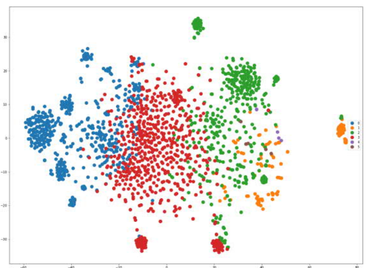
(c) When cluster= 5,silhouette_score= 0.08511972114364305



(d) When cluster= 6,silhouette_score= 0.08693534769581839



(e) When cluster= 7,silhouette_score= 0.08103773832126528



(f) When cluster= 8,silhouette_score= 0.08374639557228714

Figure 17: (a) – (f). Clustering results

4.4.7 External quality measure

After the above analysis, the clustering results (with the best cluster numbers) were provided to the IT helpdesk team and they checked the tickets email content in each group and provided the most relevant IT problem label for each category.

Experiment 1: For email text clustering result, 3 is the optimal cluster number and IT team went through tickets in each group and labeled groups with IT problem names of *"Linux"*, *"Infra"* and *"Device management"* (Fig. 18).

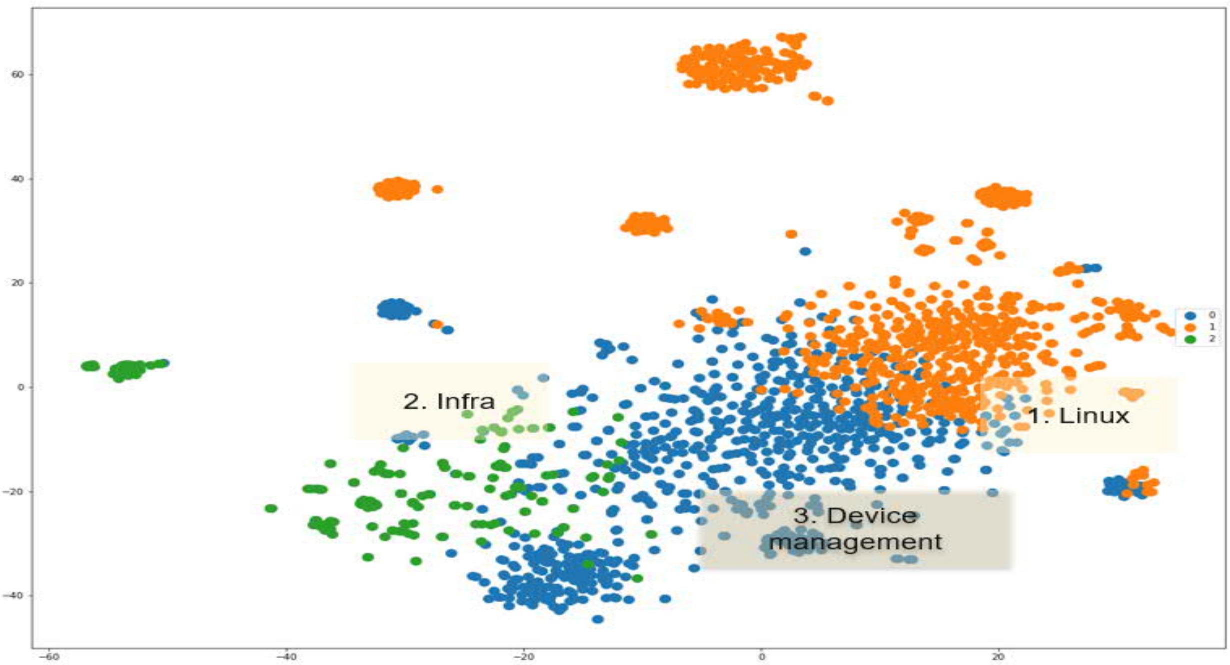


Figure 18: IT issue groups

Experiment 2: For multiple features clustering of results, 3 is also the optimal number. It includes timestamps and sentiment as additional features. After the IT helpdesk team reviewed, they could find a group with a relatively long or short requester wait time, relatively long or short processing time, and different numbers of the positive, negative and neutral sentiment of customers.

Finally, they labeled 3 groups with names of *"long wait time, long processing time with low average satisfactions - Linux"*, *"short wait time, short processing time with high average satisfactions - Device management"*, and *"normal wait time, normal processing time with normal average satisfactions -Infra"* (Fig. 19). After manual validation, the accuracy of clustering was only 85%.

However, after manual validation, the accuracy of clustering was only 34%, and the rest of the tickets could not be categorized into correct groups. The reason is most of the email content is unstructured text and customers cannot express accurately what they expect, resulting in most of the content cannot be identified by the clustering algorithm.

