Oriented Development of Enterprise Message Management: Study on Visual Attention of Email Topic Inference (AttLDA for Email) and Integration of ECS and ERP (SuccERP)

Yung-Yu Lin

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Oriented Development of Enterprise Message Management: Study on Visual Attention of Email Topic Inference (AttLDA for Email) and Integration of ECS and ERP (SuccERP)

エンタープライズメッセージ管理の指向的開発：電子メールトピック推論の視覚的注意（電子メールの AttLDA）および ECS と ERP の統合（SuccERP）に関する研究

Yung-Yu Lin

Supervisor: Yukari Nagai

Graduate School of Advanced Science and Technology

Japan Advanced Institute of Science and Technology

Knowledge Science

March, 2021

ABSTRACT

Our dissertation is mainly focusing on several topics for improving collaboration and communication in an enterprise. Come with considering two features of collaboration, unstructured collaboration (information collaboration) and structured collaboration (process collaboration); we primarily focus on two representative applications: email and Enterprise Resource Planning (ERP) System.

In terms of an enterprise, most of the current research result struggles to achieve specific and practical goals by proposed theoretical findings in the ERP domain. To allow the managers to get a fuller picture of all the messages generated from an ERP system with the Enterprise Collaboration System (ECS) and improve collaboration and communication, we propose a complete method to develop an artifact-SuccERP based on the Design Science approach to carry out the integration. Based on exploring multiple ERP systems, we summarize our tasks into three aspects before implementing the integrations: authentication, data initialization, and specific procedures implementation; we also explain how the data-processing and integrations between the ERP and ECS.

In our perspective, we can distinguish our contribution of the proposed SuccERP into two parts; 1) We present a complete demonstration of how to get the architecture and database schema of an existing ERP system and address the internal and external hosting issues. 2) According to a series of literature reviews, we implement the integration based on the critical success factors and existing issues discussed in the previous studies. In other words, we attempt to fill up the gap in communication and collaboration capabilities by enhancing the ERP and ECS systems' integrations. Meanwhile, we fulfill the data-processing and data-sharing from an ERP system to the external resources. Given the context of the increasing demands of custom ERP, it is reasonable to provide elaborate research as a guideline to those enterprises that plan to upgrade and enhance their ERP systems. Furthermore, based on our results, follow-up research can explore the implementation with other external resources for improving different issues.

Next, the definition of information collaboration is employees applying IT tools to communicate and request assistance (answer); email is the most standard documentation tool for communication. Although existing studies use the topic model

to support users for classifying emails, they disregard that humans are not like a machine that can focus on all the words in an email to determine the distribution of email topics. The Latent Dirichlet Allocation (LDA) model forms a basis for inferring topics; our work aims to discover how each word's visual attention influences the topic inference and estimates attention to a word according to its location features.

By reviewing the visual-spatial research and the state-of-the-art visual attention models, we select the Bayesian Models to estimate attention and proposing a novel model-Attention orientation Latent Dirichlet Allocation model (AttLDA). In AttLDA, each email can regard as encoded into a two-dimensional space, taking the line length (the characters per line in an email) and window size (the number of lines in an email) into account. After that, draw the optimal display size as a visual space and assigning each word's location. Besides location, attention estimation also considers the Term Frequency and Inverse Document Frequency (TFIDF) and inferred topics for each iteration. Our aim is as follows; the readers can not completely capture all the hidden topics behind each word in an email, especially the context in the forwarded message. Moreover, our result shows each email's topic distribution and including the distribution of related words' attention in each topic by considering the visual attention as the significance of an email's topic distribution. In our experiment, we consider the public Enron email corpus as a dataset and apply the Perplexity metrics to measure the performance of AttLDA. AttLDA is outperforming the previous research on the perplexity evaluation.

Advanced technology has made the communication distance between people shorter than ever before and accumulates the number of messages quicker and quicker. People might quickly out of control for managing their messages owing to their negligence. Our research proposes the SuccERP, which builds a platform to manage ERP and ECS messages through definite guidelines to keep communication efficiency. On the other hand, we also proposed the AttLDA to effectively extract the email topics to improve email message management performance. The research findings can be regarded as a strategy for settling further tasks relating to collaboration in an enterprise.

Keywords: Latent Dirichlet Allocation (LDA), Visual Attention, Email Management, Enterprise Resource Planning (ERP), Enterprise Collaboration Systems (ECS), Design Science (DS).

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Prof. Yukari Nagai for the continuous support and guide of my Ph.D. study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank my mentor and supporter Prof. Tzu-Hang Chiang for his encouragement, insightful comments, and providing explicit direction at a time when I may have gotten my directions wrong. I am also grateful to Prof. Tsutomu Fujinami, who taught us the cognition science for developing a research more focus on human demands. To inspire our imagination and innovation, we often have a group discussions with research students from different labs and areas, it opens a big door to meet the requirements and questions from different viewpoints. Also, it is my honor to let Prof. Tsutomu Fujinami to be my advisor for the Minor Research Project, he provides me timely advice and many valuable suggestions while I do the internship for my minor research.

Further, I wish to extend my gratitude to my lab mate Mr. Zhao Ching, who is the excellent Ph.D. candidate in the Nagai lab. During the COVID-19 pandemic, all of the lab discussion and meeting is considering the remote way, I can't imagine what would have happened without his kindly help.

Finally, my sincere thanks also extend to my family—especially to my mother for her long-lasting encouragement and eternal support in my study and life—and I would like to thank my husband and son for their understanding and recognition. I thank my other family members for their support throughout my doctoral study and my life in general as well.

Thank you for all your encouragement and support!

Yung-Yu Lin

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 General Introduction

With more and more employees working in complex and knowledge-intensive environments, the emphasis on demand for collaboration is increasing [1]. To this end, Prakash et al. [2] indicated that IEEE had proposed an explicit definition of collaboration, which implies the capability of two otherwise more systems to share and use the knowledge, especially emphasizes the information flow aspect.

In this context, collaborative application development is gaining attention. Enterprise Collaboration System (ECS) is a system designed for the enterprise side to implement collaboration conception, which has been recognized as a vital enabler of the modern digital workplace in the current environment [3]. Prakash et al. [2] emphasized that collaborative systems address administrative work by facilitating the sharing and diffusion of information; however, two fundamental features need to be customized and paid attention to depending on the organization's goals. These two features are unstructured collaboration and structured collaboration.

Unstructured collaboration, also known as information collaboration, is used to find answers to unknown questions by utilizing IT tools, such as emails. Instead, structured collaboration is also known as process collaboration, allowing business processes to be shared by sharing common information, structured, written rules, and set workflows. Specifically, for the collaboration and business process, a portion of the researchers have identified the integration between Enterprise Resource Planning (ERP) and external resources as one of the most critical fields. They also claim that system integration is crucial in digitizing business processes and creating business value for future IT value [4]–[7].

Hence, regarding the two notable features of the collaboration mentioned above, for the part of the unstructured collaboration (information collaboration), we consider email as the target for exploring how to improve the integration process in collaborative systems. On the other hand, for structured collaboration (process collaboration), the issues related to improving an enterprise's business process through system integration have attracted our attention.

1.1 Motivation for focusing on the email domain

According to the Email Statistics Report [8], they indicate a clear sign about how the issues of huge amount and complexity from email. The amount of email will exceed 93 billion in 2019 and continue to grow at a rate of over 4 percent per year. While email usage from ubiquitous to regular and grew into our life, we are not aware of how important it is in daily life. Hence, managing and maintaining those emails in our daily life has become a primary concern subject.

Whittaker and Sidner [9] present that people consider email as a tool for communication application in the beginning when it comes to email. However, people apply the email for task management and personal archiving that were not initially designed. Hence, the problem of email overload becomes a well-known issue in this domain. Especially in an organization, users often with inboxes containing hundreds of email messages, including tasks, partially read documents, and conversational threads. The story must proceed to the result where important messages get overlooked or lost in archives. Therefore, regarding email as an asynchronous communication application, the critical requirements for improving the management of communication are:

1. The management of email threading to support context and conversational history of email.
2. To track the status of a conversation from the email collection.
3. Find out the exact subjects (topics) from the messy email threads.

Moreover, some scholars consider task management to the email to apply the conversational threading and semantic clustering techniques to reduce the amount of inbox. In short, keeping important things (email) “in-view” [9].

A subsequent study considered the utility of various techniques for managing email and redefined email overload after a decade. Whittaker and Sidner [9] regard the primary reason to meet the email overload as the reliance on the email client for personal achievement and tasks beyond communication. However, Dabbish and Kraut [10] define email overload as users' perceptions that their email use has gotten out of control (p. 431). They further revealed that white-collar workers would attempt to moderate the stress from high email volume in the inbox via checking email boxes frequently for reducing email stress. As email is fast becoming the primary means of

communication for sharing information in an organization, it is necessary to propose a more passive solution and organize email to impact this domain significantly [10].

Afterward, Fisher et al. [11] review Whittaker and Sidner's work [9] to estimate the change of email overload circumstances in the previous decade. The increasing familiarity with management techniques would reduce the number of emails keeping in the inbox. Interestingly, the likelihood of people adopting or keeping an email management strategy is not increasing over time. Even the archiving behavior is more general than before, and there is no evidence to support the relationship with the reduction in the inboxes size. Fisher et al. [11] further show that email overload will continue growing and need more research to put effort and attention into this domain to determine the most effective email management method.

Jerejian et al. [12] present the findings that several studies investigate various email management ways to reduce employees' stress from organizing emails such as filing, filtering, periodic checking, or constant monitoring. However, the result of findings has been mixed as to the success of email management strategies. Individual differences may explain it among participants in the experiment. However, in part, we should also consider the necessity of email management that needs to look for advanced techniques to improve the efficiency of organizing emails.

Skyrme [13] gets down to practicalities in a set of tools to introduce how you get people to share their knowledge and the best start with knowledge management. Regarding email as a tool, Skyrme further points out that the use of email in business must be more focused on effective communication improvement and suggested that it would be better to restrict an email's content to only one topic. Like the above suggestion, we have mixed feelings about it. On the one hand, we agree that the restriction on topics of an email. On the other hand, it is relatively ideal for an email with only one topic, but there are mixed topics in an email in most cases.

Our intention in the email domain is to show the way how to effectively improve communication and collaboration activities in an enterprise by identifying email topics. The intention is first motivated by the issues of huge amount and complexity from email; and then how, under the reliance on the email client for personal achievement and tasks beyond communication, it has caused another issue-email overload. Researchers have begun to examine white-collar workers' behavior and analyze management approaches' effectiveness to overcome email overload. However, the result of findings has been mixed as to the success of email management strategies. They suggest looking for

advanced techniques to improve the efficiency of organizing emails. In this way, considering collaborative work and email overload issues would be a worthwhile research question to help employees effectively identify topics from email and further improve communication in an organization.

1.2 Motivation for focusing on ERP integration domain

Subsequently, we address another feature of collaboration – structured collaboration (process collaboration). For the discussion of ERP and beyond, Langenwalter [14] has created a term, Total Enterprise Integration (TEI), which is used to describe the process of integrating all the information and actions between the manufacture and supply chain sides. TEI aims to create a strategic advantage for crossing the entire manufacture and out through the customers and suppliers. In short, it is in communication to let every member know the whole process of a new customer order from received to complete, such as the current status of supplier shipment or the discussion on a machine for the unscheduled maintenance. It makes the workflow allocate the organization's responsibility to the most appropriate decision-makers while maintaining proper budgetary. TEI is an intelligent use of an enterprise, while internal costs continue to rise, and competitive pressures threaten manufacturers' ability to survive by decreasing prices. The solution is reducing costs by eliminating ineffective communication, collaboration, and coordination throughout the manufacturing company, suppliers, and customers.

Zheng et al. [15] present the integration between Supply Chain Management (SCM) and ERP Systems, aiming to generate a single and seamless system with higher efficiency and productivity. People conduct the business process related to suppliers and manufacture within the same system rather than two independent ones. Thus, integration regards as a complete information-sharing platform for improving the performance of respond from customers to the supply chain. The result shows the comparison between SCM and ERP system as an evaluation which comprises application, functionality, relationship, and explore their potential integration. Samiei and Ehsan [16] present a review study on the relation between knowledge management (KM) and enterprise resource planning (ERP) after the implementation stage. They also discuss the possible synergy between KM and ERP and the various integrations that could strengthen knowledge generation and information sharing in organizations.

TEI describes building a strategic advantage by integrating all the organization's information to eliminate ineffective communication, collaboration, and coordination. According to both Zheng et al. [15] and Samiei and Ehsan [16] findings, it is more straightforward to understand the purpose and goal of achieving system integration, aiming to generate a single and seamless system to allow business processes within the same system rather than two independent ones. The integration is a further step towards sharing information within an enterprise and organization. Simultaneously, as we previously mentioned, structured collaboration allows business processes to be shared by sharing common information, structured, written rules, and a set of workflows. They claim that system integration is crucial in digitizing business processes and creating business value. Referring to these relevant literature studies, we have a better understanding and confidence that improving business processes through ERP system integration can effectively enhance collaborative capability within an organization or enterprise.

With the emergency of global value chains and competitive business environment, we further elaborate on the ERP system. Tarn et al. [17] indicated that ERP systems have been regarding as the fundamental element of an enterprise. However, today enterprises are no longer concerned with the issue of whether they need an ERP system but how to build an effective ERP system. Admittedly, they believed that enterprises should further extend their ERP systems to integrate inter-company and collaborative operations across the entire industry processes rather than inter-department, inter-office, and inter-site integration within a single organization. In short, from existing ERP systems to shift toward Extended ERP (EERP) system. Meanwhile, the market and demands for ERP post-implementation upgrades and services, or called the second wave of ERP, are continually growing, and only a relatively small amount of study is currently focused on this field [5], [17], [18]. Accordingly, as things are, we summarize several keywords for the motivation on the ERP domain, including structured collaborative, business process, integration, and ERP post-implementation.

From the structured collaboration feature mentioned above, we follow the business process, ERP system integration, to which life cycle stage is an appropriate one for ERP system integration and the noteworthy issues. First, an ERP project's life cycle consists of three stages: ERP adoption, implementation, and post-implementation [19]. Second, Kosalge and Motwani [7] argue that the implementation stage's life cycle is not the end

in the ERP implementation; timely adding new capabilities, modifications, updates, and new improvement should be applied in ERP systems and the changes in processes. Third, some of the studies, however, point out that existing ERP literature tends to focus on the topics related to the implementation of the Critical Success Factors (CSFs) and implementation methodologies [21]–[23]. The scholars further indicate that only very few studies focus on other aspects of ERP applications [24]–[27], based on referring the current popular ERP research focuses on identifying the CSFs to ERP implementation stage as the factor research [20].

These literature results and suggestions open the door to studies that achieving the integration between ERP system and external resource for improving the collaboration by integrating the business process into a single and seamless way. Our intention starts with the feature of structured collaboration; subsequently, considering a portion of works claims that ERP integration with external systems is a required field and can create business process value. Hence, we explore different ERP system integration studies, including integrating SCM and ERP [15] and integrating KM and ERP [16]. Based on that, we answer two questions first below:

1. How to improve the collaboration within an enterprise: ERP system integration.
2. What is the purpose to implement the ERP system integration with external resources: Generate a single and seamless system to enable business processes within the same system rather than two independent ones.

Finally, from the ERP implementation life cycle, we further ensure the post-implementation stage is an appropriate one to implement the ERP system integration. Besides, we found that merely exploring CSFs in the implementation and post-implementation stage is necessary but not enough.

Hence, recent research convinces us that aiming to build the integration between ERP and ECS to improve collaboration and communication that such an approach can produce improvements for the business. Moreover, under unavoidable circumstances, each independent system is hard to share their local information and data with others; thus, the issues of Information Island [28], [29], poor communication, collaboration, and coordination [14] receiving increasing attention for exploring the possibility of integration with two or more systems. According to that, we believe there is a gap in providing a practical solution and guidelines to let the entrepreneur achieve their demands as far as we know. It makes them share the data from the ERP system side to

the external resources to fix the issues, including Information Island, poor communication, collaboration, and coordination.

We make a short conclusion to sum up this chapter; chapter one describes this dissertation's general content. Go through a series of literature reviews to establish the research background for interpreting aims, including:

1. Why are we concerned about improving email management, and how does this relate to enterprise collaboration?
2. Why do we attempt to create the artifact to achieve the integration between an ERP and ECS?

The above two questions are the primary research aims we will focus on and implement in the following section. First, people consider applying email as a tool for communication application. We explore how to improve email management performance based on Natural Language Processing (NLP) techniques. On the other hand, the ERP system becomes fundamental to an enterprise; besides the theoretical result, we expect to provide guidelines and steps to integrate with external resources. Before outlining the aims in detail, we present the motivation and purpose of this study. Moreover, we list out the tasks we expect to cover through this research. Finally, we will briefly review each chapter at the end of this chapter.

1.3 Research Aims

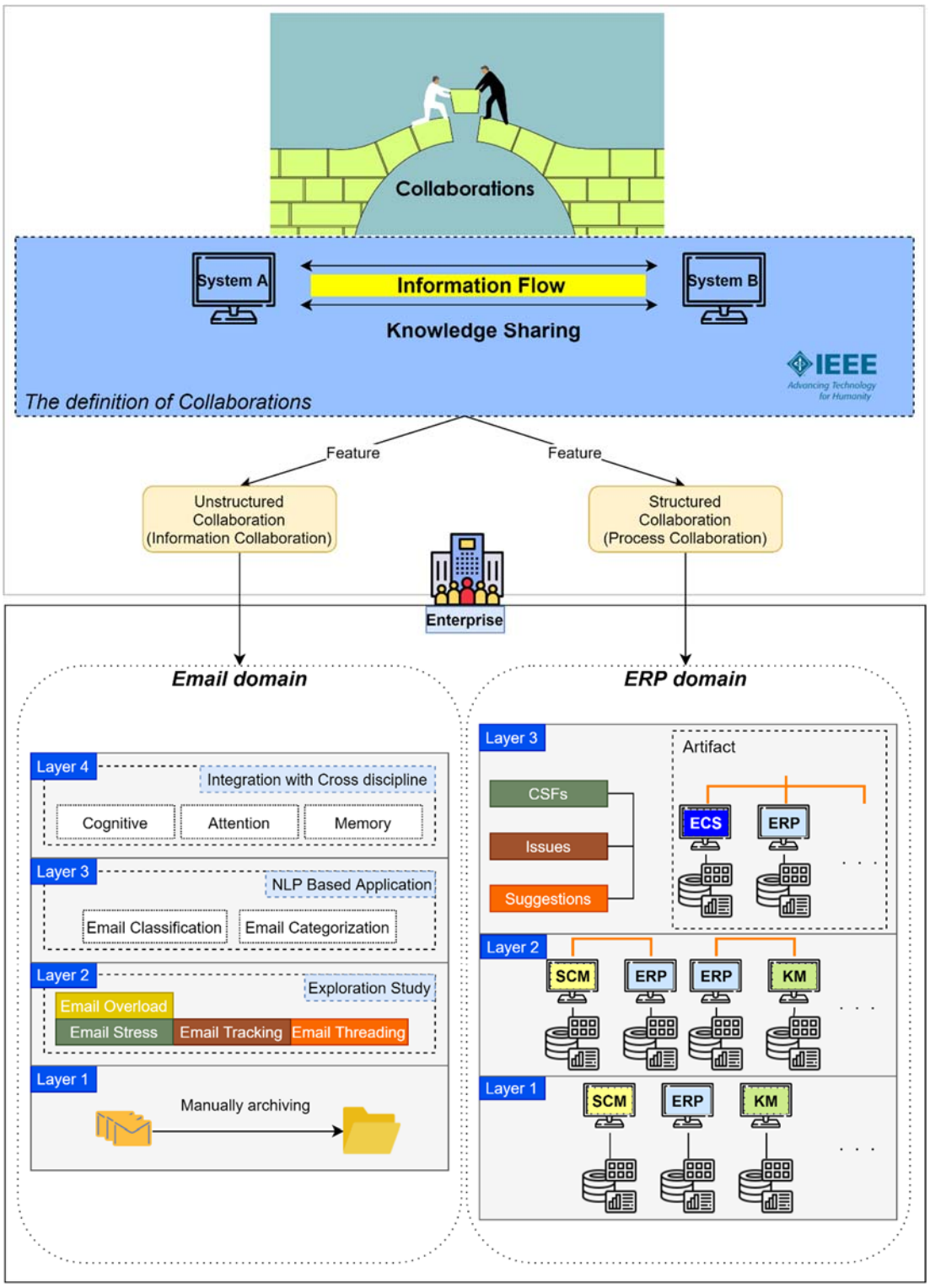


Figure 1.1 Research aims from the state of existing studies.

We discuss the state of the current study on both email and ERP, showing our work's relevance to a broader problem. We draw out this illustration to describe our

research in Figure 1.1, which aims to review studies beginning from the collaboration demands, collaboration definition, and two features of collaboration until the email and ERP domain. We expect to carry out the top layer (Layer 4 in the email domain and Layer 3 in the ERP domain) in our research. In the below description, we interpret the layer one by one to show how we construct our research aims and purposes.

About the email domain:

1. Layer 1: This is the fundamental stage for the exploration study, while many issues come out by the inconvenience of manual actions. The researchers begin to be involved in this domain and exploring the key factors that cause these inconveniences.
2. Layer 2: The exploration study differs by issues from layer one that including:
   (1) Review the definition of term or conception to describe the current issues of email management.
   (2) To inspect the influence of current email management issues and present the findings from questionnaires or interviews.
   (3) Highlight advanced solutions as a suggestion to the current issues of email management based on experimental results.
3. Layer 3: In addition to the discussion on management, some studies began to focus on the data of the email, including the subject, content, recipients, and other information to go through the data preprocessing, features extraction, and analysis to achieve specific research purposes such as spam and ham email classification, email folder classification, and email categorization. Although their studies' purpose is not exclusively on improving communication, the findings and results are meaningful for communication improvement.
4. Layer 4: For building an application that can serve a social purpose, we need to take more factors into accounts, such as attention, cognitive science, and memory aspects, with NLP to achieve our aims. Our idea is while we try to figure out the topic distribution of an email, we need to consider the limitation of attention; in our opinion, people will allocate their attention to those important words or sentences they think of in an email document. To a certain extent, the following study supports our viewpoint, Dehaene et al. [30] indicated that visual attention is a prerequisite for consciously seeing the information. In general, people consciously shift their attention to perceive those events. However, only a few studies considering what they need to serve are humans, not machines.

About to the ERP domain:

1. Layer 1: Each system likes an isolated island, neither unable to share the data and information nor to implement the communication and coordination in an organization. However, these issues are not given much attention at this layer since many studies are focused on choosing an appropriate ERP system for the organization and the CSFs of the ERP system.
2. Layer 2: The researchers noticed the problems arising from Layer one and began to discuss the possibility of interaction between different systems based on each system's characteristics. To a certain extent, the findings and suggestions of CSFs could form a research basis for further study. For example, Zheng et al. [15] present the integration between SCM and ERP systems; the findings show that the comparison between SCM and ERP system as an evaluation comprises an application, functionality, relationship, and explore additional potential integration. However, merely exploring the potential for integration is not enough to satisfy entrepreneurs' and developers' demands. Meanwhile, some studies explore system integration for different industries. Lancharoen and Suksawang [31] discuss the information system integration in a hospital, and the goal of the integration is to improve the capabilities of communication in a health care organization. The findings demonstrated that system integration could have a significant impact on service efficiency and physician decision-making.
3. Layer 3: For constructing an application that can integrate with the ERP system and other external resources and systems, we focus on showing which necessary steps to be taken for achieving the purpose of interaction. Therefore, the findings can meet business owners' or developers' actual needs and enable data and information sharing with multiple legacy systems. At the same time, a company adopts a new system in the future. The study in this layer belongs to the post-implementation study of ERP system, considering the suggestion from Amid et al. [32] that current research has focused on software selection, implementing process, and CSFs rather than the success, actual benefits, and performance improvements of the ERP post-implementation stage. Hence, proposing a practical and effective way to improve ERP systems' performance is a primary research direction in this layer.

1.4 Research Objectives

From the perspective of enterprise collaboration, we focus on two collaboration features: structured collaboration and unstructured collaboration. The overall purpose of this study summarized into four objectives as listed below.

(1) Identify the state-of-the-art model based on the NLP techniques, and discuss the possibility of inferring topic distribution by applying one of the models. The inferred topic distribution can also consider as a feature for further tasks such as email classification, email categorization, and email thread clustering.

(2) Design a topic model based on the NLP techniques and integrate with visual attention and cognitive science to serve the human demands on email management in an enterprise.

(3) Go through the complete steps and rigorous methods to build and proposing a system as an artifact to integrate an ERP system with other external resources and systems.

(4) Exploring multiple ERP systems to summarize the differences and presenting practical cases to prove the proposed artifact's functions can working well. Also, considering evaluation metrics to ensure the artifact is easy to design, implement, and maintain.

1.5 Structure of this Dissertation

1. Chapter 1: Introduce an overview of this dissertation that we concentrate on two specific domains, comprising email and ERP. Besides, it also includes motivation, research aims, objectives, and structure. This chapter briefly explains the problems, goals, and structure of this dissertation.
2. Chapter 2: Review the studies at the management level of email and explain our motivation based on the literature. Subsequently, from the ham and spam email research works, we highlight the LDA model's utilization and application. Further, some research begins to apply the term weighting function to improve the performance while applying the LDA model to achieve the topic inference. In the end, we consider visual attention as a kind of term weighting function and review the necessary literature for building and implementing visual attention by proposing the AttLDA model.
3. Chapter 3: After deciding the LDA model as our basis, we recognize the attention as a weighting function to form the AttLDA model during topic inference. According to the literature reviews of visual attention, we have ensured our attention estimation implementation follows the theoretical approach.
4. Chapter 4: By considering the open data set, we demonstrate how the AttLDA works and applies the heat map to show the value of attention on each topic's words. Before showing the visual attention distribution and topic distribution, we present how we determine the number of topics from the Coherence Score results. Further, applying and demonstrating the perplexity measurement as the evaluation metrics for further comparison.
5. Chapter 5: We are first introducing the life cycle of ERP implementation and then explain why we focus on the post-implementation stage. We also introduce the Design Science (DS) to create an artifact and provide complete guidelines for constructing the utility application. Subsequently, we review the highlights and issues of ECS that ECS supports collaborative work. People have examined it as an enabler of the modern digital workplace for longitudinal work; however, ineffective ECS content is a serious issue. Hence, we also review the suggestion and critical idea of ERP in the post-implementation stage and the DS approach

from the previous works to consider building research to overcome the issues we encountered.

6. Chapter 6: We refer to the six activities and seven guidelines from previous research suggestions of DS, including problem identification and motivation, define the objectives for a solution, design and development, and the demonstration and evaluation. Not only concentrates on the steps of DS, as well as considering the Software Engineering into part of activities to improve the quality and more explicit for building an information system as an artifact. Later, we collect the post-implementation issues of an ERP system and provide the practical steps in this chapter as a solution. Moreover, we consider multiple ERP systems to summarize the differences between them and prove our solution can work well on different ERP systems.
7. Chapter 7: After a series of DS activities in the previous chapter, we focus on the findings of ERP integrations and show the metrics to ensure our proposed artifact-SuccERP is easy to design and maintain, which is the part of the demonstration and evaluation.
8. Chapter 8: Discusses findings from all analyses and highlights original contributions to the enterprise on the Email and ERP domain, including theoretical and practical implications. Limitations and recommendations for future studies, and we also consider the inspiration for the subsequent studies.

CHAPTER 2 Introduction and literature reviews of Email Topic Identification

Email is the most popular and standard tool for communication over the business. By employing NLP techniques, Email text mining is becoming a popular research topic [33]–[36]. From the study of trends review presented by Mujtaba et al. [37] that we have a basic understanding of email research, nearly 65% of the email studies focus on exploring the spam and phishing email classification, about 20% concentrate on the multi-folder categorization and email clustering. In the rest, only limited studies consider how to overcome the issues of losing control and management, no matter of individual or organization.

Recently, Latent Dirichlet Allocation (LDA) has received increasing attention from researchers in the email research domain, in which applying the topic model to automatically clustering email into each group to improve the efficiency of email management. They applied it to extract the topic distribution as features for achieving some objectives, such as multi-folder categorization, spam detection, and email thread identification. For example, McCallum et al. [33] explore the relationship and roles between authors and recipients in an email; they regard the LDA model with different roles to identify positively related topics and keywords. Sharaff and Nagwani [36] propose a two-stage clustering approach, including the LDA model and the Non-negative Matrix Factorization (NMF) to automatically identify the email thread. However, multiple mechanisms equip our brain that grants us to filter out irrelevant information to prioritize relevant information, referred to as “attention” generally [38].

Differ from previous works, this paper intends to explore a new problem: We consider the visual-attention allocation while applying the LDA model to inference the latent topics from an email document. Our work aims to propose a novel model-Attention orientation Latent Dirichlet Allocation (AttLDA). Each email is a two-dimensional space to estimate each word's attention value for inferring the latent topics according to state-of-the-art visual attention models. Each word's attention estimation is mainly determined by three factors: location, inferred topics, and the Term Frequency and Inverse Document Frequency (TFIDF). Based on the relevant literature on visual attention, our research makes the topic inference closer to reality. We identify the

attention allocation has the nature of a theoretical basis with numerous scientific examinations that have investigated that visual attention is a prerequisite for conscious retrieving information. Simultaneously, people in general only consciously perceive those events, onto which they direct their attention at a given time [30].

