ENHANCING PUBLIC SECTOR ORGANIZATION KNOWLEDGE RETENTION WITH SOCIAL NETWORK ANALYSIS, TEXT MINING AND MACHINE LEARNING

A Dissertation
Presented to
The Academic Faculty

By

Yuzhi Guo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Computational Science and Engineering

Georgia Institute of Technology
August 2020

ENHANCING PUBLIC SECTOR ORGANIZATION KNOWLEDGE RETENTION WITH SOCIAL NETWORK ANALYSIS, TEXT MINING AND MACHINE LEARNING

Approved by:

Dr. David Frost, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Tuo Zhao
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Umit Catalyurek
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Wei Deng
Google Cloud
Alphabet Inc.

Dr. Polo Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Date Approved: [May 07, 2020]

[To My Lovely Family]

ACKNOWLEDGEMENTS

First of all, I would like to acknowledge my advisor, Dr. David Frost, for giving me the opportunity to work on the impactful research, for providing me invaluable guidance and continuous support during my PhD study. His wisdom, vision and patience have greatly inspired me not only for my research work but also for my life henceforward. I am extremely grateful for what he has offered me.

I would also like to express my sincere gratitude to my committee members: Dr. Umit Catalyurek, Dr. Polo Chau, Dr. Tuo Zhao, and Dr. Wei Deng, for their insightful comments and constructive suggestions to my work, allowing for great improvements to my research accomplishments.

I thank all my colleagues in the Sustainable Geotechnical System Laboratory and all my friends at Georgia Tech, for the stimulating discussions and their encouragements. I would like to express my special thanks to Dr. Mahdi Roozbahani and Dr. Fikret Atalay, who have contributed a lot to making my research work more successful and impactful.

A number of individuals provided valuable assistance during the course of this study. They included individuals from various GDOT units who provided critical insights into current practices and experiences, and participated in valuable project-related discussions. In particular, the contributions of Monica Ivey, Brian Robinson, and Supriya Kamatkar throughout the study period are sincerely appreciated. A number of other GDOT employees from various offices also contributed to the work through participation in informal and formal discussions. Their engaged participation is recognized.

Last but not the least, I’m very grateful to my family members. I would like to thank my parents Zhuquan Guo and Xiurong Zhao for their unconditional love and support, my wife Lin Xiao for her ongoing understanding and encouragement, and my son Alan Guo for all the happiness he has brought to my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The technical knowledge and expertise possessed by employees are considered amongst an organization's greatest assets, but are also most vulnerable and can be easily impacted or lost. The loss of experienced employees and important knowledge can put an organization's competency in great jeopardy. Thus, it is critical to address the challenge of proper knowledge transfer and retention proactively rather than reactively. Public sector organizations have their unique characteristics and are facing emerging HR challenges due to market changes. Most of the current knowledge retention approaches are either outdated and ineffective or developed without considering the features of public sector organizations. A study that overlaps computational and data science techniques with HR data management in light of these features is considered to be a strategic and systematic development that advances existing methods in knowledge retention and overcomes the emerging HR challenges faced by many large public organizations.

In the scope of this work, several data tools are studied for their applications to HR databases, with the objectives of enhancing perception on organization-wide attrition risk distribution, identifying critical knowledge at risk of being lost, and choosing the most suitable provider and recipient for a set of knowledge sharing programs. Moreover, an integrated computational system is developed for Georgia DOT. The system uses an existing HR database and provides modular tools to assist HR personnel strategically plan for a range of activities, aiming for increased level of knowledge transfer and lower employee turnover rate, among other benefits. The system is further evaluated by both user experience feedback, as well as a few "use cases" discussed with the end users.

The main contributions of this thesis are: 1) This thesis proposes an unprecedented way to transform the classic organizational chart into a more informative network style chart. Several network metrics are developed to inventively describe knowledge management-related attributes of employees; 2) This work develops an innovative approach for systematically and quantitatively evaluating a range of knowledge retention metrics. An inventive workflow is also designed and implemented for conducting various knowledge transfer activities; and 3) A novel computational system, which integrates both existing data tools and the newly developed approach and framework, is designed, developed and deployed at Georgia DOT to assist human resource data management and to enhance organization knowledge retention.

CHAPTER 1. INTRODUCTION

1.1 Overview and Objectives

Knowledge and experience within organizations are both a major investment and a valuable resource. The technical knowledge and expertise possessed by employees are considered amongst an organization's greatest assets. However, these are also some of the most vulnerable assets that can be easily impacted or lost. Employee departures, whether due to retirement attrition or career opportunities elsewhere, cause not only the loss of critical personnel but also valuable knowledge, and are of particular concern for many organizations. Especially for public sector organizations, this issue is receiving more and more attention due to their organizational characteristics and the change of market. Moreover, the loss of knowledge and experience are often associated with a decrease in the number of employees, leading to a situation where existing employees are experiencing a significant increase in responsibility with minimal preparation. Such a scenario can surely place an organization's ability to effectively accomplish its initially established mission statement in jeopardy, and if this adverse effect influences public sector organizations, it can negatively impact people's life. Therefore, there is this need for this challenge of proper knowledge transfer and retention to be addressed proactively rather than reactively.

Employee databases, essentially existing in every organization and usually being managed by a human resource (HR) department, often contain a lot of useful information. However, such rich information is often hidden in the data without rigorous analysis due to budgetary, technical, or other reasons. With the emergence of computational science and engineering fields, data science techniques like data visual analytics (DVA), social network analysis (SNA), text mining and machine learning can be applied to employee databases to reveal constructive information lying behind the data for an organization to take advantage of when making important decisions. This research work amasses a set of studies and designs into an automated computational system with the aims of: (1) quantitatively identifying the positions that are critical not only strategically but also tactically to the organization and the ones with high risk of knowledge and experience loss; and (2) assisting an organization's human resource department with decision making about planning, training, and other HR activities, consequently achieving a higher level of knowledge retention.

1.2 Background and Significance

In today's economic environment, intellectual assets, such as knowledge and experience, are among the key elements that make an organization sustainably competitive among others (Doan, Rosenthal-Sabroux, & Grundstein, 2011). Organizations with dedicated programs supporting knowledge retention tend to maintain competitiveness, even during employee downsizing (Schmitt, Borzillo, & Probst, 2012). On the other hand, without proper management in knowledge retention, organizations are extremely vulnerable to knowledge loss risk, leading to potential workforce weakening (Liebowitz, 2008). According to a knowledge-based theory of the firm, one of the primary missions of the firm is integrating employees' specialist knowledge into merchandise and services (Grant, 1996). Being a crucial factor in organization development, knowledge retention is not receiving as much attention as its importance level (Marsh & Stock, 2006), and very few organizations are devoting enough effort into implementing knowledge retention strategies (Liebowitz, 2008). Furthermore, this situation is especially severe among public sector organizations than private ones. More importantly, due to the complex nature of knowledge resident in individuals rather than existing independently, the proper management, transfer and retention tasks are much more complex and cumbersome to perform (Grant, 1996; Zhang, 2017).

Recent research on organization knowledge management has been conducted from multiple perspectives. The concept of organizational memory, consummated by Walsh and Ungson (Walsh & Ungson, 1991), classified knowledge that can be managed and retained by organizations into five categories: individual, culture, transformation, structures and ecology. We are mostly interested in technical knowledge possessed by individual employees, as it is directly acquired externally and critically influences organization development. Through a multi-case research study, Levy (Levy, 2011) mentioned that successful knowledge retention can be achieved in three main stages: defining scope, documenting knowledge, integrating it back into the organization. Doan (Doan et al., 2011) introduced a reference model for knowledge retention specifically tailored for small and medium size enterprises. Schmitt (Schmitt et al., 2012) proposed knowledge retention considerations especially during strategic downsizing, when organizations are subject to the most severe personnel and knowledge loss. While knowledge management and retention has been studied over a few decades, the following items summarize some challenges that still need to be addressed:

- Successful knowledge transfer and technical experience retention cannot be done in an instant, so HR activities with such purposes should be planned proactively rather than reactively. In this regard, identifying critical positions and personnel with relatively high risk of knowledge loss is just as important as conducting knowledge transfer activities. Although knowledge loss risk can be assessed in a more comprehensive way, the lack of an interactive visualization tool, such as pivot charts / tables, impede HR managers efficiently obtaining an intuitive overview of critical positions within the organization with respect to attrition risk.

- When recognizing position importance, managers without domain knowledge tend to assess solely based on the position level. However, such an evaluation should take into consideration not only strategic planning but also tactical planning. Without looking at how information flows among employees within an organization, position criticality and personnel knowledge loss risk cannot be comprehensively evaluated. Jennex (Jennex & Durcikova, 2013) proposed a methodology to assess knowledge loss risk considering a number of metrics including age, skill uniqueness, years of services, etc., yet the information network of an employee among the organization is not considered.
- Once an important position or a critical person is identified and is subject to knowledge transfer activities, the traditional way to choose a candidate is directly from those who are just one or two levels below the one identified with potentially high knowledge loss risk. However, there are scenarios where the best candidate may come from other teams or even from another department. Therefore, there is this key knowledge gap on selecting the best candidate for knowledge retention across the whole organization.
- Many research and development work studies on data management and knowledge retention / sharing are conducted by those with concentration on some specific types of industries (e.g. high tech companies, consulting firms, etc.), and typically focus

on private organizations. However, less studies are done with concentration on public sector organizations, taking into consideration their unique characteristics. This is usually due to factors like limitation on technology available in house, as well as budgetary concerns from many regulations.

1.3 Intellectual Merits

To advance the practical results of knowledge retention in public sector organizations, the proposed work will allow for: (1) applying data science technologies, such as data visualization, social network analysis, text mining and machine learning, on an organization's human resource database, which extracts insightful information such as the geographical distribution of organization-wide attrition risk; (2) using network analysis and statistical methods (multivariate analysis) to quantitatively evaluate employee knowledge loss risk and position importance / criticality; and (3) developing an integrated computational system that uses a real world employee database, which can practically assist HR personnel with efficient and comprehensive decision making, allowing for preventative measures to be taken towards the goal of knowledge retention.

1.4 Broader Impacts

In addition to facilitating the optimal knowledge and experience retention in organizations and their related benefits, the proposed work can reveal even more insightful information from a data-rich HR database. For example, through coupling network analysis and visualization, a traditional org-chart can be easily constructed, a network view with meaningful node size can also display metrics, such as criticality, uniqueness, etc., with a concentration on information flow. Moreover, the application, assuming successfully developed, can serve as a robust and powerful motivation tool for employees to keep them engaged and excited about their work environment and career path opportunities, resulting in a high-resilient organization in the situations like staff crisis and employee downsizing.

1.5 Thesis Overview and Organization

The research plan is envisaged with a twofold theme: (I) discussing current knowledge management approaches being widely applied, proposing a new human resource data and knowledge management framework, as well as addressing the fundamental techniques to be applied to tasks like identifying critical positions and selecting best candidates for knowledge retention activities; and (II) designing, developing and evaluating an automated computational system with a set of working modules that could practically assist an organization's human resource department with decision making related to hiring, job shadowing, workforce planning, succession training, and other activities, to achieve increased knowledge transfer rate, higher levels of employee knowledge and motivation, among other potential benefit.

The thesis consists of seven chapters. Chapter 1 introduces the overall work and Chapter 7 summarizes the conclusions and any recommendations for future work.

- Chapter 2 studies the current human resource data and knowledge management approaches widely used across different types of industries.
- Chapter 3 proposes at a level a new HR data management framework especially tailored for public sector organizations including a discussion about their characteristics.

- Chapter 4 summarizes the fundamental techniques and data tools applied in the proposed framework, including spatial analysis and visualization, social network analysis, text mining, multivariate analysis and temporal analysis.
- Chapter 5 presents an automated and scalable integrated HR data management system for Georgia DOT, which is built upon the techniques and tools that were studied in Chapter 4.
- Chapter 6 summarizes user feedback from the end users of the working system developed in Chapter 5, and evaluates the system's effectiveness and capability.

CHAPTER 2. CURRENT HR DATA MANAGEMENT APPROACHES

This chapter summarizes employee turnover rate in industry sectors, and studies current HR data and knowledge management approaches that are being implemented in large organizations, with the objective of understanding how organizations in different industry sectors manage their intellectual assets for optimal knowledge retention.

2.1 Introduction

Knowledge loss, having been a significant and expensive issue for decades faced by most organizations, can come from multiple sources such as employee departures, staff resistant to learn, outsourcing, etc. (Levallet & Chan, 2019). Attrition, no matter whether due to retirements, transfers, or departures, is undoubtedly causing the most severe loss of experienced employees, and moreover, valuable knowledge. Employee turnover rate, defined as the percentage of employees in an organization’s workforce that leave during a certain period of time and are replaced by new employees, is an important statistic for an organization as it not only determines how often an organization should hire new employees, but also can serve as a reasonable indicator of an organization’s overall knowledge loss risk (KLR). However, for different industries, average turnover rates can be vastly different. To better support the general objectives of this work, it is appropriate to study different industries’ turnover rates, as well as typical approaches for knowledge retention while dealing with employee turnover.

Turnover rate, on the other hand can be calculated from the employees' average tenure with an organization. A short tenure typically indicates high turnover rate where the organization needs to hire new employees more often, and potentially suffer higher KLR. In the same manner as turnover rate, employee tenure is varied by industry to a high degree. For example, according to the 2018 report from Bureau of Labor Statistics, the overall median years of tenure is 4.2 years. Workers in the leisure and hospitality, including food service, industry have the lowest median tenure of 2.2 years. Utility workers have the highest median tenure of 9.5 years. Government employees, including federal, state, and local government, have relatively high median tenures of 6.8 years. Table 1 shows the median years of tenure by industry from 2012 to 2018, cited partially from BLS news release. (Source: US department of Labor, Bureau of Labor Statistics).

Table 1 – Median years of tenure by industry, 2012 - 2018.

Industry	2012	2014	2016	2018
Total, 16 years and over	4.6	4.6	4.2	4.2
Private sector	4.2	4.1	3.7	3.8
Agriculture and related industries	4.1	3.6	4.5	4.6
Nonagricultural industries	4.2	4.1	3.7	3.8
Mining, quarrying, and oil and gas extraction	3.5	4.0	4.6	5.1
Construction	4.3	3.9	4.0	4.1
Manufacturing	6.0	5.9	5.3	5.0
Durable goods manufacturing	6.1	6.0	5.4	5.3
Nonmetallic mineral products	7.0	7.6	5.1	5.2
Primary metals and fabricated metal products	5.6	6.1	6.0	6.0
Computers and electronic products	7.7	5.1	5.3	5.8
Electrical equipment and appliances	5.9	5.8	4.7	4.5
Transportation equipment	7.1	7.1	6.1	5.7
Wood products	5.3	4.6	4.7	3.5
Furniture and related product manufacturing	6.5	5.9	4.8	4.8
Miscellaneous manufacturing	4.8	5.1	5.0	4.8
Nondurable goods manufacturing	5.8	5.9	5.1	4.7
Food manufacturing	4.9	4.7	4.5	3.9
Beverages and tobacco products	6.4	4.8	4.3	4.1
Paper and printing	9.7	9.7	5.3	5.4

Table 1 continued.

Industry	2012	2014	2016	2018
Petroleum and coal products	6.4	6.1	6.6	5.0
Chemicals	6.1	7.1	5.3	4.7
Plastics and rubber products	6.1	6.5	5.3	5.0
Wholesale and retail trade	3.7	3.6	3.3	3.2
Wholesale trade	5.5	5.8	5.2	5.1
Retail trade	3.3	3.3	3.0	3.0
Transportation and utilities	5.6	5.1	4.6	4.8
Transportation and warehousing	5.3	4.7	4.4	4.2
Utilities	9.5	9.2	7.4	9.5
Information	5.4	4.8	4.3	4.4
Publishing, except Internet	6.6	5.3	5.7	4.1
Motion pictures and sound recording industries	2.6	2.4	2.4	2.9
Radio and TV broadcasting and cable subscriptions programming	4.9	4.1	3.6	5.0
Telecommunications	7.4	7.8	6.0	5.2
Financial activities	4.9	5.0	4.8	4.7
Finance and insurance	5.0	5.3	5.0	5.0
Finance	4.7	5.0	5.0	4.8
Insurance	5.7	6.0	5.2	5.4
Real estate and rental and leasing	4.5	4.4	3.8	3.6
Real estate	4.5	4.6	3.9	3.7
Rental and leasing services	4.2	3.5	3.4	3.4
Professional and business services	3.8	3.6	3.4	3.6
Professional and technical services	4.4	4.2	3.9	3.9
Management, administrative, and waste services	3.1	3.1	2.8	3.3
Administrative and support services	3.0	3.0	2.6	3.1
Waste management and remediation services	4.4	4.7	4.6	5.8
Education and health services	4.4	4.5	3.9	3.9
Educational services	4.3	4.8	4.0	4.2
Health care and social assistance	4.4	4.4	3.9	3.9
Hospitals	6.0	5.7	5.6	4.9
Health services, except hospitals	3.8	3.9	3.4	3.5
Social assistance	3.1	3.2	2.6	3.0
Leisure and hospitality	2.4	2.3	2.2	2.2
Arts, entertainment, and recreation	3.1	3.0	3.2	3.0
Accommodation and food services	2.3	2.1	2.0	2.1
Accommodation	3.8	3.5	3.0	3.1
Food services and drinking places	2.1	2.0	1.8	2.0
Other services	3.8	4.0	3.9	4.0
Other services, except private households	3.8	4.2	4.1	3.9
Repair and maintenance	3.7	4.0	3.5	3.3
Personal and laundry services	3.5	3.7	3.8	3.6
Membership associations and organizations	4.3	4.9	4.9	4.5
Other services, private households	3.3	3.0	3.3	4.5

Table 1 continued.

Industry	2012	2014	2016	2018
Public sector	7.8	7.8	7.7	6.8
Federal government	9.5	8.5	8.8	8.3
State government	6.4	7.4	5.8	5.9
Local government	8.1	7.9	8.3	6.9

While there are not statistics for tech companies' average tenure data in this report, other credible sources like PayScale, Paysa, etc., have comprehensive statistics of employee tenure in tech companies. For example, Google has a tenure of 1.9 years, Apple with 1.85 years, Amazon with 1.84 years, and Uber with only 1.23 years, all of which are public tech companies with an IPO over 10 years ago and a current valuation of over $100 billion. Overall, the IT industry has a very low average tenure of less than 2 years (Source: Paysa). Without a doubt, high attrition can have great impact on an organization, increasing its KLR to a high degree.

2.2 Organization Classification based on KLR level

Short employee tenure brings high KLR to an organization. In turn, high KLR brings not only increased burden on workforce planning, but also elevated risk of losing critical knowledge, as well important social network connections for specific industries. This is especially of concern for large organizations where unstructured knowledge are more prone to be lost. Therefore, it is critical that organizations implement proper and targeted knowledge retention protocols to mitigate the KLR, and is just as important as their investment in research and development. In this section, organizations will be divided into three categories for the purpose of studying current knowledge retention approaches by different types of organization according to their KLR levels. Specifically, high-KLR organizations with short term average employee tenure (less than 3 years); medium-KLR organizations with medium term average employee tenure (3 to 4 years); and low-KLR organizations with long term average employee tenure (more than 5 years). Figure 1 shows the schematic correlation between the KLR and employee tenure in an organization at a high level.

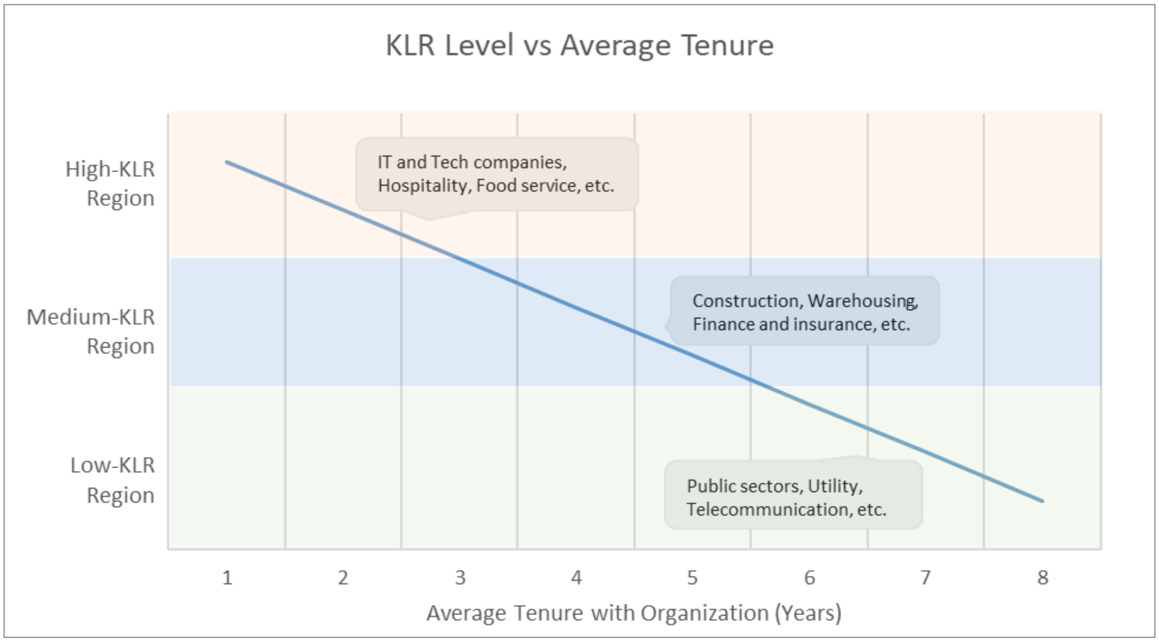


Figure 2.1 – Correlation between KLR and employee tenure. Organizations are divided into three groups: high-KLR organizations; medium-KLR organizations, and low-KLR organizations.

2.3 Current Knowledge Retention Approaches

As mentioned above, it is important that organizations with different levels of KLR and / or with different domains should develop proper and targeted knowledge retention approaches respectively. This section has three subsections, each of which will cover a category of the industries based on the KLR level. In each subsection, applicable knowledge retention approaches implemented by some typical industries of the category will be studied.

2.3.1 Approaches for high-KLR organizations

High-KLR organizations feature the shortest employee tenure (less than 3 years) and highest turnover rate. Accommodation, food services, leisure, hospitality, retail, and media industries, as well tech companies usually fall into this category. The relatively short tenures in the industries of services, retail and media industries are due to seasonal and low-level natures or project based reasons, which is quite different from the case of tech companies. The fact that employees in tech companies tend to have the lowest tenure among industries normally comes from multiple facets: Concerns about the career opportunities; Dissatisfaction with the leadership; Dissatisfaction with the working environment or culture; Dissatisfaction with the compensation, benefits, etc. (Source: LinkedIn). Above all, the high turnover rate of the roles in tech companies including software engineers, data scientists, and UX designers is being driven by the extremely high demand, differentiating such organizations from other industries with high-KLR.

More importantly, potential knowledge loss tends to bring the most severe impact to tech companies, not only because the intellectual assets are always the most valuable ones in such organizations, but also because it is very expensive to train tech professionals. Thus, the retention of knowledge and experience in tech companies deserves more attention and study. As a matter of fact, knowing that employee retention might not be feasible, such organizations as Google, Facebook, Linked, Apple, etc. have spent a lot on retaining the intellectual assets, and have developed sophisticated knowledge management systems. Therefore, tech companies will be used in this section as a meaningful example for high-KLR organizations to study their applicable knowledge retention approaches.

Under the reality that large numbers of employees are departing every day across the whole organization, tech companies must integrate knowledge retention protocols into everyday work. One of the most important characteristic is standardization. Big tech companies not only keep the coding style, documentation format, and readability criteria standardized, but also unify all the training classes, code review processes, and product releasing procedures. Detailed approaches include comprehensive and unified onboard training, plentiful and same personal growth curriculums, standardized coding library and style, etc. With these approaches, and taking Google as an example, all the codes (in the same programming language) across thousands of terabyte of code base appear similar, and every design documentation has similar format. Thus, every employee can learn the "knowledge" from every other employee through reading these standardized documents. In this way, the impact of employee departure, even including the interns, is minimized.

Another important approach, which is very effective in knowledge retention for not only tech companies but also other industries, is materialization. Materialization tries to record learned knowledge into documentation. In tech companies, codes are required to be well commented such that the underlying business logic can be recorded and learned by others, design document need to be well written with team reviews recorded, such that everyone else is aware of what is going on. Materialization can potentially harm employee productivity in the short term, however, it is very helpful with retaining professional knowledge in long term consideration.

There are many other approaches implemented in tech companies to help minimize the impact of employee departure. For example, keeping active communication among teams such as daily or weekly standup meetings, scheduled one on one meetings, and whole team review sessions, all of which can ensure that personal knowledge are proactively shared and transferred to others. Anyway, tech companies being a typical industry with especially high employee turnover rate, have put knowledge retention as a high priority and have developed many effective human resource management approaches that make organizations sustainably outperforming others and deserve study.

2.3.2 Approaches for medium-KLR organizations

Medium-KLR organizations normally have medium term employee tenure (3 to 5 years). There are a lot of organizations that fall into this category, including construction, manufacturing, financial, education, healthcare, and a lot of other industries. There are various approaches that have been researched and developed to deal with knowledge loss, and these approaches are usually general applicable to any kind of industries.

Knowledge capturing interviews, which start with asking questions to determine important and critical area, and followed up with questionnaires extracting tacit knowledge, have proven useful for eliciting knowledge from professionals to an organizational knowledge pool (Taylor, 2005). Interviews can assist respondents to articulate unorganized knowledge by probing their thoughts, further revealing and recording them.

Mentoring and storytelling, being one of the most effective knowledge transfer strategies, has been widely used for the purpose of knowledge retention. Specifically, not all kinds of knowledge can be easily materialized or verbalized to be captured into system or to be learned and transferred to others (Levy, 2011). In such cases, mentoring as an informal teaching format, not only benefit the employees who are mentored, but also improve organizational knowledge level (Swap, Leonard, Shields, & Abrams, 2001).

Job rotation is another effective way of transferring tacit knowledge among an organization. One step further from mentoring, job rotation provides rotators with a real learning scenario since rotators will be qualified to independently work on the job. Job rotation can ensure that all the rotators will eventually own sufficient knowledge to operate on the important positions, thus valuable knowledge and experience will potentially have a much higher chance to be retained in the organization (Lu & Yang, 2015).

Many other general approaches have also been proposed by researchers and entrepreneurs to help with knowledge retention including group learning (Li & Cheng, 2013), establishing organizational knowledge pool and database (Wang & Jin, 2004), seminars, coffee hour chatting, etc. All of these approaches can assist with increased level of knowledge and experience retention rate, further bringing sustainable competitiveness to organizations.

2.3.3 Approaches for low-KLR organizations

Unlike high-KLR and medium-KLR organizations, low-KLR organizations feature the longest term of employee tenure and are very stable in organizational structure, staff structure and knowledge pool. Although there are some industries like utilities and telecommunications falling into this category, we will focus on the most typical and important kind of organizations – public sector ones, or governments, including federal, state, and local ones. One of the most unique features of public sector organizations when compared to private sector ones is that public sector ones are operating in a much less competitive environment (Kaplan, 2013). Such a feature leads to public sector organizations paying less attention to knowledge management and tending to follow suit on knowledge retention approaches developed by private sectors. Another critical problem that public sector organizations are facing is employee aging. People working for governments are more prone to stable life and are less inclined to turnover aggressively. In this context, employee aging is rather considered a knowledge risk for the future. Without doubt, there are many other characteristics that distinguish public sector organizations from private sector ones, and these will be studied in detail in the next chapter. Admittedly, some government organizations have already been implementing general knowledge management protocols mentioned in the previous section. But here, we will focus on some simple knowledge retention approaches more frequently adopted by governments as low-KLR organizations.

One approach as pointed out by Liebowitz is to keep using retirees, because retirees are not only key experience holders, but also have plentiful resources established (Liebowitz, 2004). This is surely effective in knowledge retention since there will be no knowledge loss in a short term. However, this is a very unhealthy approach that does not actually realize knowledge transfer in the long term.

Other simple and naïve knowledge retention approaches such as presentation, ad-hoc conversion, etc. are also being implemented, which are usually done within the two weeks window period. However, these activities implemented once in a while are not systematic, inefficient and are reactive for the purpose of knowledge retention.

As a typical example of a public sector organization, Georgia DOT has been managing the organization's human resource data and intellectual assets in a relatively passive way. For more than ten years, Georgia DOT has been keeping an annually updated spreadsheet of around 4000 rows (employees) by 120 columns (attributes) for all their employees' data. On one hand, the data covers a wide variety of information on each of the employees. On the other hand, there were no adequate tools or systems to properly access such corpus with a large amount of data. Without rigorous analysis or proper transformation on those data, important information that can be potentially helpful for knowledge retention is only contained in the spreadsheet and cannot be exploited or utilized.

2.4 Concluding Remarks

Organizational knowledge loss can be very costly, especially for large organizations. Knowledge loss can come from multiple facets, among which employee departure is the most impactful one. Employee turnover rate can be a meaningful indicator for KLR level. On the other hand, employee turnover rate can be derived from average employee tenure.

Organizations in different industries usually have different employee tenure, thus suffer from different levels of KLR. Among all the industries, tech companies usually have the lowest employee tenures while governments usually have the highest employee tenures. There is a negative correlation between employee tenure and KLR level, which can be shown on a graph. Furthermore, organizations can be classified based on the KLR level, specifically high-KLR, medium-KLR and low-KLR organizations.

There are plenty of current knowledge retention protocols implemented in organizations. Some are general approaches that are suitable for most organizations such as mentoring, job rotation, etc. Tech companies, with highest level of KLR, have developed systematic and effective approaches for optimal knowledge transfer, such as unified coding style, readability review, comprehensive design documentation, etc. Governments, being in a less competitive environment, are least progressive with respect to knowledge retention activities. These current approaches are studied as a foundation for the following chapters to propose new framework for human resource data management.

CHAPTER 3. PROPOSED HR DATA MANAGEMENT APPROACH

This chapter summarizes the key features of public sector organizations in the context of knowledge management. An emerging challenge in knowledge loss will be discussed, and Georgia Department of Transportation (DOT) will be studied as a typical example of state government. This chapter also introduces a human resource data management framework at a high level, with the objective of addressing the knowledge loss challenge in response to the market change.

3.1 Introduction

One of the main and the most important functions of public sector organizations is serving the people in its community. While knowledge loss can significantly jeopardize an organization's work quality, productivity, and reputation, it can negatively impact people's life quality when the scenario is a government organization losing its critical knowledge. Moreover, public sector organizations not only process but also generate important knowledge (Dewah & Mutula, 2016). Therefore, approaches to knowledge management and retention ought to be put at high priority in public sector organizations (Cong & Pandya, 2003), however it has yet to be given sufficient attention (Cong, Li ‐ Hua, & Stonehouse, 2007).

Unlike most privately owned companies, public organizations have some unique differentiating characteristics. Many have argued that public and private sector organizations are becoming more similar than in previous decades (Lawton & Rose, 1991) and that some management techniques from private sectors can be adopted to public sector organizations (Mintzberg, 1973). However, there is research showing that blind copying from private sector organizations without taking into consideration public sector ones' characteristics can lead to failure (Cong et al., 2007). First of all, in contrast to private organizations who are driven by profit and expansion, public organizations are more focused on serving the community, thus leaders cannot always use monetary incentives for employee rewards. Another fact is that public organizations are less exposed to market variations and experience much less competition compared to private organizations. Moreover, public organizations are constrained by more formal legal factors when considering recruiting, promoting, and other human resource management issues. (Calo, 2008; Doan et al., 2011; Izard-Carroll, 2016; Rehman, 2012).

In the context of knowledge management, most of the current strategies implemented in public sector organizations are classical approaches. With the emergence and development of computational ability and data science techniques, knowledge management can be realized in a more advanced way. With taking characteristics of public sector organizations into consideration, and the application of modern technologies, a study in systematic and automatic knowledge management not only is beneficial to organizational long term development, but also advances employee overall knowledge level and motivation.