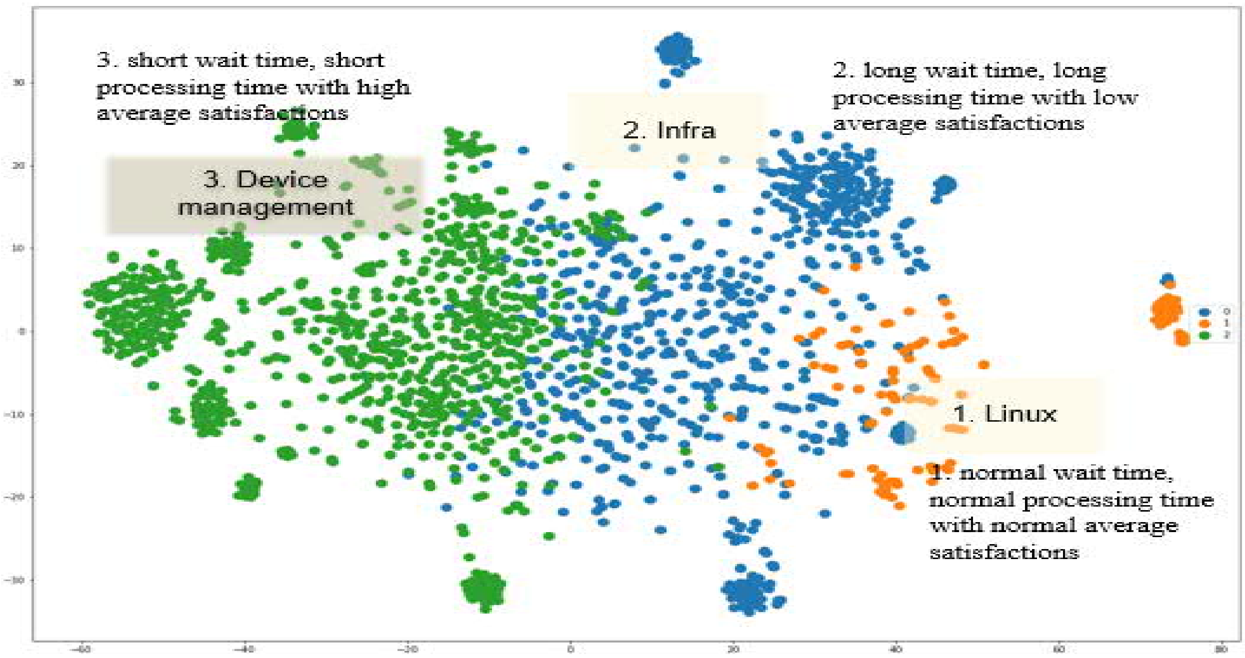


Figure 19: IT problem groups with more features

4.5 Summary of insights

As is known, the ultimate goal of lean production is to reduce lead time (processing time plus requester wait time). Based on Reinertsen's (1997) research, a lead time is a time required for a single part to complete the entire process or system. This is an important indicator of how quickly the system responds to change. In this case, the lead time is the (average) time our customers have to wait for their problems resolved. Obviously, the main stem is to control unsolved tickets and manage queues during workflows. From the above workflow data analysis results, there are some insights.

4.5.1 Why is managing the queue important

Perhaps the most underrated is the cost of delayed response. The delivery time increases proportionally with the queue. This relationship is known as Little's Law. The more inventory we have, the more sluggish the system becomes. Richardson (1995) explained that inventory is expensive. Inventory can cost anywhere from 30 to 65 percent of the entire value, depending on the environment and the cost is opportunity cost. For a pure service system as an IT helpdesk, the inventory refers to the cumulative queues in the system. The ultimate goal of an IT helpdesk is to solve the issues as soon as possible. However, Richardson also (1995) emphasized that the queue will grow and cumulative when utilization or workload reaches 80%, queue always exists and inevitable. Thus, it is important to manage the queue with a balance between workload and queue length.

Trade-off between too much and too little

It is also necessary to reach a balance between inventory and system capacity. We have to find it by lowering inventories and watching how the workflow responds. For a pure-service system, it is obvious to have little inventory rather than too much meanwhile we should consider the capacity of the system and the workload of employees.

Reinertsen (1997) explains further about the relationship between queue length and capacity utilization. He found some interesting properties in the classic queuing curve of an M/M/1 queue model (a type of common queue). The queue time will be quite low from the beginning but start increasing while levels of utilization reach about 60% to 70%. It doubles as moving from 60% to 80% utilization. It doubles again as moving from 80% to 90% utilization and keeps growing as moving more utilization (Reinertsen, 1997). Reinertsen (1997) expresses his ideas that people would not expect the system to become overloaded until they reach full utilization. Meanwhile, people expect no delays as long as they had excess capacity. However, in the stochastic world of queuing theory, delays could not be avoided even when we have excess capacity in the system. Therefore, in real optimizing practices, how to trade off queue time, and utilization of a system need to be considered carefully. In Reinertsen's research, for most processes, the waiting time in queues is much more costly than that of the process itself.

4.5.2 Workflow visibility and automation improve productivity

Through workflow data visibility, IT agents can monitor the queue and adjust their behaviors based on real-time workflow data.

Firstly, statistical data including inventories that can reflect cumulative unsolved tickets is extracted, and it also computes dynamic data such as utilization, queue length in a real-time system in a queuing model. Secondly, workflow data analysis uses data mining method to cluster tickets with the most relevant issues, which helps IT agents to determine which group of tickets to be prior solved through different metrics. If features are pure text, the clustering process splits tickets based on text contents. For example, it produces a group called Infra, which contains all Infra related problems. If adding sentiment feature and time features inside the clustering algorithm, tickets with the least satisfactions/long wait time/long processing time will be identified and can be solved timely.

Through workflow data automation, IT issues can be grouped automatically. IT agents do not need to tag tickets manually. The text categorization and clustering processes help them to do it automatically. Incoming tickets with a tag will be categorized into its group. The ticket without a tag will be clustered into a group. Tickets with the same features can directly split the long queue into sub-queue. It improves IT helpdesk workflow automation and eliminate unnecessary waste for tagging and splitting tickets manually.

4.5.3 Workflow visibility and automation improve operational management

Workflow visibility and automation help operational managers justify automation investments, new templates, service plan upgrades based on IT helpdesk performance. Managers can determine what resources need to be invested based on which type of problem has the longest or toughest queue. Managers can also monitor real-time traffic of queue, adjust resource strategies promptly, maintain a balance between workload and tickets (in queue), avoid overworking employees, and ensure their physical and mental health. The final goal is to reduce waste and maximize profits at the lowest cost.

4.5.4 Text mining helps to improve knowledge management

Vishal and Gurpreet (2009) argue that text mining is considered to have high commercial potential since more than 80% of the information was stored as text. Although knowledge could be found in many forms of information, the unstructured text remains the biggest available source of knowledge (Vishal and Gurpreet, 2009).

Figure 24 illustrates the process of extracting knowledge from text mining. It begins with a collection of documents from resources. Then, the tool retrieves specific documents and preprocesses them. It works by checking formats and feature sets. After that, it would go through a phase of text analysis, usually repeat until the information is extracted. The results can be stored in a management information system, providing a wealth of knowledge to the key stakeholders of the system.

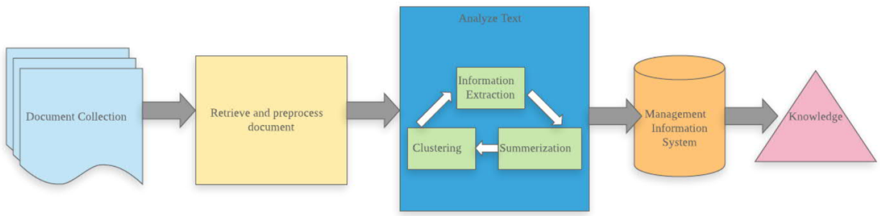


Figure 20: Knowledge is extracted from text mining (Vishal and Gurpreet, 2009)

Omer et al. (2016) propose that a potential source of knowledge acquisition is the use of the service system (Omer et al., 2016). Actually, in our research, the use phase can be provided with information and knowledge, through which the iteration of service could be improved constantly.