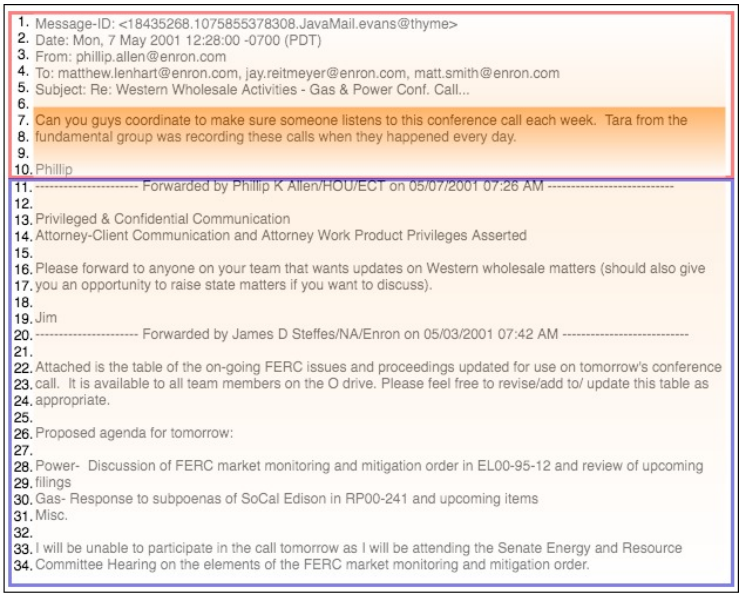


Figure 2.1 The conception for the attention allocation when reading an email.

In Figure 2.1, we describe the brief conception of the proposed AttLDA model. Further, the red box includes the header and actual email content, and the purple box is the quotation content. Here we apply gradient color to indicate how we assumed that the ideal attention allocation works while reading an email for topic inference. Our motivation comes from an email that brings a large amount of information, but people read only part of it practically. The preliminary work also supports this viewpoint; they indicate the header, signature, and quotations (such as a forwarded message or original message) are typical email noises [39]. As illustrated in Figure 2.1, from the line 1 to 5 are a header, line 10 is a signature, and a quotation lies from line 11 to line 34; only lines from 7 to 8 are actual email content. As far as we know, existing email studies did not address these issues while implementing topic inference. Therefore, in this paper, we assume people will allocate their limited attention to a part of an email's words to identify the topics that are latent variables behind the words. More precisely, if we recognize the topic distribution as a feature without considering the effect of attention allocation, there is a significant deviation from the actual cases.

The experimental dataset is the Enron Email Dataset, which is a public email dataset. Some well-known organizations provide this public dataset in CSV format by going through email cleaning and data normalization processes. However, in our case, we download the original files in UTF-8 encoding, which keeps all the original formats for drawing a two-dimensional space for each email and calculating each word's location. Our codes will release at GitHub.

From the general introduction until here, we can summarize our aims and purposes into a more explicit description below:

1. As pointed out in the trends review [37], nearly 65% of email studies focus on spam and phishing email detection. Another 20% of email studies focus on multi-folder categorization and email clustering. We have identified our research aim focuses on the relatively lacking part.
2. From the previous study's suggestion [39], most of the email content is noise data. Based on that, we propose a visual attention allocation approach to simulate that people will put less attention to the noise data practically and make the topic inference on an email by applying the LDA topic model.
3. In order to derive the attention distribution, the dataset we applied is as similar as possible to the actual format people read. Instead of downloading processed CSV format, we consider the original text files and conduct various processes to extract the sentences and words for the attention estimation.

2.1 Literature reviews for implementing the AttLDA model

Firstly, we review the current email research from ham and spam email categories. To our knowledge, the email research can roughly divide into two fields: one is to filter out those misuse/abuse emails from the dataset; they classify email into a specific class, such as spam, phishing by creating a classifier. The other one focuses on a ham email that tries to improve the user experience by considering how to manage and arrange those useful emails, such as topic extraction, email thread identification, and multi-folder categorization. Subsequently, we review the relevant studies that have used LDA models to implement email analysis and discuss some extension applications. In the end, we review the studies that motivate us to consider visual attention as a mechanism to combine with the LDA model for topic inference.

2.1.1 The existing works on spam email detection

In the first field, there is plenty of research focused on spam and phishing email detection.

Youn et al. [40] propose two ontology levels for the spam filter, including Global and User customized ontology filter. They consider the TFIDF value as an email's feature and apply it to different classification algorithms to generate the RDF file for creating ontology via Jena. The C4.5 gets the best result in the experiment than the naïve Bayesian, neural network, or SVM classifier. The result shows the Global ontology filter successfully detected 91% of spam emails. They further take into account the different background of each user to implement the User customized ontology as well as the second level ontology filter to achieve a more user-customized, scalable, modularized email spam filter.

Smadi et al. [41] present a framework to explore new phishing behavior from the online mode's newly received email. The motivation is trying to handle zero-day phishing attacks and solving the problem of the limited dataset by automatically adding email data to the offline dataset in the online mode. The proposed framework can dynamically select features from fifty features by applying the Feature Evaluation and Reduction (FEaR) to rank feature weights and then dynamically change the extraction of the number of features.

Ramanathan and Wechsler [42] explore the field of phishing email detection by proposing a multi-layered approach - PhishGILLNET, which comprises three layers. The first layer employs the Probabilistic Latent Semantic Analysis (PLSA) to build a topic model; the second layer considers the feature from PLSA to build a robust classifier by AdaBoost. The third layer applies Co-training to handle the labeled and unlabeled email data. In this study, the topic model's use attracted our attention, and there are several processes before they apply the topic model that is worth describing here. First of all, they consider Porter’s stemming algorithm [43] to remove the inflectional endings from certain words after a series of processes, including parser, tokenize, and stop words removal. Subsequently, they employ the Part-Of-Speech (POS) extractor from the WordNet dictionary [44] to look up the words in the dictionary and identify verbs, nouns, adverbs, and adjectives. Afterward, words in an email are verified

by WordNet and utilized Google's spell check Application Program Interface (API) to retrieve similar words to the misspelled word. The results obtained from the above process form a component to build the Term Document Frequency (TDF) matrix. In the end, they conduct the topic model PLSA to identify the synonym, polysemy, and other linguistic variations found in a phishing email.

Méndez et al. [45] declare a feature selection method that considers the topic from an email and applies semantic ontology to get the semantic relation and lexical conceptions. The goal is to improve the topic model by considering the taxonomic relations between synsets (hypernym and hyponymy) in the semantic ontology. They further argue that words with similar lexical meanings should be identified as the same topic, allowing the topic model to provide more consistent results for the synonymous words in topic inference. In the experiment, besides the proposed method of Semantic-based feature selection, they also apply the Information Gain and Latent Dirichlet Allocation (LDA) model as feature selection methods to compare and evaluate spam filtering by eight popular classifiers. As a result, the proposed method, Semantic-based feature selection, has a better performance than the other two methods.

2.1.2 The existing works on ham email research

Except for the concentration on excluding spam and phishing emails, part of the works is trying to achieve more effective management of ham emails by clustering or categorizing technical methods.

Alsmadi et al. [34] define five classes for folder categorization that email users' nature may consist of Personal, Job, Profession, Friendship, and Others. Subsequently, they apply K-means clustering to group email into five pre-defined clusters, randomly selected the centroids, and show magnificent performance in the evaluation stage; however, the dataset is a private dataset from Gmail. Besides, considering time-consuming, they only apply the top 100 most frequent words to execute the experiment with the entire email set. Hence, the principal weaknesses were the following researcher unable to compare the performance with their result while their dataset is not public. Several studies [46]–[48] also consider custom or private email datasets for their research that is a serious problem encountered in this research domain.

On the contrary, many studies use public email datasets to conduct experiments and evaluations, which are more rigorous and provide a performance comparison for subsequent studies.

Dehghani et al. [35] propose a solution with heterogeneity and dynamism mechanism in email categorization. Their approach constructs two unique structures for email conversation threads. One is linear; all the emails belong to one conversation thread and organize in chronological order. The other is a tree structure, which calculates the similarity between replies that extracting their relationship to create the tree. As discussed in this work, they assumed that the user might be based on content-aware, time-aware, or participants-aware structural aspects to categorize an email. By grouping features into several sets, a specific structural aspect can represent by a set of features. For simulating user behavior, the study investigates an email from its different structural aspects to the categorization problem. It tests the similarity score between an email and a folder by defining the functions that can access the belonging feature space as input for each aspect. Each function generates a probability to indicate the folder assignment. Meanwhile, they agree with the emails' text content is the best representative of the email's topic since the content-aware made a large proportion of contribution.

Sharaff et al. [36] agree that emails are the most popular and effective way to communicate over the internet. They further indicate that it is difficult for users to identify the related topics between the received email and relate newer incoming email. Hence, the authors consider email threading for email management and communication improvement, which provides a mechanism for users to sort emails within a certain time frame. For constructing an email thread that this work goes through a two times clustering method. In the first part, they apply Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) to form an email cluster. Afterward, calculate the similarity of people and subjects from the previous part's clustering result to generate the email thread.

McCallum et al. [33] propose the Author- Recipient-Topic (ART) model after reviewing the Latent Dirichlet Allocation (LDA) and Author-Topic (AT) model. The core concept is that each email's topic distribution will own a different topic distribution depending on the sender and recipient. In other words, there is a significant difference in the probability of topic distribution for different senders and recipients. They make

some improvements and modifications to the original LDA model. From the proposed ART model, the email is no longer only related to the observed word in an email but will consider the author and a set of recipients from an email to form a topic distribution. Moreover, the experiment is conducted on the public email dataset – Enron Email Dataset, and it demonstrates a better performance with lower perplexity value on previously unseen emails.

2.2 Literature reviews for enhancing LDA model

From the existed research described above, there are interesting trends in the study of both ham and spam email works that we make a short summarization below. Firstly, in the feature selection of emails, some studies investigate a specific feature in-depth, such as TFIDF. Simultaneously, some works consider a wide range of features and then select a more appropriate one for email analysis by the dynamic feature selection algorithm. Subsequently, the topic model techniques begin to gain attention in email analysis research, as well as some scholars focus on the processing of words in email's content, including WordNet dictionary check, Stemming algorithm, POS extractor, and Google's spell check API, besides, LDA model is one of the popular ones in the topic modeling techniques. Moreover, there are various extensions and applications based on the LDA model. Part of the works applies LDA in combination with other algorithms to achieve email clustering or email categorization. On the other hand, a portion of studies considers the other features of an email to estimate the specific probability of the topics distribution depending on certain constraints, such as sender and recipient set. Based on the LDA model's topics inference, we intend to combine the mechanism of attention allocation to each word while inferring the topic distribution of each email. There is a degree of similarity between this approach and the term weighting function. Hence, we will focus on reviewing the relevant studies that use term weighting to improve the performance of the LDA model below.

2.2.1 The extension of LDA on term weighting function

To our knowledge, discovering and collecting the features from email by the topic model has become more popular and common. Blei et al. [49] proposed the Latent Dirichlet Allocation (LDA), a topic model for automatically summarizing corpus. In

LDA, each document is represented as a mixture over an underlying set of topics, and each topic is selected from a probability distribution over the terms in the vocabulary. It assumes that the author has a bag full of topics and will randomly draw a topic from the bag when writing the document. Afterward, each topic has a bag full of words; the author will draw a word randomly from this bag to form a specific topic in the document. They selected each word from the word bag related to the topic after determining the topic. The way to represent each document is to consider the bag-of-words; however, a well-known problem with bag-of-words is that it does not acknowledge the order or location of the words in a document. While an email is not an ordinary document; instead, it has unique characteristics, involving multiple topics, lengthy replies, high variance [50]. Each email consists of multiple parts, a header line, a signature part, a quotation line, and actual text content. However, except for text content, most of them are typical noise [39].

Li et al. [51] argue that some of the inferred topics from the LDA model are hard to interpret, even unexplainable, because the words are given equal weight during the topic inference. They propose the Entropy Weighting (EW) scheme to measure the word co-occurrences by conditional entropy to address this issue. The primary concept is to prevent the high-frequency words from dominating the top topic word lists and giving less weight to those meaningless words, such as stop words. The experiment section applies the proposed scheme to the LDA and evaluates the topic quality, document clustering, and classification tasks based on real-world data. This study found that the more informative word, the more likely it to bring together a smaller set of words. The proposed EW scheme aims to identify the more informative words among the words with a similar frequency of occurrence for the LDA model's topic inference.

Pion-Tonachini et al. [52] propose the Crowd Labeling Latent Dirichlet Allocation (CL-LDA) for gathering crowd labels from the unlabeled datasets based on the LDA model. This study also recognized the necessity of giving higher weight value to certain words when making label (topic) inferences to achieve better performance. They apply the workers and weighting votes to equalize each work's influence, afterward, according to the change of weights and the prior distribution to estimate the instance-class probabilities for allowing to perform comparable to or better than other Crowd Label (CL) algorithms.

Wilson and Chew [53] suggest that the term weighting function is necessary when inferring the topic because the higher frequency words tend to be scattered across too many latent topics without explainable reason. They believe that the purpose of applying an appropriate term weighting function is to deal with those high-frequency words (which might be eliminated as stop words) in a more elegant way. According to the findings, they argue that removing stop words is not a necessary pre-processing step after considering the term-weighting function. However, the removal of stop words is usually unstated but widespread. To our knowledge, this approach is more realistic, while humans do not remove stop words when reading a document but instead assign a lower weight to these stop words.

2.2.2 The fundamental for building visual attention

As discussed above, most scholars have tried to apply the term weighting function to assign a lower weight to the high-frequency words as a kind of punishment. Thus, making the topic inference of the LDA model more meaningful. In our research, we highlight that the topic model has a weak point in processing email data. It ignores the original structure of an email, and the characteristic words will have different meanings according to their specific location. Therefore, we consider the visual attention allocation to assign a specific weight value according to its feature and location information.

Lovelace et al. [54] propose that we can regard the information from the text page as a two-dimensional encoded space and assign each piece of information like an address in the form of (x, y) coordinates. In other words, the 'WHERE' may serve to recall the 'WHAT'. More precisely, our purpose is to interpret and implement how to allocate attention to the words based on their location and features to infer hidden topics among these words between the various email structures.

Before applying visual attention, the first thing to do is confirm the definition and ambiguous terms. While we discuss attention that it is generally separated into two mechanisms Top-down and Bottom-up. The expectations and task requirements are the primary resources for deriving Top-down attention, and salient stimuli from the scene are the resource for deriving Bottom-up attention [55]. Besides, there are ambiguous concepts and meanings regarding the terms of attention and saliency.

Borji and Itti [56] conclude that the term "attention" is a general concept; it includes all the factors that affect the selective process of attention, whether Bottom-up or Top-down mechanisms. Furthermore, the Bottom-up computation applies the term "saliency" often, which considers the scene's part as a factor intuitively. The above mechanisms and definitions will consider understanding the intention of each calculation of probability in the formula as we calculate the attention probability on each word in the following section. Second, besides the definition, we review the previous works about how attention influences selection.

White et al. [38] describe that attention is a selective process; it will preferentially select information from specific locations in the scene since they are relatively important to the task. Based on previous research results, Rayner [57] shows that it needs to shift or allocate attention to select a single word on the page while reading. Furthermore, many studies also support that attention will allocate to those most informative, most surprising scenarios or locations that can return the maximum reward for the current task [58]–[60]. From the above review works, we can conclude that attention can effectively select the content that brings more information. However, as discussed in the first paragraph of this subsection, an email is not an ordinary document that includes multiple structures. Hence it is necessary to prove how people allocate or shift their attention to those structures and locations concurrently.

The previous research has likened the attention to a spotlight [61] or a zoom lens type [62]. The spotlight imitates the attention as a beam; it is allowed to move but process nothing outside of it. However, the zoom lens's way can define as a target stimulus that is more likely to be detected if near another stimulus [63]. Most of the discussions focus on whether attention can be split or not based on these two types. Castiello and Umiltà [64] suggest that humans can split the attention across locations that allocate it to across two to four locations in the visual field. Moreover, Cave et al. [65] propose a theoretical framework to explain human can adjust the attention in both ways which include multiple locations can be selected simultaneously and a single location only.

In our work, we propose the AttLDA model, which models email as a mixture of latent topics and considers each word's attention value as a weighting function. For implementing the AttLDA model that we summarize three critical factors from the above review works.

(1) Location: with building a two-dimensional space, each word will assign to a specific coordinate, and we can determine and estimate the attention allocation of shift based on it.

(2) Attention types: as we introduce two kinds of mechanisms: Top-down and Bottom-up, both mechanisms will consider in our experiment. Further, we agree with the idea that it describes attention as a zoom lens; another stimulus will be more easily detected while it is near the target stimulus.

(3) Split attention: an email is not like an ordinary document that includes multiple structures. Hence, no matter if we likened attention to a spotlight or zoom lens types, we ensure the attention can split across multiple locations simultaneously from the above review.

The point is, we consider the allocation of visual attention as a weight function in the process of topic inference on each word from an email to try to filter out those meaningless words. We will present the method in the following section.

CHAPTER 3 Methodology for the proposed AttLDA model

The Latent Dirichlet Allocation (LDA) is a well-known model applied to a wide variety of fields. Throughout this work, there are some terms mentioned frequently. Thus, before deep into our work, here we conclude the notation and terminology in the following to provide a clear intuition:

1. A *word* is the basic unit of a document from a vocabulary of size V (includes V distinct words). We represent words using unit-basis vectors that only a single component will equal to one and other components equal to zero. The vth word is represented by a V-vector w, $\mathrm{w} = [0 \cdots 010 \cdots 0]^T$ such that $w^v = 1$ and $w^u = 0$ for u ≠ v, where both u and v are auxiliary index over vocabulary and we also define the auxiliary index as *word_id*, which is considered as a unique identifier for each word.
2. A *document* is a sequence of N words denoted by $\mathbf{w} = \{w_1, w_2, \cdots, w_{N_d}\}$, where w_{N_d} is the nth word in the sequence of document d, and N_d indicates the number of words in document d, where $n \in \{1,2, \cdots, N_d\}$, and d $\in \{1,2, \cdots, D\}$. In our case, we recognize an email as a *document*. For estimating visual attention of word w_{N_d}, we define that within each email consists of K_d lines, and there are M_{dk} cumulative characters in kth line, where $k \in \{1,2, \cdots, K_d\}$. Based on that, the cumulative characters in k th line of document d are denoted by $\mathbf{m} = \{m_1, m_2, \cdots, m_{X_{dk}}\}$, where$m_{X_{dk}}$ indicates the cumulative amount of characters until the *x*th word of the *k*th line in document d and $x \in \{1,2, \cdots, X_{dk}\}$.
3. A *corpus* is a collection of *D* documents denoted by $\mathrm{Y} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_D\}$; in our case, the corpus is a collection of all the email documents.
4. The *visual attention* of each word is denoted by $\mathbf{h} = \{h_1, h_2, \cdots, h_{N_d}\}$, where h_n represents the attention value given by the *n*th word in document *d*; h_x represents the value of attention given by the *x*th word of *k*th line in document *d*. For example, if there are only two sentences in a document:

   “Check your inbox to complete the verification.”

   "And we will activate your account."

   In the first sentence, the word "your" $h_n = 2$, and $h_x = 2$. While in the second sentence, the word "your" $h_n = 12$, and $h_x = 5$.

5. A *topic* is a latent variable; during the sampling and inferring processes, each topic distribution of a document is according to the word probabilities, where $t \in \{1,2,\cdots,T\}$.

Table 3-1 the notation of this work.

Symbol	Meaning
v	Auxiliary index over vocabulary of the corpus *Y*.
V	Number of words in vocabulary of the corpus *Y*.
d	Auxiliary index over documents (emails).
D	Number of documents (emails) in the corpus *Y*.
t	Auxiliary index over topics.
T	Specified number of topics.
k	Auxiliary index over lines in a document.
K_d	Number of lines in document *d*.
M_{dk}	Value of cumulative characters of the *k*th line in document *d*.
n	Auxiliary index over the words in document *d*.
N_d	Number of words in document *d* (document length).
x	Auxiliary index over the words of *k*th line in document *d*.
X_{dk}	Number of words of *k*th line in document *d*.
w_n	It represents the *n*th word in a document.
w_{dn}	It represents the *n*th word in document *d*.
m_x	Value of cumulative amount of characters until the *x*th word in *k*th line.
α	A prior T-vector represented the topic distribution of documents.
β	A prior V-vector represented the word distribution of topics.
β_{z_n}	A prior V-vector represented the word distribution conditioned on the topic z_n.
θ_d	A topic proportions for document *d*, which subject to the $Dir(\alpha)$.
$Dir(\alpha)$	A T-dimensional Dirichlet distribution with a symmetric parameter α for providing the topic proportions.
$\phi_{z_{dn}}$	A topic-word proportions for topic z_{dn}, which subject to the $Dir(\beta)$.
$Dir(\beta)$	A V-dimensional Dirichlet distribution with a symmetric parameter β for providing the vocabulary proportions.

z_{dn}	Topic index over topics for *n*th word in document *d*; $z_{dn} = t$ means that the *n*th word in the *d*th document is assigned to topic *t*, which subject to $\text{Multinomial}(\theta_d)$.
h_{dn}	It represents an attention value of the *n*th word in document *d*.
f_{dn}	Value of visual feature of *n*th word *w* in document *d*.
f_v	Value of visual feature of vocabulary word *v*.
l_{dn}	Value of location of the *n*th word in document *d*.

3.1 Data collection of Enron Email Corpus

The Enron Email Corpus is alluring and of particular interest with much academic value for applying as the experimental dataset. It is a rich temporal record of internal communication within an organized, real-world corporation. This corpus contains a large amount of raw data on communication, knowledge sharing, relationships, perceptions, resources, and events in a company. In addition to the above considerations, the Enron Email Corpus has been widely used in academic works, including Klimt and Yang [66] exploring the distribution of emails and each user's time series from the Enron Email Corpus extract the characteristics to form an email discussion thread. Bekkerman et al. [67] consider that workers in an organization have to deal with many spam emails and large quantities of legitimate emails. Hence, he presents a classifier to automatically classify emails into pre-defined folders by applying two corpora, including Enron Email Corpus and SRI Corpora. Duan et al. [68] analyze and identify the importance of the Enron employees by extracting the Enron Email Corpus features, including the number of emails sent and received.

Before delving further into the proposed AttLDA model, we describe how we acquire the mail dataset and process it into a type that can be used for visual attention estimation on each word, as illustrated in Figure 3.1. Some of the agencies provide the Enron email data through a series of data pre-processing to facilitate the research work, such as Kaggle [69]. A large percentage of scholars download this dataset directly from those organizations or agencies. However, in our case, we need to preserve the original structure and format for estimating the visual attention on each word, according to the

word's location, feature, and inferred topics. Hence, we download the original resource in zip file format from https://www.cs.cmu.edu/./enron/enron_mail_20150507.tar.gz.

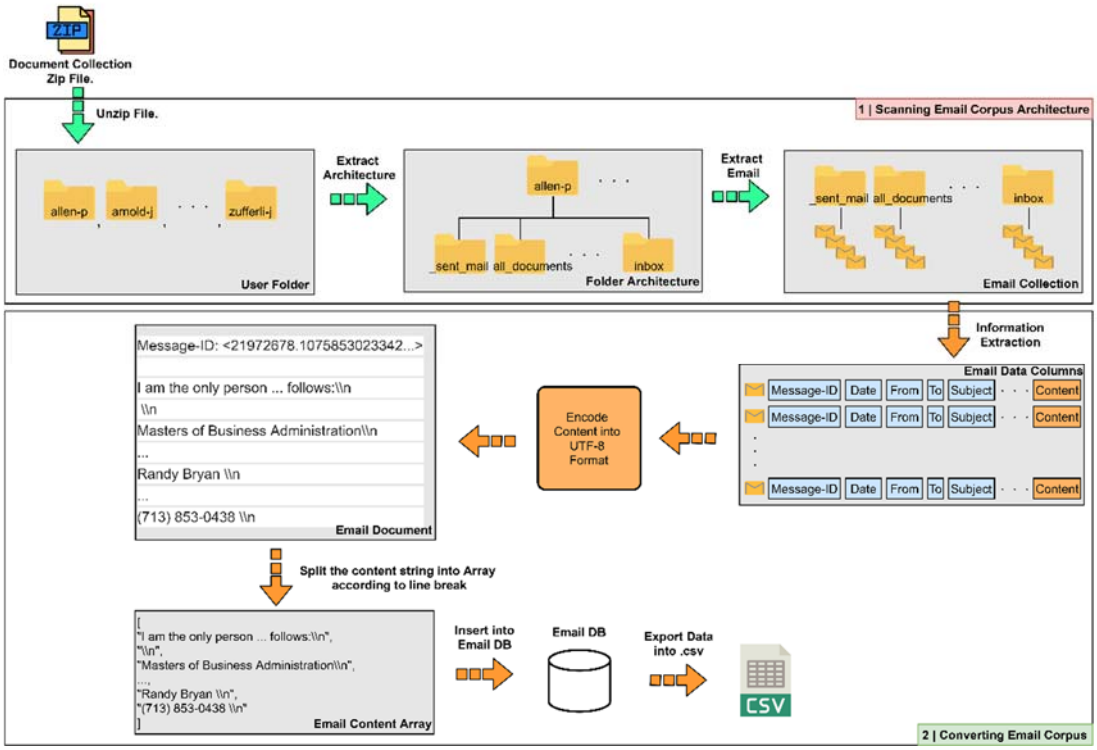


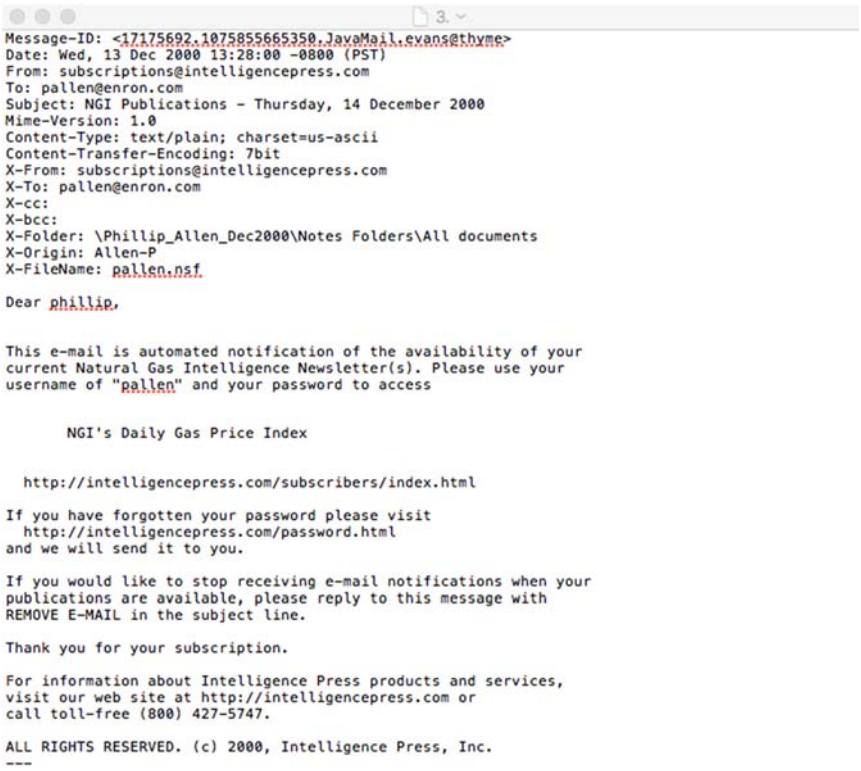
Figure 3.1 The collection process of Email data.

In Figure 3.1, there are two main aspects to this illustration, which we divided into the upper and bottom parts. In the upper part, the process focuses on scanning email corpus architecture, and we describe it from left to right. In either part, we will automate the work by building a python tool.

First and foremost, we decompress the downloaded file. After decompressing, several folders will be created in the pre-created folder. In our case, we store all the decompressed folders within the folder with the name "EnronEmail". Each decompressed folder is regarded as a personal mailbox of an Enron employee, and the folder name represents their user name. Once completed, we overview how many Enron employees' mailboxes were collected in this corpus.

Secondly, going through each Enron employee's folder, there are several customized folders beyond the inbox folder. Also, there is no fixed number of customized folders over each folder. Therefore, the python tool we built must support a dynamic approach to scanning the customized folders' structure.

Last but not least, each customized folder may hold several emails and might be an empty folder as well. The emails in the folders are named numerically. Thus, before accessing the content of an email, the only information we identified is that the email is stored in a specific employee's mailbox and to which defined folder it is classified.



```
Message-ID: <17175692.1075855665350.JavaMail.evans@thyme>
Date: Wed, 13 Dec 2000 13:28:00 -0800 (PST)
From: subscriptions@intelligencepress.com
To: pallen@enron.com
Subject: NGI Publications - Thursday, 14 December 2000
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: subscriptions@intelligencepress.com
X-To: pallen@enron.com
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\All documents
X-Origin: Allen-P
X-FileName: pallen.nsf

Dear phillip,


This e-mail is automated notification of the availability of your
current Natural Gas Intelligence Newsletter(s). Please use your
username of "pallen" and your password to access


       NGI's Daily Gas Price Index


  http://intelligencepress.com/subscribers/index.html

If you have forgotten your password please visit
  http://intelligencepress.com/password.html
and we will send it to you.

If you would like to stop receiving e-mail notifications when your
publications are available, please reply to this message with
REMOVE E-MAIL in the subject line.

Thank you for your subscription.

For information about Intelligence Press products and services,
visit our web site at http://intelligencepress.com or
call toll-free (800) 427-5747.

ALL RIGHTS RESERVED. (c) 2000, Intelligence Press, Inc.
---
```

Figure 3.2 The snapshot for an original Enron email (text file).