3.2 Business of Public Sector Organizations

The business of any organization consists of three parts: Organization environment, including market, customers, society and structure; Organization mission, its commitment and testimony; Core competency to fulfill the established mission statement (Drucker, 1994). Such business theory well fits both private and public organizations. However as mentioned, public sector organizations have its own characteristics due to its different market environment, structure, mission and many other factors. In this section, public organizations will be studied in detail regarding their unique features in the context of knowledge management. Moreover, a current trend about knowledge loss in public organizations will be discussed. Although Georgia Department of Transportation (DOT) will be used as a typical example for the research work, the findings and conclusions can be easily adopted to a lot of other public organizations, and even to private ones with similar organizational structures.

3.2.1 Features of public sector organizations

First of all, most public sector organizations have a stovepipe-type organizational structure, which is hierarchical, as shown in Figure 3.1. For example, Georgia DOT deployed the headquarter in Atlanta, and then the organization is divided into seven district offices with each one covering several counties. Each district office is then divided into several area offices with each branch covering a certain area. Such a district-area organized structure makes Georgia DOT a highly hierarchical organization. Moreover, every branch office has its own complete functional system as a local administration department.

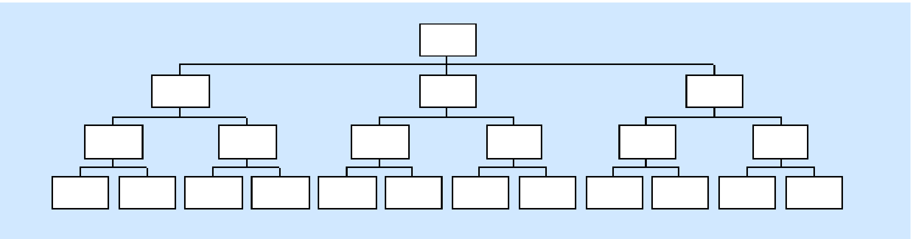


Figure 3.1 – Stovepipe-type organizational structure.

One of the main characteristics of the stovepipe-type organization is that the flow of information within the organization is restricted to a great extent along the top-down line of management. Although such kind of structure can help simplify administrative management, it also brings many disadvantages that largely prevent cross-department communication and organization-wise knowledge sharing since staff from multiple sites do not have the chance for regular interaction. The stovepipe-type structure also creates a long reporting lines, especially for large organizations with multiple intermediate management layers, impeding efficient flow of information. Many traditional large organizations, especially governments, fall into this type, while modern companies operate differently. For example, most tech companies have multiple branches across different states, and usually a functional department can span two or more geo locations. Staff transfer between different geo locations, and cross-department cooperation are also more frequent than in government organizations.

Another downside of this structure is that recruiting, promoting, and retiring are mostly conducted and planned on the local level. As a result, most knowledge transfer protocols are also implemented in the local department. There are limited chances for organization-wise knowledge transfer activities to be conducted even when the most willing and suitable candidate is in another branch office at a different location. Things can be even worse in certain rural areas, where there is a limited-size application pool to fulfill the local offices' hiring requirements and the stovepipe-type structure can cause a higher attrition risk. Therefore, it can be potentially much more effective and efficient if such a structural feature is taken into consideration when developing a knowledge management system for governmental organizations.

Another feature of public sector organizations is the longevity and stability of the positions, as well as the pay levels, which is unlike the case in profit-driven private organizations. The revenue for public sector organizations mainly comes from taxpayers as opposed to that of private ones from customers. The longevity and stability of positions make the employees well informed about their own duties, but may be not engaged with those of other positions at all. Especially when many employees have spent many years in the same position, this makes it even harder to transfer a large amount of knowledge. Further, the stable salary and lack of monetary incentives make staff less motivated towards learning new knowledge, preventing organizations from creating a healthy environment for knowledge gain that can benefit long term development.

Last but not the least, the issue of employee aging is faced by most public sector organizations. As research conducted by others has shown, the oil and gas industry is also bothered by such issues (Sumbal, Tsui, See-to, & Barendrecht, 2017). Taking Georgia DOT as an example, the average age of the workforce is 44 years old, and approximately 25 percent of the employees are eligible to retire within five years. More importantly, the organization has seen a shift in the generational makeup of its staff as Millennials have begun entering the workforce (GDOT, 2018). As a result, this group of staff will continue to make up a large portion of the workforce as the older generations continue to retire in the next decade, thus employee aging is not only of concern for current situation but also for the future.

On one hand, the large number of senior employees possess a large amount of valuable knowledge. On the other hand, knowledge possessed by senior employees has evolved over many years during their tenure in the organization, which is not easy nor fast to be transferred to someone else (Levy, 2011). For example, mentoring would take a longer than usual period, and the candidate must be properly selected due to the level of domain knowledge that may be required. With this regard, employee aging not only creates fast-paced attrition, but also make knowledge transfer more difficult. Moreover, the lack of incentives and rewards in public sector organizations as mentioned earlier can make senior employees even less motivated, further impeding knowledge retention progress.

3.2.2 Emerging challenges

Having studied the features of public sector organizations, this section will explore some emerging challenges in the context of knowledge retention, especially those faced by government organizations. The first challenge is that government agencies are being faced with an increasing demand for work along with and a decreasing number of employees.

Figure 3.2 shows the Population Growth in the state of Georgia over the last few decades (Source: US Census Bureau, 2018). Georgia DOT, with its mission of "Delivering a transportation system focused on innovation, safety, sustainability and mobility", is hence facing an increasing demand on providing and maintaining the state transportation systems as the population keep growing. In the meantime, Figure 3.3 shows the number of employees in Georgia DOT since a decade ago (Source: Georgia DOT Office of Human Resource, April 2018), where the dramatic decrease from more than 5,500 to less than 4,000 can be seen. Under the situation of increased demand and decreased workforce, many employees are required to perform job tasks with minimal training. This has placed Georgia DOT's ability to effectively accomplish the mission statement established by the organization in jeopardy. Therefore, while actively hiring more employees is important, it is also critical to let current employees work more efficiently through innovate strategies including effective knowledge retention and transfer.

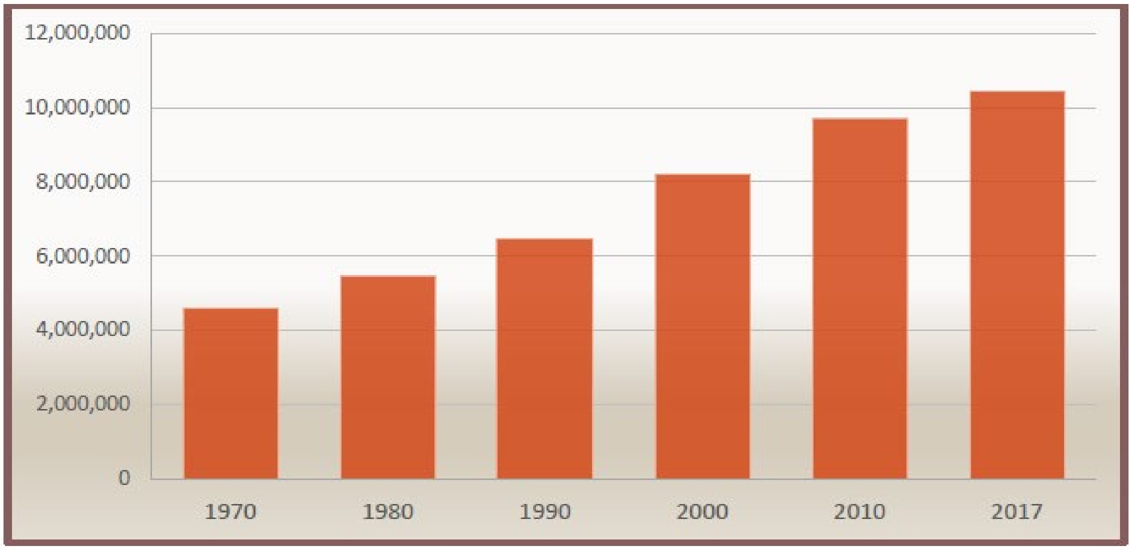


Figure 3.2 – Georgia's Population Growth (Source: US Census Bureau, 2018).

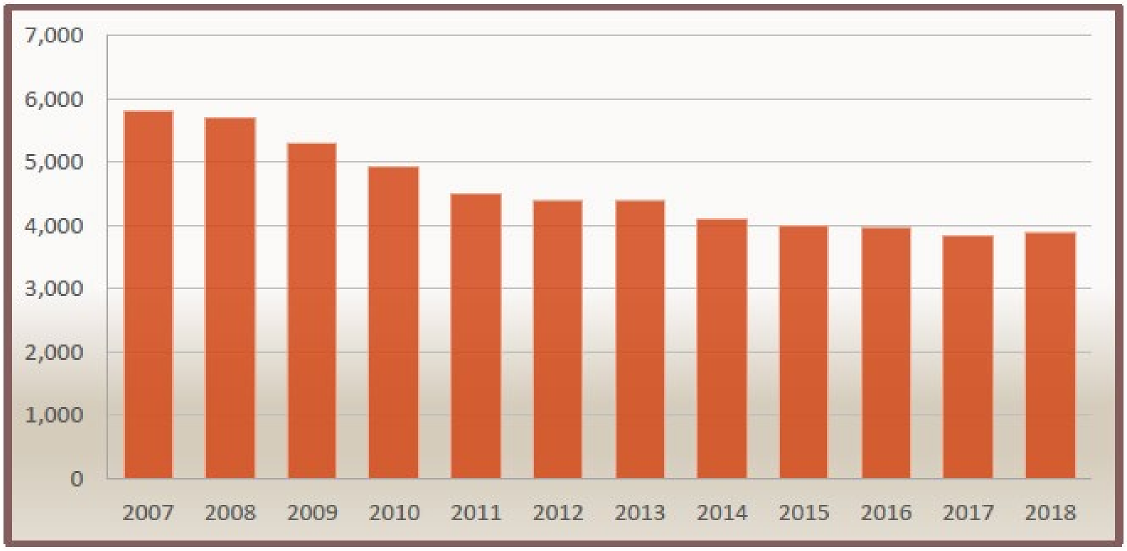


Figure 3.3 – Georgia DOT Employees (Source: Georgia DOT Office of Human Resource, April 2018).

The decreasing number of employees in Georgia DOT is not due to a single reason. Admittedly, employee aging, as mentioned in the previous section, is an important factor.

Moreover, state budgetary restrictions, retirement plan restructuring, and other external factors are also causing a result that the number of retiring employees has exceeded the number of new individuals being hired. More importantly, an emerging challenge of increased turnover rate, not only to Georgia DOT but also many other government organizations, is playing a critical role in workforce attrition. From Table 1, we can witness a decreasing median tenure in public sector organizations from 7.7 years to 6.8 years. And the trend is even more severe in local governments. The cause of the shortening employee tenure here is more due to younger employees than to retiring ones. With years of work experience in government agencies, many employees prefer to seek opportunities elsewhere, for example in consulting companies, or other private firms. Furthermore, career development considerations are also contributing to this trend. It can be seen from Figure 3.4 that the decreasing average tenure is pushing public sector organizations from the zone of low KLR level toward the zone of medium KLR level.

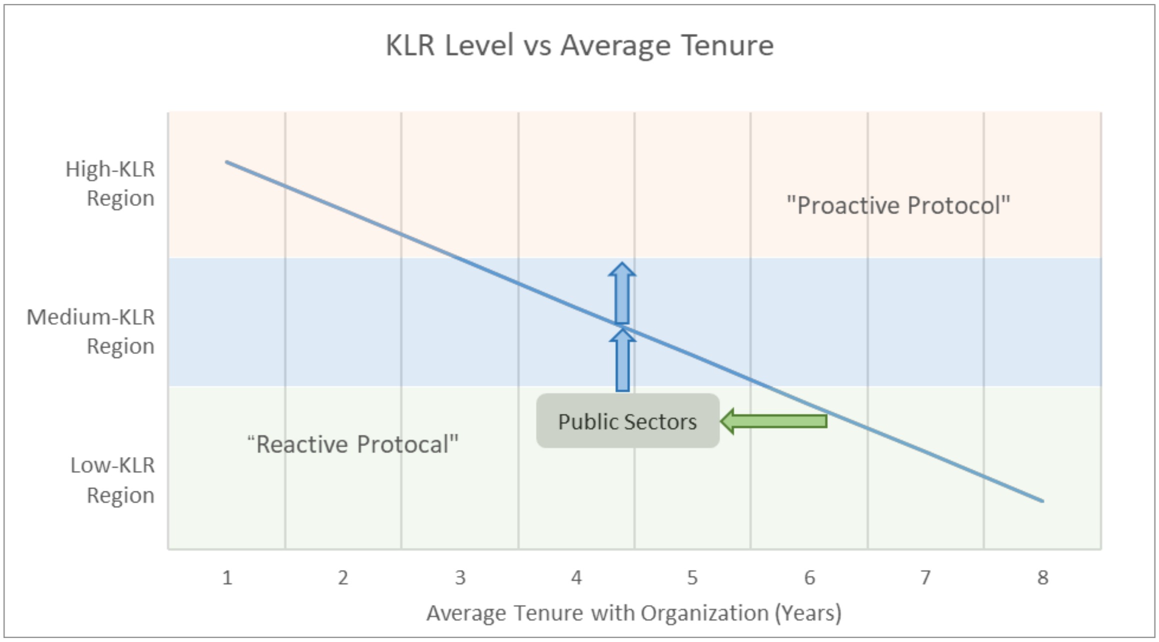


Figure 3.4 – Showing the emerging challenge, higher turnover rate, on the KLR Level vs Average Tenure table.

If we divide the chart into two regions with the line, the lower region can be considered “reactive” while the upper region would be “proactive” with regard to knowledge retention. As public sector organizations being pushed towards the lower average tenure, its current knowledge retention protocols would be more reactive than proactive such that less knowledge are effectively transferred, and more are lost during attrition. Therefore, new knowledge retention tools need to be developed and implemented in public sector organizations to push upwards in the chart. With proper knowledge management tools and a range of strategic activities, public sectors organizations can be moved to the balance line, or even towards the region of proactive. As a result, new systems can potentially lead to benefits such as increased knowledge transfer, lower employee turnover, and higher levels of employee knowledge and motivation, amongst others.

3.3 Proposed Knowledge Retention Framework

Due to the features of public sector organizations and the emerging challenges faced that were studied in the previous section, it is critical to propose a proactive systematic framework for an organization’s human source department to achieve optimal knowledge retention. The framework should primarily address two facets: (i) identifying and (ii) transferring critical knowledge at risk of being lost. This section will propose the framework in the context of each of the two facets at a high level, followed by the next chapter that describe each computational and data science technique in detail.

3.3.1 Identification of critical knowledge at risk

First of all, it is a fact that not all knowledge is critical to organizations, as well that not all knowledge is at risk of being lost. With limited budget and workforce that can be devoted to knowledge retention activities, it is important to identify critical knowledge which is most at risk of being lost at the very beginning (Frigo, 2006). Thus, the identification will be discussed in two parts: the identification of critical knowledge; the identification of knowledge that is at risk of being lost.

The identification of critical knowledge requires the recognition of the positions or employees who are critical to the organization. Many managers or human resource personnel recognize those at high level as important ones. However, criticality should be considered not only strategically but also tactically, and position level should not be the only factor for evaluation. There are many other factors such as the uniqueness of the position, years of tenure, resource availability, skill sets, etc. that also contribute to position criticality. Moreover, it is also necessary to consider how impactful it is when a position is absent due to unforeseen reasons. Furthermore, as organization being a network of social groupings (Katz & Kahn, 1966), the relationships and patterns among its members should not be ignored. Especially considering that information flow during interaction among employees forms an important way of knowledge sharing, social network analysis is considered a useful computational technique in the context of knowledge retention. To sum up, multivariate analysis on the combination of various metrics of employees, and social network analysis on the relationships amongst, can evaluate knowledge criticality comprehensively.

The identification of knowledge that is at risk of being lost requires the recognition of the employees who are expected to leave the organization within a specific amount of time. It is unlikely for organization managers to be able to predict those who will leave the organization to seek opportunities elsewhere. However projected retirement on the other hand is a foreseen workforce attrition that can be determined by employee's age and years of tenure. Pivot table, as a dynamic summarization tool that allows for sorting, counting, averaging and many other statistical analysis, is particularly powerful in recognizing near retirement employees. Further, it can also be integrated with pivot chart to visualize attrition risk. Moreover, for large organizations that have many branch offices in different locations across the state, the country, or even the world, geographical information visualization can deliver an intuitive overview of positions in each office with the number of years to retirement, across the whole organization. To sum up, statistical analysis and visualization of employee's number of years to retire can give an effective indication on the knowledge that is at risk of being lost.

3.3.2 Approach for knowledge transfer

Once critical knowledge at risk has been identified, the important employees at the critical positions are also recognized. Thus, the items remaining would be approaches to retain the critical knowledge. Knowledge transfer in organizations is the process through which the knowledge-taker is affected by the experience of the knowledge-giver (Argote & Ingram, 2000). Therefore, knowledge transfer activities basically have two main parts: choosing the most proper and practical knowledge transfer activity; determining the most suitable candidates to be the knowledge-taker.

Organizations usually have various training and development strategies, for example training classes, job rotation, job shadowing, desk side reviews, etc. Choosing the most proper and practical knowledge transfer strategy not only depends on the willingness, availability and locations, of the trainer and trainee, but also external factors like IT infrastructure availability. A coefficient matrix can be established by an organization's human resource personnel with each element in the matrix indicating the suitability score of the evaluating criteria associated with each strategy. With the matrix established, questionnaires can be answered by the trainer and the trainee for analyzing, ranking, and visualizing the suitability of each knowledge transfer strategy.

Optimal knowledge transfer not only relies on the choice of the most proper knowledge transfer activity, but also on the selection of the most suitable candidates for the position of interest to conduct the activity. For most of the knowledge transfer activities including job shadowing, cross training or job rotation, succession planning, etc., it is important that the selected candidate have a similar job content as the position being planned. Most managers and human resource personnel tend to select those who work in the same department as or even directly under that planned position, largely due to the aforementioned stovepipe-type structure of the organization. However, it is not always true that the individuals directly under a position share a lot in common with the planned one in job content. Moreover, there are scenarios that the most suitable candidate can even be in another branch office not co-located as the planned one, thus the stovepipe-type structure needs to be "broken" to achieve organization-wise knowledge transfer rather than department-wise. Therefore, it is critical to perform similarity analysis on the job contents prior to the selection of the most suitable candidate. Furthermore, factors other than job content similarity, such as years of tenure, past experience and performance, leadership, etc., can also play important roles on candidate suitability. All related factors should be evaluated and analyzed for comprehensive assessment of the suitability, and a suitability score can be calculated for quantitative evaluation. To sum up, knowledge transfer should not be limited to a single department or a single branch office, but rather on organization-wise level. Multivariate analysis on all related factors can comprehensively evaluate candidate suitability.

3.4 Concluding Remarks

Public sector organizations have many features distinguishing them from private sector ones including a stovepipe-type structure, as well as employee aging. Further, they are also facing emerging challenges such as increasing workload along with weakening workforce, higher turnover rate among younger employees. Georgia DOT has been studied as a typical example to illustrate the business approach of a public sector organization.

The characteristics and challenges of public organizations have a great impact on the methodology and implementation of knowledge retention strategies, thus there is this need to propose a new human resource and knowledge management framework for them. The framework has two main tasks to ensure optimal knowledge retention. The first task is the identification of critical knowledge that is most at risk of being lost. With limited budget and workforce, knowledge transfer strategy should focus on the knowledge not only critical to the organization but also at risk of being lost, and KLR score can be used for quantitative evaluation. The second task is the selection of knowledge transfer activity and the identification of the best candidate. Multiple factors need to be taken into consideration when making selections. With multivariate analysis, calculated candidate suitability scores can be used for quantitative assessment. This chapter proposes a human resource data management framework on a high level, while the next chapter will study each of the computational and data science techniques in detail.

CHAPTER 4. DATA TOOLS AND TECHNIQUES

This chapter describes the data tools and techniques that to be used in the proposed human resource data management approach. With the objective of developing an automated and scalable knowledge retention framework, several computational tools and data science techniques, including geographical information visualization, social network analysis, text mining and machine learning, multivariate analysis and temporal analysis will be applied in the proposed approaches. Their effectiveness and applicability will be studied in detail.

4.1 Geographical Information Visualization

4.1.1 Overview

Databases, data warehouses, and other data repositories, which contain a large amount of data, are growing rapidly and are considered data-rich (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). In order to take them from data-rich to information-rich, data mining in databases has become an important research field. Among all kinds of data mining technologies, data visualization is an effective way, which enhances users' perception and interaction with complex real-world data, regardless of their prior experience. For abstract data such as categorical, geographical, spatial, and other non-numerical information, data visualization is also useful for reinforcing users' cognition.

In recent years, many organizations have seen a dramatic decrease in the number of employees. Many long-time employees are nearing retirement age, creating a potential critical loss of personnel, knowledge and experience for the organization. In some cases, the number of retiring employees has exceeded the number of new individuals being hired as a result of budgetary restrictions, retirement plan restructuring, and other external factors. Such retirement information needs to be of concern for organization human resource personnel, thus deserve particular attention. In the meanwhile, projected retirement data can be calculated from an organization's human source database, where there are thousands of employees' information, including their basic bio info, profession-related info, and location-related info. With proper analysis and transformation on the database, useful information including retirement statistics can be revealed. Foreseen retirement, along with office location information (especially for large organizations with multiple branch offices across different geo locations) can be disclosed through geographical information visualization techniques, giving an intuitive organization-wide overview distribution of attrition and knowledge loss risk.

4.1.2 Methodology

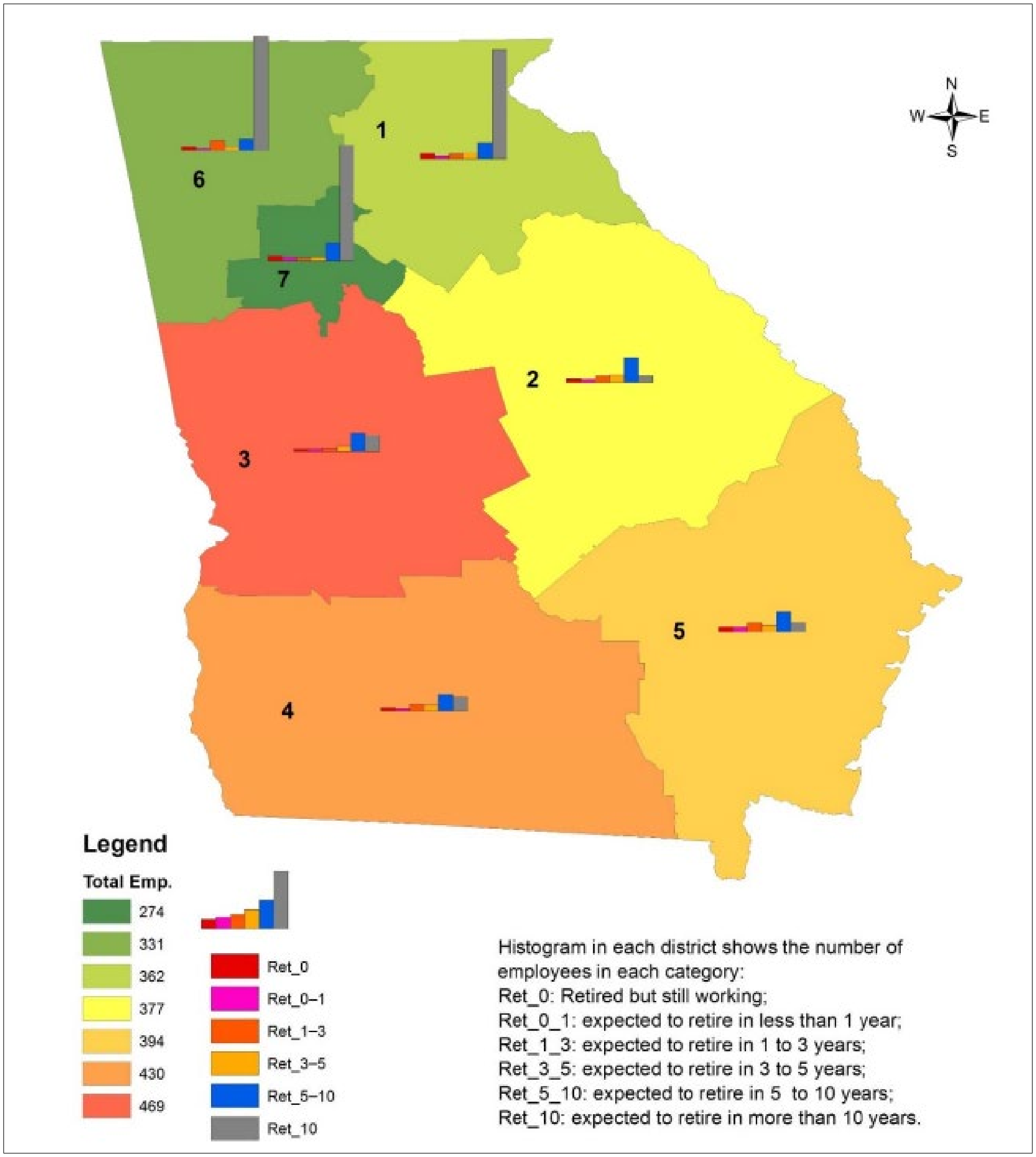
Above all, the first step for the any kind of data visualization work should be data preprocessing. A lot of the information that is critical for data visualization tools to render stays in the form of abstract data in the database, thus it must be extracted. Similarity, some numerical data that potentially needs to be visualized as categorical information also needs some form of preliminary categorization. For example, the district and area of a branch office can be read from the department code and the detailed address can be associated for spatial analysis, tenure can be calculated through the job entry date and then categorized into intervals, years to retire can be calculated from the job entry date and organization policy projected retirement date, etc.

After the data has been preprocessed to the format suitable for analysis and visualization, analytics and visualization tools can be applied. A pivot table is a dynamic data summarization and analysis tool which can allow sorting, counting, averaging, and performing many other mathematical and statistical operations, and a pivot chart enables bar chart style visualization on top of the results from the pivot table. The combination of pivot tables and charts allow for analysis and visualization of the data extracted from a database in an interactive and dynamic way. Additionally, in the case of organizations with many offices across a state, country or even the world, it is beneficial to have an understanding of the spatial distribution of employee information across all locations. In this regard, the use of geoinformatics can allow for management, analysis and visualization of employee data spatially, to understand patterns, trends, and relationships. This can be accomplished by integrating the pivot analysis results with geographical analysis tools.

Interactive pivot tables and charts can be realized by a variety of visual analytics tools. JavaScript libraries such as d3.js and c3.js, can provide interactive visualization for web browser-based application. Visual analytics software such as MS Power BI and Tableau can provide comprehensive functionalities, with a variety of plugins. MS Excel, as a general spreadsheet program, can also serve as a powerful visualization and analysis tool. Experiments on different tools are performed for contrast and comparison.

4.1.3 Results and Conclusions

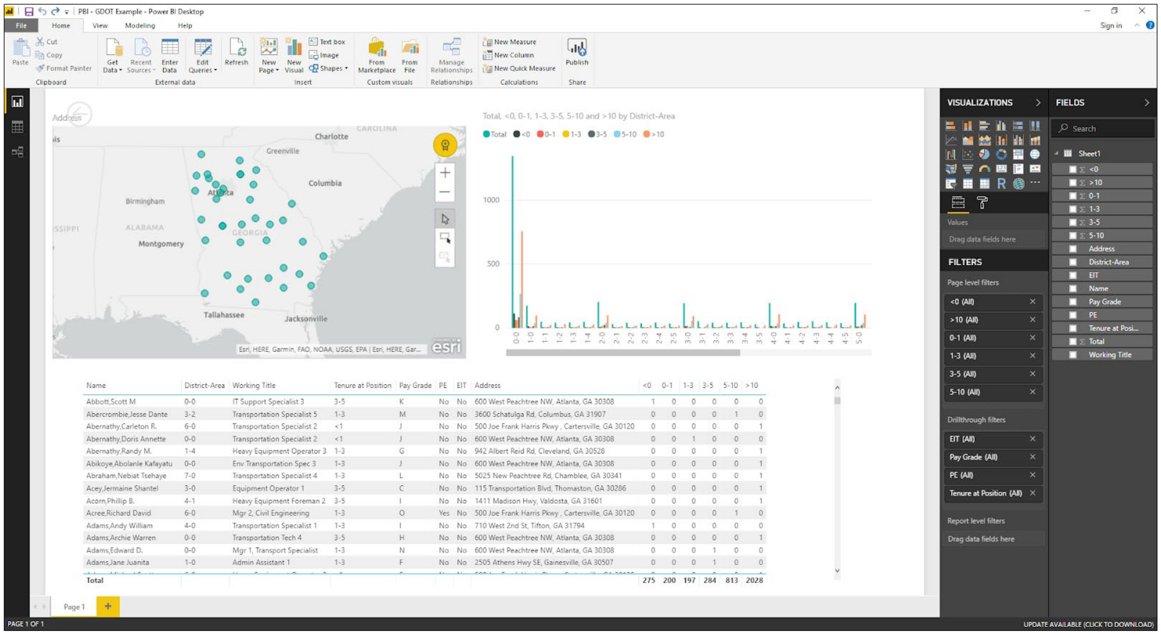
With preprocessed human resource data from Georgia DOT database as an example, interactive visualization studies are built on the web with d3.js, on MS Power BI, and on MS Excel, respectively. Figure 4.1 shows a web visualization of Georgia DOT attrition data, aggregated by each district using d3.js library. It enables user interaction by simply clicking on each district to show detailed attrition data for the areas.



Histogram in each district shows the number of employees in each category:
Ret_0: Retired but still working;
Ret_0_1: expected to retire in less than 1 year;
Ret_1_3: expected to retire in 1 to 3 years;
Ret_3_5: expected to retire in 3 to 5 years;
Ret_5_10: expected to retire in 5 to 10 years;
Ret_10: expected to retire in more than 10 years.

Figure 4.1 – Web visualization view of Georgia DOT attrition data for each district with d3.js.

Figure 4.2 shows an interactive visualization of Georgia DOT human resource information using MS Power BI. It not only supports dynamic pivot chart and summarization pivot table, but also has a geographic visualization panel provided by ESRI plugin. Moreover, all the visualization panels are interlinked with each other, supporting dynamic selection, filtering, as well as ranking. Furthermore, users are able to create their own visualization on top the pivot table, allowing for customization.