In the IT helpdesk workflow, email exchanges between IT agents and customers create high potential and meaningful information. Through the workflow data analysis process, meaningful knowledge is acquired from those unstructured texts. It helps us to understand customers' emotions, dissect problems, and identify needed skills for solving those problems. IT workers perceive and learn from an iteration of knowledge, constantly updating their own knowledge and share with other members. In the meantime, the workflow automation improves the knowledge extraction process, and accelerates entire knowledge flow.

Furthermore, Ambos and Schlegelmilch (2009) argue that the ultimate goal and appropriate assessment of KM should be the return on investment and time and money saved by knowledge management. Therefore, the application of workflow visibility and automation improves knowledge extraction and acquisition efficiency, saving money and resources, and eliminating unnecessary waste.

4.5.5 Lean production conversion at IT helpdesk

This thesis tries to apply lean production to IT helpdesk workflow. The ultimate goal is eliminating "waste" while increasing productivity. The "waste" at the IT helpdesk is detected firstly. The root cause of "wastes" are from the non-transparency of the IT helpdesk workflow. There are two categories of "waste": time waste and resource waste. IT helpdesk agents spend mostly on processing issues when they cannot understand the prioritization of tickets. Actually, the long time pending issues should be prioritized. Another is resource waste, IT helpdesk agents cannot understand which skills and effort needed for issue resolution. The issues are allocated to the wrong person who has not such kind of experience, which will make a waste of resources. As May (2005) suggests, for lean knowledge work, that information needs to flow to the right person at the right time in the right form with the lowest cost and highest quality. Therefore, the right issue needs to go to the right agent who can solve the problem. After understanding the root cause of waste, a solution is proposed and focused on two parts: workflow visibility and workflow automation. The workflow visibility can help to visualize workflow and monitor the health of holistic workflow. Work automation can help to divide issues into the same group by category. IT helpdesk agents can know the issue type directly, ensuring the right person can be assigned to the corresponding issue. The solution provides a way of converting an IT helpdesk workflow to lean production by reducing processing time while increasing productivity.

4.6 Limitations and validity

The experiments' data only contain a small volume of data and cannot fully reflect a real situation.

K-means algorithm is sensitive to the initially selected centroid points, and the clustering results gained from different random seed points are completely different, which has a great impact on the results.

The accuracy of clustering depends on data quality. Some of email texts were poorly recognized and categorized. The accuracy of experiments was not ideal. The experimental results show that only a small part of the data is separated, but there are still some data that are not properly divided.

First-in, first-out (FIFO) rule, which is inappropriate in some service systems, especially those handling urgent services. The used M/M/s queuing model has some limitations and weaknesses. This model uses the factor of FIFO and there is no hamster time, which cannot reflect the real case. In reality, the queuing process is stochastic. However, it is roughly right, more precise models need to be validated in future research.

There is no need labeling of samples manually in the emotion dictionary-based sentiment analysis, which is much easier to implement. However, the quality of sentiment analysis depends on the quality of sentiment dictionaries. Sentiment dictionary lack of field sentiment vocabulary may affect the final results.

4.7 Future research

More research could be done in the future about improving clustering processes and improve the accuracy of clustering. Trying different clustering algorithms not only K-means and comparing their outcomes. It is also necessary to change to another way of dividing queue into different issue groups. For example, it could be done by labeling firstly and then categorize it by the supervised learning algorithm. Then, comparing its outcome with unsupervised clustering outcome. Furthermore, future research will be a transformative attempt at applying this tool to other application scenarios.

5 Prototype design and evaluation

In Section 4, problems and "wastes" were identified. Insights out of workflow data indicate that different IT issues can be categorized into groups. Bad satisfactions of customers issues are extracted from their email text. Utilization of system that reflects real-time queue health is calculated from the queuing model.

Therefore, this phase proposes a prototype design based on the above analysis results, which is an interactive data-based tool, aiming to convert the Futurice IT helpdesk workflow to be a lean production workflow (eliminating "waste"). Overall, it can help to improve the IT helpdesk workflow in two ways: *workflow visibility* and *workflow automation*. The design prototype was created using Figma including interactive functions that users can interact with on the prototype.

5.1 Design objectives

The objective is to design a lean tool. A lean tool is a tool that can identify and remove waste to improve processing speed. In short, it aims to create maximum value for customers while minimizing waste of all kinds.

IT helpdesk agents and operation manager can monitor the queue health from different perspectives. It can help change the behaviors of IT workers by this data-driven tool. Moreover, the design framework can also be used in other scenarios e.g. GitHub issues, client projects, and other similar operations.

5.2 Hypotheses

The value hypotheses of this design would be if IT helpdesk team can understand and monitor the health of workflow by using this design tool, they might easily eliminate three "wastes" of long processing time, long wait time and inventories when they might easily:

H1. Understand existing problems, e.g. overload problems, thus improving the health and safety of employees (employee's view).

H2. Understand prioritizing, thus avoiding too long queues and satisfaction failures of customers (customer's view).

H3. Justify the automation investments/new templates/service plan upgrades (company's view).

5.3 Workflow visibility

Workflow visibility defines visible data shown on a dashboard from multiple dimensions. It could reflect a real-time queue health and system performance. IT agents can also check states of different categories of IT problems. The workflow visibility was achieved by three different views.

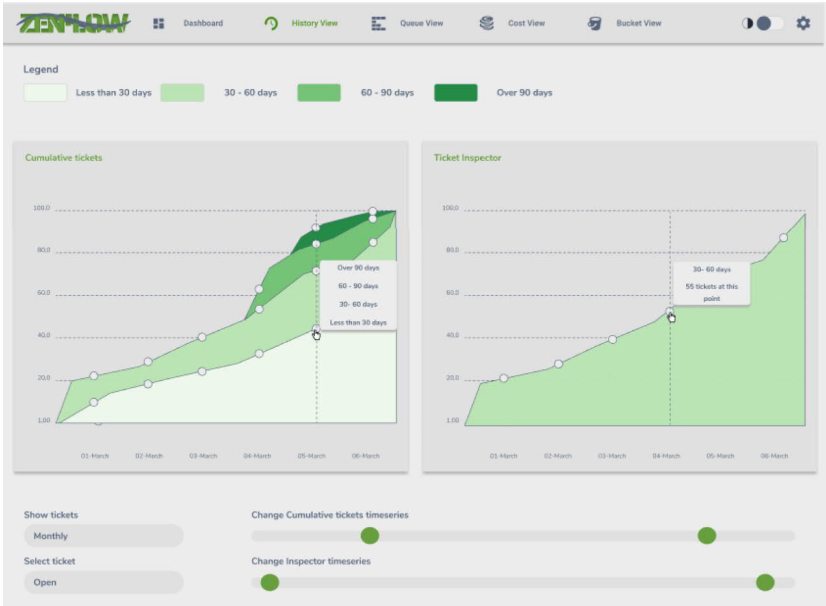


Figure 21: Inventory view dashboard

Inventory view. It (Fig. 20) addresses the problem of the non-visibility of unsolved tickets (inventories). It helps to keep track of pending issues lasting over 30 days, 30-60 days, 60-90 days, and over 90 days. Moreover, it alerts operations management people, and also IT agents be aware of the increasing numbers of unsolved tickets.