Next, proceeding to the bottom part, each email is an unstructured text file as well as the presented snapshot in Figure 3.2. The definition of unstructured text is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Hence, using the Regular Expression (RegEx), we identify each column of an email and extract the content, including Message-ID, Date, From, To, Subject, Mime-Version, Content-Type Content-Transfer-Encoding, X-From, X-To, X-cc, X-bcc, X-Folder, X-Origin, X-FileName, and Content.

Afterward, the email content is encoded in UTF-8 format. The purpose is to identify the line breaks' location, which is display as "\\n". According to that, we can perform the sentence segmentation in the subsequent process that we split the encoded content string into an array according to the line breaks.

Lastly, we insert the email data column by column into the MySQL database for the permanent application and export it as a CSV file as shown in Figure 3.3 to Figure 3.5, which will be used as input data for subsequent actions. To sum up, we have listed the actions we carried out below:

1. Decompress the Enron Email Corpus.
2. Identify and access the structure of Enron Email Corpus.
3. Identify and access the customized folders from each Enron's employee.
4. Identify the email text file over each customized folder and access its content by using the RegEx.
5. Identify the line breaks' location by encoding email content in UTF-8 format for sentence segmentation process.
6. Insert the email data into the MySQL database for the permanent application.
7. Export the data into CSV from the MySQL database for the subsequent analysis process.

MySQL connection is closed

	ID	MessageID	EmailDate	EmailFrom	EmailTo	Subject
0	835	<17449361.1075855672476.JavaMail.evans@thyme>	1979-12-31 16:00:00	phillip.allen@enron.com	['maryrichards7@hotmail.com']	Re:\n
1	1096	<5722531.1075855698441.JavaMail.evans@thyme>	1979-12-31 16:00:00	phillip.allen@enron.com	['c@enron.com']	\n
2	1920	<12860336.1075855675177.JavaMail.evans@thyme>	1979-12-31 16:00:00	phillip.allen@enron.com	['john.lavorato@enron.com', 'beth.perlman@enron.com', 'hunter.shively@enron.com', 'tscott.neal@enron.com', 'thomas.martin@enron.com', 'john.arnold@enron.com']	systems wish list\n
3	1930	<23724536.1075855675391.JavaMail.evans@thyme>	1979-12-31 16:00:00	phillip.allen@enron.com	['stephen.harrington@enron.com', 'mary@enron.com']	\n
4	662	<12929996.1075855668941.JavaMail.evans@thyme>	1979-12-31 16:00:00	phillip.allen@enron.com	['muller@thedoghousemail.com']	Re: (No Subject)\n

Figure 3.3 The snapshot for the Enron email inserted into database.

XFrom	XTo	Xcc	Xbcc	XFolder	XOrigin	XFileName
Phillip K Allen\n	['"mary richards" <maryrichards7@hotmail.com> @ ENRON\n']	[]	[]	\Phillip_Allen_Dec2000\Notes Folders\All documents\n	Allen-P\n	pallen.nsf\n
Phillip K Allen\n	['C\n']	[]	[]	\Phillip_Allen_June2001\Notes Folders\All documents\n	Allen-P\n	pallen.nsf\n
Phillip K Allen\n	['John J Lavorato', ' Beth Perlman', ' Hunter S Shively', ' Scott Neal', ' Thomas A Martin', ' John Arnold\n']	[]	[]	\Phillip_Allen_Dec2000\Notes Folders\Discussion threads\n	Allen-P\n	pallen.nsf\n
Phillip K Allen\n	['Stephen Harrington', ' Mary\n']	[]	[]	\Phillip_Allen_Dec2000\Notes Folders\Discussion threads\n	Allen-P\n	pallen.nsf\n
Phillip K Allen\n	['muller@thedoghousemail.com\n']	[]	[]	\Phillip_Allen_Dec2000\Notes Folders\All documents\n	Allen-P\n	pallen.nsf\n

Figure 3.4 The snapshot for the Enron email inserted into database.

EmailContent	EmailUser	MimeVerb	ContentType	ContentTransfer	folderPath	FullPath
['Mary,\\n', ' It is OK to buy a carpet shampooer.\\n', ' About the W-2\\'s, how would you '\\n']	allen-p	1.0\n	text/plain; charset=us-ascii\n	7bit\n	allen-p/all_documents	./EnronEmail/Test_Folders/allen-p/all_documents/319.
['George,\\n', 'In response to your ideas\\n', 'Time and cost\\n', '1. I realize that asking for a fixed price contract would result in the \\n', "builder using a higher estimate to cover uncertainty. That '\\n']	allen-p	1.0\n	text/plain; charset=us-ascii\n	7bit\n	allen-p/all_documents	./EnronEmail/Test_Folders/allen-p/all_documents/559.
['attached is the systems wish list for the gas basis and physical trading\\n', ""\\n']	allen-p	1.0\n	text/plain; charset=us-ascii\n	7bit\n	allen-p/discussion_threads	./EnronEmail/Test_Folders/allen-p/discussion_threads/87.
['EOL report for TV in conference on 33\\n', 'Cash\\n', ' \\n', '-Hehub\\n', '-Chicago\\n', '-PEPL\\n', '-Katy\\n', ' -Waha\\n', ' \\n', "Prompt Month Nymex'\\n']	allen-p	1.0\n	text/plain; charset=us-ascii\n	7bit\n	allen-p/discussion_threads	./EnronEmail/Test_Folders/allen-p/discussion_threads/97.
['How is your racing going? What category are you up to? \\n', "I'\\n']	allen-p	1.0\n	text/plain; charset=us-ascii\n	7bit\n	allen-p/all_documents	./EnronEmail/Test_Folders/allen-p/all_documents/157.

Figure 3.5 The snapshot for the Enron email inserted into database.

3.2 Latent Dirichlet Allocation (LDA)

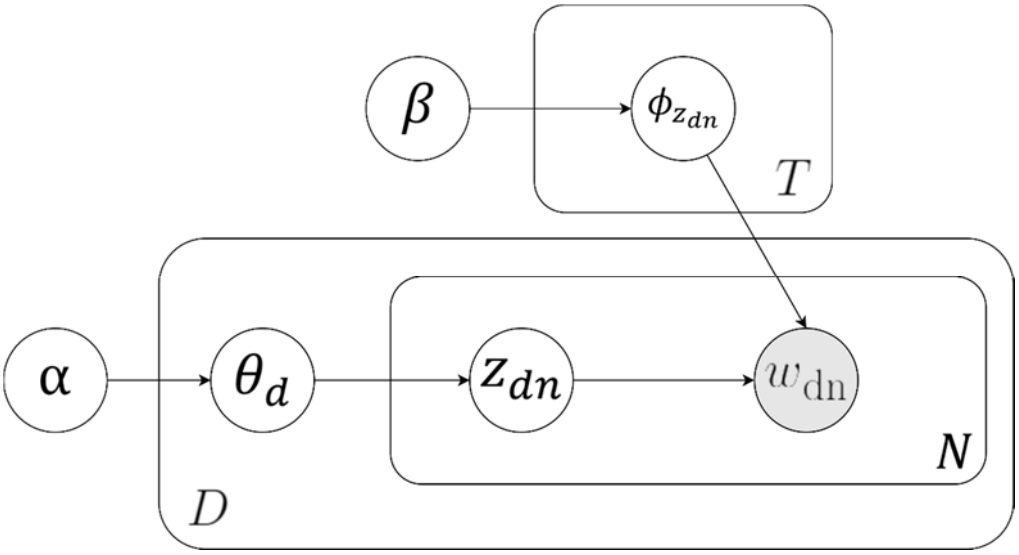


Figure 3.6 The graphical model of LDA.

LDA is a generative hierarchical Bayesian model that represents the documents as random mixtures over latent variables. Each document conducts as a random mixture of corpus-wide topics, and each word in a document is drawn from one of the specified topics. It assumes topics exist outside of the document collection, and each topic is distributed over fixed vocabulary. In Figure 3.6, we illustrate the graphical LDA model that the bottom-outer box represents documents in the corpus, and the inner one represents the selection of topics and words within a document. The upper one represents the selection of topics. One of the advantages of LDA is that it provides well-defined inference procedures for previously unseen documents.

As shown in Figure 3.6, a corpus Y consists of D documents, and each document d has N_d words. Except for words $\boldsymbol{w}_d$ in document d, other variables are not directly observable. Here we describe the different meanings of representatives over the

graphical model. The plates' boxes represent the replicated structure, nodes are random variables, edges represent possible dependence, and the gray shaded one is the observed variable. The documents are unsupervised data, and applying this model to explore the latent variables to identify a document's topics. LDA represents topic by word probabilities, and all the probability generation starts from Dirichlet prior. We conclude the steps for the generative process of LDA:

1. Choose $N \sim Poisson(\xi)$, where ξ is the average number of words in a document over the corpus. This step can be omitted while the number of words is a known parameter.
2. Choose $\theta_d \sim Dirichlet(\alpha)$, θ and α are T-dimension and α is a prior as regard as a pseudo counts of a topic in a document.
3. Afterward, for each word w_n, choose a topic $z_n \sim Multinomial(\theta_d)$ where z_n is a topic index over a T-dimension. A multinomial distribution with parameter θ_d over topics, drawn from a Dirichlet prior with hyper-parameter α.
4. Choose $\phi_{z_{dn}} \sim Dirichlet(\beta)$, ϕ and β are V-dimension and β is a prior as regard as a pseudo counts of word-topic pair. A multinomial
5. Choose a word $w_n \sim Multinomial(\phi_{z_{dn}})$, which is a multinomial probability conditioned on the topic z_n. A multinomial distribution with parameter ϕ over words, drawn from a Dirichlet prior with hyper-parameter β.

In short, given the prior parameters α and β, we choose the random variables θ and ϕ. For each word *w* in document *d*, we are sampling a topic *t* as a word-topic pair from a multinomial distribution with parameter θ_d, represented as z_{dn}, clearly indicating the word w_{dn} belongs to which topic. To generate the word w_{dn} in document *d*, we are sampling a topic *t* as z_{dn} from the multinomial distribution with parameter θ_d firstly. After that, according to the specific topic z_{dn} and the parameter $\phi_{z_{dn}}$, we select the corresponding multinomial distribution to sample the word w_{dn} into document *d*.

Lastly, we summarize the above description into the equation below:

$$\theta_d \sim \text{Dirichlet}(\alpha)\ \forall d \in D \quad (1)$$

$$\phi_t \sim \text{Dirichlet}(\beta)\ \forall t \in \{1,2,\cdots,T\} \quad (2)$$

$$z_{dn} \sim \text{Multinomial}(\theta_d)\ \forall d \in D, \forall n \in \{1,2,\cdots,N_d\} \quad (3)$$

$$w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})\ \forall d \in D, \forall n \in \{1,2,\cdots,N_d\} \quad (4)$$

3.3 Attention integration based Latent Dirichlet Allocation (AttLDA)

Describing the LDA model in linear-algebraic terms that the product of θ (the $D \times T$ column-stochastic topic-by-document matrix) and ϕ (the $T \times V$ column-stochastic topic-by-vocabulary matrix) is the $D \times V$ term-by-document matrix. The standard LDA model presented above assumes each word is equally important (equal weighting) in calculating the conditional probabilities. However, this is not the case from both an information-theoretic and linguistic point of view. For example, in English, the word "the" is meaningless compare with other low-frequency words. Besides, in the case of an email, contents located in a forward message are often considered less meaningful.

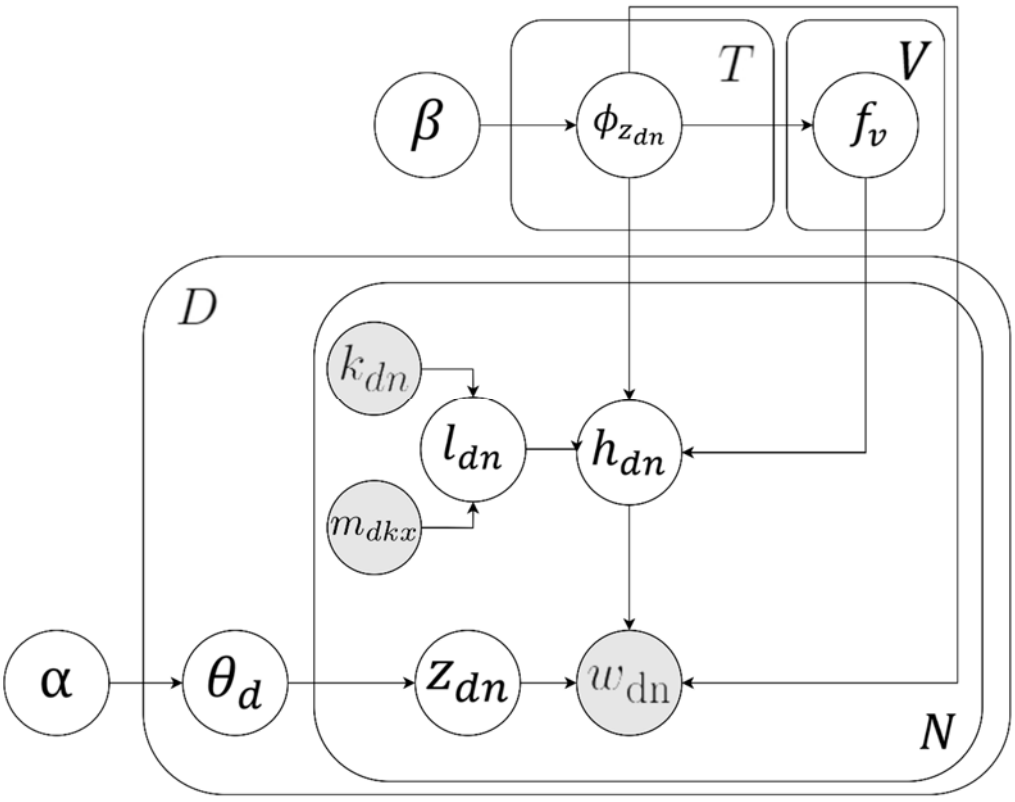


Figure 3.7 Graphical representation of proposed AttLDA model.

Our proposed model – AttLDA infers the email topic distribution by considering the visual attention, which is calculated by the variance among the topic distribution, word location, and visual feature per word. For each vocabulary v, we estimate the visual feature from the Term Frequency (TF) and Inverse Document Frequency (IDF) as a visual feature. Jones [70] proposed the IDF idea firstly in 1972 and conceived a statistical interpretation of term-specificity (IDF) which became a cornerstone of term weighting. Salton and Yu [71] present the TFIDF algorithm in their work by combing IDF's idea and demonstrating the effectiveness of this algorithm in information retrieval. We consider the TFIDF for estimating the visual feature, as in Eq. (5).

$$tfidf(v,d,D) = tf(v,d) \cdot idf(v,D) \tag{5}$$

$$tf(v,d) = \mathcal{F}_{v,d} \Big/ \sum_{v'} \mathcal{F}_{v',d} \tag{6}$$

$$idf(v,D) = \log\left(\frac{D}{|\{d \in D : v \in d\}|}\right) \tag{7}$$

In Eq. (6), a term frequency is to use the raw count of a term v in document d, e.g., the raw frequency of term v divided by all the occurring terms in the document. In Eq. (7), the inverse document frequency is a measure of how much information from a word, i.e., the information provided by a word, is rare or common across the corpus. The logarithmically scaled inverse fraction is obtained by dividing the total number of documents by the number of documents containing the term.

A location parameter l_{dn} is constructed by the parameter k_{dn} (value of line until nth word over lines in document d) and the parameter m_{dkx} (value of cumulative amount of characters until the xth word in kth line). In the part of word location estimation, we have consider to several literatures to determine other necessary parameters, which will be described with more details in a later section.

Given a Dirichlet parameters α and β and a sequence of N words in a document which denoted by **w**, the generative process works as follows:

1. The occurrence rate of each topic t in document d,

$$Draw\ \theta_d \sim Dirichlet(\alpha) \tag{8}$$

2. For each document d, number of times the nth word in document d occurred in each topic,

$$z_{dn} \sim Multinomial(1, \theta_d), z_{dn} \in \{1,2,\cdots,K\} \quad (9)$$

3. The occurrence rate of the words over the topics,

$$\phi_{z_{dn}} \sim Dirichlet(\beta) \quad (10)$$

4. The visual feature for each vocabulary word v,

$$f_v = VisualEstimation(tfidf(v,d,D)) \quad (11)$$

5. For each document d,

(a) The word location l_{dn} for each word w_{dn},

$$Draw\ l_{dn} = LocationEstimation(k_{dn}, m_{dkx}) \quad (12)$$

(b) The visual attention h_{dn} for each word w_{dn},

$$Draw\ h_{dn} = AttentionEstimation(l_{dn}, f_v, \phi_{z_{dn}}) \quad (13)$$

(c) For each word w_{dn},

$$Draw\ w_{dn} \sim Multinomial(\phi_{z_{dn}}) \cdot h_{dn}, w_{dn} \in \{1,2,\cdots,V\} \quad (14)$$

After defining the structure of the AttLDA model, we can derive the probability of observing a word w_{dn} in the Eq. (15) below:

$$p(w_{dn}|z_{dn}, \phi_t, h_{dn}) \quad (15)$$

Where w_{dn} is the observed word, z_{dn} is an index from 1 to T, ϕ_t are the topics, and h_{dn} is the visual attention.

Next, we can derive the mechanism of data generation, i.e., the joint probability distribution of all the hidden and observed variables according to the Eq. (16) below:

$$p(\theta, \phi, h, z, \boldsymbol{w}, f|\alpha, \beta, l) = \left(\prod_{t=1}^{T} p(\phi_t|\beta)\right) \cdot \left(\prod_{v=1}^{V} p(f_v|\phi_t)\right) \left(\prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{dn}|\theta_d)\, p(h_{dn}|l_{dn}, \phi_t, f_v) p(\boldsymbol{w}_{dn}|z_{dn}, \phi_t, h_{dn})\right)\right) \quad (16)$$

Where $p(\phi_t|\beta)$ is topics distribution over words which comes from Dirichlet distribution with parameter β. $p(\theta_d|\alpha)$ is the topic proportion over documents from Dirichlet distribution with parameter α. $p(f_v|\phi_t)$ is used to estimate the visual

features by TFIDF value of the words given ϕ_t. $p(z_{dn}|\theta_d)$ is used to indicate within each document, we have the words drawn from the topic assignment from θ_d. $p(h_{dn}|l_{dn}, \phi_t, f_v)$ is used to indicate that the visual attention of each word determined by location l_{dn}, topic-word proportion ϕ_t, and visual feature f_v. $p(w_{dn}|z_{dn}, \phi_t, h_{dn})$ is the probability of observing this word, conditioned on z_{dn}, ϕ_t and h_{dn}.

Subsequently, in our discussion of joint probability, we describe how we determined the visual attention, the word location, and the visual features of each word. Afterward, we describe how AttLDA model learning topics and attention via Collapsed Gibbs Sampling in the next section. Although LDA has been widely used in latent topics exploration, there is rare research to consider visual attention among words.

The review works of state-of-the-art visual attention modeling [56], introducing attention models including Cognitive model, Bayesian model, Decision Theoretic model, Information Theoretic model, Graphical model, Spectral Analysis model, Pattern Classification model, and other model. They further explain their mechanism to obtain saliency and attention. Also, all of the introduced models can implement in software. In this study, we select the Bayesian model, which combines the prior knowledge (e.g., scene content or gist) and sensory information (e.g., target features) to detect an object of interest probabilistically. According to that, we consider it as a weighting function into the LDA.

As mentioned above, LDA is a generative probabilistic model that identifies the documents as a mixture of latent topics and a distribution over the words determined by each topic. The basic idea is that if a word with higher probabilities to a topic, it gets a higher probability of indicating a specific topic than other words. Nevertheless, we assume people will allocate variance attention to a word according to the location, feature, and target of each word (topic) in our work. We depict the graphical model of AttLDA in Figure 3.7. Specifically, given a corpus *Y* consist of documents *D*, where document d is an email that composed of the $\boldsymbol{w}_d$ words, and there are five steps for explaining the parameters of our proposed AttLDA in detail:

1. The occurrence rate (pseudo count) of each topic t in document d is drawn from a Dirichlet distribution with parameter α. For each document d, randomly choose a topic z from the multinomial distribution $Multinomial(\theta_d)$. According to the suggestions from previous research [33], we define the $\alpha = 50/T$.

2. For each word in the vocabulary, we first determine the occurrence rate (pseudo count) of the words over the topics which draw from the Dirichlet distribution with parameter β. According to the previous research [33], we define the parameter β equal to 0.1. In the proposed AttLDA model, the proportions of the words generation not only consider the multinomial distribution $Multinomial(\phi_z)$, but also the visual attention $h_{z_{dn}}$.

For estimating the word location l, we consider two parameters to represent a document into two dimensional space. We apply k_{dn} to indicate the word in kth line, and m_{dkn} is the cumulative characters until nth word from kth line.

3. Regarding the word location determination, we first deliberate on determine the coordinate of a document for defining the word location later. Thus, we arranged the detail in the following and summarize the algorithm for assigning word location as described in Figure 3.8 and Figure 3.9.
   (1) As discussed in the chapter of literature reviews above, Castiello and Umiltà [64] suggest that we can split the attention across locations that include two to four locations. They expressed that people can put their attention to different locations at the same time while reading a document. Further, Dillon et al. [72] present that people will extract information from text using the optimal display size according to the target even though some differences might exist between individuals. For achieving the task and take into account the display size, there are two factors that we need to define: line length (the character per line) and window size (the number of lines or window height). Previous research has several suggestions about the line length, including 132 characters, between 25 characters to 100 characters, less than 70 characters, between 40 characters to 60 characters, 55 characters, and 100 characters [73]–[78]. Some of the views consider reading rate or ease of reading purposes to present the suggestions on line length; however, the line length with 55 characters is suggested by Chaparro et al. [77] which is based on reading comprehension. Therefore, we used 55 characters as the value for the line length because we believe that the purpose of reading comprehension is more closely related to our research objectives. We define the parameter *CharsLimit* in Algorithm1, as shown in Figure 3.8.

(2) The next factor that needs to take into account is the window size. From the previous research, Richardson et al. [79] show that people prefer the larger window size for reading or writing on a text page. Dyson and Kipping [74] present a study of the minimum window size, and the number of lines that are easiest to read must be at least 15 lines. However, it only provides the lower bound for determining the window size. Hence, we also refer to the work proposed by Duchnicky and Kilers [80] that they apply different variables for the evaluation, including five different values (1 line, 2 lines, 3 lines, 4 lines, and 20 lines). We select 20 lines for the window size from the above five values and considering the recommendations of a relevant study on the lower bound value that the minimum window size must be at least 15 lines. Therefore, we define the parameter *WindowLimit* in Algorithm 1 with the value 20.

(3) Some extra factors might also affect location segmentation, such as margin size, interlinear spacing and font size, Etc. However, the result proposed by Duchnicky and Kilers [80] states that there is no difference in incomprehension with the various sets of font sizes. Gould and Grischkowsky [81] found the interlinear spacing and visual angles that have no significant impact on reading accuracy. After considering the recommendation and suggestions from previous research, we apply two parameters (windows size and line length) to create a visual space representing the best display size for people to extract text information from a document (an email). Subsequently, the sentence segmentation will automatically process according to the rules below to form a visual space. If the number of characters in a sentence exceeds the threshold of line length and there is no line break symbol, we will create a new line as a new sentence in the current document (email) automatically; and simultaneously update the *WindowLimit* in Algorithm 1. The intention of doing so is because we only consider two factors to form our visual space of an email, including line length and window size; other factors will not include in this work for any further consideration.

Algorithm 1 Assign target c_{dn} and location l_{dn} to each word w_{dn}

```
procedure PREPARE(email_raw_dataset)        ▷ From generated .csv file.
    CharsLimit = 55, WindowLimit = 20
    DistinctWordsArray, WordWithourStopWordsArray = new Array
    Adj_row_counts, Adj_convert_times = 0
    email_dataset ← removePunctutation(email_raw_dataset)
    email_dataset ← removeHTMLTag(email_dataset)
    for All emails from email_dataset do
        email_sentences ← contentArraySplit(emails)
        email_sentences ← removeNumber(email_sentences)
        email_sentences ← removeLineBeak(email_sentences)
        Array extra_sentences = new Array
        for All sentences from email_sentences do
            Array email_words ← Tokenize(sentences)
            CharsCount ← CalculateTotalChars(email_words)
            CharsCountPopLast ← CalculateTotalChars(email_words.pop()
            Array extra_breakword = new Array
            for word from email_words do
                if CharsCount > CharsLimit and CharsCountPopLast >
  CharsLimit then
                    Array extra_breakword ← PushBreakwords(word)
                    Array extra_sentences ← PushSentences(breakword)
                end if
            end for
            if Len(extra_sentences) > 0 then
                call procedure allocate(extra_sentences)
            else
                call procedure allocate(email_words)
            end if
        end for
    end for
end procedure
```

Figure 3.8 Algorithm 1: Preprocessing for assigning location l_{dn} to each word w_{dn}.

(4) By determining the line length and window size, we define the visual space as above mentioned. Castiello and Umiltà [64] indicate that people could split attention across two to four positions. Depending on that, we divide the words into five locations; and we further define the additional fifth location to indicate those words do not display within the pre-defined visual space where the line length of the word is larger than 55 characters, and window size is larger than 20 lines. Besides, we also consider whether the location of a word is under some prefixes (e.g., "forward" or "original message" or

"forwarded message," Etc.), when such a scenario exists, we will identify the following content (words) as located in the additional fifth location.

Algorithm 2 Assign target and location to each word w_{dn}

```
procedure ALLOCATE(sentences, horizontal_index, sentence_index, adj_index, adj_time)
    for sentence from sentences do
        for word from sentence do
            if word not in DistinctWordsArray then
                DistinctWordsArray ← PushBreakwords(word)
            end if
            if word not in StopWords then
                WordWithourStopWordsArray ← PushBreakwords(word)
            end if
            if word in ForwardMsg then c_dn ← 0
            else
                c_dn ← 0
            end if
            horizontal_value ← CharsCountEachSentence//(CharsLimit/2
            sentence_estimate ← sentence_index + adj_index − adj_time
            if sentence_estimate > 20 then l_dn = 4
            else
                if horizontal_value == 0 then
                    if sentence_estimate <= 10 then l_dn = 0
                    else l_dn = 2
                    end if
                else
                    if sentence_estimate <= 10 then l_dn = 1
                    else l_dn = 3
                    end if
                end if
            end if
            Adj_row_counts ← CalculateAdjustRowCount()
        end for
        Adj_convert_times ← CalculateAdjustTimes()
    end for
end procedure
```

Figure 3.9 Algorithm 2: Assigning target and location to each word w_{dn}.

4. In the third step, we completed the definition of the word's location and explained that literature reviews which we refer to achieve our purpose. Next, after we reviewed 65 state-of-the-art attention models from the previous work [56], and we choose the Bayesian model for estimating the visual attention by Eq. (17) below. The reason we select the Bayesian model is because this model considers overall

factors, including the Bottom-up salient, prior knowledge, and Top-down knowledge (location prior). As we discussed the types of attention in the literature reviews chapter that the Bayesian model consider both types (attention and saliency), which is the primary reasons we select it as our model basis. Although most studies apply this model for image-related research, we allocate the location to words in the document through the previous steps, which allowed us to integrate this model with the LDA model into our research. We make an explanation below.

$$\begin{aligned}
&\log s_z = -\log P(F = f_z) + \log P(F = f_z|C = 1) + \log P(C = 1|L = l_z) \\
&\log s_{w_{dn}} = -\log P\left(F = f_{w_{dn}}\right) + \log P(F = f_z|C = 1) + \log P\left(C = 1|L = l_{w_{dn}}\right) \\
&h_n = h_z = exp(\log s_z)
\end{aligned} \tag{17}$$

(1) First, let us briefly review the original equation and definition of the Bayesian model as shown in Eq. (17). They define s_z as the overall saliency of a pixel point z in an image, where F is the visual features of a point, C is used to determine whether a point belongs to a predefined class or not, and L is the location of a point (pixel coordinates), where f_z is the visual features on point z and l_z is the location on point z.

(2) Next, we define the notations and the parameters for applying the Bayesian model into our proposed AttLDA model. We define point z as a word w_n in an email in our experiment since it is the basic unit in a document similar to a pixel point of an image. Therefore, we further explain the $VisualEstimation$ method here as we describe in the above Eq. (11). About the F of the $-\log P(F = f_z)$, we calculate the TFIDF value first by applying $tfidf(w_{dn}, d, Y)$ and define ten groups of TFIDF values by the percentile function. Each word w_{dn} will estimate the nearest index from the ten groups as their feature $f_{w_{dn}}$, and we calculate the probability of each features based on it later. The target class C of the $\log P(F = f_z|C = 1)$ is defined as a binary value like Boolean logic in the original formulation to indicate whether find the target or not. In our case, our target is discovering the hidden topic behind a word; therefore, we define the target class C as a topic.