Name	District-Area	Working Title	Tenure at Position	Pay Grade	PE	EIT	Address	<0	0-1	1-3	3-5	5-10	>10
Abbott,Scott M	0-0	IT Support Specialist 3	3-5	K	No	No	600 West Peachtree NW, Atlanta, GA 30308	1	0	0	0	0	0
Abercrombie,Jesse Dante	3-2	Transportation Specialist 5	1-3	M	No	No	3600 Schatulga Rd, Columbus, GA 31907	0	0	0	0	1	0
Abernathy,Carleton R.	6-0	Transportation Specialist 2	<1	J	No	No	500 Joe Frank Harris Pkwy , Cartersville, GA 30120	0	0	0	0	0	1
Abernathy,Doris Annette	0-0	Transportation Specialist 2	<1	J	No	No	600 West Peachtree NW, Atlanta, GA 30308	0	0	1	0	0	0
Abernathy,Randy M.	1-4	Heavy Equipment Operator 3	1-3	G	No	No	942 Albert Reid Rd, Cleveland, GA 30528	0	0	0	0	0	1
Abikoye,Abolanle Kafayatu	0-0	Env Transportation Spec 3	1-3	J	No	No	600 West Peachtree NW, Atlanta, GA 30308	0	0	0	0	0	1
Abraham,Nebiat Tsehaye	7-0	Transportation Specialist 4	1-3	L	No	No	5025 New Peachtree Rd, Chamblee, GA 30341	0	0	0	0	0	1
Acey,Jermaine Shantel	3-0	Equipment Operator 1	3-5	C	No	No	115 Transportation Blvd, Thomaston, GA 30286	0	0	0	0	0	1
Acorn,Phillip B.	4-1	Heavy Equipment Foreman 2	3-5	I	No	No	1411 Madison Hwy, Valdosta, GA 31601	0	0	0	0	0	1
Acree,Richard David	6-0	Mgr 2, Civil Engineering	1-3	O	Yes	No	500 Joe Frank Harris Pkwy , Cartersville, GA 30120	0	0	0	0	1	0
Adams,Andy William	4-0	Transportation Specialist 1	1-3	I	No	No	710 West 2nd St, Tifton, GA 31794	1	0	0	0	0	0
Adams,Archie Warren	0-0	Transportation Tech 4	3-5	H	No	No	600 West Peachtree NW, Atlanta, GA 30308	0	0	0	0	0	1
Adams,Edward D.	0-0	Mgr 1, Transport Specialist	1-3	N	No	No	600 West Peachtree NW, Atlanta, GA 30308	0	0	0	1	0	0
Adams,Jane Juanita	1-0	Admin Assistant 1	1-3	F	No	No	2505 Athens Hwy SE, Gainesville, GA 30507	0	0	0	1	0	0
Total								275	200	197	284	813	2028

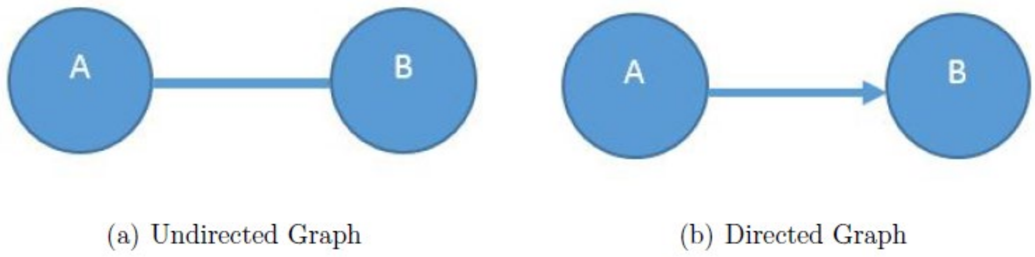
Figure 4.2 – Interactive visualization view of Georgia DOT human resource information on MS Power BI.

While MS Excel also supports interactive pivot charts, it’s not easy to implement the geographical view with plugins. Considering development complexity, ease of use, and data security, integrating MS Power BI into the proposed knowledge retention framework was considered to be a superior choice for web visualization. Moreover, since the potential users will be human resource personnel in public organizations, MS Power BI also provides a more user friendly, and more comprehensive interactive functionalities.

4.2 Social Network Analysis

4.2.1 Overview

Social Network Analysis (SNA) examines social structures through relationship between subjects under investigation (SUI) using network theory and graph theory (Otte & Rousseau, 2002). Research on SNA has been an interest for a few decades in the area of social science (Borgatti, Mehra, Brass, & Labianca, 2009). Moreover, as a multidisciplinary field, SNA has also been widely used in many other domains including friendship networks (e.g. Facebook), business networks (e.g. LinkedIn), knowledge networks (e.g. Google Scholar), disease transmission (e.g. CDC), etc. In a graph view, the objects in a network are often referred to as nodes and the relationship between objects as edges. In general, a graph can be directed graph or undirected graph depending on whether there is an orientation in the edges, as can be seen from Figure 4.3. Directed graph is used in hierarchical structures such as topological dependencies, website links, etc. On the other hand, undirected graph is usually applied in bidirectional relationships such as social network where friendship relations are mutual.



(a) Undirected Graph

(b) Directed Graph

Figure 4.3 – Basic structures of graph.

The technique of SNA can be employed in any fields where as long as objects are connected with each other through some sort of relationship. In the context of organizations, as they contain a set of employees with patterns, interactions and information flow amongst them, thus can be viewed as a network of social groupings (Katz & Kahn, 1966). An effective way to study the social structure of such a grouping is to investigate the patterns of the relationship between its members (Wellman, 1983), which is exactly what SNA does. Through the application of SNA theory, great advances can be made in the field of organization theory, for example the sub clusters within an organization, especially important individuals, career patterns and succession, etc. (Tichy, Tushman, & Fombrun, 1979). In organizations consisting of multiple sub-units, professional networks and the relationships between the network members can be especially important in determining the flow and sharing of data and knowledge through the network (Hansen, 1999; Hansen, Mors, & Løvås, 2005; Reagans & McEvily, 2003).

4.2.2 Methodology

First of all, the most critical task in SNA is establishing the graph. Given the human resource database of an organization, the "reporting to" information which describes "who reports to whom" in the organization, can serve as the "edge" in the graph. These edges are connecting all the "employees" which serve as the "nodes" in the graph. Moreover, since the research is focusing on knowledge sharing, which is realized by the information flow between employees, and the information flows are considered two-way, thus the established graph should rather be established as an undirected graph.

In terms of this study's topic, i.e. knowledge retention, SNA can be potentially helpful on one of the main tasks: recognition of the critical knowledge. This can be achieved by evaluating employees' criticality in terms of information flow, as critical employees are usually the holders of knowledge and experience that is critical to the organization. Relative importance, or "centrality", of members often has meaningful implications on the criticality for professional networks. There are a number of ways for centrality measurement, such as betweenness centrality, closeness centrality, stress centrality, PageRank centrality, etc. For example, betweenness centrality (Butts, 2008; Freeman, 1978) can serve as an indicator of a member's relative importance in a network, with high BC individuals acting as "bridges" between different groups that may otherwise be loosely connected. In Figure 4.4, node A is such a member with relatively high betweenness centrality. High centrality individuals in the employee graph tend to have a large influence on the sharing and transfer of resources or knowledge through edges assuming that the transfer takes place along the shortest paths associated with a given network member (Barthelemy, 2004). As for graph algorithms that calculate betweenness centrality, the most famous Brandes' Algorithm (Brandes, 2001) calculates the exact value in O (|V| |E|). Furthermore, faster approximation algorithms (Bader, Kintali, Madduri, & Mihail, 2007) or incremental algorithms (Green, McColl, & Bader, 2012) have also been developed.

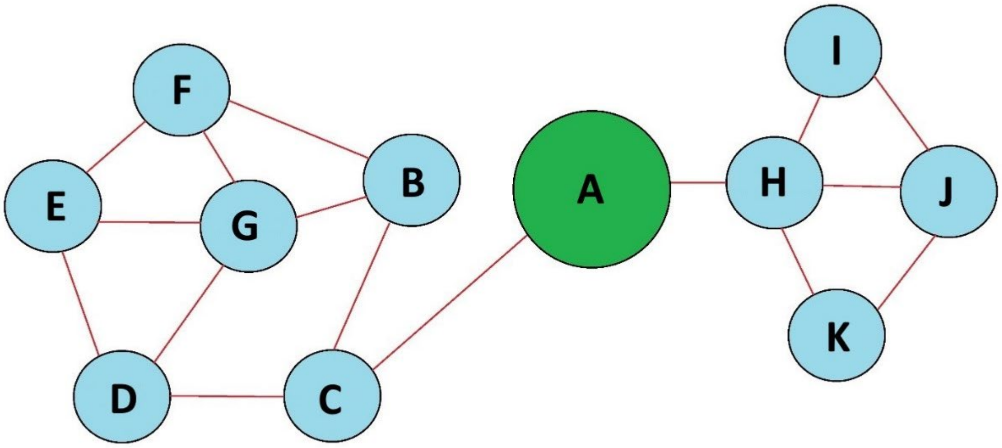


Figure 4.4 – An example of betweenness centrality. Node A has the highest centrality value as it acts as a bridge connecting left cluster and right cluster of nodes.

Apparently as can be seen, betweenness centrality will be an important metric to be used in this study to evaluate personnel relative criticality in an organization. However, betweenness centrality only takes into account the immediate ties that a member has in a network. There are scenarios where one node tied to a large number of other nodes, but those others can be disconnected from the network as a whole. In such a case, the node may have high centrality, but only in a local neighborhood (Hanneman & Riddle, 2005). Closeness centrality is a metric that tackles this limitation. Closeness, calculated by the inverse of the farness sum (Bavelas, 1948), effectively measure how long it will take to spread information from a specific node to all other nodes. Moreover, Katz centrality, which takes into account the total number of walks between a pair of nodes, makes another improvement through penalizing distant neighbors by an attenuation factor (Katz, 1953). Furthermore, PageRank centrality, developed by Google founders (Brin & Page, 1998), is similar to Katz centrality, however, overcomes the problem that nodes directly linked with a high centrality node get high centrality. PageRank centrality considers three factors that together determine the centrality of a node: a) the number of links it receives; b) the link propensity of the linkers; and c) the centrality of the linkers. After all, if compared to betweenness centrality, PageRank centrality considers not only the edges of a node, but also the quality of those edges. For example in Figure 4.5, it is obvious that node E has a higher betweenness centrality than node C, however, node C has a higher PageRank centrality than node E because it receives link from node B. Lastly, although PageRank centrality was firstly developed for measuring the importance of website pages which are usually modeled as directed graphs, it is perfectly applicable to undirected graphs since we can always treat undirected graph as directed graph with two edges between each pair of nodes. Therefore, this centrality measurement can be used to assess the importance of employees in an organization.

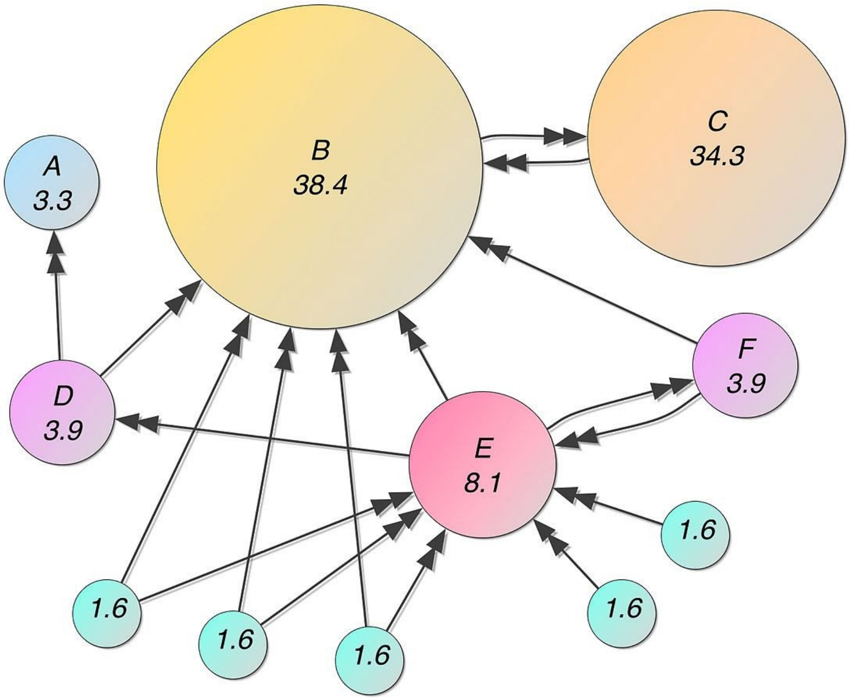


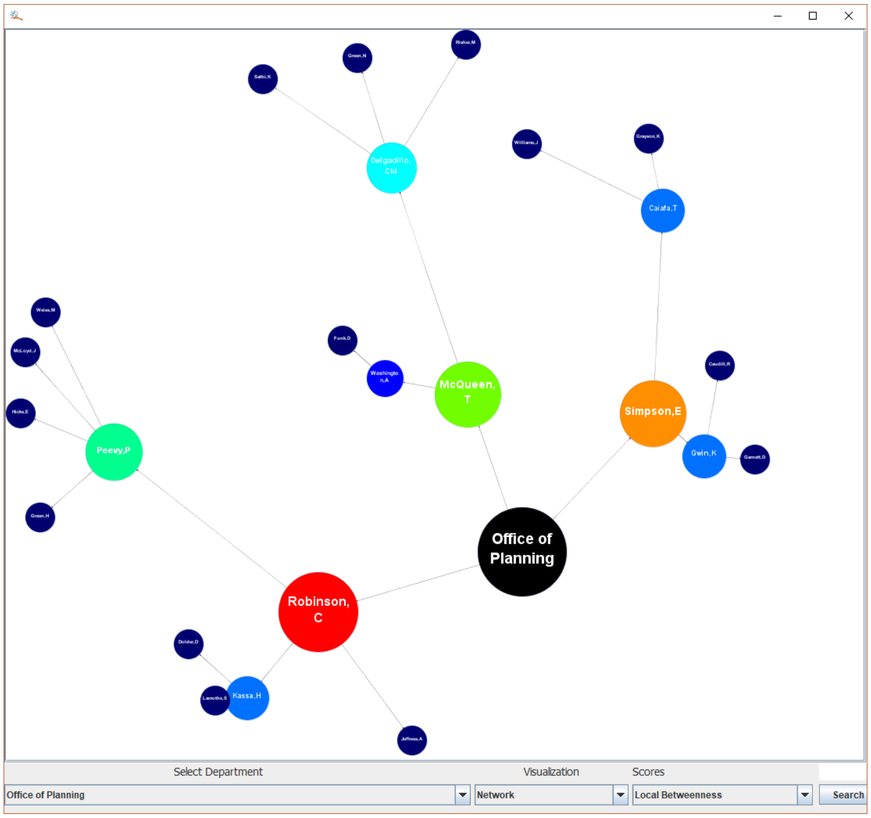
Figure 4.5 – An example of PageRank centrality results. Although node E has higher betweenness centrality than node C, however, node C has higher PageRank centrality than page E.

To sum up, both betweenness centrality and PageRank centrality are important measurements for employee importance, but from different perspectives. As betweenness centrality focus on the information bridge among employees, it can be helpful on identifying employees strategically important to the organization. On the other hand, while PageRank centrality also has a focus on the neighborhood quality, it is potentially useful in identifying employees tactically important to the organization. Therefore, both centrality measurements will be applied to the employee network established with "who report to whom" relationship, in order to assess the relative importance of employees, in regard to critical knowledge identification.

4.2.3 Results and Conclusions

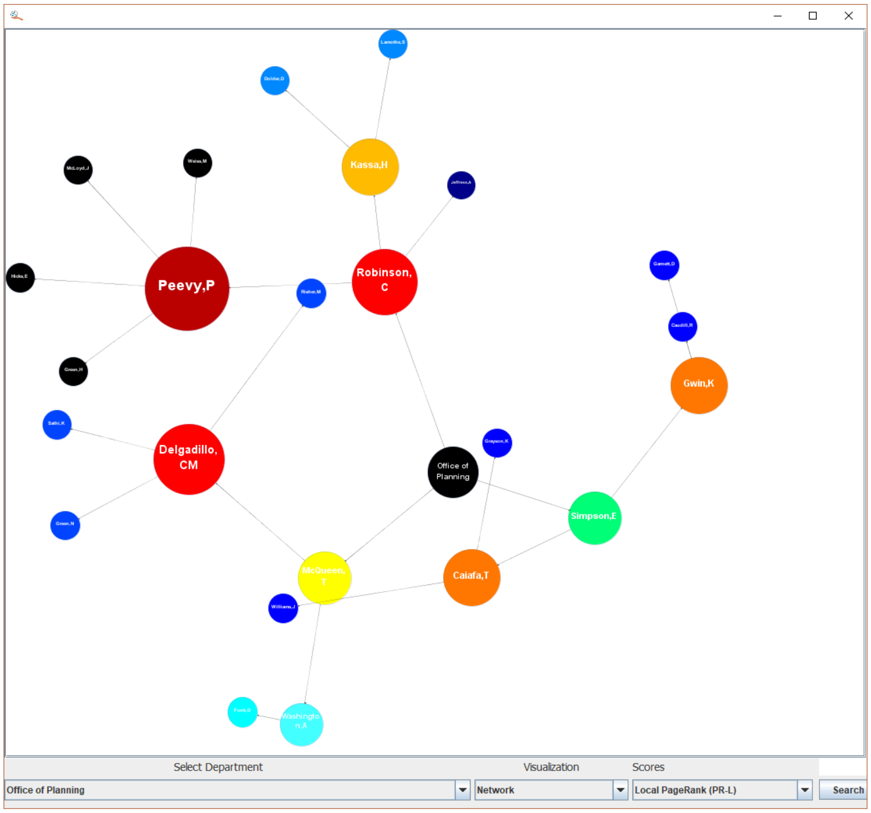
Georgia DOT's human resource database is used here as an example to study the effectiveness of using betweenness and PageRank centrality for employee criticality measurements. The experiments, including graph metrics analysis and intuitive visualizations are conducted using Java and Java GUI.

First of all, an undirected graph is established with the employee data from one sub office in Georgia DOT, i.e. the Office of Planning. Breadth first search (BFS) algorithm is applied here to build the graph from the connections. JGraphT, a Java library of graph theory data structures and algorithms, is then used for efficiently computing graph metrics including betweenness centrality and PageRank centrality. For assisting with perception, JUNG (Java Universal Network / Graph Framework) is then applied for visualization. The node size is set to be proportional to the centrality of each measurements. Figure 4.6 shows the network view of the employees in the Office of Planning, with the node size indicating the betweenness centrality; Figure 4.7 shows that of the same employees in the same office, while the node size indicating the PageRank centrality.



Select Department: Office of Planning
Visualization: Network
Scores: Local Betweenness

Figure 4.6 – Network view of the employees in a department (with node sizes indicating betweenness centrality).



Select Department: Office of Planning

Visualization: Network

Scores: Local PageRank (PR-L)

Search

Figure 4.7 – Network view of the employees in a department (with node sizes indicating PageRank centrality).

Based on the comparison of the two figures, it can be seen that betweenness centrality measurement evaluates those with higher levels as having higher centrality, thus more critical. While PageRank centrality measurement evaluates those at middle levels have higher centrality, thus more critical. From this point of view, betweenness centrality evaluation can assist managers and HR personnel with more strategic decision making in terms of critical position identification, while PageRank centrality evaluation could be more helpful in tactical decision making. Moreover, as such an approach works well on the small portion of data, it is believed that the application of SNA to the overall database of the organization would also lead to a meaningful result in the context of critical knowledge identification for knowledge management.

4.3 Text Mining and Machine Learning

4.3.1 Overview

Text mining is defined as the process of discovering patterns and extracting useful information from the data that is in the form of natural text. It has been widely used in many areas including business, education, military, etc. As an integral part in the field of knowledge management, text mining has been used for talent planning, scenario forecasting, human talent predictions, and strategic decision making (Jantan, Hamdan, & Othman, 2010; Ranjan, Goyal, & Ahson, 2008; Sadath, 2013). Moreover, text mining has been a useful tool in the hiring process, such as keyword extraction, resume mining, vacancy mining, etc. (Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018).

In terms of organization knowledge retention, when human resource personnel have identified the critical knowledge that at risk of being lost, and are trying to plan for some activities that allow for knowledge transfer, such as succession planning, cross training etc., optimal knowledge transfer can only be effectively achieved if the most appropriate candidates for the position of interest are identified. Although sometimes it is a good choice to make such selections from those directly under the position of critical knowledge, or from someone that the manager is familiar within the same department, however, those employees might be unavailable, unwilling, or even unqualified in many scenarios since they share very little in common with the critical employee. Moreover, for scenarios like job rotation or cross training, the most suitable candidate should not have the exact same title as the critical employee. With this regard, it is important to analyze the similarity of job content between the one being planned and the potential candidate. Job title, a direct indication of what a job does, can be used for similarity analysis, but is often not sufficient. Detailed job descriptions and job entry qualifications, on the other hand can describe a job content in a more comprehensive and accurate way. To extract useful information from those job details in the form of natural texts to perform similarity analysis, text mining, should be applied.

For some scenarios such as filling the position of interest with suitable employees, recommender systems can be useful as a plus to the job content similarity analysis. Valuable recommendations are an important component in the natural process of human decision making (Melville & Sindhwani, 2010). Recommender system, such as software tools that provide suggestions for items to users, has been widely used in various areas such as YouTube videos, Netflix movies, Amazon products, etc. In the system, "items" usually refer to what the system recommends to the users (Ricci, Rokach, & Shapira, 2011). In the context of candidate recommendation, since we are recommending employees whose features fit to that of the critical position, the "items" would refer to "employees", and "users" to "positions". Word embedding, a Natural language processing (NLP) technique that maps words or phrases from natural texts to vectors of real numbers can be applied to job content texts to improve syntactic analysis, after which supervised learning models such as neural network, random forest, etc. can be trained to accomplish the recommendation work. Although such an approach can be potentially powerful in specific scenarios of recommending new employees to some positions, however, within the scope of this research work, only employees within the organization are considered as candidates for knowledge sharing. With this regard, the capability of recommender system cannot be fully used thus will not be studied in detail for this specific objective.

4.3.2 Data pre-processing

For almost all scientific works that deal with natural texts, data pre-processing is one of the most critical processes. With regard to text mining for job similarity analysis, before any of the analytics tools can be applied, features need to be exacted from the texts that describe job contents, which requires the texts in a uniform format to facilitate features extraction. Natural texts in the form of sentences must be processed since the sentence as read by humans is different from the text features used by the analytics tools (Sun, Liu, Hu, & Zhu, 2014).

Several steps need to be taken in data pre-processing prior to features extraction, which include: Checking misspelled words; Special informative terms extracting based on application domain; Removing stopping words, and removing uninformative words based on domain knowledge; Lowercasing, stemming and lemmatization; and splitting and tokenization as features to use (Jivani, 2011).

In the very first step, all misspelling cases should be checked and corrected to ensure the subsequent similarity analysis results in consistent and correct findings. Although there exist advanced algorithms for fast misspelling detection (Popescu & Vo, 2014), with regard to the scope of this work, which focuses on moderate size of texts, any misspelling detection tools such as MS Word could achieve satisfactory result. Then, there is the step of special informative terms extraction, which is largely based upon the application domain. Detailed examples and use cases for a specific domain are discussed in the next chapter. Such a step needs to be performed before the removal of any stopping words and uninformative words, since there can be uninformative words, such as under, with, etc. in the special terms. As for uninformative words, both general ones such as a, an, the, for, etc. and domain-specific ones, can be put into a library, after which text mining codes can be developed to remove the words that exist in the library from the sentence. The step of lowercasing, stemming and lemmatization concentrates on unifying words format. Stemming transforms all words that having the same stem to a common form and lemmatization removes inflectional endings of words such as -s, -es, -ed, -ing, etc. (Balakrishnan & Lloyd-Yemoh, 2014). These tasks can be achieved using open source NLP libraries such as Apache OpenNLP with Java. The very last step of data pre-processing is splitting and tokenization, where the sentence is split into words and phrases. These terms will eventually be used as features that describe and define the job content, for the purpose of text similarity analysis. In this step, tokenization can be realized by splitting the sentence into words (except for the special terms extracted in previous step). More advanced, NLP tokenization models can be applied to each sentence for splitting sentences into terms. The effectiveness of both approaches will be tested and evaluated. Moreover, Wordle can be created to visualize all the features (text terms) and their respective frequencies, as an additional procedure to ensure the tokenized terms, especially those with high frequency make sense in the domain of the application.

4.3.3 Similarity analysis

Text similarity measurement is one of the most critical components in many research and application areas including information retrieval, topic detection, summarization, etc. There are three widely known approach categories in the context of text similarity analysis: string-based approaches, corpus-based approaches and knowledge-based approaches (Gomaa & Fahmy, 2013). In the scope of this work, a subcategory of string-based approaches, specifically term-based similarity measurements, will be studied for measuring the similarity between pairwise of jobs based on the text of job description and entry qualification.

For a String-based approach, there are two main subcategories: character-based approaches, such as Longest Common Subsequence, N-gram, etc., and term-based approaches, such as, Jaccard Similarity, Cosine Similarity, Euclidean Similarity, etc. Since it makes more sense to evaluate based on terms in sentences rather than characters in sentences, term-based approaches will be considered and studied. Specifically, Jaccard Similarity and Cosine Similarity, the two most common methods will be implemented to test the results. Jaccard similarity measures the similarity between two texts by treating them as two sets of attributes and computing the intersection of the two sets divided by the union of the two sets. Cosine similarity measures the similarity between two texts by treating them as two vectors of attributes and computing the cosine of the angle between the two vectors. A value of 0 in cosine similarity indicates that there are no similarities between the two vectors (i.e. the angle between the two vectors is 90 degrees), while a value of 1 indicates that the two vectors are identical (i.e. the angle between the two vectors is 0 degrees). From another point of view, one of the main differences between Jaccard similarity and Cosine similarity relates to duplicate words in texts: Jaccard similarity only considers unique set of words for each term, while Cosine similarity considers the frequency of each term.

With pairwise text similarity evaluated using both Cosine method and Jaccard method, statistical study on the analysis result can be performed to compare the two methods: The value of the all pairwise similarity results (between 0 and 1) can be summarized and the distribution can be further visualized. Such a statistical study can be helpful to evaluate how the two methods perform on the text features in the domain of job descriptions.

4.3.4 Results and Conclusions

The text similarity analysis techniques studied above, including those ones for data pre-processing, are implemented to an example use case with a total 184 job titles from Georgia DOT, and results in 16836 pairwise comparisons. The texts that describe job contents come from two parts: job description and job entry qualification, which are retrieved from the Georgia Department of Administrative Services' website. For "Transportation Specialist 3" working title as an example, the job description is recorded as *"Under broad supervision, provides advanced level professional support in one or more of the following areas: OMAT, maintenance, construction, intermodal, planning, utilities, surveying, environmental or traffic. May also serve as a lead worker providing training to lower level staff."*, and the job entry qualification is recorded as *"Bachelor's degree from an accredited college or university in a related field AND One year of related experience OR Associate's degree in a related field from an accredited college or university AND Two years of related experience OR One year of experience required at the lower level Transportation Specialist 2 or position equivalent."*. The two parts for all other job titles are all recorded in a similar manner. Features extracted from both parts of texts will be used for similarity analysis, while features from the job entry qualification can further serve as criteria features for advanced filtering tasks.

For the texts in job descriptions and entry qualifications of all the job titles, the sentences are firstly pre-processed following the steps mentioned above, generating text terms. In the example of the job title Transportation Specialist 3 above, the text terms extracted from the job description include “under broad supervision”, “maintenance”, “construction”, “survey”, “traffic”, “lead”, etc., and the text terms extracted from the job entry qualification include “Bachelor’s degree”, “One year experience”, etc. These text terms are to be used as features for similarity analysis. Pre-trained NLP model was applied to job description and entry qualification for the step of splitting and tokenization for testing, in which case the uninformative words are not removed to prevent irregular sentence structure. However, the resulting terms are not stable enough for similarity analysis. For example, some terms are significantly long, and some long term features even contain other short term features. Therefore, NLP model is not to be used for tokenization in data pre-processing.

Pairwise similarities are computed with both the Cosine approach and the Jaccard approach, which is implemented in Java. Again, the job title of Transportation Specialist 3 is used as an example use case. The top 15 most similar job titles based on Cosine approach and Jaccard method, and the respective similarity value are listed in Table 2.

Table 2 – Top 15 job titles similar to “Transportation Specialist 3” based on both Cosine method and Jaccard method.

Top similar job titles based on Cosine method		Top similar job titles based on Jaccard method	
Job Title	Similarity	Job Title	Similarity
Transportation Specialist 4	0.91	Transportation Specialist 4	0.81
Transportation Specialist 5	0.84	Transportation Specialist 5	0.74
Transportation Specialist 1	0.75	Transportation Specialist 1	0.59
Transportation Specialist 2	0.75	Transportation Specialist 2	0.59
Transportation Tech 2	0.68	Transportation Tech 4	0.53
Transportation Tech 4	0.66	Mgr 2, Transport Specialist	0.51

Table 2 continued.

Top similar job titles based on Cosine method		Top similar job titles based on Jaccard method	
Job Title	Similarity	Job Title	Similarity
Mgr 2, Transport Specialist	0.65	Transportation Tech 2	0.5
Transportation Tech 1	0.64	Mgr 3, Transport Specialist	0.49
Mgr 3, Transport Specialist	0.62	Sr Mgr 1, Transport Specialist	0.49
Sr Mgr 1, Transport Specialist	0.62	Sr Mgr 2, Transport Specialist	0.49
Sr Mgr 2, Transport Specialist	0.62	Mgr 1, Transport Specialist	0.47
Transportation Tech 3	0.62	Transportation Tech 3	0.47
Env Transportation Spec 3	0.61	Transportation Tech 1	0.46
Mgr 1, Transport Specialist	0.61	Env Transportation Spec 3	0.39
Env Transportation Analyst 2	0.56	Transport Planning Spec 3	0.37

A statistical study is then conducted to the result for both evaluation methods. Figure 4.8 shows the distribution of pairwise similarity for the 16836 pairs of job descriptions with Cosine method and Jaccard method. Moreover, a visualization is created in Java to visualize the similarities between each pair of jobs. Figure 4.9 shows the pairwise similarities of all the job titles, where a pair of jobs are only connected if their respective similarity (Cosine and Jaccard) is greater than 60%.

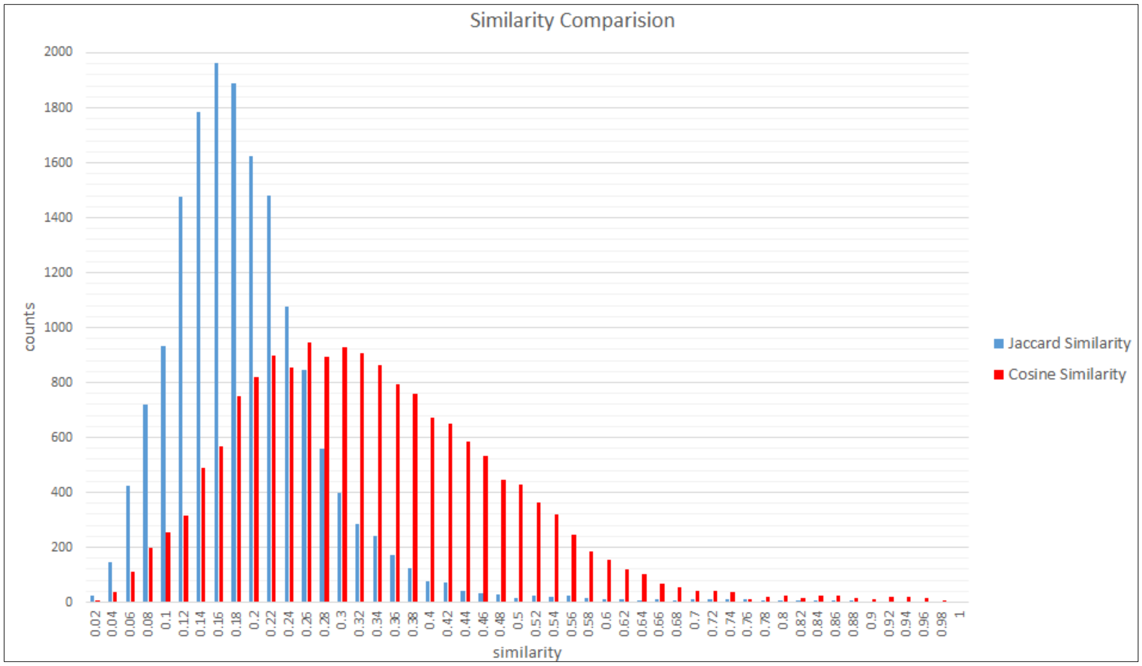


Figure 4.8 – The distribution of pairwise similarity with Cosine method and Jaccard method.

