Multi-Level Knowledge Extraction and Modeling to Support Job Hazard Analysis Process for Oil and Gas Pipeline Projects

by

Shadi N A Altawil

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In
Construction Engineering and Management

Department of Civil and Environmental Engineering
University of Alberta

ABSTRACT

Construction projects are impacted negatively by construction safety incidents. Job hazard analysis (JHA) process is a critical process component of safety management system in the construction industry. The JHA process is a planning process that aims to address potential hazards associated with execution of construction activities. It involves collecting knowledge from several safety knowledge resources. Explicit resources such as safety manuals, safety codes and regulation, and safety best practices are the primary input knowledge. In addition, tacit safety knowledge that is related to the experience of construction professionals is a critical knowledge component that feeds into the JHA process. JHA documents is the output of each JHA process for each construction activity.

The construction industry is a very dynamic and complex environment. Collecting knowledge to perform JHA process requires time and significant efforts. Construction personnel do not have the same experience and ability in identifying construction hazards. In addition, new construction manpower is continually joining the workforce and they lack sufficient experience and knowledge required for hazard identification. Previous JHA documents, which were prepared in previous projects, contain valuable knowledge related to construction hazards. Currently, documents are scattered and not reused for future JHA processes.

Oil and Gas Pipeline Projects consist of risky construction activities that involve dynamic interaction between humans, heavy construction equipment, heavy material, and the complex surrounding environment. Currently, safety research related to nonbuilding construction projects is not sufficient. Nonbuilding projects such as

pipeline construction and complex infrastructure need research focus due to their execution complexity and high potential risks.

This research aims to introduce a method for hazard knowledge extraction and modeling to assist and make the JHA process more consistent and systematic. To reuse the hazards knowledge embedded in JHA forms, multi- levels of knowledge extraction are performed. Text mining is used to organize documents in classes by adopting two stages of machine learning algorithms, clustering and classification. Moreover, JHA forms' contents were analyzed to extract hazards' concepts and relationships to build a hazard dictionary and knowledge schema. Text mining for concept extraction is used along with qualitative approach to build hazard dictionary. Ontology modeling is used to model the extracted knowledge schema. The model aims to represent the knowledge concepts, taxonomies, and semantic relationships. The knowledge model will support the JHA process by enabling retrieval and communication of hazards knowledge in future projects.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

Construction projects involve processing a large amount of important knowledge that is embedded inside different construction documents. Knowledge is produced over time as a normal output of many management processes such as planning, contract administration, quality control, and safety management. Most of the knowledge content is scattered and not structured in an organized way that can enable reusing it in future management processes. The construction industry is still far from utilizing knowledge assets in efficient ways. Management systems in organizations suffer leaking and losing diverse types of valuable knowledge (Rezgui, 2007).

Safety is a critical subject in construction projects due to its high impact on construction manpower, construction progress, and quality. The construction industry suffers a high rate of fatalities caused by incidents occurring during construction project execution (Hallowell, 2012). Construction incidents cost construction companies a significant amount of money and time and have an adverse impact on execution productivity (Kartam, 1997). Also, construction incidents have major impacts on safety reputation of companies, which could affect construction companies' qualifications in winning future projects.

Safety knowledge flow through safety management systems and processes, and they are dynamically evolving over time. Safety knowledge assets are critical for supporting safety processes that we count on for decreasing construction incidents and preventing construction workers from repeating past mistakes (Carter and Smith, 2006). Safety knowledge is either explicit or tacit. Explicit knowledge is in the form of written documents that is stored on companies' servers. Tacit knowledge is the knowledge related to workers' experiences and it is expressed in their actions and decisions (Carrillo and Chinowsky, 2006). Tacit and explicit safety knowledge are crucial to identify hazard in different hazard identification stages. Construction hazard identification is the main process in any safety management system, and it requires suffiecint knowledge input to identify hazards associated with construction execution.

This research aims to introduce new and integrated methodologies of knowledge extraction and modeling to support the hazard identification process for construction pipeline projects. Enabling retrieval of hazard knowledge during the JHA process will improve the efficiency of the process and contribute to decreasing construction incidents.

1.2 Research background

Although construction organizations have accumulated different and diverse types of safety knowledge through the execution of the previous projects, they do not reuse this knowledge for future safety management processes. This situation prevents construction workers from benefiting from past knowledge to eliminate or mitigate risks in current or future projects (Hallowell, 2012). Not learning from previous

knowledge is one of the main reasons behind the repetitive occurrences of construction incidents due to repeating the same past mistakes.

Because hazard identification process is the backbone of any safety management system , slight improvements in this process will significantly impact safety pe vcrformance in construction projects. Safety knowledge resulting from construction projects, explicit and tacit, exists mostly in separated and isolated management systems. These systems do not support proper information storing, retrieving, and sharing.

Construction projects lack an efficient identification of hazards associated with construction activity execution. By comparing the number of hazards identified for construction activities and what should have been identified and assessed, hazard identification level is low and required significate enhancement (Carter & Smith, 2006). Moreover, the hazard identification process is time-consuming and mostly depends on performing brainstorming sessions to quantify potential hazards associated with construction activities (Wang & Boukamp, 2009).

Researchers recognized the importance role of the JHA analysis process in improving safety performance. Hazard identification improvement strategies were developed using several approaches. Case based reasoning was used to retrieve similar cases of JHA documents and relevant incident cases to assist in future hazard identification process (Goh & Chua, 2010). A web-based system was introduced to store safety information which is related to regulation and best practices to enable retrieval of safety rules required for execution of construction activities (Kamardeen, 2013).

JHA knowledge stored in previous JHA documents was represented by ontology to support the hazard identification process (Wang & Boukamp, 2011). Chi, et al. (2014) created an ontology model that includes safety documents resources such as safety regulation and standards. Moreover, ontology modeling is used to model safety knowledge using safety regulations and safety best practices and enables integration with building information modeling (BIM) to automate safety planning using a BIM environment (Zhang, et al., 2015).

However, the knowledge structure used for building job hazard knowledge models consisted of activity, activity steps, hazards, and controls. It is based on the explicit structure found in JHA documents and not built based on knowledge domain analysis. Knowledge analysis is crucial to extract knowledge schema that can be modeled using knowledge modeling tools such as ontology (Gasevic, et al., 2009). Addressing embedded knowledge concepts and extracting semantic relationships between them, can leverage the structure of the knowledge and improve the performance of an ontology model. Extraction of semantic relations between concepts and entities can improve the retrieval process of hazards and its related control measures information.

Safety knowledge is mostly a text-based knowledge and is represented and communicated in text format. Text mining is a promising approach that can help in processing text of safety knowledge to perform classification, retrieval, and concepts extraction. Using text mining that includes natural language processing and Machine learning can help in quickly extracting structured useful information from unstructured texts (Tixier, et al., 2016).

Text classification was used to identify safe approaches using knowledge in documents stored in databases and contained predefined safety violation scenarios (Chi, et al., 2014). Zhang, et al. (2016) used text classification to identify classifications of hazardous actions explicitly recorded in crash reports based on the narratives given in the reports. Project document grouping is one of the key areas of application of text mining (Al Qady & Kandil, 2015). Knowledge concept extraction is another area of applications of text mining. Although text mining is very useful in extracting concepts, entities, and semantic relations, it cannot be fully automated and requires checking and filtering by knowledge domain experts during the extraction process.

Oil and gas pipeline projects consist of fewer activities in comparison with building projects. However, pipeline projects involve dynamic and complex execution environments. For example, activities like excavation, backfilling, pipe stringing, and hydro-testing cannot be done in a controlled environment and are highly affected by weather and ground conditions. Safety research related to nonbuilding construction projects are not sufficient. Nonbuilding projects such as pipeline construction and complex infrastructure need more research focus due to their execution complexity and high potential risks (Zhou, et al., 2015).

1.3 Research objective

The main objectiveof this reserch is to enhance and support the JHA process through reusing previous knowledge stored in the past JHA documents. To reuse the safety knowledge, knowledge extraction and knowledge modeling are essential. Enhancing the JHA process aims to decrease repeated incidents, improve safety performances,

and promote a safety culture. The following objectives were developed and updated periodically during the development of the research:

1- Develop and evaluate documents categorization model to organize the JHA documents in classes that are related to pipeline construction activities;
2- Analyze the JHA documents to extract knowledge concepts, semantic relations, and build knowledge schema;
3- Develop and evaluate ontology models to represent extracted knowledge schema and enable hazard knowledge communication for more consistent and systematic JHA process.

1.4 Research methodology

To accomplish the research objectives, a six stages methodology (Figure 1-1) was developed and implemented as follows:

Stage 1: Survey the literature and investigate pipeline projects incidents to identify the research gap.

Stage 2: Collect the JHA documents for oil and gas pipeline projects, and explore the text inside the forms.

Stage 3: Build integrated model to categorize JHA documents using clustering and classification algorithms. Clustering process output is an input into the classification model.

Stage 4: Use text mining and qualitative approaches to analyze and extract knowledge concepts, semantic relationships, and knowledge taxonomies.

Stage 5: Representation of extracted knowledge schema and hazard dictionary using ontology model to enable knowledge communication and retrieval to assist the hazard identification process.

Stage 6: Present research conclusion, contribution, limitation and future recommendation.

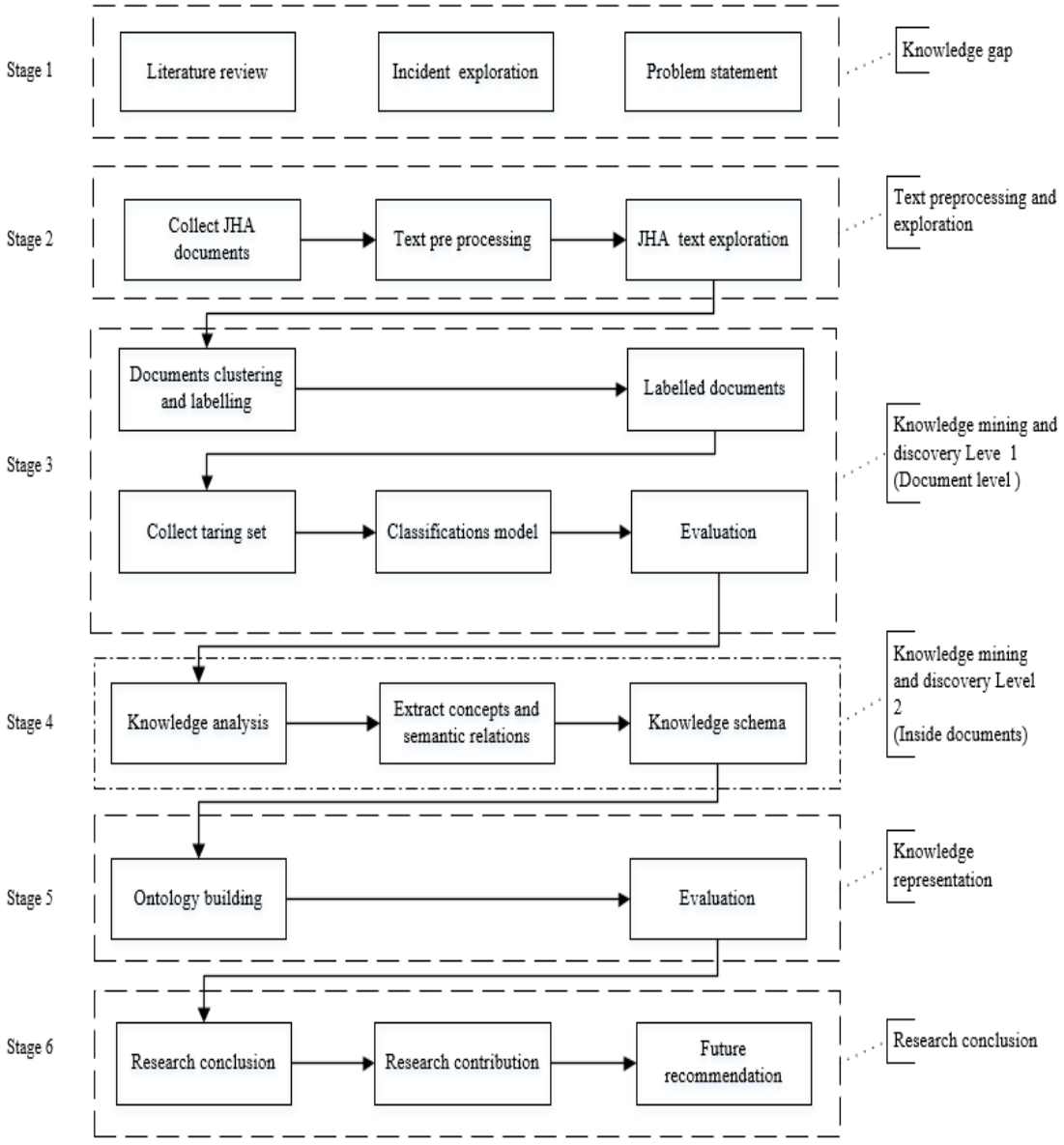


Figure 1-1 Methodology of the research

1.5 Research organization

Chapter 2 presents a literature review of related safety research. Safety knowledge extraction and representation related research is highlighted. Past incident cases, collected from previous pipeline projects construction projects were explored. A review of research concerned with assisting JHA process was presented. Text mining approach and its different area of applications is reviewed.

Chapter 3 focuses on classifying JHA documents using two stages of categorization: The first stage uses clustering technique to group and label unknown documents; the second one uses labeled documents from stage one as a training set of data to build a classification model. Related research to text documents clustering and classification are presented. Models evaluation and outputs are analyzed and discussed.

Chapter 4 represents knowledge concepts extraction from documents for the purpose of hazard knowledge mapping. A mix of quantitative and qualitative techniques is used for knowledge concept extraction. Natural language processing, tokenization, is used to automatically extract words and collocation from a collection of JHA forms. Co-occurrence analysis is used to extract semantic relations between hazards' concepts.

Chapter 5 illustrates hazard knowledge representation and modeling. The Ontology modeling approach is illustrated. Semantic web technology and its advantages are presented. Related research about safety knowledge representation is reviewed. The implementation method of ontology are explained. Finally, validation of ontology is discussed.

Chapter 6 contains research conclusions, research contributions, and potential future research directions.

CHAPTER 2

Safety knowledge management and the job hazard analysis process

2.1 Introduction

The construction industry suffers a high rate of fatalities caused by incidents occurring during construction projects execution. Construction incidents cost construction companies a significant amount of money and time and have a negative impact on execution productivity (Kartam, 1997). Also, construction incidents have major impacts on companies' safety reputation, which could affect construction companies' qualifications in winning future projects.

Construction companies allocate enormous efforts and finances to enhancing safety performance and decreasing incidents rate. Efforts are distributed to safety training and education, safety tools, hiring safety specialists, and implementing advanced safety management systems.

Safety information and knowledge flow through Safety management systems and processes, which dynamically evolve over time. Safety Knowledge assets are critical for supporting safety processes that we count on for decreasing construction incidents and preventing construction workers from repeating past mistakes (Carter & Smith, 2006).

The construction industry is still far from utilizing knowledge assets in an efficient way. Management systems in organizations suffer leaking and losing valuable different types of knowledge (Rezgui, 2007). Safety knowledge is either explicit or tacit knowledge. Explicit knowledge is scattered all over the system in the form of documents stored in companies' servers. Tacit knowledge is knowledge related to workers experience that is expressed through their actions and decisions (Carter & Smith, 2006). To utilize tacit knowledge, capturing mechanism of the knowledge should be embedded in the safety management system to enable knowledge retrieval and sharing.

Most of safety knowledge is in text formats such as safety manuals and/or different safety forms and checklists. JHA documents are one of the most important safety documents in the construction field. they are the output form of the JHA process and contain identified hazards and their controls associated with construction activities. Hazard identification is the primary process in any safety management system. It requires adequate knowledge supporting to identify construction hazards.

Unfortunately, the uses of hazard knowledge embedded in older JHA documents for benefiting current and future JHA processes is very limited. To take advantage of past JHA forms, documents content must be extracted, mapped and formalized in a way that can enable hazard communication among construction professional.

In this chapter, knowledge management and text mining is explored. An overview of job hazard analysis processes in pipeline construction projects is reviewed. Exploration of JHA forms is conducted using a text mining approach. Examples of extracted patterns from texts are presented.

2.2 Safety in Canada

The construction industry is one of the most significant contributors to Canada's Gross domestic product (GDP). The construction industry suffers a significantly high rate of fatalities. Incident records and statistics produced by the Association of Workers' Compensation Board of Canada for the years 2013, 2014, and 2015 showed that the construction industry has the highest rate of fatality incidents in comparison to other industry sectors (see Figure 2-1).

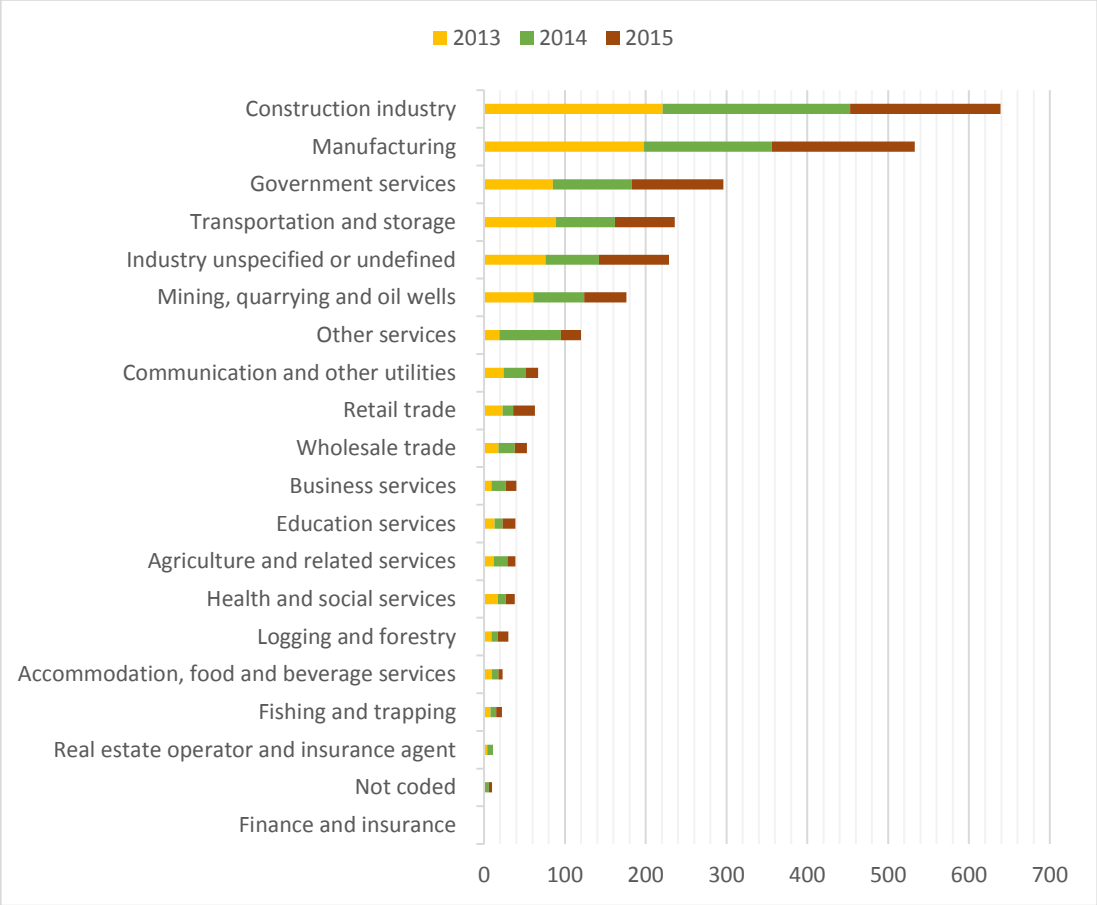


Figure 2-1 Fatality statistics by sector (Association of Workers' Compensation Boards of Canada, 2015)

2.3 Knowledge and construction

Construction Knowledge management is a discipline that focus on integrating the processes of creating, capturing, sharing, and retrieving knowledge in a particular domain of construction (Lin & Lee, 2012). Construction and engineering knowledge are essential input to construction management process operations.

Construction project consists of management processes and operations. Several knowledge resources are necessary to commence planning and execution of construction project phases. Construction knowledge can be classified as follows (Tatume, 2011):

- Technical fundamental (including design aspects and technical specification of construction processes)
- Construction material (includes knowledge of material specification which covers technical knowledge related to design, construction, and maintenance processes)
- Construction applied resources (includes knowledge related to equipment, humans, and methods).
- Construction operations (includes all operations and process management that are implemented to achieve construction product).

Knowledge is classified as explicit and tacit; explicit knowledge is formal and presented in an organized way. Tacit knowledge is embedded in people minds and appears in personal experience, decisions, and skills. Knowledge management is crucial and considered as important driving factor for sustainable growth of

construction companies (Yusof & Bakar, 2012). The information technology evolution will help construction organizations in automation of knowledge management processing to enable reliable advanced methods of knowledge capturing, transferring, and sharing.

2.4 Construction safety knowledge

The construction safety domain has attracted many knowledge management researchers. Construction safety is an important division in the construction industry due to its critical impact on construction performance and human lives. Safety knowledge consists of several resources such as safety manuals, safety codes and regulation, safe work practices, incidents reports, and other safety documents related to different safety reports and forms.

Accessing accurate safety knowledge in a quick and efficient manner can improve safety performance and promote safety culture. Safety knowledge is one of the essential inputs for the construction execution stage. Prior to starting any construction activity, a hazard analysis process must be conducted to make sure all construction workers are well informed about risks involved during the execution (Goh & Chua, 2010).

Knowledge management comprises of essential process stages required for building a robust knowledge system (Hallowell, 2012).These stages are Knowledge acquisition, storing, and transfer (see Figure 2-2). Knowledge acquisition is the process of knowledge mining to extract, filter, and refine information from diverse knowledge resources. Knowledge storing is formalizing the knowledge in a storage using specific structures to represent knowledge concepts.

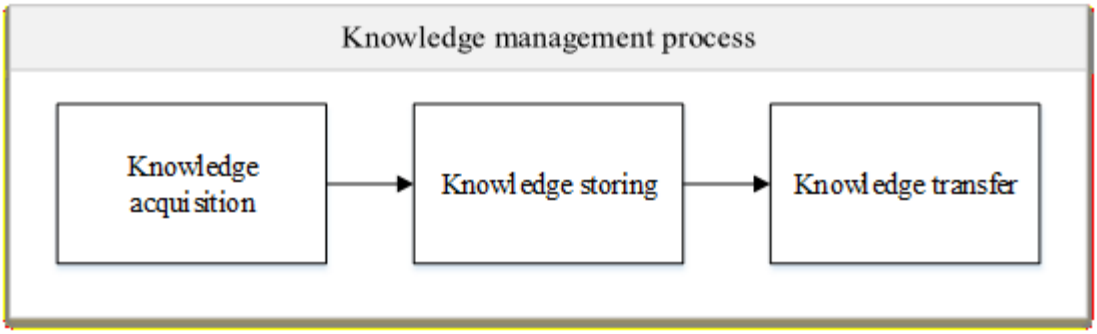


Figure 2-2 Knowledge processing steps

Knowledge transfer is the process of accessing, reliving, sharing, and communicating knowledge among either people or computer machines. Proper storage of safety knowledge can enable efficient retrieval of safety knowledge and integrate it into other construction management processes.

Improved safety knowledge communication contributes to promoting safety culture and enhancing construction projects performance and a company's reputation (Rezgui, 2007). Kartam (1997) presented a model system to represent safety codes and regulations using a relational database. The information model of safety knowledge was integrated into a CPM schedule to enable contractors to be able to plan, monitor, and control their projects' safety performance.

Chua & Goh (2002) used case-based reasoning to represent incident cases and safety plans cases to benefit safety planning process. Hadikusumo & Rowlinson, (2004) designed a tool to capture the safety tacit knowledge. The tool depends on capturing the interaction between safety officers and a 3D model of site construction. Captured Knowledge contributes to educating and enriching the safety knowledge of construction workers.

Hallowell (2015) investigated how the construction industry is employing knowledge management strategies. He concluded that construction contractors could acquire safety knowledge and store it in paper or software format. However, construction contractors lack an efficient knowledge management system for storing and sharing safety data with construction workers.

Wang and Boukamp (2011) developed a frame of work to represent hazards associated with construction activity to enhance the job hazard analysis process. The frame is limited to describing safety rules related to each activity or job step. Chi, et al. (2014) developed a text classification model to assist the job hazard identification process using an ontology concept. The model depends on storing and retrieving explicit safety knowledge that exists on OSHA standards and NIOSH face reports.

Le, et al. (2014) developed a social network system for sharing safety information with the construction team. The system consists of a data information model, a knowledge base model, and a social network model to mimic social communication platforms to enhance safety knowledge sharing and transferring.

Integrating a knowledge management system in an OHS system to enable knowledge transfer and sharing, is a vital process to control, mitigate, and eliminate construction hazards (Fargnoli, 2011).

Several researchers developed models to integrate knowledge management systems into hazards identification and planning processes (Kim, et al., 2015).

Some researchers integrated designed safety ontology into BIM to link safety knowledge to 3D models and to add the ability of visualization during the hazard identification task (Zhang, et al., 2015; Melzner, et al., 2013). Kamardeen, (2013)

developed a web-based tool to retrieve safety information stored in the database to assist in hazard identification for construction activities.

Table 2-1 is tabulating recent research related to safety knowledge management formalization and integration into other management systems to assist in safety management planning and support construction hazard identification process. Hazard identification was supported by different methodologies such as retrieval of safety codes, manuals, safety best practices, and past incidents records.

Researchers recognize the importance of the JHA analysis process and its potential impact on improving safety performance. Hazard identification improvement strategies were introduced using several approaches. Case base rezoning was used to retrieve similar cases of JHA documents and relevant incident cases to assist in future hazard identification process (Goh & Chua, 2010). The web-based system was introduced to store safety information related to regulation and best practices to enable retrieval of safety rules required for execution of construction activities (Kamardeen, 2013).

Table 2-1 Recent research related to Safety Knowledge

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process, store, and retrieve knowledge	Objective of developed system/ model
1	Information Retrieval Framework for Hazard Identification in Construction (Kim, et al., 2015)	Built retrieval system to retrieve past incident cases to benefit current work risk assessment, integrating BIM and PMIS, DB	Retrieval	Text	Historical incidents information	BIM, DB	Assist safety planning
2	Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA) (Zhang, et al., 2015)	Formulated safety knowledge (regulation, hazards) using ontology and integrate with BIM to help analyze hazards associated with construction activity	Representation and revival	Text	OSHA	Ontology, BIM	Assist safety planning

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process store, and retrieve knowledge	Objective of developed system/ model
3	Using ontology-based text classification to assist Job Hazard Analysis (Chi, et al., 2014)	Using text classification to retrieve useful safety document to support the JHA Process	Representation and revival	Text	OSHA, historical fatality cases	Ontology	Job Hazard Analysis
4	A social network system for sharing construction safety and health knowledge (Le, et al., 2014)	a system to integrate ontology, social networks, and the web to represent, refine, and transfer safety knowledge.	Representation and revival	Text	Historical incidents information	Ontology	Assist safety planning
5	The use of ontologies for enhancing the use of accident information (Batresa, et al., 2014)	Represent past accidents access by ontology and enable reasoning retrieval for a chemical project	Representation and revival	Text	Historical incidents information	Ontology	Assist safety planning

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process store, and retrieve knowledge	Objective of developed system/ model
6	A case study on automated safety compliance checking to assist fall protection design and planning in building information models (Melzner, et al., 2013)	This paper presents a comparative case study based on an automated rule-based checking system for building information modeling (BIM). The scope of the work focuses on safety rule implementation of fall protection standards from two countries: Germany and the USA.	Representation and revival	Text	Safety regulation limited to text	DB	Assist safety planning
7	Automated Information Retrieval for Hazard Identification in Construction Site (Kim, et al., 2013)	Built retrieval system to retrieve past incident cases to benefit current work risk assessment, integrating BIM and PMIS, DB	Representation and revival	Text	Historical incidents information	DB	Hazard identification

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process store, and retrieve knowledge	Objective of developed system/ model
8	Exploratory Case Study of Pictorial Aids for Communicating Health and Safety for Migrant Construction Workers (Hare, et al., 2013)	Explore the influence of using safety pictorial aids to help new migrants in understanding safety rules	pictorial aids (visualization)	Images	Created to represent safety rule	-	Assist safety planning
9	EHS Electronic Management Systems for Construction (Kamardeen, 2013)	Developed a safety management system to support the job hazards analysis process using a web-based tool to enable communication between the project team and OHS	Representation and revival	Text, pictures, videos	Australian codes and regulation, work practices	DB	Assist safety planning
10	Developing a Versatile Subway Construction Incident Database for Safety Management (Zhou, et al., 2012)	Develop a relational DB to represent past incident, near miss, and unsafe behavior to help quantify risks in a subway construction project	Representation and revival	Text	Historical incidents information	DB	Assist safety planning

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process store, and retrieve knowledge	Objective of developed system/ model
11	Knowledge management integration in Occupational Health and Safety systems in the construction industry (Fargnoli, 2011)	Developed KM (relational database) and aim to support OHS systems	Representation and revival	Text	Multiple resources	DB	Assist safety planning
12	Ontology-Based Representation and Reasoning Framework for Supporting Job Hazard Analysis (Wang & Boukamp, 2011)	Built a safety representation model using ontology and reasoning mechanism to represent safety knowledge	Representation and revival	Text	78 JHA documents	Ontology	Job hazard analysis
13	Case-Based Reasoning Approach to Construction Safety (Goh & Chua, 2010)	Retrieve past hazard identification and incident case scenarios to benefit the current hazard identification process using case-based reasoning approach	Representation and revival	Text	Historical incidents information	Case- Based Reasoning Techniques	Job hazard analysis

S.N	Research Title	Description	Knowledge Operations	Knowledge Format	Safety Knowledge Resources	Tools used to process store, and retrieve knowledge	Objective of developed system/ model
14	Capturing Safety Knowledge Using Design-for-Safety-Process Tool, (Hadikusumo & Rowlinson, 2004)	Safety planning and training, using safety database, capture tacit knowledge from safety engineers	Representation and revival	Text and visualize virtual construction components	Safety regulation	DB	Assist planning

JHA knowledge stored in previous JHA documents was represented by ontology to support the hazard identification process, (Wang & Boukamp, 2011). Chi, et al. (2014) created an ontology model that included more safety documents resources such as safety regulation and standards. Moreover, Ontology modeling is used to model safety knowledge using safety regulations and safety best practices, and enables integration with building information Modeling (BIM) to automate safety planning using the BIM environment (Zhang, et al., 2015).

However, the hierarchy structure used for building job hazard knowledge models consists of activity, activity steps, hazards, and controls. It is based on the explicit structure found in JHA documents and not built on sufficient knowledge domain analysis. Knowledge analysis is crucial to extract knowledge schema that can be modeled using knowledge modeling tools such as ontology (Gasevic, et al., 2009). Addressing embedded knowledge concepts and extracting more semantic relationships between hazards entities can leverage the structure of the knowledge domain and improve the performance of ontology models. Moreover, extracting semantic relations between concepts and entities can improve the retrieval process of hazards and their related control measures information.

Most of the safety knowledge is in the text form. Using text mining for text knowledge extraction will enable improving automatic text extraction and pattern recognitions of safety knowledge. In this research, text mining will be employed to explore safety knowledge in JHA forms. The objective is to extract knowledge structure which can be used to build knowledge model. The model aims to improve the retrieval of

construction hazards to support JHA process in the domain of pipeline construction project.

2.5 Incidents data exploration

As one of the points of departure for the research, Pipline projects' incidents records were investigated. 412 incident reports were collected from several pipeline projects in Alberta. Incident events occurred in 2012 and the first half of 2013. Incidents causes are classified into two categories: immediate/direct causes and root causes. Direct causes are defined as risky or unsafe acts or situations that lead directly to incident events.