Queue view. Queue view (Fig. 21) is a dynamic dashboard that monitors the queue health and queue flow. From the live data, the user can monitor and inspect the flow of queues, to make adjustments of task resolution prioritization based on different metrics.

Each queue is separated by each group of IT issues (that generated from clustering), which is also called a *"Bucket"*. IT issue types, the heat of tickets, lead time, resolution rate, workload, and satisfaction rate of each queue can be seen from this queue view. The heat of tickets represents the number of emails (including replies and requester' email). Moreover, from the queue view, the recent trend of workload and resolution rate can also be tracked.

Bucket view. It (Fig. 22) represents the complexity of different categorizes of IT issues. Each circle represents one *"Bucket"*, referring to one IT issue type that is generated from clustering. Moreover, it shows the Word Cloud of each group based on overall text content. IT agents and operations managers can fully understand the keywords of group tasks and know which specific skill needed to be input for a certain group of tasks.

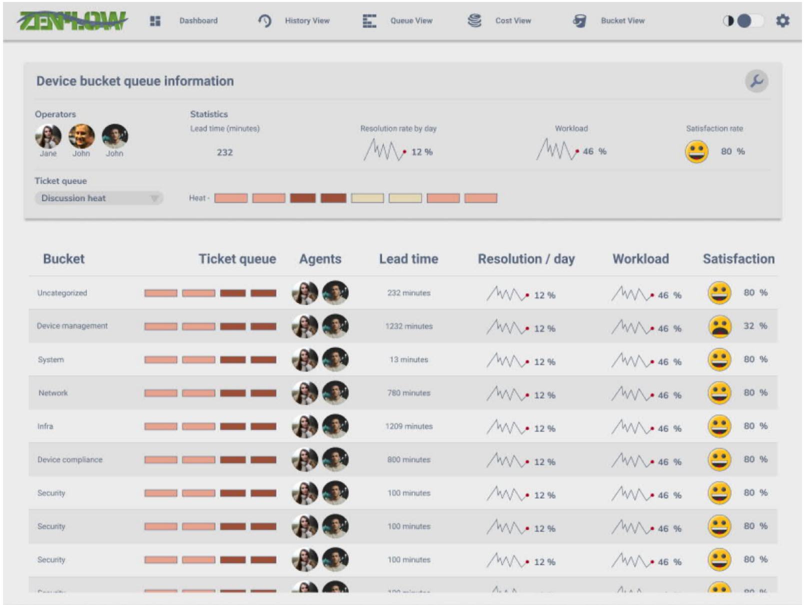


Figure 22: Queue view dashboard

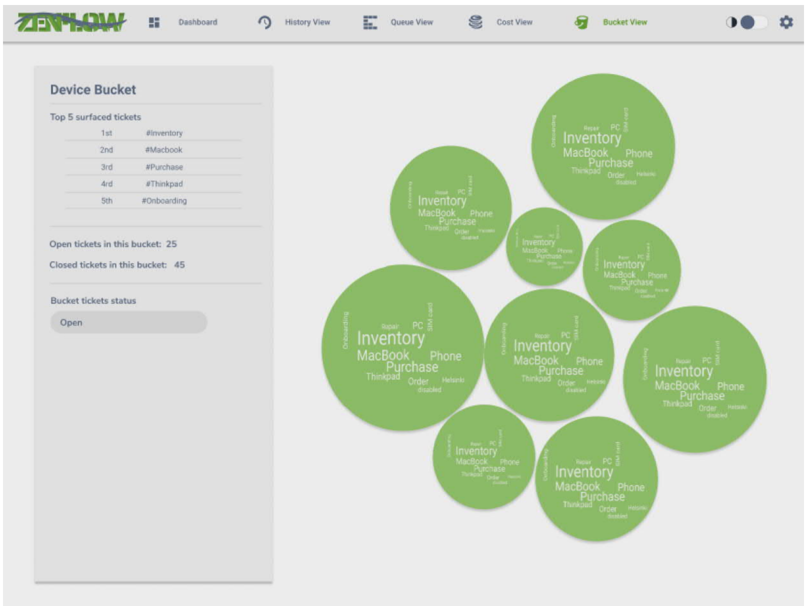


Figure 23: Bucket view dashboard

5.4 Workflow automation

```
predict_classification("./fasttext_model.bin", "Hey It access FutuHours Could contract expiring I renewed end March Best Aleksi", 5)
```

```
access right 0.021291660144925117
licence 0.021288273856043816
okta 0.021288208663463593
laptop 0.021288076415657997
FutuHours 0.021288050338625908
```

Figure 24: FastText model to category one ticket

Workflow automation refers to the back end of this tool. The function is that an incoming ticket can go to its own category automatically with the corresponding label by clustering process. It can divide different IT issues into groups. Features for clustering can be adjusted based on business needs. There provides another alternative. Since clustering cannot fully meet our needs to classify tickets because of its relatively low accuracy, the back-end algorithm can be changed to use a supervised learning algorithm. The categorization process above refers to a supervised classification of tickets based on text. When the class scale and data volume became relatively large, fastText for text classification is used to realize the purpose of fast training prediction and memory saving (Joulin et al., 2016). The core of fastText was to average the words and n-gram vectors of the whole document to obtain the document vector and then use the document vector to do Softmax multi-classification. In the fastText model, it is a document vector realized by embedding words. An important feature of the term vector is that vector distance can be used to measure the semantic similarity between words. The use of term embedding rather than the term itself as a feature could produce a good effect. The introduction of character-level N-gram features could help improve the classification. The fastText model scores the incoming text and predicts the categorization of a ticket (Fig. 23) based on Futurice IT helpdesk tickets data.