(3) Lastly, about the $l_{w_{dn}}$ in $\log P(C = 1|L = l_{w_{dn}})$, we employ the K_d lines as Y-axis and m_{dkn} character as X-axis to form a coordinate to measure the $l_{w_{dn}}$ of each word as the location in a document.

In Eq. (17), the first term on the right side is the self-information (bottom-up saliency), and it depends on estimate the probability of each index of a word w_{dn} from the whole index by discovering the nearest index of TFIDF value groups; the word w_{dn} is same as the point z. The self-information will take a higher value for a rare TFIDF index of a word, that is, the more familiar TFIDF index, the less interest (attention). The second term is the log-likelihood, which considers the TFIDF index of a word consistent with the prior knowledge of the target (topic) C. The purpose here is to figure out if the feature (TFIDF index) of a word is rarely co-occurred with related prior knowledge of the target C; in that case, it will get a smaller value. Finally, as we previously presented, we assign a particular location to a word, and the third term of the equation is the top-down knowledge of the location at a word point z. We use it to calculate the probability that a word co-occurred with a particular C in the case of the given location l_z.

The attention value of each word is primarily based on Eq. (17), in which word point, saliency s_z, feature f_z, target C, and location l_z are concluding as follows:

$$\begin{cases} s_z\left(z \in 1,2,\cdots,w_{N_d}\right) \\ f_{w_{dn}} = \{f_1, f_2, \cdots, f_V\} \\ C = \{t_1, t_2, \cdots, T\} \\ l_{w_{dn}} = \begin{cases} 0, & 55 > m_{dkx} > 27.5 \; and \; k_{dn} \le 10 \\ 1, & m_{dkx} < 27.5 \; and \; k_{dn} \le 10 \\ 2, & m_{dkx} < 27.5 \; and \; 20 \ge k_{dn} > 10 \\ 3, & 55 \ge m_{dkx} \ge 27.5 \; and \; 20 \ge \; k_{dn} > 10 \\ 4, & otherwise \; or \; under \; predixes \; content \end{cases} \end{cases} \tag{18}$$

5. We then talk about how to integrate with the LDA to form our AttLDA. For each document d, only both of a word w and the location l can observe in the generative process. For each document d, form a location l_{dn} by considering k_{dn} and m_{dkx}, and a topic z of document d is chosen from the multinomial distribution with

parameter θ_d. To generate a word w_{dn} in document d, it then selects the corresponding distribution $\emptyset_{z_{dn}}$ and the value of attention h_{dn} depending on the topic z_{dn}, where the h_{dn} is a term weighting schema to represent how much attention on a word. In the AttLDA model, given the Dirichlet prior α and β, the location $l_{w_{dn}}$, the topic distribution $\text{Multinomial}(\theta_d)$, the mixture distribution $\text{Multinomial}(\phi_t)$ for the topic z, a visual attention h, a set of topic T, and a set of words in the corpus Y to learn the topic distribution of a document d. Although the AttLDA model with several latent variables that need to estimate by posterior distribution, in practice, once the topic z assignment and visual attention allocation of each word are derived, the distribution θ and ϕ can be estimated accordingly. As shown in Eq. (19), it is a further illustration based on Eq. (17).

$$p(\theta_d, \phi_{1:T}, h_d, z_d, \boldsymbol{w}_d, f_v | \alpha, \beta, l) = \prod_{n=1}^{N_d} \left\{ \begin{array}{c} p(\phi_{1:T}|\beta) \cdot p(\theta_d|\alpha) \cdot \\ \left(\prod_{v=1}^{V} p\left(w_{dn}|\phi_{z_{dn}}, h_{z_{dn}}\right) \cdot p\left(h_{z_{dn}}|l, \phi_{1:T}, f_v\right) \right) \cdot \\ \left(\prod_{v=1}^{V} p(z_{dn}|\theta_d) \right) \cdot \left(\prod_{v=1}^{V} p(f_v|\phi_{1:T}) \right) \end{array} \right\} \quad (19)$$

However, the above joint probability distribution is not our ultimate goal, our objective is to learn from a large number of emails in an email corpus Y to obtain the distribution of topics of each email, the distribution of words of each topic and the visual attention allocation of each word. In other words, try to get the posterior distribution of θ, ϕ, z, and h.

$$p(\theta, \phi, z, h | \alpha, \beta, l, Y, f) = \frac{p(\theta, \phi, z, h, Y, f | \alpha, \beta, l)}{p(Y, f | \alpha, \beta, l)} \quad (20)$$

The subsequent step is called inference in Bayesian statistics. However, the challenge in the inference step is in writing an email $w_n \in D$, while both the topic z_n and the associated words proportion ϕ_{z_n} of the topic z_n are predefined. We are not aware of the generation mechanism of an email without given z_n and ϕ_{z_n}. Therefore, some works apply the common and well-known Bayesian estimation method –

Variation Inference. Our research employs a more efficient method, Collapsed Gibbs Sampling for the inference process, and describes the process in the section below.

3.4 Inference by Collapsed Gibbs Sampling on AttLDA

Before presenting how we apply the Collapsed Gibbs Sampling in our works, we also review another well-known estimation algorithm - Expectation Maximization algorithm. Blei et al. [49] present the variational Expectation-Maximization algorithm by executing the E-step and M-step iteratively. This algorithm considers the E-step to estimate each document's topic distribution by current model parameters in the training stage and applying the M-step to update the model parameters. However, as Xiao and Stibor [82] indicate, that Expectation-Maximization algorithm approach is prone to local optima due to the wiggly likelihood function, and Minka and Lafferty [83] further argue that variational Expectation-Maximization algorithm can lead to inaccurate inferences and biased learning. Subsequently, Griffiths and Steyvers [84] proposed the Collapsed Gibbs Sampling, a Markov-chain Monte Carlo method and as present by Xiao and Stibor [82] that the Collapsed Gibbs Sampling is an efficient and straight-forward approach to rapidly converges in a known ground-truth. Besides, it has been widely used in many LDA research works. The description above is the primary reason why we select Collapsed Gibbs Sampling in this work.

For obtaining a practical result from the proposed model, we use collapsed Gibbs sampling [84] to perform the inference process. After defining the AttLDA model's architecture, we also obtained a data generation mechanism, that is, the joint probability distribution between parameters and email data as shown in Eq. (19). As a consequence of the joint probability distribution, our ultimate target is to learn how to identify the topic distribution of words in an email by considering attention's influence. Gibbs sampling through the Markov Chain Monte Carlo process, repeat sampling to approximate the unknown joint probability distribution. Meanwhile, Gibbs sampling is a Markov chain with stable distribution; strictly speaking, while simulated times enough, the sample generation’s distribution will be equal to the unknown joint probability distribution. Before implementing the inference, we also refer to the previous studies [51]–[53], which apply term weighting schemes into LDA to filter out meaningless words. They consider multiple schemes such as Log weighting to achieve

the purpose; Log weighting is used to concern a word frequently appears in a document or not. In the AttLDA model, we used attention to emphasize a word. While a word gets a higher value of attention, it will provide the rich word for the topic inference and give a clear explanation. Hence, we assume it will more effectively capture the topic from a document.

Accounting for attention and weighing the probability of a word to simulate the influence of attention when selecting a word for a particular topic, the probabilities in the AttLDA are calculated based on Eq. (20). In practice, we only need to estimate the visual attention of each word and the topic distribution of each word in an email, due to that, the distribution of θ and ϕ can be estimated accordingly. Hence, when we conduct Gibbs Sampling and only sample for z without considering other latent variables, we call Collapsed Gibbs Sampling.

$$\begin{aligned} & p(z_d|\alpha,\beta,l,Y,f,h) \\ &= \frac{p(z_d,z_{-d},Y|\alpha,\beta,l,Y,f,h)}{p(z_{-d},Y|\alpha,\beta,l,Y,f,h)} \\ &\propto p(z_d,z_{-d},Y|\alpha,\beta,l,Y,f,h) = p(z,Y|\alpha,\beta,l,Y,f,h) \end{aligned} \tag{21}$$

As shown in Eq. (21), we begin with the known distribution and given the topic distribution from the inferred emails z_{-d} to approach the topic distribution of the email d. In Eq. (22) below presents the probability that topic z of nth word in email d is assigned to topic t by applying the Collapsed Gibbs Sampling derived from the joint probability distribution.

$$\begin{aligned} & p(z_{dn}=t|\alpha,\beta,l,Y,z_{-dn},h_{-dn}) \\ & \propto \frac{h_{dn}N_{t,-n}^{(v)}+\beta}{\sum_{v'=1}^{V}h^{(v')}N_{t,-n}^{(v')}+\beta}\cdot\frac{N_{d,-n}^{(t)}+\alpha_k}{\sum_{t'=1}^{T}N_{d,-n}^{(t')}+\alpha_{k'}} \end{aligned} \tag{22}$$

Here, $N_{d,-n}^{(t)}$ indicates the summation of word count are assigned to another topic t expect topic z_{dn}, and z_{dn} is the topic that the model tries to sample in the current document d Also, $N_{t,-n}^{(v)}$ denotes the summation of word count except for the word itself. The below Eq. (23) and Eq. (24) show the derivation of document and topic and document pair distribution θ and the topic word pair distribution ϕ from z.

$$\phi_{t,v}=\frac{h_v N_t^{(v)}+\beta}{\sum_{v'=1}^{V}h^{(v')}N_t^{(v')}+\beta} \tag{23}$$

$$\theta_{d,t} = \frac{N_d^{(t)} + \alpha_t}{\sum_{t'=1}^{T} N_d^{(v')} + \alpha_{t'}} \tag{24}$$

To review all the processes we perform, we use the flow chart to describe, as we illustrated in Figure 3.10. We divide the current workflow into five stages:

(1) Loading and initializing email data;

(2) Removing noise data and assigning word positions;

(3) Estimating the value of attention for each word;

(4) For training and testing process, split the data set into training and test data, and define the number of topics based on topic consistency metric;

(5) Infer the topic of each word and used perplexity to evaluate the performance of the proposed AttLDA.

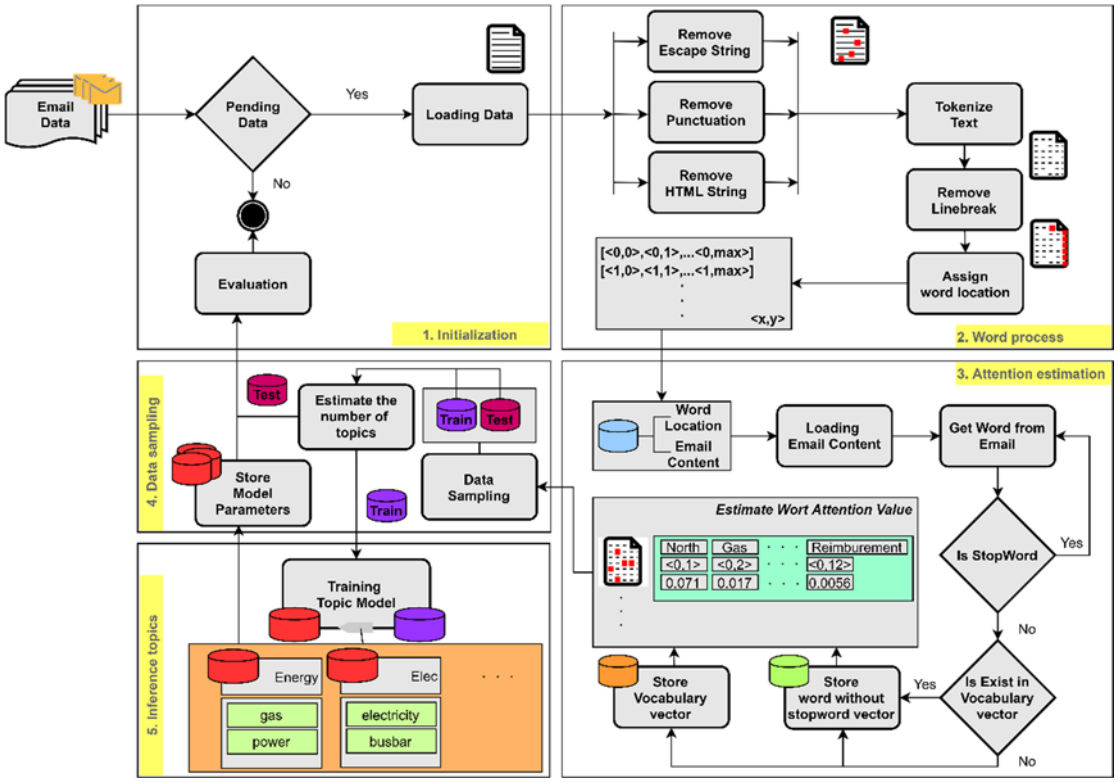


Figure 3.10 The workflow of AttLDA model.

As described in Figure 3.10, the mission in the first stage is loading the Email dataset. About the Email dataset, there are two critical things. We consider the Enron Email Dataset to our experiment to make our result and performance more reliable because it is a public Email dataset. The other one is that for calculating each word's

location, we need to keep all the format of an email; therefore, we download the original ZIP files, UTF-8 encoded process, segmentation, insert into a database, and export to CSV format. After that, the first stage will be loading the CSV format and deliver the dataset to the second stage-Word process.

In the second stage, we remove the noise data but not including the Stopwords. After we complete the Tokenize process, remove the line break from each document encoded in UTF-8 format. We keep these line breaks until the second stage because we can keep the word location even through the Tokenize process and then deliver the Tokenize word-location pairs to the next stage.

Before deep into the third stage, we first download two packages from Natural Language Toolkit (NLTK), including the words package and the stopwords package. We employ the words package to verify that the words in an email are verified and without the mis-spelling situation. On the other hand, we apply the stopwords package to verify a word is a meaningless word or not, such as to, would, with, by, and may, Etc. In the third stage, we divide the pair data delivered from the previous stage and extract each word from the Email content data to determine whether it exists in the Stopwords list or not. If so, the process will ignore and loading the next word. If a word does not belong to the Stopwords, we will insert it into the Word count vector for the model evaluation.

Moreover, if a word is never seen before, it will insert into the Vocabulary vector. By downloading the NLTK words package, besides the Stopwords list, we also form a legal word list to verify the words without misspelling issues; if a word does not exist in the legal word list, we will treat it as a stop word. After we run a loop of all the words, calculate each word's attention value based on its location, features, and the inferred topics.

In the fourth stage, in addition to dividing the data set into training and test data, it is worth mentioning that we use the Coherence model here to determine the description of the number of topics for the AttLDA model. Namely, we select the number of topics with the highest Coherence Score from the Coherence model results. Finally, training and testing our topic model, and according to the result, define each topic's name and show the evaluation to demonstrate the performance of AttLDA.

CHAPTER 4 Experimental results of AttLDA

In this section, we present the experiments to demonstrate the performance of the proposed AttLDA model. First, we discuss how to carry out the data splitting, i.e., we apportion the collected data into the training set and test set. Next, we describe how we implement the proposed AttLDA model, including the programming language and experiment environment. Third, we consider the Coherence model to determine the number of topics T. Fourth, we present how we evaluate the performance of the proposed AttLDA model. Lastly, based on the results of the proposed topic model, we conduct a further discussion.

4.1 Experimental Data set

As described in the previous chapter, we consider the Enron Email Corpus (https://www.cs.cmu.edu/./enron/enron_mail_20150507.tar.gz) to accomplish our experiments. This data set is consists of 490,745 emails from 150 users. After download it, the raw data set goes through the process shown in Figure 3.1 from the database to generate the CSV file for the subsequent experiments. We choose this data set for the experiment is because it is the most used data set by researchers in the email classification domain. For demonstrating the AttLDA model's performance, we divide the dataset into two parts, that is, randomly divided the data set into a training dataset with 90% and a test dataset with 10%. Besides, in the LDA experiment, the number of topics should be defined first. Hence, we determine the number of topics based on the Coherence model results and give a simple topic name to identify it, e.g., Topic 1, Topic 2 and Topic T.

4.2 Experimental setup

We build and deploy Docker containers to the server with python files and Enron email corpus for running the experiments. Our experiments were performed to evaluate the performance on a VPC cluster server [85], each node is equipped with an Intel Xeon $\mathrm{E5}-2680\mathrm{v2@2.80GHz}\times 20$, 64 GB of RAM, running Red Hat Enterprise Linux 6.4. The proposed AttLDA model was implemented with Python.

The definition of the number of topics is what research generally disapproves of; that is, it is vital to define the number of topics in an objective way. Therefore, before

beginning to implement the experiments and the evaluation, we apply the Coherence model [86] to determine this parameter-the number of topics among all the corpus. Coherence is used to measure a set of statements or facts that could able to support each other. They will consider being coherence; thus, a coherence fact set can regard as an interpreted context that covers all or most of the facts. In other words, the words under the topic are consistent in interpretation. We apply Java-based Mallet to create the Coherence model. According to the coherence measurement to determine the number of topics, we describe results generated from the Coherence model in Figure 4.1 to Figure 4.6. As we can see, while we set the number of topics as 20, the coherence score is the highest. Based on that, we employ 21 as the number of topics. The extra one topic is a pre-define topic for the stop words since we did not remove stop words from a corpus for calculating attention value; hence all the Stopwords will conclude into this topic in our experiment. Estimating the number of topics is the main task we perform in the fourth stage of the workflow, as described above in Figure 3.6.

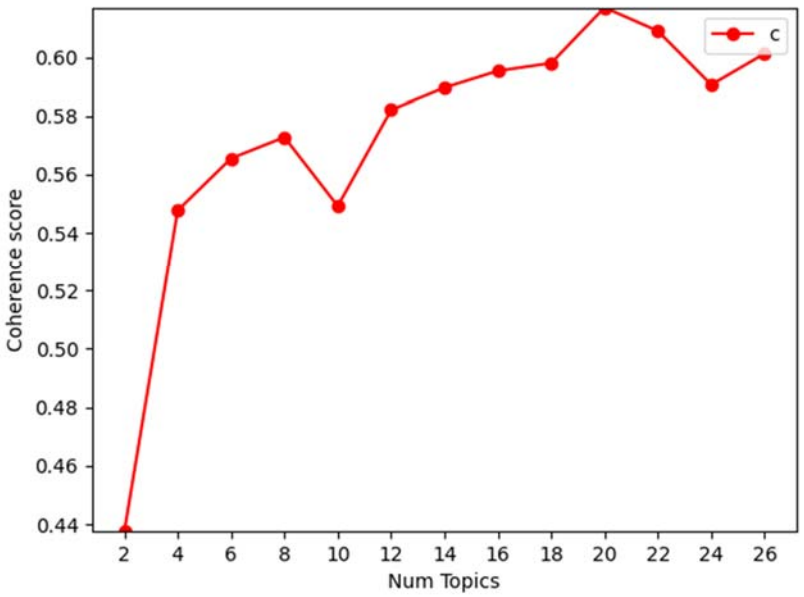


Figure 4.1 The result of the coherence score with the number of topics from 2 to 26.

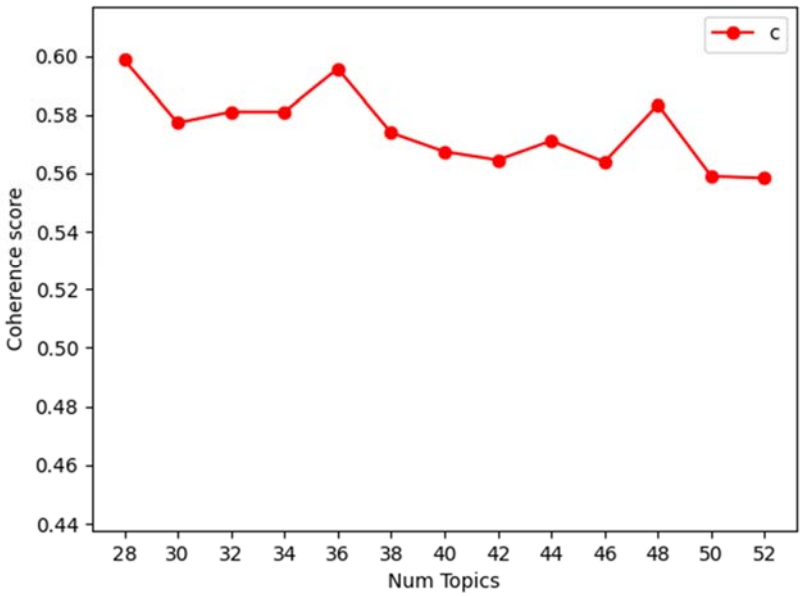


Figure 4.2 The result of coherence score with the number of topics from 28 to 52.

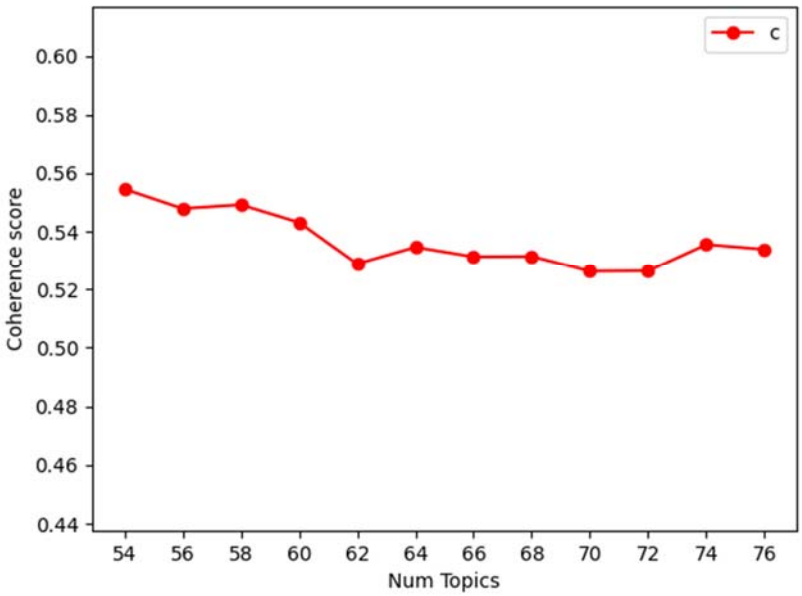


Figure 4.3 The result of coherence score with the number of topics from 54 to 76.

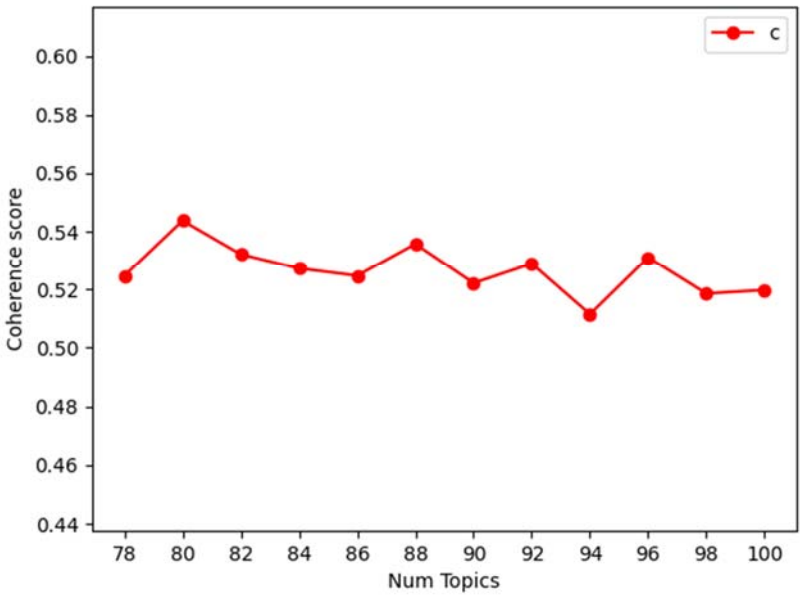


Figure 4.4 The result of coherence score with the number of topics from 78 to 100.

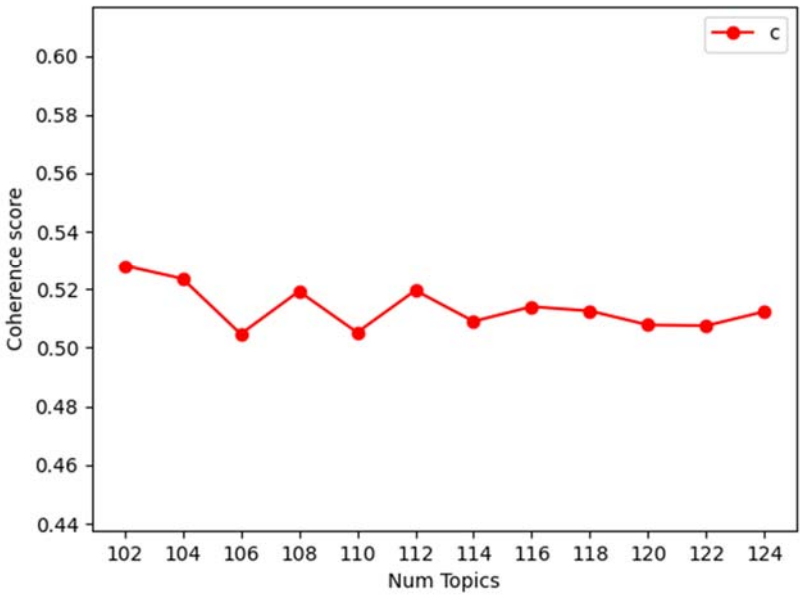


Figure 4.5 The result of coherence score with the number of topics from 102 to 124.

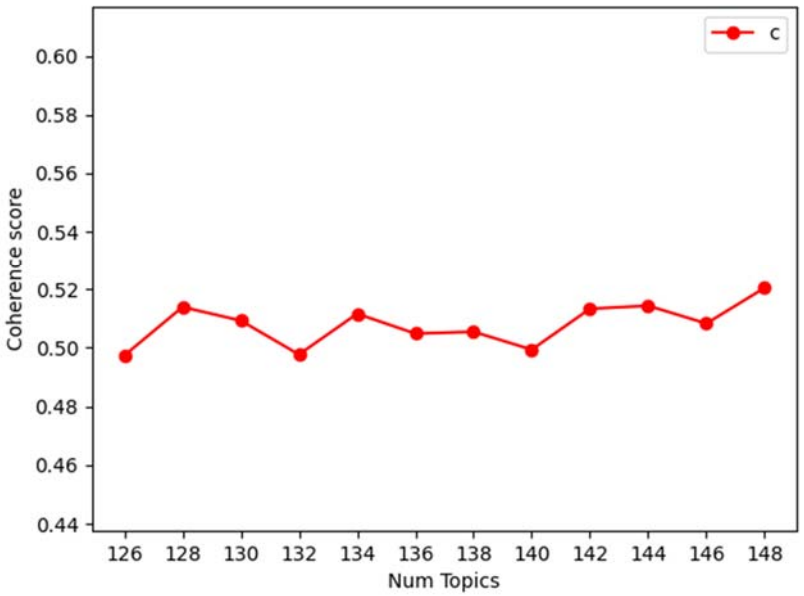


Figure 4.6 The result of coherence score with the number of topics from 126 to 148.

4.3 Evaluation metrics and comparison

Most natural language studies evaluate their proposed model by perplexity to demonstrate the performance. The most common way to evaluate a probabilistic model is to measure a held-out test dataset's log-likelihood. In our experiment, the test dataset is a collection of unseen emails. The perplexity is used to measure a model's uncertainty, based on that to evaluate the goodness of model fitting. If the value of perplexity is lower, in our case, it indicates the proposed model can able to make a correct prediction of a topic for the words effectively, and those words have never seen before. In the following perplexity formulation as Eq. (25), it is necessary to take into account the location of each word. The definition of the location of each word is described as the above Eq. (18), the location $l_{w_{dn}}$ of word w_{dn} is calculated based on the *n*th word of the *k*th line k_{dn} and the cumulative characters m_{dkx} until the *x*th word w_{dx} of *k*th line appears in document *d*.

In the case of training stage that AttLDA model generates a new document $\mathbf{w}$ by firstly drawing a document-specific topic distribution. Next, drawing a topic assignment for each word. Lastly, observed the words as describe in below Eq. (25).

$$\begin{aligned} &\theta \sim Dirichlet(\alpha) \\ &z \sim p(z|\theta) \\ &\mathbf{w} \sim p(\mathbf{w}|z, \phi, h) \end{aligned} \tag{25}$$

However, a test dataset is a collection of unseen emails, the parameter θ is not taken into account as it represents the topic distribution for the document of the training dataset. Therefore, we will ignore the θ to estimate the log-likelihood as below Eq. (26).

$$\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{w}|\phi, \alpha, h) = \sum_{d} \log p(\boldsymbol{w}_d|\phi, \alpha, h) \tag{26}$$

We first evaluate the log-likelihood of a set of unseen emails given the topics ϕ and the hyper parameter α for topic distribution. The likelihood of unseen emails can be used to compare models; higher likelihood implying a better model. The measurement of perplexity of held-out emails $\boldsymbol{w}_d$ defined as below Eq. (27).

$$perplexity(test\ set\ \boldsymbol{w}) = exp\left(-\frac{\mathcal{L}(\boldsymbol{w})}{count\ of\ tokens}\right) \tag{27}$$

Which is a decreasing function of the log-likelihood $\mathcal{L}(\boldsymbol{w})$ of the unseen emails $\boldsymbol{w}$; the lower the perplexity, the better the model. We also define the information rate to make a comparison with previous works as below Eq. (28).

$$Infomation\ Rate(test\ set\ \boldsymbol{w}) = log\big(Perplexity(test\ set\ \boldsymbol{w})\big) \tag{28}$$

In Table 4.1, we summarize the number of words, perplexity performance, and the information rate (logarithm of perplexity) in our experiment. Here we compare with the previous study, McCallum et al. [33] proposed the ART model that distribution over topics conditioned on both the sender and recipient. Although their results show that the information rate is roughly between 8 and 9, there is no basis for defining the ART model's number of topics. If we consider the number of topics less than 30, the ART model's information rate is falling between 9.4 and 9.6. In addition to the ART model, the result of the information rate from other previous studies is larger than 10 [87], [88].