Hazardous conditions/situations are defined as the condition in which site location, physical layout, status of tools, material, or equipment are in violation of safety rules and standards. Direct causes of incidents include reasons such as failure to follow procedures, organization policy, and professional practices. Table 2-2 shows a list of different potential direct causes whch was extracted from incident records.

Indirect or root causes are factors that are initiated by a human or due to job nature and contribute directly to the cause of incidents (Abdelhamid & Everett, 2000). Root causes such as lack of knowledge or inadequate communication are common hidden reasons of construction incidents. Table 2-3 shows more examples of root causes that were extracted from collected incidents records.

Table 2-2 Incidents direct causes

Incident direct causes
Failure to Identify Hazard/Risk
Failure to Follow Procedure/Policy/Practice
Improper Position for Task
Defective Tools, Equipment or Materials
Failure to Check/Monitor
Improper Placement
Road Conditions
Failure to Secure
Failure to React/Correct
Weather Conditions
Using Equipment Improperly
Improper Lifting
Inadequate Guards or Barriers
Congestion or Restricted Action
Inadequate Information/Data
Improper Loading
Failure to Communicate/Coordinate

Table 2-3 Incidents root cases

Incidents root causes
Lack of Knowledge
Improper Motivation
Inadequate Engineering
Excessive Wear and Tear
Inadequate Maintenance
Lack of Skill
Inadequate Communications
Mental or Psychological Stress
Inadequate Tools and Equipment
Inadequate Leadership and /or Supervision
Inadequate Physical/Physiological Capability
Abuse or Misuse

The most frequent direct cause of incidents was failure to identify hazards (see Figure 2-3). Identifying hazards in construction is a critical in any construction safety management system. The level of hazard identification in construction projects related to the nuclear industry, railway industry, and projects within both railway and general construction industry were studied and evaluated (Carter & Smith, 2006). The hazard identification level was 89.9 % within the nuclear industry, 72.8 % within the railway industry, and 66.6% for railway and general construction.

Failure to identify hazards is a result of many factors related mainly to construction personnel. Hazard identification requires knowledge and experience of construction scope of work. In addition, it requires skills that are gained by training or performing construction jobs at site.

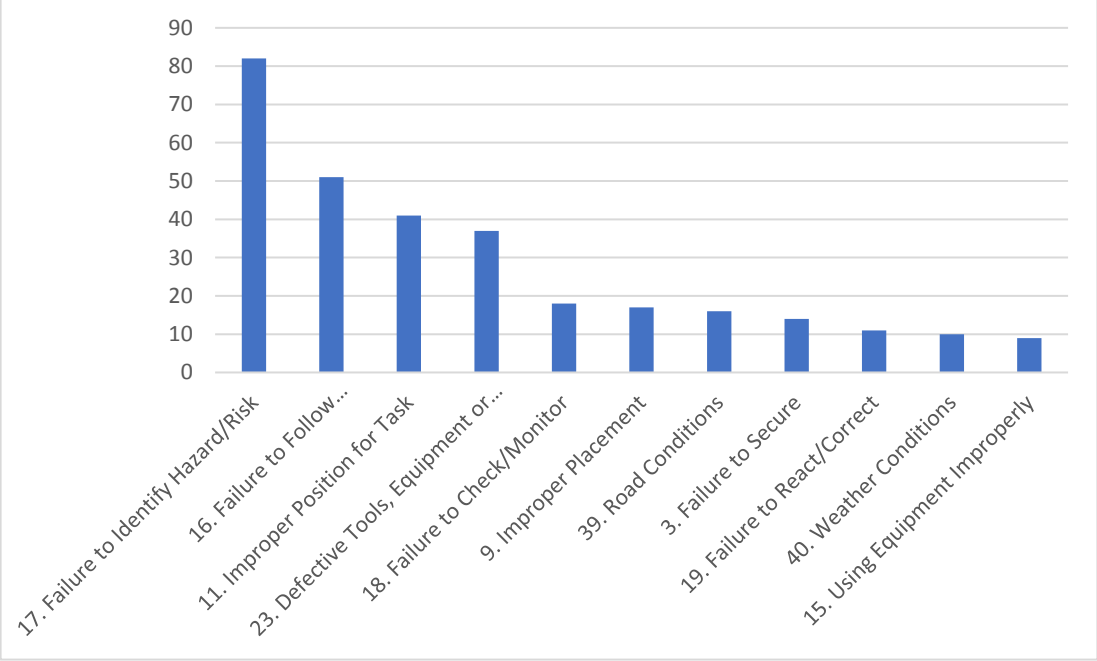


Figure 2-3 Direct causes of incidents

Investigating the root causes of the incidents records support this understanding since root cause lack of knowledge was the most frequent root cause of incidents (see Figure

2-4). Lack of knowledge is an explanation of the direct cause of failure to identify hazards.

The pipeline construction environment is very dynamic, where experienced workers and construction professionals are moving continuously from site to site or project to project. Consequently, their knowledge is also moving with them. At the same time, new people who enter the construction industry need to be oriented, trained, and supported by adequate safety knowledge prior to engaging with construction execution.

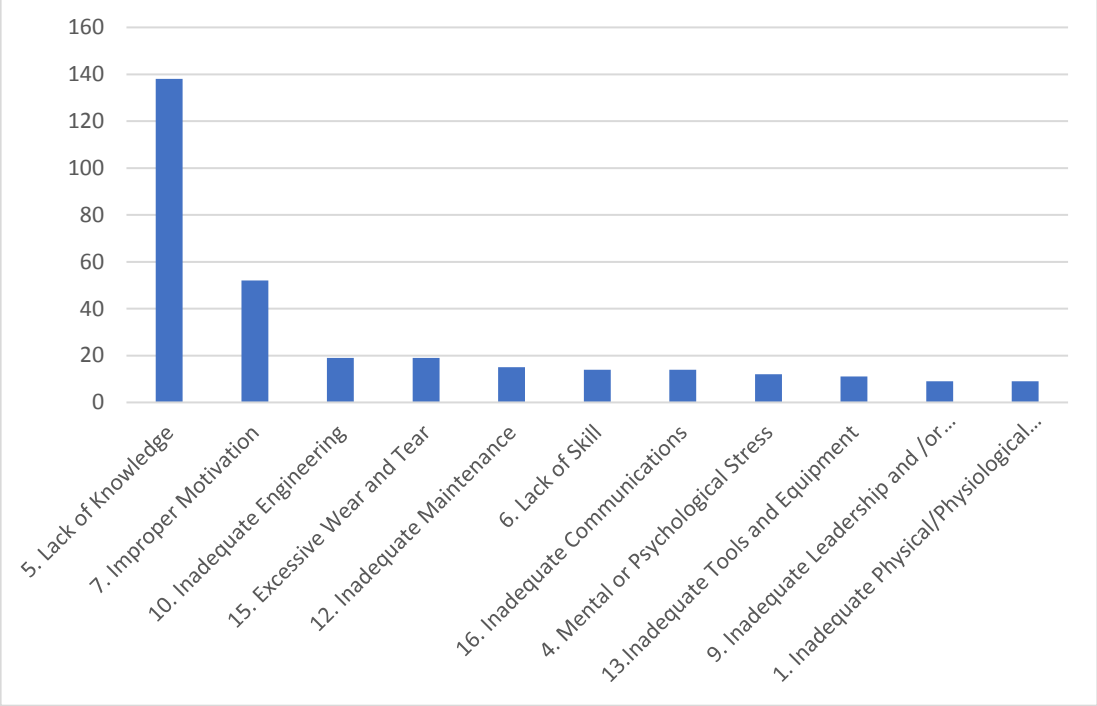


Figure 2-4 Root causes of incidents

2.6 Job Hazard assessment procedure

An effective hazard identification process can help workers avoid an accident that could cause injuries to workers and negatively impact construction productivity (Goh & Chua, 2010). Well-designed hazarded identification processes could contribute to

continuing education and safety awareness of different hazards associated with construction activities,

Hazard identified according to Occupational Health and Safety Code, (2009) is "*a situation, condition or thing that may be dangerous to the safety or health of workers*." Another definition of hazards associated with oil and gas pipelines by Canadian Standards Association, (2012) is "*a condition or event that might cause a failure or damage incident or anything that has the potential to cause harm to people, property, or the environment.* " The hazard identification process is a method of identifying potential hazards and defining their characteristics (Canadian Standards Association, 2012).

Hazard identification levels, generally, can be classified as formal assessment or site-specific hazard assessment (Alberta Construction Safety Association, 2013). Formal hazard assessment is a high-level assessment for all potential hazards that can be associated to tasks within a company, shop, or site. Site specific hazard assessment is specific to a site and includes hazards that were not identified in the formal hazard assessment. Hazard assessment levels for construction project can be as follows:

- Project hazard assessment: it is usually done at the beginning of the project to form an idea of how to construct a customized project safety plan. The project manager owns the responsibility of this plan. Identifying hazards at this level takes into consideration criteria such as the following:
    - Project logistic site plan
    - Project type and complexity
    - Project surrounding environment

- Special equipment or material shall be used in the project
- Existing overhead and underground utilities

- Job hazard analysis: it is conducted when starting a new construction job. Jobs could be classified based on the type of construction activities or construction zones; it depends on project type and also on the project breaking down strategy. JHA is safety planning for construction jobs prior to commencing any activity. Several aspects are taken into consideration when doing JHA, including:
  - Workers' ability to perform specific construction activities.
  - Equipment and material.
  - Different activities overlap and space availability
  - Government regulation.
  - Site location and conditions
  - Activities durations and delivery strategy
- Field level hazard assessment:

  FLHA is to be performed at the site level to encounter other hazards related to changing site or environment conditions. Also, it is conducted on an ongoing basis, daily or when task related conditions change. It is the responsibility of the Foreman and crews at the site.

The job hazard identification process is time-consuming and mostly depends on performing brainstorming sessions to quantify potential hazards associated with construction activities (Wang & Boukamp, 2009). Also, it involves collecting knowledge from several safety knowledge resources. Explicit Resources such as safety

manuals, safety codes and regulation, and safety best practices are used in performing the JHA process. In addition, tacit safety knowledge related to the experience of construction professional is a critical knowledge component that feeds into the JHA process. JHA documents are the final physical output of each JHA process for each construction activity in every construction project.

Construction projects lack an efficient identification of hazards associated with construction activity execution. Comparing the number of hazards identified for construction activities and what should have been identified and assessed showed that the hazard identification level is low and required significate enhancement, (Carter & Smith, 2006).

2.7 Overview of pipeline construction operation

Oil and gas pipeline projects consist of less in comparison with the building project. However, pipeline projects involved dynamic and complex execution environment. For example, activities like excavation, backfilling, pipe stringing and hydro-testing are directly affected by weather condition, equipment integrity, ground condition and competency of equipment operators.

Safety research that is related to nonbuilding construction projects are not sufficient. Nonbuilding projects such as pipeline construction and complex infrastructure need more focus due to their execution complexity and high potential risks (Zhou, et al., 2015).

Typical pipeline construction projects consist of several activities as shown in Figure 2-5. These construction activities have hazards associated with construction execution

and need to be addressed in advance through implementing effective hazard identification processes.

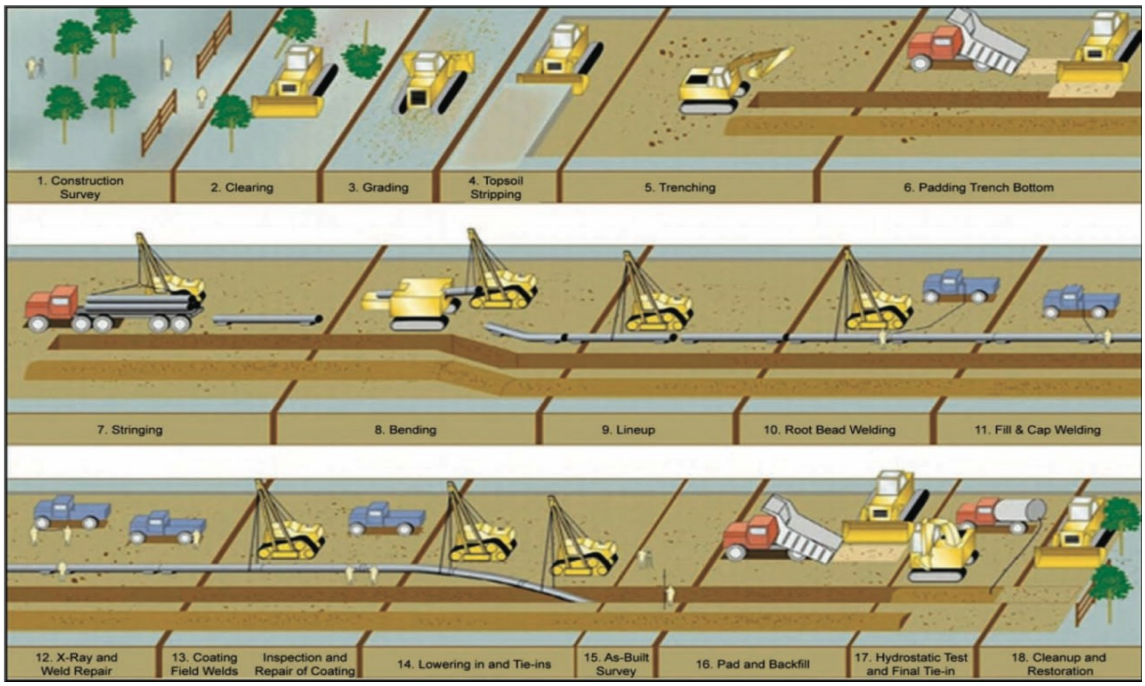


Figure 2-5 Typical pipeline construction activities (United States Department of State, 2014)

2.8 Text mining for knowledge discovery

The rapid development of technology in both hardware and software has enabled fast production of different type of information. This information can be in a structured format which can be managed by the database system. Text information is in an unstructured format, which is managed by search engine because of lack of structure inside text (Agarwal & Yu., 2009). Text is the one of the main forms of recording information and knowledge, and has been used in recording and transferring knowledge from ancient history till now. The text is the last output forms of any science or art. Text represents from 78 to 80 percent of information and knowledge resources (Gajzler, 2010). The text mining field is a growing field due the need for

exploring patterns and extracting information from large amounts of textual information. Text mining is a knowledge intensive process that enables users to interact with unstructured textual information to extract useful information by exploring patterns inside texts using an integrated suite of analysis tools (Feldman & Sanger, 2007).

One of the main and earliest text mining application areas is text summarization and classification for library catalogues. Summarization was used either to produce library catalogues or generate abstracts. Since 1990, text mining has changed significantly due to the advancement of technology and integration of advanced data mining and statistical learning, Primarily, text mining development was triggered as a response to the need to catalog text documents. Computers and the internet introduced new types of information such as web pages, emails, and electronic documents. In addition, the volume of data is growing massively. From 2001 to 2009, the rate of growth of web page volume was 40 percent per year (Miner, et al., 2012).

Text mining is using several techniques integrated together such as natural language processing, machine learning, and information retrieval (Feldman & Sanger, 2007; Sebastiani, 2002). Text mining is not too different from data mining. The Data mining concept was initiated to extract patterns and has further insight inside data to better understand the trends and relationships. Data mining focuses on structured numerical data in the database. In contrast, text as data exists in an unstructured format in its container, which can be the web, documents, or database (Weiss, et al., 2015).

Text mining can be done at two main levels. The two levels are: document level and inside document level. The document level is about organizing the documents in

similar groups or classifying documents according to predefined classes. The inside documents level is about extracting information represented by words, collocation, or phrases. In addition, inside document mining can be conducted to explore the semantic relation between extracted words and terms.

Text mining has many applications based on the goals of conducting analysis and can be categorized into six areas. First area, information retrieval and search for documents from large collections and by using keyword queries. Second, document clustering which is used to group relative or similar documents using data mining methodology. Third, document classification which is used to assign documents to predefined labels or classes based on user interest. Fourth, web mining which is developing according to advancement the use of internet and its massive of information rate growth. Fifth, information extraction that can be explained by extracting structured text from unstructured one. Sixth, natural language processing which is about giving sufficient information about specific natural language such as English. Sixth, concept extraction is an area of practice that use the interaction or integration between machine processing and human understanding to extract useful information, (Miner, et al., 2012).

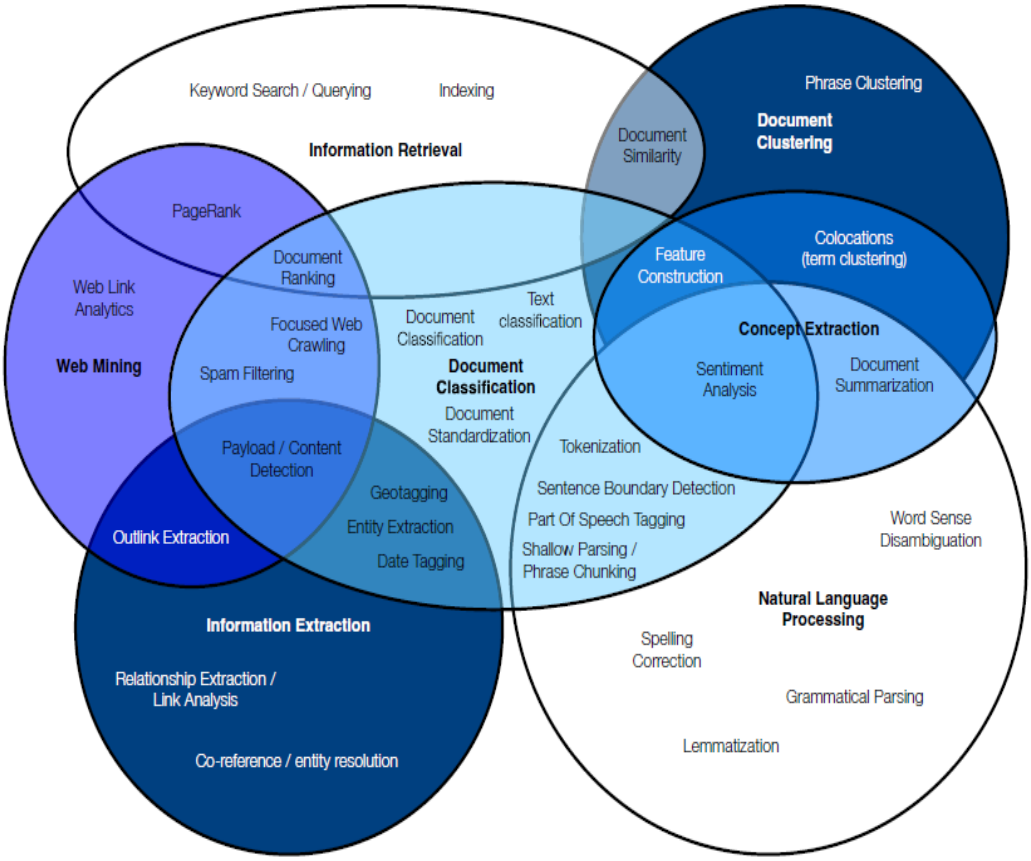


Figure 2-6 Applications of text mining (Miner, et al., 2012)

Documents represent the basic element of the text mining process. It can be defined as discrete textual data inside the collection. Documents' examples can be correlated with new types of documents such as emails, reports, manuscripts, articles, or research reports (Feldman & Sanger, 2007).

Project documents clustering and classification was used to categorize project documents (Al Qady & Kandil, 2015; AL Qady & Kandil, 2014). Text clustering is a useful technique for document automatic categorization and generating document taxonomy (K.Sruthi & Reddy, 2013). Using text mining, which includes natural language processing and Machine learning, can help in quickly extracting structured

useful information from unstructured documents (Tixier, et al., 2016). Fan & Li, (2013) discussed the issue of retrieving similar incident cases from electronic storage to help analyze current incident cases as a strategy to help in resolving disputes between parties involved in incidents claim process. Another researcher used NLP and machine learning (classification technique) to predict hazardous action from incident car crash reports. The method saved the time and efforts of checking narrative text reports manually to classify hazardous actions (Zhang, et al., 2016).

2.9 Text documents pre-processing

Prior to transforming text documents into a structured format, the document preprocessing stage should be done. Document preprocessing consists of several sophisticated processes to transform unstructured text inside the document to a structured format that enables machine learning algorithms application.

Structured data is the data stored in fields in files such as relational databases or spreadsheets. Unstructured data is related to information that is not residing in defined row and column databases or tables. The text pre-processing objective is to transform the text content of documents into structured spreadsheet tables, called a document-term matrix. Text preprocessing consists of several steps that are required to process text inside documents prior to application of machine learning.

2.9.1 Text tokenization

Tokenization is the process of dividing the stream of text into individual tokens. Tokens can be one word or multi-word (phrases or collocation). Tokenization is the primary step to extract features for clustering and classification and concept extraction.

In tokenization, token boundaries are space and punctuation. However, defining the boundaries of tokens or text features is sometimes problematic. For example, in the English language it is difficult to distinguish between the period that is ending the sentence and the period that should be considered part of the previous token such as Mr. or Dr. (Feldman & Sanger, 2007).

In other languages like Chinese and Thai, the words are unsegmented. Tokenizing such languages requires additional lexical and morphological information. To enhance the output of the tokenization step, a tokenization algorithm must be built based on text available for analysis.

2.9.2 Lemmatization (stemming)-English

After tokenization of the text, the next step is transferring tokens to its English standard forms. This process, referred to as lemmatization or stemming, aims to decrease the number of attributes or features and increase the frequency of some keywords by using one English word form. For example, all plurals are transformed to singular forms and past tense verbs are transformed to simple tense verbs. Also, increasing the frequency of keywords may influence the performance of the machine learning algorithms (classification and clustering) (Weiss, et al., 2015).

2.9.3 Stop word exclusion

Stop word elimination is part of natural language processing. In text documents, most of the frequently-used words have no meaning that contributes to core text meaning such as articles, preposition, and pronouns. For example, common words like “and,” “are,” and “the” are high frequency words in the texts.

Removing these types of word will reduce the dimensionality of text documents. Reducing dimensionality will improve the performance of machine learning algorithms.

2.9.4 Feature selection and weighting

The feature selection process consists of removing attributes or features that are not contributing to an efficient performance of text analysis. Three main parameters are calculated for each token: frequency, document frequency, and TF-IDF weight.

Frequency is the count of each token (keywords) in text documents. High frequency keywords are important to aid in discovering a pattern in the text. Also, high-frequency words are critical for concept extraction as will be discussed in chapter 4. However, high frequent tokens may be not efficient for text document classification and clustering due to their weakness as discriminative attributes.

Document frequency (DF) is the number of documents where a specific word has occurred. Document frequency is another measure for the strength of tokens and can be used in a vector space model as an input for, machine learning.

Term frequency and document frequency are key factors that describe the weight of a text feature in text corpora. However, a term's frequency may not help improve clustering performance. High frequency words that exist in many documents will not help clustering algorithm in discriminating between documents.

TF-IDF is a factor derived from both the token frequency and document frequency of the term. It takes into consideration in how many documents the term occurred. For terms occurring in a high number of documents, its discrimination power between

documents is less than other terms that have the same frequency but occurred in fewer documents. TF-IDF can be calculated by multiplying the frequency of a term by the inverse document factor (IDF) as follows:

$$\text{TF-IDF=TF * IDF}$$

$$IDF = \log\left(\frac{N}{DF}\right)$$

Feature selection and filtering can be accomplished using one or a combination of frequency, document occurrence, and TF-IDF. The first filter for the feature is done using stop word elimination. Feature selection depends on its purpose; if it is for text prediction, TF-IDF is efficient for filtering the highest discriminative features that can improve machine learning applications.

For pattern exploration or concept extraction, a mix of filtering strategies can be used under human supervision. In pattern recognition, frequency plays a critical role in exploring texts for a better understanding of underlying concepts in the text.

2.9.5 Vector space model

Text mining is different from numerical data mining and most of Machine learning algorithms were built for numerical data. However, in text mining each document is described by keyword or phrase attributes. The representation of the vector is in the form of a spreadsheet table. The rows represent the document vector and columns represent attribute vectors.

In the vector space model, the relation between features and documents may be described by the frequency of the token in a document (see Table 2-4). Also, it can be

represented by the binary representation zero or one; zero means the feature word did not appear in the document and 1 means the feature did appear in the document (see Table 2-5).

Table 2-4 Vector space model using frequency tabulation

	Feature 1	Feature 2	.	Feature m
D1	2	2		5
D2	0	9		3
D3	3	0		1
.				
Dn	1	10		0

Table 2-5 Vector space model using occurrence tabulation

	Feature 1	Feature 2	.	Feature m
D1	0	1		0
D2	1	0		1
D3	0	1		0
.				
Dn	1	0		0

2.10 Text visualization

2.10.1 Text visualization background

In data mining, the complex mathematical operation is executed to process the data and extract patterns and trends. It is usually done by computer and the end user is not involved in the analysis process. The information visualization primary goal is to help users in exploring data, understanding pattern and trends inside data to have better strategies for analysis.

Text extraction needs the support of visualization during analysis to help in understanding text patterns and trends. Understanding text features will help in

updating and controlling text mining strategies. Cao & Cui (2016) attached text visualization to three types of text information as follows:

Documents: such as articles, published paper, and web page. Visualization is mainly concerned about summarizing the content of documents.

Corpus: a collection of documents and visualization used to explore similarity and shared topics between documents.

Text stream: such as Twitter, a massive number of texts is produced continuously over time. Displaying the trend of texts visually can help in understanding the texts' subject directions.

Information visualization using computer-supported techniques can amplify human cognition (Card, et al., 1999). Using visualization to represent the output of extraction, summarizing, and abstraction processes will increase the rate of cognition of the knowledge receiver. The digitization of texts in the form of electronic books will help to analyze text information and extract visual abstract (John, et al., 2016).

2.10.2 Text visualization techniques

Text visualization is either related to the document's level or information related to inside the document's level. Visualizing inside documents will help in extracting important words, phrases, or topics in order to extract valuable information or map important concepts. Several visualizations were introduced recently to explore the content of documents such as word clouds (also known as a tag cloud), word trees, and words in context.

2.10.3 Word cloud

Word cloud is a commonly used technique for visualizing the existence and frequency weight of words in documents. A word cloud is used to explore high level information about documents (John, et al., 2016). A dynamic word cloud is developed from a text stream to show the rate of changes of words' frequencies in documents over time (Cui, et al., 2010).

Figure 2-7: Word cloud

2.10.4 Word tree

A word cloud cannot show the associates of the words. A word tree was found to show the related words of the text terms (Cao & Cui, 2016). It is the process of summarizing text via construction of trees based on syntax as shown in figure 2.2.

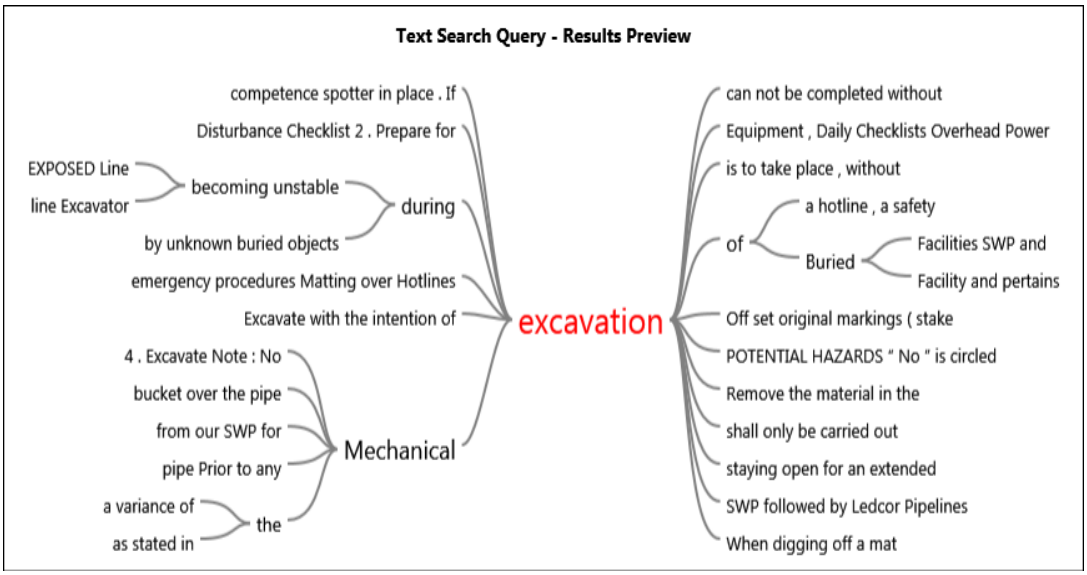


Figure 2-8: Example of word tree

2.10.5 Key word in context (KWIC):

KWIC is a technique to check the consistency of using a specific word or phrase in a collection of text documents. It enhances the understanding of the context of the words and phrases (Stemler, 2001). It helps to discover other words co-occurring with keywords. The KWIC technique enables rapidly extracting concepts from text and moving quickly between different levels of text concepts and abstractions (Wegerif & Mercer, 1997).

2.10.6 Events and storyline:

Texts can have several dimensions other than similar words or its co-occurrences. Other dimensions can be related to time, space, characters, and actions. These dimensions can be referred to as events. Most of the text is composed of serial events such as stories, incident reports, and crime reports (Buetow, et al., 2003). These

dimensions give more information about the text and can be used to visualize text for the objective of summarization and increasing human cognition.

For example, (Cho, et al., 2016) developed a visual system for exploring the history of Rome in Europe based on a large collection of Wikipedia articles. Text exploration takes into consideration knowledge related to places, characters, time, and events. For example, an interval of historical time can be explored using different attributes (linked information) such as historical characters, geographical places, and events. The system aims to enable knowledge learning, extraction, and formalization. Figure 2-9 illustrates ViaRoma interface.

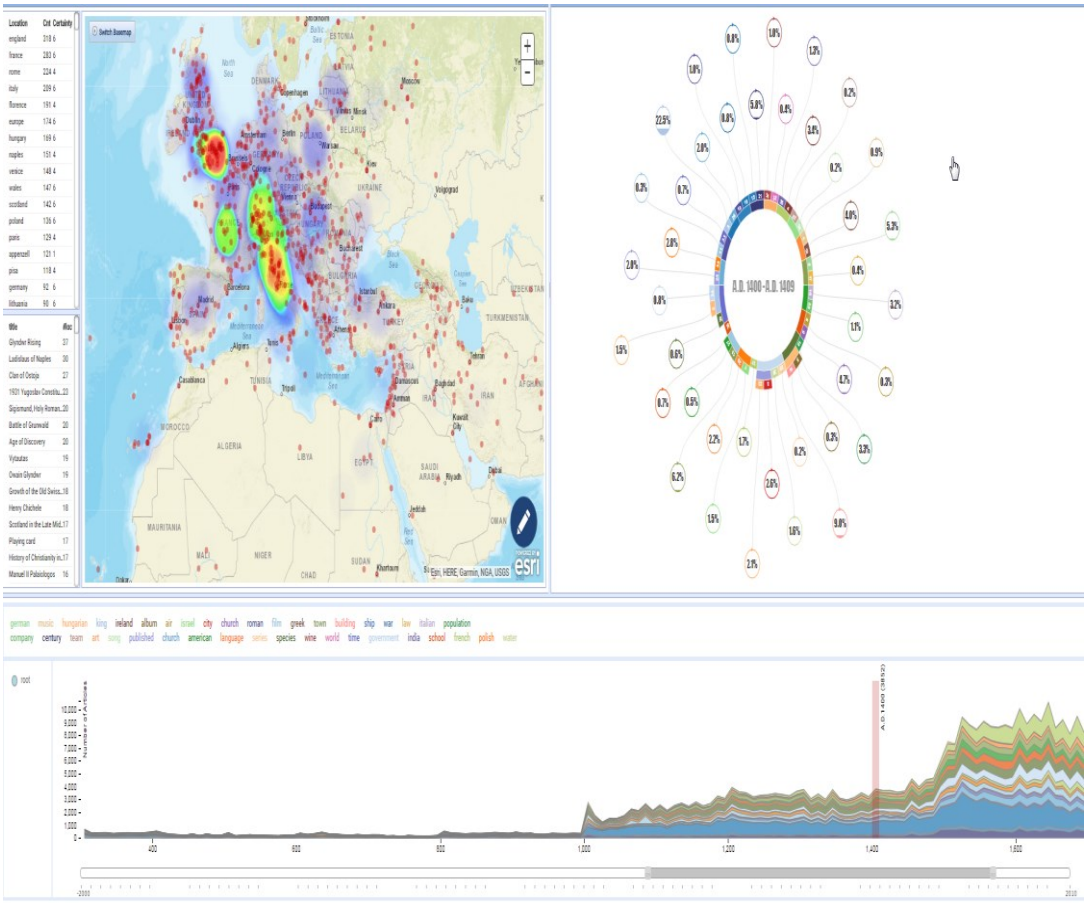


Figure 2-9 ViaRoma interface, (Cho, et al., 2016)

2.11 JHA documents: text exploration

JHA documents were collected for text analysis. JHA forms represent pipeline construction activities. The text exploration process is a continuous process that is used during the different stages of analysis such as document grouping, classification, and concept extraction.