The categorization rule should be defined. For example, email text with *"PC"* or *"Purchase"* will go to the *"PC purchase bucket"*. Tickets that cannot be identified well will go to an *"Unknown-categorization bucket"*, and the user can label them with the correct label, then click *"Categorizing"*, it will go to the right *"Bucket"* category. Users can keep training this model with feeding tickets with accurate labels, for ensuring more accurate of the categorization process.

5.5 Usability testing

Usability testing aimed to gather feedback and ideas from the IT helpdesk team (10 employees) about their feeling and experience of using the prototype and also validated value hypotheses of solutions. When prototype design was ready, a usability testing was conducted by firstly sending users an online testing link. Brief user guidance was also provided to them. It required users to test the prototype from different angles, including layout, color, and term clarity, use experience, findings, and overall feeling. After that, a focus group discussion was held for collecting feedback on the prototype design. The team communicated and fully expressed their thoughts in the group discussion. In the final, a 1 on 1 interview was held with 2 key users including 1 team leader and 1 operation manager, clarifying more details about this prototype design. They provided insights into:

- Whether the prototype could be possible to improve their work productivity.
- Whether the prototype could partially or fully validate the above three hypotheses.
- Suggestions to the prototype.

From focus group discussion and interviews, IT helpdesk workers expressed their thoughts about prototype solutions, proving some evidence for three proposed hypotheses. Here summarized:

H1. Understand existing problems, e.g. overload problems, thus improving the health and safety of employees (employee's view).

For the inventory view, all IT team members expressed that they could easily see the cumulative tickets and reminded them to deal with those tickets for avoiding long-term and too much accumulation of inventories. From the queue view, over 90% of IT agents expressed that they could easily see which category of tickets had a relatively long lead time. Several IT helpdesk workers also expressed that they could fully understand the workload rate (Percentage of time agent is being utilized by tickets) on this queue, which reminded them utilization of system for the queue and also workload of agents. They realized it was important to avoid workload rate to reach to nearly 100%, which would cause queue increase and also impact IT helpdesk agents' health. Moreover, they could also see customer's satisfaction rate from the queue dashboard, which could remind them to improve their work quality. But few of them still could not understand the meaning of terms. Moreover, the operation manager also expressed the importance to monitor real-time queue health, which could help her to adjust the workforce timely and avoid the overloaded situation. For the bucket view, they expressed that they could directly distinguish categories of different IT issues. The word cloud in each group provided them ideas about the next actions and helped them to partially understand the difficulty and needed skills to handle those groups of tickets.

H2. Justify the automation investments/new templates/service plan upgrades (company's view).

The operation manager (Tiina) expressed that she saw opportunity cost from this tool, which would largely impact the decision-making processes of investments and plan upgrades. She also said she could see huge potential from this tool of helping them to improve their operational management level. She could evaluate and allocate the resources more reasonably to arrange the workload of employees, which would help to improve the health of the whole team and also improve the work quality and efficiency of employees.

H3. Understand prioritizing, thus avoiding too long queues and satisfaction failures of customers (customer's view).

Almost all IT agents were aware of the urgency of tickets in the queue when they inspected the queue view dashboard. They understood the prioritization of tickets based on the shown metrics, which could help improve work efficiency, preventing dissatisfaction of customers caused by long lead time.

6 Discussion and recommendation

The section provides discussion and recommendations from literature reviews, for improving IT helpdesk operational management and knowledge work.

6.1 Strategies to manage queue at IT helpdesk

Control time

Obviously, it can be seen that waiting time (queuing time, hamster time, agent wait time, everybody's wait time) and processing time (processing time, rework time) are two important indicators that impact heavily system efficiency and workflow stability from both customers experience and agent experience. Therefore, to improve workflow efficiency, it is critical to shortening both of them. The workflow data analysis provides a way of monitoring workflow, enabling IT agents to know the states of the queue and understand prioritization and shorten time accordingly.

Monitor queue

It is not necessary to solve all issues at one time. The workflow data analysis results reveal that the clustering process of splitting issues to different issue topics can be easily monitored and evaluated. It divides the obstacles hidden in queues into smaller pieces. IT agents can decide to put more effort into which category with a relatively long lead time. By knowing the type of issue, the expertise could be reasonably utilized and allocated to deal with the toughest queue, preventing a waste of resources.

Control utilization

Most importantly, if we want to have a reasonable lead time or queue length, we need to get away from 100 percent utilization. The lower the utilization, the shorter the lead time. Higher utilization also means better use of investment capital. Thus, it needs to balance the effective use of a high-utilization system against low utilization lead time. Using this lean tool, utilization (workload) could be monitored for adjusting actions properly.

Inspect variation

Lower changes allow us to leave with lower inventory. Many research reveals that it is easier said than done. In lean production, the concept of levels is designed to reduce this discrepancy for faster delivery times and other benefits. If variability is high, try to use low utilization, and if it has a high demand for flexibility, we can reduce the stress by reducing utilization.If utilization is high, minimize variability.If it has high utilization, it may help reduce variability, although this is often easier said than done. Reinertsen (1997) explains that reducing variability is much harder than non-repetitive operations, but it is still available. 1) measuring variation 2) increase reuse 3) standardize the process. Although the workflow analysis cannot help to minimize variations, it can alert IT people that there might be variations causing a long lead time. IT agents can exploit variations behind the queue.

Control batch

It is also necessary to make customers know the process, sending emails to tell them about waiting time or a reservation system. Capacity planning to prevent the formation of what is called *"predictable queues"*. For managing batch size, moving to a small process batch size, it can achieve a dramatic reduction in the queues without changing either capacity or demands (Reinertsen, 1997).

6.2 Combining TOC and lean production

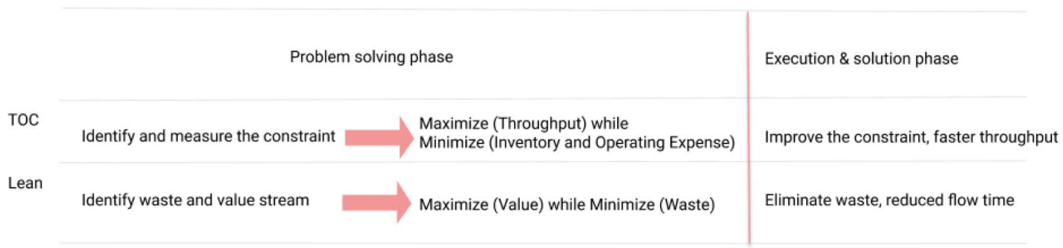


Figure 25: Combining TOC and lean production

If converting IT helpdesk to be more lean production, it is efficient of applying both theory of constraints (TOC) and lean production to eliminate waste at IT helpdesk workflow (Fig. 25).