Table 4-1 the summarized results of perplexity measurement.

$\sum_{d=1}^{D_{test}} \sum_{n=1}^{N_d} -\log p(w_{dn}\|\theta_d, \phi_{z_{dn}}, h_{z_{dn}})$	*Perplexity*	*Word Count*	*Information Rate*
13706727.076438876	13845.903286	1437405	9.5357446763

4.4 The discussion of visual attention and keywords

After comparing perplexity evaluation, we selected eight topics with their highest conditional probability words from 21 topics to run for the Enron email corpus, respectively, as shown in Table 4.2 and Table 4.3.

Table 4-2 the top 10 probability words in four of eight topics.

Topic 2		Topic 3		Topic 5		Topic 8	
"Accountant"		"Evaluation"		"Transaction"		"Tax"	
blond	0.002003	srrs	0.007141	auction	0.011979	deserved	0.010099
xls	0.000239	unreadable	0.002585	shout	0.004261	turbotax	0.005857
accountants	0.000185	finley	0.002339	counterparty	0.004146	interrogatories	0.002221
offsetting	0.000130	policies	0.001846	userid	0.002188	expense	0.002019
worksheet	0.000119	captured	0.001846	wreck	0.001843	invoiced	0.001514
word	0.000119	using	0.001477	pamela	0.001727	coordinated	0.001413
heather	0.000098	egrb	0.001109	shoemaker	0.000576	nicolevancroft	0.001211
mpeg	0.000087	shout	0.000985	tactic	0.000576	update	0.000808
invoiced	0.000054	fledgling	0.000985	ditches	0.000576	ledford	0.000808
corman	0.000043	fischer	0.000861	warman	0.000460	shireman	0.000808

The quoted title in the table is we inference and summarize from the keywords. For example, the words "xls", "worksheet", "word" in Topic 2 are very relevant to word processing. The words "blond", "offsetting", "invoiced" give us the confidence that Topic 2 has a high degree of similarity with the Accountant. In Topic 3, the word "srrs" is the abbreviation of the Social Readjustment Rating Scale, which is a questionnaire used to evaluate, and we estimate the word "unreadable", "captured" is an action that people use the "srrs" for the evaluation. Finally, the word "policies" might be the results derived from the result of "srrs", and "fledgling" might be the target to do "srrs" investigation. Also, in Topic 8, although the word "invoiced" appeared in Topic 2 as

previous mentioned, however, since the word "deserved", "interrogatories", "coordinated" also appeared in Topic 8, we thought that this topic should be a more serious topic than Topic 2. Furthermore, according to the word "turbotax" and "expense", this serious topic might be related to tax.

Table 4-3 the top 10 probability words in four of eight topics.

Topic 12		Topic 15		Topic 18		Topic 19	
"Downtime"		"Sport"		"Recruitment"		"Marketing"	
twister	0.009180	game	0.005460	pics	0.007868	emarket	0.003823
drinks	0.004590	avi	0.001260	contract	0.006141	beep	0.001982
adventurevillage	0.004590	team	0.001260	fribble	0.004989	ask	0.001416
furches	0.004590	erica	0.001260	headhunter	0.004797	duplicate	0.001132
mead	0.004339	ballpark	0.001120	comments	0.004350	customers	0.000991
burn	0.002336	assistants	0.000980	hreich	0.004222	nemec	0.000849
clint	0.002003	queue	0.000840	hello	0.003838	hpln	0.000849
babysitter	0.001418	madhup	0.000700	schoene	0.003710	ads	0.000849
jean	0.001251	furches	0.000700	letter	0.002175	laporte	0.000849
interacting	0.001251	crowd	0.000560	variance	0.001983	concur	0.000708

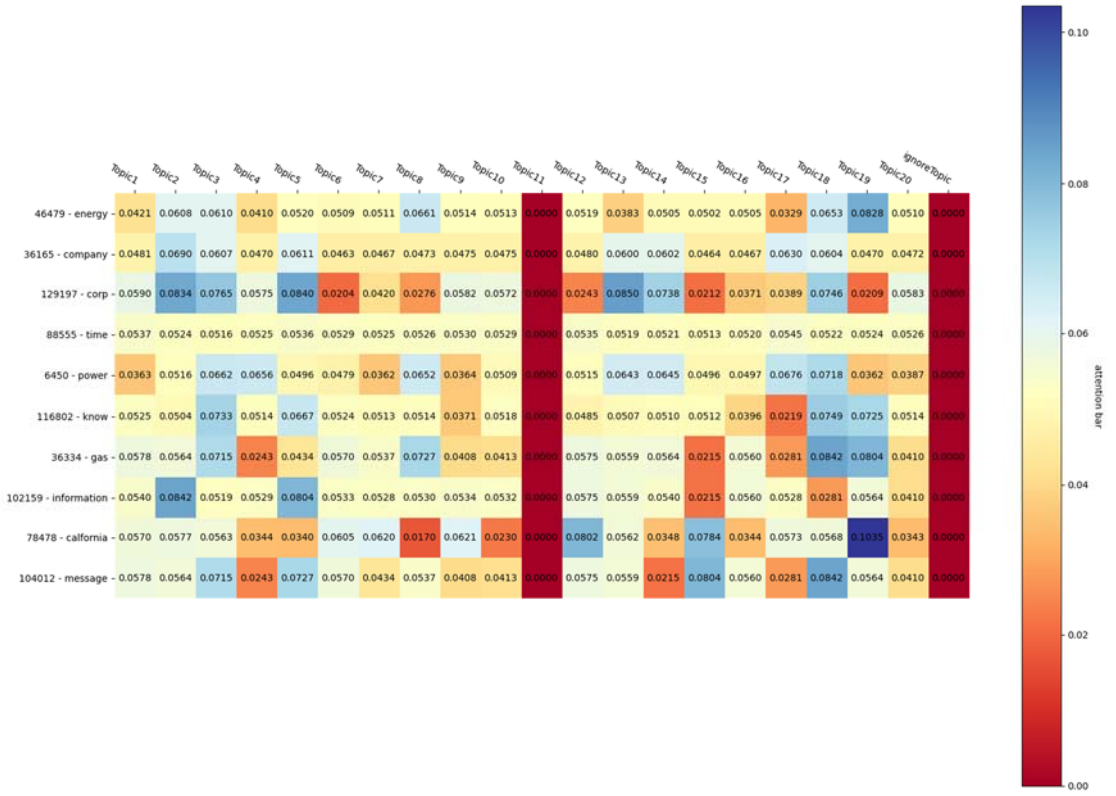


Figure 4.7 the attention value of the top 10 words from all the topics.

We identify divergent findings from those words with a high probability of a topic, and those with high attention value can discover a topic from another perspective. Some of the words with higher attention value than others but did not hold a high probability in a specific topic as we present in the above table, although their attention value higher than other words significantly. In that case, from the conclusions of the previous attention research [89], that attention influences the selection of stimuli of interest. Hence, we apply the heat maps to demonstrate the value of several words' attention value in each topic in Figure 4.7; the vertical axis is the word, where the number is the *word_id* defined by the vector w. There are several interesting patterns we found from this Figure. As we can see in Table 4-3, Topic 12 and 15 are not entirely related to the work; we assume it might relate to the company trip or the club activities. Therefore, only the word "calfornia" and "message" get a higher value of attention. On the other hand, the word "calfornia" also gets greater attention while talking about Topic 19 "Marketing".

Additionally, there is another interesting pattern while we focus on Topic 2 "Accountant" and Topic 8 "Tax". To our knowledge, both topics are severe issues in an organization. At the same time, we notice that the words "energy", "company", "corp", and "information" get higher attention value on these topics. Here we think the accountant covers broad business fields that focus not only on a single product or service. However, the Topic 8 "Tax", it might get a higher probability to focus on a specific target, that is why the word "energy", "power" and "gas" get greater attention in this topic.

Further, it is also reasonable that the word "calfornia" gets lower attention since it gets a higher value of attention while talking about Topic Downtime and Sport. Moreover, the word time get the average value of attention among topics but do not provide significant assistance. In the subsequent works, we can consider removing it to get better performance.

CHAPTER 5 Introduction and literature reviews for implementing the proposed SuccERP

As stated above, Prakash et al. [2] emphasized that collaborative systems address administrative work by facilitating the sharing and diffusion of information; however, two fundamental features need to be customized and paid attention to depending on the organization's goals. Thus, besides the email message, managing the messages and information from an ERP system is a noteworthy study. Also, the necessary extension of the business and an ERP system need to integrate with external systems and share information keeping a comprehensive manner has become an essential issue in the ERP post-implementation stage. An ERP system comprises multiple software modules to facilitate the flow of information in organizations. According to the American Production and Inventory Control Society (APICS) [90], ERP is a financial accounting-oriented information system that improves manufacturing firms' overall performance by effectively integrating and arranging all the resources, including resources for purchasing, production, distribution, and logistics operations. Mraz [90] further indicates that the strategic IT investment of ERP system can help sustain operational efficiencies in the long run.

However, some scholars have highlighted the considerable disadvantages of ERP systems. Carton et al. (p. 159) [91] commented that "ERPs are good for storing, accessing and executing data used in daily transactions, but it is not good at providing information" and "Many [organizations] experience frustration when they attempt to use their ERP system to access information and knowledge". While we try to improve the shortcomings by considering the integration approach, the implementation does not seem easy. There are quite different characteristics from legacy systems to ERP systems, including complex, large-scale, integrated, and interdependent [92], [93].

When it comes to ERP, most of the researchers will readily agree to divide the ERP project life cycle into multiple implementation stages and focus on each stage's different issues to propose their research. On the one hand, Esteves and Bohórquez [94] claim that we can classify the lifecycle into three implementation stages, including pre-implementation, implementation, and post-implementation. On the other hand, Swanson and Ramiller [19] propose that the lifecycle consists of three stages: ERP

adoption, implementation, and post-implementation. Others even defined various stages rather than three stages: scope and commitment, analysis and design, acquisition and development, implementation, and operation [95]. However, several studies have suggested that existing studies are primarily focusing on the adoption and implementation stages. The topic of maintenance and upgrading in the post-implementation has received less attention compared to others [96]–[100]. Esteves and Bohórquez [94] further present the relevant statistics from the existing research, with about 47% of ERP implementation studies, compared with only 15% of ERP post-implementation studies. Willis and Willis-Brown [5] point out that once an ERP system has been successfully built, it has a 'go-live' date; however, implementation is not the end of the ERP journey; the post-implementation or exploitation stage is where the real challenges begin.

It is clear that research in the post-implementation stage is necessary but lacking. To that end, scholars have given different opinions on the issues of the inadequacy of the post-implementation study. Some studies suggest that the post-implementation study should toward exploring the Critical Success Factors (CSFs) [97], [100]–[103]. Thus, a portion of the study identified several key factors critical to the post-implementation stage, including support from top management, user training, internal ERP team's competency, continuous process improvement, collaboration, communication, continuous systems integration/extension. Whereas, on the other hand, other scholars hold different views on exploring CSFs. They claim that the relevant literature remains limited to exploring the success factors and learning issues on the post-implementation stage [103], [104]. More precisely, Ram et al. [105] suggest that the researcher should focus more on improving performance to achieve real benefits as the success of ERP post-implementation rather than the current literature, which focuses on software selection and implementation processes CSFs.

Our objective corresponds to the suggestions about post-implementation from previous research; that is, Hsu et al. [96] assert that for effectively monitoring employee usage of the adopted ERP system in post-implementation, the managers should select meaningful information stores in the ERP system. However, companies usually encounter great difficulties maintaining or enhancing (to meet organizational requirements) ERP systems in post-implementation. To our best knowledge, only a few studies have put their effort into collecting meaningful data and information from an

ERP system and integrating with an advanced information system to let the manager precisely control the performance and situation of the adopted ERP system. Although ERP replaces some legacy systems (e.g., accounting, billing, order entry, etc.), it is challenging to integrate ERP systems seamlessly with other resources [28].

As our expected result, we focus on the post-implementation stage to allow the managers to realize the effective way to monitor and improve employees' efficiency. This paper presents the artifact - SuccERP, which can extract completely and convert data from an ERP system and achieve the interaction between the ERP system and the Enterprise Collaboration Systems (ECS). More precisely, instead of operating in the original ERP system, the user can carry out the ERP procedures in the ECS directly based on the support of the SuccERP, such as the inventory inquiries, order processing, and bill processing. Meanwhile, we will synchronize all the operations and data of the existing ERP system with the ECS. Each completed procedure has a corresponding message in the ECS to allow the managers and employees to do more in-depth communication. Furthermore, as Kwon and Lee [106] demand, the development of SuccERP will be along with the algorithms to present the details of our proposed system. In short, our demonstration also includes the pseudo-code to provide detailed steps and functions.

Afterward, we review the ECS since the SuccERP implements the integration between ERP and ECS. Later, a brief overview of the ERP system mostly focuses on the post-implementation stage. Finally, we review Design Science (DS) to interpret how we apply it to achieve the design, implementation, and evaluation of the proposed artifact-SuccERP.

5.1 Literature reviews of Enterprise Collaboration Systems (ECS)

From the 8C Model for Enterprise Information Management description, we can explicitly get the ECS outline. Employees can achieve collaborative works such as information and content sharing, communication, cooperation, and coordination through all areas of collaboration covered by ECS. On the whole, ECS is a software system that supports employees' collaborative work [107]. More specifically, the

composition of ECS consists of Enterprise Social Software (ESS) components (e.g., social profiles, tags, and blogs) with traditional groupware components (e.g., email, calendars, and document libraries). We can apply the ECS to support internal business communication, collaboration, and content and knowledge sharing activities [108].

Previous studies think highly of ECS, as Schubert and Glitsch [3] suggest that ECS has been seen as an essential enabler of the modern digital workplace. Unfortunately, research using ECS for further exploration is relatively fragmented and rarely provides in-depth empirical studies. Besides, ECS challenges are multifaceted and, therefore, often require to be addressed in different ways [108].

Schubert and Glitsch [3] depict ECS implementation projects' scope and possible conditions considering 26 case studies on ECS introduction projects. Although the small scope of cases limits the study's findings, they ensure that ECS is a suitable tool for longitudinal work according to identifying cases and scenarios. In other words, by using ECS, from the manager to the employees that the members of an organization can work effectively.

Greeven and Williams [109] identify the challenges and issues while organizations encounter during the introduction and use of an ECS. They propose five adoption challenge areas, including culture, business/operation, technology in use, benefits, and attitude/behavior according to academics literature and interviews. Significantly, they indicate the lack of activity in the period of ECS adoption that the ineffective content causes it. While ineffective content contributes to this problem primarily, low quality and insufficient usable content are also important factors. As a result, it leads the employees to move to other alternatives (e.g., email, previous generations of groupware software) that demonstrate the weak points from adopting the new system.

Diehl et al. [110] suggest that the laissez-faire approach cannot get the full potential to project success. Besides avoiding the laissez-faire approach, ECS is social software that social software presents cultural rather than technical challenges. These cultural challenges predictable, and we should manage and control them beforehand, not ad hoc. Most importantly, the adoption of trading systems such as ERP is almost mandatory, and the use of ECS is usually voluntary.

Schwade and Schubert [111] first refer to the findings from Herzog et al. [112] work, which introduces four methods with metrics to measure the success of ESS from the usage analysis. Later, based on that, Schwade and Schubert suggest considering the

database queries and log file analysis to measure the usage of ECS, for database queries and log file analysis, the number of posts, visits, creates pages per day or the average time per user per visit from Web Analytics are suggested.

From the above literature reviews of ECS, we highlight the views below:

1. Although ECS effectively supports collaborative work, the content of the ECS determines the user's attitude, either animated in using ECS or moving to alternatives. Hence, how to make the ECS content more significant is an important issue.
2. In the attitude's discussion of the users between ERP and ECS, ERP is almost mandatory, and ECS is usually voluntary. These outcomes open the door to studies that the possibility of combination or integration with ERP and ECS.

5.2 Enterprise Resource Planning (ERP)

Hasan et al. [103] propose an ERP success measurement model with seven interrelated, interdependent success factors that investigate the critical factors and evaluate the ERP's post-implementation stage performance. The research is motivated by the research trend that an enormous amount of research has been devoted to finding various factors for successfully implementing or adopting ERP.

Equally, from Ha and Ahn proposed model [113], they have identified six factors that influence the performance of ERP in the post-implementation stage, including top management support, competency of an internal ERP team, user training, inter-department collaboration and communication, continuous process improvement and continuous systems integration/extension. They also emphasize that continuous improvement in the post-implementation stage has a positive impact on ERP performance.

As we mentioned earlier, many studies have noticed that post-implementation research is lacking, and the enterprises have been struggling under this stage. The following studies explicitly point out the issues; Ali and Miller [100] indicate critical gaps in current research that only a few works concentrate on the post-implementation stage. There is no standard methodology that has been devised. Besides, the cost is also a significant limitation. Willis and Willis-Brown [5] argue that once the ERP system is set up, it has a 'go-live' date. However, this is not the end; instead, the post-

implementation stage's development process is a big challenge. Such proposals are analogous to other studies. Peng and Nunes [114] figure out that organizations often encounter various risks such as technical pitfalls, emergent business needs, inadequate user behavior, and poor system design when developing, maintaining, and enhancing the ERP system.

In discussing the literature review on post-implementation, one controversial issue has been what kind of studies should emphasize. On the one hand, Osnes et al. [102] argue a need for more research on Critical Success Factors (CSFs) in the ERP post-implementation stage. On the other hand, Amid et al. [32] contend current research has focused on software selection, implementing process, and CSFs rather than the success, actual benefits, and performance improvements of the ERP post-implementation stage. Others even directly claim the relevant literature is still limited to explore CSFs and learning issues in the post-implementation stage [103]. Our view aims to improve ERP performance by data integration and the sharing of information between the organization's systems and business processes. We propose a complete development process to let those enterprises lack experience and guidelines for continuous improvement and enhancement in the post-implementation stage, preventing them from encountering various risks appropriately, such as technical aspects. We have arranged the descriptions of how to develop, enhance, and upgrade the ERP system after implementation in the following Table 5-1.

Table 5-1 the arranged suggestion for the post-implementation stage of ERP.

Keywords / Source	Short conclusion
Internal hosting. / [115]	From this report, the ratio of hosting options between the external and internal is 22.6% and 77.4%. The reasons why organizations do not select cloud ERP include: (i) Risk of security branch (27.27%); (ii) Risk of data loss (31.82%); (iii) Lack of information/knowledge about offerings (40.91%). Another point is that the more employees use cloud ERP, the more licenses must purchase.
Functional upgrade, Enhancement. / [116]	There are two functional roles of an ERP in organizations: automating and informating. Automating is considered to

Automating, Informating. / [117]–[120]	be an essential function and has been thoroughly utilized to date. Informating role is defined as all the information generated during the work process using an ERP system and translating the description of activities, events, and objects into data to support work integration and decision support. The potential of an ERP depends on the informating role and data integration throughout the enterprise.
Information sharing, Communication, Monitoring. / [96], [101], [121], [122]	In order to solve unexpected projects, we need to develop an ERP system continuously. Meanwhile, both the continuous system development and ongoing business process monitoring were also identified as the ultimate success factor. Part of the studies involves the achievement of inter-department information sharing, as poor communication is considered one of the reasons for failure in the post-implementation stage. Besides, ERP systems are complex in nature, and the utilization of such complex systems often necessary to rely on external IS services in the post-implementation stage. Therefore, it is reasonable to assume that for higher extended use, the managers should be mindful of selecting metrics and information to monitor the employee's use of an ERP system. Moreover, keep these metrics and information accurate, updated, consistent, relevant, complete, and format is easy to understand.

5.3 Design Science (DS)

As stated above, most ERP researchers have focused on the subject of study in behavioral science research. We consider the theory-oriented cumulative knowledge base as a set of statements that summarized the critical factors and lessons learned from ERP research. The theoretical basis is the logical concern and product of both design and behavioral science scholars. Aboulafia [123] specifies the relationship between theory and artifact that the truth (justified theory) and utility (practical artifacts) are two sides of the one coin, and the scientific research should consider its practical implications for evaluation. Lee [124] elaborates that technology and behavior are not dichotomous in an information system; they are inseparable. To some degree, it also supports the perspective that the theory and artifacts are equally important.

Simon [125] advocates the natural sciences and social sciences' attempt to understand reality. Design science attempts to create things that serve a human purpose. The DS paradigm has its roots in engineering and the sciences of the artificial.

Hevner et al. [126] describe the DS research in information systems by proposed a conceptual framework that consists of seven guidelines: Design as an Artifact, Problem Relevance, Design Evaluation, Research Contribution, Research Rigor, Design as a Search Process and Communication of Research. Besides, they also summarized two fundamental questions for DS research: 'What utility does the new artifact provide?' and 'What demonstrates that utility?'.

Peffers et al. [127] propose and develop a design science research methodology (DSRM) for implementing the Information System. There are six activities in the DSRM includes: Problem identification and motivation, Define the objectives for a solution, Design and development, Demonstration, Evaluation, and Communication. Besides, there are multiple possible entry points for DS research; they show several cases start from different activities to carry out.

CHAPTER 6 Design and the Development of SuccERP

This section will refer to the six activities of the DSRM proposed by Peffers et al. [127] and seven guidelines addressed by Hevner et al. [126]. Davenport [6] further explains that an ERP system's heart is a central database used to draw the data and feed them into a series of applications to support the organization. The flow of information could throughout a business in an organization based on a single database. In order to provide more substantive guidelines of how we built our artifact based on the activities and guidelines to integrate with an ERP database, this section begins with the problem identification and motivation to the evaluation and demonstration. Also, some activities and guidelines that will moderately incorporate software engineering.

6.1 Problem Identification and Motivation

Software engineering is a widely used discipline, especially in the improvement and development of information systems. The purpose of the software specification function and the problem identification and motivation of DSRM is similar; therefore, we apply the user requirements definition and system requirements specification to achieve this work more systematically, which present in Table 6-1 and Table 6-2.

Table 6-1 user requirements definition.

Item	Descriptions
1.	How to enable an ERP system to communicate with external resources regardless of internal hosting or external hosting.
2.	How to provide an efficient way for a manager to monitor the operating performance of ERP.
3.	How to make ECS content more compelling (improving ineffective issues).
4.	For SMEs, how to strike a balance the enhancements and costs.

The items of Table 6-1 mainly refer to the literature review in the previous chapter, where we outline the items one-by-one below.

For item 1, we first consider the ratio of hosting options between the external and internal, 22.6% and 77.4% from the ERP report [115]. Hence, the post-implementation stage studies are impractical if without considering the issues of the ratio of internal

hosting still high. Further, although Liu et al. [128] identify the necessity for integrating an ERP system with other external resources, we have no choice but to face a growing challenge for developing and enhancing an ERP system. When implementing integrating an ERP system with other external resources, applying to both internal and external hosting is a more practical and complete solution.

Next, we consider both item 2 and item 3 at the same time. To some degree, the ECS and ERP can support the weak points of each other. Most organizations are struggling to share information effectively from an ERP for communication. Oseni et al. [129] show that for attaining ERP-based operational effectiveness, it needs to rely on the present advances in information and communication capabilities. Al-Mashari [121] presents that considering ongoing business process monitoring of ERP as a significant stage for completing a full process of ERP implementation.

Meanwhile, Hsu et al. [96] also suggest that managers consider meaningful metrics from the ERP system to monitor and test the performance from employees' use of an ERP system. As a result, selecting practical information for the managers to communicate and monitor is crucial in the post-implementation stage. Although the ECS can achieve collaborative works such as communication and suit for the longitudinal work, the ineffective content is the primary reason for the lack of activity in ECS adoption [3], [108], [109].

For item 4, Weston [130] and Elragal and Hassanien [131] present that the cost matters in determining whether the SMEs should carry out upgrade/enhancement or not. Peng and Nunes [114] further figure out that organizations often encounter a variety of risks of techniques when developing or enhancing an ERP system in the post-implementation stage. Hence, for small and medium enterprises (SMEs), both enhancements and functional upgrades require development costs and extra license costs with the increase in the number of users. Therefore, an explicit guideline of upgrades and enhancements can prevent encountering the risks of techniques; also, the cost of self-exploring and consultant charges can significantly declaim.

After proposing the user requirements definition, we summarize the system requirements specification to indicate how we define the proposed artifact's functions in Table 6-2. In the latter section, by going through the practical snapshots to show how we implement the system requirements.

Table 6-2 system requirements specification.

Item	Descriptions
1.	Regardless of internal or external hosting, the artifact needs able to access an ERP database.
2.	To collect the database schema and information of tables from the corresponding procedure of an ERP system.
3.	Provide the user interface in the artifact to allow the users to carry out the ERP procedures and deliver the data to external resources that allow the managers to monitor and evaluate the performance of employees.
4.	For SMEs, how to strike a balance the enhancements and costs.
5.	Synchronize the data between the external service and an ERP system to let all the members get information timely, whether the members access external service or an ERP system.

6.2 Define the objectives for a solution

Our work focuses on enhancement and functional upgrades in the post-implementation stage of an ERP system, especially communication and monitoring. The target is the SMEs because compared with a large enterprise, they have fewer resources. This research expects to create integration with the external resource to strengthen the communication and monitoring of an ERP, which led us to consider internal and external hosting issues. Also, synchronize the data between ERP and an external resource via artifact, and provide the user interface to enable users to carry out ERP procedures in an external resource.

We select ECS as the external resource and interpret the entire development process for creating the integration between ECS and ERP system by the proposed artifact. The reason to select ECS is describing below.

(1) ECS is a suitable tool for longitudinal work [3]; such a characteristic is ideal when an organization manager needs to monitor employees' performance.

(2) The adoption of an ERP system is almost mandatory, and the use of ECS is usually voluntary [110]. To our knowledge, with voluntary that the resulting content is relatively casual. It might be the primary reason for an

ECS with lower quality and not enough usable content [109]. However, ERP data sources are mandatory and valid that can improve these shortcomings of an ECS.

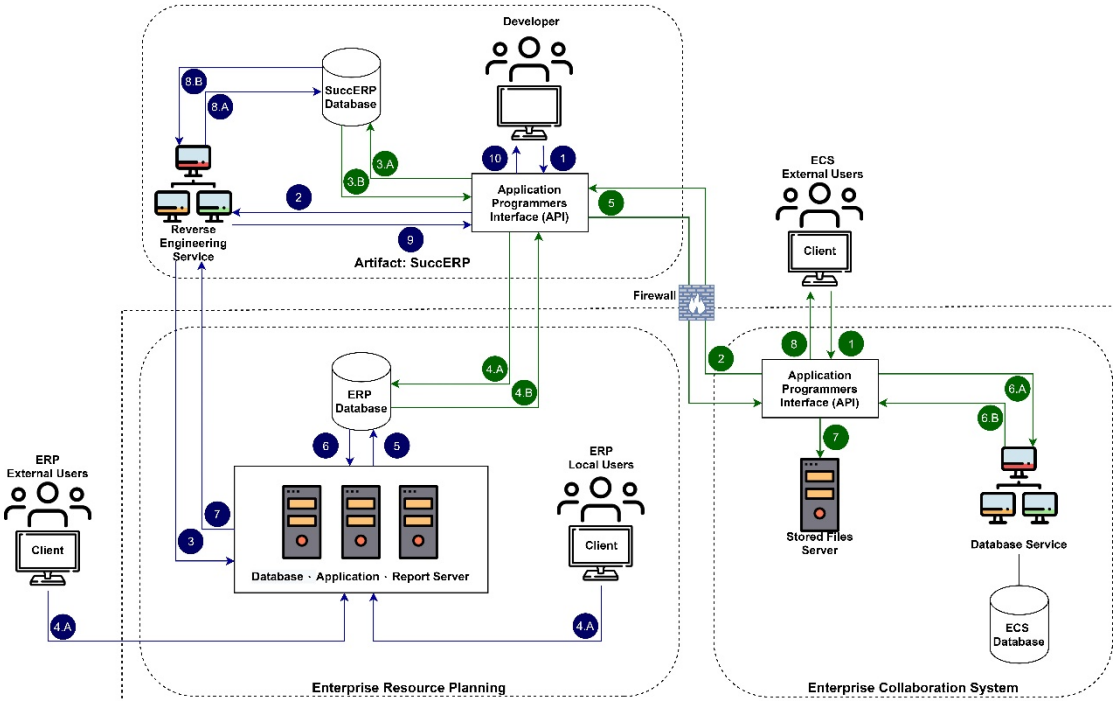


Figure 6.1 the architecture of proposed artifact-SuccERP.