Text exploration is necessary to discover a useful pattern of words in a text. Text patterns can support extracting different levels of concepts. Levels of concepts can be either general information about text knowledge domain or detailed about specific information in the text.

JHA exploration methodologies consist of several steps. Text preprocessing consists of document standardization, text tokenization, lemmatization, and stop word exclusion steps.

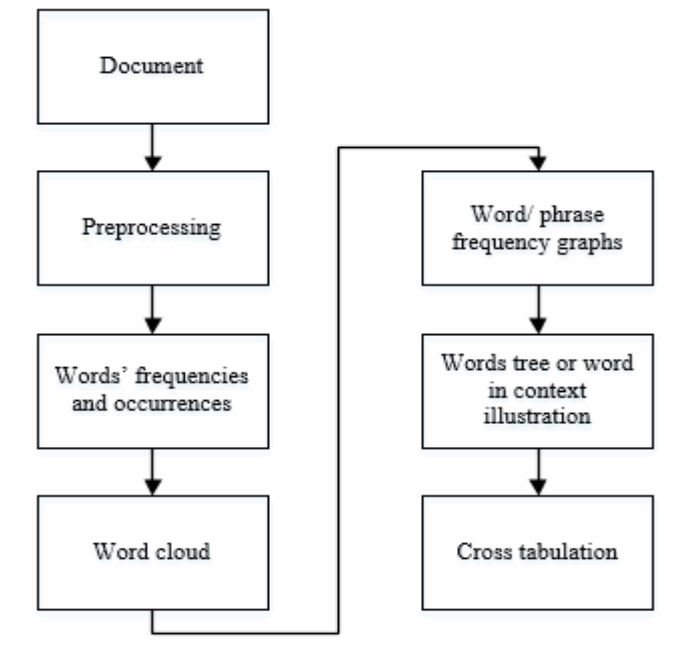


Figure 2-10 Text exploration for JHA forms

Extracting general information about text in JHA forms can be achieved using word clouds. A JHA document word cloud is illustrated in Figure 2-11. Also, word clouds can be developed for phrases instead of words. Phrases have more meaning than single words and can support understanding more about JHA form domain (see Figure 2-12). Other representations of the highest common words can be illustrated by frequency histograms as shown in Figure 2-13.

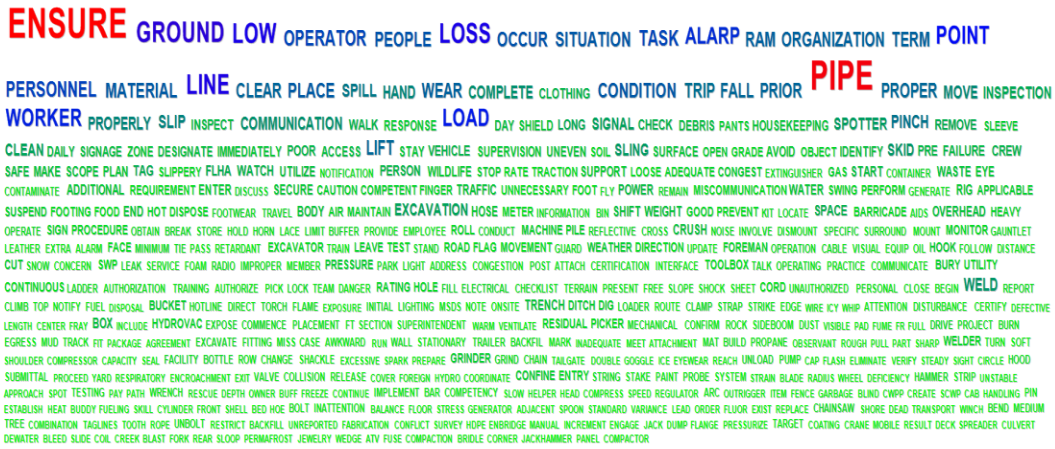


Figure 2-11 Word cloud for a single word

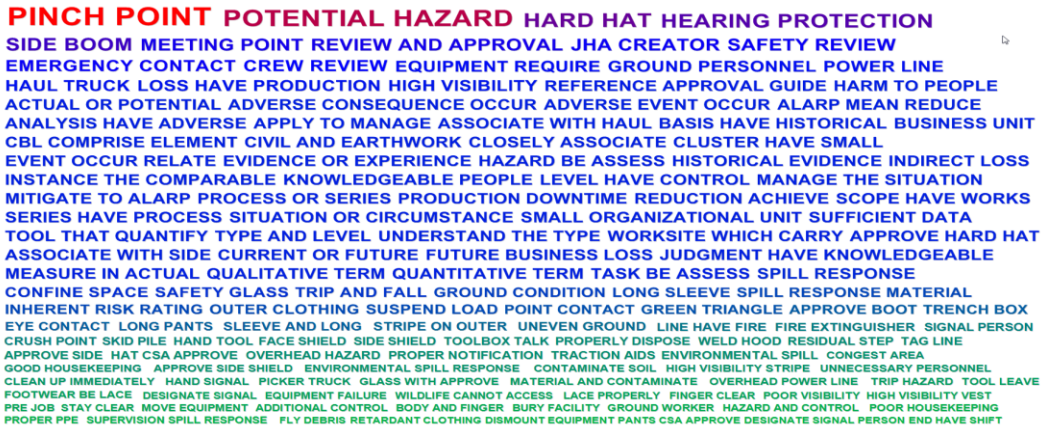


Figure 2-12 Word cloud for phrases

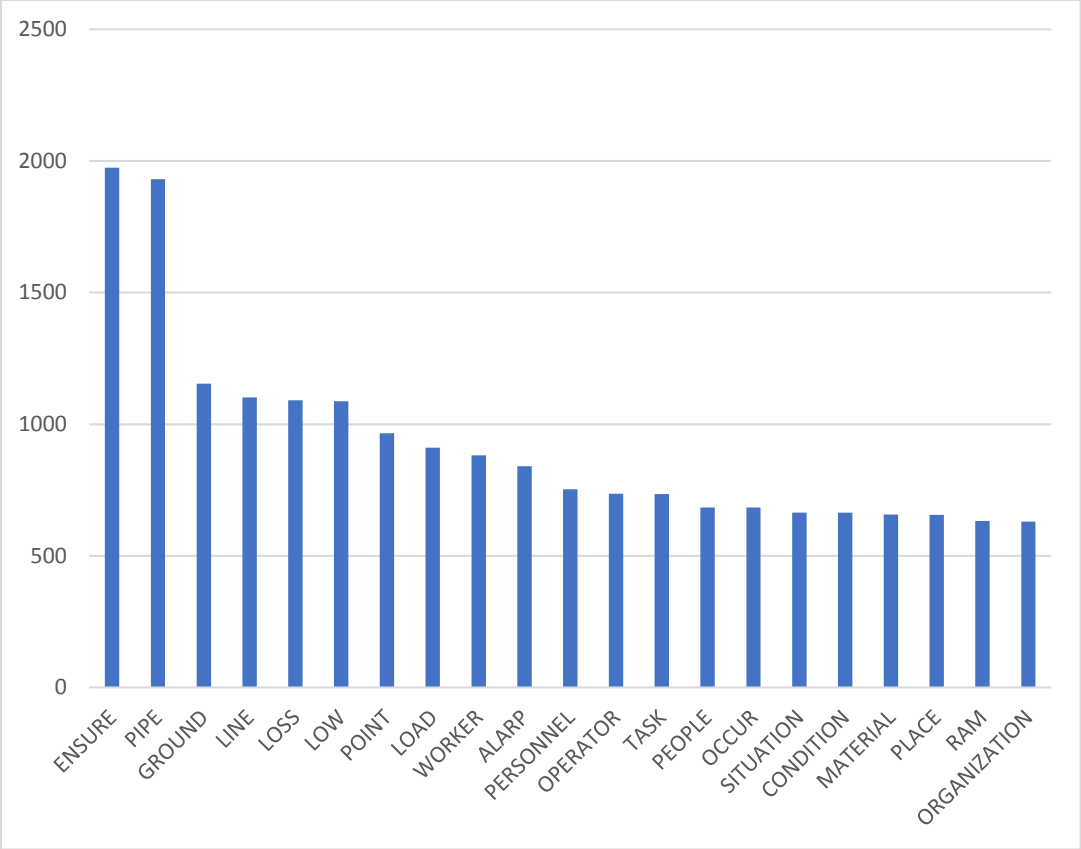


Figure 2-13 Highest frequent words in the collection of JHA documents

Sometimes words can have different meanings depending on the context of the text. Going to a further level of exploration of specific words, word trees, or word in context tool can be used to understand the background of using the word in different sentences or paragraphs. For example, the usage of the word “slip” in texts of JHA forms can be illustrated in the shape of a tree as shown in Figure 2-14.

Also, the word “overhead” could have several meanings depending on the context. Using QDA Miner software, keywords in a context table can be extracted to expose the targeted meaning of using the word. Table 2-1 illustrates the use of “overhead” in context.

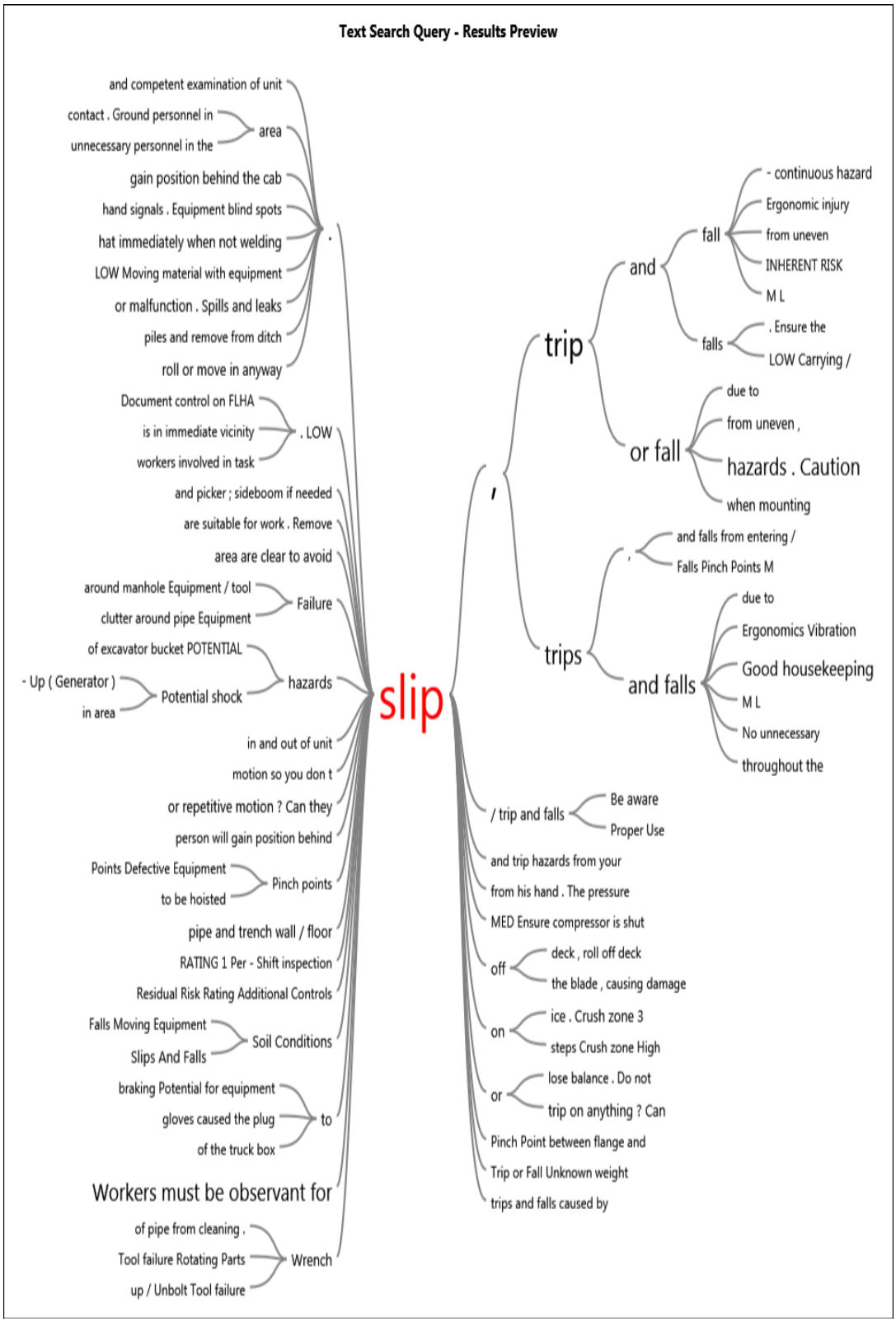


Figure 2-14 word tree for “slip”

Table 2-6 Keyword in context table for word “overhead”

Words before	KEYWORD	Words After
Of	overhead	utilities must be assessed and a
	Overhead	Power Lines Identified Safety glasses with approved side shields or over glasses
when working in close proximity of	overhead	
non-energized	overhead	Lines
An	overhead	power line and encroachment
An	overhead	power line an encroachment permit
	Overhead	power line. Do not release the straps until the picker is fully
For areas where the	overhead	hazard is not in
	Overhead	hazards Where overhead hazards within the immediate
Overhead hazards Where	overhead	hazards within the immediate
For areas where the	overhead	hazard is not in
	Overhead	hazards Where overhead hazards within the immediate
Overhead hazards Where	overhead	hazards within the immediate
For areas where the	overhead	hazard is not in
	Overhead	hazards Where overhead hazards within the immediate
Overhead hazards Where	overhead	hazards within the immediate
- Each falling operation in the immediate vicinity of	overhead	utilities must be assessed and a plan put in place to prevent felled trees or reached equipment from making contact.
	Overhead	Power Lines Identified
V in close proximity of	overhead	power line
Ensure spotter is utilized when traveling under non-energized	overhead	lines.
	Overhead	Hazards
Where	overhead	hazards within the immediate work area exist, welders must wear a hard hat/welding hood combination.
	Overhead	Hazards
Where	overhead	hazards within the immediate work area exist, welders must wear a hard hat/welding hood combination.

In addition, word frequency changes can be graphed against documents to show how words are used in different documents (see Figure 2-15 and Figure 2-16). However, frequency calculation may reflect false information about word or phrase weight. Although the word has high frequency, the word may occur in a small number of documents. This is important to consider especially if the word represents a concept. Word occurrence represents the distribution of word over all documents. It enables extracting a high-level concept from all documents. See an example of a word occurrence graph for "Debris" in Figure 2-17.

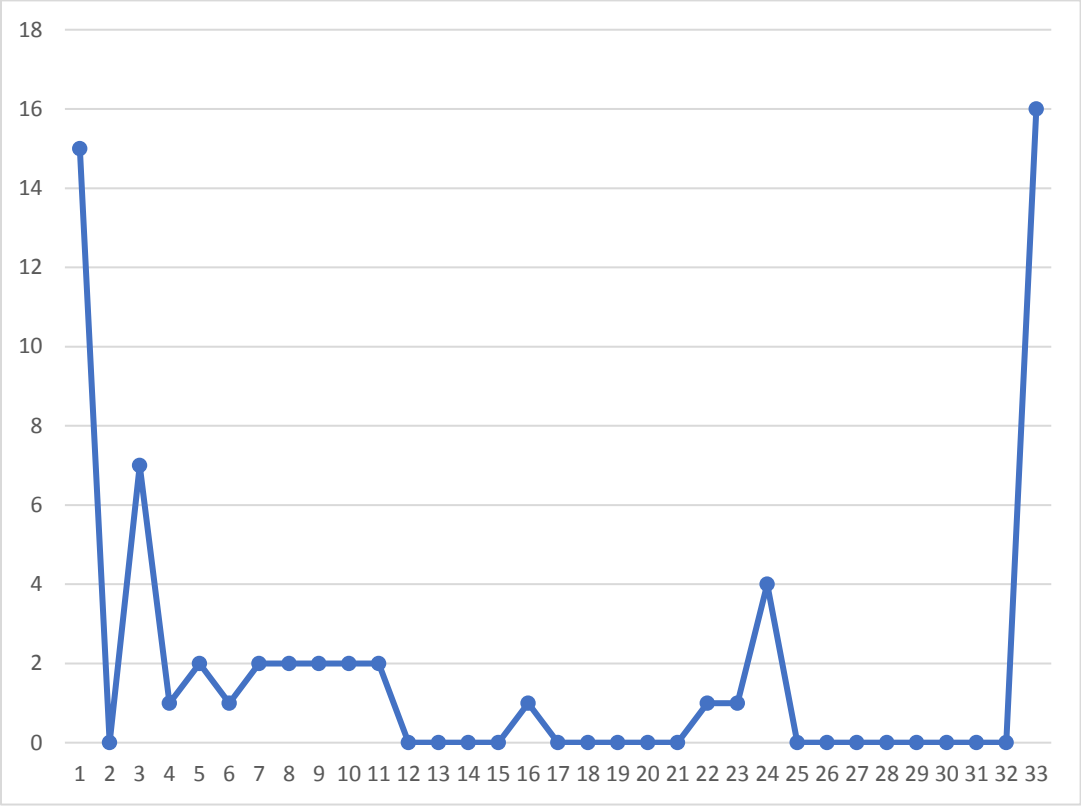


Figure 2-15 word "Ditch" frequency in documents

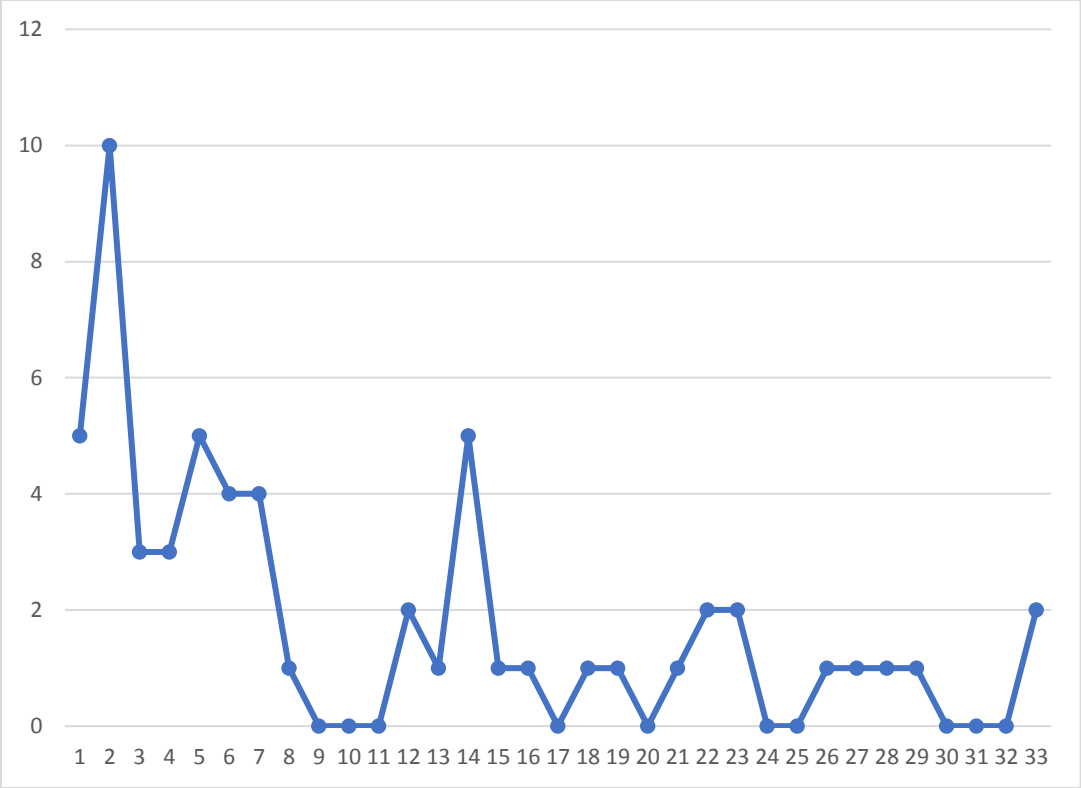


Figure 2-16 Word “Debris” frequency in documents

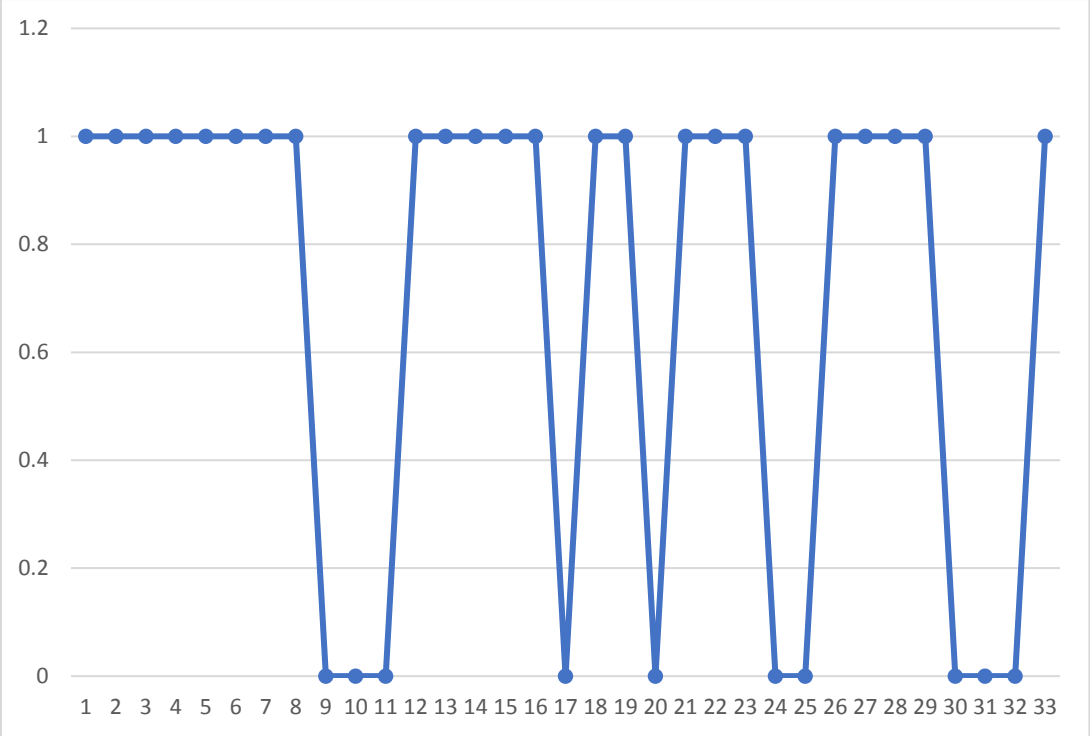


Figure 2-17 Word “Debris” occurrences in documents

Text exploration can go further in detail by extracting general concepts related to the domain of the text. For example, pipeline construction projects consist of construction activity such as excavation, backfilling, stringing, welding, hydro testing, hydro testing, pipe cutting, and coating. Text related to these activities is extracted and words' frequencies are calculated as shown in Table 2-7 . Figure 2-18 illustrates cross tabulation of some extracted words with construction activity classes of pipeline project. Cross tabulation highlights how the text words are used and related to each construction activity.

Table 2-7 Example of words representing pipeline project domain

Words	Frequency	Doc. Occurrence
LOAD	911	164
LIFT	607	139
MOVE	477	180
WELD	426	52
EXCAVATION	374	89
DIG	183	48
CUT	201	63
CONFINE	199	31
HYDROVAC	166	44
UNLOAD	116	34
STRING	105	31
BEND	85	18
UNBOLT	68	17
BACKFILL	57	16
HYDRO	55	32
COATING	47	12
SURVEY	41	15

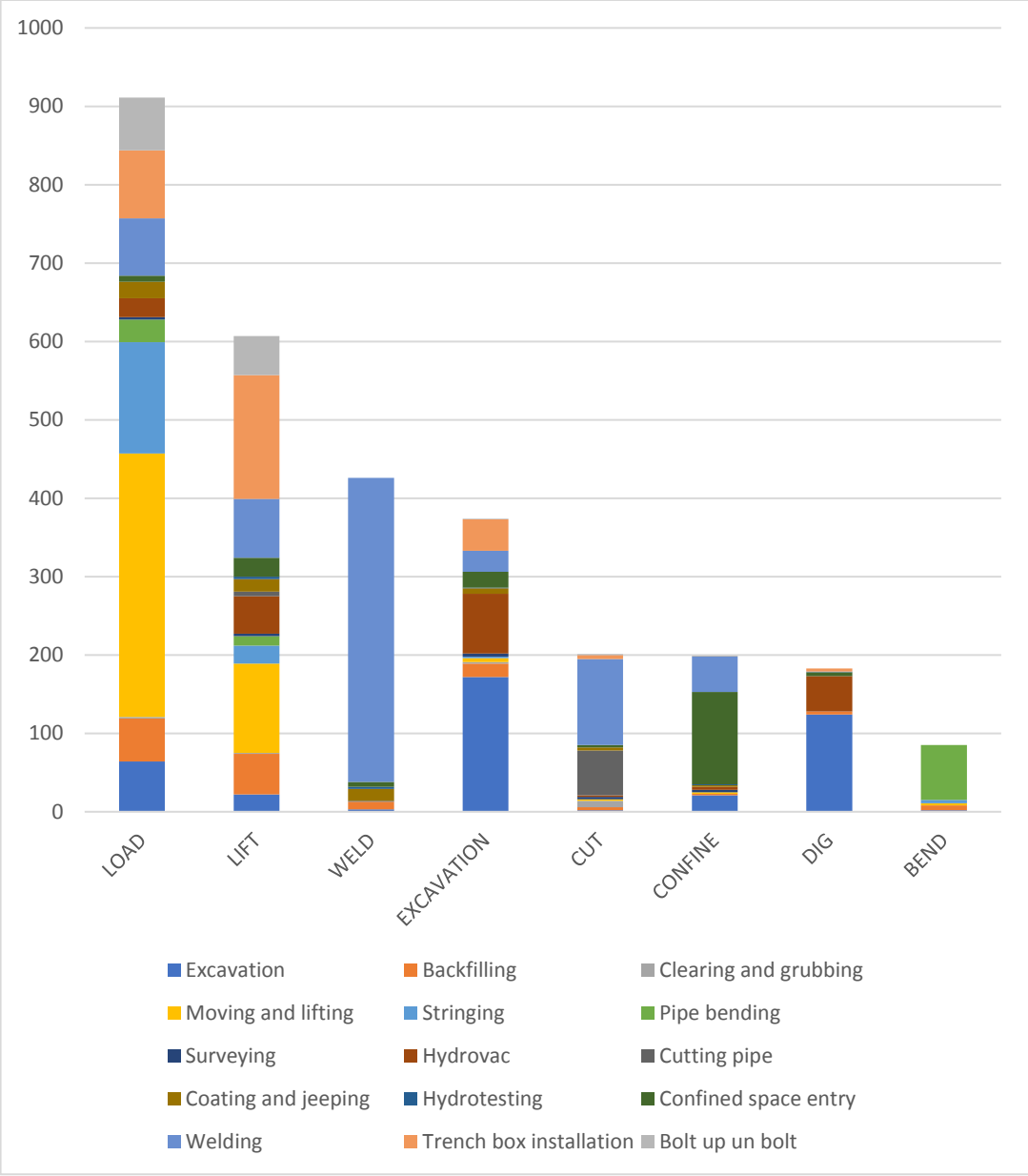


Figure 2-18 Crosstabulation of words with activity classes

Another example is exploring how and in what text form people are mentioning words in JHA documents. The word "worker" is the most frequent word used in JHA documents. The word "Personnel" is a synonym for "worker" and is frequently used in the forms. It is okay to find usage diversity of words' synonyms in

JHA forms because different people are filling out the forms. Since pipeline construction projects are heavy equipment based activities, equipment operators are the main part of the workforce in these projects (Figure 2-19).

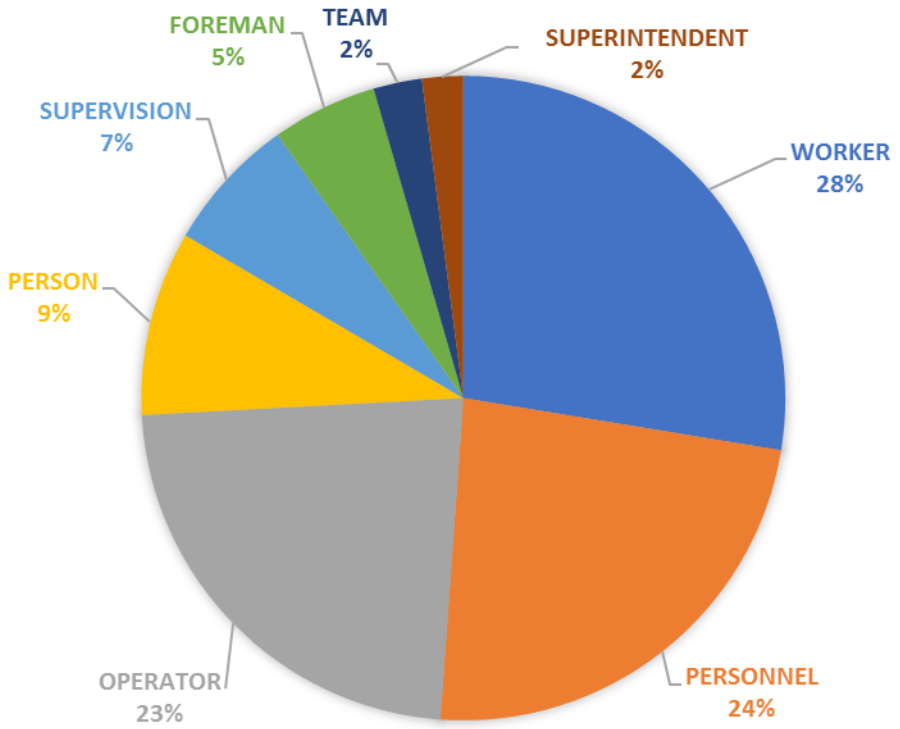


Figure 2-19 People entities in JHA forms

Another example is exploring words related to communication concepts. Communication is a critical component of project management and crustal subject for hazard management. Table 2-8 shows terms related to communication concepts such as miscommunication, information, and report.

Table 2-8 Terms related to communication concept

Terms	Frequency	Doc. Occurrences
COMMUNICATION	446	168
MISCOMMUNICATION	139	97
INFORMATION	118	87
TALK	113	58
REPORT	113	52
UNREPORTED	50	16

2.4 Conclusion

Investigating of pipeline construction incident reports indicated that failure to identify hazards is the most direct cause of incidents. Many researchers introduced different methods to reuse knowledge inside previous JHA documents to support hazard identification process. However, knowledge structures that were used as the bases for developing of the knowledgebase models, lack sufficient knowledge domain analysis.