Constraint theory (TOC) is a holistic management philosophy published in 1984 by the use of Goldratt and Jeff Cox Eliyahu M. Knowledge goals identify the most important constraints, called constraints, that limit or prevent the realization of a goal, and then systematically improve that constraint until it is no longer a constraint (Sergio, 2006). This limitation is often referred to as a bottleneck in manufacturing. TOC focuses on system improvement (Nave, 2002).

The reason for using TOC is that in the real world, for consulting firms like Futurice, compromise is necessary because all companies have limited resources. Not every aspect of every process is worth optimizing, and not every waste is worth eliminating. Therefore, TOC can be used as an efficient mechanism for optimizing workflow, while lean could provide a rich toolbox for improving technologies. By eliminating waste in the parts of the system that are most restrictive to opportunity and profitability, the result will produce a significant increase in manufacturing efficiency.

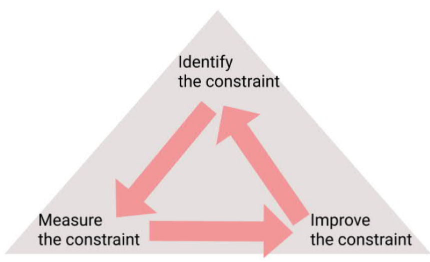


Figure 26: 3-step of analysis

Since the non-transparency of the workflow and information in Futurice IT helpdesk, it is hard to decide where to focus resources. Should all workflows be improved? Where should the focus be? Any breakdowns? The answer to questions is focused on and measuring the constraint, which is the fulcrum for the entire workflows. Focusing on constraints ensures the best use of resources and was the fastest way to get there, increasing productivity and profitability.

The ultimate goal of successful implementation of constraint theory is :1) increase profit 2) rapid improvement 3) increase capacity 4) reduce delivery times 5) reduce inventory. The process of analysis is focused on the first two steps of the five-step focusing model (Goldratt,1984; Mahesh and Lynn, 2008), which is identifying and exploiting. For the IT helpdesk workflow, three main simplifications of processes can be applied as Figure 26.

Since the focus of lean manufacturing is eliminating waste while the emphasis of the Theory of Constraints is increasing throughput, it is powerful to implement a combination of two methods to achieve better outcomes (Power and Conboy, 2015).

6.3 Strategies to improve knowledge work

Gonzalez et al. (2020) describe this current Futurice IT helpdesk approach as an agent-centric approach, which is a typical model of a normal call center with an agent-centric collection of knowledge and information. The disadvantage of the agent-centric approach is that if an agent solves a problem, the knowledge resides only in that agent. If the problem arises again but is assigned to another agent, the second agent must learn and solve the problem without prior knowledge from the other agents. Therefore, each agent follows its learning curve and does not benefit from the organizational knowledge acquired (Gonzalez et al., 2020).

The knowledge can be mined from interactions between IT agents and customers. It includes the knowledge acquires from different channels when an IT agent processes with a ticket and also experiences, feedback, and replies from customers. The results reveal that it is critical for an organization to use knowledge fully.

Invest in training and cross-training employees

Investing in training is an efficient method for improving productivity. Experienced employees are more productive than new employees. From the workflow data analysis, different types of tasks are extracted and the IT team can fully understand which skillful talents are needed to deal with the tickets. It provides evidence for input knowledge for unsolved tickets. Furthermore, cross-training is valuable because most critical bottlenecks occur in specific areas. IT helpdesk team can be supplemented by cross-training to assist the team in heavy periods.

Understand your customers' emotions

From the above sentiment analysis from customers' replies, it is interesting to see that there were still about 11% of customers who expressed bad experiences in the service. Customer experience is an important indicator that can help us to improve the service quality. Furthermore, it can also help to identify the problems behind a certain ticket where it got stuck.

Use Knowledge-centric approach than agent-centric approach

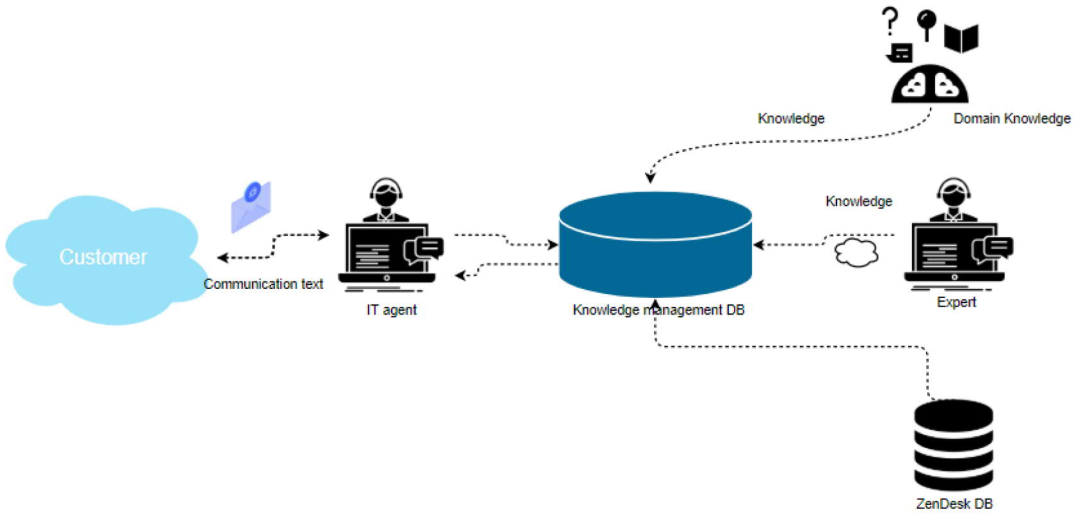


Figure 27: The knowledge-centric approach (González, 2005)

The knowledge-centric approach (Gonzalez, 2005) refers to a solution that consists of a knowledge management system that acts as an intermediary between the help desk agent and all information, data, and knowledge resources. From the above insights, it is necessary to build a knowledge-centric solution (González, 2005). Figure 27 shows the entire concept of a knowledge-centric system based on workflow automation. Thus, based on this approach, knowledge extracted from workflows can be centralized as a workflow knowledge intermediary between the helpdesk agent and data from workflows, to accelerate the input knowledge of employees. The advantages of this knowledge-centric approach to management are obvious. First, the key knowledge extracted from the workflow, including different metrics and complexity of knowledge, is transmitted through the system, thus speeding up the knowledge acquisition function. Knowledge acquisition is often a challenge (Lee and Keefe, 1996) because busy knowledge workers may neglect to capture knowledge into the system. Through the workflow analytical tool, the results from automation centralize the information and provide IT agents the skill needed and resource requirements. Based on the knowledge extracted from text mining, an IT agent can quickly integrate information, quickly and purposefully find solutions and integrate ideas. Furthermore, the management level can also make decisions based on workflow knowledge.