We illustrate the architecture diagram in Figure 6.1, which provides a more applicable consensus of how the integration between the proposed artifact-SuccERP, an ERP system, and an ECS interact substantially. There are four kinds of roles in the architecture diagram, including the ECS external user, ERP external user, ERP internal user, and developer. It is worth mentioning that there are two primary actions in this architecture, and we used the arrow with different colors for recognizing.

The blue lines are used to show an action that grasps the database schema and corresponding tables info, then inserts into the SuccERP database for achieving the actions represents with green lines. We describe it step by step below.

- **Step 1 and Step 2**: The developer initializes the Reverse Engineering Service (RES) by invoking the SuccERP Application Programmers Interface (API).
- **Step 3**: The RES is ready to grasp and get the database schema and corresponding table info with the Database Profiler after initialization. We applied the Database Profiler to grasp the Database and Application server's

information and rules after the ERP user carried out the ERP procedures in Step 4.

- **Step 4**: In this step, the ERP system's specific procedure is carried out by the client in both external hosting with a physical address and internal hosting with a dynamic address.
- **Step 5 and Step 6**: After receiving the ERP users' requests, there is a series of data processing between the ERP database and server.
- **Step 7**: Then, RES grasps and receives those necessary schema and table info.
- **Step 8**: The RES insert the schema and tables info of the ERP database into the SuccERP database.
- **Step 9 and Step 10**: The RES returns the confirmation message to the SuccERP API and the developer.

We turn to the green part of Figure 6.1, where we describe how the ECS users with the proposed artifact-SuccERP to carry out the ERP procedures with the ECS directly.

- **Step 1 and Step 2**: The ECS users will deliver the request for executing the specific ERP procedures to the SuccERP API via the ECS API.
- **Step 3 and Step 4**: While the SuccERP API receives the request from the ECS API, then the SuccERP API will request the database schema and tables info from the SuccERP database to execute the corresponding procedures in the ERP database.
- **Step 5**: After completing the data processing from the SuccERP to an ERP database, sending a confirmation message to the ECS API.
- **Step 6**: Create a corresponding message in the ECS system for further communication and the practical content show in Figure 7.6.
- **Step 7**: Upload the relevant report files to the stored files server.
- **Step 8**: Return the confirmation message to the ECS users.

The blue and green parts are two independent actions. Typically, the blue part executed at the beginning only, and we describe the subsequent operations from the ECS users as the green part, such as order creation, inventory inquires, and bill of purchase.

6.3 Design and Development: The solution for the hosting issues

Regarding the keywords of Internal hosting from Table 5-1, as well as item 1 of Table 6-2 for defining the system requirements, we apply Figure 6.2 to interpret the productive solution. First of all, the limitation on the ERP for internal hosting is without a physical address for identification. In other words, SuccERP API must be able to move to the ERP local environment and hold the ability to connect with the ECS API.

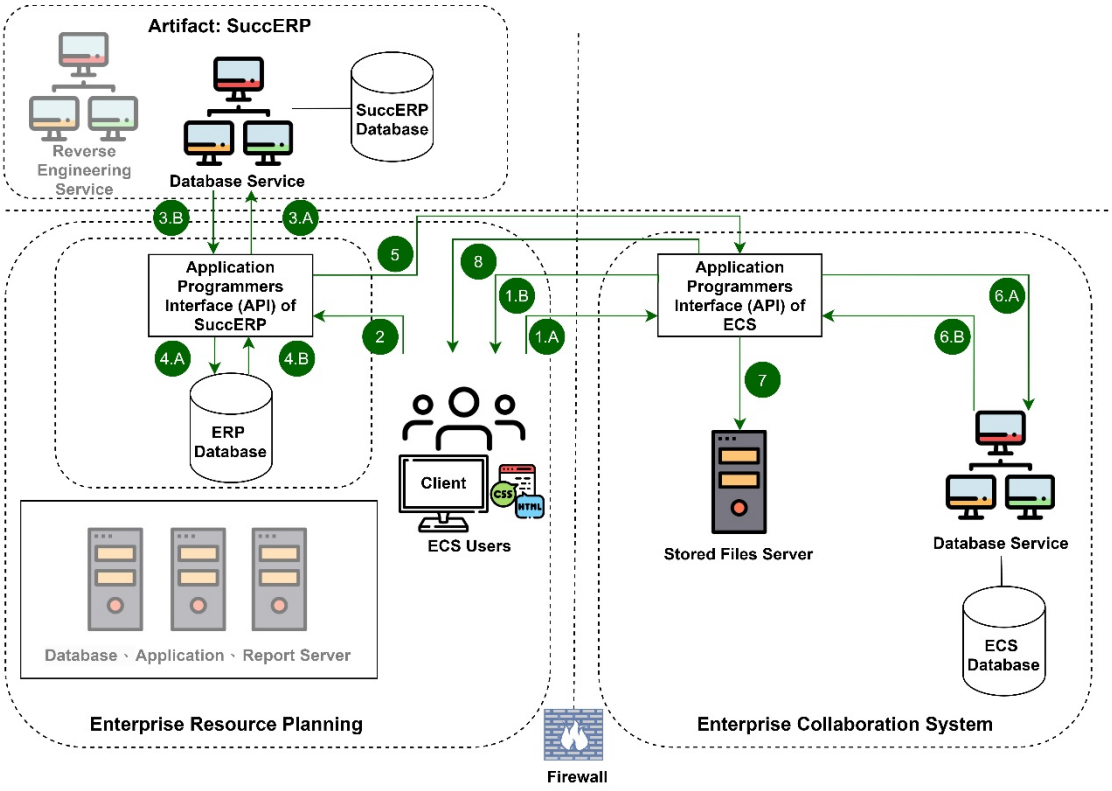


Figure 6.2 the architecture for internal hosting of SuccERP.

The key point in Figure 6.2 is the SuccERP API that deploys in the internal hosting environment with the ERP system. In internal hosting, users are only allowed to access the ERP with a dynamic address. As mentioned earlier, in Figure 6.1, the action constructed by blue lines is only executed at the beginning to recognize the database schema and tables info of the ERP database. Hence, in Figure 6.2 that we only describe the action constructed by green lines. Besides, the servers and services relevant to the blue lines turn to the semi-transparent color in the diagram that they have no connection with this solution.

The users will access ECS and SuccERP API via the web browser as a local client. Here we describe the steps that differ from Figure 6.1.

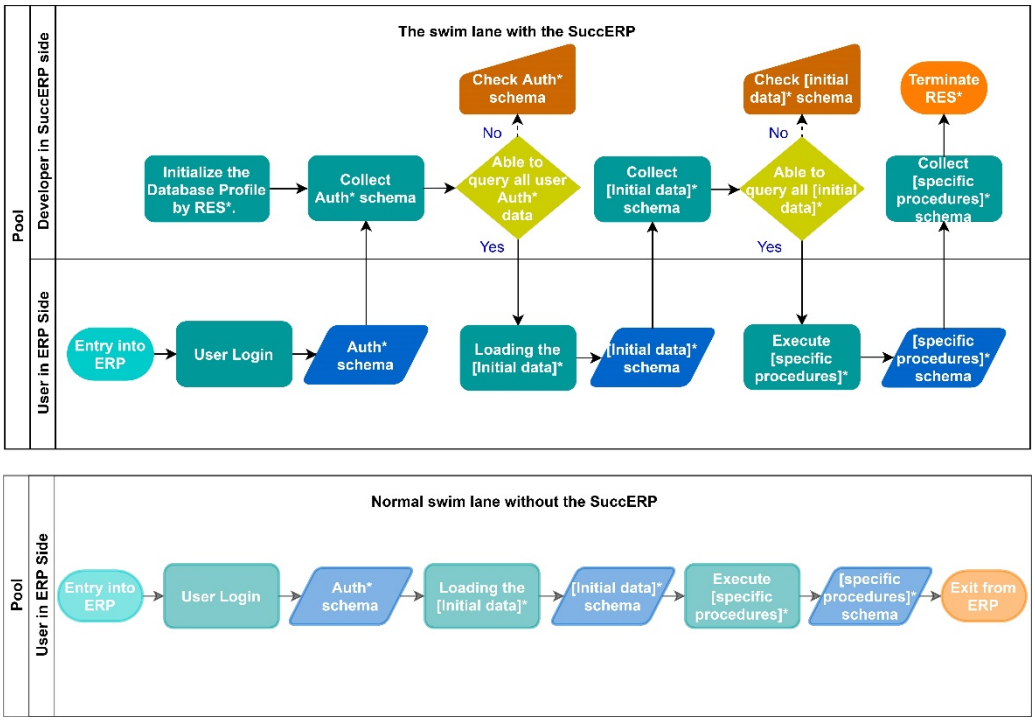
- **Step 1**: We divide this step into two parts, mainly; Step 1.B: which is used to request the initial data from a client user of ECS for keeping those data in the local environment.
- **Step 2 to Step 4** is used to describe the ECS user's requests to carry out the ERP specific procedure in ECS.
- **Step 5 to Step 7** is used to create the ECS's corresponding message after the data-processing in the ERP database is completed.
- **Step 8** will send a confirmation message to the ECS user after completing the ECS message creation.

In summary, the SuccERP API can change the role according to the environment; the role of SuccERP API in Figure 6.1 is as a bridge that ECS API delivers the data to the ERP database by the SuccERP API. On the other hand, under the local environment's limitation, it can be treated as the only pathway for communicating with external services, as we describe in Figure 6.2.

6.4 Design and Development: Collect the outline and schema of ERP Database

The Reverse Engineering System (RES) is an initial point under the SuccERP architecture; also, it is the heart of the action constructed by blue lines in Figure 6.1. By invoking the Database Profiler, the RES used to request and grasp the statements executed in the ERP database from the ERP users login into the ERP system to carry out the specific procedures.

We illustrate in Figure 6.3, which consist of two kinds of swim lane for the comparison. The bottom part of Figure 6.3 is the general ERP process without SuccERP. On the contrary, the upper part is in the case with the SuccERP. The essential part is verifying the collected schema and table info able to select the relative data from the ERP database according to the two decision processes in the swim lane.



RES*: Reverse Engineering System
Auth*: Authentication
[Initial data]*: (Company, Sheet Rule, Items, Warehouse, Customer, TaxType, etc.)
[Specific procedures]*: (order creation, bill of purchase creation, inventory inquiries, warehouse operation, etc.)

Figure 6.3 the swim lane diagram of Reverse Engineering System (RES).

We consider multiple ERP systems to verify whether the RES is well working or not and mention those substantial parts below.

- **Part 1**: First, before the user login into an ERP system, the developer initializes the Database Profiler to waiting for getting the statements and corresponding authentication schema by invoking SuccERP API. After that, collect the authentication schema and relevant table info into the SuccERP database during user login into an ERP system. The schema and tables info is the basis for subsequent integration between SuccERP API and ERP system. There are two common scenarios of authentication schema.
  - We describe the first kind of scenario in Figure 6.4. The authentication data is collected in an independent database Auth*, and the login data of each user is preserved in a table of the database Auth*. The point is, they summarize all the authentication data in a database and determine which companies that a user allows accessing according to the specific column (i.e., MA003) of the table dbo.Table N. Also, each company has an independent database.

- In Figure 6.5, the second scenario is distinct; they separate the authentication data into two parts. First, they collect all the gathered company list in table dbo.Table M in the independent database System. Subsequently, it determines whether a user is allowed to access by table dbo.comPerson from a database of each company.

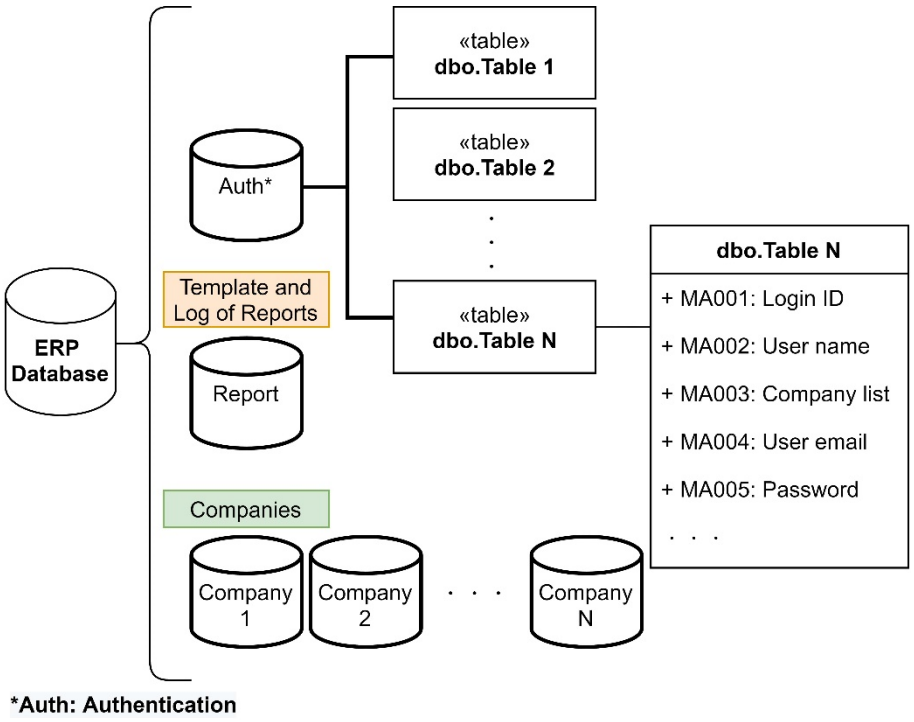


Figure 6.4 the schema of authentication in an ERP system.

- **Part 2**: Then, we make the short conclusion of the initialize data schema in Table 6-3. We can categorize the initial data into ten categories: user department, company info, warehouse, items type, items collection, sheet rule, invoice, currency, customer & supplier, and tax. To our knowledge, most ERP systems are following this category to initialize the necessary data. Therefore, we summarize it to let enterprises have these rules in hand for upgrading or enhancing the post-implementation stage.
- **Part 3**: Finally, regarding the execution of the specific procedures, we considered the four most frequently used procedures, including order creation, bill of purchase creation, bill of sale creation, and inventory inquiries. Equally, we arrange the comparative tables associated with these four procedures in Table 6-4.

Based on RES, researchers and developers can get the scope of an ERP database. In our perspective, this is an essential and first step for the post-implementation stage. Especially in the authentication part, there is a significant difference between different ERP systems. Next, we describe how ECS integrates with ERP while carried out ERP procedures.

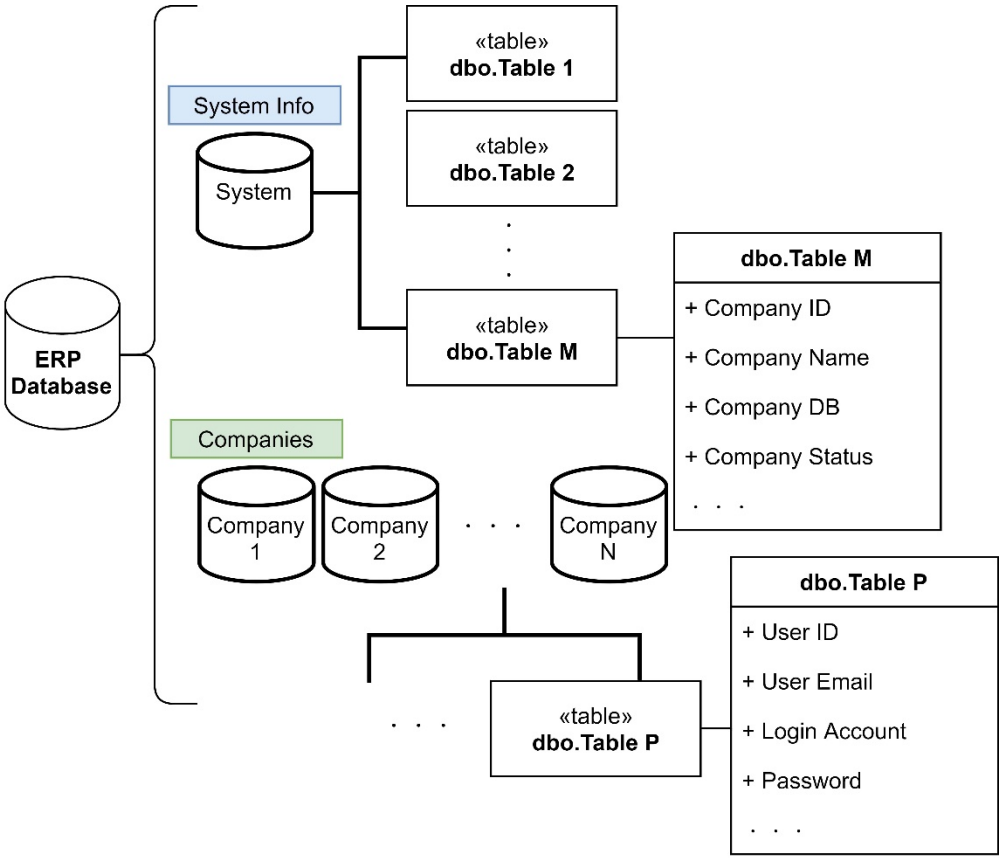


Figure 6.5 The schema of authentication in an ERP system.

Table 6-3 the relative tables and descriptions in initialize data step.

Category	Short descriptions
User department	An ERP system goes through the department of a user to govern the permission.
Company info	The company info typically includes the company name, unique identifier code of the company, telephone, fax, address, and the unique identifier code of administrator columns.
Warehouse	There are two dependent tables relevant to the warehouse. One is the collection of warehouse includes a unique identifier code of

	warehouse and name. The other is used to collect the inventory of each item in a warehouse. While we try to carry out the procedure of inventory inquiries, we must identify how many warehouses in a company first confirm the inventory of each warehouse.
Items type	The items typically comprise finished goods inventory, semi-finished products, and defective products, Etc. Besides, a single warehouse might comprise multiple types of items.
Items collection	The items collection table typically includes a unique identifier code, name, specification, items unit, item type, and unit price columns. In some cases, they summarize all the inventory from each warehouse as an on-hand inventory in this table.
Sheet rule	This table is used to indicate how to generate the unique identifier code of each sheet.
Invoice	This table is used to preserve invoice records of purchases and sales, including invoice type, tax rate structure, and tax name columns.
Currency	Each company might have multiple options on the currency. This table typically includes the currency name, rate, and a unique identifier code column.
Customer, Supplier	These two data are similar; the customer and supplier data are preserved in a single table in many cases. It typically comprises a unique identifier code, name, liaison, tax ID number, telephone number, fax number, and address columns.
Tax	This table collects a list of tax types. It comprises the tax rate value, name of the tax, and a unique identifier code column.

Table 6-4 the relative tables and descriptions in the step of four specific ERP procedures.

Category	Short descriptions
Header of an order, Header of a bill.	Each sheet with a single header-only and the important columns includes order serial number, date, and identifier code of customer or supplier, currency, an address for delivery, tax type, tax fee, sum total, and sum quantity.
Content of an order, Content of a bill.	At least one record per sheet in the content and the critical columns include order serial number, index number, relevant data of items, and relevant data of warehouse, quantity, unit, unit price, and subtotal.
Log data of an order, Log data of a bill	While the user inserts or updates a sheet, there is a record insert into this table. Typically, the crucial columns are similar to the tables of the header category, as described above.
Inventory	There are two kinds of the scenario for recording the inventory of each item. One only records the inventory of items in each warehouse. The other is beside the inventory of each warehouse's inventory, and record all the inventory of each item on-hand in another table.

6.5 Design and Development: The integration between ECS and ERP.

Let us briefly look back on why we want to integrate with ECS and ERP from the conclusion and suggestion of previous works. First, the ERP system contains much information sufficient for management. However, collaborating and communicating with the ERP system is partly ineffective, much less testing employees' performance. While ECS has the advantages of facilitating communication and collaboration, it usually contains ineffective and useless content. We believe that ECS and ERP offset each other's lack and settle ERP's severe issues in the post-implementation stage.

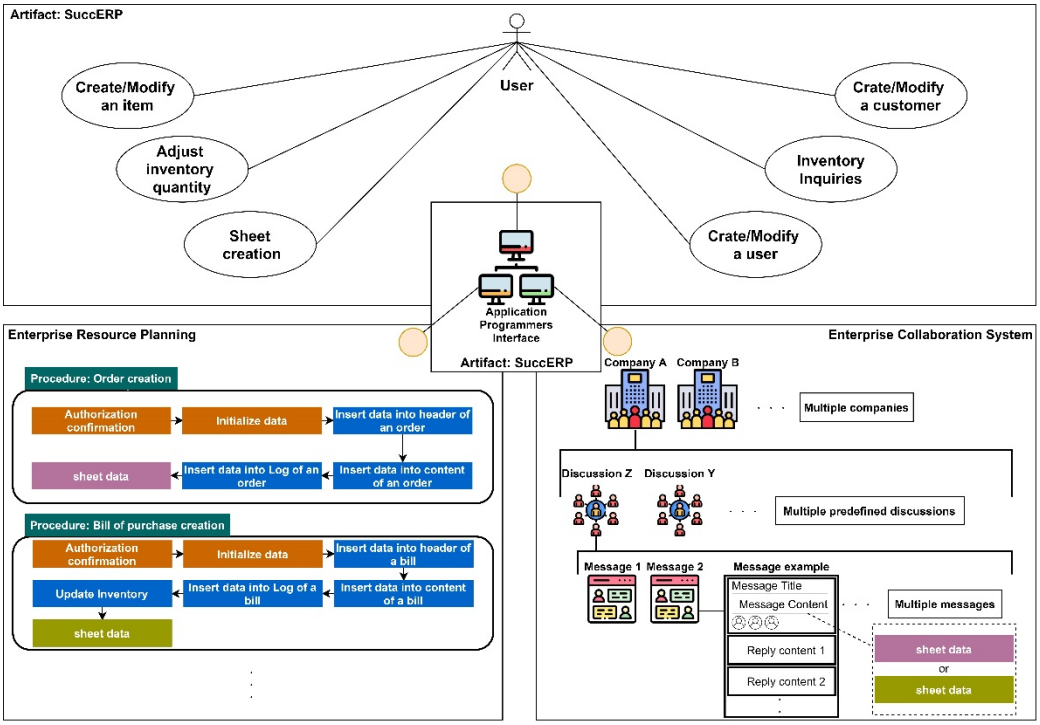


Figure 6.6 Use case diagram and the illustration of SuccERP.

In Figure 6.6, we present the detailed integration between ECS and ERP system via the proposed artifact that is the more interpretive version of Figure 6.1, especially in terms of processes and data flow. From our previous work, Lin et al. [132] presented SuccMail (https://succmail.com) as an ECS to achieve the research purpose of the integration between ECS and ERP system as an initial exploration. Regarding the use of ECS, we can discover some interesting features. The application of an ECS is hard to prescribe, and it is not easy to develop manuals or guidelines for its utilization. From the theory on the social construction of technology (SCOT), we can explain that the evidence of the ECS is open to interpretative flexibility [133]. The practical cases define part of the features. For example, the software IBM Connections provides users with multiple options for creating templates of activities with a list of tasks. Creating such a template is simple, but the areas of use are endless.

Equally, during our research, we found evidence for such a task by integration with the ERP system and the SuccMail. We identify these areas for the use case to implement the specific purposes in an enterprise and require the employees' individual creativity. As illustrated in the right-bottom corner of Figure 6.6, the ECS software SuccMail provides managers and employees with the possibility of communicating/collaborating with each in the lists of enterprises they took part in.

With the SuccMail, the managers will pre-define several discussion groups that give an initial name to reveal the group's expected topics. People assign each discussion group with different task orientations that compose members according to what they are in charge of in an enterprise or organization. Each member can assign the receiver while they create a message within a discussion. Each message comprises a subject (title and content) and multiple replies similar to the structure of an email.

By supporting from the SuccERP, the relevant data of the specific ERP procedures will be fully synchronized to a message of a discussion in the SuccMail, such as order data. For communication, it allows the managers and employees can create replies to do further discussion and collaboration. Considering the suggestions from Kwon and Lee [106], we present the pseudo-code below and the functions of each table of the ERP database.

Algorithm 1 Synchronize data while accessing a message in ECS

procedure SYNCHRONIZE DATA WITH ERP VIA SUCCERP(sheet)
  $C \leftarrow \{comp\_id, user\_id, ip\_address, db\_name, user\_name, password, host\_type\}$
  $D \leftarrow \{order\_id, bill\_id, flag, message\_id\}$ ▷ params for synchronize
  $targetHost \leftarrow GENERATE\_HOST\_URL(host\_type)$
  $connRes \leftarrow CONN\_SERVER(C, targetHost)$
  if $connRes.IsSuccess == FALSE$ **then**
    return FAIL;
  end if
  $bindingUserRes \leftarrow USER\_BINDING(user\_id, user\_email, targetHost)$
  if $bindingUserRes.IsSuccess == FALSE$ **then**
    return FAIL;
  end if
  $bindingCompRes \leftarrow COMP\_BINDING(comp\_id, targetHost)$
  if $bindingCompRes.IsSuccess == FALSE$ **then**
    return FAIL;
  end if
  $sheet\_data \leftarrow GET\_SHEET\_DATA(D, targetHost)$ ▷ whether an order or bill data depending on the case
  $UPDATE\_MESSAGE\_DATA(sheet\_data)$
  if $sheet == ORDER$ **then**
    $UPDATE\_MSG\_BINDING\_INFO(order\_id, targetHost)$
  end if
  if $sheet == BILL$ **then**
    $UPDATE\_MSG\_BINDING\_INFO(bill\_id, targetHost)$
  end if
  $REFRESH\_REPORT\_DATA(sheet\_data)$ ▷ remove and upload a new one
end procedure

Figure 6.7 Pseudo-Code: synchronize data while accessing a message in ECS.

The Pseudo-Code scenario: synchronize data while accessing a message in ECS in Figure 6.7. While a user accesses a created message in the SuccMail, the ECS API will send multiple requests to the SuccERP API to determine whether it needs to update the message. Although we conduct the SuccMail as the example of the ECS system in our research, the Pseudo-Code: synchronize data while accessing a message in ECS we present in Figure 6.7 is not limited to specific systems and programming languages.

It begins at the user accesses a created message; the ECS API will apply the function *CONN_SERVER ()* to confirm whether it can create a connection with the ERP database and ensure the ECS user has permission to access the company of an ERP. It is worth mentioning that considering internal hosting. We apply the function *GENERATE_HOST_URL ()* to determine the way of connection with the ERP database.

Next, the ECS API delivers an ECS user's user data as a parameter to the function *USER_BINDING ()* for requesting the user binding, typically depends on the email address. Each created message in the ECS will record the unique, identical code of ERP company, which is the parameter for applying the function *COMP_BINDING ()* for identifying which ERP company is the target and accessing it for subsequent requests.

Then, according to the result of the request from the function *GET_SHEET_DATA ()* to update ECS's corresponding message and the binding info. Finally, each sheet (bill or order) report will remove and generate a new one by the function *REFRESH_REPORT_DATA ()*, if the sheet has changed. The format of the report is in PDF file.

Compared with the Pseudo-Code: synchronize data while accessing a message in ECS in Figure 6.7, the Pseudo-Code: sheet creation in Figure 6.8 is more complicated than the previous one. Hence, we will specifically explain the function with corresponding table categories, summarized in Table 6-3 and Table 6-4. Besides, our description will begin from function *SELECT_DISC ()* because we skip the same parts in Figure 6.7. The Pseudo-Code: sheet creation in Figure 6.8 mainly describes sheet creation, including order creation, bill of purchase, and bill of sales creation.

Algorithm 2 Sheet creation

```
1: procedure CREATE SHEET( )
2:     C ← {comp_id, user_id, ip_address, db_name, user_name, password, host_type}
3:     D ← {flag, message_id}                          ▷ params for sheet creation
4:     targetHost ← GENERATE_HOST_URL(host_type)
5:     connRes ← CONN_SERVER(C, targetHost)
6:     if connRes.IsSuccess == FALSE then
7:         return FAIL;
8:     end if
9:     bindingUserRes ← USER_BINDING(user_id, user_email, targetHost)
10:    if bindingUserRes.IsSuccess == FALSE then
11:        return FAIL;
12:    end if
13:    bindingCompRes ← COMP_BINDING(comp_id, targetHost)
14:    if bindingCompRes.IsSuccess == FALSE then
15:        return FAIL;
16:    end if
17:    discGroup ← SELECT_DISC()
18:    initData ← INITIALIZE(C)
19:    sheetCode ← GENERATE_SHEET_CODE(flag, initData)
20:    sheetDate ← ASSIGN_DATE()
21:    sheetObject ← SELECT_OBJECCT(initData)
22:    sheetCurrency ← SELECT_CURRENCY(initData)
23:    sheetTax ← SELECT_TAX(initData)
24:    INIT_EMPTY_SHEET_ITEM()
25:    sheetData ← FILL_SHEET()                        ▷ the action from operator
26:    while CREATION_SUBMIT do
27:        MsgTitle ← GENERATE_TITLE(sheetData)
28:        MsgContent ← GENERATE_CONTENT(sheetData)
29:        TaxInfo ← CALCULATE_TAX(sheetData, sheetTax)
30:        TotalQuantity ← CALCULATE_QUANTITY(sheetData, sheetTax)
31:        ReportInfo ← GENERATE_REPORT(sheetData, sheetCode, sheetDate)
32:        CREATE_SHEET(sheetData, sheetCode, sheetDate, TaxInfo, TotalQuantity)
33:        CREATE_MESSAGE(discGroup, sheetCode, sheetData, MsgTitle, MsgContent)
34:        UPDATE_INVENTORY(TotalQuantity)
35:        CREATE_SHEET_LOG(sheetData, sheetCode, sheetDate, TaxInfo, TotalQuantity)
36:        UPLOAD_REPORT(ReportInfo)
37:    end while
38: end procedure
```

Figure 6.8 Pseudo-Code: sheet creation.