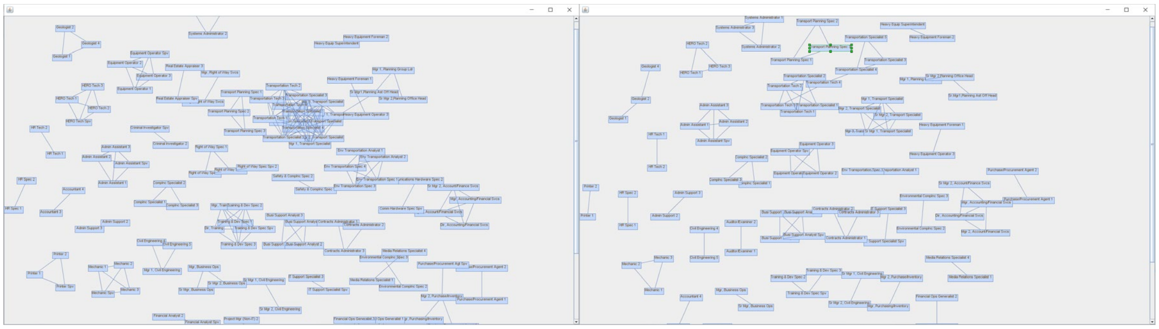


Figure 4.9 – Network visualization of the connectivity between jobs. A pair of jobs are connected if the similarity (left: Cosine; right: Jaccard) is larger than 0.6.

From both figures above, as well as the table that showing the top similar job titles for the example one, it can be seen that although the two evaluation approaches result in different similarity values, they have similar trends, i.e. the comparative similarity, or similarity ranking are following a similar pattern. Moreover, it is a fact that Cosine similarity, which account for the frequency of each features in the sentences, tends to evaluate a pair of jobs more similar in terms of the similarity value, due to which the statistical bar chart for Cosine similarity results in a less steep curve compared to the bars for Jaccard similarity.

4.4 Multivariate Analysis

4.4.1 Overview

Multivariate Analysis (MVA), originally intended to perform statistical study over more than one statistic variants, has evolved and been applied in many fields like multivariate hypothesis testing, multivariate regression, measurements of relationship, variable selection, etc. (Olkin & Sampson, 2001). In the real world, most systems are defined by multivariate data rather than univariate date. MVA can be particularly useful to extract important information from the data to understand system mechanism, as well to find relationships to model variants and response. With regard to human resource management in organizations, an employee database is such a system that is characterized by multivariate data. For example, each item in the database describes an employee from multiple aspects including age, license, tenure, etc. Therefore, MVA can be applied to better understand the metrics in the database, as well to assist human resource personnel with decision making.

In the context of knowledge retention, several data tools have been studied in the previous sections, as well many measurements have been mentioned that can be useful in various assessments, such as the assessment of knowledge criticality, risk of knowledge being lost, candidate suitability, etc. However, these data tools cannot perform properly, nor can the evaluations be measured comprehensively without the application of MVA. For example, we have introduced the idea of knowledge loss risk (KLR). In the evaluation of KLR, the sole use of the computed "years to retirement" cannot evaluate the metric comprehensively since other factors such as the uniqueness of the knowledge, or the available resource need to be considered as well. Similarly, in the evaluation of knowledge criticality, the only use of the computed centrality values from the network cannot account for the results comprehensively while other parameters such as the uniqueness of the knowledge, or the impact level of the knowledge being absent among others are also playing important roles. Moreover, in the evaluation of mentor or candidate suitability, it is not enough to only consider the job content similarity, but rather other factors like past experience, willingness, leadership, etc. should also be taken into account. These are the scenarios where the use of MVA is beneficial.

4.4.2 Methodology

As mentioned above, important decision making requires the consideration of more than one factor, and this is referred as multi-criteria decision making (MCDM) or multi-criteria decision analysis (MCDA). The weights of different factors are usually most critical to the correctness of the resulting decision (Roszkowska, 2013). With regard to human resource decision making within the scope of the study, all the assessments depend on multiple factors. It is obvious that there is no such specific formula to precisely calculate the assessment result from the impact factors. But rather, these factors are not equal in the degrees of contribution to the determination of the resulting decision, thus should have various weights according to their respective ranking on the relative importance. For the consideration of the tradeoff between the applicability and the performance of the methodology on MCDM, rank-based approaches that convert such relative importance-based ranking into respective numerical weights are shown to be most effective (Sureeyatanapas, 2016). And there are many different rank-based approaches that could achieve this goal, for example, rank sum (RS), rank exponent (RE), rank reciprocal (RR), and rank order centroid (ROC). Many studies have shown that ROC, due to its steepness and non-linear function to the weights being consistent to the process of decision making behavior, tend to result in the highest performance among most of the rank-based approaches (Ahn, 2011; Sureeyatanapas, 2016).

ROC calculates the weight of each factor using equation (1), and then the total score of the outcoming result is determined by equation (2). One of the features of ROC is that it can account for two or more factors at the same level of relative importance by ranking them at the lowest level, which will be illustrated by the below example.

$$W_i = \left(\frac{1}{n}\right) \times \sum_{j=i}^{n} \left(\frac{1}{j}\right) \tag{1}$$

$$S_t = n \times \sum_{i=1}^{n} (W_i \times S_i) \tag{2}$$

The following example shows the detailed procedure of using ROC to calculate the total score from 4 factors. Denote the rank of ith factor as R_i, and the score of ith factor as S_i, which is a natural number from 1 to 3.

For example, the rank of each factor is: $R_1 = 1$; $R_2 = 3$; $R_3 = 3$; $R_4 = 4$ (the second factor and the third factor have the same level of relative importance). The weights of each factor W_i will be calculated as followed:

$$W_1 = \left(\frac{1}{1} + \frac{1}{3} + \frac{1}{3} + \frac{1}{4}\right) \times \left(\frac{1}{4}\right) = 0.479$$

$$W_2 = \left(\frac{1}{3} + \frac{1}{3} + \frac{1}{4}\right) \times \left(\frac{1}{4}\right) = 0.229$$

$$W_3 = \left(\frac{1}{3} + \frac{1}{3} + \frac{1}{4}\right) \times \left(\frac{1}{4}\right) = 0.229$$

$$W_4 = \left(\frac{1}{4}\right) \times \left(\frac{1}{4}\right) = 0.063$$

For example, the score of each factor is: $S_1 = 3$; $S_2 = 3$; $S_3 = 2$; $S_4 = 3$. Then the total score S_t will be calculated as followed:

$$S_t = 4 \times (0.479 \times 3 + 0.229 \times 3 + 0.229 \times 2 + 0.063 \times 3) = 11.1$$

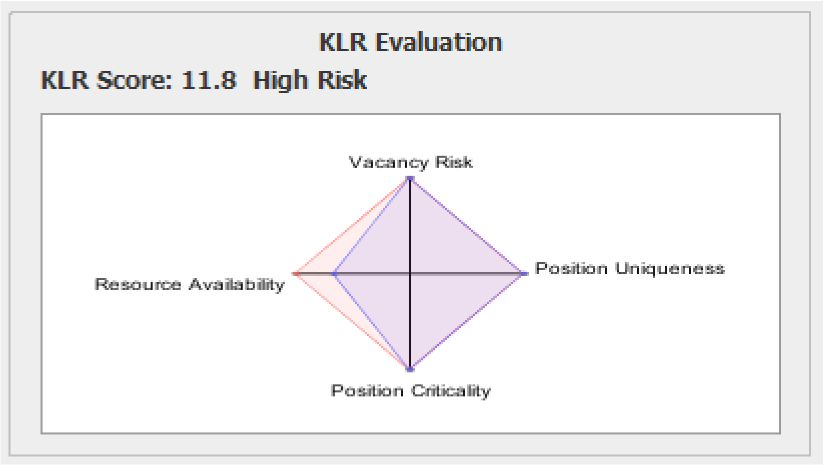
With regard to the assessments in the context of knowledge management, each impact factor can be firstly mapped to a numerical value, for example 1 to 3, to indicate the factor score. The relative importance ranking of each score can be determined through domain knowledge within the organization. Then ROC can be applied to calculate the weight of each factor, after which the total score can be calculated using the weighted summation. With this approach, assessments including KLR, knowledge criticality, mentor / candidate suitability can be evaluated quantitatively, which can further assist with human resource decision making.

4.4.3 Results and Conclusions

An example of the KLR evaluation will be used here to illustrate the application of ROC. In the evaluation of KLR, a total of four factors need to be taken into consideration, including employee vacancy risk (years until retirement), position criticality (determined from network analysis), position uniqueness (determined from text mining), and resource availability (requires user input). The evaluation of each factor is mapped to a numerical score from 1 to 3 with 1 indicating low, 2 indicating medium and 3 indicating high. For example, one employee has less than one year until retirement, then the vacancy risk is mapped to the score of 3, i.e. high risk; Their centrality value is higher than 90% of all the positions in the department, then the position criticality is mapped to the score of 3, i.e. high criticality; Their position job content is similar to a small number of other positions, then the position uniqueness is mapped to the score of 3, i.e. very unique; Their resource availability is set to 2, medium availability. Moreover, according to the domain expert, the vacancy risk is the most important factor in determining KLR (rank = 1), followed by the position criticality and uniqueness (both rank = 3), and finally the factor of resource availability (rank = 4). Therefore, the total score can be calculated as followed, where a score of 11.8 indicates a very high KLR, thus requiring immediate attention.

$$S_{KLR} = 4 \times (0.479 \times 3 + 0.229 \times 3 + 0.229 \times 3 + 0.063 \times 2) = 11.8$$

In addition to the application of ROC, the use of radar charts (or spider web charts) is also an approach to visually interpret the multivariate data analysis by intuitively showing all quantitative variables at the same time in the form of a two-dimensional chart with axes starting from the same point. Figure 4.10 shows the score and radar chart of the KLR evaluation example. Each of the four variables' scores is projected to one of the axes in the radar chart visualization.



KLR Evaluation

KLR Score: 11.8 High Risk

Figure 4.10 – Use radar chart to visualize the multiple factors scores of a KLR evaluation example.

As studied in this section, MVA can be especially useful for decision making in organization knowledge retention related activities which falls into MCDM. Rather than traditional decision making approaches implemented by human resource in most public organizations, MVA can quantitatively evaluate multiple factors according to their relative importance and domain scoring system, resulting in a scientific and accurate evaluation result. The combination of ROC and radar chart visualization further allows for intuitive enhancement that assists user decision making procedures.

4.5 Temporal Analysis

4.5.1 Overview

Temporal analysis allows for analyzing the behavioral parameters of the system under investigation (SUI) over time. As a commonly used statistical analytic tool, temporal analysis has been widely applied in various fields of multiple industries that have time-related data analysis. For example, a typical use case is in the field of big data for healthcare where temporal analysis can be applied on patients' vital data over time, for meaningful results and patterns to be derived. More commonly, temporal analysis is sometimes combined with spatial analysis for a more complex analytic. For example, criminal activity data can be analyzed by temporal approach for pattern recognition over hours of a day, or days of a week, as well with geographical datapoints, which then allows for prediction that can be used by law enforcement agencies for tactical optimization (Brantingham & Brantingham, 1984). Another example is the check-in function of many social network applications, where spatial-temporal analysis can be applied to the past check-in records with timestamp and location information for human behavioral patterns analysis as well as for future check-in predictions (Bannur & Alonso, 2014).

In the real world, almost any system evolves over time, and analyzing the important parameters that defining or describing the system over temporal elapse provides significant insights for monitoring and evaluating the system behavior. Especially, temporal analysis can be applied for protocol evaluation by analyzing system parameters and how they change before and after the deployment of a protocol. With regard to organization knowledge management, many factors can be potentially useful in assessing how well an organization is handling knowledge retention, such as the average turnover rate, employees' average knowledge loss risk (KLR), even employee engagement level, etc. These parameters can be further analyzed by temporal method, for example measured at a fixed time interval, say every quarter of a year, and analyzed on a time series basis after the implementation of the proposed human resource data and knowledge management framework. The resulting effectiveness of the framework or system can then be evaluated. More specifically, temporal analysis allows the system users to keep monitoring those parameters in the organization, which can in turn reveal a trend over time. Although market environment and other external factors could have potential impact on the workforce statistics, these parameters can still yield a significant indication on the effectiveness of the implemented system.

4.5.2 Methodology

Many approaches can be incorporated for analyzing data with temporal features. For example, the basic linear regression analysis, which models the relationship between a response parameter with one or more variables (explanatory variables), and the date and time here can be one of the explanatory variables for temporal analysis. More general statistical method such as ANOVA can also be utilized for purposes like temporal correlation analysis, time related patterns identification, etc. For a more straightforward temporal analysis such as the one for knowledge retention framework evaluation, which involves few responding parameters with time (or date) as only explanatory variable, a more intuitive visualization approach can be potentially more suitable for such an evaluation task. data visualization (line chart, heat map, etc.). Figure 4.11 is a line chart that mocks the visualization of how average employee years of tenure and average KLR level would be expected to change over time, resulting from the deployment of the innovative knowledge retention protocol. Figure 4.12 represents another form of data visualization, heat map, which mocks the changes on the number of employees with KLR higher than a certain percentile amongst them. Both types of visualization have their own benefits and drawbacks which can then be traded off by the user.

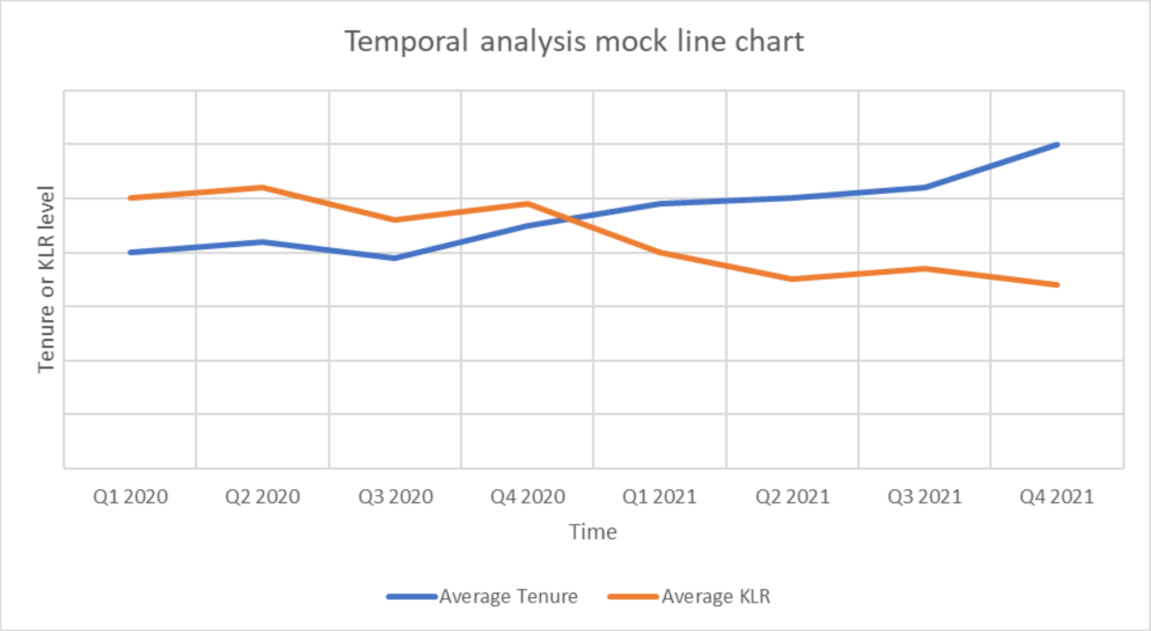


Figure 4.11 – Mock line chart plotting average turnover rate and average KLR change over time.

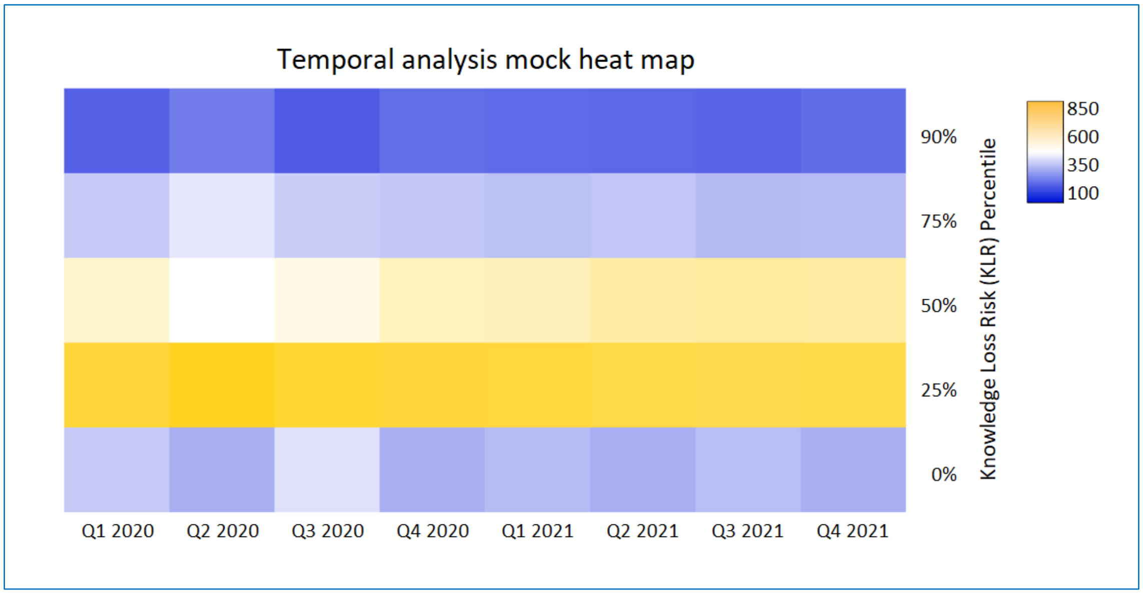


Figure 4.12 – Mock heat map showing the change of each part of the average KLR (those with KLR >= 90% percentile, 75% percentile, etc.) over time.

Temporal analysis, especially combined with data visualization, can be a powerful tool for system behavior assessment. However, such an approach requires enough data points over a long period of system monitoring, especially for such descriptive metrics in the context of knowledge management. While Chapter 6 will discuss system evaluation with use cases, this approach of temporal analysis can be an effective evaluation method for long-term use.

4.6 Concluding Remarks

With the development of computational ability, many data science techniques and tools can be applied to the development of an integrated, automated and scalable human resource data management system. This chapter has studied some of the tools that can be useful to such a system.

Geographical information visualization can enhance user perception and interaction with the human resource database that can assist with visualization of attrition risk distribution across different geolocations of the whole organization, and with identification of knowledge most at risk of being lost. Network analysis on the human resource database using who reports to whom relationship can analyze the positions in organization as a connected undirected graph and reveal the information flow patterns amongst them that allow for the identification of critical positions with high centralities. Moreover, various centrality assessment approaches can have different emphasis that evaluate not only strategical but also tactical criticality to the organization. Text mining techniques analyze job descriptions and entry qualifications in a format of natural texts regarding the job content similarity and allow for the recognition of the employees whose daily work are most similar to the critical position of interest. Multivariate analysis allows for comprehensive consideration of all the impact factors in multi-criteria decision making by calculating factor weights using rank order centroid approach. In the scope of this chapter, MVA can integrate evaluation results from other computational tools mentioned above, to generate quantitative assessment results to further help with optimal knowledge transfer.

With the study of the data science tools that can be potentially helpful on the development of the integrated system for human resource data management and knowledge retention, the next chapter will study the detailed design and implementation of this computational system using real world database from Georgia DOT as a case study.

CHAPTER 5. INTEGRATED COMPUTATIONAL SYSTEM

This chapter introduces the integrated human resource data management system that was developed with computational tools and data science techniques studied in Chapter 4. The system is built with the human resource database from Georgia DOT, and is to be used and evaluated with Georgia DOT human resource department personnel. The context in this chapter will present the overall system structure, details of each functional module, as well as the application user interface design.

5.1 System Design Overview

Human Resource Data Tools (HRDT) is an integrated computational system that is developed for an organization's HR departments to visualize HR database, to manage organizational knowledge, as well as to support decision making. The ultimate objective is to integrate the system in most of the public sector organizations, serving as a useful framework for knowledge management and retention. As a large portion of this research work is funded by Georgia DOT, and the HR database used in the development of this system comes from Georgia DOT with approximately 4000 employees across the state of Georgia, the current version of the HRDT system is built and optimized for HR department in Georgia DOT's benefit.

As is similar to the development of any computational systems, the main programming language must be chosen, where multiple factors need to be taken into consideration. First of all, the most important factor needs to be considered is information security. The system is designed to query from the organization's HR database directly, which contains a lot of sensitive personal information such as social security, birthday, salary, etc., and should never allow unauthorized access. Especially considering the users will be government personnel, security concerns should be put as the top priority. Due to the vulnerable security nature of web application, web programming should not be used for this system, rather desktop application will be the format of the HRDT system. Another consideration is the ease of development and future maintenance. Although HRDT is designed to interact with, and only with the HR database from the organization as the data source, the development simplicity and the ease of expanding functionalities for future maintenance is still playing an important role in the design of any systems. Moreover, as HRDT will support friendly user interface, the programming language is desired to have the graphic user interface development capability. With considering the various factors mentioned above, Java, a general-purpose programming language that supports graphic user interface design with both native and third-party libraries, was chosen as the main language. Furthermore, Java is also a cross-platform compatible programming language that provides extra versatility for various users with different operating systems.

HRDT is designed as a desktop application that communicates with the organization’s HR database on the backend, and interacts with users’ operations on the front end, in a way that the system serves as an interface between the users and the tacit organizational knowledge pool. The structure of HRDT mainly constitutes computation engines and functional modules, where the computation engines perform the database queries, data manipulations, graph analyses, etc. on the backend, and each of the functional modules involves a workflow of data tools designed for knowledge management, as well as actual graphic interactions with users on the frontend.

Namely there are a total of four computation engines: attrition analysis engine, network analysis engine, text mining engine, and multivariate analysis engine. There are three functional modules in total: attrition distribution analysis and visualization, network analysis and visualization, and comprehensive knowledge management modules which contains five sub modules. Each of the functional modules or sub modules is driven by one or more of engines. Figure 5.1 shows the overall structural design of the proposed HRDT system, where green boxes represent the engines on the backend and blue boxes represent the functional modules on the front end.

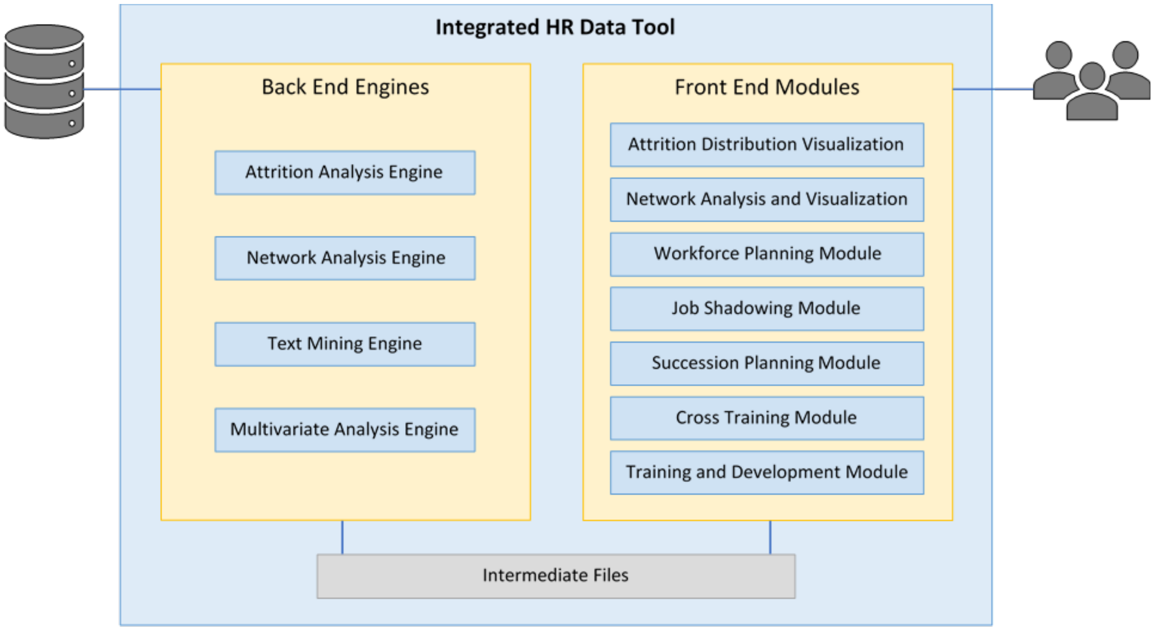

Integrated HR Data Tool
Back End Engines
Attrition Analysis Engine
Network Analysis Engine
Text Mining Engine
Multivariate Analysis Engine
Front End Modules
Attrition Distribution Visualization
Network Analysis and Visualization
Workforce Planning Module
Job Shadowing Module
Succession Planning Module
Cross Training Module
Training and Development Module
Intermediate Files

Figure 5.1 – The overall structural design of the HRDT system.

Upon startup of the system, computation engines will perform their respective tasks and generate any necessary intermediate files for use by the functional modules. Figure 5.2 shows a screenshot of the system landing page. Each clickable button will take users to the corresponding functional modules that will be introduced later.

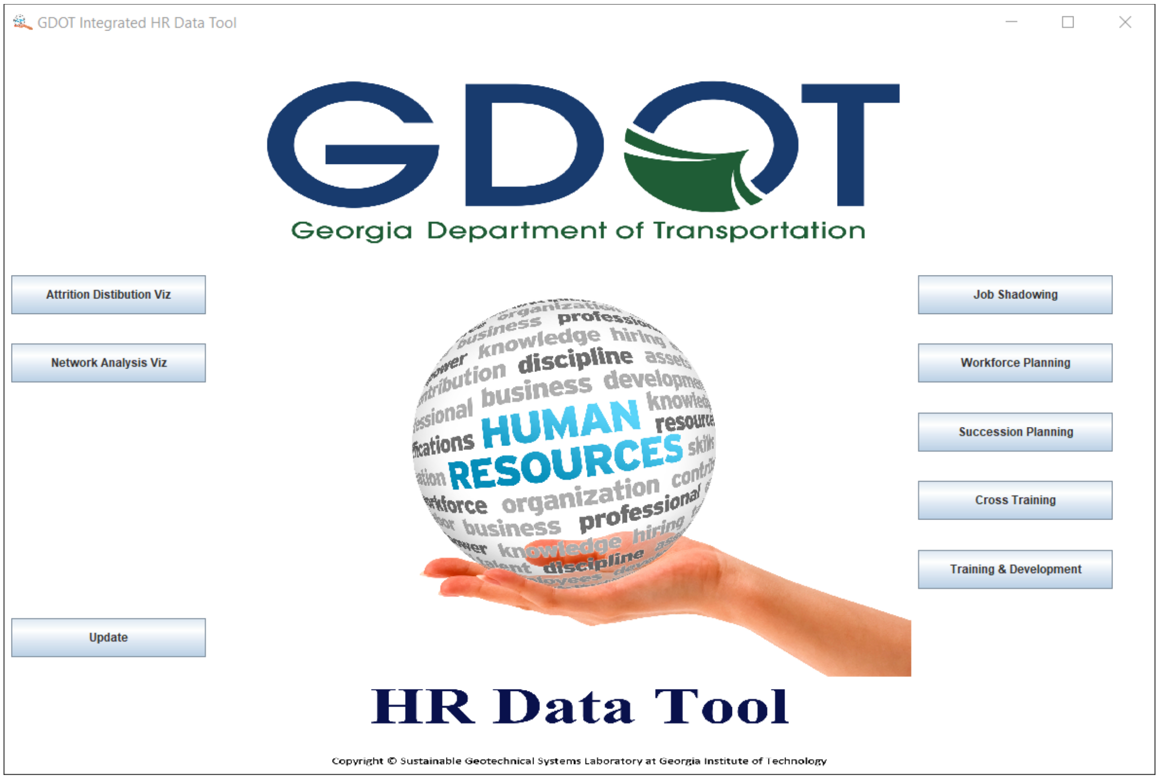

GDOT Integrated HR Data Tool
GDOT
Georgia Department of Transportation
Attrition Distibution Viz
Network Analysis Viz
Update
Job Shadowing
Workforce Planning
Succession Planning
Cross Training
Training & Development
HUMAN RESOURCES
HR Data Tool
Copyright © Sustainable Geotechnical Systems Laboratory at Georgia Institute of Technology

Figure 5.2 – Landing page of the HRDT system.

5.2 Computation Engines

This section focuses on the mechanisms of the four backend computation engines. The design and workflow of each functional module or sub module, as well as the graphic user interface will be introduced in the following sections of this chapter.

As mentioned in the design overview, computation engines are backend services that communicate with the HR database in Georgia DOT. Moreover, HRDT is designed in the way that the backend services communicate only with the HR database, such that the data resource for the system is only the HR database rather than any other resources. On one hand, HRDT's target users are Georgia DOT human resource personnel or managers, and this design makes it easier and simpler for users to deploy and use the system on a regular basis, since they don't have to deal with any protocols or request any additional information from other departments. On the other hand, using only HR database as the data resource also makes the development and any future maintenance easier and more straightforward.

5.2.1 Attrition Analysis Engine

The main objective of the attrition analysis engine is to extract all relative information of employees from Georgia DOT’s HR database, and make pivoting-related transformation and analysis to it such that the information is in a proper format that is ready for filtering, ranking, and other pivoting analyses. The analysis results will be mainly used by the functional module of attrition distribution visualization, while the knowledge management modules as well as other computation engines will use part of the transformed results for evaluation and calculation. Considering the detailed functionalities of the modules, various information needs to be extracted: besides employees’ basic information, such as id, name, gender, age, etc., other work-related information is also needed, including working department, job title, starting date, technical license, etc. After the required information is extracted from the database, it will then be saved into a temporary table for further text mining and transformation.

In the process of text mining and data transformation, one of the most important features that has to be acquired is the physical location of the employee as it is not explicitly given, which will be critical in the module of attrition distribution visualization for the purpose of displaying employee information with geographical related feature. In order to acquire the physical location of an employee, we first need to apply text search on the working department of the employee to extract the district and area code, which is used by Georgia DOT to index the branch offices across the state. Then, we have to map the extracted district and area codes into actual physical locations. Table 3 shows the address information of all the 39 branch offices of Georgia DOT with district and area code.

Table 3 – Branch office location of Georgia DOT by district and area code.

District-Area Code	Name	Address
0-0	GDOT Headquarters	600 West Peachtree NW, Atlanta, GA 30308
1-0	Gainesville (Main)	2505 Athens Hwy SE, Gainesville, GA 30507
1-1	Gainesville (Area 1)	2594 Gillsville Hwy, Gainesville, GA 30507
1-2	Athens	450 Old Hull Rd, Athens, GA 30601
1-3	Carnesville	301 Conger Rd, Carnesville, GA 30521
1-4	Cleveland	942 Albert Reid Rd, Cleveland, GA 30528
2-0	Tenille (Main)	643 Hwy 15 S, Tennille, GA 31089
2-1	Milledgeville (Area 1)	161 Blandy Rd, Milledgeville, GA 31061
2-2	Dublin	2003 US Hwy 441 S, Dublin, GA 31021
2-3	Louisville	2791 US Hwy 1 N, Louisville, GA 30434
2-4	Augusta	4260 Frontage Rd, Augusta, GA 30909
2-5	Madison	1570 Bethany Rd, Madison, GA 30650
3-0	Thomaston (Main)	115 Transportation Blvd, Thomaston, GA 30286
3-1	Thomaston (Area 1)	101 Transportation Blvd, Thomaston, GA 30286
3-2	Columbus	3600 Schatulga Rd, Columbus, GA 31907
3-3	Perry	200 Julianne St, Perry, GA 31069
3-4	Macon	4499 Riverside Dr, Macon, GA 31210
3-5	LaGrange	1107 Hogansville Rd, LaGrange, GA 30241
4-0	Tifton (Main)	710 West 2nd St, Tifton, GA 31794
4-1	Valdosta (Area 1)	1411 Madison Hwy, Valdosta, GA 31601
4-2	Douglas	1835 S Peterson Ave, Douglas, GA 31535
4-3	Donalsonville	734 W Crawford St, Donalsonville, GA 39845
4-4	Moultrie	120 Veterans Pkwy N, Moultrie, GA 31788
4-5	Albany	2060 Newton Rd, Albany, GA 31701
5-0	Jesup (Main)	204 North Highway 301, Jesup, GA 31546
5-1	Baxley (Area 1)	740 Oakdale Cir, Baxley, GA 31513
5-2	Waycross	104 N Nichols St, Waycross, GA 31502
5-3	Brunswick	128 Public Safety Blvd, Brunswick, GA 31525
5-4	Statesboro	17213 US Hwy 301 N, Statesboro, GA 30458
5-5	Savannah	630 West Boundary St, Savannah, GA 31401
6-0	Cartersville (Main)	500 Joe Frank Harris Pkwy, Cartersville, GA 30120
6-1	Cartersville (Area 1)	874 Peeples Valley Rd NW, Cartersville, GA 30120
6-2	Dalton	1313 North Tibbs Rd, Dalton, GA 30720
6-3	Buchanan	4323 US Hwy 27, Buchanan, GA 30113
6-4	Rome	533 East 20th St, Rome, GA 30161
7-0	Chamblee (Main)	5025 New Peachtree Rd, Chamblee, GA 30341
7-1	Chamblee (Area 1)	5025 New Peachtree Rd, Chamblee, GA 30341
7-2	Marietta	1296 Kennestone Cir, Marietta, GA 30066
7-3	College Park	4125 Roosevelt Hwy, College Park, GA 30349

Another important factor for the functional module is how many years are expected before the retirement of each employee. This feature can be calculated using a formula provided by Georgia DOT which considers both employee's age and years of tenure at current position, and obviously the tenure can be easily calculate from the job entry date that was previously extracted. Basically, an employee is considered approaching retirement when either they have stayed at Georgia DOT for 30 years, or they have reached the age limit. With this feature obtained, every employee can then be tagged by the category of years to retire, for example, less than 0 year (which mean that the employee should have already retired), 0 to 1 years, 1 to 3 years, etc. And this feature can be critical in the determination of the vacancy risk and the evaluation of KLR in other parts of HRDT. Moreover, whether the job title of an employee falls into the special categories of foreman, superintendent, supervisor, manager, or director, information can be extracted from the job title and saved into the temporary table, which will be one of the contributing factors to the evaluation of position criticality and KLR.