Safety knowledge is mostly a text-based knowledge. Text mining tools are effective in organizing, classifying, and retrieving text documents. In addition, Text mining is used for extraction of knowledge concepts and taxonomies from unstructured texts. Using text mining in analysis of JHA documents domain can be effective for extracting knowledge structure that is essential input in to knowledge modeling process. Safety research that is related to nonbuilding construction projects are not sufficient. Nonbuilding projects such as pipeline construction and complex infrastructure need more research focus due to their execution complexity and the existence of high potential risks.

CHAPTER 3

JHA Documents Categorization

3.1 Introduction

Past JHA documents that were prepared for previous projects contain valuable knowledge about hazards associated with construction activities of pipeline projects. Extracting previous knowledge can benefit future JHA processes in future projects (Goh & Chua, 2010). There are two main levels of knowledge extraction. The first level is related to organizing and retrieving JHA documents. The second level is related to extraction of concepts, entities, and relationship patterns from text contents of JHA documents.

JHA forms are stored in the server, scattered and unused. Retrieving documents related to specific construction activity is challenging and takes a considerable amount of time. Identifying documents labels or classes is crucial to obtain the high-level general information about documents content. This information is important for further knowledge analysis, mapping, and extraction. Exploring documents manually costs time and significant human effort. Text mining and Machine learning (ML) are promising approaches for automatically grouping and organizing scattered JHA documents.

3.2 High- level methodology

JHA documents were collected from several pipeline construction projects. The documents do not have any file structure that enables retrieving documents based on their activity classes. There are two stages for categorizing the JHA documents (see Figure 3-1). The first stage is grouping the documents based on their text content similarity. Text similarity will be used to cluster documents in similar groups. Clustering is a useful approach when we do not have much information about the classes of documents (Feldman & Sanger, 2007).

After grouping documents in related groups, labels will be assigned to each cluster based on construction activity type. Manual Labelling takes significant effort. However, clustering documents in groups can help simplifying the process of labeling. Labeling documents' groups shortens the time and effort required to label each individual document manually. The second stage is building classification model to organize and categorize future JHA documents automatically. Training documents are necessary to build a classification model. Clustering stage output will be used as training data input to train the classifier.

Two objectives can be achieved by classifying JHA documents based on their construction activity classes. The first one is that organizing documents will enable JHA documents retrieval to be used as similar cases to support future JHA processes. The second objective is enabling further analysis to extract more detailed knowledge concepts, taxonomies and semantic relationships. The analysis and knowledge extraction represents the first step toward knowledge modeling to enable automatic retrieval and communication to support JHA process.

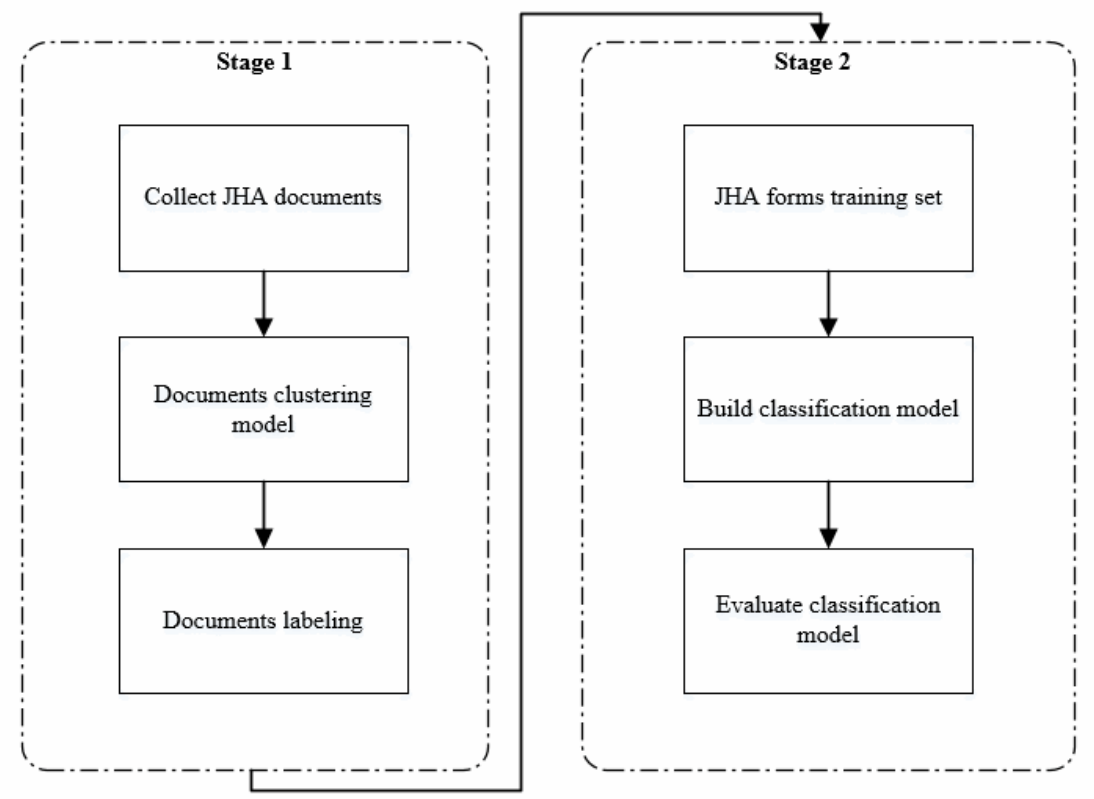


Figure 3-1 Document organization methodology

3.3 JHA Documents clustering and labeling stage

Document grouping based on text similarity was explored by several researchers using text mining techniques (Meziane & Rezgui, 2004; AL Qady & Kandil, 2014; Jun, et al., 2014). Document Grouping can be accomplished using machine learning technique, clustering methods. Clustering is an unsupervised machine learning technique. It does not require training data sets like other machine learning such as classification (Nagarajan & Aruna, 2016).

Manual labeling of enough documents for creating training documents required for classification model takes much time and effort (Moens, 2006). Clustering is a useful technique for information retrieval and data summarization and can be used as an input

process for other Machine learning process.

The document clustering process starts with the document pre-processing phase. Document pre-processing is a critical step for preparing textual data inside documents before starting clustering process. Feature extraction is the process of extracting terms (words or phrases) and their frequencies in documents. Document-term matrix is a form of representation that is required as an input for clustering algorithms. It consists of rows, also called examples, which represent documents and columns that represent text features (attributes) (AL Qady & Kandil, 2014). After grouping all documents in similar groups, documents groups are labeled based on pipeline construction activity classes. The output of this process is labeled group of JHA documents and categoried according to pipeline construction activity classes (see Figure 3-2).

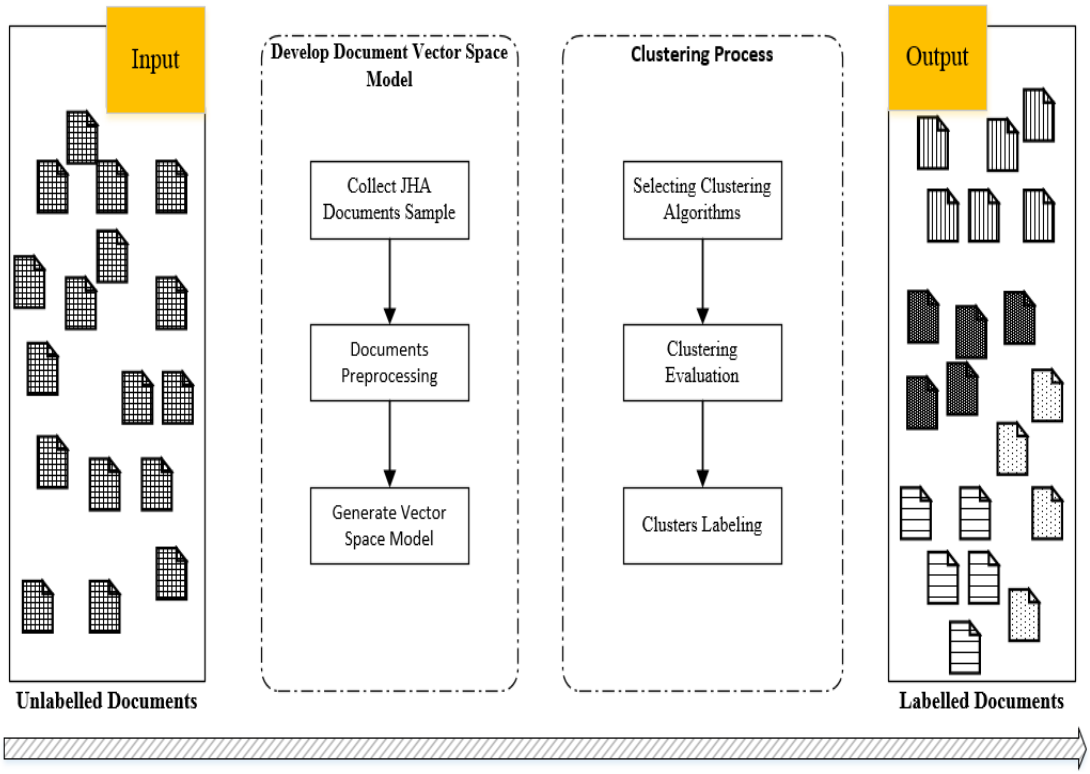


Figure 3-2: Document grouping and labeling process

3.3.1 Clustering method

Clustering is an unsupervised machine learning process that aims to segment a large number of examples into clusters by measuring similarity or association between different examples. So, examples in the cluster are similar to each other and dissimilar to examples in other clusters. The output of the clustering process is clusters of data sets, and this contributes to form a high-level description of each cluster which is essential for the further deep analysis of data (Kantardzic, 2011).

The clustering technique is used in many applications related to different areas such as:

- Data summarization: grouping and extracting main concepts is an important objective of data mining. Clustering approach enables abstracting process by grouping data in clusters based on similarity (Aggarwal, 2015).
- Social network analysis: understanding human community depends on the analysis of the social relationship. The digitized information of social networks makes networks analysis possible using artificial intelligence techniques. Clustering techniques are used in social network analysis to group users in similar groups. These groups share similar relations, habits, and interests (Aggarwal, 2015).
- Clustering is input to other data mining technique: sometimes clustering is considered as a preprocessing step for other data mining processes such as classification and outlier detection models (Aggarwal, 2015).

- Clustering in business: clustering is widely used in business for grouping customers based on their interests, location, income, or age (Han & Kamber, 2006).
- Documents clustering into topics: by using text similarity inside documents, clustering is used to group documents in classes (Kotu & Deshpande, 2015).

3.3.2 Similarity measure:

The similarity measure is one of the most important key factors in the clustering process. It is the mathematical method of measuring distance or similarity between document vectors. The similarity measure mechanism has a major influence on the performance of clustering algorithms and plays a significant role in the success or failure of the clustering process (Anon., 2013). To choose proper metrics, several conditions must be satisfied as follows:

- The distance between any data objects must not be negative
- The distance is zero when the data points are identical
- The distance between data points are symmetrical

Three common types of similarity measures are presented in this section. The three-similarity measures are Euclidean distance measure, cosine similarity measure, and Jaccard coefficient.

1- Euclidian distance

The Euclidean distance measure is simple and used widely in clustering problems. The Euclidean distance between two data objects X and Y can be calculated as per equation (3-1):

$$Distance = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots (x_n - y_n)^2} \quad (3\text{-}1)$$

Where attributes of X and Y are expressed as follows:

$$X = (x_1, x_2, \ldots \ldots x_n)$$

$$Y = (y_1, y_2, \ldots \ldots y_n)$$

2- Cosine similarity

Cosine similarity is a well-known and popular measure for similarity in text clustering problems (Han & Kamber, 2006; Huang, 2008). Cosine similarity is the cosine angle between two document vectors, $d_i \; and \; d_j$, and can be calculated by equation (3-2):

$$\cos(d_i, d_j) = \frac{d_i . d_j}{\|d_i\| \|d_j\|} \quad (3\text{-}2)$$

The cosine similarity has a value bounded between zero and 1, more similarity when close to 1 and less similarity when close to zero. Cosine similarity does not depend on document length, which means two documents that have the same main content but differ in length may be considered identical (Huang, 2008).

3- Jaccard coefficient

Jaccard coefficient measures the similarity by dividing the intersection of the objects by the union. In document domain, it is defined as the division of the total number of shared text terms between two documents A and B and the total number of terms or attributes of both documents A and B as shown in equation (3-3):

$$Sim\ (A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3\text{-}3)$$

Jaccard coefficient ranges from 0 to 1. A zero means that document A and B have no relation and are entirely different, but one means that the documents are identical.

3.3.3 Clustering algorithms

There are many clustering algorithms that can be used to perform clustering for different data types. Clustering algorithms can be classified into four categories: the partition method, hierarchal method , density-based method, and grid-based method (Han & Kamber, 2006). This research utilizes two methods, the partition method and hierarchal clustering method. The partition method (K-Mean clustering) will be used for documents clustering and the hierarchal clustering method will be used for text terms clustering to extract semantic relationships (see Chapter 4).

The partition method organizes a set of data in K clusters based on similarity. K-mean clustering is a well-known partition clustering algorithm and used broadly in many types of research to demonstrate clustering problems (Zhang, et al., 2010; Li, et al., 2008; AL Qady & Kandil, 2014).

1- Partition methods (K- Means algorithm)

K- Means is a straightforward and well-known clustering method for numerical data clustering. It is also widely used for document clustering due to its simplicity and relative high efficiency (Weiss, et al., 2015; AL Qady & Kandil, 2014). It is based on calculating the centroid of data points. Each cluster contains centroid points that are

updated iteratively based on new data points that have recently joined the cluster (see Figure 3-3).

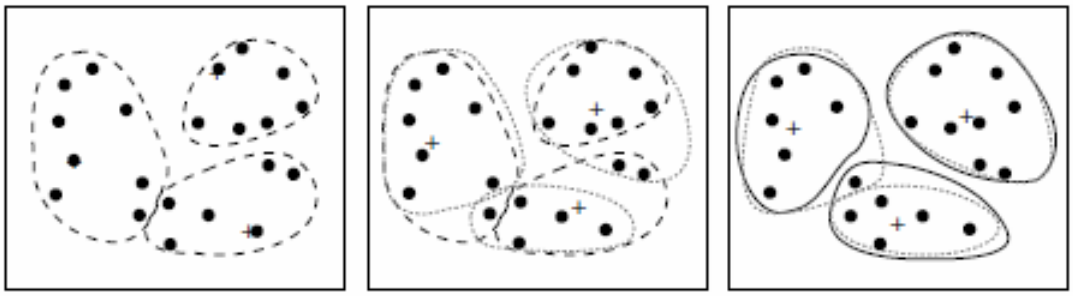


Figure 3-3 K-mean clustering iterative steps (Han & Kamber, 2006)

The algorithms initially and randomly select number k of documents as centroids of the k clusters, and then assign the rest of the documents to cluster in an iterative manner. In each iteration, the algorithm calculates new centroids for each cluster based on the distance between documents and the cluster mean (Huang, 2008).

The iteration is repeated until criterion function converges and this is based on the squared error distance between each point in the cluster and the cluster mean. This will lead to a denser oriented cluster and more clear separation between other clusters. The square error criterion is calculated as per equation (3-4):

$$E = \sum_{i=1}^{K} \sum_{P \in C_i} |P - m_i|^2 \qquad (3\text{-}4)$$

E : sum of squared error

K : number of clusters

P : data object in the space

m: is the cluster i centroid

2- Hierarchal clustering methods:

The hierarchal technique depends on hierarchal decomposition of data set, and the output can be represented in the form of a dendrogram or tree structure. There are two types of hierarchal methods based on the method of breaking down of the clusters. They are an agglomerative method and divisive method. Agglomerative clustering, also known as bottom-up clustering, is an approach of building the hierarchy cluster from bottom to top by starting with each data object as a primary cluster and then merging similar clusters in upper levels until merging all clusters in one level or until specific termination conditions are met. Divisive clustering, also known top-down clustering, starts from one big cluster of all data objects, and then breaks it down into sub clusters based on similarity until each data object forms clusters by itself or satisfies specific termination conditions such as the required number of clusters. The agglomerative clustering approach is more employed in a real-world application than the divisible method (Han & Kamber, 2006; Kantardzic, 2011).

Dendrogram is a tree structure used to represent the hierarchal clustering as shown in Figure 3-4. Considering agglomerative clustering, the first level is showing each data object in a single cluster. The next levels represent more clusters of objects grouped together by horizontal lines based on a similarity scale.

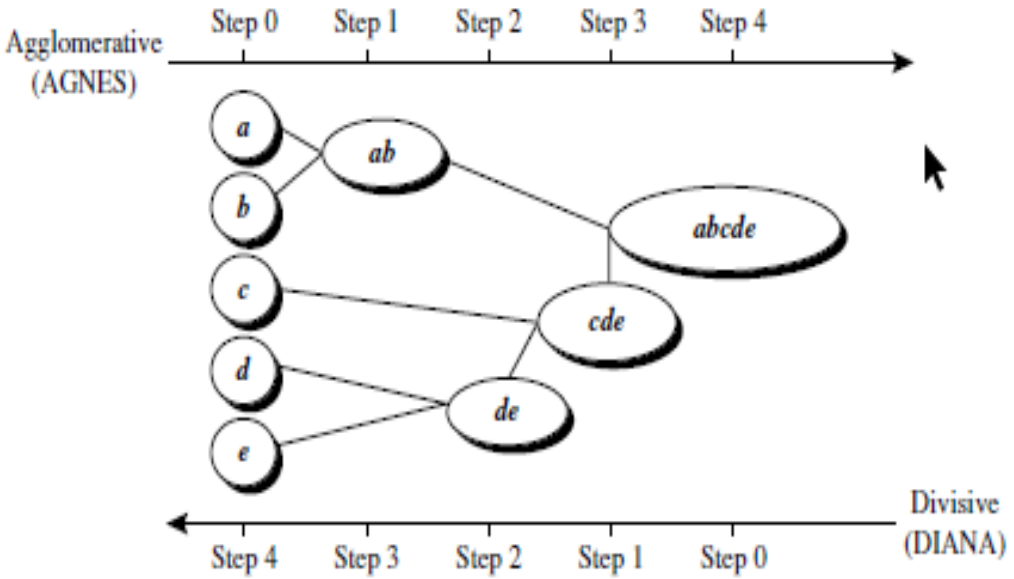


Figure 3-4 Hierarchal clustering example for data (Abcde) (Han & Kamber, 2006).

To evaluate clustering algorithms, the purity measure is used in present research. The purity measure is simply checking the coherence of outcome clusters by counting dominant documents types in each cluster and dividing by the total number of documents. Clustering purity is calculated for each cluster as per equation (3-5):

$$Purity = \sum_{j} \frac{1}{nj} \max(nij) \qquad (3\text{-}5)$$

Where *ni* is the number of documents in class *i, nj* is the number of documents in cluster *j* and *nij* is the number of documents of class *i* in cluster *j*

3.3.4 JHA documents pre-processing for clustering stage

The document preprocessing stage is essential for preparing text data for application of machine learning algorithms. Text in JHA documents is found in an unstructured

format and needs to be transformed to a state where the application of machine learning becomes possible. Document preprocessing consist of document sampling, spell check, word lemmatization, text tokenization, feature selection, and building a document -term vector space model (see Figure 3-5). Document term matrix is the initial input for clustering algorithm.

142 JHA documents were collected from 9 classes to apply the document clustering methodology (see Table 3-1). Document sampling and selection take into consideration the quality and completeness of the JHA forms.

Table 3-1 JHA forms classes used for clustering process

Construction activity class	No of JHA forms	Total percent
Excavation	34	23.90%
Stringing	13	9.20%
Pipe Bending	9	6.30%
Cutting pipe	11	7.70%
Coating and jeeping	10	7.00%
Hydro testing	9	6.30%
Welding	32	22.50%
Trench box installation	9	6.30%
Bolt up unbolt	15	10.60%
Total	142	

JHA documents have spelling errors and need to be checked to enhance the feature selection step. Correcting word spelling improves word frequency that is required for feature selection. In addition, misspelled words could be considered by the machine learning algorithm as good discriminative attributes due to its low frequency. Tokenization breaks the stream of character into individual words or collocations features to use them as attribute dimensions that describe JHA documents in a space vector model. Table 3-2 shows examples of extracted text word tokens.

Table 3-2 Example of extracted tokens

Extracted tokens			
WELD	SUPPORT	TRENCH	DESIGNATE
PINCH	WELDER	FLHA	START
LIFT	CUT	PRE	MAKE
CONDITION	SHIFT	POWER	OPEN
SKID	UTILIZE	SIGNAL	IMMEDIATELY
SLING	POOR	GRINDER	VEHICLE
EXCAVATION	PERSON	LONG	ACCESS
COMMUNICATION	PRESSURE	DAILY	STAY
CLEAN	END	OVERHEAD	SCOPE

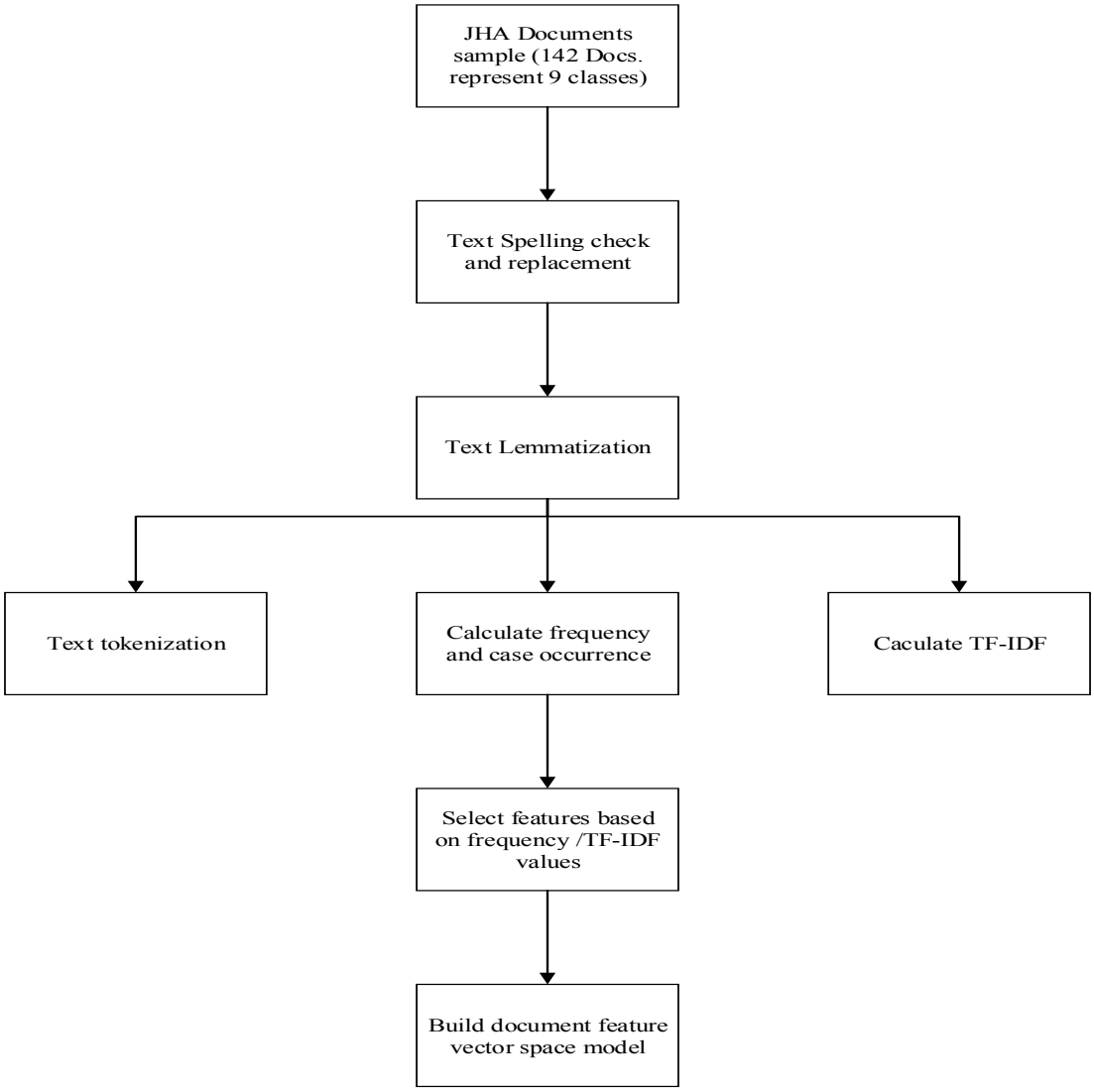


Figure 3-5 Document preprocessing steps

Not all tokens are qualified for the selection and they may have no discriminative power between documents. For example, word types such as conjunctions and pronouns are repeated a lot in the documents and have high frequencies in all documents. Three main parameters were calculated for each token to be used in the Vector space model (see Table 3-3). Frequency and case occurrence for each token have been computed and used to calculate the TF-IDF factor using the equation as per equation (3-6) and equation (3-7):

$$\text{TF-IDF}=\text{TF} * \text{IDF} \tag{3-6}$$

$$IDF = \log\left(\frac{N}{DF}\right) \tag{3-7}$$

Where:

TF: is the total frequency of the text term

IDF: inverted document frequency

N: total number of documents

DF: document frequency (number of documents where text term occurred)

The TF-IDF take into consideration the document frequency where the word occurs. When text term token occurs in large number of documents, the reduction factor $\log\left(\frac{N}{DF}\right)$ shall be increased.

Table 3-3 Example of Token parameters

Tokens	Frequency	No. of cases	TF - IDF
WELD	410	46	200.7
PINCH	393	97	65
LIFT	354	79	90.2
CONDITION	329	90	65.2
SKID	313	76	85
SLING	300	89	60.9
EXCAVATION	250	45	124.8
COMMUNICATION	249	97	41.2
CLEAN	235	91	45.4
REMOVE	224	90	44.4
SPOTTER	215	67	70.1
BOX	209	23	165.2
TAG	208	80	51.8
TRENCH	201	29	138.7
POWER	195	67	63.6
SIGNAL	195	80	48.6
GRINDER	190	33	120.4
HOSE	180	72	53.1
WALK	176	95	30.7
SURFACE	175	90	34.7
HOOK	174	48	82
DEBRIS	162	93	29.8
SUPPORT	161	77	42.8
WELDER	160	34	99.3
CUT	159	46	77.8
SHIFT	157	52	68.5
UTILIZE	157	76	42.6
FAILURE	147	59	56.1
DESIGNATE	146	80	36.4
START	144	62	51.8
CRUSH	133	36	79.3
SAFE	131	61	48.1
SUSPEND	130	60	48.6
UNEVEN	130	77	34.6
DIG	128	20	109

For example, the word “Lift” has a frequency of 354 and occurred in 79 out of a total of 142 documents. Since the word occurred in a high number of documents, its frequency reduced to 90.2. Another example is the word “Trench” which has a frequency of 209 and a case occurrence of 29. Since “Trench” occurred in less documents, its frequency reduced to 138.7. Comparing the two examples, “Trench” has more discriminative powered than “Lift” because “trench” occurred in less documents. Figure 3-6 shows examples of word tokens and their calculated frequencies, occurrences, and TF-IDF values.

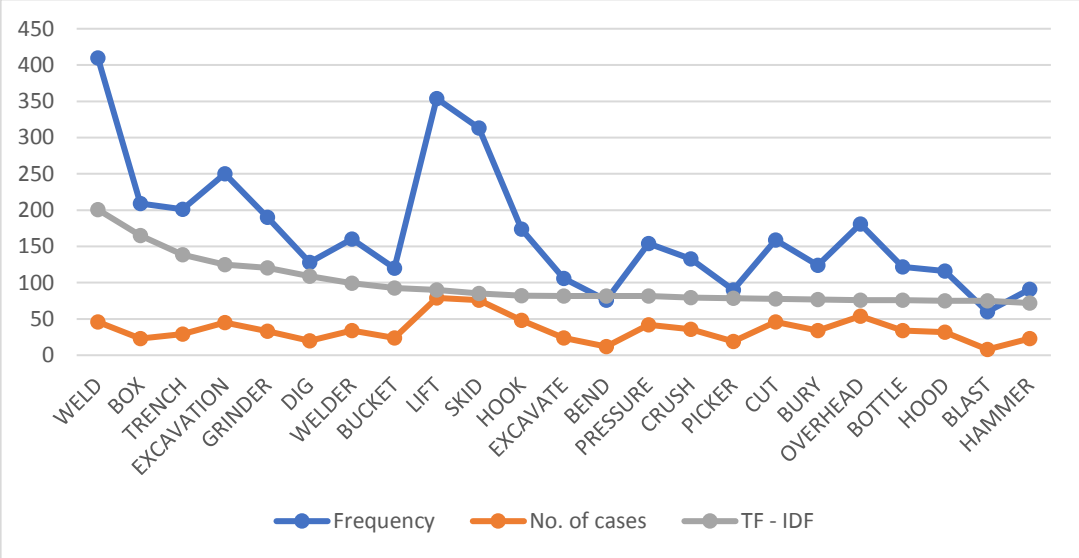


Figure 3-6 Tokens frequency, case occurrence, and TF-IDF

Word Lemmatization is the process of transforming words to their original form. This means increasing the frequency and TF-IFD weighting factors and hence improving the performance of clustering algorithms. For example, words like “spills,” “trips,” and “welding” are substituted by “spill,” “trip,” and “weld,” respectively (See Table 3-4).

Table 3-4 Example of words lemmatization

Original	Substituted	Length	Freq.
REQUIRED	REQUIRE	8	958
HAZARDS	HAZARD	7	600
CONTROLS	CONTROL	8	497
APPROVED	APPROVE	8	411
WELDING	WELD	7	401
TOOLS	TOOL	5	392
CONSEQUENCES	CONSEQUENCE	12	359
POINTS	POINT	6	358
CONTAINED	CONTAIN	9	347
WORKERS	WORKER	7	324
CONDITIONS	CONDITION	10	304
CONTRACTORS	CONTRACTOR	11	295
STEPS	STEP	5	287
TIMES	TIME	5	281
LINES	LINE	5	258
SLIPS	SLIP	5	253
ASSESSED	ASSESS	8	250
OCCURRING	OCCUR	9	248
TRIPS	TRIP	5	247
CONSIDERED	CONSIDER	10	240
ASSESSING	ASSESS	9	239
WORKING	WORK	7	239
REDUCING	REDUCE	8	239
ASSETS	ASSET	6	238
HAPPENED	HAPPEN	8	238
TERMS	TERM	5	237
AREAS	AREA	5	224
MOVING	MOVE	6	211
GLASSES	GLASS	7	208
FALLS	FALL	5	205
SLINGS	SLING	6	192
SKIDS	SKID	5	189
GLOVES	GLOVE	6	180
TRUCKS	TRUCK	6	175
BOOMS	BOOM	5	148
BOOTS	BOOT	5	145
DESIGNATED	DESIGNATE	10	145
SPILLS	SPILL	6	137
CONTACTS	CONTACT	8	133
MEANS	MEAN	5	132
CREATORS	CREATOR	8	131
HOSES	HOSE	5	131

Tokens can comprise more than one word, consisting of two or three words. Tokens consisting of two or three consecutive words are called collocations. Collocations are extracted based on the frequency of terms occurring together frequently in JHA documents. Table 3-5 shows examples of collocations extracted from JHA documents.

Table 3-5 Example of collocation extracted from JHA documents

Collocation	Frequency	No. of Cases	Length	TF-IDF
TRENCH BOX	142	10	2	163.6
OVERHEAD HAZARD	89	16	2	84.4
POWER LINE	165	46	2	80.8
HOOK PERSON	52	6	2	71.5
WELD HOOD	108	31	2	71.4
CONFINE SPACE	70	15	2	68.3
MECHANICAL EXCAVATION	59	13	2	61.3
EYE CONTACT	71	20	2	60.4
TARGET LINE	56	12	2	60.1
PINCH POINT	346	97	2	57.3
STRING PIPE	48	10	2	55.3
RESIDUAL STEP	57	17	2	52.5
SKID PILE	81	33	2	51.3
JOB STEP	59	20	2	50.2
FOLLOW JOB STEP	54	17	3	49.8
CONTINUOUS HAZARD	54	17	2	49.8
GROUND PERSONNEL	130	59	2	49.6
FRONT HOOK	36	6	2	49.5
BODY AND FINGER	66	26	3	48.7
TOOLBOX TALK	77	34	2	47.8
SUSPEND LOAD	115	56	2	46.5
TRIP AND FALL	133	66	3	44.3
PICKER TRUCK	49	18	2	44
CRUSH POINT	63	29	2	43.5

Frequency, occurrence, and TF-IDF are calculated for each collocation. Collocation features are used in clustering experiments to estimate clustering performance. Figure 3-7 describes the relationship between frequency, TF-IDF, and document frequency. If a collocation has high frequency in many documents, the collocation has less

discriminative power. For example, although “pinch point” has the highest frequency in the 142 JHA documents, it has relatively low TF-IDF weight due to its occurrence in many documents.