- Function *SELECT_DISC()* is used to provide a list of discussion groups for the ECS user, and the user assigns a discussion group to the creating message associated with the ERP sheet in the ECS.
- Function *INITIALIZE()*is used to read all the necessary initial data from the ERP database via the API of SuccERP. These data are summarized into categories, as shown in Table 6-3, and we discuss the critical differences from various ERP systems while initial the data.
  (1) **Company_info:** It is used to show the information on the current ERP Company connected. In our case, reading the information of ERP Company is not only at a particular point in this time (initializing) and

beginning to generate the report data after the sheet creation. Other users might modify the ERP system between invoking the *INITIALIZE()* function until submitting the sheet creation request. In other words, our artifact will send the request to get the company info again to ensure the data consistency between the ERP system and ECS.

(2) **Warehouse:** Regardless of order, bill of purchase, or bill of sales creation, the content items must be associate with a warehouse. Hence, a sheet might relate to multiple warehouses. While we do the inventory, inquiries, or adjustments need to refer to the warehouse data. As describe in Table 6-4, there are two kinds of scenarios. While an ERP system is with recording all inventory of each item on-hand in another table, that the ECS API will sum the inventory from each warehouse and update the corresponding record in that table.

(3) **Items type:** The items type contain finished goods inventory, semi-finished product, defective product, Etc. A single warehouse might comprise multiple types of items.

(4) **Sheet rule:** The rules of identical code of an order or bill. Generally, the code consists of 4-digits for the year, 2-digits for the month, 2-digits for the date, and 4-digits indicating how many orders or bills were created in a day. In most ERP systems, they record their rule in a table. Some defined the rules in the ERP system instead of the table; in that case, the developer needs to check the rule manually.

(5) **Customer, Supplier:** The object of an order is a customer, and the object of a bill is a supplier. In some ERP systems, they used different tables to collect the customer and supplier data separately. However, others apply an additional column as a flag to identify the collected data according to the customer and the supplier based on a single table. Hence, we need to deliver one more parameter while an ERP system collects both customer and supplier into a single table.

- Function *GENERATE_SHEET_CODE()* is used to generate the latest unique identifier code according to the sheet rule and the number of sheets created on that day. According to the sheet type (order or bill), consider the corresponding table to estimate the number of the created sheet on that day.

- Function *SELECT_OBJECT()* is according to the sheet type (order or bill) to provide the customer or supplier list options. While the user selects an object that the artifact will load the relevant data, including address, phone number, and liaison.
- Function *SELECT_TAX()* most ERP systems provide several options such as taxable, exemption, and untaxed. These tax options need to take into account while calculating the total of a sheet.
- Function *INIT_EMPTY_SHEET_ITEM()* is used to generate a new item in the sheet's content after initializing the data. Our artifact's user interface allows the user to add a new item in the sheet's content by themselves.
- Both functions *GENERATE_TITLE()* and *GENERATE_CONTENT()* are used to generate the message of ECS, and the generated message looks like the example we illustrate in the right-bottom corner of Figure 6.6. We organize the sheet from the ERP database into completed message information via these two functions. These two functions provide more possibilities to the managers and developers in the post-implementation stage, such as transferring the ERP data into other languages, building evaluation up with more metrics/indicators, or keeping the prompting messages for further communication ECS.
- Function *CALCULATE_TAX()* is according to the user's selected tax option to calculate the total and the Value-added Tax (VAT). The tax rate value is a record in a table as the category tax we describe in Table 6-3.
- Function *CALCULATE_QUANTITY()* outputs a pair of data with the item's quantity and the corresponding warehouse. It is the necessary parameter for the following function *UPDATE_INVENTORY()*.
- Function *GENERATE_REPORT()* is used to generate the corresponding report for each sheet, usually in PDF format.
- After the report is generated, the artifact uploads the report with function *UPLOAD_REPORT()* to an ECS side's stored file server.
- Function *CREATE_SHEET()* is used for the sheet creation. The parameters will separate into two parts: the header and the other is the content. As the categories of a table that we mentioned in Table 6-4, the SuccERP API will submit the parameters for insert, update, or query the relevant table.

- Function *CREATE_MESSAGE()* is used to create a message of the ECS for further communication and monitoring. The two most important parameters (title and content) are generated by functions *GENERATE_TITLE()* and *GENERATE_CONTENT()*.
- After completing the message creation and sheet creation, the artifact will insert the log data to the ERP system according to the type of sheet by function *CREATE_SHEET_LOG()*.

CHAPTER 7 Results, Demonstration and Evaluation of SuccERP

In this chapter, it begins by using the sequence diagram to describe how those functions of the Pseudo-Code: sheet creation in Figure 6.8 work between API services, servers, and database sequentially. Besides, the demonstration also arranges snapshots in pairs with each function.

7.1 The Demonstration of proposed SuccERP

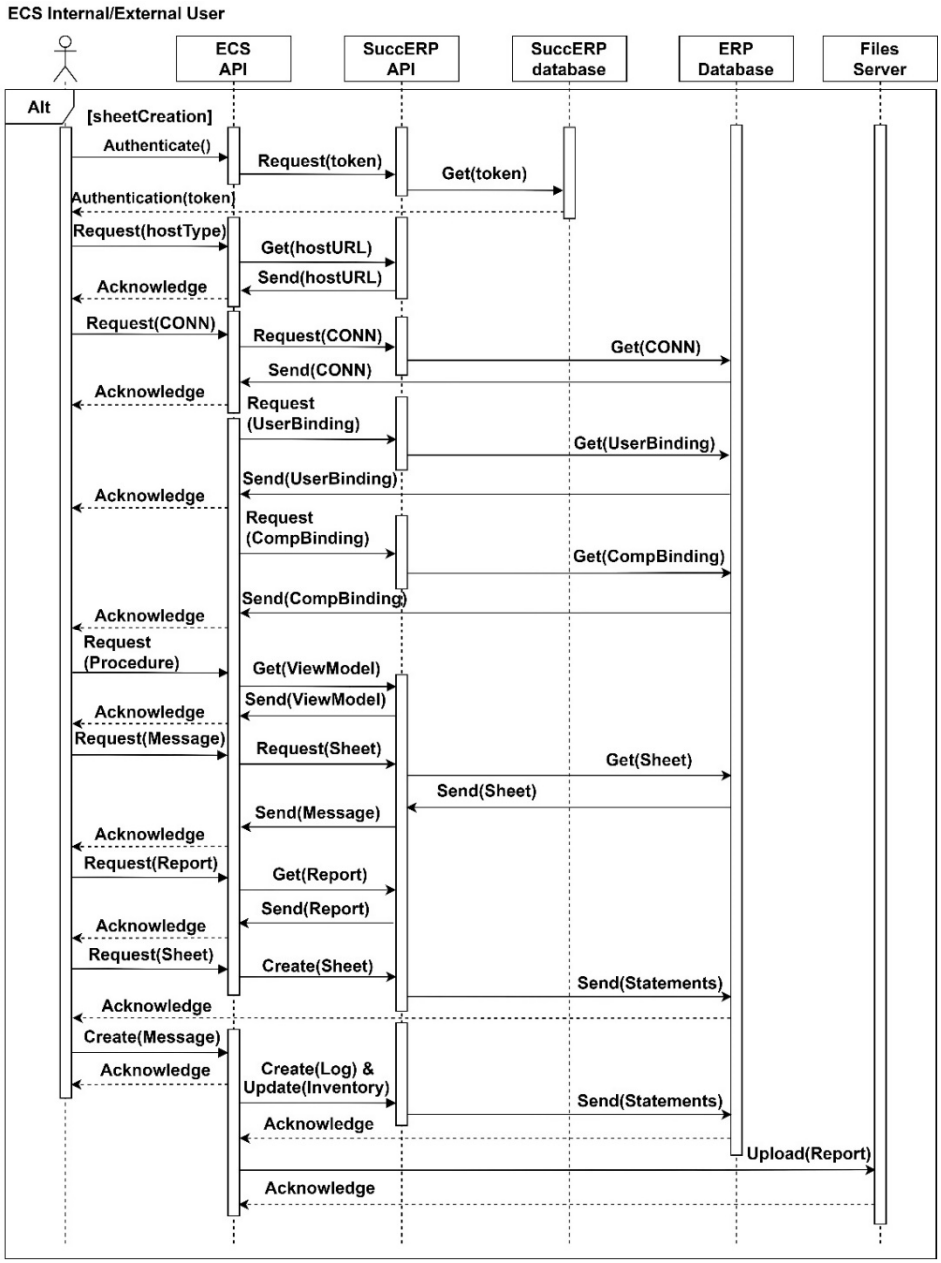


Figure 7.1 the sequence diagram for the case of sheet creation.

In Figure 7.1, we first provide the physical and dynamic IP options that users can decide how to operate ERP depending on their current environment. Our proposed artifact - SuccERP has been deployed in the public server. However, as we discussed earlier, some enterprises struggle with internal hosting issues that might not build a connection with the public server. Hence, we also provide the guideline about installing and deploying SuccERP within a server in the local environment instead of the case by accessing the public server. As shown in Figure 7.1, we provide a red label for each solution to let the user download the SuccERP installation files and guidelines.

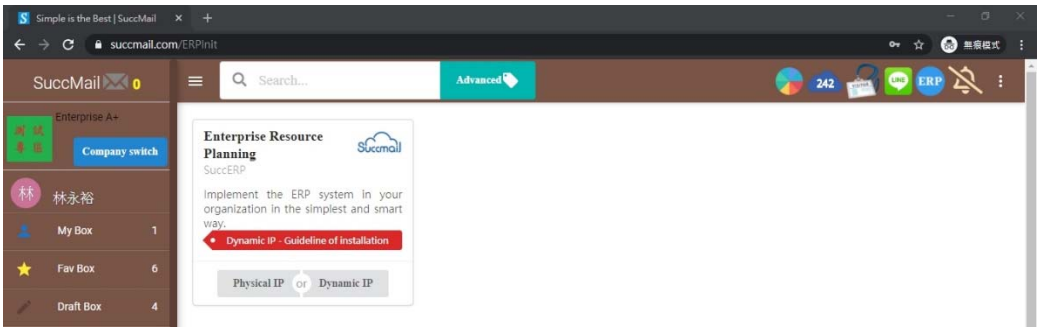


Figure 7.2 snapshot of SuccERP for providing options to decide which type of connection prefer in the current environment.

After confirming the connection between each service and servers also work well, the ECS API will send a request to the SuccERP API for the user binding that the purpose is to create and assign the association between these two user data. As we mentioned earlier, typically, the association is according to the email of the user. As shown in Figure 7.3, the snapshots show that three companies existed in an ERP system. The user will choose one of the companies to connect and bind their user information with the existing user in the ERP system by email address afterward. Then, provide users with the options of ERP procedures, bill creation, and order creation. Finally, return the ViewModel of sheet creation as a user interface and related initial data to the ECS user for subsequent processes according to the selected option.

Figure 7.4 is the user interface for the sheet creation, which allows the user to carry out the specific procedure of the ERP system within our artifact. In this example, the user has complete most of the content of this sheet. There are three items in the sheet's content, and it will create the message in the BILL OF SALE group of the Enterprise A+ company. The user can create more or fewer items by clicking the add button or the remove button. While the user fills all the necessary data of the sheet, after they click

the submit button, the ECS API will send several requests to the SuccERP API and ERP database, including creating a sheet, generating a report, creating the logs, and update the inventory.

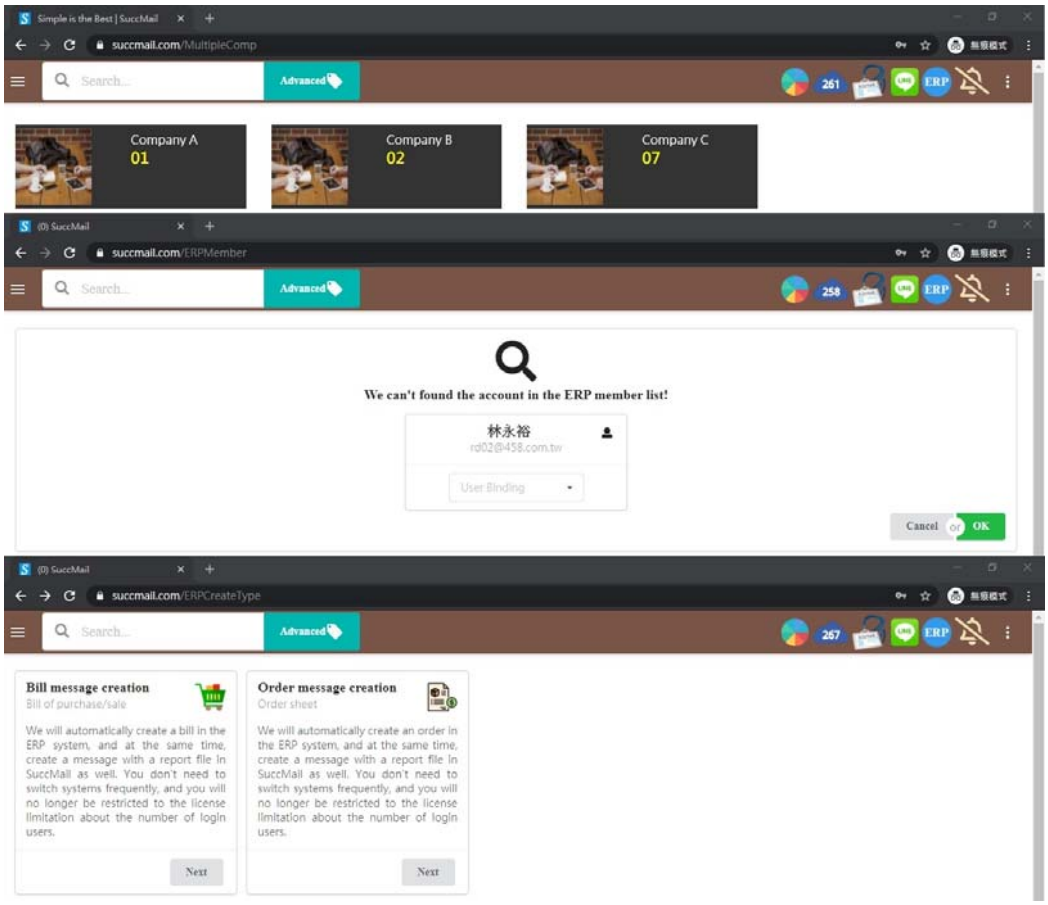


Figure 7.3 snapshots including company binding, user binding, and specific procedures selection.

As shown in Figure 7.5, the ERP system can correctly read the data from its database after the sheet data submit from ECS API to the ERP database. It confirms that the artifact SuccERP able to deliver the data to the ERP database correctly. Through SuccERP, the ECS owns the ability to provide a User Interface to carry out ERP procedures within SuccMail.

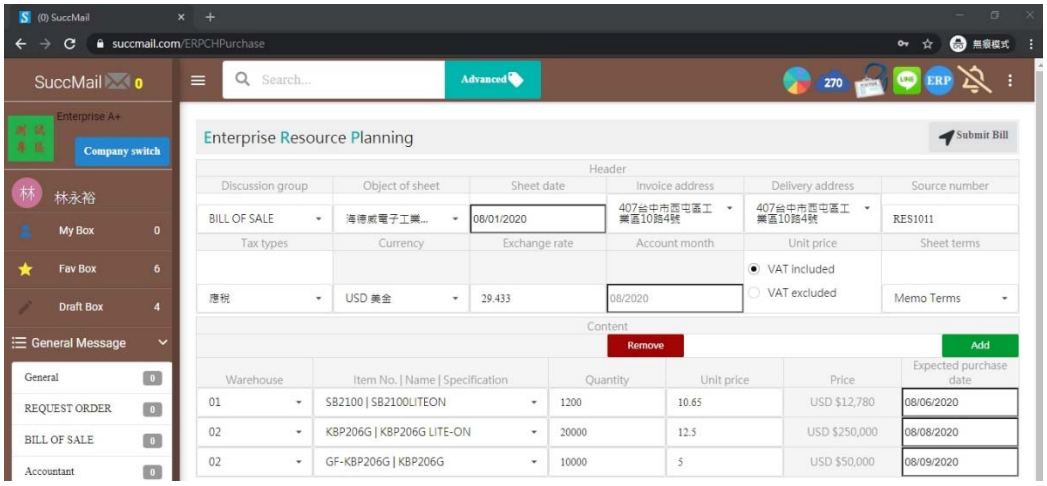


Figure 7.4 snapshot of the user interface for the bill message creation.

Figure 7.5 snapshot of created sheet in an ERP system.

Simultaneously, the related message will be created in the SuccMail, as described in Figure 7.6. On the left-hand side, it is the discussion list of the Enterprise A+, and the created message is on the right-hand side.

While the user clicks the submit button as illustrated in Figure 7.5, our artifact will complete the data-processing in both ERP and ECS, and the results present in Figure 7.5 and Figure 7.6. Moreover, as we illustrate an example of the right-bottom corner in Figure 6.6, Figure 7.6 is the practical case.

In this example, the message comprises a single subject, and two replies allow the employees and managers to do further discussion and communication after creating a sheet. The subject's content is equal to the corresponding sheet data in the ERP system. The members (employees or managers) can create the reply to do further communication on this sheet; as we can see, there are two replies in this example. The generated report data is right on the bottom of the subject in PDF format that can

preview online or download. Each subject and reply will show the read or unread status of all the receipts. The username with a red color shows the user has not read this subject or reply yet; on the contrary, the username with a black color shows the user read the subject or reply. The managers can grasp the status of message delivery based on the status of reading. Based on the integration between ECS and proposed artifact, each sheet expands as information flow in an enterprise instead of only stored in the database.

7.2 The Evaluation of proposed SuccERP

Next, we get into the evaluation part. To verify that the proposed artifact-SuccERP is workable and straightforward to design, we apply the metrics to discuss it further.

Table 7-1 the notation for the evaluation of SuccERP.

Notation	Descriptions
n_1	The number of distinct operators.
n_2	The number of distinct operands.
n	Program vocabulary.
N_1	The total number of operators.
N_2	The total number of operands.
N	Program length.
D	Halstead difficulty.
V	Halstead volume.
CC	Cyclomatic complexity.
CCD	Cyclomatic complexity density.
MI	Maintainability Index.
LOC	Lines of Code.

$$n = n_1 + n_2 \tag{29}$$

$$N = N_1 + N_2 \tag{30}$$

$$D = N \times \log_2 n \tag{31}$$

$$V = \frac{n_1}{2} + \frac{N_2}{n_2} \tag{32}$$

$$MI = MAX\big(0, \big(171 - 5.2 \times ln(V) - 0.23 \times CC - 16.2 \times ln(Lines\ of\ Code)\big) \times 100/171\big) \quad (33)$$

We apply both the Cyclomatic Complexity and the Maintainability Index to evaluate the performance of the proposed artifact. We consider Cyclomatic Complexity to count the number of independent paths through the source code. McCabe [134] recommends that developers consider the Cyclomatic Complexity of each module and divide module with the value higher than ten into the smaller modules. Furthermore, there is another research to suggest the value of Cyclomatic Complexity could accept as fifteen. However, it should provide a written explanation for exceeding limitations [135]. The Maintainability Index is an index value between 0 and 100. Welker and Oman suggest the Maintainability Index value between twenty and one hundred indicates the program has good maintainability. Between ten and nineteen indicates the program is moderately maintainable, and between zero and nine, that means low maintainability [136]. We divide the evaluation of SuccERP into two parts: the front end and the back end. We regard the back end to process the database operation, and the front end is the ViewModel in our case, which is used to process the user interface and the relative logic.

Table 7-2 the metrics of functions in the back end part of SuccERP.

Function	*MI*	*CC*	*LOC*
InitConnection()	43	8	37
GetWarehouseList()	44	8	35
GetTermsList()	44	8	35
GetItemsList()	36	13	55
GetItemsTypeList()	44	8	33
GetMemberList()	42	8	39
GetIdentifyCode()	44	9	34
GetMaxSheetCount()	44	9	34
GetInvoiceInfo()	41	8	43
GetDepartmentList()	44	8	35
GetObjectList()	36	8	62

GetCurrencyList()	41	8	43
GetCompDeailInfo()	33	5	73
GetAddressList()	40	8	43
CalculateTotal()	63	1	8
InsertOrderSheet()	29	9	98
InsertBillSheet()	29	9	97
UpdateUserBinding()	49	5	22
SynchronizeSheetInfo()	63	2	11
GenerateSheetReport()	47	4	22
GenerateReportHeader()	82	1	3
GenerateSheetPageContent()	54	6	13
GenerateSheetPageContext()	50	4	20

As shown in Table 7-2, the back-end part's evaluation shows all the functions have good maintainability by considering the MI metrics. However, both functions *InsertOrderSheet()* and *InsertBillSheet()*re lower than thirty on MI's value. The reason is that the function also includes inventory adjustment, log creation parts. We recommend the two functions for future research into smaller modules if there are further requirements. Then, about the Cyclomatic complexity, the function *GetItemsList()*is higher than ten but lower than fifteen; the primary reason is this function scrutinizes each item's value and process those invalid values, including null value invalid date-time string and extra space. Next, the evaluation of the front-end part, as shown in Table 7-3. It presents all the functions have good maintainability by considering the MI metrics, and the Cyclomatic complexity value is lower than ten. The result shows that the ECS is relatively straightforward to maintain and low complexity for expanding applications via the support from SuccERP.

Table 7-3 the metrics of functions in the front end part of SuccERP.

Function	*MI*	*CC*	*LOC*	*CCD*	*D*	*V*
onSheetHeaderView()	47.46	1	30	5.88%	7.57	762.25
handleDiscChange()	62.76	1	10	50%	2	152.29
handleObjectChange()	37.79	4	62	16.67%	19.72	1673.73

handleDateChange()	70.5	1	6	33.33%	2.85	58.81
handleAddressChange()	66.43	1	8	33.33%	2.8	91.37
handleTaxChange()	78.82	1	3	100%	1.16	33
onSettingItemTaxChange()	78.39	1	3	100%	1.14	38.03
onSheetContentView()	57.87	1	15	14.29%	5	215.22
onAddItem()	67.41	1	7	33.33%	5.35	100.37
onRemoveItem()	62.78	2	10	40%	7.5	144.67
handleWarehouseChange()	69.29	1	6	33.33%	2.8	87.56
onQuantityChange()	68.3	2	7	33.33%	2.16	71.69
onPriceChange()	58.57	3	16	75%	6	128
onGenerateECSMsg()	33.34	2	84	4.35%	10.4	3066.71

CHAPTER 8 Conclusion, Implication, Limitation and Future Work

Some research topics are progressively gaining importance as more and more employees work in complex and knowledge-intensive environments, such as Communication, Knowledge Management, and Knowledge Transfer. These topics may not be linked to each other at first glance; however, they are associated with the collaboration concept.

There is a noteworthy keyword – information flow, which covers Knowledge Management and Knowledge Transfer. Alavi and Leidner [137] repressed that the main challenge for Knowledge Management (KM) is to facilitate the information flow, based on the assumption that the information flow created by individuals is valuable and can improve performance. In order to achieve the maximum amount of knowledge transfer, individuals can share knowledge during decision making and also access the stored information flow to develop group knowledge if necessary. On the other hand, achieving knowledge transfer with organizations can be organized informally in email communications.

It is imperative to have a more appropriate and explicit definition of collaboration, IEEE's definition of collaboration also emphasizes the concept of information flow and knowledge sharing among systems. The diverse work environment leads to numerous collaboration challenges; in particular, two features of collaboration, unstructured collaboration (information collaboration) and structured collaboration (process collaboration). First, the definition of information collaboration is used to find answers to the unknown question by utilizing IT tools. An email is undoubtedly the most popular and general tool in information collaboration. Thus, the first direction we took is to explore the issues encountered in the email domain and improve them. Second, the definition of process collaboration is used to allowing business processes to be shared by sharing common information, structured written rules, and set workflows. More precisely, scholars have identified the integration between ERP and external resources is one of the most critical fields for strengthening the collaboration and business process. Thus, the second direction we took is implementing and building the integration between ERP and ECS systems. Such integration problems become more complex and

challenging, especially for Small and Medium-sized Enterprises (SMEs). When dealing with such integration requires a considerable investment of resources, including technology, finance resources, and workforce. On the technical level, when dealing with different ERP systems, integration needs to consider the database structure and system architecture, as well as the attributes from different data that need to be treated in a heterogeneous way.

Generally speaking, this dissertation has focused on several problems related to collaboration and communication, especially in the context of email and ERP systems. Firstly, in Chapter 1, we have introduced the definition and description of collaboration and identified two features: information collaboration and process collaboration. For information collaboration, we select email as the primary target by considering the following ingredients in order: discovering one of the features in collaboration: information collaboration, identifying the most popular and general approach to the information collaboration: IT tool (email), identifying the crucial issues in email management domain: email overload, lastly, exploring the potential solutions and research direction: topic identification able to enhance the communication and collaboration, and improve the email overload issue. On the other hand, for the process collaboration, we select the ERP system as the primary target by considering the following ingredients in order: Identify the orientation of business process improvement: conducting business process within the single and seamless system rather than two independent ones, Identify the concept an purpose for implementing system integration: Eliminating effectiveness communication, coordination and collaboration, Identify the most fundamental components of an enterprise before implementing system integration: ERP system.

Second, in Chapter 2, by conducting a review of email research, in which 65% of the research is concentrating on spam and fishing email detection and 20% of the research is focusing on email clustering and relative applications, where the remaining part is a minority of research considering cross-disciplinary and management related issues. Next, after acknowledging the current research trend, we found that whether in spam, phishing, and ham email literature, the topic model's application is showing an increasing trend. Further, a part of the works conducts the weighting function to assign varying weights to words to give more attention to essential words during topic inferences. At the same time, we also take into account the findings from cross-

disciplinary. The theories of cognitive science that each piece of information like an address in the form of (x, y) coordinates and the 'WHERE' may serve to recall the 'WHAT' while reading a document. The above theoretical grounds in Chapter 2 are what motivate us to conduct a more in-depth study, which is to propose the Attention orientation Latent Dirichlet Allocation (AttLDA) model.

Third, in Chapter 3, many email studies have used Enron Email Corpus data that have been processed into CSV format. However, what we need is as similar as possible to the format that people read. Thus, we designed a framework for processing the data, from downloading the compressed data to insert it into a MySQL database until exporting it to CSV format for the following analysis. Subsequently, we review the original LDA model and associated statistical distributions before presenting the proposed AttLDA model's graphical representation. We further present the data generation mechanism, i.e., the joint probability distribution of the hidden and observed variables. As for the estimation of visual attention, we refer to the previous literature to define the relevant parameters, including the Window size, Line length, and the pseudo-count definition. Lastly, we conduct the Collapsed-Gibbs-Sampling to carry out the inference process.

Fourth, in Chapter 4, we describe the preliminaries of the experiment, including the execution environment, the programming language, and how we split the dataset into training and test datasets. In the inference of the topic model, how to define the specific number of topics is a considerable issue. Therefore, we conduct the Coherence model to estimate the number of topics in our experiment. Lastly, we calculate the perplexity and information rate to measure the performance of proposed AttLDA model, and further interpret it based on the results of topic inference and the heat map of visual attention map.

Fifth, in Chapter 5, we introduced another feature of collaboration – process collaboration and identified the integration of ERP systems with external resources/systems could enhance the performance of the business process of an enterprise. We also consider the life cycle of the ERP system to get a concrete idea of when to implement the integration between ERP and external resources/systems. Finally, scholars suggest that the implementation process should be presented as a pseudo-code to provide a practical way for subsequent researchers and developers. After identifying the demand for building ERP system integration, we reviewed the

relevant literature. First, we focused on the ECS section to explain why it is the target to carry out ERP integration, including two primary advantages: it is an essential enabler of the modern digital workplace and a suitable longitudinal work tool. Meanwhile, several shortcomings, including ineffective content, are easily generated, making user's usage gradually inactive. The laissez-faire approach cannot get the full potential to project success. However, we have found that the ERP system's standardized content is a complementary component to the ECS. Afterward, we reviewed the literature related to ERP systems. It is worth mentioning that we have organized the findings into a keyword list, such as internal hosting, functional upgrade, enhancement, automating, informating, information sharing, and monitoring. Lastly, Design Science's belief is the relationship between theory and artifact that the truth (justified theory) and utility (practical artifacts) are two sides of the one coin. The scientific research should consider its practical implications for evaluation; hence, we employ the Design Science methodology to implement and build the proposed artifact-SuccERP.

Sixth, in Chapter 6, we follow the framework and methodology based on DS proposed by Hevner et al. [126] and Peffers et al. [127] to carry out our artifact – SuccERP. First, we conduct the user requirements definition and system requirements specification from the Software Engineering to achieve the "Problem identification and motivation" section. Then, in the "Define the objectives for a solution" section, we demonstrate the system architecture of SuccERP to make a further interpretation. Afterward, in the "Design and development" section, we present how to address the internal hosting issue by proposing two different system architectures, proceeding to collect the outline and schema from the ERP database. Lastly, the integration between the ERP system and ECS is shown by the pseudo-code.