Besides the features mentioned, others such as whether an employee holds Professional Engineer (PE) license, Engineer in Training (EIT) license, etc. are also extracted and stored. The temporary table can then be saved into the file system for use by the functional modules and computation engines, such that the analysis processes do not have to be re-performed every time users start HRDT. However, for the situation where the organization has experienced workforce changes, due to either new employees' hiring or existing employees' departure, it will be taken care by the update function of HRDT which will be introduced later.

5.2.2 Network Analysis Engine

The main mission of the network analysis engine is to use the “who reports to whom” relationship to establish an undirected graph as mentioned in Section 4.2, as well as to derive multiple important metrics from the graph. The graph and the relating metrics will be mainly used by the functional module of the network analysis and visualization, while the knowledge management modules and the multivariate analysis engine will also use part of the derived metrics for assessments and analyses.

Before going into the details of the network analysis engine, there is the important consideration that different departments in an organization usually have different sizes in term of workforce and are playing different roles to the organization. Therefore, in addition to the large graph of the employees across the whole organization, sub graphs can be established for each departmental level for local metrics evaluation. In the context of the HRDT system, Georgia DOT consists of 56 departmental units. Table 4 summarizes each department name obtained from the “Master Organizational Chart” from Georgia DOT.

Table 4 – Departmental units of Georgia DOT.

Department Name	Department Name
Office of Planning	Office of Utilities
Office of Human Resources	Office of Traffic Operations
Office of Legal Services	Office of Maintenance
Office of Equal Employment Opportunity (EEO)	Office of Innovative Delivery
Office of Strategic Communications	Office of Program Delivery
Office of Procurement	Office of Program Control
Information Technology	Office of Engineering Services
Office of Application Support	Office of Transportation Investment Act (TIA)
Office of Infrastructure	Office of Performance-Based Management
Office of Local Grants	Office of Budget Services
District 1	Office of Financial Management
District 2	Office of General Accounting
District 3	General Counsel Division of Administration

Table 4 continued.

Department Name	Department Name
District 4	Division of Local Grants
District 5	Division of Engineering
District 6	Division of Intermodal
District 7	Division of Construction
Office of Equipment and Facilities Management	Division of Permits and Operations
Office of Environmental Services	Division of Public-Private Partnerships (P3)
Office of Roadway Design	Program Delivery
Office of Bridge Design	Division of Finance
Office of Right of Way	Deputy Commissioner
Office of Design Policy and Support	Chief Engineer
Office of Intermodal	Treasurer
Office of Materials and Testing	Division of Planning
Office of Construction	Office of Audits
Office of Bidding Administration	Government and Legislative Relations
Office of Transportation Data	Commissioner

With all the departmental units configured, each department chair can then be used as the starting point to build the topological graph of the employees in local departments. Therefore, using the extracted "reporting to" information and the Breadth First Search (BFS) algorithm, employee relationship graphs can then be established both at organization level and at department level. Moreover, during the process of traversing the employee graphs with BFS, employee levels (both globally and locally) can be calculated and saved for use by the following contents as well. After all graphs have been successfully established, other graph-oriented metrics can then be computed.

Employee Uniqueness Assessment

As noted, employee uniqueness can be defined at two different levels: (i) Local, meaning it is assessed at the departmental unit level (for each of the 58 departmental units considered), and (ii) Global, meaning it is assessed at the level of the entire organization. Furthermore, employee uniqueness can be assessed at two different operational levels: (i) at the basic level, which uses the working title as a simple indicator, or (ii) at the network level, that uses both the working title and the position level of an employee. In this regard, there are a total of four different possible definitions of employee uniqueness that can be given as follow:

- Basic local uniqueness: the inverse of the number of employees with the same working title in the same departmental unit. For example, if there are five employees in a given departmental unit with the title "Engineer", the basic local uniqueness of Engineer in that departmental unit can be calculated as 1/5.
- Network local uniqueness: the inverse of the number of employees with the same working title AND the same position level in the same departmental unit. For example, if there are five employees in a given departmental unit with the title "Engineer", and three of them are at a "Junior" position level, the network local uniqueness of Junior Engineer can be calculated as 1/3.
- Basic global uniqueness: the inverse of the number of employees with the same working title in the entire organization. For example, if there are 10 employees in the entire organization with the title "Accountant", the basic global uniqueness of Accountant in the entire organization can be calculated as 1/10.
- Network global uniqueness: the inverse of the number of employees with the same working title AND the same position level in the entire organization. For example, if there are 10 employees in the entire organization with the title "Accountant", and five of them are at a "Senior" position level, the network global uniqueness of Senior Accountant in the entire organization can be calculated as 1/5.

The employee uniqueness rating is defined in a way based upon the discussion with the stakeholder of the HRDT system at Georgia DOT. It is assumed that if uniqueness is less than a value of 0.25 (i.e., greater than 1 in 4), there is enough redundancy for the employee to be considered non-unique and hence assigned a uniqueness rating of 1. If uniqueness is greater than or equal to 0.25 but less than 0.5 (i.e., greater than 1 in 2), the employee is assigned a uniqueness rating of 2. Lastly, if uniqueness is 0.5 or greater, the employee is assigned a uniqueness rating of 3, indicating the employee is unique.

Employee Criticality Assessment

As mentioned in the previous chapter, one of the most important graph-based metrics for HRDT is the centrality, which will be used for employee criticality assessment. While it is true that such an assessment approach is more rigorous and comprehensive, a simple first-order assessment of criticality of an employee can also be made based solely on the information contained in his/her job title. The basic assumption is that employees classified as Director, Manager or Supervisor have a higher level of criticality. Employees classified as Foreman and Superintendent are considered critical if the person has been at the job for at least two years. This is a naïve but functional method for criticality assessment, which will be referred to as Basic Criticality (as oppose to Network Criticality in the following paragraphs). Users will be given the option to choose this simple assessment approach. In this regard, the following criteria are used to evaluate Basic Criticality:

- Basic Criticality = 3 if “Working Title” contains any of the following:
  Dir [Director], Mgr [Manager], Spv [Supervisor]
- Basic Criticality = 3 if “Working Title” contains any of the following:

Foreman, Superintendent AND tenure at current position ⩾ 2 years

- Basic Criticality = 2 if "Working Title" contains any of the following: Foreman, Superintendent
- Basic Criticality = 1 for all else

Network centrality-based criticality assessment, as a more rigorous approach, can then be developed by evaluating node centrality in the graph. This approach will be referred to as Network Criticality. PageRank centrality, which was briefly introduced in Section 4.2, with tactical importance emphasis, will be applied here. The general equation for the Page Rank scoring can be expressed as equation (3)

$$PR_m = \sum_{n \in E_m} \frac{PR(n)}{L(n)} \tag{3}$$

This equation shows that Page Rank (PR) for the employee m is dependent on PR for any employee n in the set E_m, and the number of outgoing edges L for employee n. This set contains all employees connected to employee m. For example, assuming there are 4 employees (A, B, C, and D) in a company, initially all employees will get the same value of PR, which in this example will be 0.25 (considering a probability distribution between 0 and 1). PageRank algorithm is an iterative process. The PR value of an employee will be equally transferred to the neighbor employees upon a new iteration. For example, if all employees B, C, and D report to employee A (Employees B, C and D are not connected to each other), the PR for Employee A would be:

$$PR_A = PR_B + PR_C + PR_D = 0.75$$

On the other hand, if Employees A, B, C, and D have the structure as shown in Figure 5.3:

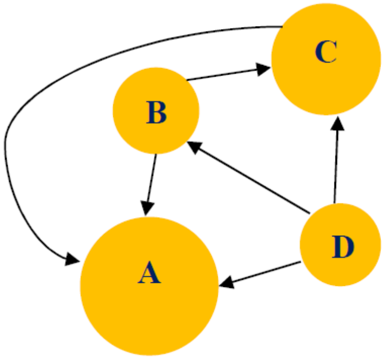


Figure 5.3 – Example case for PageRank calculation.

Then, the PR for each employee upon first iteration would be:

$$PR_A = \frac{PR_B}{2} + \frac{PR_C}{1} + \frac{PR_D}{3} = 0.458$$

$$PR_B = \frac{PR_A}{3} + \frac{PR_D}{3} = 0.167$$

$$PR_C = \frac{PR_A}{3} + \frac{PR_B}{2} + \frac{PR_D}{3} = 0.292$$

$$PR_D = \frac{PR_A}{3} = 0.083$$

In the above example, Employee A is considered as a “dangling node” and its effects are equally distributed to other employee nodes in order to have a probability distribution definition over the results of PR values for all the employees. In order to avoid the occurrence of any sinks (dangling nodes), a damping factor is added to the general equation as shown in the new equation (4), where N is the total number of the nodes and d is the damping factor which is usually taken as 0.85.

$$PR_m = \frac{1-d}{N} \sum_{n \in E_m} \frac{PR(n)}{L(n)} \tag{4}$$

Furthermore, employee uniqueness is also a contributing factor to employee criticality assessment, thus should be put into the formula. With this regard, we bring another network-based evaluation to this Network Criticality to make it more comprehensive, that is the cosine similarity. Equation (5) show the general expression:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5}$$

where, A_i and B_i are attributes of vector A (Employee A) and vector B (Employee B). A cosine similarity with zero value represents orthogonality, decorrelation or independency of data meaning that two vectors are not similar to each other. Alternatively, a cosine similarity of 1.0 indicates that two vectors are exactly the same.

To assess criticality, cosine similarity calculations were performed considering both the employee uniqueness (across four dimensions of uniqueness as defined in previous section) as well as Page Rank centrality performed at both the departmental unit (local) level and the organizational (global) level. In this approach, six different indicators are considered in order to measure distance similarity among different instances containing more quantitative information. As different features in one instance can have different weights and scales, data over each feature is normalized to avoid any type of bias toward a specific feature in a vector. Lastly, because employees with higher independence relative to their peers can be assumed to have higher criticality, the calculated cosine similarity value was subtracted from unity to calculate “cosine dissimilarity” in order to identify the employee with higher scores, and therefore, higher criticality.

Based on an evaluation of the histogram of the calculated values, it was assumed that employees with a cosine dissimilarity value less than that corresponding to the 70th percentile are considered non-critical, and hence assigned a criticality rating of 1. Those employees with cosine dissimilarity values corresponding to between the 70th and 90th percentiles are assigned a criticality rating of 2. Lastly, if the cosine dissimilarity value falls into the 90th percentile or above, the employee is assigned a criticality rating of 3, indicating the employee is critical.

Furthermore, we introduce a concept to measure how much impact an employee is expected to bring if he or she is absent, referred to as Absence Impact. The graph metric of betweenness centrality, which was introduced in section 4.2, is used here to quantify absent impact of an employee. The reason is that betweenness centrality is based upon the shortest path analysis. In the context of HRDT, for every two employees in a network, there is at least one shortest path, where this path is defined as the minimum number of nodes that is needed to be traversed in order to reach from one node to another one, or if there is a weight for each edge, it is defined as the shortest path from one node to another node which leads to the lowest weight. With this regards, betweenness centrality is used to simulate the information flow among employees in a public organization, specifically Georgia DOT. Betweenness centrality of a node in a graph can be calculated using equation (6).

$$BC(m) = \sum_{m \neq n \neq t} \frac{\sigma_{nt}(m)}{\sigma_{nt}} \tag{6}$$

where σ_{nt} is the total number of shortest paths from Employee n to t in the whole graph. $\sigma_{nt}(m)$ is the number of shortest paths passing through employee m. For evaluating absence impact, betweenness centrality is calculated at the departmental unit (local) level, since a calculation at the organizational (global) level places too much emphasis on the persons who are located higher in the organization structure from a hierarchical perspective.

Based on an evaluation of the histogram of the calculated values, it is assumed that employees with a betweenness centrality value less than that corresponding to the 70th percentile are considered non-essential, and hence assigned an absence impact rating of 1. Those employees with betweenness centrality values corresponding to between the 70th and 90th percentiles are assigned a rating of 2. Lastly, if the betweenness centrality value falls into the 90th percentile or above, the employee is assigned a rating of 3, indicating the employee is essential, thus his / her absence would be highly impactful to the organization.

To sum up, the network analysis engine has derived six important attributes for each employee, i.e. basic uniqueness, network uniqueness, basic criticality, network criticality, employee position level and absence impact. A table containing employee's basic information, together with all the derived attributes are then saved into the file system for use by the functional modules and computation engines, such that the network analysis doesn't have to be re-performed every time users start HRDT. Similar to the attrition analysis engine, the update function will take care of the situation that workforce changes in the organization.

5.2.3 Text Mining Engine

The main task of the text mining engine is to apply text mining techniques that were studied in Section 4.3, to the texts in job titles, descriptions, and entry qualifications of every employee in Georgia DOT, and evaluate how similar the content of a job is compared to every other one. There are a total of 184 jobs among all the employees in Georgia DOT for text mining analysis. Instead of querying data from the employee database, the job description and entry qualification of all the jobs are collected from the official government website of Georgia Department of Administrative Services (Georgia DOAS) for each job title excluding commissioners and temporary positions. In the HRDT system, the similarity evaluation results from the text mining engine are to be used in some of the knowledge management modules to provide recommendation and / or filtering on the job titles. Although there are cases such as job shadowing, where job similarity is not considered critical, or even jobs in two very different fields can be matched, however, some knowledge transfer activities such as workforce planning or succession planning, require that the knowledge provider and the knowledge recipient are ideally working contents similar to each other, so as to allow for optimal knowledge retention.

Starting with data-preprocessing, the text mining engine applies each of the steps as mentioned to the texts. The texts are firstly checked for any misspelling words, and are auto corrected if there are any. It turns out that all the job description and entry qualification texts collected from the official website are in great shape and uniform format without misspelled words. Then, each part of the texts (job title, description and entry qualification) are respectively read into the text mining engine of HRDT system for the remain procedures of pre-processing.

In the step of special terms extraction, there are certain text terms in both entry qualification and job description that need to be specifically taken care of. In the entry qualification texts, the special terms include degree requirements, such as “Bachelor’s degree in Business”, “Master’s degree in Civil Engineering”, “Juris Doctorate”, “High school diploma”, etc., work experience requirements, such as “Three years of related experience”, “Seven years of experience in purchasing”, etc., and moreover, the leadership requirements, such as “One year of demonstrated experience in a leadership role”, “One year of which is in a lead/supervisory role”, “Two years of which in a supervisory, administrative or lead worker role”, etc. These terms are carefully parsed into features such as degree level, degree field, experience, experience field, leadership, etc. These features are not only used as descriptive features for similarity analysis, but also for filtering for job title recommendation in the functional modules. Furthermore, in the job description texts, there are also some special terms such as “Under direct supervision”, “Under broad supervision”, etc. These terms should be viewed as phrases in a whole as a descriptive feature rather than split into each word as several features. One important fact to note is that such a step is performed before the removal of any uninformative words due to the fact that some of the phrases actually contain some of the uninformative words.

In the next step of stopping words and uninformative words removal, two hash sets of words are created respectively. While stopping words are fairly universal, containing “a”, “an”, “the”, “again”, “only”, etc., and a set can usually be acquired from libraries, uninformative words sets are more domain specific, and require scientific collection. In this text mining engine, the uninformative words set contains such words that do not directly describe the job content of any job titles, such as “item”, “make”, “need”, “take”, etc. A Word Cloud visualization can be applied here to check any high-frequency words that are deemed to be uninformative for domain in the HRDT system. Upon the completion of this step, the only words that are left in the text sentences are those directly describing job contents. Just as important is the step of stemming and lemmatization, which aims at reducing inflectional and derivative forms of each word into its base format. Inflectional and derivative forms sometimes have context specific meaning, however, those words have similar meanings with their stemming words for most of the scenarios. Therefore, prior to applying lemmatization library to the texts, inflectional words with domain-specific meanings are firstly extracted, such as: “transportation” rather than stemmed to transport, “maintenance” rather than stemmed to maintain, “construction” rather than stemmed to construct, etc., since all these words are specifically descriptive to the job contents. The final step of data pre-processing in the text mining engine is splitting and tokenization, where the remaining words in the sentences are tokenized into each word, which serves as a qualified feature for defining job content, as well as for job similarity analysis in the following context.

For pairwise similarity analysis, as studied in Section 4.3, the common evaluation approaches are Cosine similarity and Jaccard similarity. Equation 7 shows the Cosine similarity evaluation of two items, and equation 8 shows the Jaccard similarity evaluation of two items.

$$CosineSimilarity = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{7}$$

$$JaccardSimilarity = \frac{Intersection(A, B)}{Union(A, B)} = \frac{A \cap B}{A \cup B} \tag{8}$$

Where A and B are two items, i.e. two jobs in the context of text mining engine. With an example to illustrate the two similarity evaluation approaches, let A be the job description features of the job with title Civil Engineer 2 after data pre-processing: "under supervision, perform, introductory, engineering, provide, engineering, support, plan, design, cost, coordination, transportation, construction, maintenance, activity, Bachelor's degree, 0 years, non-manager role", and B be the features of job title Transportation Specialist 2: "under supervision, provide, working level, engineering, support, OMAT, maintenance, construction, intermodal, plan, utility, survey, environment, transportation, Bachelor's degree, 0.5 years, non-manager role". Part of the term frequency used by Cosine similarity is shown in the following table:

Item	under supervision	introductory	working level	engineering	survey	maintenance	…
A	1	1	0	2	0	1	
B	1	0	1	1	1	1	

Then the term frequencies need to be normalized with regard to the L2-norm of each item, i.e. each job features vector. L2-norm of vector-A is 4.47 and L2-norm of vector B is 4.12. Part of the normalized term frequency as below:

Item	under supervision	introductory	working level	engineering	survey	maintenance	…
A	0.224	0.224	0	0.447	0	0.224	
B	0.243	0	0.243	0.243	0.243	0.243	

CosineSimilarity = (0.224*0.243) + (0.447*0.243) + (0.224*0.243) + … = 0.51

Jaccard on the other hand, ignore the term frequency and treat the terms as set, calculated as: JaccardSimilarity = intersect size / union size = 0.33

Due to the fact that Cosine similarity measures the cosine angle of two vectors where it takes into account the frequencies of each term (feature) while Jaccard similarity treats features as set and the term frequency does not affect the evaluation, we see a more flat trend with Cosine approach in the similarity distribution in Figure 4.8. In this text mining engine, Cosine approach will be used for pairwise similarity analysis. The pairwise similarity analysis result will be utilized by the functional modules in the way that when users search candidates for knowledge transfer activities, text mining engine are called to provide recommendation or filtering on the working titles of potential employees. For example, if the user is trying to search knowledge recipient for a position with title “Sr Mgr 2, Transport Specialist”, then the top 15 similar job titles will be automatically selected for potential candidates list. However, there are always the cases where the critical position is highly unique where the most similar job title is not very similar based on evaluation. Therefore, users always have the option to choose any job titles for filtering based on their own judgement. This part will be further discussed in the functional modules where the result from the text mining engine is used.

5.2.4 Multivariate Analysis Engine

The main mission of the multivariate analysis (MVA) engine is to feed multiple parameters into a Multi-Criteria Decision-Making (MCDM) model with the parameter weights calculated by Rank Order Centroid (ROC) approach as mentioned in section 4.4. To be consistent across the whole system, multivariate analysis engine uses a four-parameter format model for all evaluations and assessments, and each parameter is defined on a three-level rating scale. All of the assessments and evaluations are used by the knowledge management functional modules which consists of five submodules to be introduced in the following section.

As for all the input parameters for MVA, some are pre-calculated from other computation engines, i.e. attrition analysis engine and network analysis engine, and additionally, there are some user-defined parameters that must be manually selected depending on the knowledge management module being used.

Pre-calculated Parameters

Pre-calculated parameters are those factors that are calculated from other computation engines, thus are automatically filled out for HRDT users. Parameters pre-calculated by the network analysis engine include employee uniqueness, criticality, position level and absence impact. Those pre-calculated by attrition analysis engine include vacancy risk, tenure at current position and certificate.

For employee uniqueness and criticality assessments, as mentioned in for network analysis engine, there is a basic version and a network version. In HRDT, users have the option to select either one for both parameters, resulting in a total of four combinations, i.e. basic uniqueness and basic criticality, basic uniqueness and network criticality, network uniqueness and basic criticality, network uniqueness and network criticality.

The position level parameter refers to the level of the employee in the hierarchical sense within the organizational chart. For example, at the departmental unit (local) level, the chair is considered to be Level 1, those reporting directly to the chair are considered to be Level 2, and so on. At the organizational (global) level, the commissioner of GDOT is considered to be Level 1, and those reporting directly to the commissioner are considered to be Level 2, and so on. By default, candidates at the same level as the employee are assigned a position level rating of 3, those one level below are assigned a rating of 2, and those two levels or greater below are assigned a rating of 1. Position level evaluation will be used as one parameter when determining the candidate suitability.

Other parameters, including absence impact, vacancy risk, tenure and certificate, have been introduced in the two computation engines before, and are fairly straightforward for users to understand in each knowledge management module.

User-defined Parameters

The following is a list of all the user-defined parameters that require manual input:

- Resource availability: this refers to the fact that there may be budget and/or time constraints for carrying out knowledge retention / transfer activities within the organization. A score of 1 should be assigned if there is little to no organizational support, 2 should be assigned if there is some organizational support, and 3 should be assigned if there is full organizational support.
- Time availability: this refers to time available as a percentage of total time that a potential candidate has for participation in a knowledge transfer program. For example, a potential candidate with only 1 or 2 hours per week available may not be as effective as someone who has 6 to 8 hours per week or more. A score of 1 should be assigned if 4 to 8 hours per week are available, 2 should be assigned if 8 to 16 hours per week are available, and 3 should be assigned if more than 16 hours per week are available.

- Time period: this is different from the time availability parameter, and refers to the time period that a potential candidate has available for participation in a knowledge transfer program, as the success of a program is related to the program's duration. A score of 1 should be assigned if there is less than 3 months available, 2 should be assigned if there is 3 to 6 months available, and 3 should be assigned if there is more than 6 months available.
- Willingness / Attitude: this refers to the willingness / attitude of a potential candidate (in Workforce Planning and Succession Planning modules), and also of either a trainer or trainee (in Cross Training Module and Job Shadowing Module). Those with a low level of willingness should be assigned a score of 1, those with a moderate level of willingness should be assigned a score of 2, and those with a high level of willingness and motivation should be assigned a score of 3.
- Leadership: this refers to a potential candidate's perceived leadership skills for the purposes of Workforce Planning and Succession Planning. Those with low level of perceived leadership should be assigned a score of 1, those with moderate level of perceived leadership should be assigned a score of 2, and those with a high level of perceived leadership should be assigned a score of 3.
- Past performance: this refers to the past performance of a candidate that can be based on annual reviews (if available for internal candidates), or based on resume (for external candidates). Those with low perceived performance should be assigned a score of 1, those with moderate perceived performance should be assigned a score of 2, and those with high perceived performance should be assigned a score of 3.

- Skill set: this refers to relevant professional skills and certifications; those with a low level of relevant skills should be assigned a score of 1, those with a moderate level of relevant skills should be assigned a score of 2 (this is also the default value), and those with a high level of relevant skills should be assigned a score of 3. For example, a person applying to become a heavy equipment operator would be considered to have a high skill set if he/she has the required certifications to operate various heavy machinery, moderate skill set if he/she is certified to operate only one or a few specific types of heavy equipment, and a low skill set if he/she does not have the relevant skills and certifications required for the position.

After introducing each of the input parameters for multivariate analysis, more importantly, we need to design a ranking among each of the four parameters for all MVA assessments, such that proper weight for each parameter can be calculated from ROC. There are total of 8 assessments in this MVA engine, and Table 5 shows a summary of the scoring matrix for each input parameter for each MVA evaluation that used in HRDT. The value in parenthesis below each parameter represents the according weight that was calculated from ROC, where the parameter importance ranking is based upon both discussions with stakeholders from Georgia DOT and the research team’s scientific judgement. Moreover, the HRDT users have full control to override any parameter’s default rating calculated from computation engine and they can do so by manually selecting what is deemed to be a more appropriate rating.

Table 5 – Scoring matrix for ranking of factors used in multivariate analysis.

Factor Evaluation	1	2	3	4
KLR Assessment (1, 3)	Vacancy Risk (0.479)	Criticality (0.229)	Uniqueness (0.229)	Resource Availability (0.063)
Position Importance Evaluation (2, 4)	Absence Impact (0.479)	Criticality (0.229)	Uniqueness (0.229)	Resource Availability (0.063)
Experience Score (2, 3)	Position Level (0.521)	Centrality (0.271)	Tenure at Current Position (0.146)	Certifications (0.063)
Candidate Suitability Evaluation (2, 3)	Experience Score (0.479)	Performance Score (0.229)	Willingness / Attitude (0.229)	Leadership (0.063)
Mentor Evaluation (1)	Willingness (0.500)	KLR (0.250)	Time Period (0.125)	Time Availability (0.125)
Protégé Evaluation (1)	Past Performance (0.521)	Willingness (0.271)	Time Availability (0.146)	Location (0.063)
Trainer Evaluation (4)	Position Importance (0.250)	Tenure at Current Position (0.250)	Willingness / Attitude (0.250)	Time Availability (0.250)
Trainee Evaluation (4)	Position Level (0.400)	Skill Set (0.200)	Willingness / Attitude (0.200)	Time Availability (0.200)

Once the scores are assigned, they are then multiplied by their respective weights summarized in the above table, after which an overall score is calculated with weighted summation approach as mentioned in section 4.4. With the regard to the fact that there are four parameters for each evaluation and each parameter is rated from 1 to 3, the range of the total score can be from 4 to 12.

This innovative algorithm for human resource metrics evaluations uses a consistent set of 4 factors for each metric. Such an evaluation design not only can calculate evaluation scores comprehensively and scientifically, but also allows for a friendly user interface for the following knowledge management modules.

The detailed algorithm, including the ranking of multiple factors and the weighting of those factors based on Rank Order Centroid approach, was first developed by the research team, after which it was presented to Georgia DOT's human resource personnel for validation from the domain experts. During the validation cycle, a few improvements were addressed to further enhance the algorithm. For example, users should have full control over the score of each factor, in which way a subjective evaluation is made possible whenever deemed necessary. Moreover, some unique constraints needed to be added to the algorithm for dealing with particular unanticipated situations.

Lastly, a "rating" system has been developed by studying all possible combinations of overall scores to define ranges, and applying additional constraints considering particular situations. For example, one constraint is that a candidate or trainee's overall score is automatically set to low if his/her willingness / attitude is chosen as low, as this indicates that he/she is not a willing participant and therefore not suitable for selection. Another constraint is that if both uniqueness and criticality are low, then the KLR score is automatically set to low. This rating system applies to KLR assessment, mentor and protégé evaluation, experience score, candidate evaluation, and trainer and trainee evaluation. For general cases, an overall score that ranges larger than or equal to 10 is considered a high score, else if a score is larger than or equal to 8, it is then considered moderate / medium score, and otherwise (below 8) the score is considered low.

5.2.5 Performing Update

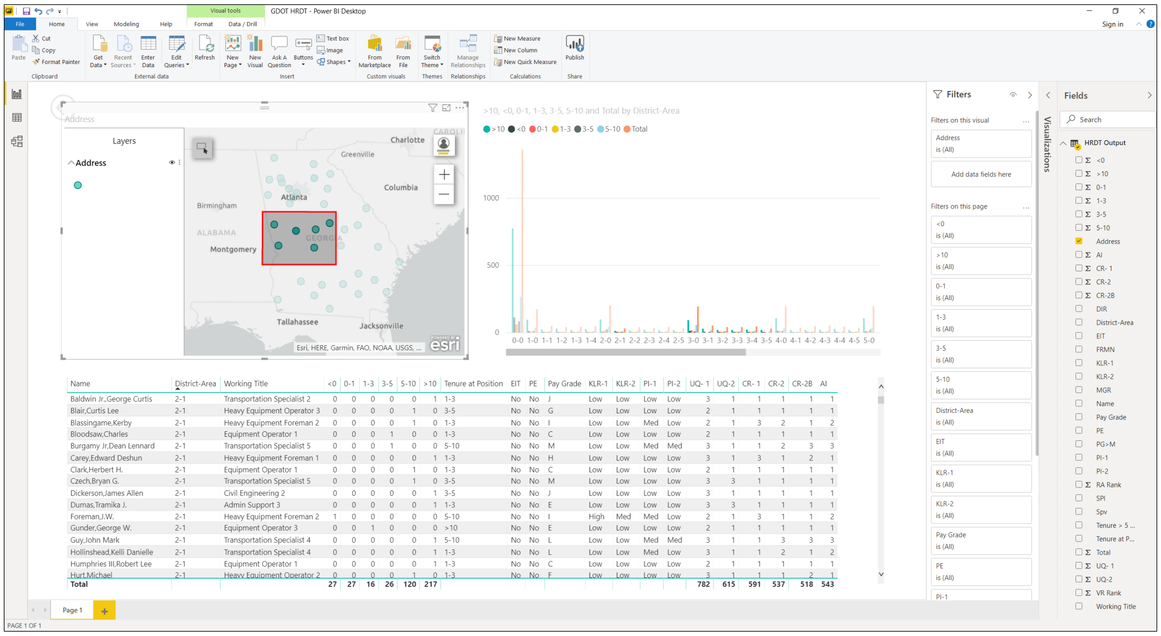
As previously mentioned in the sections about the attrition analysis engine and the network analysis engine, there are common situations such as Georgia DOT having new external employees hired, existing employees departed, employees promoted or transferred, etc. Since the two computation engines store calculated attributes in the file system to prevent any unnecessary recalculations, HRDT has to take care of these scenarios by implementing the functionality that checks any inconsistency between the HR database and the stored files. Therefore, there is this update functionality that upon the startup of the system, it checks the employee id and basic employee information in the previously stored attrition analysis and network analysis files. If there is any inconsistency with the HR database, the two computation engines will perform update processes which rerun all the analyses to generate all the essential attributes for HRDT to function correctly, during which process, users will be notified upon update completion.

5.3 Organization-wide Decision-Making Tools

This section and the next section introduce the design of user interaction and the workflow of HRDT functional modules. This section focuses on the two organization-wide tools, i.e. attrition distribution visualization module and network analysis and visualization module, that aim to expand Georgia DOT's ability for HR decision making from a "person to person" mode to an "organization wide" one. The next section focuses on the knowledge management modules, that contain five submodules, with the objective to assist Georgia DOT users in HR department to design and conduct a set of knowledge retention and transfer activities through an intuitive graphical user interface and smooth workflow.