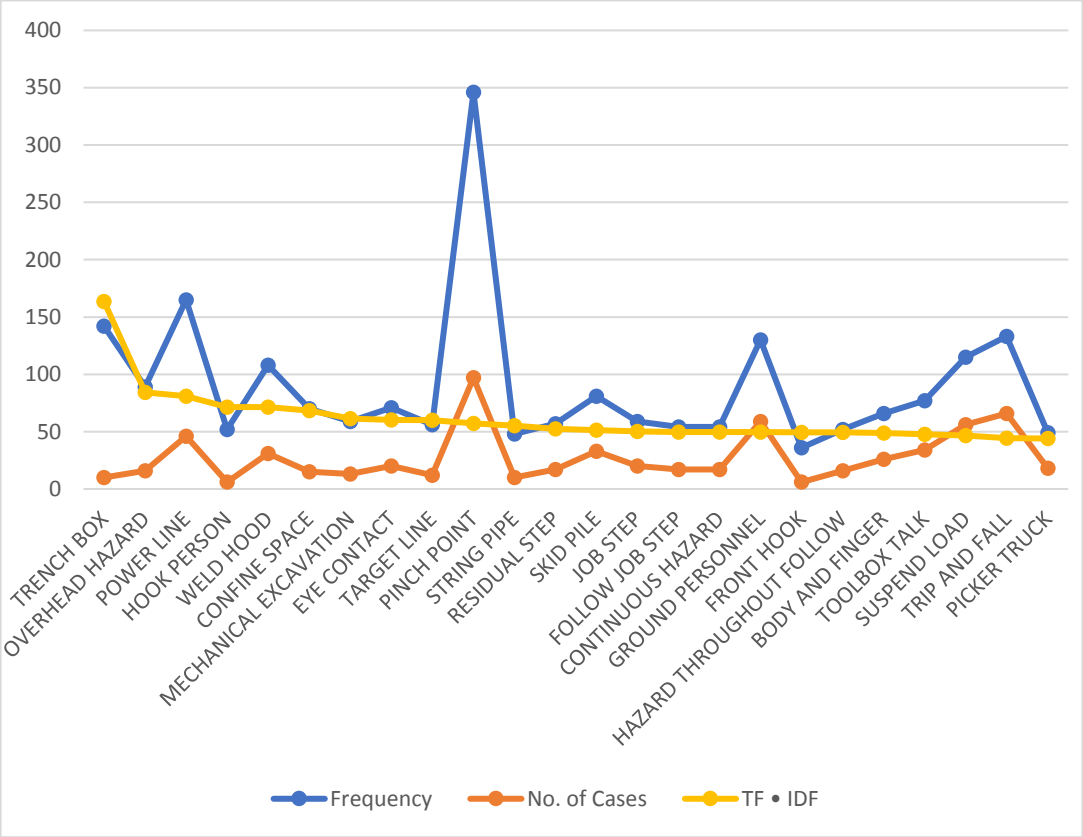


Figure 3-7 Collocation frequency, case occurrence, and TF-IDF

Documents – term (D-T) matrix, also known as a vector space model, is widely used in document clustering. It is developed to represent document vectors by their text term features (Bharti, 2014). In a D-t matrix, a document represents an example (row) associated with several attributes/features (columns) (see Table 3-6). A D-T matrix developed for clustering is a large multi-dimensional sparse matrix. A Sparse matrix means that many values in the matrix are zeros.

Table 3-6 D-T matrix (word tokens)

Doc.	AIR	ALARM	ALIGNMENT	ARC	ATTACH	ATTACHMENT	ATTENTION	AVOID
E37	1	0	0	0	1	1	0	2
TB4	0	0	0	0	4	0	0	0
TB3	0	0	0	0	0	0	0	0
TB5	0	0	0	0	0	0	0	0
TB7	0	0	0	0	5	0	0	3
E28	4	0	0	0	1	0	0	2
W41	0	0	0	0	4	0	0	1
HY9	1	1	0	0	0	0	0	1
W42	0	0	1	2	2	0	0	2
HY3	3	0	0	0	0	0	0	0
HY6	3	0	0	0	0	0	0	0
E22	0	0	0	0	2	0	3	1
E29	0	2	0	0	2	1	0	1
PB1	0	0	0	0	0	0	3	0
PB7	0	0	0	0	0	0	3	0
PB8	0	0	0	0	0	0	3	0
PB9	0	0	0	0	0	0	3	0
BU3	0	0	0	0	1	0	0	2
HY7	1	1	0	0	0	0	0	1
CJ1	13	0	0	0	0	0	0	1
E6	1	1	2	0	0	2	1	2
ST16	0	1	0	0	1	0	0	4

3.3.5 Clustering experiments and results

Experiments were done on both one-word tokens and collocation using K- mean clustering algorithms. K- mean clustering is selected in this research to demonstrate the problem and because of its simplicity and reported adequate efficiency in previous research such as (AL Qady & Kandil, 2014). RapidMminer software was used for executing K-mean clustering algorithms for all experiments (see Figure 3-8).

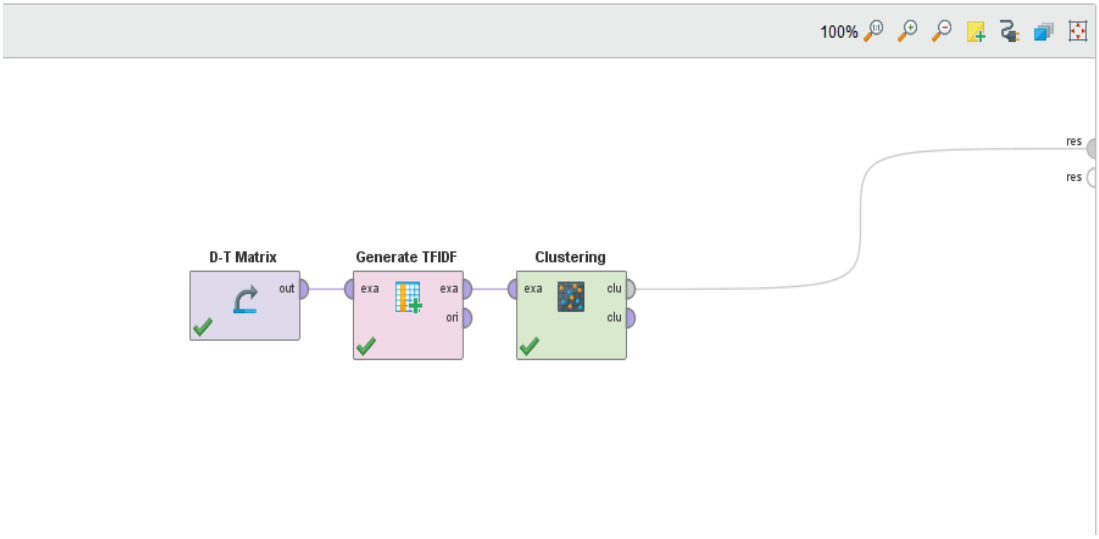


Figure 3-8 RapidMiner software environment

Three D-t matrices were developed for experiments. Documents were tabulated with features using frequency, case occurrence, and TF-IDF to measure which one has more influence on the performance of clustering algorithms. 526 single word tokens were selected based on the highest TF-IDF value. Similarity parameter is another input of a clustering algorithm. Cosine similarity is used to measure the similarity between documents. Another input to K-Mean clustering algorithm is the number of cluster K. The number of cluster is chosen to be the same as the number of document classes or

groups of selected document samples (9 classes of pipeline construction activities). The outcomes of each clustering experiment are shown in Table 3-7,Table 3-8, and Table 3-9. D-t matrices using collocation feature are developed the same way as those done for single word D-t matrices. Examples of clustering outcomes using TF-IDF are shown in Table 3-10.

Table 3-7 Cluster output using frequency (single word feature)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
W1	E1	E2	C1	TB1	W33	E28	ST1	E37
W4	E3	E4	C2	TB3	W35	W2	ST2	HY1
W10	E9	E6	C3	TB4	W36	W6	ST3	HY2
W12	E13	E7	C4	TB5	W37	W8	ST8	HY3
W14	E15	E8	C5	TB6	W38	W11	ST10	HY4
W17	E16	E10	C6	TB7	W39	W15	ST11	HY5
W18	E20	E11	C7	TB9	W40	W24	ST12	HY6
W19	E21	E12		TB10	W42	W41	ST13	HY7
W22	E22	E14		TB12	W44	W47	ST14	HY8
W23	E24	E17			W45		ST15	HY9
W25	E25	E18			BU1		ST16	
W43	E31	E19			BU2		ST17	
W46	E33	E23			BU3		ST18	
W49	E34	E26			BU4		PB1	
		E27			BU5		PB2	
		E29			BU6		PB3	
		E32			BU7		PB4	
					BU8		PB5	
					BU9		PB6	
					BU10		PB7	
					BU11		PB8	
					BU12		PB9	
					BU13		E36	
					BU14			
					BU15			
					CJ1			
					CJ2			
					CJ3			
					CJ4			
					CJ5			
					CJ6			
					CJ7			
					CJ8			
					CJ9			
					CJ10			
					C8			
					C9			
					C10			
					C11			

Table 3-8 Clustering output using TF-IDF (single word feature)

Cluster 1	Cluster2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
E1	E13	BU1	ST1	C1	E28	E2	HY1	E12
E3	E22	BU2	ST2	C2	W1	E4	HY2	E20
E9	E24	BU3	ST3	C3	W2	E6	HY3	TB1
E15	E31	BU4	ST10	C5	W4	E7	HY4	TB3
E16	E33	BU5	ST11	C6	W6	E8	HY5	TB4
E21	E34	BU6	ST12	C7	W8	E10	HY6	TB5
E25	ST8	BU7	ST13		W10	E11	HY7	TB6
PB1	PB4	BU8	ST14		W11	E14	HY8	TB7
PB2	E36	BU9	ST15		W12	E17	HY9	TB9
PB3		BU10	ST16		W14	E18	CJ1	TB10
PB5		BU11	ST17		W15	E19	CJ2	TB12
PB6		BU12	ST18		W17	E23	CJ3	
PB7		BU13			W18	E26	CJ4	
PB8		BU14			W19	E27	CJ5	
PB9		BU15			W22	E29	CJ6	
		E37			W23	E32	CJ7	
		C4			W24		CJ8	
		C8			W25		CJ9	
		C10			W33		CJ10	
		C11			W35			
					W36			
					W37			
					W38			
					W39			
					W40			
					W41			
					W42			
					W43			
					W44			
					W45			
					W46			
					W47			
					W49			
					C9			

Table 3-9 Clustering output using case occurrence (single word feature)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
W2	E2	W1	ST1	C1	E6	BU1	E1	CJ2
W6	E4	W4	ST2	C2	E7	BU2	E3	CJ3
W8	E10	W10	ST3	C3	E8	BU4	E9	CJ4
W11	E11	W12	ST10	C4	E24	BU5	E12	CJ5
W15	E14	W14	ST11	C5	E28	BU6	E13	CJ6
W24	E17	W17	ST12	C6	E29	BU7	E15	CJ7
W33	E18	W18	ST13	C7	E32	BU8	E16	CJ8
W35	E19	W19	ST14		ST8	BU9	E20	
W36	E23	W22	ST15		W41	BU10	E21	
W37	E26	W23	ST16		TB1	BU11	E22	
W38	E27	W25	ST17		TB3	BU12	E25	
W39		W46	ST18		TB4	BU13	E31	
W40		W49	HY1		TB5	BU14	E33	
W42			HY2		TB6	BU15	E34	
W43			HY4		TB7	HY7	PB1	
W44			HY5		TB9	HY8	PB2	
W45					TB10	C8	PB3	
W47					TB12	C9	PB4	
					BU3	C10	PB5	
					E36	C11	PB6	
					E37		PB7	
					HY3		PB8	
					HY6		PB9	
					HY9			
					CJ1			
					CJ9			
					CJ10			

Table 3-10 Clustering output using TF-IDF (collocation feature)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
E2	E12	E20	E24	ST1	W33	W1	E22	E1
E4	ST8	W25	CJ1	ST2	W35	W2	E28	E3
E6	PB1	HY3	CJ2	ST3	W36	W4	W37	E9
E7	PB2	HY6	CJ3	ST10	W40	W6	W38	E13
E8	PB3	C1	CJ4	ST11	W44	W8	W39	E15
E10	PB4	C2	CJ5	ST12	W45	W10	W41	E16
E11	PB5	C3	CJ6	ST13	BU1	W11	W42	E21
E14	PB6	C5	CJ7	ST14	BU2	W12	TB1	E25
E17	PB7	C6	CJ8	ST15	BU4	W14	TB6	E29
E18	PB8	C7	C4	ST16	BU5	W15	TB9	E31
E19	PB9			ST17	BU6	W17	TB10	E33
E23	TB3			ST18	BU7	W18	TB12	E34
E26	TB4			HY1	BU8	W19	BU3	
E27	TB5			HY2	BU9	W22	HY9	
E32	TB7			HY4	BU10	W23	CJ9	
	E36			HY5	BU11	W24	CJ10	
	E37			HY7	BU12	W43		
				HY8	BU13	W46		
					BU14	W47		
					BU15	W49		
					C8			
					C9			
					C10			
					C11			

3.3.6 Clustering validation

To evaluate clusters, purity measures are calculated for each clustering experiment using equation (3-5). Evaluation of clustering experiments shows that a one-word token strategy has more accurate clustering output than a collocation token strategy. One-word tokens are more powerful in representing all JHA documents than collocation due to their wide frequency distribution. Experiments show that using TF-IDF has a significant impact on clustering algorithm results as shown in Figure 3-9 and Figure 3-10. Considering document occurrences in token weighting can enhance the performance of a machine learning algorithm in discriminating between documents.

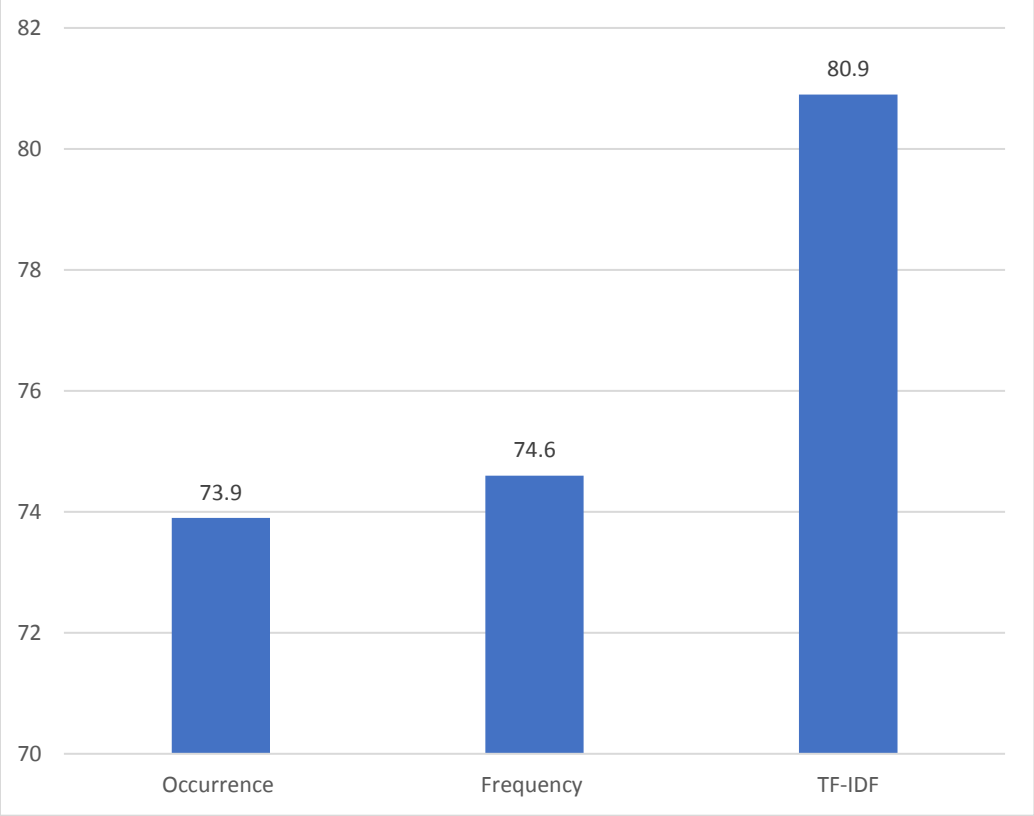


Figure 3-9 Purity measures for documents clustering using word tokens

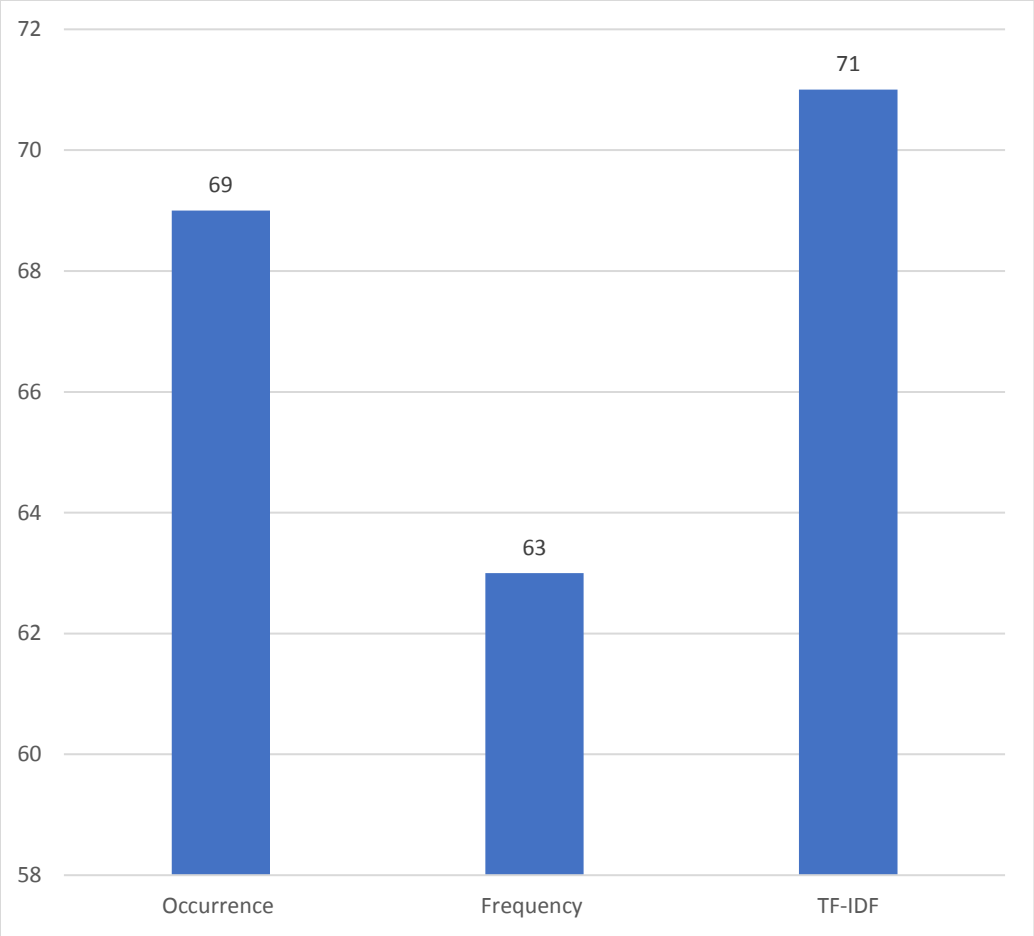


Figure 3-10 Purity measures for documents clustering using collocation tokens

JHA document clustering helps identify classes of the documents without any previous information about the documents. In addition, the clustering process makes labeling of JHA documents easier. Labeling groups of documents in clusters is much convenient than the labeling of hundreds of documents individually. Manual labeling for documents takes much time and significant manpower (see Figure 3-11). Transferring JHA documents from unstructured and scattered status to organized class groups will enable building of a classification model using pre-defined classes.

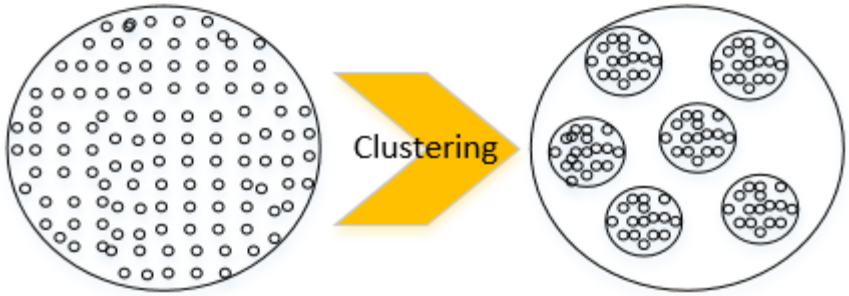


Figure 3-11 Labeling cluster groups

3.4 JHD Classification model for future JHA documents

3.4.1 Introduction

The machine Learning approach has been used by construction research for more than two decades but has still not been applied to the safety domain (Tixier & Hallowell, 2016). Classification is one of the important machine learning approaches and is used in classifying data based on predefined labels. Classification is a supervised learning process because it is based on learning from previous cases or examples (training data set). Classification is generally more powerful than clustering because it predicts classes based on previous identified categories based on client requirements or targets (Aggarwal, 2015).

Classification modeling is used in different domains. It is used in business marketing by using a previous transaction as an example to predict specific labels that reflect customer groups or interests. Medicine is another field of application of classification, where they use patient records (tests or treatment plans) as an example to extract features and predict either diagnosed diseases or treatment outcomes. Classification modeling is used also in different multi-media analysis. Nowadays, data production is

enormous due to the existence of the internet and social networks (Aggarwal, 2015). The objective of job hazard documents classification is to build classification models to classify future JHA documents, using outputs of document clustering and labeling stage as a training set to create a robust classification model.

3.4.2 Background

Classification is the primary step of organizing the documents in targeted classes prior to further knowledge mining. Caldas, et al. (2002) used text classification to predict project documents. The documents represent the construction topics of minutes of meeting (MOM). Several algorithms were tested in the classification of documents; the best performance was by using support vector machines with an accuracy 91.12 %.

Tixier & Hallowell, (2016) developed an injury report prediction model to predict injury type, energy type, and body part. The model extracted text features or attributes from injury reports using natural language processing (NLP) and predicted certain classes using classification algorithms.

Williams & Gong, (2014) developed a prediction model to predict the level of cost overrun or under run for highway projects in California in the United States. The model integrates two types of data, numerical and textual, as an input to the model. The text was collected from sources like work description of significant items in bidding documents and text from project summary. Text data was transformed to a numerical data matrix to be combined with numerical cost data. It was found that specific words and word pair are associated with the level of cost overrun of construction projects. Despite the small amount of text used in the prediction model, text data shows a

positive impact on the prediction results. Zhang, et al. (2016) used text mining and a classification method for solving problems of inconsistency between narrative crash incident descriptions and recorded hazard actions in crash reports. Three algorithms were used to build classifiers for experiments. The accuracy of classifiers was 83.94% using Naïve Bays, 67.88 % using Decision tree and 80.32% using SVM.

3.4.3 Classification methods

1- Decision tree

In early 1980, a decision tree algorithm was developed by J. Ross Quinlan. A Decision tree model is a supervised learning method for classifying data and generating a decision tree based on testing the attributes associated with data examples. For example, the Decision tree in Figure 3-12 was generated based on testing attributes X and Y. If X >1 is true, then the value of Y will define which class the tuple belongs to. If X>1 is false, then the tuple belongs to class 1.

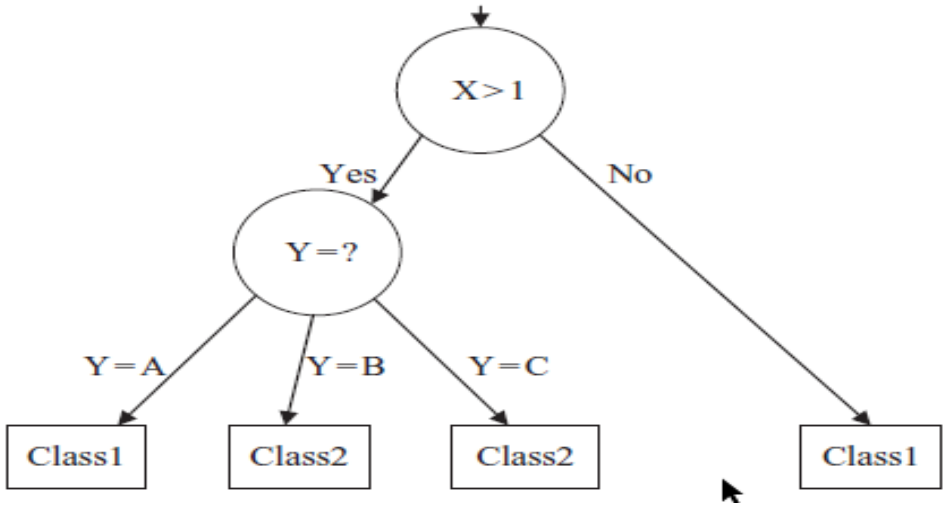


Figure 3-12 Decision tree example (X and Y are attributes) (Kantardzic, 2011)

Decision tree algorithms require three main inputs as follows:

- Data examples and their associated classes;
- Features that are describing examples (documents);
- Attribute selection method, a procedure for selecting best attributes that can discriminate between data tuples based on their classes.

A decision tree is first built using a training set of data tuples and is based on a top-down strategy. The training set is divided recursively into smaller subsets based on attribute tests to form the tree. Figure 3-13 shows the basic form of a decision tree algorithm (Han & Kamber, 2006).

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Figure 3-13 Basic algorithm for decision tree (Han & Kamber, 2006)

One of the commonly used attribute selection methods is information gain (IG). IG is used for dimension reduction that is required for classification. The information gain

approach minimizes the number of tests needed to classifying tuples of data sets (Dash & Liu, 1997).

2- Naïve Bayes

Bayes classifier is statistical classifier and built on Bayes theory, which is built on conditional probability. Bayes theorem estimates the conditional probability of a class variable given known observations about the feature variables (Aggarwal, 2015). Naïve classifier assumes the influences of features in a specific class are independent, called class conditional independence. It is designed in this way to simplify computation and because of that it is called Naïve (Han & Kamber, 2006). Although this assumption is generally not true especially for text feature associated with documents, Naïve Bayes showed sufficient accuracy and was very efficient in classification data (Caldas, et al., 2002). Given C is corresponding to a class variable and F is corresponding to feature, calculating the conditional probability P(C/F) can be estimated using following equation:

$$P(C|F = \frac{P(F|C)\ P(C)}{P(F)} \tag{3-8}$$

3- K-NN (Nearest neighbor)

K-nearest neighbor classifier depends on the calculation of the distance between documents and uses high ranked similar K documents predicting the classes of unlabeled document, (YANG, 1999). The simple algorithm of K-Nearest Neighbor is shown in:

1-Compute similarity of new documents in collection {D (1)}
2-Select k documents that are most similar to new Doc
3-The answer is the label that occurs most frequently in the k selected documents.

Figure 3-14 K-NN algorithm for documents classification (Weiss, et al., 2015)

3.4.4 Classification performance measurement

Measuring performance of classification is the assessing process of how accurate classification model is in predicting document labels. Classification accuracy and error rate are calculated according to the following equations:

$$Accuracy = \frac{TP + TN}{P + N} \tag{3-9}$$

$$Error\ rate\ = \frac{FP + FN}{P + N} \quad or \quad 1 - Accuracy \tag{3-10}$$

Where:

True positive (TP): Number of positive documents that were correctly labeled

True negative (TN): Number of negative documents that were correctly labeled

False positive(FP): Number of negative documents that were incorrectly labeled as positive

False negative(FN): Number of positive documents that were incorrectly labeled as negative.

Classification effectiveness is usually measured in terms of recall and Precision measures, (Sebastiani, 2002). Recall is defined as the percentage of positive documents classified as such. Precision is the percentage of documents that are correctly labeled. Recall and precision can be estimated as per the following equations:

$$Recall = \frac{TP}{TP + FP} \tag{3-11}$$

$$Precision = \frac{TP}{TP + FN} \tag{3-12}$$

Combining recall and Precision measurements in one single measure is an alternative way for using recall and precision factors. Recall and precision can be combined using F-measure (F-score), a harmonic means of precision and recall and can be calculated as follows:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{3-13}$$

$$or\ F_\beta = \frac{(1 + \beta^2) * precition * recall}{\beta^2 * prccision + recall} \tag{3-14}$$

Where: β is a positive real number. F_β is a weighted measure of both recall and precision. It assigns the same weight β to both recall and precision.

3.4.5 JHA Documents classification methodology

Labeled JHA documents are selected to build classification model. Collected labeled documents are the output of the clustering and labeling stages. Part of JHA documents will be used as training data to train the classifier in predicting classes of future stream documents. Test documents shall be used to evaluate the classifiers.

Figure 3-15 shows the classification methodology of JHA documents. 250 JHA labeled documents were collected for building a classification model. The documents belong to 15 classes of pipeline construction activities as shown in Table 3-11. The document pre-processing phase consists of the same steps discussed in the clustering part. Tokenization, Lemmatization, and exclusion were applied during preprocessing of JHA documents. A document-term matrix was developed and TF-IDF were used for weighting and filtering the features. Several classification algorithms were used in experiments, and classifiers' performances were evaluated.