Finally, in Chapter 7, the sequence diagram explains how SuccERP implements the integration between ERP system, ECS API service, SuccERP API service, SuccERP database, and Files server. Then, we show a series of snapshots of the practical operation of the artifact-SuccERP. In the end, we apply the Cyclomatic complexity and Maintainability Index to verify the proposed artifact is easy to maintain and develop. Also, we attempt to answer two fundamental questions proposed by [126] below:

(1) What utility does the new artifact provide?

First, artifact-SuccERP enables the architecture of both internal and external hosting. Meanwhile, allow the enterprise to expand the applications with legacy ERP systems via our artifact. The SuccERP keeps the consistency between the ERP system and the ECS, not only implement the enhancement of communication but also meticulously sustain all the data in the legacy ERP systems. In short, SuccERP can be regarded as a bridge to connect with the external systems. We proposed the complete architecture and implementation process in this research.

(2) What demonstrates that utility?

Exploring the CSFs and learning issues is still in the post-implementation stage; however, most of the results are academic. For the enterprise, they need more practical cases as a guideline. Hence, we present the complete steps and architecture to construct the artifact and present the snapshots and complexity metrics to show how it works and integrates the legacy ERP system and external systems.

Moreover, in our previous research works [138], we implemented and released the ECS system – SuccMail and employing the questionnaire survey to investigate how the ECS improves communication and e-management capabilities and performance improvement of enterprises in Taiwan, especially under the industry 4.0 trend. Likewise, after released the ECS system, we further discuss and explore the possibility and conception of building integration between the ECS and ERP systems [132]. We highlight the contributions as below:

(1) We believe that this is the first research to construct a post-implementation and show how to integrate the ERP system and ECS using a complete method.

(2) Instead of providing case studies, we present a complete process for investigating various ERP systems' structure. It will make our research results widely available to develop communication between different ERP systems and ECS.

(3) Truth and utility (practical artifacts) are two sides of the one coin. We list the weaknesses of both ECS and ERP systems from the literature review to ensure this research is based on the theoretical foundation. Also, we consider

the guidelines of DS to demonstrate how to develop and design the artifact completely.

8.1 Implication

In this dissertation, we focus on improving the performance of an enterprise's collaboration and communication, and we work on two features of the collaboration. To the best of our knowledge, the AttLDA model is the first to take visual attention into account when the topic inferences. Compared with previous studies, we provide the list of keywords related to each topic and the ratio of attention of the words under each topic. Meanwhile, in Chapter 3, we present how we define a document as a two-dimensional space with knowledge of numerous works of literature, and then further estimate the location of words and calculate the visual attention by the state-of-the-art visual attention model – Bayesian model. Besides, in the ERP and ECS integration section, we have collected numerous keywords and noteworthy issues in the post-implementation stage in Chapter 6. More importantly, we explore multiple ERP systems to integrate Chapter 7 to identify vital distinctions between different ERP systems. We believe that our results will help the managers and developers enhance their existing ERP systems and provide various ERP system architecture to the scholars for carrying out further research works. Furthermore, due to the COVID-19 pandemic, Byrnes et al. [139] emphasize that the global business urgently needs to conduct new technologies to enhance communication and collaboration whilst maintaining a physical distance.

8.2 Limitation

There are several inherent limitations of this dissertation as shown in the following:

(1) It is possible to receive emails consisting of multiple languages; however, our proposed AttLDA model is not currently implemented.

(2) In the visual attention model, we merely conduct the TFIDF for the visual feature part, while more visual features should be considered for future research evaluation.

(3) In ERP system integration, we focused on Collaboration and Communication issues, which also limited our integration target to the ECS system, and the

way to deal with this issue can be discussed from the exploratory stage towards other factors and develop system integration for numerous issues.

8.3 Future Work

As future work, two ideas can be considered. The first is to make AttLDA generate the probability of topics in each email, including the value of each word's attention, and summarize it to create feature vectors. Based on that, to make further approaches such as email classification, categorization, or thread identification. The second idea is to improve the experience of an Enterprise Collaboration System (ECS). While the ECS is increasingly choosing and follow by the entrepreneur or manager, more and more sharing information and conversation keep accumulating. It will be out of control soon if without considering how to manage it. We consider our proposed AttLDA as the solution for dealing with the management of messy messages. In other words, they provide the suggestion (topic) and give the comprehensible description (attention distribution) to provide the hint and direction for the users to manage and classify their message in an ECS.

Bibliography

[1] C. M. Fisher, J. Pillemer, and T. M. Amabile, “Deep Help in Complex Project Work: Guiding and Path-Clearing Across Difficult Terrain,” *Academy of Management Journal*, Aug. 2018, doi: 10.5465/amj.2016.0207.

[2] S. Prakash *et al.*, “Characteristic of enterprise collaboration system and its implementation issues in business management,” *International Journal of Business Intelligence and Data Mining*, vol. 16, no. 1, pp. 49–65, 2020.

[3] P. Schubert and J. H. Glitsch, “Use cases and collaboration scenarios: How employees use socially-enabled enterprise collaboration systems (ECS),” *International Journal of Information Systems and Project Management*, vol. 4, no. 2, pp. 41–62, 2016.

[4] N. Melville, K. Kraemer, and V. Gurbaxani, “Information technology and organizational performance: An integrative model of IT business value,” *MIS quarterly*, vol. 28, no. 2, pp. 283–322, 2004.

[5] T. H. Willis and A. H. Willis-Brown, “Extending the value of ERP,” *Industrial management & data systems*, 2002.

[6] T. H. Davenport, “Putting the enterprise into the enterprise system,” *Harvard business review*, vol. 76, no. 4, 1998.

[7] P. Kosalge and J. Motwani, “Understanding the subcultures key to ERP implementation: an empirical investigation,” *International Journal of Business Excellence*, vol. 1, no. 1–2, pp. 55–70, 2008.

[8] S. Radicati and J. Levenstein, “Email Statistics Report, 2015-2019,” *Radicati Group, Palo Alto, CA, USA, Tech. Rep*, 2015.

[9] S. Whittaker and C. Sidner, “Email overload: exploring personal information management of email,” in *Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI ’96*, Vancouver, British Columbia, Canada, 1996, pp. 276–283, doi: 10.1145/238386.238530.

[10] L. A. Dabbish and R. E. Kraut, “Email overload at work: an analysis of factors associated with email strain,” in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, New York, NY, USA, Nov. 2006, pp. 431–440, doi: 10.1145/1180875.1180941.

[11] D. Fisher, A. J. Brush, E. Gleave, and M. A. Smith, “Revisiting Whittaker & Sidner’s ‘email overload’ ten years later,” in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, New York, NY, USA, Nov. 2006, pp. 309–312, doi: 10.1145/1180875.1180922.

[12] A. C. M. Jerejian, C. Reid, and C. S. Rees, “The contribution of email volume, email management strategies and propensity to worry in predicting email stress among academics,” *Computers in Human Behavior*, vol. 29, no. 3, pp. 991–996, May 2013, doi: 10.1016/j.chb.2012.12.037.

[13] D. Skyrme, *Knowledge Networking: Creating the Collaborative Enterprise*. Routledge, 2007.

[14] G. A. Langenwalter, *Enterprise Resources Planning and Beyond: Integrating Your Entire Organization*. CRC Press, 2020.

[15] S. Zheng, D. C. Yen, and J. M. Tarn, “The New Spectrum of the Cross-Enterprise Solution: The Integration of Supply Chain Management and Enterprise Resources Planning Systems,” *Journal of Computer Information Systems*, vol. 41, no. 1, pp. 84–93, Sep. 2000, doi: 10.1080/08874417.2000.11646980.

[16] E. Samiei and J. Habibi, “The Mutual Relation Between Enterprise Resource Planning and Knowledge Management: A Review,” *Glob J Flex Syst Manag*, vol. 21, no. 1, pp. 53–66, Mar. 2020, doi: 10.1007/s40171-019-00229-2.
[17] J. M. Tarn, D. C. Yen, and M. Beaumont, “Exploring the rationales for ERP and SCM integration,” *Industrial Management & Data Systems*, 2002.
[18] M. K. McGee, “ERP Services Solution.,” *InformationWeek*, no. 712, pp. 161–163, 1998.
[19] E. B. Swanson and N. C. Ramiller, “Innovating mindfully with information technology,” *MIS quarterly*, pp. 553–583, 2004.
[20] A. M. Aladwani, “Change management strategies for successful ERP implementation,” *Business Process management journal*, 2001.
[21] H. Akkermans and K. van Helden, “Vicious and virtuous cycles in ERP implementation: a case study of interrelations between critical success factors,” *European journal of information systems*, vol. 11, no. 1, pp. 35–46, 2002.
[22] J. Esteves and J. Pastor, “Enterprise resource planning systems research: an annotated bibliography,” *Communications of the association for information systems*, vol. 7, no. 1, p. 8, 2001.
[23] K.-K. Hong and Y.-G. Kim, “The critical success factors for ERP implementation: an organizational fit perspective,” *Information & management*, vol. 40, no. 1, pp. 25–40, 2002.
[24] P. Ifinedo, G. Udo, and A. Ifinedo, “Organisational culture and IT resources impacts on ERP system success: an empirical investigation,” *International Journal of Business and Systems Research*, vol. 4, no. 2, pp. 131–148, 2010.
[25] C. Yoon, “The effects of organizational citizenship behaviors on ERP system success,” *Computers in Human Behavior*, vol. 25, no. 2, pp. 421–428, 2009.
[26] Z. Zhang, M. K. Lee, P. Huang, L. Zhang, and X. Huang, “A framework of ERP systems implementation success in China: An empirical study,” *International Journal of Production Economics*, vol. 98, no. 1, pp. 56–80, 2005.
[27] M. Zviran, N. Pliskin, and R. Levin, “Measuring user satisfaction and perceived usefulness in the ERP context,” *Journal of computer information systems*, vol. 45, no. 3, pp. 43–52, 2005.
[28] R. E. Giachetti, “A framework to review the information integration of the enterprise,” *International Journal of Production Research*, vol. 42, no. 6, pp. 1147–1166, 2004.
[29] G. C. Peng and M. B. Nunes, “Surfacing ERP exploitation risks through a risk ontology,” *Industrial Management & Data Systems*, 2009.
[30] S. Dehaene, J.-P. Changeux, L. Naccache, J. Sackur, and C. Sergent, “Conscious, preconscious, and subliminal processing: a testable taxonomy,” *Trends in cognitive sciences*, vol. 10, no. 5, pp. 204–211, 2006.
[31] S. Lancharoen, P. Suksawang, and T. Naenna, “Readiness assessment of information integration in a hospital using an analytic network process method for decision-making in a healthcare network,” *International Journal of Engineering Business Management*, vol. 12, p. 1847979019899318, 2020.
[32] A. Amid, M. Moalagh, and A. Z. Ravasan, “Identification and classification of ERP critical failure factors in Iranian industries,” *Information Systems*, vol. 37, no. 3, pp. 227–237, 2012.
[33] A. McCallum, X. Wang, and A. Corrada-Emmanuel, “Topic and role discovery in social networks with experiments on enron and academic email,” *Journal of Artificial Intelligence Research*, vol. 30, pp. 249–272, 2007.

[34] I. Alsmadi and I. Alhami, “Clustering and classification of email contents,” *Journal of King Saud University - Computer and Information Sciences*, vol. 27, no. 1, pp. 46–57, Jan. 2015, doi: 10.1016/j.jksuci.2014.03.014.
[35] M. Dehghani, A. Shakery, and M. S. Mirian, “Alecsa: Attentive Learning for Email Categorization using Structural Aspects,” *Knowledge-Based Systems*, vol. 98, pp. 44–54, Apr. 2016, doi: 10.1016/j.knosys.2015.12.013.
[36] A. Sharaff and N. K. Nagwani, “Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques,” *Journal of Information Science*, vol. 42, no. 2, pp. 200–212, 2016.
[37] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, “Email classification research trends: Review and open issues,” *IEEE Access*, vol. 5, pp. 9044–9064, 2017.
[38] A. L. White, G. M. Boynton, and J. D. Yeatman, “The link between reading ability and visual spatial attention across development,” *Cortex*, vol. 121, pp. 44–59, Dec. 2019, doi: 10.1016/j.cortex.2019.08.011.
[39] J. Tang, H. Li, Y. Cao, and Z. Tang, “Email data cleaning,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 489–498.
[40] S. Youn, “SPONGY (SPam ONtoloGY): Email classification using two-level dynamic ontology,” *The Scientific World Journal*, vol. 2014, 2014.
[41] S. Smadi, N. Aslam, and L. Zhang, “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning,” *Decision Support Systems*, vol. 107, pp. 88–102, Mar. 2018, doi: 10.1016/j.dss.2018.01.001.
[42] V. Ramanathan and H. Wechsler, “PhishGILLNET-phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training,” *EURASIP Journal on Information Security*, vol. 2012, Dec. 2012, doi: 10.1186/1687-417X-2012-1.
[43] “Porter Stemming Algorithm.” https://tartarus.org/martin/PorterStemmer/ (accessed Jan. 09, 2021).
[44] “WordNet | A Lexical Database for English.” https://wordnet.princeton.edu/ (accessed Jan. 09, 2021).
[45] J. R. Méndez, T. R. Cotos-Yañez, and D. Ruano-Ordás, “A new semantic-based feature selection method for spam filtering,” *Applied Soft Computing*, vol. 76, pp. 89–104, 2019.
[46] I. Alberts and D. Forest, “Email pragmatics and automatic classification: A study in the organizational context,” *Journal of the American Society for Information Science and Technology*, vol. 63, no. 5, pp. 904–922, 2012.
[47] A. Krzywicki and W. Wobcke, “Incremental e-mail classification and rule suggestion using simple term statistics,” in *Australasian Joint Conference on Artificial Intelligence*, 2009, pp. 250–259.
[48] K. Coussement and D. Van den Poel, “Improving customer complaint management by automatic email classification using linguistic style features as predictors,” *Decision Support Systems*, vol. 44, no. 4, pp. 870–882, 2008.
[49] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
[50] S. Liu, K. Lee, and I. Lee, “Document-level multi-topic sentiment classification of Email data with BiLSTM and data augmentation,” *Knowledge-Based Systems*, p. 105918, 2020.

[51] X. Li, A. Zhang, C. Li, J. Ouyang, and Y. Cai, “Exploring coherent topics by topic modeling with term weighting,” *Information Processing & Management*, vol. 54, no. 6, pp. 1345–1358, 2018.
[52] L. Pion-Tonachini, S. Makeig, and K. Kreutz-Delgado, “Crowd labeling latent Dirichlet allocation,” *Knowledge and information systems*, vol. 53, no. 3, pp. 749–765, 2017.
[53] A. Wilson and P. A. Chew, “Term weighting schemes for latent dirichlet allocation,” in *human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 465–473.
[54] E. A. Lovelace and S. D. Southall, “Memory for words in prose and their locations on the page,” *Memory & Cognition*, vol. 11, no. 5, pp. 429–434, 1983.
[55] T. J. Buschman and E. K. Miller, “Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices,” *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.
[56] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.
[57] “The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search - Keith Rayner, 2009.” https://journals.sagepub.com/doi/abs/10.1080/17470210902816461 (accessed Sep. 28, 2020).
[58] N. Sprague and D. Ballard, “Eye movements for reward maximization,” in *Advances in neural information processing systems*, 2004, pp. 1467–1474.
[59] N. Bruce and J. Tsotsos, “Saliency based on information maximization,” in *Advances in neural information processing systems*, 2006, pp. 155–162.
[60] L. Itti and P. F. Baldi, “Bayesian surprise attracts human attention,” in *Advances in neural information processing systems*, 2006, pp. 547–554.
[61] M. I. Posner, “Orienting of attention,” *Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.
[62] C. W. Eriksen and Y. Yeh, “Allocation of attention in the visual field.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 5, p. 583, 1985.
[63] J. E. Hoffman, B. Nelson, and M. R. Houck, “The role of attentional resources in automatic detection,” *Cognitive Psychology*, vol. 15, no. 3, pp. 379–410, 1983.
[64] U. Castiello and C. Umiltà, “Splitting focal attention.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, no. 3, p. 837, 1992.
[65] K. R. Cave, W. S. Bush, and T. G. Taylor, “Split attention as part of a flexible attentional system for complex scenes: comment on Jans, Peters, and De Weerd (2010).,” 2010.
[66] B. Klimt and Y. Yang, “The enron corpus: A new dataset for email classification research,” in *European Conference on Machine Learning*, 2004, pp. 217–226.
[67] R. Bekkerman, “Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora,” 2004.
[68] Y. Duan, J. Wang, M. Kam, and J. Canny, “A secure online algorithm for link analysis on weighted graph,” in *Proceedings of the Workshop on Link Analysis, Counterterrorism and Security, SIAM Data Mining Conference*, 2005, pp. 71–81.
[69] “The Enron Email Dataset.” https://kaggle.com/wcukierski/enron-email-dataset (accessed Jan. 14, 2021).

[70] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, 1972.
[71] G. Salton and C. T. Yu, “On the construction of effective vocabularies for information retrieval,” *Acm Sigplan Notices*, vol. 10, no. 1, pp. 48–60, 1973.
[72] A. Dillon, J. Richardson, and C. McKnight, “The effects of display size and text splitting on reading lengthy text from screen,” *Behaviour & Information Technology*, vol. 9, no. 3, pp. 215–227, 1990.
[73] M. C. Dyson and M. Haselgrove, “The influence of reading speed and line length on the effectiveness of reading from screen,” *International Journal of Human-Computer Studies*, vol. 54, no. 4, pp. 585–612, 2001.
[74] M. C. Dyson and G. J. Kipping, “Exploring the effect of layout on reading from screen,” in *International Conference on Raster Imaging and Digital Typography*, 1998, pp. 294–304.
[75] H. Spencer, *The visible word*. Hastings House Book Publishers, 1969.
[76] R. S. Grabinger and D. Amedeo, “CRT text layout: Perceptions of viewers,” *Computers in Human Behavior*, vol. 4, no. 3, pp. 189–205, 1988.
[77] B. S. Chaparro, M. Bernard, M. Fern, and S. Hull, “The Effects of Line Length on Children and Adults’ Online Reading Performance,” 2002.
[78] M. Youngman and L. Scharff, “Text width and margin width influences on readability of GUIs,” *Southwest Psychological Association*, 1998.
[79] J. Richardson, A. Dillon, and C. McKnight, “The effect of display size on reading and manipulating electronic text,” 1989.
[80] R. L. Duchnicky and P. A. Kolers, “Readability of text scrolled on visual display terminals as a function of window size,” *Human Factors*, vol. 25, no. 6, pp. 683–692, 1983.
[81] J. D. Gould and N. Grischkowsky, “Does visual angle of a line of characters affect reading speed?,” *Human Factors*, vol. 28, no. 2, pp. 165–173, 1986.
[82] H. Xiao and T. Stibor, “Efficient collapsed gibbs sampling for latent dirichlet allocation,” in *Proceedings of 2nd Asian Conference on Machine Learning*, 2010, pp. 63–78.
[83] T. P. Minka and J. Lafferty, “Expectation-propogation for the generative aspect model,” *arXiv preprint arXiv:1301.0588*, 2012.
[84] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
[85] “Research Center for Advanced Computing Infrastructure: Computing Servers.” https://www.jaist.ac.jp/iscenter/en/mpc/ (accessed Jan. 20, 2021).
[86] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.
[87] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, “Learning author-topic models from text corpora,” *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 1, pp. 1–38, 2010.
[88] A. K. McCallum, “Multi-label text classification with a mixture model trained by EM,” 1999.
[89] M. Carrasco, “Visual attention: The past 25 years,” *Vision research*, vol. 51, no. 13, pp. 1484–1525, 2011.
[90] S. J. Mraz, “Keeping Up With ERP.,” *Machine Design*, vol. 72, no. 14, pp. S56–S56, 2000.

[91] F. Carton, D. Sammon, and F. Adam, “Benefit realisation through ERP: the re-emergence of data warehousing,” *Electronic Journal of Information Systems Evaluation*, vol. 6, no. 2, pp. 155–164, 2003.
[92] M. Y. Yi and F. D. Davis, “Developing and validating an observational learning model of computer software training and skill acquisition,” *Information systems research*, vol. 14, no. 2, pp. 146–169, 2003.
[93] H.-W. Chou, H.-H. Chang, Y.-H. Lin, and S.-B. Chou, “Drivers and effects of post-implementation learning on ERP usage,” *Computers in Human Behavior*, vol. 35, pp. 267–277, 2014.
[94] J. Esteves and V. W. Bohórquez, “An updated ERP systems annotated bibliography: 2001-2005,” *Instituto de Empresa Business School Working Paper No. WP*, pp. 07–04, 2007.
[95] E. R. Mahendrawathi, S. O. Zayin, and F. J. Pamungkas, “ERP post implementation review with process mining: A case of procurement process,” *Procedia Computer Science*, vol. 124, pp. 216–223, 2017.
[96] P.-F. Hsu, H. R. Yen, and J.-C. Chung, “Assessing ERP post-implementation success at the individual level: Revisiting the role of service quality,” *Information & Management*, vol. 52, no. 8, pp. 925–942, 2015.
[97] S. V. Grabski, S. A. Leech, and P. J. Schmidt, “A review of ERP research: A future agenda for accounting information systems,” *Journal of information systems*, vol. 25, no. 1, pp. 37–78, 2011.
[98] Y. B. Moon, “Enterprise Resource Planning (ERP): a review of the literature,” *International journal of management and enterprise development*, vol. 4, no. 3, pp. 235–264, 2007.
[99] C. C. Law, C. C. Chen, and B. J. Wu, “Managing the full ERP life-cycle: Considerations of maintenance and support requirements and IT governance practice as integral elements of the formula for successful ERP adoption,” *Computers in Industry*, vol. 61, no. 3, pp. 297–308, 2010.
[100] M. Ali and L. Miller, “ERP system implementation in large enterprises–a systematic literature review,” *Journal of Enterprise Information Management*, 2017.
[101] T. C. McGinnis and Z. Huang, “Rethinking ERP success: A new perspective from knowledge management and continuous improvement,” *Information & Management*, vol. 44, no. 7, pp. 626–634, Oct. 2007, doi: 10.1016/j.im.2007.05.006.
[102] K. B. Osnes, J. R. Olsen, P. Vassilakopoulou, and E. Hustad, “ERP systems in multinational enterprises: A literature review of post-implementation challenges,” *Procedia computer science*, vol. 138, pp. 541–548, 2018.
[103] N. Hasan, S. J. Miah, Y. Bao, and M. R. Hoque, “Factors affecting post-implementation success of enterprise resource planning systems: a perspective of business process performance,” *Enterprise Information Systems*, vol. 13, no. 9, pp. 1217–1244, 2019.
[104] C.-S. Yu, “Causes influencing the effectiveness of the post-implementation ERP system,” *Industrial Management & Data Systems*, 2005.
[105] J. Ram, D. Corkindale, and M.-L. Wu, “Implementation critical success factors (CSFs) for ERP: Do they contribute to implementation success and post-implementation performance?,” *International Journal of Production Economics*, vol. 144, no. 1, pp. 157–174, 2013.

[106] O. B. Kwon and J. J. Lee, “A multi-agent intelligent system for efficient ERP maintenance,” *Expert Systems with Applications*, vol. 21, no. 4, pp. 191–202, 2001.
[107] S. P. Williams, “Enterprise 2.0 and collaborative technologies,” *Koblenz: Working Report of the Research Group Business Software, University of Koblenz-Landau*, 2011.
[108] C. S. Greeven and S. P. Williams, “Enterprise collaboration systems: addressing adoption challenges and the shaping of sociotechnical systems,” *International Journal of Information Systems and Project Management*, vol. 5, no. 1, pp. 5–23, 2017.
[109] C. S. Greeven and S. P. Williams, “Enterprise collaboration systems: An analysis and classification of adoption challenges,” *Procedia Computer Science*, vol. 100, no. 100, pp. 179–187, 2016.
[110] R. Diehl, T. Kuettner, and P. Schubert, “Introduction of enterprise collaboration systems: In-depth studies show that laissez-faire does not work,” 2013.
[111] F. Schwade and P. Schubert, “Social Collaboration Analytics for Enterprise Collaboration Systems: Providing Business Intelligence on Collaboration Activities,” 2017.
[112] C. Herzog, A. Richter, M. Steinhüser, U. Hoppe, and M. Koch, “Methods and metrics for measuring the success of enterprise social software-what we can learn from practice and vice versa,” 2013.
[113] Y. M. Ha and H. J. Ahn, “Factors affecting the performance of Enterprise Resource Planning (ERP) systems in the post-implementation stage,” *Behaviour & Information Technology*, vol. 33, no. 10, pp. 1065–1081, 2014.
[114] G. C. Peng and M. B. Nunes, “Identification and assessment of risks associated with ERP post-implementation in China,” *Journal of Enterprise Information Management*, 2009.
[115] “The 2020 ERP Report [Panorama’s Independent Analysis],” *Panorama Consulting Group*, Feb. 11, 2020. https://www.panorama-consulting.com/resource-center/2020-erp-report/ (accessed Sep. 28, 2020).
[116] T. Oseni, S. V. Foster, R. Mahbubur, and S. P. Smith, “A framework for ERP post-implementation amendments: A literature analysis,” *Australasian Journal of Information Systems*, vol. 21, 2017.
[117] H.-W. Chou, Y.-H. Lin, H.-S. Lu, H.-H. Chang, and S.-B. Chou, “Knowledge sharing and ERP system usage in post-implementation stage,” *Computers in Human Behavior*, vol. 33, pp. 16–22, 2014.
[118] J. Hisnanick, “In the age of the smart machine: The future of work and power,” *Employ Respons Rights J*, vol. 2, no. 4, pp. 313–314, Dec. 1989, doi: 10.1007/BF01423360.
[119] O. Lorenzo, “Human, Contextual, and Processual Issues Influencing Enterprise System Use,” p. 5.
[120] A. Boza, L. Cuenca, R. Poler, and Z. Michaelides, “The interoperability force in the ERP field,” *Enterprise Information Systems*, vol. 9, no. 3, pp. 257–278, 2015.
[121] M. Al-Mashari, “Constructs of Process Change Management in ERP Context: A Focus on SAP R/3,” p. 5.
[122] J. W. Ross and M. R. Vitale, “The ERP Revolution: Surviving vs. Thriving,” *Information Systems Frontiers*, vol. 2, no. 2, pp. 233–241, Aug. 2000, doi: 10.1023/A:1026500224101.

[123] M. Aboulafia, *Philosophy, social theory, and the thought of George Herbert Mead*. SUNY Press, 1991.
[124] A. Lee, "Systems Thinking, Design Science, and Paradigms: Heeding Three Lessons from the Past to Resolve Three Dilemmas in the Present to Direct a Trajectory for Future Research in the Information Systems Field,"Keynote Address," 2000.
[125] H. A. Simon, *The sciences of the artificial*. MIT press, 2019.
[126] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly*, pp. 75–105, 2004.
[127] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.
[128] Q. Liu and G. Liu, "Research on the framework of decision support system based on ERP systems," in *2010 Second International Workshop on Education Technology and Computer Science*, 2010, vol. 1, pp. 704–707.
[129] T. Oseni, M. M. Rahim, S. Smith, and S. Foster, "Exploring ERP post-implementation modifications and their influence on business process outcomes: A theory driven model," 2013.
[130] F. C. Weston Jr, "ERP implementation and project management," *Production and Inventory Management Journal*, vol. 42, no. 3/4, p. 75, 2001.
[131] A. Elragal and H. E.-D. Hassanien, "Augmenting advanced analytics into enterprise systems: A focus on post-implementation activities," *Systems*, vol. 7, no. 2, p. 31, 2019.
[132] Y. Lin, Y. Nagai, T. Chiang, and H. Chiang, "Design and Develop Artifact for Integrating with ERP and ECS Based on Design Science," in *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, 2020, pp. 218–223.
[133] D. MacKenzie and J. Wajcman, *The social shaping of technology*. Open university press, 1999.
[134] T. J. McCabe, "A complexity measure," *IEEE Transactions on software Engineering*, no. 4, pp. 308–320, 1976.
[135] A. H. Watson, D. R. Wallace, and T. J. McCabe, *Structured testing: A testing methodology using the cyclomatic complexity metric*, vol. 500. US Department of Commerce, Technology Administration, National Institute of …, 1996.
[136] K. D. Welker, P. W. Oman, and G. G. Atkinson, "Development and application of an automated source code maintainability index," *Journal of Software Maintenance: Research and Practice*, vol. 9, no. 3, pp. 127–159, 1997.
[137] M. Alavi and D. E. Leidner, "Knowledge management and knowledge management systems: Conceptual foundations and research issues," *MIS quarterly*, pp. 107–136, 2001.
[138] H.-K. Chiang, Y. Nagai, and Y.-Y. Lin, "Link up Industry 4.0 with the Enterprise Collaboration System to Help Small and Medium Enterprises," *Mathematical Problems in Engineering*, vol. 2020, 2020.
[139] K. G. Byrnes, P. A. Kiely, C. P. Dunne, K. W. McDermott, and J. C. Coffey, "Communication, collaboration and contagion: 'Virtualisation' of anatomy during COVID-19," *Clinical Anatomy*, vol. 34, no. 1, pp. 82–89, 2021, doi: https://doi.org/10.1002/ca.23649.