5.3.1 Attrition Distribution Visualization

The attrition distribution visualization module in HRDT is an interactive dashboard which allows users to use spatial, tabular and graphical tools to rapidly visualize employee information across all the branch offices in the state of Georgia to aid in decision making. This module utilizes the geographical information visualization data tool that was previously studied in section 4.1, and as discussed, Microsoft Power BI is integrated into this module in order to leverage the pivot table and chart, as well as spatial analysis and visualization capabilities. The files that were stored in the file system by the attrition analysis engine and the network analysis engine will act in turn as the input data sources for Power BI's subsequent analysis. Figure 5.4 shows the interface of Power BI in the attrition distribution visualization module, which consists of three main panels, i.e. spatial information panel (top left), pivot chart panel (top right), and pivot table data panel (bottom). The three panels are interlinked, meaning that performing map-based, chart-based or data-based analysis also updates the results shown in the other panels and vice versa. As shown in the chart, when users box-select some dots in the map panel, chart panel automatically highlight bars for the district-areas that correspond to the selected ones, and the data panel also updates the contents accordingly.



Name	District-Area	Working Title	<0	0-1	1-3	3-5	5-10	>10	Tenure at Position	EIT	PE	Pay Grade	KLR-1	KLR-2	PI-1	PI-2	UQ- 1	UQ-2	CR- 1	CR-2	CR-2B	AI
Baldwin Jr.,George Curtis	2-1	Transportation Specialist 2	0	0	0	0	0	1	1-3	No	No	J	Low	Low	Low	Low	3	1	1	1	1	1
Blair,Curtis Lee	2-1	Heavy Equipment Operator 3	0	0	0	0	1	0	3-5	No	No	G	Low	Low	Low	Low	2	1	1	1	1	1
Blassingame,Kerby	2-1	Heavy Equipment Foreman 2	0	0	0	0	1	0	1-3	No	No	I	Low	Low	Med	Low	2	1	3	2	1	2
Bloodsaw,Charles	2-1	Equipment Operator 1	0	0	0	1	0	0	1-3	No	No	C	Low	Low	Low	Low	2	1	1	1	1	1
Burgamy Jr,Dean Lennard	2-1	Transportation Specialist 5	0	0	0	1	0	0	5-10	No	No	M	Low	Low	Med	Med	3	1	1	2	3	3
Carey,Edward Deshun	2-1	Heavy Equipment Foreman 1	0	0	0	0	0	1	1-3	No	No	H	Low	Low	Low	Low	3	1	3	1	2	1
Clark,Herbert H.	2-1	Equipment Operator 1	0	0	0	0	1	0	1-3	No	No	C	Low	Low	Low	Low	2	1	1	1	1	1
Czech,Bryan G.	2-1	Transportation Specialist 5	0	0	0	0	1	0	3-5	No	No	M	Low	Low	Low	Low	3	3	1	1	1	1
Dickerson,James Allen	2-1	Civil Engineering 2	0	0	0	0	0	1	3-5	No	No	J	Low	Low	Low	Low	3	1	1	1	1	1
Dumas,Tramika J.	2-1	Admin Support 3	0	0	0	0	0	1	1-3	No	No	E	Low	Low	Low	Low	3	3	1	1	1	1
Foreman,J.W.	2-1	Heavy Equipment Foreman 2	1	0	0	0	0	0	5-10	No	No	I	High	Med	Med	Low	2	1	3	1	1	2
Gunder,George W.	2-1	Equipment Operator 3	0	0	1	0	0	0	>10	No	No	E	Low	Low	Low	Low	2	1	1	1	1	1
Guy,John Mark	2-1	Transportation Specialist 4	0	0	0	0	0	1	5-10	No	No	L	Low	Low	Med	Med	3	1	1	3	3	3
Hollinshead,Kelli Danielle	2-1	Transportation Specialist 4	0	0	0	0	0	1	1-3	No	No	L	Low	Low	Med	Low	3	1	1	2	1	2
Humphries III,Robert Lee	2-1	Equipment Operator 1	0	0	0	0	0	1	1-3	No	No	C	Low	Low	Low	Low	2	1	1	1	1	1
Hurt,Michael	2-1	Heavy Equipment Operator 2	0	0	0	0	1	0	1-3	No	No	F	Low	Low	Low	Low	3	1	1	1	2	1
Total			**27**	**27**	**16**	**26**	**120**	**217**									**782**	**615**	**591**	**537**	**518**	**543**

Figure 5.4 – Power BI interface for attrition distribution visualization module. (When users box-select dots in the map panel, others interlinked panels auto update).

The attrition distribution visualization module also provides powerful filtering and analysis tools, and allows the users to perform "what-if" scenario analyses in order to evaluate specific organization needs. Table 6 provides a brief description of all the fields and parameters, including basic employee information and those for the module to perform filtering, ranking, counting, and other statistics analyses.

Table 6 – Summary of all the parameters available in attrition distribution visualization module, including both basic fields and those for statistical analysis.

Parameter	Description
Name	The name of the employee.
District-Area	The district-area code to which the employee is assigned
Address	The physical address to which the employee is assigned (from district-area code)
Working Title	The job title of the employee.

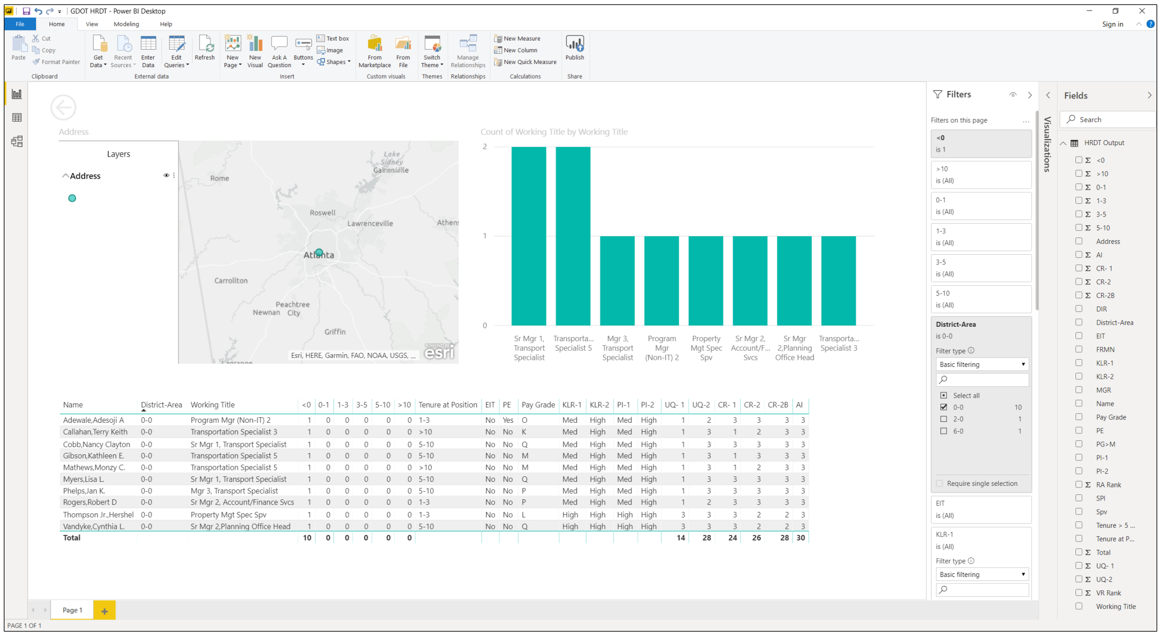
Table 6 continued.

Parameter	Description
< 0	A value of "1" is assigned to each employee who has already met the retirement requirements (based on age and/or tenure), but has not retired yet.
0 – 1	A value of "1" is assigned to each employee who is expected to retire within the next one year (based on age and/or tenure).
1 – 3	A value of "1" is assigned to each employee who is expected to retire in the next one to three years (based on age and/or tenure).
3 – 5	A value of "1" is assigned to each employee who is expected to retire in the next three to five years (based on age and/or tenure).
5 – 10	A value of "1" is assigned to each employee who is expected to retire in the next five to ten years (based on age and/or tenure).
> 10	A value of "1" is assigned to each employee who is not expected to retire within the ten years (based on age and/or tenure).
PE	"Yes" if the employee has the registered professional-engineer license.
EIT	"Yes" if the employee has the registered engineer-in-training license.
Pay Grade	The alphabetical pay grade of the employee as obtained from the database.
Tenure at Position	The tenure of the employee (in years) at his/her most recent position, based on the position entry date in the database.
VR	Vacancy risk rating that was calculated in attrition analysis engine.
UQ-1	Basic uniqueness rating of the employee as computed in the network analysis engine.
UQ-2	Network-based uniqueness rating of the employee as computed in the network analysis engine.
CR-1	Basic criticality rating of the employee as computed in the network analysis engine.
CR-2	Network-based criticality rating of the employee using PageRank on local level as computed in the network analysis engine.
CR-2B	Network-based criticality rating of the employee using Cosine dissimilarity as computed in the network analysis engine, which was referred as network criticality.
AI	Absence impact of the employee as computed in the network analysis engine.
KLR-1	Knowledge loss risk of the employee that was calculated using basic uniqueness and basic criticality.
KLR-2	Knowledge loss risk of the employee that was calculated using network uniqueness and network criticality.
FRMN	"True" if the employee's job title contains "FRMN" (Foreman).

Table 6 continued.

Parameter	Description
SPI	“True” if the employee’s job title contains “SPI” (Superintendent).
SPV	“True” if the employee’s job title contains “SPV” (Supervisor).
MGR	“True” if the employee’s job title contains “MGR” (Manager).
DIR	“True” if the employee’s job title contains “DIR” (Director).
Tenure ≥ 2	“True” if the employee has been at his / her most recent position for at least two years.

With these parameters extracted from the database and calculated from the engines, users are able to access the interacting selecting and filtering functions from the filtering area to the right of the three panels. For example, a use case can be selecting all the employees in the Atlanta HQ (select district-area code 0-0), that have already met the retirement requirement (select “< 0”), who have high ratings of network based KLR and network based position importance (select KLR-2 and PI-2 as “high”). The visualization will yield the result as shown in Figure 5.5, which also shows a distribution of the working titles of the employees with the “risk” criteria.


GDOT HRDT - Power BI Desktop
Address
Layers
Address
Rome
Lake Sidney Lanier
Gainesville
Roswell
Lawrenceville
Athens
Atlanta
Carrollton
Peachtree City
Newnan
Griffin
Esri, HERE, Garmin, FAO, NOAA, USGS, ...
Count of Working Title by Working Title
2
1
0
Sr Mgr 1, Transport Specialist
Transporta... Specialist 5
Mgr 3, Transport Specialist
Program Mgr (Non-IT) 2
Property Mgt Spec Spv
Sr Mgr 2, Account/F... Svcs
Sr Mgr 2,Planning Office Head
Transporta... Specialist 3
Filters
Fields
HRDT Output
Page 1
PAGE 1 OF 1

Name	District-Area	Working Title	<0	0-1	1-3	3-5	5-10	>10	Tenure at Position	EIT	PE	Pay Grade	KLR-1	KLR-2	PI-1	PI-2	UQ- 1	UQ-2	CR- 1	CR-2	CR-2B	AI
Adewale,Adesoji A	0-0	Program Mgr (Non-IT) 2	1	0	0	0	0	0	1-3	No	Yes	O	Med	High	Med	High	1	2	3	3	3	3
Callahan,Terry Keith	0-0	Transportation Specialist 3	1	0	0	0	0	0	>10	No	No	K	Med	High	Med	High	1	3	1	2	3	3
Cobb,Nancy Clayton	0-0	Sr Mgr 1, Transport Specialist	1	0	0	0	0	0	5-10	No	No	Q	Med	High	Med	High	1	3	3	3	3	3
Gibson,Kathleen E.	0-0	Transportation Specialist 5	1	0	0	0	0	0	5-10	No	No	M	Med	High	Med	High	1	3	1	3	3	3
Mathews,Monzy C.	0-0	Transportation Specialist 5	1	0	0	0	0	0	>10	No	No	M	Med	High	Med	High	1	3	1	2	3	3
Myers,Lisa L.	0-0	Sr Mgr 1, Transport Specialist	1	0	0	0	0	0	5-10	No	No	Q	Med	High	Med	High	1	3	3	3	3	3
Phelps,Jan K.	0-0	Mgr 3, Transport Specialist	1	0	0	0	0	0	5-10	No	No	P	Med	High	Med	High	1	3	3	3	3	3
Rogers,Robert D	0-0	Sr Mgr 2, Account/Finance Svcs	1	0	0	0	0	0	1-3	No	No	P	Med	High	Med	High	1	2	3	3	3	3
Thompson Jr.,Hershel	0-0	Property Mgt Spec Spv	1	0	0	0	0	0	1-3	No	No	L	High	High	High	High	3	3	3	2	2	3
Vandyke,Cynthia L.	0-0	Sr Mgr 2,Planning Office Head	1	0	0	0	0	0	5-10	No	No	Q	High	High	High	High	3	3	3	2	2	3
Total			**10**	**0**	**0**	**0**	**0**	**0**									**14**	**28**	**24**	**26**	**28**	**30**

Figure 5.5 – View of the example use case from conducting selecting and filtering functionalities.

As this simple use case demonstrates, by identifying such employees and their locations, HR personnel can initiate succession planning activities to ensure that the knowledge of these employees is not lost due to attrition. Similarly, running such what-if scenarios can be used as a powerful decision-making tool at both the departmental unit and the organizational levels.

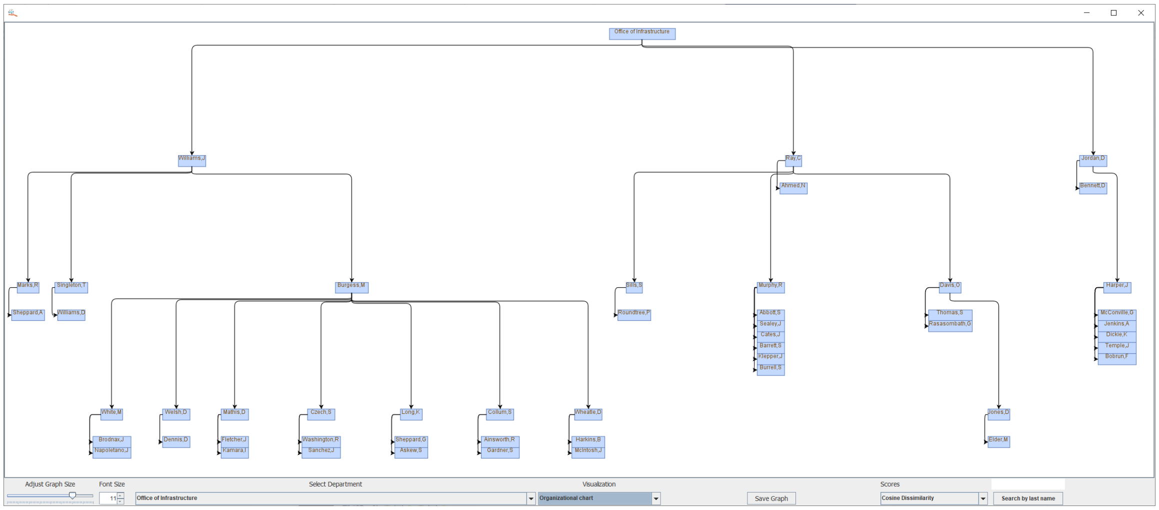
5.3.2 Network Analysis and Visualization

The network analysis and visualization module in HRDT is an interactive window that was built with the native Swing library of Java. With the HR database and the network file that is output and stored by the network analysis engine as the inputs, this module allows users to visualize the graph layout as well as graph analysis results of all the 56 predefined departmental units through interactive dropdown list operations. Users are able to visualize each of the departments in multiple views, including back-constructed traditional organizational charts and network views. Moreover, in the network view options, users can further select various graph metrics to be denoted through the node sizes. Table 7 shows a summary of all the ten derived metrics of the employee that are available to be shown through differentiating the node sizes in the network graph view.

Table 7 – Summary of all the metrics that can be shown through graph node sizes.

Parameter	Description
Position Level	The position level the employee based on the departmental level.
Local Betweenness	Betweenness centrality of employee as computed on departmental level.
Global Betweenness	Betweenness centrality of employee as computed on organizational level.
Local PageRank	PageRank centrality of employee as computed on departmental level.
Global PageRank	PageRank centrality of employee as computed on organizational level.
Cosine Dissimilarity	Cosine dissimilarity of the employee as computed in the network analysis engine, which was used for network criticality rating.
Basic Local Uniqueness	Uniqueness of employee as computed using basic definition (without considering position level) on departmental level.
Basic Global Uniqueness	Uniqueness of employee as computed using basic definition (without considering position level) on organizational level.
Network Local Uniqueness	Uniqueness of employee as computed using network definition (considering position level) on departmental level.
Network Global Uniqueness	Uniqueness of employee as computed using network definition (considering position level) on organizational level.

In the following context, a use case example is demonstrated for network analysis and visualization module. Figure 5.6 shows the organizational chart for the employees in the Office of Infrastructure. This chart is automatically constructed, using the “who reports to whom” information in the HR database. In the meantime, Figure 5.7 shows the network views of the employees from the same office, and instead of focusing on the reporting relationship, these views have a concentration on visualizing the connectivity of individuals and information flow within the employee network.



Adjust Graph Size

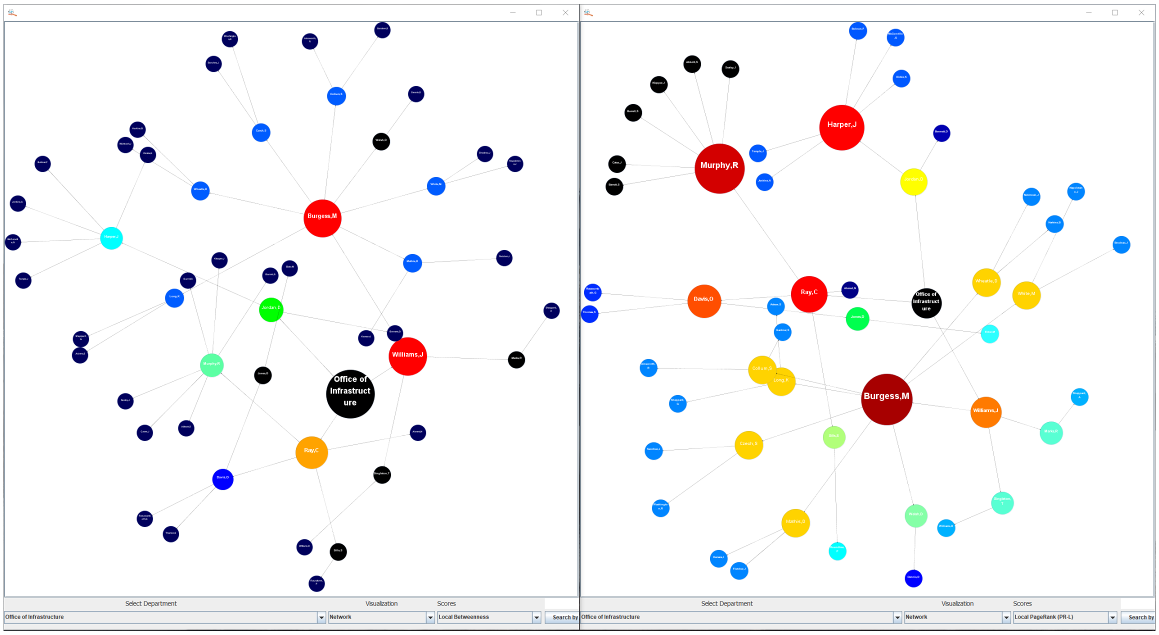
Font Size: 11

Select Department: Office of Infrastructure

Visualization: Organizational chart

Scores: Cosine Dissimilarity

Figure 5.6 – An example of back-constructed traditional organizational chart using the "reporting to" information from the HR database.



Select Department: Office of Infrastructure

Visualization: Network

Scores: Local Betweenness

Select Department: Office of Infrastructure

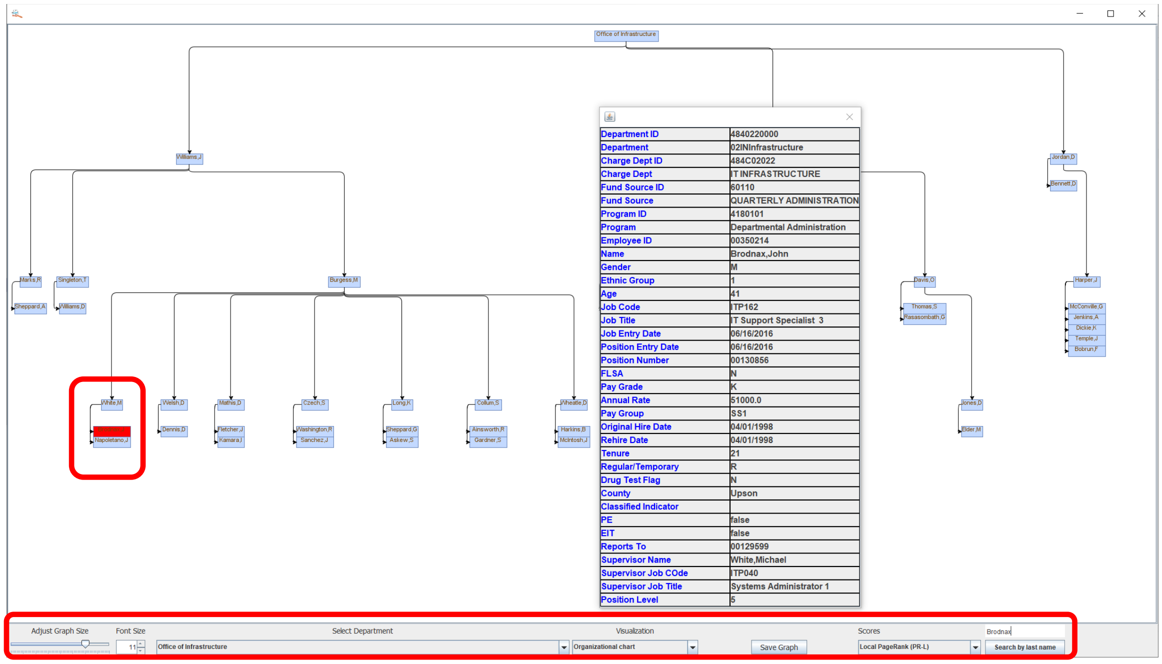
Visualization: Network

Scores: Local PageRank (PR-L)

Figure 5.7 – An example of network representations of the same department as in Figure 5.6. In Figure 5.7a (left), node size denotes the local betweenness centrality; In Figure 5.7b (right), node size denotes the local PageRank centrality.

As mentioned above, various attributes can be denoted by the node size in the network view representation. HRDT users can easily select from ten options (per Table 6) from a dropdown list within the visualization window, to fulfill specific needs for Georgia DOT. As an example, Figure 5.7a shows the network view with node size denoting the local betweenness centrality of each employee, and Figure 5.7b shows the same view with node size denoting the local PageRank centrality of each employee. As have been noted in section 4.2, betweenness centrality has a concentration of the strategically critical employees while PageRank centrality focuses on tactically important employees. Moreover, the network analysis and visualization module also allows users to visualize the data for the whole organization, which enhances perception of employee attributes on the organizational level.

Furthermore, the graphic user interface design of this module enables users to search for a specific employee by their last name, allowing for more effective searches. Upon the employee being found, his / her name box will be highlighted in the red color, and the detailed information shows up as shown in Figure 5.8. Upon clicking on the text box of an employee, his / her detailed HR information will be shown in a pop up window. Additionally, as can be seen in the figure, users can adjust the graph size using the slider, adjust the font size using the spinner, and can also save the visualization view as a PNG file for purposes such as generating a workforce planning report.



Department ID	4840220000
Department	02INInfrastructure
Charge Dept ID	484C02022
Charge Dept	IT INFRASTRUCTURE
Fund Source ID	60110
Fund Source	QUARTERLY ADMINISTRATION
Program ID	4180101
Program	Departmental Administration
Employee ID	00350214
Name	Brodnax,John
Gender	M
Ethnic Group	1
Age	41
Job Code	ITP162
Job Title	IT Support Specialist 3
Job Entry Date	06/16/2016
Position Entry Date	06/16/2016
Position Number	00130856
FLSA	N
Pay Grade	K
Annual Rate	51000.0
Pay Group	SS1
Original Hire Date	04/01/1998
Rehire Date	04/01/1998
Tenure	21
Regular/Temporary	R
Drug Test Flag	N
County	Upson
Classified Indicator	
PE	false
EIT	false
Reports To	00129599
Supervisor Name	White,Michael
Supervisor Job COde	ITP040
Supervisor Job Title	Systems Administrator 1
Position Level	5

Adjust Graph Size

Font Size: 11

Select Department: Office of Infrastructure

Visualization: Organizational chart

Save Graph

Scores: Local PageRank (PR-L)

Brodnax

Search by last name

Figure 5.8 – Additional UI functionalities in this module to help users.

5.4 Knowledge Management Modules

This section introduces the five functional modules that assist HR personnel at Georgia DOT with practical knowledge retention and knowledge transfer activities. The five functional modules are: workforce planning, job shadowing, succession planning, cross training, training and development. With computation engines' output files, i.e. the attrition file, the network file and the similarity file, as the input data sources, the five knowledge management modules implement very similar interface design in term of analysis and evaluation workflow. In the following subsections, the interface and procedure workflow of each module are introduced, and the first one will be discussed in detail with an example use case for instruction demonstration.

5.4.1 Workforce Planning

Workforce planning module helps HR personnel identify high-importance positions, as well as select the most suitable candidate for the important position that was identified in case the position needs to be replaced or supplemented. The module uses derived attributes such as uniqueness, tenure, leadership, etc., to perform multivariate analysis as mentioned in section 5.2, and generates evaluation results aiding knowledge retention. Additionally, users also have the option to use pairwise job similarity computed from the text mining engine for narrowing the search. The structure and evaluation workflow for the workforce planning module is summarized in Figure 5.9. An example use case will be demonstrated in the following context to better illustrate the mechanism of the module.

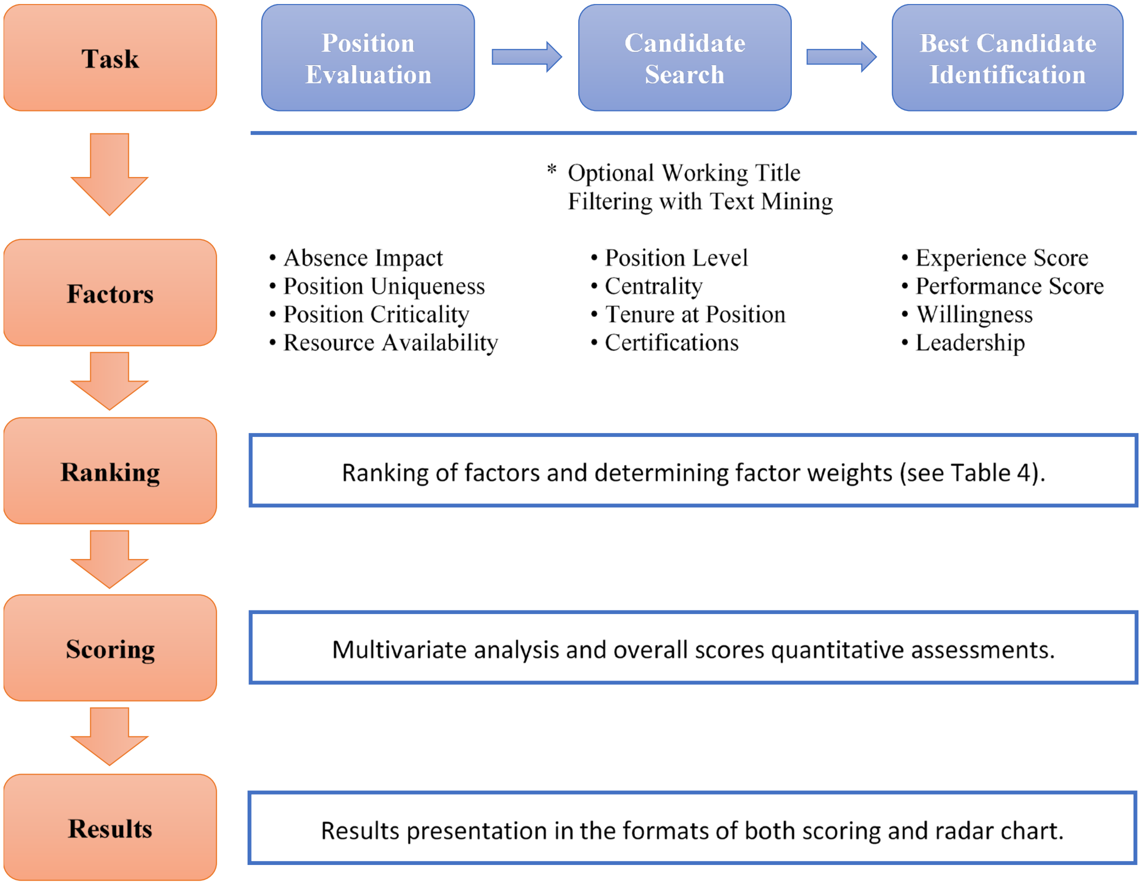
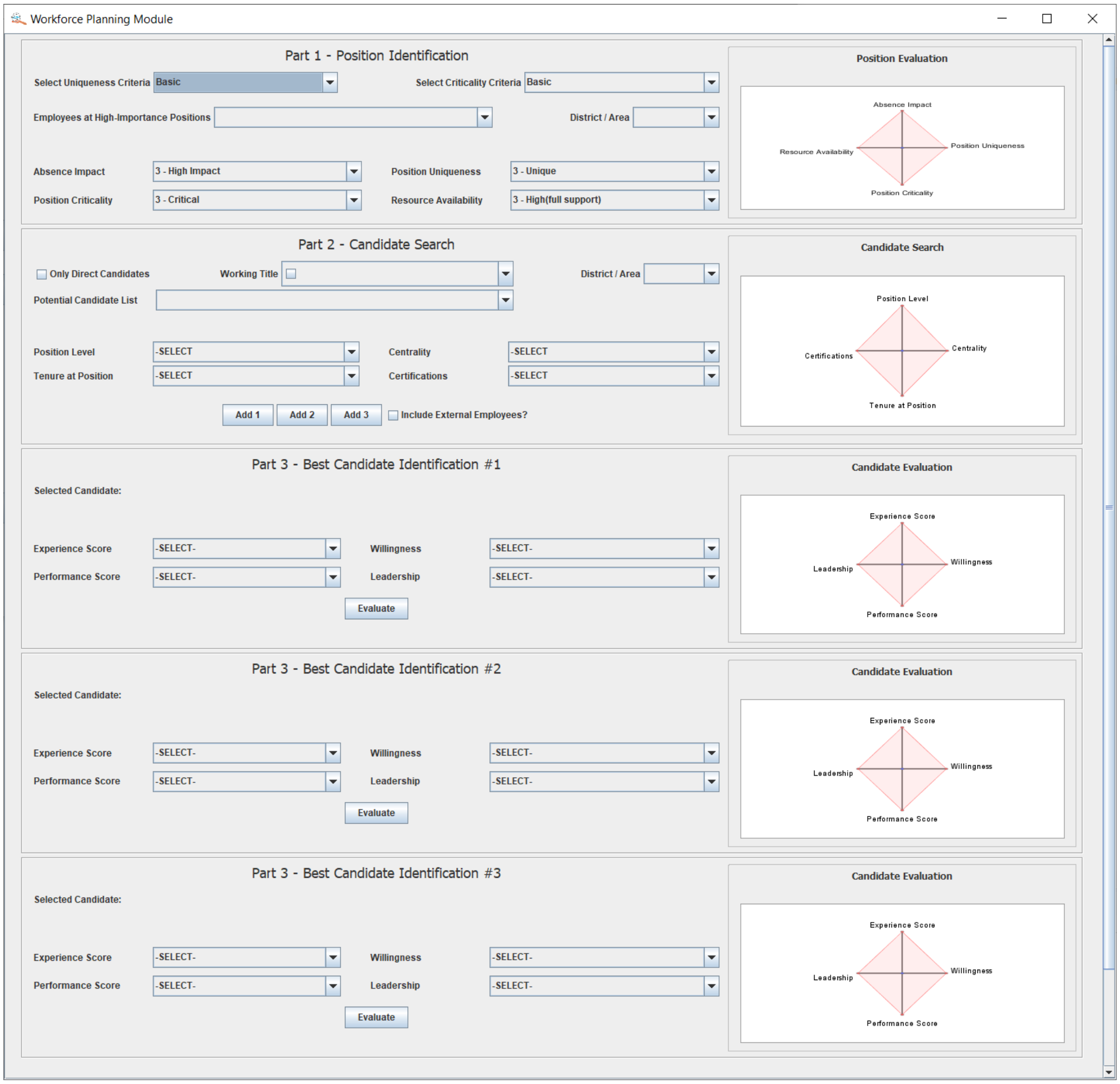

Task
Position Evaluation
Candidate Search
Best Candidate Identification
* Optional Working Title Filtering with Text Mining
Factors
• Absence Impact
• Position Uniqueness
• Position Criticality
• Resource Availability
• Position Level
• Centrality
• Tenure at Position
• Certifications
• Experience Score
• Performance Score
• Willingness
• Leadership
Ranking
Ranking of factors and determining factor weights (see Table 4).
Scoring
Multivariate analysis and overall scores quantitative assessments.
Results
Results presentation in the formats of both scoring and radar chart.