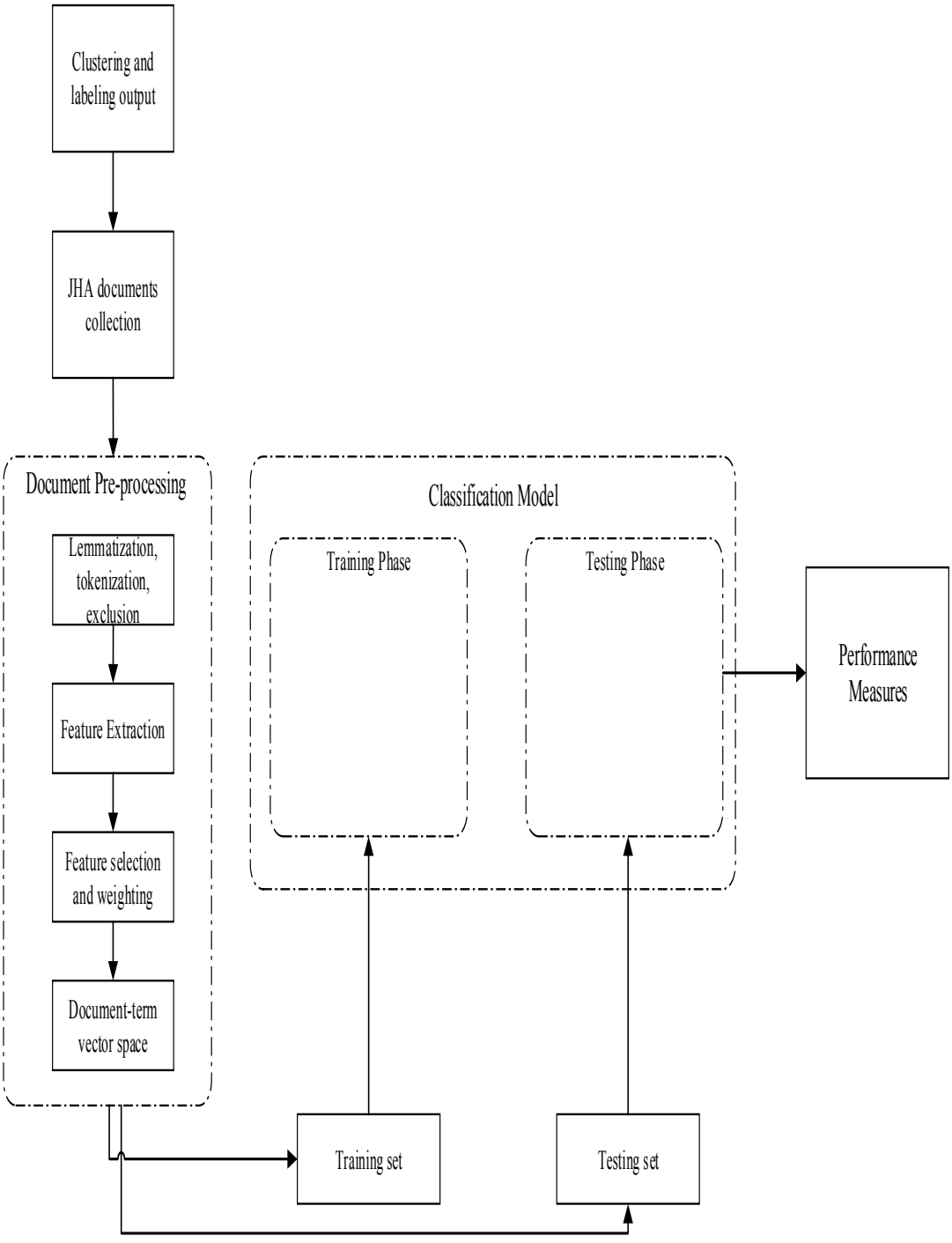


Figure 3-15 JHA documents Classification methodology

Table 3-11 JHA documents classes

Classes	No. of Documents	Doc. Percentage
Excavation	34	13.60%
Backfilling	20	8.00%
Clearing and grubbing	12	4.80%
Moving and lifting	24	9.60%
Stringing	13	5.20%
Pipe bending	9	3.60%
Surveying	10	4.00%
Hydrovac	22	8.80%
Cutting pipe	11	4.40%
Coating and jeeping	10	4.00%
Hydro testing	9	3.60%
Confined space entry	14	5.60%
Welding	38	15.20%
Trench box installation	9	3.60%
Bolt up un bolt	15	6.00%
Total	250	100%

3.4.6 Classification Experiments and results validation:

JHA documents were divided randomly into a training set and a testing set. 175 (70%) JHA documents were assigned to the training set. 30 documents (30%) were allocated to the testing set. Since the JHA documents are not uniformly distributed over the 15 classes, testing documents were sampled based on stratified sampling to make sure the testing documents had the same distribution of the JHA documents collection.

Document – term matrix was developed as an input for the classification algorithms. The matrix involved documents as examples (rows) and 1652 features (columns). In addition, the matrix had a label column that contains classes of JHA documents. TF-IDF parameters were used to tabulate documents with features.

Naïve Bayes, Decision Tree, and K-NN classification algorithms were selected to apply the classification methodology. Algorithms were chosen based on previous

research results (Zhang, et al., 2016; Caldas, et al., 2002; Al Qady & Kandil, 2015). The results of classification are shown in confusing matrices in Table 3-12, Table 3-13 and Table 3-14.Table 3-12 Confusion matrix for Naïve Bayes classifier

The confusion matrix contains information about true and false predicted documents using previously mentioned classification algorithms. In the confusion matrix, columns represent true or actual class and rows represent predicted classes by classifiers. The diagonal cells in the matrix represent the documents that were labeled correctly by the classification algorithm, while the off-diagonal cells are the wrong classification.

Precision and recall are calculated for each of JHA document classes. Precision is in the left column and recall in the bottom row as shown in the confusion matrices. Average recall and precision were calculated and tabulated in Table 3-15.The analysis of classification outputs shows satisfactory performance of K-NN. K-NN was built using different similarity measures: Euclidean distance, Jaccard similarity and cosine similarity. Using cosine similarity produces the best results which is 93.33 % accuracy. The accuracy of K-NN using Euclidean distance and Jaccard are 92 %. Naïve Bayes classification accuracy was 89.33 %, and Decision Tree accuracy is 77.33 % as shown in Table 3-15.

Table 3-12 Confusion matrix for Naïve Bayes classifier

Predict \ True	Excavation	Stringing	Pipe bending	Welding	Trench box installation	Bolt up un bolt	Hydrotesting	Coating and jeeping	Cutting	Backfilling	Confined space	Surveying	Clearing and grubbing	Moving and lifting	Hydrovacing	**Class precision**
Excavation	8	0	0	0	0	0	0	0	0	0	0	0	0	0	2	80.0%
Stringing	0	4	0	0	0	0	0	0	0	1	0	0	0	1	0	66.7%
Pipe bending	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Welding	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	100.0%
Trench box installation	1	0	0	0	3	0	0	0	0	0	0	0	0	0	1	60.0%
Bolt up un bolt	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	80.0%
Hydrotesting	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	100.0%
Coating and jeeping	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	100.0%
Cutting	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	100.0%
Backfilling	1	0	0	0	0	0	0	0	0	5	0	0	0	0	0	83.3%
Confined space	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	100.0%
Surveying	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	100.0%
Clearing and grubbing	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	100.0%
Moving and lifting	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	100.0%
Hydrovacing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	100.0%
Class recall	80.0%	100.0%	100.0%	90.9%	100.0%	100.0%	100.0%	100.0%	100.0%	83.3%	100.0%	100.0%	100.0%	85.7%	57.1%	

Table 3-13 Confusion matrix for Decision Tree classifier

	True															
	Excavation	Stringing	Pipe bending	Welding	Trench box installation	Bolt up un bolt	Hydrotesting	Coating and jeeping	Cutting	Backfilling	Confined space	Surveying	Clearing and grubbing	Moving and lifting	Hydrovacing	**Class precision**
Excavation	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Stringing	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Pipe bending	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Welding	0	0	0	10	0	1	0	0	0	0	0	0	0	0	0	90.9%
Trench box installation	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	50.0%
Bolt up un bolt	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	75.0%
Hydrotesting	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	100.0%
Coating and jeeping	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	75.0%
Cutting	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	100.0%
Backfilling	2	0	0	1	1	0	0	0	0	5	0	0	1	0	2	41.7%
Confined space	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	100.0%
Surveying	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	75.0%
Clearing and grubbing	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	100.0%
Moving and lifting	0	2	0	0	0	0	1	0	0	0	0	0	0	6	1	60.0%
Hydrovacing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	100.0%
Class recall	60.0%	50.0%	100.0%	90.9%	66.7%	75.0%	33.3%	100.0%	100.0%	83.3%	100.0%	100.0%	75.0%	85.7%	57.1%	

Table 3-14 Confusion matrix for K-NN classifier (cosine similarity)

Predict \ True	Excavation	Stringing	Pipe bending	Welding	Trench box installation	Bolt up un bolt	Hydrotesting	Coating and jeeping	Cutting	Backfilling	Confined space	Surveying	Clearing and grubbing	Moving and lifting	Hydrovacing	**Class precision**
Excavation	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Stringing	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Pipe bending	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	100.0%
Welding	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	100.0%
Trench box installation	1	0	0	0	3	0	0	0	0	0	0	0	0	0	1	60.0%
Bolt up un bolt	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	80.0%
Hydrotesting	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	100.0%
Coating and jeeping	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	75.0%
Cutting	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	100.0%
Backfilling	1	0	0	0	0	0	0	0	0	5	0	0	0	0	0	83.3%
Confined space	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	100.0%
Surveying	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	100.0%
Clearing and grubbing	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	100.0%
Moving and lifting	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	100.0%
Hydrovacing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	100.0%
Class recall	80.0%	100.0%	100.0%	90.9%	100.0%	100.0%	100.0%	100.0%	100.0%	83.3%	100.0%	100.0%	100.0%	100.0%	85.7%	

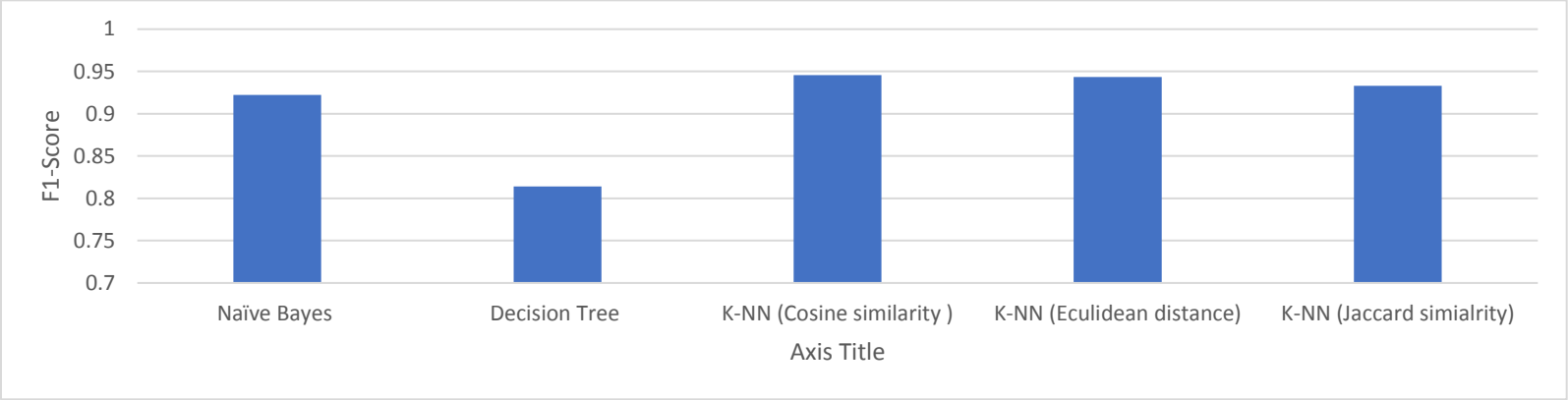


Figure 3-16 F1 Score for different classification algorithms

Table 3-15 Classification algorithms performance measures

Classification Algorithm	Recall	Precision	F1-Score	Accuracy	Error rate	Kappa statistics
Naïve Bayes	93.1%	91.3%	0.922	89.33	10.67%	0.884
Decision Tree	78.50%	84.50%	0.814	77.33%	22.67%	0.753
K-NN (Cosine similarity)	96%	93.20%	0.946	93.33	6.67%	0.928
K-NN (Euclidean distance)	94.43%	94.28%	0.944	92%	8%	0.918
K-NN (Jaccard similarity)	93.27%	93.33%	0.933	92%	8%	0.913

3.5 Conclusion

At the beginning of this research JHA forms were collected from several pipeline constructions projects. However, the JHA documents were unlabeled and did not belong to any predefined categories based on construction activities. Text mining was used to categorize JHA documents. Using clustering techniques enabled organizing documents in similar groups. The K- mean clustering method was used and performed well in clustering JHA documents. Using word tokens as features for clustering is better than using collocations.

Clustering process made labeling of JHA documents easier and faster. Labeling groups of documents is much convenient than the labeling of hundreds of documents individually. Manual labeling for documents takes much time and significant manpower. Clustering stage output was used as an input for the classification model. it is used as training and testing data sets for Classification model. Several classification algorithms were used in the classification experiments. K-NN and Naïve bays performed well in classifying JHA documents.

Document organization can enable JHA documents examples retrieval to assist in performing future JHA processes. Furthermore, organizing JHA forms in construction activities classes will help for further hazard knowledge analysis and extraction. knowledge analysis will enable exploration and discovery of the pattern and relationships between hazard concepts in the pipeline project domain. In the next chapter, further knowledge extraction to build hazard knowledge schema will be discussed in detail.

CHAPTER 4

JHA content analysis and concepts extraction

4.1 Background

Organizing construction JHA documents can improve the job hazards identification process by enabling accessing similar past JHA documents during future JHA preparation (Goh & Chua, 2010). However, construction professionals still need to explore documents and extract the best combination of knowledge to prepare a new JHA for new construction activity.

Although organizing documents improves document retrieval for future use, not all previous documents have the same quality in terms of hazard knowledge content. To overcome this problem, construction professionals must scan and explore several JHA forms to extract the best knowledge of hazards identified for previous construction jobs. Pipeline Construction execution strategy depends on dividing projects into repetitive segments and each segment consists of repetitive activities.

Knowledge in JHA forms are all in text format. Extracting knowledge concepts is important part of text mining approach. However, extracting full knowledge schema using fully automated knowledge extraction strategy is still far from reality. Human efforts are still needed for validating and evaluating extracted knowledge concepts.

Tixier & Hallowell, (2016) developed a model to automatically extract incidents precursor attributes, energy sources, and body parts using natural language processing (NLP). The model is aided by a dictionary of keywords for matching process with scanned injuries reports. The model can quickly scan reports and extract the targeted keywords for further analysis. the research highlights the potential power of text mining for construction management in term of automatic text extraction.

Machine learning is an important part of text mining and can be used at different levels of text mining stages. Generally, text has many levels of abstraction. Text abstraction levels can be presented in documents, paragraphs, sentences, and words. Clustering and classification is used widely in document organization and retrieval (Caldas, et al., 2002; Jun, et al., 2014). Also, machine learning is used in keyword retrieval and text summarization. Clustering is a very common method to perform text abstraction and topic modeling (LI, et al., 2014; Singthongchai & Niwattanakul, 2013).

4.1.1 Collocation extraction

Collocation is a consecutive text expression which consist of two or more words happened together frequently and explain one idea. collocation include noun phrases such as “personal computer,” “project management,” and “electrical plant” or phrasal verb such as “looking for”. Extraction and selection collocations from text can be done using several approaches. These approaches are frequency method, mean and variance method, hypothesis testing method, and mutual information method (Manning & Schutze, 2001).

The frequency method is the simplest method for finding collocation in texts. If two words occur together frequently, that means they have a special meaning that can explain a concept. Frequency is perfect for fixed phrases, where the distance between consecutive words is zero. For example, the following sentences contain the two words knock and door:

a. She knocked on her door.
b. He knocked at the door.
c. Smith knocked on Ronald's door.

In the previous sentences the distance between knock and doors is not zero; hence, the frequency method is not applicable in this situation. To improve the extraction process, text preprocessing steps are essential to exclude high-frequency meaningless collocation such as "of the" or "he said." Such collocations can be eliminated by using a stop word list step. Also, using stemming or lemmatization can reduce the redundancy of words.

4.1.2 **Text co-occurrence analysis**

Generally, co-occurrence analysis is the process of counting of a co-occurred pair of text words (Buzydlowski, 2015). Keywords Co-occurrence analysis is one of the text content analysis techniques. It is used to identify relationships between extracted text entities and concepts from a textual data format (He, 1999). Co-occurrences as relationships between text concepts, is the most commonly studied relationship in the field of text mining and natural language processing (Cao & Cui, 2016). Typically, the

Co-occurrence matrix can be produced by aggregating all co-occurrences of pairs of keywords or phrases (collocation) as shown in Table 4-1.

Table 4-1 Example of co-occurrence table

	W1	W2	W3	W4	W5	W6
W1		2	3	3	4	5
W2	2		2	1	2	2
W3	3	2		3	3	0
W4	3	1	3		2	3
W5	4	2	3	2		1
W6	5	2	0	3	1	

Co-occurrence analysis for text has more advantages than qualitative text analysis. Qualitative text analysis is very expensive. Also, it takes a lot of time and effort to extract knowledge concepts. Co-occurrence method is a solution for the drawbacks of the qualitative method. Co-occurrence analysis uses different methods of visualization that amplify its advantage over qualitative text analysis, (He, 1999).

4.1.3 Link analysis

Link analysis is based on visualizing links between extracted concepts and entities. It is built based on the co-occurrence of lexical terms in text units such as paragraphs, sentences and documents. Link analysis can assist in investigating underlying patterns and structures of text terms in the text body (Feldman & Sanger, 2007).

Link analysis is presented in a network graph, where nodes represent text terms (words or collocation) and the relationships between text terms are represented by lines. The

strength of the relationships between entities is demonstrated by the thickness of the connecting line or by using similarity measures posted as line descriptions.

4.2 Content analysis and concepts extraction methodology

After document clustering and labeling, JHA were scanned to extract all hazard associated with construction activities. The methodology consists of integrated methods of text mining and qualitative approach. Text mining is used to speed up the process of hazard concepts extraction. Tokenization of words and collocation are used to extract concepts and entities. Text terms frequencies and document occurrences are used to detect the weight of text terms in the documents. Co-occurrence analysis, hierarchal clustering and link analysis were used to exposed underlying pattern and semantic relationships between text concepts in JHA forms (see Figure 4-1).

Extracted hazards' concepts are coded in a code book (also called a hazard dictionary). Adding any hazard to a code book is based on previous steps and manual evaluation. Building a hazard dictionary is a repetitive process and the final output of the process is a code book that contains all hazards concepts associated with each construction activity. Hazards concept have semantic relationships that are represented by Co-occurrences and similarity matrices. Computer programs are used to execute this methodology, QDA Miner is a qualitative text analysis program and WORDSTAT 7 is a text mining tool software. Both are products of Provalis Research company (Provalis Research, 2015).

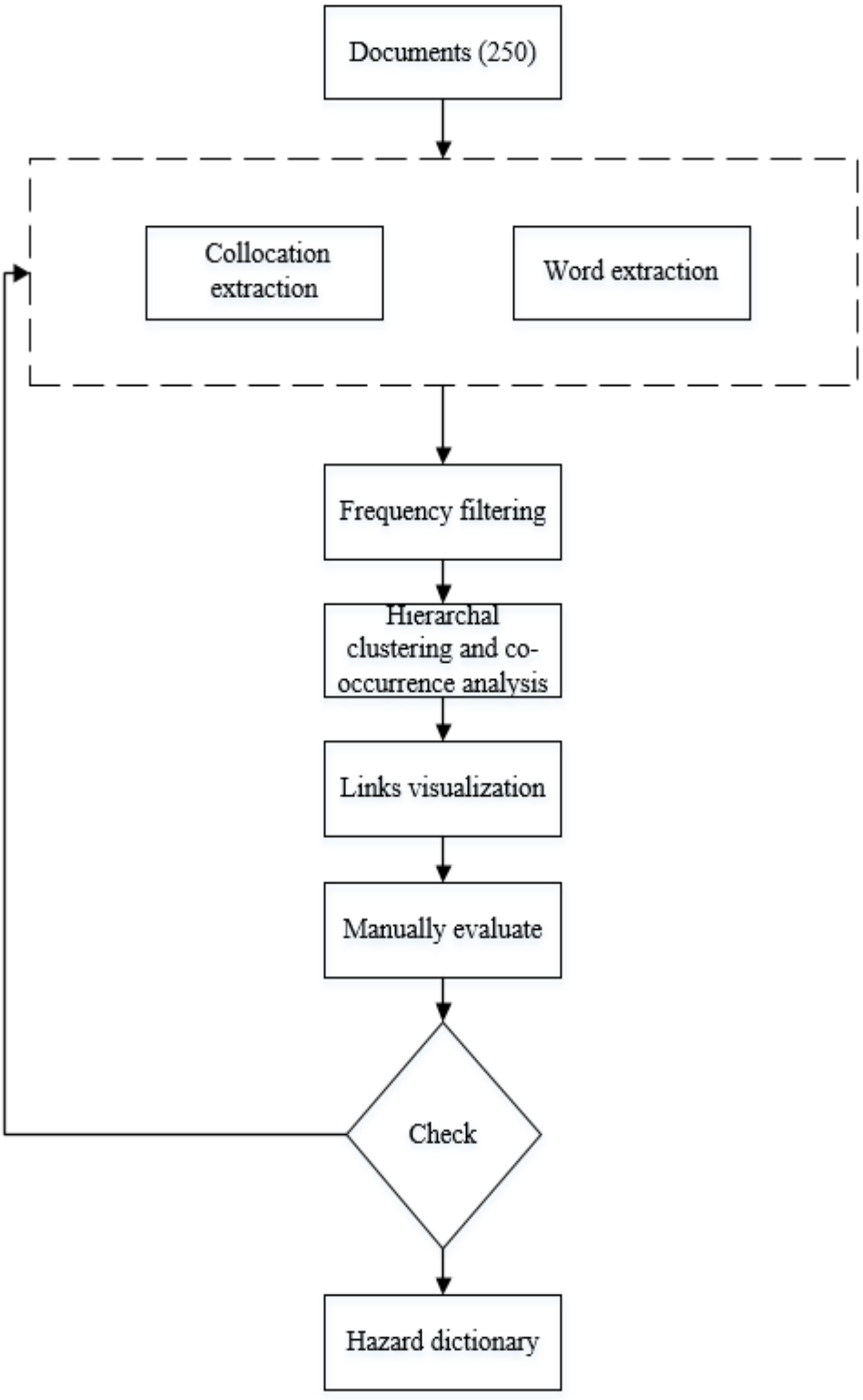


Figure 4-1 Content analysis methodology

4.3 Hazard concepts extraction process

250 JHA forms belong to 15 classes of pipeline construction projects were used for analysis, (see Figure 4-2). The process used text mining to extract text words and collocations from text using text tokenization. Frequencies and case occurrences were calculated for all words and collocations. In this stage, high-frequency words and collocations were filtered for further analysis.

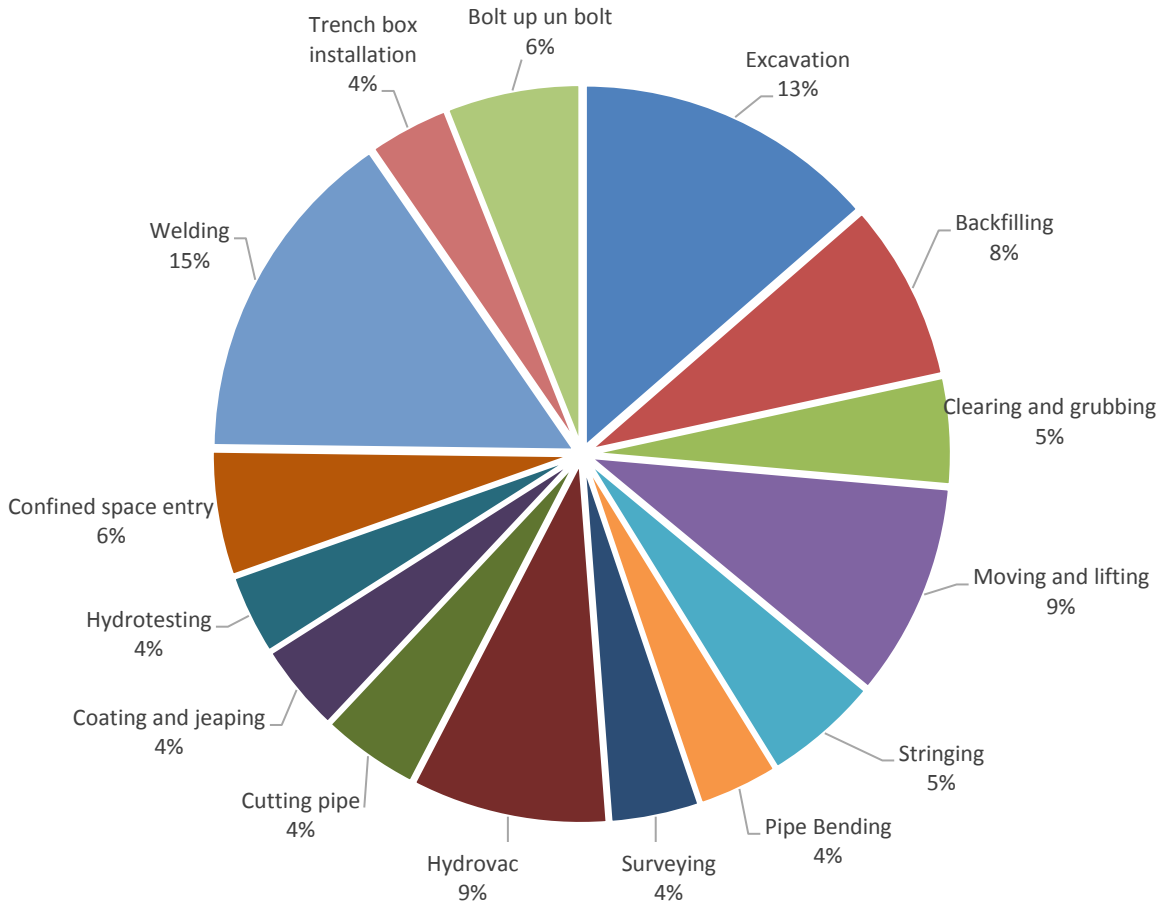


Figure 4-2 Construction activity classes

4.3.1 Single word extraction

Single word extraction is the first step in exploring the text in JHA forms for each class of pipeline construction activity. The text inside each document is segmented into single word tokens. Each word frequency is calculated to explore the word pattern among all documents and activity classes. Several words that have no meaning or have very low frequency were excluded from targeted list. For example, words that have high frequency but no meanings to add to the analysis (e.g., the, does, that) were excluded from the text.

To investigate extracted text words against construction activities, tabulation with document classes has been constructed (see Figure 4-3). Tabulating words can give information about the distribution of extracted words in an activity class and help in identifying which words are associated with a specific construction activity.

For example, text words associated with JHA documents belonging to an excavation are shown in Figure 4-4. It is observed that the word "line" is the most frequent word occurred in documents that belong to excavation class. The word "line" does not have much meaning until it is examined against activity classes. Excavation in pipeline projects is associated strongly with hazards related to existing underground utilities and overhead power lines. The words "ground," "equipment," and "condition" are other examples of words attached to the excavation class. JHA documents for pipe stringing class has many words occurred frequently as shown in Figure 4-5. Words such as "pipe," "ground," "point," "truck," and "load" are extracted from JHA forms belonging to the pipe stringing activity class.

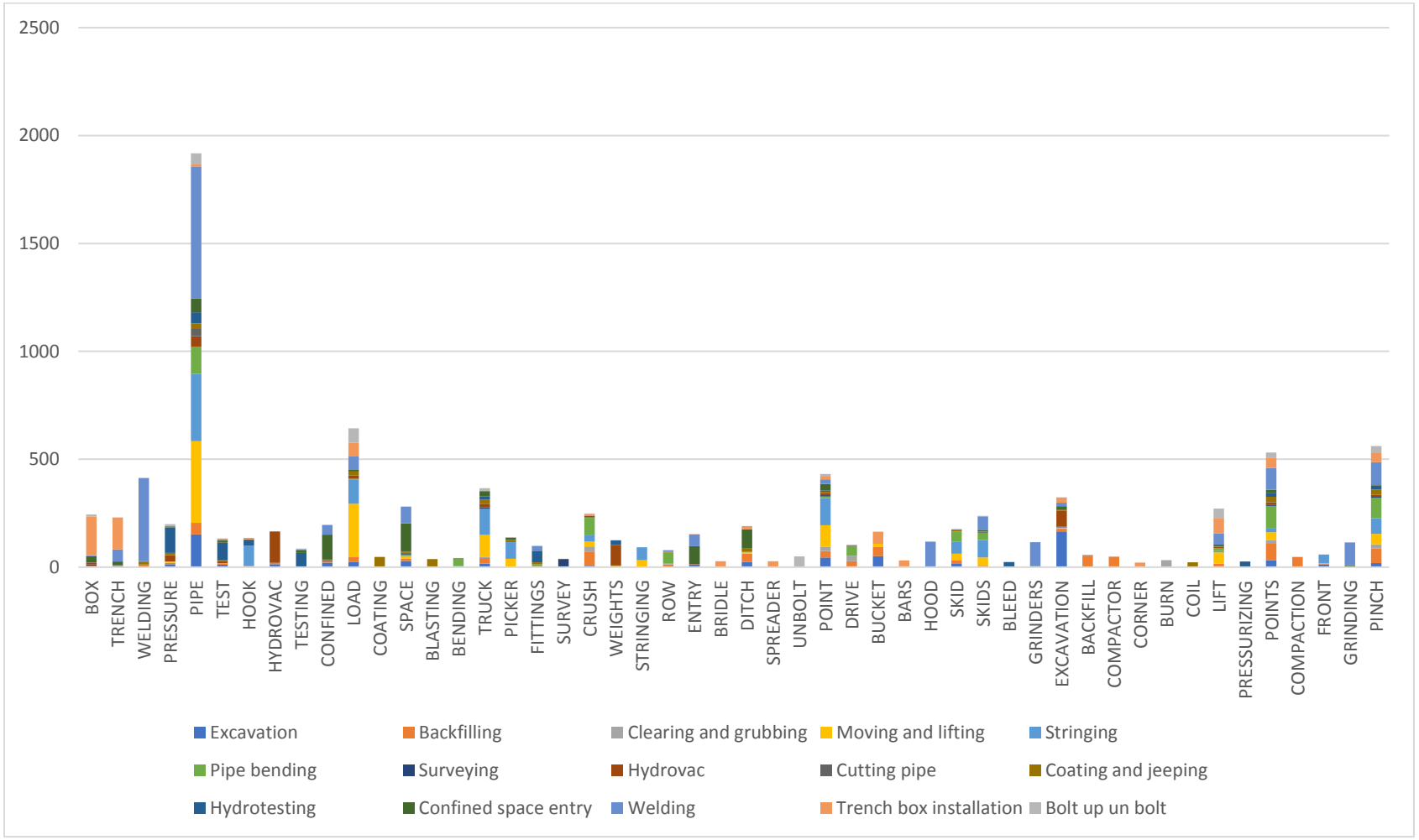


Figure 4-3 Example of extracted words tabulated with different activity classes

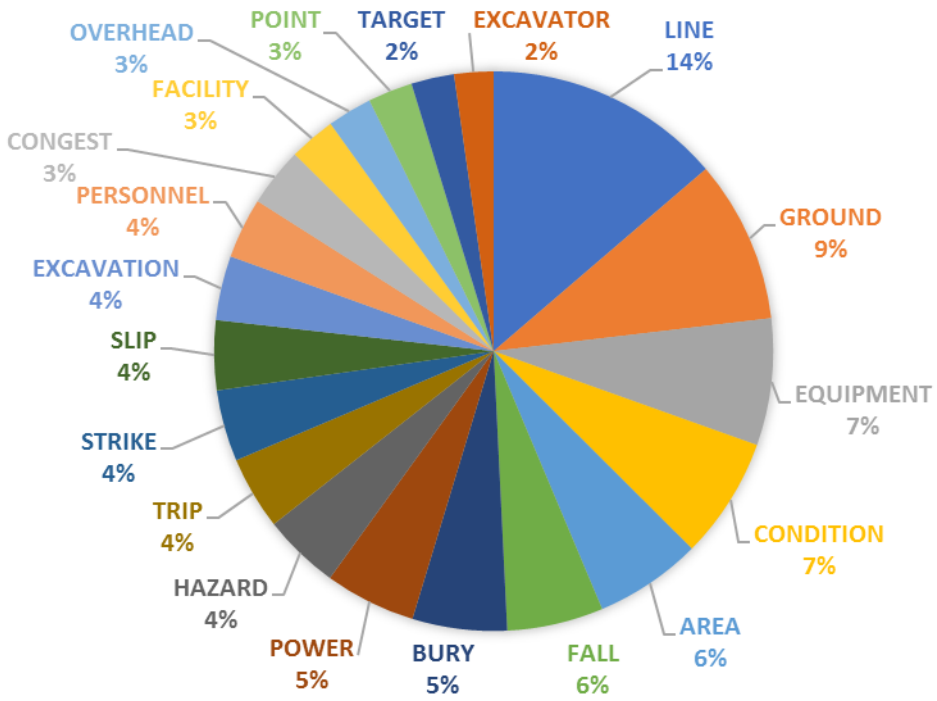


Figure 4-4: Words associated with excavation class

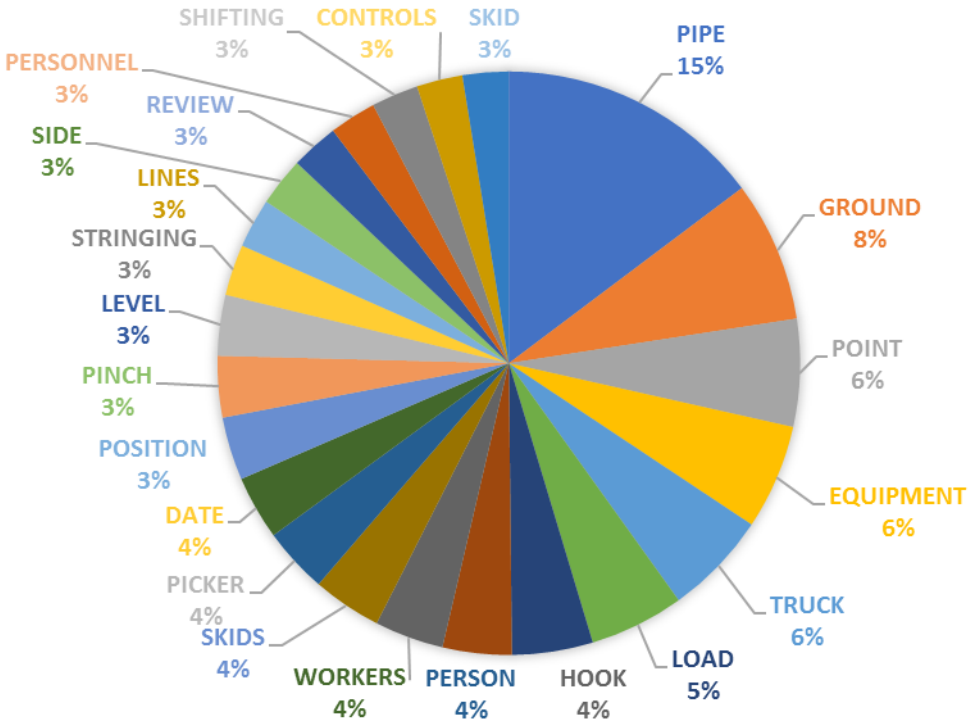


Figure 4-5 Words associated with pipe stringing activity

Not all words can provide sufficient meaning about document content. Single words are a starting point of investigation of hazard concepts. However, some words represent by itself concepts of hazards that were identified for several construction activities. These words include “wildlife,” “noise,” “competency” and “sparks,” (see Figure 4-6). These hazard concepts are added to the hazard dictionary during the present analysis step. Single word extraction did not contribute much to the hazard extraction process because most of the hazard concepts were written in the form of phrases (collocations).