7 Summary

This thesis research is data-driven, trying to mine insights out of Futurice IT helpdesk workflow data. The main existing challenge at Futurice IT helpdesk is overwhelming workloads caused by increasingly cumulative issues. The thesis focused on facilitating better IT helpdesk operational management and knowledge management by applying workflow data analysis. Next, the thesis proposed three research questions: investigating how workflow data analysis result improve productivity (RQ1); finding ways to improving IT helpdesk operational management and knowledge work (RQ2), and finding ways of applying lean production to IT helpdesk operations and knowledge work (RQ3).

This thesis combined a literature review and data analysis as methods. For literature reviews, it went through literature about known theories for operational management, knowledge management, and data mining, and tried to align methods into practice. For data analysis, it focused on analyzing workflow data extracted from the ZenDesk system by applying multiple data analysis techniques, including data visualization, queuing model, text mining, and K-means clustering. After that, problems and "wastes" of time and resources in the IT helpdesk workflow were identified. Data visualization and queuing model were applied, aiming to estimate a current state of workflow and providing users indicators of workflow performance. Data analysis experiments were applied for mining more insights from customers' text: sentiment analysis aimed to mine customers' emotions while text mining aimed to separate queues and relevant issue groups.

In the discussion and recommendation section, it combined literature reviews with data analysis results to discuss insights and provided possible recommendations. Workflow health monitoring and IT issues categorization could help to improve productivity (RQ1). Text mining could help to improve operational management and knowledge work by accelerating knowledge extraction (RQ2). Workflow visibility and workflow automation were proposed solutions for converting IT helpdesk to lean production (RQ3). From the workflow data analysis, Futurice IT helpdesk could easily monitor workflow health and understand the root causes of overloaded work and long queue. It could help to impact IT helpdesk agents' decision-making process of solving issues while promoted IT helpdesk operator's decision-making process of adjusting resources and controlling cost.

This thesis also provided possible suggestions for managing IT helpdesk based on literature reviews and analysis results: 1) strategies of managing queue at IT helpdesk include controlling time, monitoring queue, controlling utilization, inspecting variations, controlling batch; 2) detecting waste in IT helpdesk workflow can be applied by combining TOC and lean production methods that could produce a better outcome. 3) strategies of improving knowledge work are investing in training and cross-training employees, understanding customer's emotions, especially bad feedback, using a knowledge-centric approach.

In summary, this thesis focuses on managing queues with multiple techniques such as data mining, queuing model, etc, for pursuing a better operation at Futurice IT helpdesk. This thesis provides insights and an example of practice for future IT helpdesk converting itself to be more lean production.

References

[1] Aranganayagi, S. & Kuttiyannan, Dr. Thangavel. (2008). Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure. 13-17. 10.1109/ICCIMA.2007.328.

[2] Ambos, Tina & Schlegelmilch, Bodo. (2009). Managing Knowledge in International Consulting Firms. Journal of Knowledge Management. 13. 491-508. 10.1108/13673270910997141.

[3] Bali, R. K. & Wickramasinghe, N. & Lehaney, B. (2009). Knowledge Management Primer. Routledge.

[4] Botha, A. & Kourie, D. & Snyman, R. (2008). Coping with Continuous Change in the Business Environment: Knowledge Management and Knowledge Management Technology, Chandos Publishing, Oxford, UK.

[5] Čiarnienė, R., & Vienažindienė, M. (2012). *LEAN MANUFACTURING: THEORY AND PRACTICE.*

[6] Donald G. Reinertsen. (1997). Managing the Design Factory - a product developer's toolkit.

[7] Davenport, T., Prusak, L. (2000). Working Knowledge: How organizations manage what they know, Harvard Business School Press, Boston, MA.

[8] Dinh DT., Fujinami T., Huynh VN. (2019) Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. In: Chen J., Huynh V., Nguyen GN., Tang X. (eds) Knowledge and Systems Sciences. KSS 2019. Communications in Computer and Information Science, vol 1103. Springer, Singapore.

[9] Frost, A. (2010). An Integrated Knowledge Management Model, Retrieved from http://www.knowledge-management-tools.net/.

[10] Fan, Weiguo & Wallace, Linda & Rich, Stephanie & Zhang, Zhongju. (2005). "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg.

[11] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu. (2019). Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary, in IEEE Access, vol. 7, pp. 43749-43762, doi: 10.1109/ACCESS.2019.2907772.

[12] Gattnar, Eva & Schneider, Jens & Dehmel, Anna & Heiland, Sebastian & Stroetmann, Brigitte & Ekinci, Okan. (2014). Knowledge Management in Consulting Activities. 255-260. 10.5220/0005086402550260.

[13] Gamble, P. R., Blackwell, J. (2001). Knowledge Management: A State of the Art Guide, Kogan Page Publishers.

[14] Goldratt, E.M. (1984). The Goal, North River Press, New York, NY.

[15] González, Luz & Giachetti, Ronald & Ramirez, Guillermo. (2005). Knowledge management-centric help desk: Specification and performance evaluation. Decision Support Systems. 40. 389-405. 10.1016/j.dss.2004.04.013.

[16] Holweg, M. (2007). The Genealogy of Lean Production. Journal of Operation and Management, 25(2), 420-437.

[17] Ho, Ricky. (2014). Common Text Mining Workflow, Retrieved from https://dzone.com/articles/common-text-mining-workflow.

[18] Haralampos, K., & Babis, T., Manchester. (2001). Knowledge Discovery in Text and Text Mining Software, Centre for Research in Information Management, UK.