Figure 5.9 – Structure and evaluation workflow for the workforce planning module.

Figure 5.10 shows the overall user interface design of the workforce planning module upon starting. As shown, the module consists of three part, and users are supposed to follow the workflow of each part in order to perform series of evaluations and identifications for workforce planning activity.



Workforce Planning Module

Part 1 - Position Identification

Select Uniqueness Criteria: Basic
Select Criticality Criteria: Basic
Employees at High-Importance Positions:
District / Area:
Absence Impact: 3 - High Impact
Position Uniqueness: 3 - Unique
Position Criticality: 3 - Critical
Resource Availability: 3 - High(full support)

Position Evaluation

Part 2 - Candidate Search

☐ Only Direct Candidates
Working Title: ☐
District / Area:
Potential Candidate List:
Position Level: -SELECT
Centrality: -SELECT
Tenure at Position: -SELECT
Certifications: -SELECT
☐ Include External Employees?

Candidate Search

Part 3 - Best Candidate Identification #1

Selected Candidate:
Experience Score: -SELECT-
Willingness: -SELECT-
Performance Score: -SELECT-
Leadership: -SELECT-

Candidate Evaluation

Part 3 - Best Candidate Identification #2

Selected Candidate:
Experience Score: -SELECT-
Willingness: -SELECT-
Performance Score: -SELECT-
Leadership: -SELECT-

Candidate Evaluation

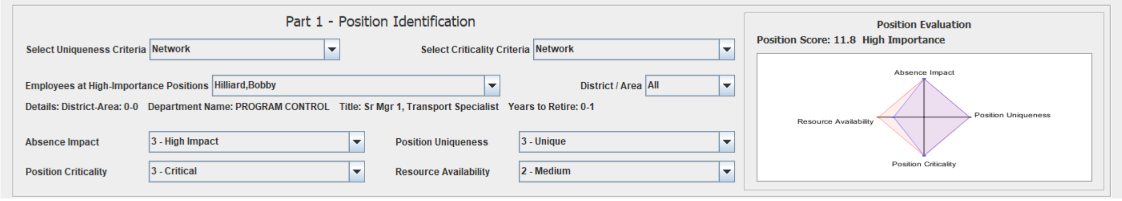
Part 3 - Best Candidate Identification #3

Selected Candidate:
Experience Score: -SELECT-
Willingness: -SELECT-
Performance Score: -SELECT-
Leadership: -SELECT-

Candidate Evaluation

Figure 5.10 – User interface of Workforce Planning module.

Starting with Part 1, users need to firstly choose the desired evaluating criteria for the uniqueness and criticality (basic or network), after which the critical employees at high-importance position, that identified based upon the selected criteria, are automatically returned in the list of "Employees at High-importance Positions". Then, users need to select one employee that of most interest from the list as the knowledge provider, whose attributes are automatically filled in the four dropdown lists. In the meantime, the calculation result of the quantitative position importance evaluation, together with a radar chart are presented on the right panel for users' reference. An illustrative example is shown in Figure 5.11.



Part 1 - Position Identification

Select Uniqueness Criteria: Network

Select Criticality Criteria: Network

Employees at High-Importance Positions: Hilliard,Bobby

District / Area: All

Details: District-Area: 0-0 Department Name: PROGRAM CONTROL Title: Sr Mgr 1, Transport Specialist Years to Retire: 0-1

Absence Impact: 3 - High Impact

Position Uniqueness: 3 - Unique

Position Criticality: 3 - Critical

Resource Availability: 2 - Medium

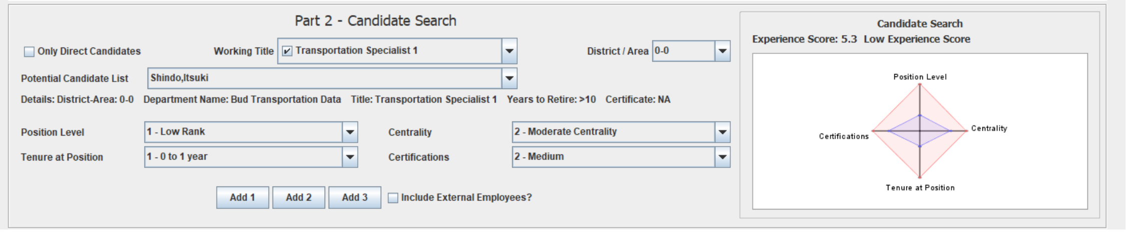
Position Evaluation

Position Score: 11.8 High Importance

Figure 5.11 – An example evaluation result of the module's part 1 which performs important position identification.

Part 2 of the module helps users perform initial search of the potential candidates for the purpose of workforce planning with respect to the position identified in Part 1, as well as assess candidate's experience score from the four attributes as shown in Part 2. To begin with this part, users have the option to choose candidates only from those who report directly to the critical employee selected in Part 1, by checking the "Only Direct Candidates" checkbox. On the other hand, by unchecking this box, the entire workforce can be searched using the "Working Title" and "District / Area" filters. The example continued from the previous part is shown in Figure 5.12.

Most importantly, a set of checkboxes in the "Working Title" filter are automatically selected based upon each respective similarity against the working title of the employee selected in Part 1, using the pairwise similarity analysis results that were computed in the text mining computation engine. Moreover, users are also given the ability to select or deselect any checkbox in the filter dropdown list. Based on the filtering condition, a list of potential candidates are returned in the list of "Potential Candidate List". Users can then select candidates from the list, whose attributes are automatically filled in the dropdown lists and the experience score is computed via multivariate analysis engine. The experience score result is also interpreted from the four dimensions with a radar chart. At the bottom of the Part 2, users can add up to three potential candidates to the panels in Part 3 for final decision. Furthermore, external employees can be included through the checkbox, whose basic information must then be manually input.



Part 2 - Candidate Search

□ Only Direct Candidates

Working Title: ☑ Transportation Specialist 1

District / Area: 0-0

Potential Candidate List: Shindo,Itsuki

Details: District-Area: 0-0 Department Name: Bud Transportation Data Title: Transportation Specialist 1 Years to Retire: >10 Certificate: NA

Position Level: 1 - Low Rank

Centrality: 2 - Moderate Centrality

Tenure at Position: 1 - 0 to 1 year

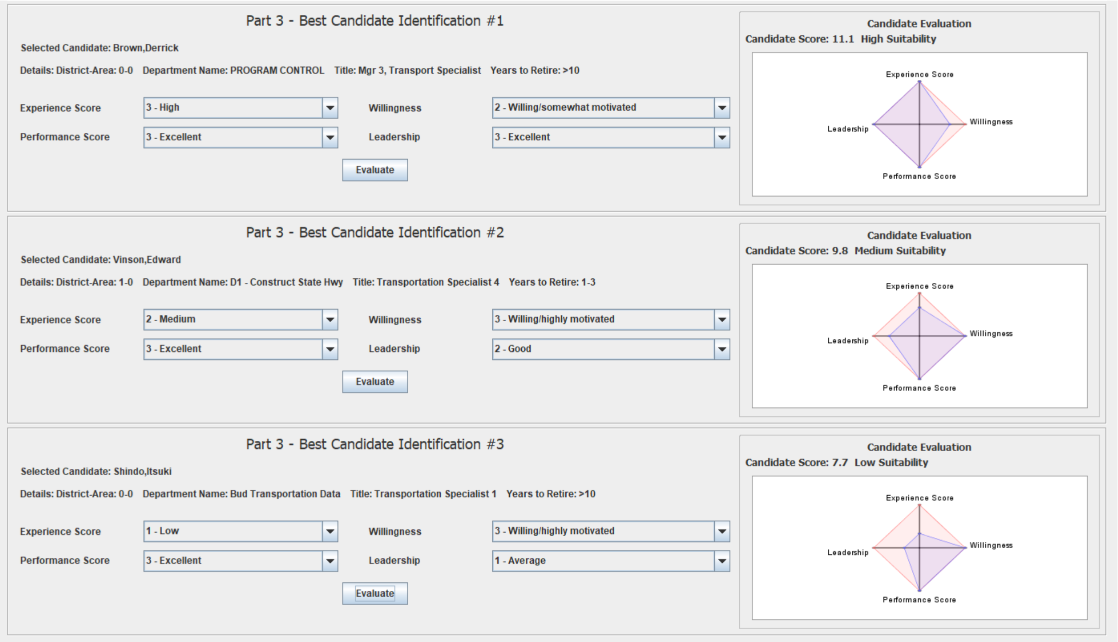
Certifications: 2 - Medium

Candidate Search

Experience Score: 5.3 Low Experience Score

Figure 5.12 – An example evaluation result of the module's part 2 which performs initial candidate search as well as candidate experience score assessment.

Part 3 has three panels, each containing the candidate information of the ones added from Part 2. The experience score attribute in each panel is automatically filled with the computed result from Part 2, while the performance score, willingness and leadership require user input so as to evaluate the candidate suitability. Once all of the four attributes are set, the "Evaluate" button enables the multivariate analysis engine to quantitatively assess the suitability of the candidate, whose evaluation result is shown on the right panel, as well as a radar chart just like the previous two parts. Figure 5.13 shows the evaluation results of three potential candidates as an example.



Part 3 - Best Candidate Identification #1

Selected Candidate: Brown,Derrick

Details: District-Area: 0-0 Department Name: PROGRAM CONTROL Title: Mgr 3, Transport Specialist Years to Retire: >10

Experience Score: 3 - High

Willingness: 2 - Willing/somewhat motivated

Performance Score: 3 - Excellent

Leadership: 3 - Excellent

Candidate Evaluation

Candidate Score: 11.1 High Suitability

Part 3 - Best Candidate Identification #2

Selected Candidate: Vinson,Edward

Details: District-Area: 1-0 Department Name: D1 - Construct State Hwy Title: Transportation Specialist 4 Years to Retire: 1-3

Experience Score: 2 - Medium

Willingness: 3 - Willing/highly motivated

Performance Score: 3 - Excellent

Leadership: 2 - Good

Candidate Evaluation

Candidate Score: 9.8 Medium Suitability

Part 3 - Best Candidate Identification #3

Selected Candidate: Shindo,Itsuki

Details: District-Area: 0-0 Department Name: Bud Transportation Data Title: Transportation Specialist 1 Years to Retire: >10

Experience Score: 1 - Low

Willingness: 3 - Willing/highly motivated

Performance Score: 3 - Excellent

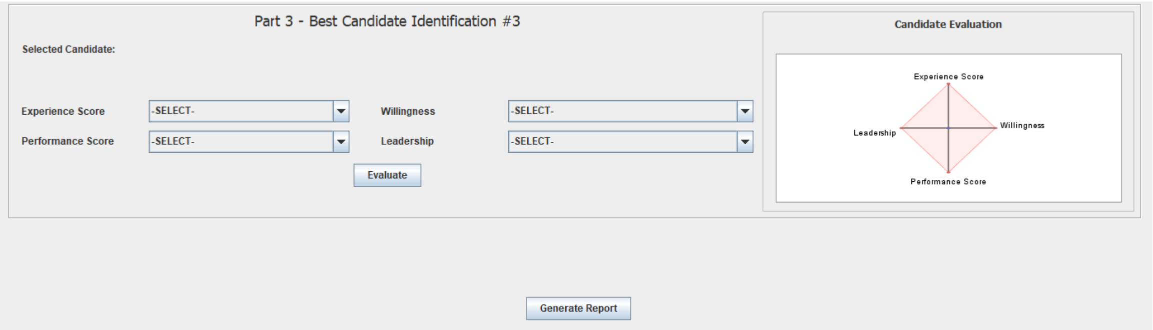
Leadership: 1 - Average

Candidate Evaluation

Candidate Score: 7.7 Low Suitability

Figure 5.13 – An example evaluation result of the module's part 3 which performs best candidate identification.

Lastly, the whole module can be output as a jpg of png image file through the "Generate Report" button at the very bottom, as shown in Figure 5.14. This enables the user of the system to prepare for a comprehensive report or presentation that assists workforce planning decision making.



Part 3 - Best Candidate Identification #3

Selected Candidate:

Experience Score: -SELECT-

Willingness: -SELECT-

Performance Score: -SELECT-

Leadership: -SELECT-

Candidate Evaluation

Figure 5.14 – Functionality of outputting evaluation result of the module as a report.

5.4.2 Job Shadowing

Job shadowing module helps HR personnel quantitatively evaluate employees' knowledge loss risk (KLR), as well as assess the suitability of employees being the mentor and protégé to allow for optimal knowledge transfer. While the derived attributes are different, the evaluation workflow of this module, as shown in Figure 5.15, is similar to that of the workforce planning module, allowing for the consistency and usability of the modules. Since the module shares a similar user interface and overall workflow, it will not be demonstrated with example use case.

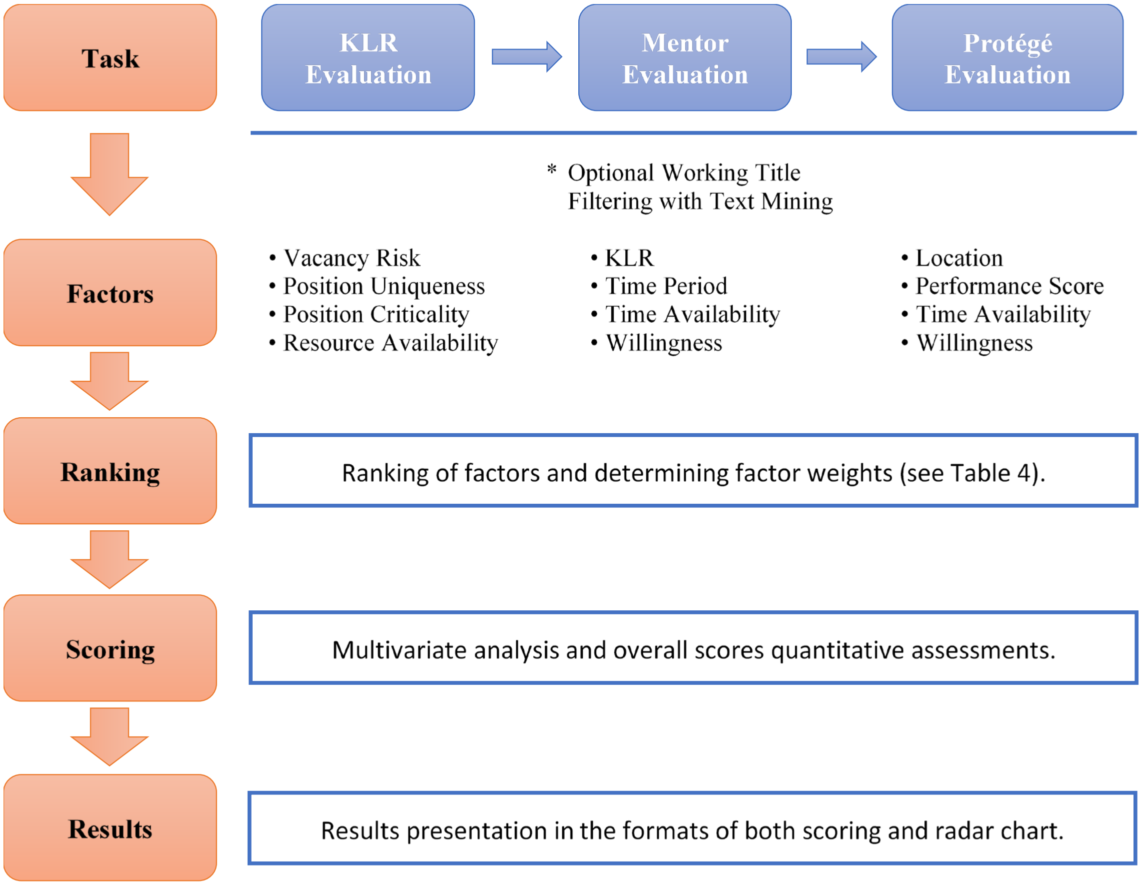


Figure 5.15 – Structure and evaluation workflow for the job shadowing module.

5.4.3 Succession Planning

Succession planning module helps HR personnel quantitatively evaluate employees' knowledge loss risk (KLR), as well as proactively select the most suitable candidate for the identified position with high-KLR, in case that the position need to be filled in the near future. The first part of this module emphasize the KLR assessment and other parts are identical to the workforce planning module. Similarly, the evaluation workflow chart of the module is shown in Figure 5.16.

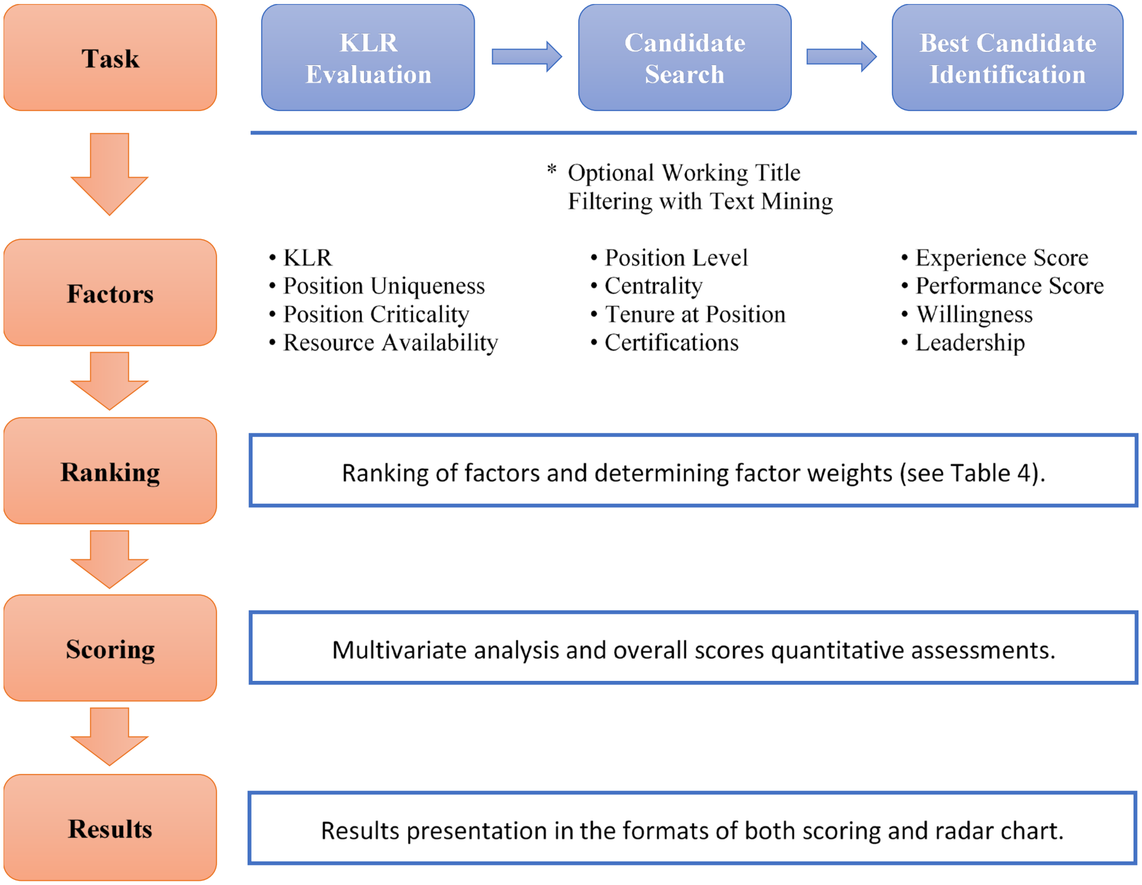

Task
KLR Evaluation
Candidate Search
Best Candidate Identification
* Optional Working Title Filtering with Text Mining
Factors
• KLR
• Position Uniqueness
• Position Criticality
• Resource Availability
• Position Level
• Centrality
• Tenure at Position
• Certifications
• Experience Score
• Performance Score
• Willingness
• Leadership
Ranking
Ranking of factors and determining factor weights (see Table 4).
Scoring
Multivariate analysis and overall scores quantitative assessments.
Results
Results presentation in the formats of both scoring and radar chart.

Figure 5.16 – Structure and evaluation workflow for the succession planning module.

5.4.4 Cross Training

Cross training module helps HR personnel identify high-importance positions, as well as select the most appropriate trainer and trainees for the important position identified in cases such as the employees at the critical position need to be emergency leave or long term vacation and the position is in great need of supplement. With a focus on assessing position importance, the module also shares a similar structure and workflow as previous modules, which is shown in Figure 5.17.

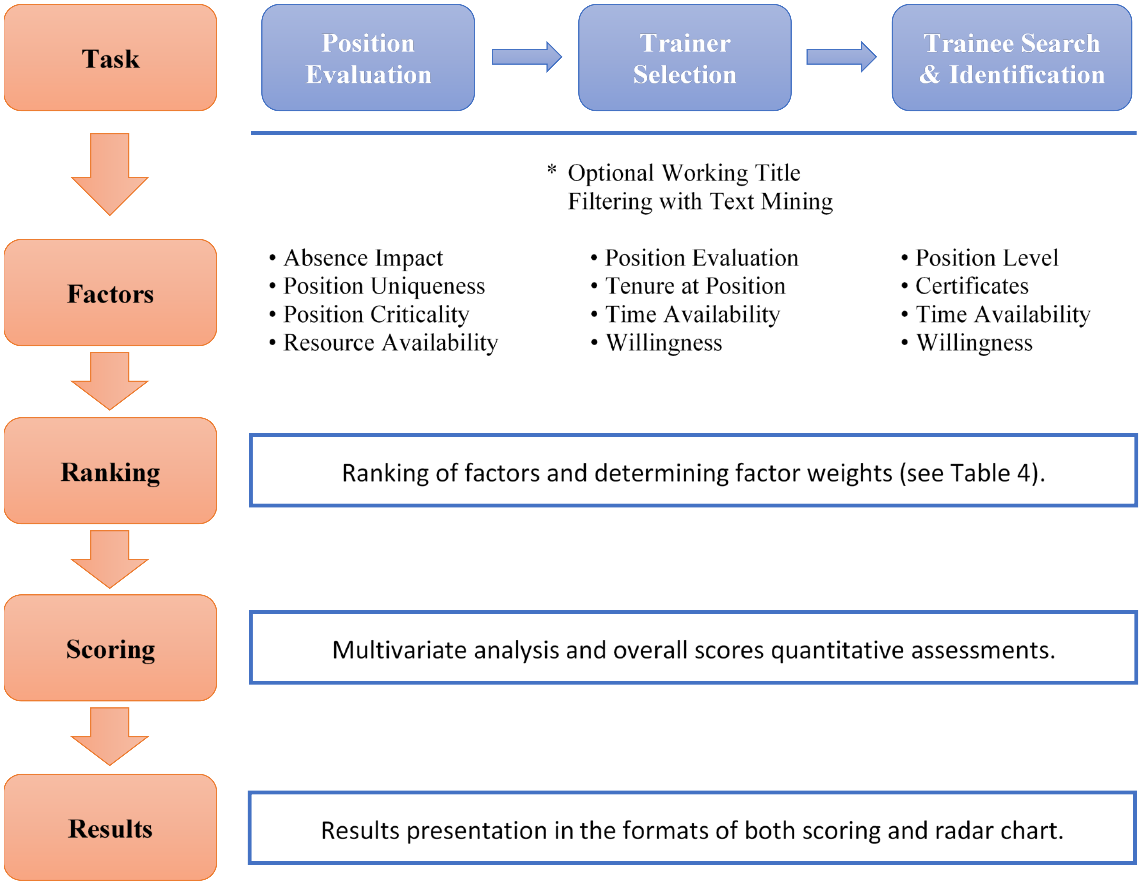
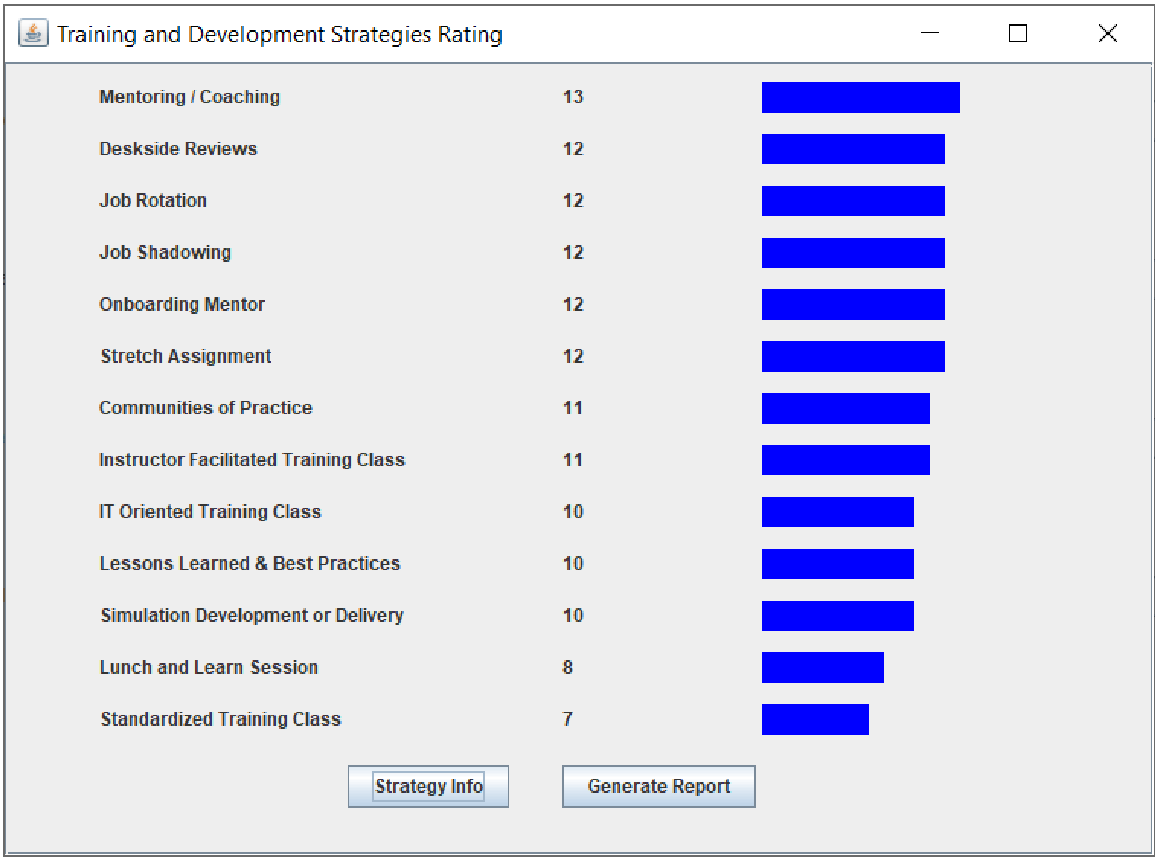


Figure 5.17 – Structure and evaluation workflow for the cross training module.

5.4.5 Training and Development

Training and development (T&D) module was developed to allow Georgia DOT HR personnel to evaluate different T&D alternatives based on various environmental characteristics and contributing parameters as shown in Table 8. Default coefficients provided in the table are set in the module to quantify the effectiveness of each T&D method, and users have the option to change these coefficients through a dedicated interface. Based on the selections made, the module provides suggestions for preferred T&D methods, ranked from most suitable (i.e. highest score) to the least suitable for the task, with a bar chart for visualization as shown in Figure 5.18.



Mentoring / Coaching	13
Deskside Reviews	12
Job Rotation	12
Job Shadowing	12
Onboarding Mentor	12
Stretch Assignment	12
Communities of Practice	11
Instructor Facilitated Training Class	11
IT Oriented Training Class	10
Lessons Learned & Best Practices	10
Simulation Development or Delivery	10
Lunch and Learn Session	8
Standardized Training Class	7

Figure 5.18 – Sample output from the Training and Development module.

Table 8 – Environmental characters for Training and Development techniques and their effectiveness.

Training and Development Strategies Environmental Characterization	Lessons Learned & Best Practices	Communities of Practice	Instructor Facilitated Training Class	Lunch and Learn Session	Standardized Training Class	Information Technology (IT) Oriented Training	Deskside Reviews	Job Shadowing	Simulation Development or Delivery	Job Rotation	Mentoring / Coaching	Stretch Assignment	Onboarding Mentor
Only one knowledge source and one knowledge receiver	2	2	2	1	1	2	3	3	2	3	3	3	3
One knowledge source and several to many knowledge receivers	2	2	3	2	1	2	2	2	2	3	2	2	2
Knowledge source is available less than 8 hours a week	1	1	1	1	1	1	2	1	1	1	1	1	3
Knowledge source is available between 8 and 16 hours a week	2	2	2	2	1	2	2	2	2	2	3	2	2
Knowledge source is available more than 16 hours a week	3	3	3	3	1	2	2	3	2	3	3	3	3
IT structure exists to support / distribute knowledge	2	3	2	2	1	2	2	2	2	2	2	2	3
IT structure does not exist to support / distribute knowledge	1	1	2	2	1	1	1	2	1	2	2	1	1
Knowledge source and knowledge receiver are co-located	2	2	3	2	2	2	3	3	3	3	3	3	2
Knowledge source and knowledge receiver are not co-located	1	1	2	1	2	2	2	1	1	2	2	2	1
There is less than 3 months available for knowledge transfer to take place	2	2	1	2	2	2	2	2	3	2	2	1	1
There is 3 to 6 months available for knowledge transfer to take place	2	2	2	1	2	2	2	2	1	2	2	2	2
There is 3 to 6 months available for knowledge transfer to take place	1	2	2	1	1	1	3	3	1	3	3	3	1

5.5 Concluding Remarks

The HRDT system, as an integrated computational system, was designed for human resource department at Georgia DOT. It has both a user-friendly interface to navigate through and a set of powerful analytic and visualization features. HRDT applies the computational techniques and data science tools studied in chapter 4, and delivers a standalone desktop application. The system consists of backend computation engines, which communicate with HR database and perform computation intensive tasks, and frontend functional modules, which interact with end users and present results from computation engines.

Following is a summarization of HRDT's main functionalities: To better manage organizational human resource data; To visualize employees information in a more intuitive way with geographical information integrated to evaluate attrition risk distribution; To evaluate criticality, uniqueness and other attributes of employees, so as to identify the important employees with high risk of critical knowledge being lost, as well as to comprehensively identify suitable knowledge providers and recipients that allows for optimal knowledge transfer. The deployment and consistent use of the system is expected to bring benefits to Georgia DOT such as extending person-to-person knowledge sharing to an organization-wide manner, solving the ongoing staff crisis with proactive planned knowledge management activities in the long-term, increasing employee engagement and knowledge level, etc. The effectiveness of the system and validation studies will be discussed in the next chapter.