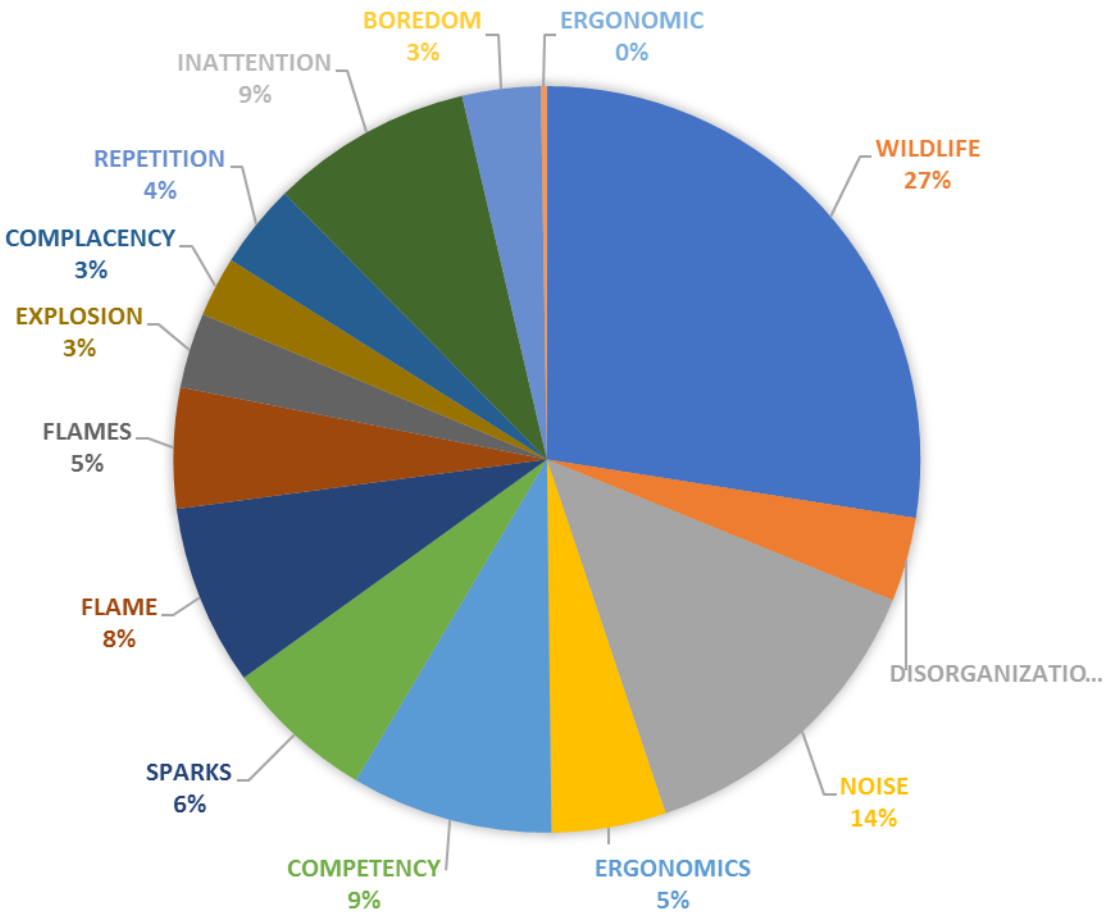


Figure 4-6 Words describe hazards’ concepts

4.3.2 Hazard collocations extraction

Collocation extraction is used to add a sufficient meaning to many single words. Extracting collocation or phrases can add more meaning to many extracted single words. Frequency and document occurrence parameter were calculated for each phrase. Example of extracted collocations are shown in Table 4-1. Collocations can be tabulated with construction activity class to examine its weight in JHA documents of construction activity classes. Figure 4-7 shows example of collocation tabulation with classes of JHA documents.

Table 4-2 Example of Collocations extracted from documents collection

Collocation	Frequency	No. Cases	Length
TRIPS AND FALLS	167	82	3
LONG SLEEVES	166	133	2
GREEN TRIANGLE	163	163	2
POINT CONTACT	163	99	2
POWER LINES	158	58	2
LONG PANTS	149	149	2
UNEVEN GROUND	138	92	2
LINE OF FIRE	129	78	3
WORK SCOPE	121	109	2
SIGNAL PERSON	120	74	2
CRUSH POINTS	113	47	2
HAND TOOLS	113	90	2
WELDING HOOD	110	32	2
RESIDUAL STEPS	109	38	2
SUSPENDED LOAD	109	62	2
TOOLBOX TALKS	109	58	2
CONTAINED AND CLEANED	108	107	3
APPROVED SIDE	107	105	2

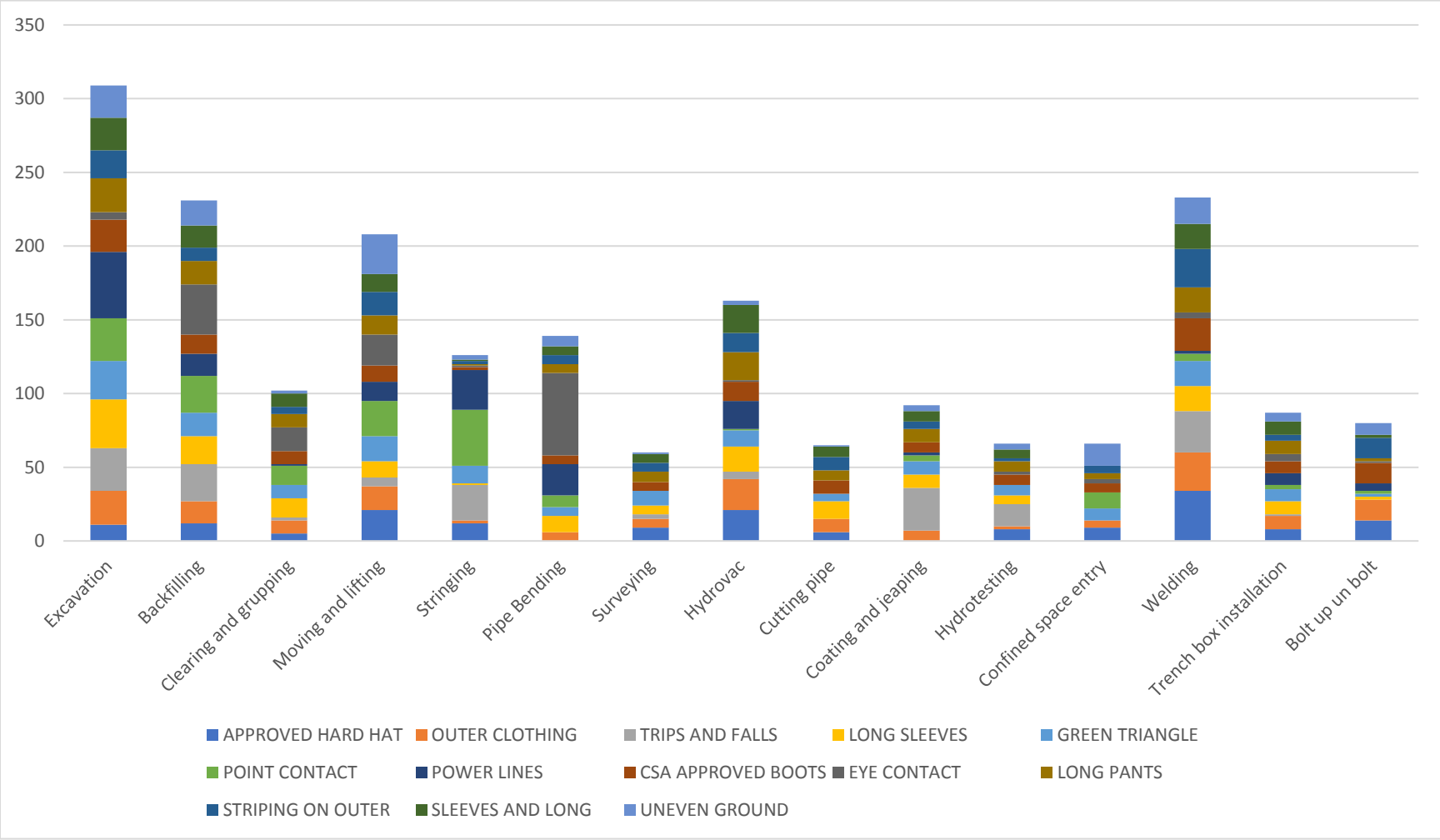


Figure 4-7 Collocations tabulated with construction activities

To extract hazard concepts from collocations, a qualitative judgment is performed for each collocation associated with each construction activity class. Figure 4-8 presents the collocations associated with pipe-bending activity. One of high frequent extracted collocation is “Pinch point “, which is a very frequent hazard associated with most pipeline construction activities. It is triggered due to the interaction between human parts such as body, hands, or foot, and other parts such as pipe, equipment, tools, and skids.

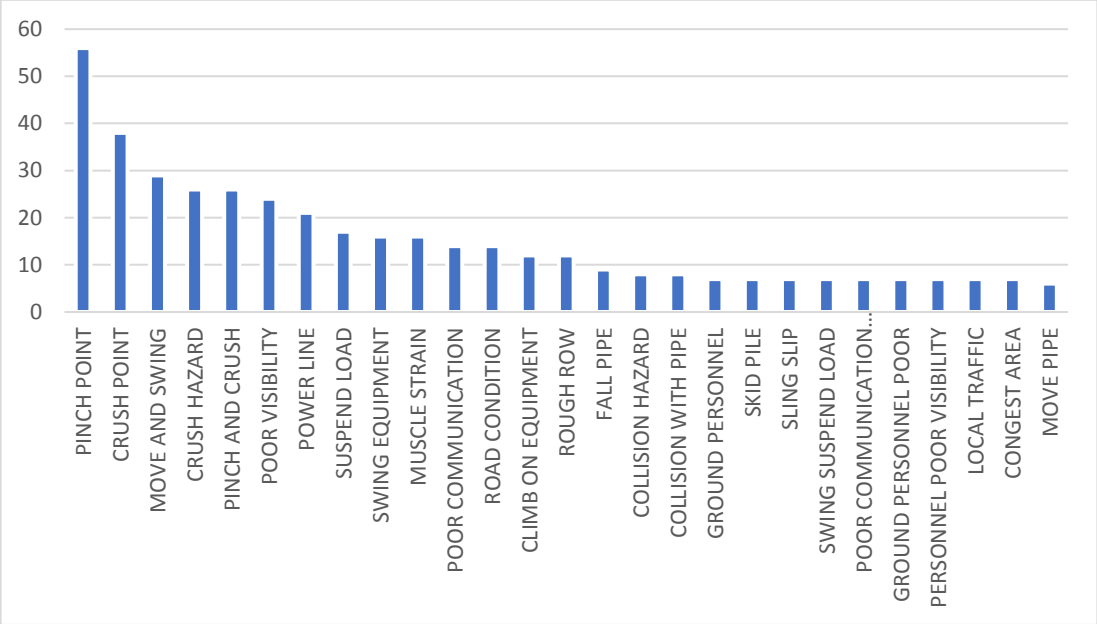


Figure 4-8 Collocations associated with pipe bending activity

Figure 4-9 presents collocations extracted from JHA forms of clearing and grubbing activity. Collocation such as “poor visibility,” “fall tree,” “slip and trip,” “road condition,” and “ground personnel” are hazards concepts and can be transferred to the hazard dictionary. Most of the hazards identified in JHA documents are in the form phrases(collocations) that consist of two or three words. Collocation extraction contributes significantly to the process of building hazards dictionary.

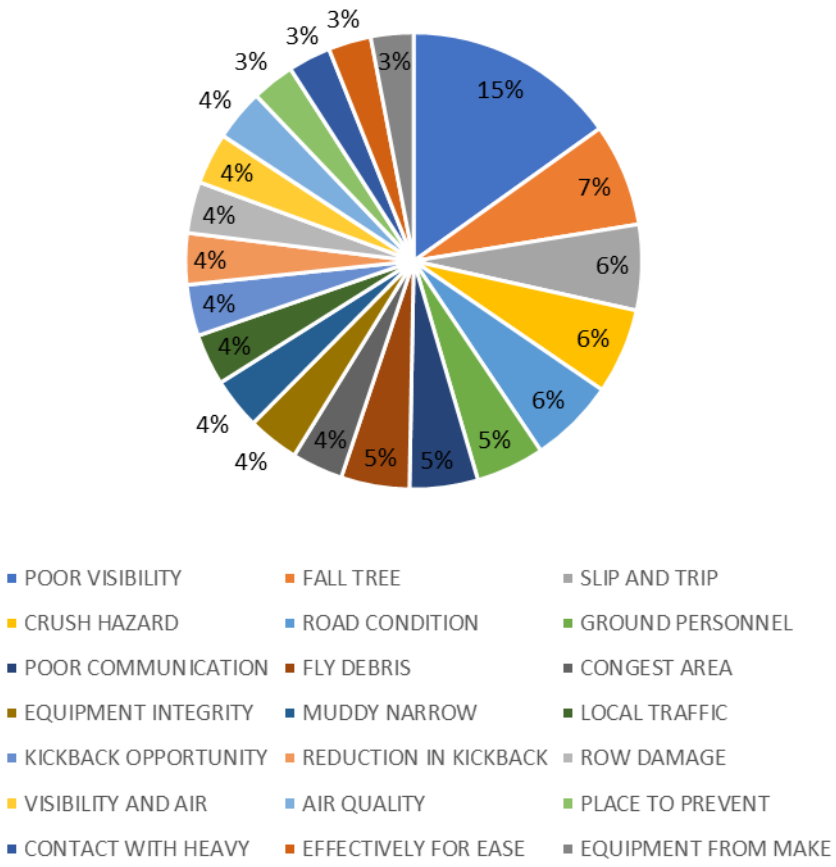


Figure 4-9 Collocations associated with clearing and grubbing

4.3.3 Co-occurrence, hierarchal clustering and link analysis for text terms

To conduct a further analysis of text words and collocations, relationships extraction using hierarchal clustering and co-occurrence analysis are performed. Relationships between text terms can help identify semantic relationships between knowledge concepts and explore the implicit patterns within the text body.

Co-occurrence analysis is conducted during concepts extraction to investigate the relationships between different words and collocations. Relations between concepts can be represented by co-occurrence matrix as shown in Table 4-3 and

Table 4-4.

Table 4-3 Words co-occurrence matrix

	ACCESS	ADDITIONAL	ADDRESS	ADEQUATE	ADJACENT	ALARM	ANKLE	ARC	ATTENTION	AUTHORIZE	AVOID	AWKWARD	BACKFIL	BACKFILL	BACKING	BALANCE
ACCESS																
ADDITIONAL	54															
ADDRESS	40	60														
ADEQUATE	84	57	46													
ADJACENT	10	17	7	7												
ALARM	33	29	6	28	10											
ANKLE	30	23	18	26	0	5										
ARC	17	6	4	25	0	8	2									
ATTENTION	16	17	7	14	4	22	0	1								
AUTHORIZE	53	39	32	40	10	19	9	0	11							
AVOID	85	69	50	71	18	28	39	11	9	41						
AWKWARD	35	19	14	21	8	9	3	10	3	16	29					
BACKFIL	12	19	4	14	11	35	0	1	21	10	17	2				
BACKFILL	7	4	2	3	1	2	0	1	9	2	4	2	4			
BACKING	12	9	6	10	0	12	0	0	11	9	7	2	10	2		
BALANCE	11	14	14	14	1	3	13	1	3	3	17	4	2	2	0	

Table 4-4 Collocation co-occurrence matrix

	ADEQUATE_SUPPORT	AID_KIT	AIR_HORN	ASSESS_GROUND_CONDITION	AUTHORIZE_PERSONNEL	AVOID_ANY_ICE	AWKWARD_LIFT	BEND_MACHINE	BENEATH_OVERHEAD	BLIND_SPOT	BUFF_AND_GRIND	BURY_FACILITY	BURY_OBJECT	BURY_UTILITY
ADEQUATE_SUPPORT														
AID_KIT	7													
AIR_HORN	18	13												
ASSESS_GROUND_ CONDITION	22	3	20											
AUTHORIZE_ PERSONNEL	18	2	20	23										
AVOID_ANY_ICE	17	0	16	17	17									
AWKWARD_LIFT	0	4	1	0	14	0								
BEND_MACHINE	0	1	1	0	0	0	0							
BENEATH_OVERHEAD	1	11	5	2	4	0	6	1						
BLIND_SPOT	1	6	4	4	2	0	3	1	6					
BUFF_AND_GRIND	19	2	0	4	0	0	0	0	1	1				
BURY_FACILITY	0	3	13	0	9	0	8	0	4	2	0			
BURY_OBJECT	0	3	12	0	1	0	0	0	3	1	0	14		
BURY_UTILITY	1	10	15	2	3	0	5	0	10	10	1	15	15	

Text terms that is co-occurred in the text body frequently, have relationships which can be explored to identify implicit pattern. Clustering is based on calculating the similarities between text terms. The similarity index between text terms indicate how often they co-occur together in the same documents. WordStat 7 from Provails Research is used to perform terms hierarchal clustering for text terms extracted from JHA documents (Provalis Research, 2015). Terms hierarchal clustering is a visual method to explore terms that have co-occurred in the same document classes (see Figure 4-10 and Figure 4-11).

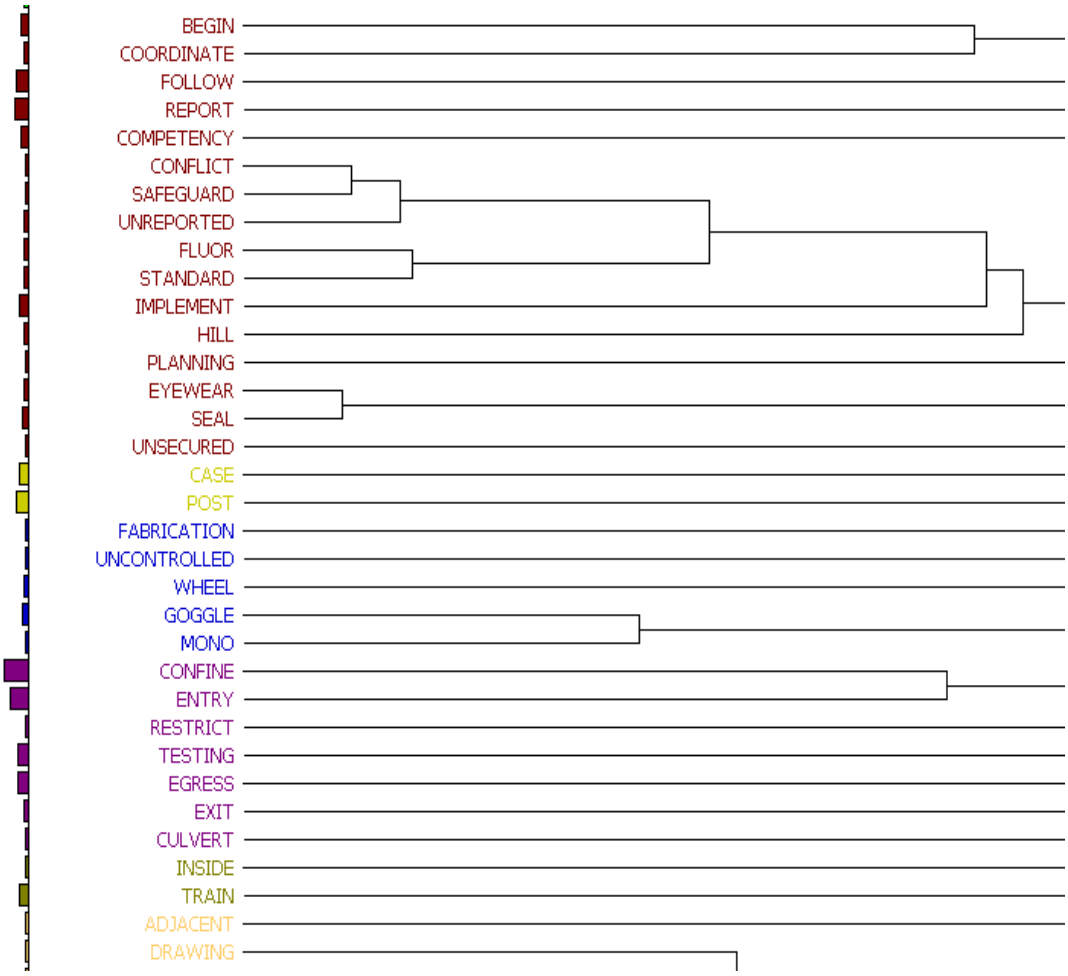


Figure 4-10 Words hierarchal clustering

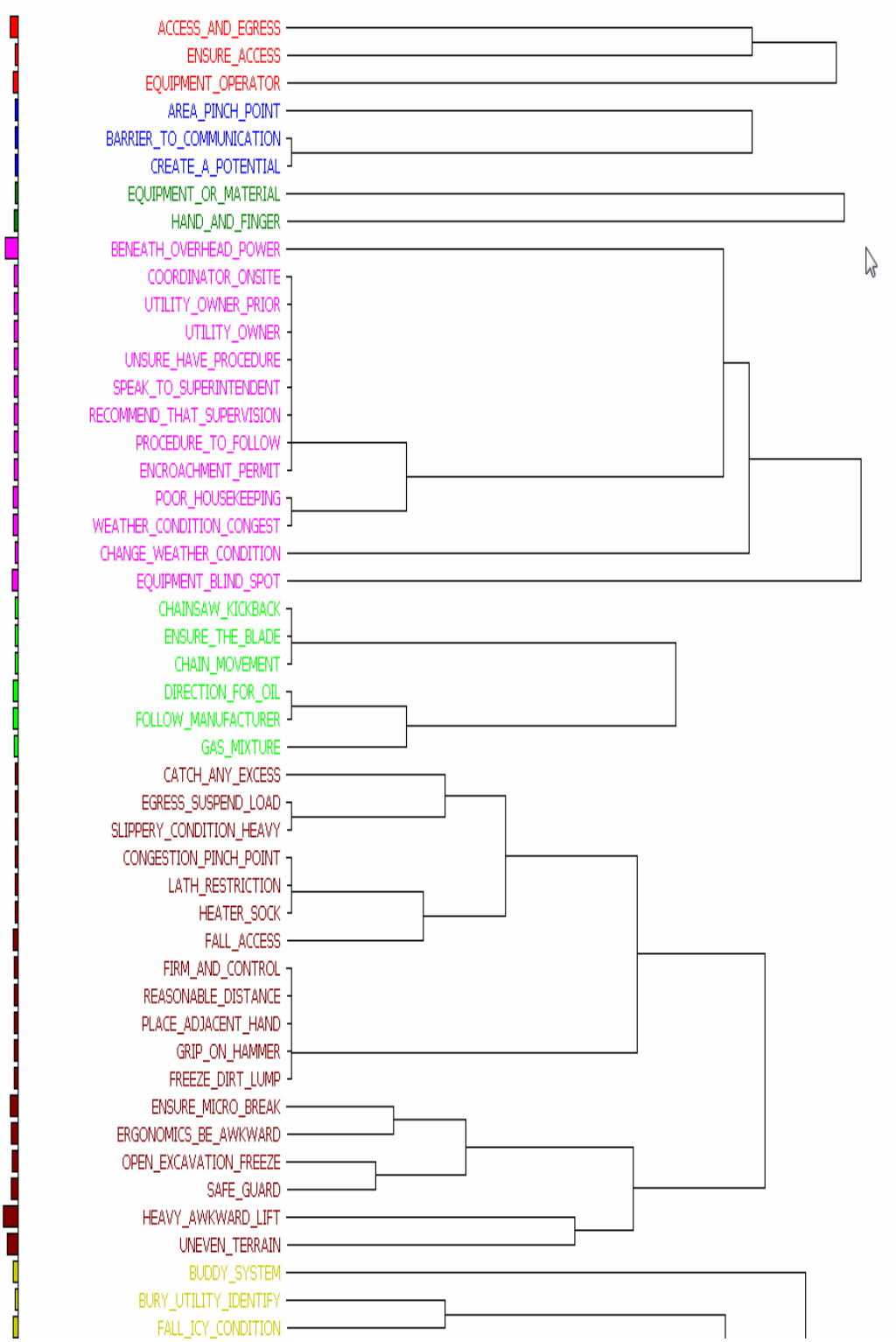


Figure 4-11 Collocation hierarchal clustering

A Jaccard coefficient is used as similarity measure for hierarchal clustering. A Jaccard coefficient is an efficient similarity measure for keyword and documents clustering (Niwattanakul, et al., 2013). Jaccard equation used to calculate the similarity between text terms (X and Y) is as follows (Provalis Research, 2015):

$$Jaccard\ cofficient = \frac{A}{A + B + C} \quad (4\text{-}1)$$

Where:

A = the numen of documents where two items (X and Y) co-occur together.

B = the number of documents where term X occur alone.

C = the number of documents where term Y occur alone

A Jaccard coefficient value is bounded by zero and one. A value close to one indicates that the terms are occurred together frequently in the same JHA document class and a value close to zero means the terms occurred together less frequently. Table 4-5 presents an example of a similarity matrix for words extracted from JHA documents.

Extracting similarity relationships enable knowledge analyst to explore relationship patterns between text terms. For example, word “ditch” has s similarity relation with many words such as “terrain,” “edge,” “utilities” and “exposed”. High ranked similar words associated with “ditch” are shown in Figure 4-12.

Table 4-5 Similarity matrix

	CAPACITY	CASE	CAUSING	CAUTION	CENTER	CERTIFICATIO	CERTIFIED	CHAIN	CHAINSAW	CHANGE	CHECK	CHECKLIST	CHECKS	CLAMPS	CLEAN	CLEANED	CLIMBING	CLOSE	COATING	COLLISION
CAPACITY																				
CASE	0.1																			
CAUSING	0.09	0.08																		
CAUTION	0.12	0.16	0.02																	
CENTER	0.36	0.08	0.06	0.18																
CERTIFICATION	0.05	0.14	0.03	0.05	0.05															
CERTIFIED	0.04	0.05	0.06	0.13	0.15	0.19														
CHAIN	0.06	0.12	0	0.01	0	0.14	0.03													
CHAINSAW	0.04	0.18	0	0.01	0	0.1	0.02	0.71												
CHANGE	0	0.09	0.16	0.08	0.1	0.12	0.07	0	0											
CHECK	0.04	0.14	0.06	0.29	0.15	0.19	0.24	0.03	0.04	0.14										
CHECKLIST	0.1	0.07	0.04	0.06	0.07	0.13	0	0.1	0.09	0.04	0.11									
CHECKS	0.01	0.11	0.03	0.19	0.07	0.18	0.11	0.03	0.03	0.1	0.28	0.03								
CLAMPS	0.17	0.1	0.07	0.21	0.12	0.1	0.05	0.1	0.08	0.07	0.2	0.07	0.27							
CLEAN	0.19	0.29	0.06	0.42	0.28	0.1	0.18	0.01	0.04	0.13	0.39	0.07	0.24	0.24						
CLEANED	0.14	0.21	0.04	0.34	0.2	0.24	0.28	0.09	0.07	0.12	0.41	0.08	0.2	0.21	0.47					
CLIMBING	0.04	0.27	0	0.25	0.04	0.16	0.02	0.13	0.17	0.07	0.26	0	0.25	0.25	0.21	0.09				
CLOSE	0.16	0.15	0.02	0.06	0.06	0.03	0.03	0	0.05	0.06	0.24	0.18	0.08	0.01	0.12	0.09	0.07			
COATING	0.02	0.04	0	0.11	0.02	0.03	0	0	0	0	0.04	0.04	0.14	0.04	0.07	0.04	0.02	0.05		
COLLISION	0	0.07	0.09	0.21	0.14	0	0.13	0	0.02	0.05	0.2	0	0.07	0.13	0.13	0.05	0.28	0.01	0	

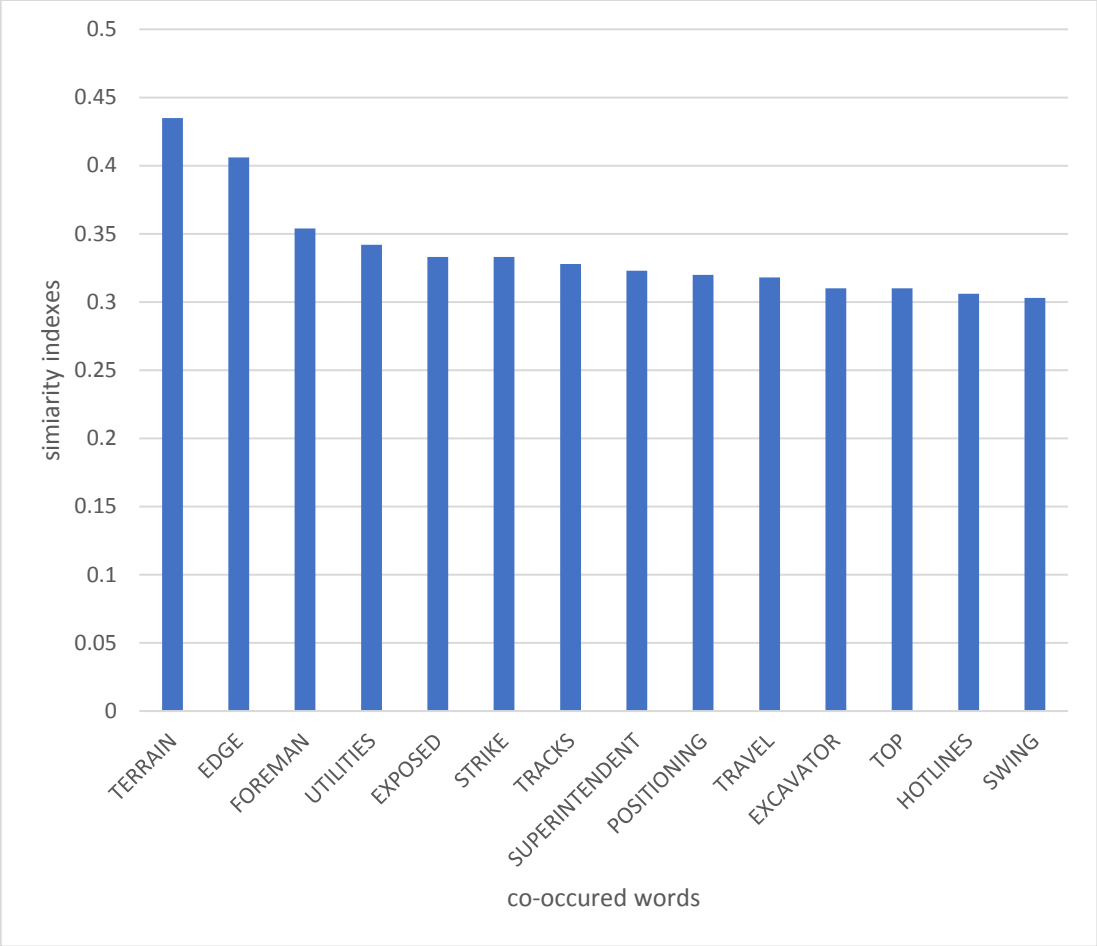


Figure 4-12 “Ditch” word and high ranked co-occurred words

A similarity matrix can be visualized using a network graph to perform links analysis among words and collocation. For example, collocation “ground conditions” and its associates are presented in Figure 4-13. Circles in the graph represent collocations such as “ground personnel,” ” loose material,” and “suspend load”. The lines between the circles are the relationships and the strength of relationships are indicated by similarity indexes over the lines.