[19] Hotho, A., Nürnberger, A. and Paaß, G. (2005). A brief survey of text mining. In Ldv Forum, Vol. 20(1), p. 19-62.

[20] Hui, L, & Tao, Y. (2000). Theory and methodology queues with a variable number of servers. European journal of operational research, 124, 615-628.

[21] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T.(2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[22] Kruger, D.J. (2014). Lean implementation in the Gauteng public health sector. In Management of Engineering and Technology (PICMET), 2014 Portland International Conference on (2699-2708).

[23] Kendall, D. G. (1953). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Embedded Markov Chain. The Annals of Mathematical Statistics. 24 (3): 338. doi:10.1214/aoms/1177728975. JSTOR 2236285.

[24] Kaufman, Leonard & Rousseeuw, Peter. (2009). Finding Groups in Data: An Introduction to Cluster Analysis.

[25] Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual Analysis, Dictionaries and 10-Ks. The Journal of Finance, 66(1), 35-66.

[26] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, (2011).Lexicon-based methods for sentiment analysis, Comput. Linguistics, vol. 37, no. 2, pp. 267-307.

[27] McManus, H., L. (2005). Product Development Value Stream Mapping (PDVSM) Manual Release 1.0.

[28] Mikolov, Tomas & Corrado, G.s & Chen, Kai & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. 1-12.

[29] Mahesh, Gupta & Lynn, Boyd. (2008). Theory of constraints: A theory for operations management. International Journal of Operations & Production Management. 28. 10.1108/01443570810903122.

[30] McDermott, C.M., & Venditti, F.J. (2015). Implementing lean in knowledge work: Implications from a study of the hospital discharge planning process. Operations Management Research, 8, 118-130. https://doi.org/10.1007/s12063-015- 0103-7.

[31] May, M. (2005). Lean thinking for knowledge work. Quality progress, 38(6), 33-40. Available at: https://www.ksmartin.com/wp-content/files/Lean-Knowledge-Worker.pdf.

[32] Marti, Hearst. (2003). What is Text Mining?, Derived from http://people.ischool.berkeley.edu/~hearst/text-mining.html.

[33] Newell, S. (2015). Managing knowledge and managing knowledge work: what we know and what the future holds. Journal of Information Technology, 30(1), 1-17. https://doi.org/10.1057/jit.2014.12.

[34] Omer, M. & Venkataraman, S. & Wilberg, J. & Hollauer, C. (2016). A knowledge management framework for identifying innovation potentials from use phase information, in Proceedings of the DESIGN 2016 14th International Design Conference.

[35] Piercy, Niall and Rich, Nick. (2009). Lean transformation in the pure service environment: The case of the call service centre. International Journal of Operations & Production Management - INT J OPER PROD MANAGE. 29. 10.1108/01443570910925361.

[36] Richardson, Helen. (1995). Control your costs–then cut them. Transportation & Distribution, Vol. 36 Issue 12, p. 94.

[37] Rahani, A.R., & al-Ashraf, M. (2012). Production Flow Analysis through Value Stream Mapping: A Lean Manufacturing Process Case Study.

[38] Richard, L. Frei & Michael A. McDaniel. (1998). Validity of Customer Service Measures in Personnel Selection: A Review of Criterion and Construct Evidence, Human Performance, 11:1, 1-27, DOI: 10.1207/s15327043hup1101_1.

[39] Rachman, A., & Ratnayake, R.M. (2016). Implementation of lean knowledge work in oil and gas industry - A case study from a Risk-Based Inspection project. In Industrial Engineering and Engineering Management (IEEM), 2016 IEEE International Conference on (675-680). https://doi.org/10.1109/IEEM.2016.7797961.

[40] Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math.

20, 53-65. Journal of Computational and Applied Mathematics. 20. 53-65. 10.1016/0377-0427(87)90125-7.

[41] R. Marcella, I. Middleton. (1996) The role of the help desk in the strategic management of information systems OCLC Systems and Services, 12 (4). pp. 4-19

[42] R.J. Sharer.(1998). Applying policy management to reduce support costs for remote and mobile computing. International Journal of Network Management, 8, pp. 211-218

[43] Staats, B.R., & Upton, D.M. (2011). Lean knowledge work. Harvard Business Review, 89(10), 100-110

[44] Staats, B.R., Brunner, D.J., & Upton, D.M. (2011). Lean principles, learning, and knowledge work: Evidence from a software services provider. Journal of Operations Management, 29(5), 376-390. https://doi.org/10.1016/j.jom.2010.11.005

[45] Sergio, Rattner. (2006). What is the Theory of Constraints, and How Does it Compare to Lean Thinking? Lean Enterprise Institute.

[46] Shah, R. and Ward, T. (2007). Defining and Developing Measures of Lean Production. Journal of Operations Management, 25, 785-805.

[47] S. Kasim, S. Deris, R.M. Othman Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data Compute. Biol. Med., 43 (2013), pp. 1120-1133

[48] Saurav, Kaushik. (2016). An Introduction to Clustering and different methods of clustering. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[49] S. Sandborn. (2001). Structuring the service desk Information World, 23 (52), p. 28

[50] Tyagi, S.K. & Choudhary, Alok & Cai, X. & Yang, Kai. (2014). Value stream mapping to reduce lead-time of a product development process. Int. J. Prod. Res.. 160. 202-212.

[51] Vishal, Gupta & Gurpreet S. Lehal. (2009). A Survey of Text Mining Techniques and Applications.

[52] Tiwari, S. K. & Gupta, V. K. Gupta & Joshi, T. N. (2016). M/M/S queuing Theory Model to Solve Waiting Line and to Minimize Estimated Total Cost.

[53] Taha, H. (1987). Operations research. Fourth Edition,Collier Mac Willam

[54] X. Wu et al., (2017). Investigation on sentiment of reviews with shopping field dictionary construction, Comput. Technol. Develop., vol. 27, no. 7, pp. 194-199.

[55] Zhu, Rui & Yang, Delu & Li, Yang. (2019). Learning Improved Semantic Representations with Tree-Structured LSTM for Hashtag Recommendation: An Experimental Study. Information. 10. 127. 10.3390/info10040127.

[56] Zhou, H. B., & Gao, J. T. (2014). Automatic Method for Determining Cluster Number Based on Silhouette Coefficient. Advanced Materials Research, 951, 227–230. https://doi.org/10.4028/www.scientific.net/amr.951.227