CHAPTER 6. SYSTEM EVALUATION USE CASES

This chapter focuses on the evaluation of the HRDT system proposed in Chapter 5, with the objective of proving the effectiveness and capability of the system. The validation of HRDT will be evaluated and studied through gathered feedback from the end users of the system, as well as through the summarization from a face-to-face meeting with the director of the human resource department at Georgia DOT, who is also the main stakeholder of the system.

6.1 Overview

To evaluate the effectiveness of a system can also be called system validation, which is defined as "*a set of actions used to check the compliance of any element such as a system, document, a service, a task, etc. with its purposes and functions*" according to IEEE. Almost for any new developed system, the validation process is necessary to check that the system meets the specifications and requirements as initially proposed, to ensure especially the core components of the system, as well as to guarantee the correctness for the expected users from various points of views. During the validation process, one of the most critical questions to ask is: "Are we developing the right system?", and the best ones to answer this question are the stakeholders and the actual end users of the system. Within this regard, collecting user feedback can be an effective and efficient way for system evaluation.

For a system where user engagement is the central part, good user experience is sometime considered the most critical aspects in system evaluation (Schrepp, Hinderks, & Thomaschewski, 2014). Questionnaire, or user experience (UX) questionnaire, as an important element in the life-cycle of a system, is also a simple, immediate but powerful approach for initial and lightweight user feedback collection (Laugwitz, Held, & Schrepp, 2008). With careful calibration, a well-designed questionnaire with meaningful results derived from user feedbacks is capable of helping system developers to verify design ideas, to avoid any confusion, as well as to keep the system usable and effective. Moreover, user experience questionnaire can also reflect how users get involved with the system and in which aspect the users expected improvements, thus further enhancing the performance of the system.

While a questionnaire may have many different types of questions such as scaling system, multiple choice questions and short answer survey, users are typically less motivated about those that take a long time to response to. With this regard, other strategies and approaches can be implemented to collect more accurate and more profound feedback from users. Reaching out directly to the user, although can be potentially difficult to conduct, however is amongst the most effective ways for feedback collection. For example, email provides a fast and instant mean to reach out to the user, but it does not make a significant difference from questionnaire due to its lack of effective communication and interaction. A face-to-face meeting, on the other hand allows for in-person communications between the system developer and the actual stakeholder. Such an approach, due to its complicated setup procedures, i.e. scheduling time, preparing interview guidelines, etc., is sometime underestimated. As a matter of fact, huge benefits can result from meeting in-person for user feedback. A better communication usually leads to a deeper understanding on the user's comments and a more confident system evaluation. Some complicated feedbacks or any further demands can only be addressed in such an in-person meeting environment. Furthermore, the user is able to demonstrate any typical use cases of the system during the meeting, giving the developer feedback from multiple aspects. In the context of the HRDT system evaluation, both a questionnaire with a list of targeted questions, sent to the end users, and a face-to-face meeting with the stakeholder of the system has been conducted for collecting user feedback and evaluating system effectiveness.

6.2 User Experience Questionnaire

The first part of the system evaluation work is developing a questionnaire and sending to the end users. The questionnaire contains a list of 15 questions that covering general usability, user interface, as well as detailed functionalities and effectiveness for different components of HRDT. It is implemented with a scaling system where the user response with the degree to which he / she agree with each statement. This questionnaire is for use to collect user feedback from a broad of aspects, but not to take very long to complete. As followed is the list of the 15 questions in the questionnaire.

- Overall, I am satisfied with the usability and functionality of the HRDT system.
- I would recommend the HR department in Georgia DOT to deploy this system for frequent use.
- The system is easy to learn and simple to use for the users in the HR department on a regular basis.
- The graphic user interface of the system is clean and straightforward enough to navigate through, even for the users with limited domain knowledge.

- Overall, I am satisfied with the helping documents and supporting information that were provided along with the system.
- I expect that the functional modules in HRDT can contribute to solving the staff crisis and knowledge loss problems that Georgia DOT is facing.
- I think Microsoft PowerBI is a useful tool to be integrated in HRDT for analyzing and visualizing attrition risk distribution across multiple branch offices.
- I think the network analysis and visualization module is a great tool to evaluate and visualize employee’s criticality to the organization.
- I think that the network analysis and visualization module can expand HR decision making ability from “person to person” level to “organization-wide” level.
- I think the integrated visualization tools in the system can clearly convey the analysis result and effectively enhance users’ perception.
- The HRDT can help users to effectively identify critical employees with knowledge that are most at risk of being lost.
- The HRDT can help HR managers to choose the best knowledge sharing activities between the provider and recipient.
- The HRDT can assist users with rational decision making when choosing knowledge transfer mentors and proteges.
- I feel confident about addressing workforce related decision making based on the analysis result generated from HRDT.
- I feel confident that positive changes in Georgia DOT can be expected upon the continued deployment of HRDT, including increased knowledge transfer rate, higher employee knowledge level, lower employee turnover rate, etc.

The first 5 questions focus on the general-wise feedbacks from end users while the remaining ones have concentrations on the detailed functionality and effectiveness of each module. The questionnaire has a scoring mechanism as: a score of 5 indicates strongly agree, 4 indicates agree, 3 indicates neutral, 2 indicates disagree, and 1 indicates strongly disagree. According to the statistics on the feedbacks, every question in the questionnaire has received an average score of 4.5 or above. Such feedback results showed an overall user satisfaction about the HRDT system not only on the user interface design, but also on the usability and functionality.

6.3 Face-to-Face Meeting

The second part of the system evaluation work is to gather more in-depth feedback for a more comprehensive evaluation on the HRDT system through a face-to-face interview meeting with the deputy HR director at Georgia DOT, who is responsible for oversight of more than 20 employees in the following areas within human resource tasks for approximately 4,000 employees at the Georgia DOT: classification and compensation, safety, employee relations and district HR operations. The deputy director is the main point of contact in the development of HRDT, and is also one of the main stakeholders of the system. As having served as a lead project manager on workforce planning efforts, he is experienced with human resource data management and knowledge sharing works, thus, is able to give profound feedback for the system evaluation tasks.

The meeting was about hour-long, and was prepared with four main themes, each of which will be summarized in each of the following subsections, including an interview guideline that drives each theme, as well as the key points taken away from the meeting.

6.3.1 Evaluation on overall functionality and usability

The first theme of the meeting concentrated on the overall feedbacks about the functionality and usability of the HRDT system, serving as an entering point of the whole meeting. The following presents an interview guideline for this theme and the summarization of the key points discussed with the user.

Interview guideline

How are the intellectual assets managed in Georgia before and after the deployment of the HRDT system? Prior to the use of the system, how are training and development programs and / or knowledge retention activities conducted, how often do the activities take place? Is the system helping Georgia DOT seek for knowledge retention more proactively through well-planned activities? Is the system helping Georgia DOT with proactive workforce planning and succession planning, reducing the risk of critical knowledge being lost? Does the system make it easier and smarter regarding the selection of the most suitable candidates for knowledge transfer / sharing activities such as job shadowing and cross training?

Overall speaking, do you think that HRDT is fulfilling or will fulfill its proposed missions, such as increasing knowledge transfer rate, lowering employee turnover rate, rising levels of employee knowledge and motivation, keeping them more engaged and motivated about their work environment and career path opportunities.

Summarization of user response

The stakeholder said that he had been using the system on a more and more regular basis, currently more often than once a week, to explore human resource data in the organization. He mentioned that Georgia DOT has performance metrics for all units, and for the HR department, it is ensuring the ongoing work with every other office or district, while HRDT is supporting this role exactly. Prior to the deployment of the system, Georgia DOT is merely holding exit interviews as employees reported retirement or departure, which is unrealistic to capture their knowledge and experience developed over years at the current position. A potential use case, as was discussed during the meeting, is the use during the current outbreak situation of coronavirus, when there can be unprecedent sick leave and / or family emergency of important employees, where HRDT can provide proactive decision making on activities, such as cross training, to mitigate the impact of any unexpected employee leave. Moreover, the information technology department at Georgia DOT had tried to develop system similar to HRDT in the past but not nearly as user-friendly and effective as this system.

6.3.2 Evaluation on attrition risk distribution

The theme 2 of the meeting focused on the feedback with regard to the functional module of attrition risk distribution analysis and visualization. The topic also includes any feedbacks for the embedded MS Power BI.

Interview guideline

How is the human resource information managed and used by Georgia DOT's HR department before and after the deployment of the HRDT system? Prior to the use of the system, how does HR personnel query human resource database? How to look for the employees that are going to retire within a specific number of year? How to filter the employees with specific criteria such as working title, branch office, etc.? Do you think the attrition risk distribution module with embedded MS Power BI analytics tool is assisting users with these tasks? Are the features provided with tool, such as various filtering and ranking functions, pivot chart and table, spatial analysis and visualization, etc., enhancing the perception on the human resource information for the HR personnel, and potentially helping with better decision making process?

Summarization of user response

As mentioned by the user, the human resource information are contained in the HR database, however, there were no effective tools in Georgia DOT that could help with those functions needed to extract useful information for decision making. For example, one mentioned use case is that there are multiple criteria to consider when determining how many years until an employee is eligible for retirement. One of the criteria is the employee's age, and another one is the year of tenure. If the employee meet either one of the criteria or a combined of both on different thresholds, he / she is considered eligible for retirement. With this regard, when an HR manager tries to find out those expecting to retire within certain years, or to get an overview of workforce retirement projections, all the employees' information needs to be queried from the database and imported into spreadsheet for manual check. The automatically calculated "projected retirement date" feature provided in HRDT is deemed to be very helpful when the users are trying to get an overall picture of retirement projections.

Another use case commented was that when the HR personnel tries to figure out how many employees with Professional Engineer (PE) license are in the organization and / or in each of branch offices, furthermore, how many of these are expected to retire within the near future, say a year or so. The convenient filtering functions, interactive pivot chart and geographical information visualization map in Power BI make it very easy to perform HR tasks as such. The number of PE's with regard to the location of PE's, and their retirement projections can all be shown at the same time. This module is considered very helpful and beneficial when grasping an overall idea of any potential attrition, and its distribution across all branch offices.

6.3.3 Evaluation on network analysis and visualization

The theme 3 of the meeting focused on the feedback regarding the module of network analysis and visualization.

Interview guideline

Does the organization leadership evaluate the risk of employees' knowledge and experience being lost, and plan for knowledge transfer activities accordingly? How do HR leaders prioritize position importance and / or the risk of knowledge being lost before and after the deployment of HRDT, for the purpose of knowledge transfer? Were information flow and connectivity among employees considered for these tasks, and do the users think the employee criticality and uniqueness evaluation work in the network analysis module makes sense with the objectives and is helpful with the tasks? Are the users familiar with the two types of centrality evaluation, i.e. betweenness centrality and PageRank centrality, and the respective suitable scenario? Do you think the network view visualization, in addition to the traditional organizational chart, is a useful feature to help the engineering leadership with decision making?

Summarization of user response

The stakeholder commented that the network analysis and visualization module has been one of the most frequently used modules. A use case of the module was given where it facilitated preparation of a presentation slides deck relating to human resource matters. The presentation was given to the leaderships in the department of Roadway Design, and the audiences were surprised by the analysis results and network visualizations derived from the module. The network view, along with the node size denoting some graph metrics, not only gave them a global perspective in addition to individual perspective, but also opened up their mind in identifying candidates pool, as in the view of a traditional organization chart, everyone looks similar without sizing and people always tend to focus on the direct reports for candidates. The module could enhance data query, account for information flows among employee networks, as well as assist important position identifications.

The users also felt that the feature of uniqueness assessment is extremely useful as it is able to help mitigate, if not prevent, the adverse impact brought by unexpected employee departures or unprecedent temporary leave, through proactive cross training for unique positions. For example, during the current outbreak of coronavirus, unforeseeable absence can happen to any employee in Georgia DOT. If the person on leave is a highly unique individual, and his / her job needs to be taken over by someone else to ensure the ongoing of any essential business, proactive training can help get through the difficult situation by placing another well trained employee on this position temporarily. Furthermore, combined with the visualization technique, user can easily spot the employees that are relatively unique to the department.

6.3.4 Evaluation on text mining and multivariate analysis

The last theme of the meeting focused on the feedback on the knowledge management functional modules of HRDT including the position evaluation and candidate identification, as well as the practical usability of the text mining features on the job content similarity analysis.

Interview guideline

What factors are taken into consideration when choosing the employees with high risk of knowledge and experience being lost, when selecting the positions that are critical to the organization, and when identifying candidates for knowledge sharing before the use of HRDT? How do the managers usually make decisions on these tasks, and what is the candidate pool for knowledge sharing activities? Do you think the knowledge management modules with multivariate analysis could give a more comprehensive consideration on these tasks? Are the job description on the government website accurately reflecting the employee's job content, and are the job entry qualifications strictly enforced for all employees? Do you think the application of the text mining on job description and entry qualification with the objective of evaluating job content similarity between knowledge provider and recipient could help to identify ones that are more suitable?

Summarization of user response

According to the feedback during the meeting, there was not a very scientific system for the discussed tasks prior to the use of the HRDT system, and most candidate selections were performed in an office-specific level and most of the candidates come from those who directly report to the employee at the position being planned. With a combination of the network visualization and the use of the knowledge management module, the users realized that the most suitable candidates simply might not be from the direct reports, or not even work in the same department as the planned one, and sometimes they are just "around the corner", which was revealed by the use of HRDT. This module not only enables knowledge sharing across an organizational level, but also allows for much more comprehensive assessments with multiple parameters and their relative importance being considered. The module can help proactively prepare multiple backup employees for a specific critical position for the emergent scenarios.

With regard to text mining for pairwise job contents similarity analysis, the director mentioned that the job entry qualifications are indeed strictly enforced. For example, if a position requires three years of work experience and a PE license, then whoever is considered for that position needs to meet such requirements. As for the job description, although part of the details are temporarily missing on the official website, they can be a reasonable indication on employees' working content. The application of text mining technique seems to be very promising, especially after the missing parts are retrieved. For some knowledge sharing activities, such as succession planning, employees working on similar topics are preferred to be considered for candidacy, however, this is not the case for job shadowing, where the similarity is not a preference, and sometimes those who work on a very different domain can be selected. Afterall, the knowledge management modules have successfully extended knowledge sharing in Georgia DOT from a department level to an organizational level.

6.4 Evaluations During System Development

Several system evaluation and validation works that are worth noting were also performed during the design and development stages of the HRDT system. Feedback was constantly collected from the stakeholders and end users. The feedback was used to improve the delivered version of the system.

- Prior to the coding development of the HRDT system, a meeting with Georgia DOT's IT department was conducted for addressing any compatibility issues. Based upon their system requirements, security consideration and compatibility, a desktop version of the application was selected, and Java was determined to be the main programming language for the HRDT system.
- During the preliminary system design stage, a few fact-to-face meetings were held with Georgia DOT HR deputy director, along with his colleagues are the system's main end users. The meetings were focused on their needs and requirements from the system, which were in turn used as a guideline for the approaches adopted in the system. For example, a pilot study on the network analysis and visualization module with minimal functionalities was showcased to the HR personnel, and received highly positive feedback. More practical evaluative metrics, along with useful functionalities were then added, allowing for an enhanced and comprehensive working module as part of the whole system.

- Discussion meetings focusing on the attrition distribution visualization module were also held between the research team and GDOT HR personnel to determine the best analytics tool to be integrated into the HRDT system. Demos versions built with MS Excel, de.js web, and MS Power BI were presented for feedback, with MS Power BI being selected as the preferred option.
- Before going into details of the other knowledge management modules, the Job Shadowing module was firstly mocked and demoed in-person to GDOT HR personnel and a few district engineers from GDOT. The design of the 4-factor mechanism for HR metrics evaluations, integrated in a 3-part structure on the module workflow, was showcased for the very first time to the HR management domain experts for evaluation. Based on the highly positive feedback and some practical suggestions, an enhanced version of the evaluation algorithm was then coded and adopted into all four knowledge management modules. Furthermore, algorithm and workflow consistencies across all modules, allowing for the best user friendliness, were also favored by the end system users.

6.5 Concluding Remarks

The effectiveness and usability of the HRDT system have been evaluated through user feedback from both user experience questionnaire and face-to-face with the main stakeholder of the system. The questionnaire showed, at the high level, the satisfaction from the users about the system in multiple aspects including friendly user interface, ease to learn for inexperienced users, functionality and usability of the system. In the face-to-face interview meeting, the main stakeholder of the system has given several use cases, not only showing his satisfaction about HRDT, but also showed with detailed examples how the system is assisting HR department in Georgia DOT with their usual tasks on employee data and workforce information management. With those valuable feedbacks from actual end users, the system is considered effective in fulfilling its missions as it was proposed in the initial place.

The system evaluation presented in this chapter was conducted approximately one year after delivery of the system to Georgia DOT. Additional longitudinal studies over the next year or two can enhance both the evaluation and future development of the system. Although the evaluations of the system addressed in this chapter is rather minimal due to the short deployment period of the system in Georgia DOT, however, this initial feedback is valuable as it not only covers performance evaluation, but also uncovers new HR opportunities for the users. It also worth mentioning that the HRDT system is a winner of the “2019 AASHTO High Value Research Award”, which shows its huge potential for extending application to other organizations than Georgia DOT, or even to other domains of industries.

CHAPTER 7. CONCLUSIONS AND FUTURE WORKS

7.1 Conclusions

The objectives of this work included: (1) understanding the current HR data management and knowledge retention approaches in different industries; (2) summarizing the typical characters and the emerging challenges for public sector organizations in knowledge management; (3) studying computational techniques and data tools that can help HR data management for public sector organizations; (4) designing and developing an integrated HR data tool system for Georgia DOT HR Department. The main conclusions of this work are summarized below:

- Different types of organizations are usually characterized with different employee turnover rates. Companies with high turnover rates, such as high tech companies, suffer from higher level of experience and knowledge being lost, than those with low turnover rates, such as public sector organizations, thus are often implementing much more comprehensive knowledge retention protocols, which might not be suitable for public sector ones.
- Public sector organizations usually have a stovepipe-type organizational structure, which makes the flow of information restricted along top-down management hierarchical routes, as well makes recruiting, promoting, and other workforce planning on a local level. Public sector organizations are also facing the emerging challenges such as increasing workload, decreasing workforce, employee aging, and shorter employee tenure.

- An innovative HR data and organizational knowledge management approach is proposed with considering the unique characters existing in, and emerging challenges faced by the public sector organizations. The proposed approach consists of two main facets: (1) Identification of knowledge, which is not only critical to the organization but also at risk of being lost; (2) Transfer of important knowledge, by selecting the most suitable candidates and approaches.
- With the development of the computing ability, many computational techniques and data tools can assist with scalable and automated HR data management and organization knowledge retention: (1) Geographical information visualization helps enhance perception on HR database and organization-wide attrition distribution; (2) Social network analysis enables graph-based analysis on information flow, position criticality and uniqueness, further expands knowledge sharing from a "person to person" mode to an "organization wide" level; (3) Text mining of job descriptions extracts job features for similarity analysis, which allows for filtering working titles on candidate selection; (4) Multivariate analysis assists with HR decision-making by ranking and weighting multiple factors with Rank Order Centroid; (5) Temporal analysis can be potentially useful in evaluating knowledge retention effectiveness by monitoring organization metrics over time.
- An integrated computational system, HR Data Tools, is developed for use by the HR Department at Georgia DOT to strategically plan for a range of activities including workforce planning, job shadowing, cross training, etc. Integrated with the data tools mentioned above, the system backend communicates with the HR database in the organization, and performs a series of data manipulation activities, while the frontend

interacts with end users through a friendly interface, and also provides a set of modular tools incorporated with various visualizations.

- The HR Data Tools is anticipated to bring benefits to Georgia DOT such as increased knowledge transfer, lower employee turnover, higher levels of employee knowledge and motivation, amongst other. Based on the feedbacks from both a user experience questionnaire and an interview meeting with the HR Department deputy director, who is also the main stakeholder, the system is considered fully satisfactory to the users from the aspects of usability, functionality, effectiveness and user interface.

7.2 Major Contributions

The scientific contributions of this work not only lie in an unprecedented manner for visualizing an organization human resource database, and the innovative approaches for quantitatively evaluating metrics in the domain of human resource data management, but also rest with the development of a working computational system that integrates several existing data tools to solve new emerging human resource challenges faced by public sector organizations. Some detailed summarizations of the main contributions are listed below:

- This work proposed an unprecedented way to transform the classic organizational chart. The organizational chart has a limited range of abilities that usually focus on “who reports to whom” relationships within an organization. By generating a network style organizational chart from this reporting to relationships by showing employees as nodes and relationships as edges, the inventive chart is able to analyze and convey a much wider range of information that was hidden before. Several graph metrics that are able to describe employee properties have also been developed in this work. By integrating the network chart, the scale of these metrics can be displayed by the node size for intuitively indicating properties. For example, local betweenness centrality is used for evaluating employee's relative importance at the department level with an emphasis on strategic planning, while local PageRank centrality is used for importance evaluation but with a focus on tactical planning. Other metrics such as network local / global uniqueness, cosine dissimilarity, etc., are also innovatively developed for assisting organization human resource data management.

- This work developed innovative approaches for systematic and quantitative evaluations on a range of metrics for assisting decision making on knowledge transfer activities. For example, the Knowledge Loss Risk (KLR) was developed to quantify the risk of the organization losing a specific employee's knowledge and experience. The numeric value of KLR metric can be determined by a 4-factor mechanism including vacancy risk, criticality, uniqueness, and resource availability, each of which are scientifically derived from the information in the HR database. Moreover, an innovative workflow was designed for assisting several knowledge transfer programs such as succession planning, cross training, job shadowing, etc. Each one of the programs includes the implementation of the designed 3-part workflow for identifying optimal knowledge provider and recipient. For example, succession planning follows the 3-part steps of 1) KLR evaluation; 2) candidate search / experience evaluation; and 3) best candidate identification. Other programs also follow a similar 3-part workflow with different assessment criteria. The consistent design of this 4-factor mechanism and 3-part workflow is also partially driven by user friendliness. Furthermore, the whole design was also implemented in Java during the development of this research work.

- This work also includes a computational system, HRDT, that was not only designed and implemented, but also practically deployed in Georgia DOT. The system was designed to not only integrate existing data tools but also developed new approaches and frameworks, to enhance organization knowledge retention. More importantly, the system design philosophy has taken into consideration unique factors, such as stovepipe organizational structure, and emerging HR challenges, such as aging employees, increasing turnover rate, faced by a public sector organization. There were no similar systems developed prior the HRDT system, thus this novel system is also one of the main practical contributions of this interdisciplinary research work.

7.3 Recommendations for Future Work

Though in this work, some computational techniques and data science tools have been studied for the application in HR data management and knowledge retention, and an integrated system has been developed for Georgia DOT, some potential improvements and future works are recommended to enhance the current work:

- With social network analysis, an unweighted graph was built using the "reporting to" information in the HR database, in considering information flows amongst them. Despite the fact that the model accounts for employee connectivity among groupings, it can be further improved to a more complex and more comprehensive weighted graph if the actual communications, such as number of emails among employees can be measured and provided by the organization.

- Text mining was applied to the job description, which was retrieved from the Georgia DOAS (Department of Administrative Services) website, for similarity analysis based on cosine distance metric. While the pairwise similarity results reflect how similar the contents of two jobs are, more detailed descriptions of the daily duties associated with working titles, if it can be obtained and organized, would allow for more reasonable and more accurate analysis comparison results. Furthermore, other text similarity evaluation approaches may be tested to improve the performance.
- While the HR Data Tools uses existing HR database in Georgia DOT, and is designed and developed specifically for use there, a potential future work can be generalizing the system framework such that it can be applied in all the DOT's in the US, and even further extend to other public sector organizations where there is the need for HR data and knowledge information management.
- Many theories, techniques and tools that were applied to HR database for knowledge data management in this work can be extended to other domains of application in the future research work. For example, if the city or state infrastructures (roadways, bridges, dams, etc.) database is of interest, where the critical information such as year built, dimensions, etc. are contained, then these data tools are able to extract useful and meaningful information for objectives such as infrastructure health condition monitoring. Graph analysis also allows for viewing the infrastructures as connected graph for tasks like criticality assessments.

REFERENCES

Ahn, B. S. (2011). Compatible weighting method with rank order centroid: Maximum entropy ordered weighted averaging approach. *European Journal of Operational Research, 212*(3), 552-559.

Argote, L., & Ingram, P. (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational behavior and human decision processes, 82*(1), 150-169.

Bader, D. A., Kintali, S., Madduri, K., & Mihail, M. (2007). *Approximating betweenness centrality.* Paper presented at the International Workshop on Algorithms and Models for the Web-Graph.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances.

Bannur, S., & Alonso, O. (2014). *Analyzing temporal characteristics of check-in data.* Paper presented at the Proceedings of the 23rd International Conference on World Wide Web.

Barthelemy, M. (2004). Betweenness centrality in large complex networks. *The European physical journal B, 38*(2), 163-168.

Bavelas, A. (1948). A mathematical model for group structures. *Applied anthropology, 7*(3), 16-30.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *science, 323*(5916), 892-895.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology, 25*(2), 163-177.

Brantingham, P. J., & Brantingham, P. L. (1984). *Patterns in crime*: Macmillan New York.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine.

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology, 11*(1), 13-41.

Calo, T. J. (2008). Talent management in the era of the aging workforce: The critical role of knowledge transfer. *Public Personnel Management, 37*(4), 403-416.

Cong, X., Li‐Hua, R., & Stonehouse, G. (2007). Knowledge management in the Chinese public sector: empirical investigation. *Journal of Technology Management in China*.

Cong, X., & Pandya, K. V. (2003). Issues of knowledge management in the public sector. *Electronic journal of knowledge management, 1*(2), 25-33.

Dewah, P., & Mutula, S. M. (2016). Knowledge retention strategies in public sector organizations: Current status in sub-Saharan Africa. *Information Development, 32*(3), 362-376.

Doan, Q. M., Rosenthal-Sabroux, C., & Grundstein, M. (2011). A Reference Model for Knowledge Retention within Small and Medium-sized Enterprises. *KMIS, 2011*, 306-311.

Drucker, P. F. (1994). *Post-capitalist society*: Routledge.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37-37.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Frigo, M. (2006). Knowledge retention: A guide for utilities. *Journal ‐ American Water Works Association, 98*(9), 81-84.

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

Grant, R. M. (1996). Toward a knowledge ‐ based theory of the firm. *Strategic management journal, 17*(S2), 109-122.

Green, O., McColl, R., & Bader, D. A. (2012). *A fast algorithm for streaming betweenness centrality.* Paper presented at the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing.

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods. In: University of California Riverside.

Hansen, M. T. (1999). The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. *Administrative Science Quarterly, 44*(1), 82-111. doi:10.2307/2667032

Hansen, M. T., Mors, M. L., & Løvås, B. (2005). Knowledge sharing in organizations: Multiple networks, multiple phases. *Academy of Management journal, 48*(5), 776-793.

Izard-Carroll, M. D. (2016). Public sector leaders' strategies to improve employee retention.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm. *International Journal on Computer Science and Engineering, 2*(8), 2526-2534.

Jennex, M. E., & Durcikova, A. (2013). *Assessing knowledge loss risk.* Paper presented at the 2013 46th Hawaii International Conference on System Sciences.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl, 2*(6), 1930-1938.

Kaplan, B. (2013). Capturing, retaining, and leveraging federal agency workforce knowledge. *Public Manager, 42*(3), 27.

Katz, D., & Kahn, R. L. (1966). *Ths Social Psychology of Organizations*: Wiley.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39-43.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods, 21*(3), 733-765.

Laugwitz, B., Held, T., & Schrepp, M. (2008). *Construction and evaluation of a user experience questionnaire.* Paper presented at the Symposium of the Austrian HCI and Usability Engineering Group.

Lawton, A., & Rose, A. (1991). *Organisation and management in the public sector*: Pitman.

Levallet, N., & Chan, Y. E. (2019). Organizational knowledge retention and knowledge loss. *Journal of Knowledge Management*.

Levy, M. (2011). Knowledge retention: minimizing organizational business loss. *Journal of Knowledge Management*.

Li, G., & Cheng, Q. (2013). Summary of Researches on Enterprise Tacit Knowledge at Home and Abroad [J]. *Information Research, 3*.

Liebowitz, J. (2004). Bridging the knowledge and skills gap: Tapping federal retirees. *Public Personnel Management, 33*(4), 421-448.

Liebowitz, J. (2008). *Knowledge retention: strategies and solutions*: CRC Press.

Lu, H., & Yang, C. (2015). Job rotation: an effective tool to transfer the tacit knowledge within an enterprise. *Journal of Human Resource and Sustainability Studies, 3*(01), 34.

Marsh, S. J., & Stock, G. N. (2006). Creating dynamic capability: The role of intertemporal integration, knowledge retention, and interpretation. *Journal of Product Innovation Management, 23*(5), 422-436.

Melville, P., & Sindhwani, V. (2010). Recommender systems. *Encyclopedia of machine learning, 1*, 829-838.

Mintzberg, H. (1973). The nature of managerial work.

Olkin, I., & Sampson, A. R. (2001). Multivariate analysis: overview.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science, 28*(6), 441-453.

Popescu, O., & Vo, N. P. A. (2014). *Fast and Accurate Misspelling Correction in Large Corpora.* Paper presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Ranjan, J., Goyal, D., & Ahson, S. (2008). Data mining techniques for better decisions in human resource management systems. *International Journal of Business Information Systems, 3*(5), 464-481.

Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly, 48*(2), 240-267.

Rehman, S. (2012). A study of public sector organizations with respect to recruitment, job satisfaction and retention. Global Business & Management Research, 4 (1), 76-88. In.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35): Springer.

Roszkowska, E. (2013). Rank ordering criteria weighting methods–a comparative overview.

Sadath, L. (2013). Data Mining: A Tool for Knowledge Management in Human Resource. *International Journal of Innovative Technology and Exploring Engineering, 2*(6), 2278-3075.

Schmitt, A., Borzillo, S., & Probst, G. (2012). Don't let knowledge walk away: Knowledge retention during employee downsizing. *Management Learning, 43*(1), 53-74.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014). *Applying the user experience questionnaire (UEQ) in different evaluation scenarios.* Paper presented at the International Conference of Design, User Experience, and Usability.

Sumbal, M. S., Tsui, E., See-to, E., & Barendrecht, A. (2017). Knowledge retention and aging workforce in the oil and gas industry: a multi perspective study. *Journal of Knowledge Management*.

Sun, X., Liu, X., Hu, J., & Zhu, J. (2014). *Empirical studies on the nlp techniques for source code data preprocessing.* Paper presented at the Proceedings of the 2014 3rd International Workshop on Evidential Assessment of Software Technologies.

Sureeyatanapas, P. (2016). Comparison of rank-based weighting methods for multi-criteria decision making. *Engineering and Applied Science Research, 43*, 376-379.

Swap, W., Leonard, D., Shields, M., & Abrams, L. (2001). Using mentoring and storytelling to transfer knowledge in the workplace. *Journal of management information systems, 18*(1), 95-114.

Taylor, H. (2005). A critical decision interview approach to capturing tacit knowledge: Principles and application. *International Journal of Knowledge Management (IJKM), 1*(3), 25-39.

Tichy, N. M., Tushman, M. L., & Fombrun, C. (1979). Social Network Analysis for Organizations. *The Academy of Management Review, 4*(4), 507-519. doi:10.2307/257851

Walsh, J. P., & Ungson, G. R. (1991). Organizational memory. *Academy of management review, 16*(1), 57-91.

Wang, J., & Jin, Z. (2004). Tacit Knowledge: The Wellspring to Sustainable Competitive Advantage [J]. *Nankai Business Review, 5*.

Wellman, B. (1983). Network analysis: Some basic principles. *Sociological theory*, 155-200.

Zhang, Z. J. (2017). Graph databases for knowledge management. *IT Professional, 19*(6), 26-32.