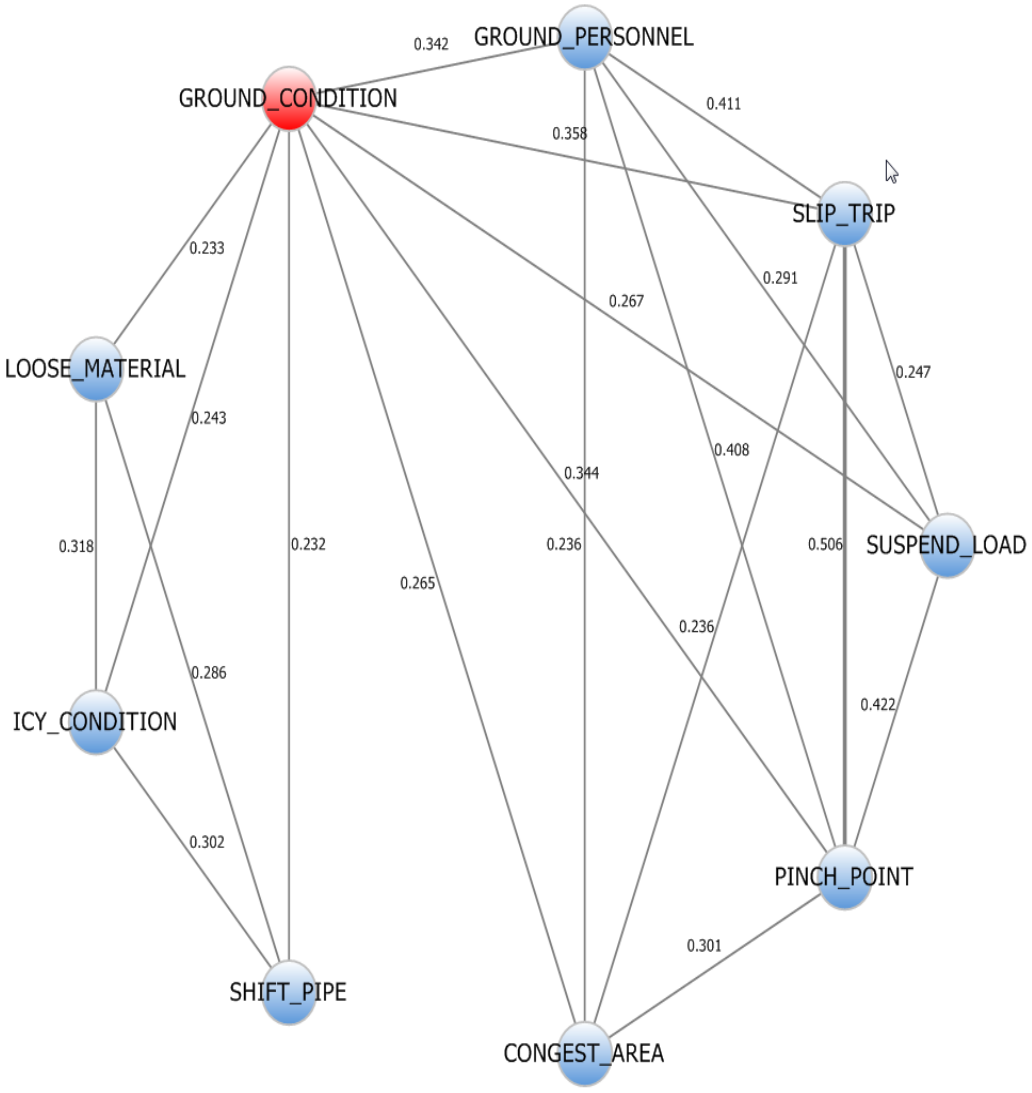


Figure 4-13 Links diagram for the collocation “Ground condition”

4.4 Construction hazard concepts analysis

Scanning the documents shows that hazards associated with each activity was in one of three categories. These categories belong to the stages of execution of construction activity on site (see Figure 4-14). The first category is the planning and setup stage. Each construction activity should go through planning and setup stage where all equipment, material and human power are prepared prior to starting core execution.

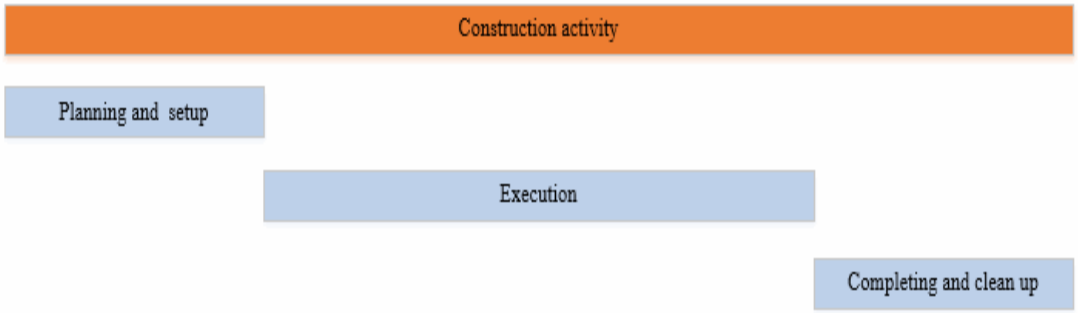


Figure 4-14 Hazards categories

The second category is the execution stage where the core execution of the construction activity is commenced. In this stage, construction professionals are implementing the safety plans that was planned in the first stage. Hazards are developed due to the interaction between several components of the construction execution process such as equipment, tools, humans, the environment, weather, and ground conditions.

The third category is completing and housekeeping. It is the process of organizing the site after the full or partial completion of the construction activity execution. It is very important because it is also preparing for the next activity or next the day's site work. It contributes to decreasing the potential of several hazards at the site such as slip and trip, disorganization, and losing tools.

4.4.1 Setup and planning stage.

The setup task prior construction activities is mandatory for safely executing the construction activity.

Figure 4-15 shows examples of hazards associated with setup stage. the setup stage mainly consists of checking required documentation, equipment, and the competency of construction workers.

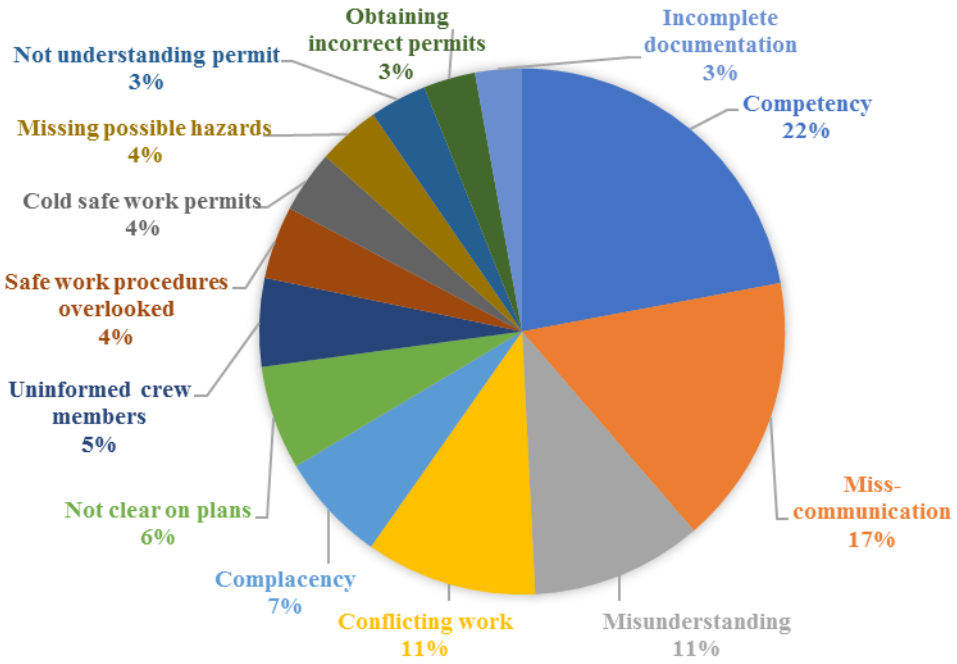


Figure 4-15 Hazards in planning and setup stage

The competency of construction workers and equipment operators is a frequently identified potential hazards in the JHA forms. Competency is crucial for some activities that require sufficient knowledge and experience such as entering a confined space. Competency is critical for some construction activity that required high technical skills to perform the job such as welding activity. Equipment operators competency is critical since pipeline projects involved using many types of construction equipment.

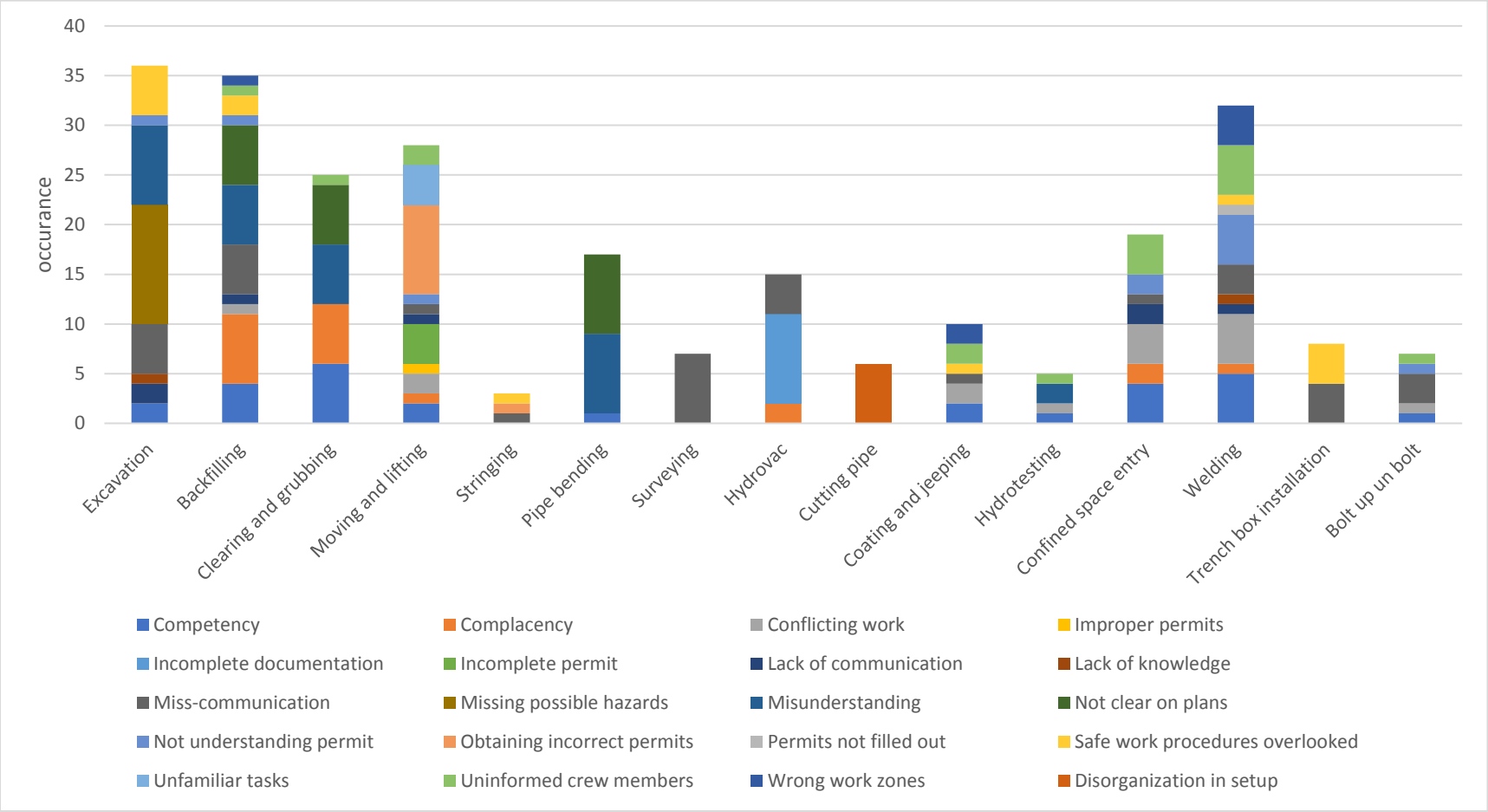


Figure 4-16 Setup hazards per activity class

Figure 4-1 shows extracted setup hazards tabulated with the construction activity classes, using document occurrences measure. Setup and planning hazards in JHA documents have different occurrences based on activity type and the number of JHA forms in each class. However, hazards are ranked by their occurrence in classes of construction activity. For example, the "miscommunication" hazard is identified in 11 classes. The "competency" hazard is identified in 10 construction activity classes as shown in Figure 4-17.

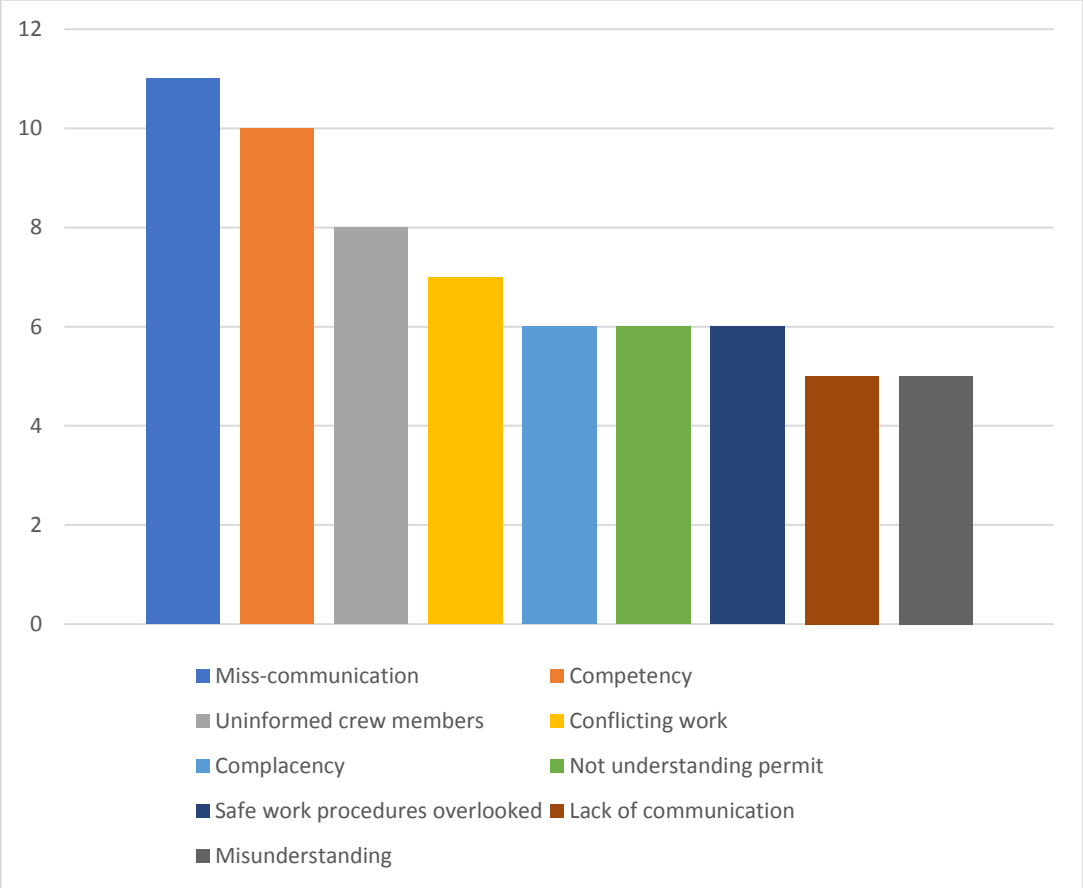


Figure 4-17 High-ranked hazards identified in planning and setup stage

Preparing and finalizing documents that is related to construction permits and safety checklist forms, is a required and important setup task. This task is contributing to safety communication to make sure all parties have a mutual understanding of safety conations on site. In addition, obtaining permits and other documentation will confirm that adequate and safe construction procedures are followed.

In pipeline construction projects, equipment plays a key role in construction execution. Checking equipment before using it in construction is a very impotent proactive action that can prevent potential damage or injuries while operating the equipment. A walk-around, warm-up and integrity check prior to operating the equipment are key tasks in the setup stage. Equipment inspection requires climbing on the equipment to check the parts. Falling from equipment is potential hazard that can cause injuries to workers and operators. In addition, Equipment damage or failure is a potential risk associated with the setup and planning stage.

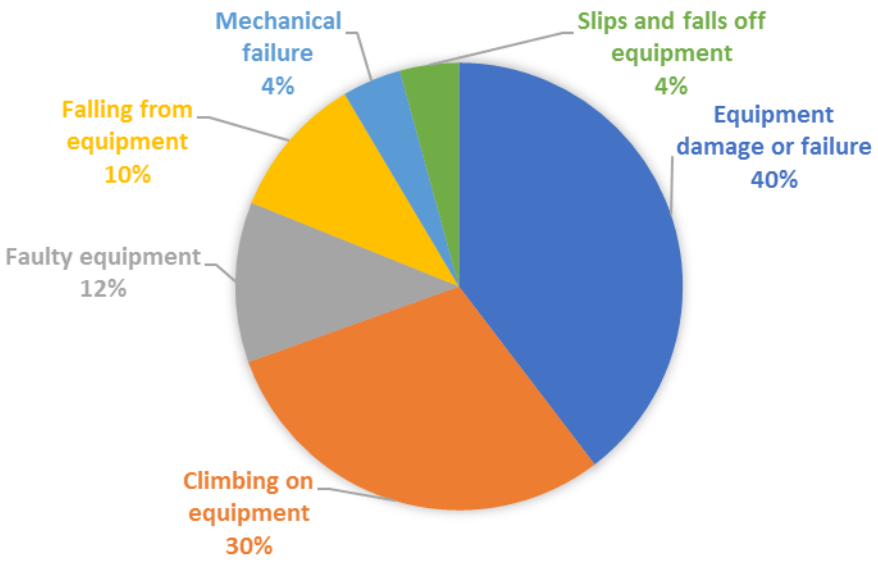


Figure 4-18 Walk around and integrity check hazards

4.4.2 Execution stage

During execution, all construction key components interact with each others. In pipeline construction projects, heavy equipment, ground conditions, weather conditions, heavy construction material (pipe), and construction personnel are all in moving state. The combination of these components is risky and can cause dangerous incidents. Hazards belonging to the execution stage were extracted for each pipeline construction activity. Figure 4-19 and Figure 4-20 show examples of hazards extracted for welding and hydro-testing activities.

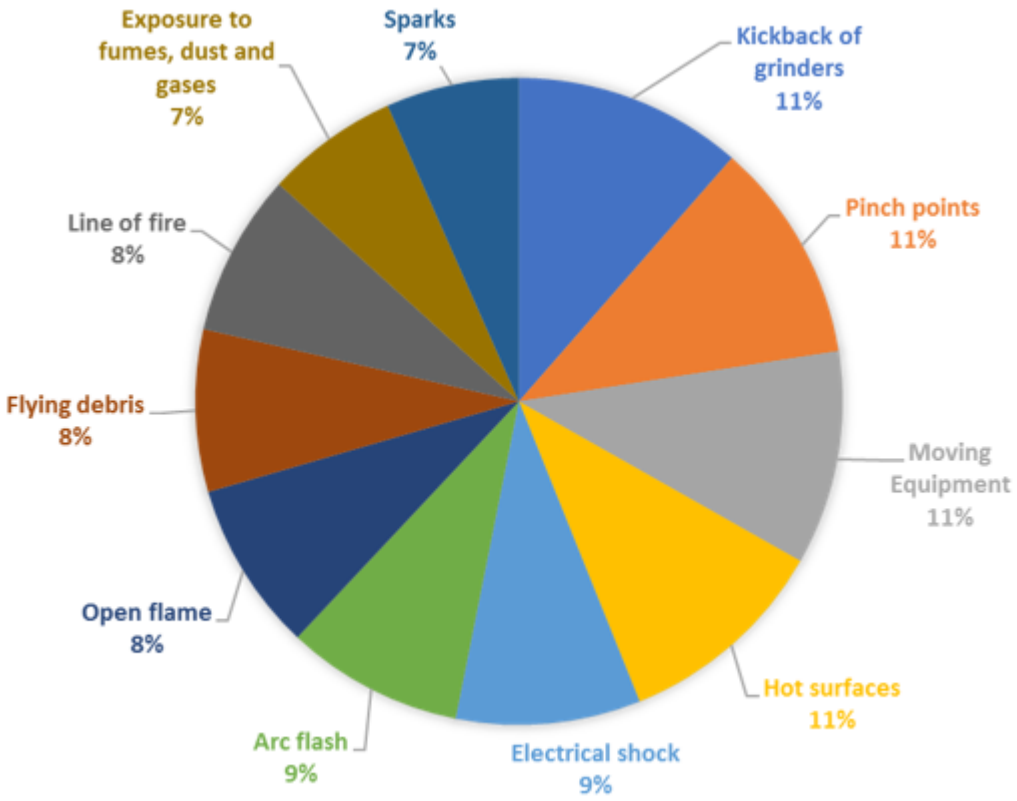


Figure 4-19 High ranked hazards associated with welding activity

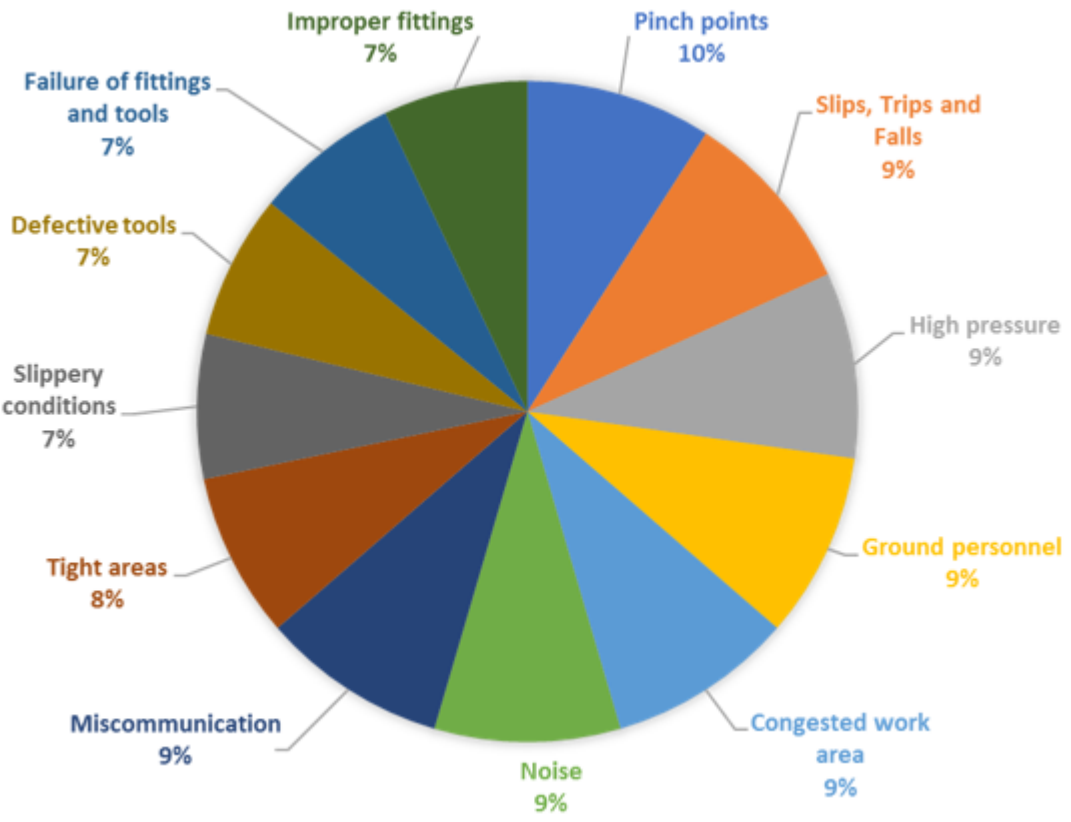


Figure 4-20 High ranked hazards associated with hydro-testing activity

A pipeline construction project is characterized by its execution complexity. Construction activities in pipeline construction share several characteristics related to the execution process. For example, excavation, backfilling, and stringing share using heavy equipment. Operating heavy equipment is associated with many hazards due to factors related to ground conditions, weather conditions, and ground personnel.

High-ranked hazards are tabulated with construction activities in Figure 4-21. The pinch point is a very common hazard in all construction activity classes. The pinch point hazard is mainly triggered due to the interaction between construction workers and construction equipment, material, or tools. The pinch point is a risky hazard

because of its potential consequences, such as damaging or causing the loss of human body parts. “Slip, trip, and fall” is a common hazard due to difficult ground conditions, which are one of the main characteristics associated with pipeline construction projects. Ground personnel are identified as potential hazards that can lead to incidents. Utilizing heavy equipment in activities like excavation, moving and lifting, pipe stringing, and backfilling can put any construction personnel working on the ground in a risky situation.

A congested work area is also a frequent hazard identified because all pipeline construction work is concentrated in or around the trench of the pipeline. This hazard is a coordination problem, and the possible control to mitigate the hazard is by removing unnecessary personnel, equipment or material from the work area. In addition, using a spotter at the site will help in coordinating and managing the equipment movement. Construction Hazards can be viewed from different perspectives and can be related to new specific concepts. Hazards is normally associated with construction activities classes. Also, hazards can be linked to the sub concepts of construction execution such as equipment, material, and surrounding environment. Sub concepts of hazards related to ground condition, equipment, weather and pipe material were extracted. Ground condition is one of the key issues that influence work sites. The heavy equipment is highly affected by ground conditions. However, this hazard is high frequent due to the nature of pipeline projects. Pipeline projects are linear projects built in remote rather than urbanized locations. pipeline construction progress is continuously moving from location to a new location that have different ground condition.

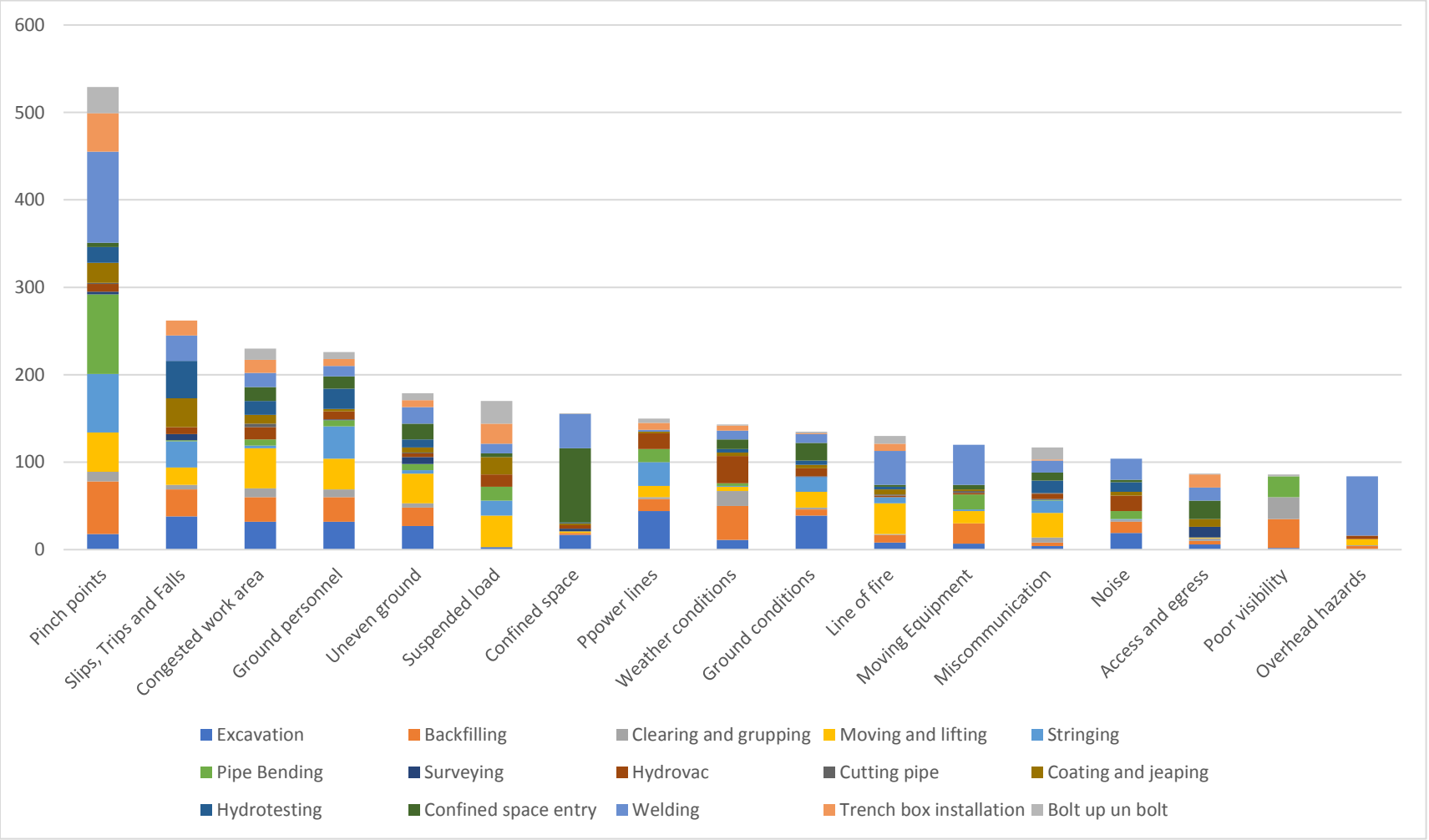


Figure 4-21 High ranked execution hazards per construction activity

In the JHA forms, hazards related to ground condition is expressed in several forms such as unstable ground, uneven ground and steep ground. Uneven ground, ground condition and icy or muddy condition are the highest ranked hazard terms extracted from JHA forms and related to ground component (see Figure 4-22). Construction activities (Excavation, moving and lifting, backfilling, and stringing) are the most activities associated with these hazards.

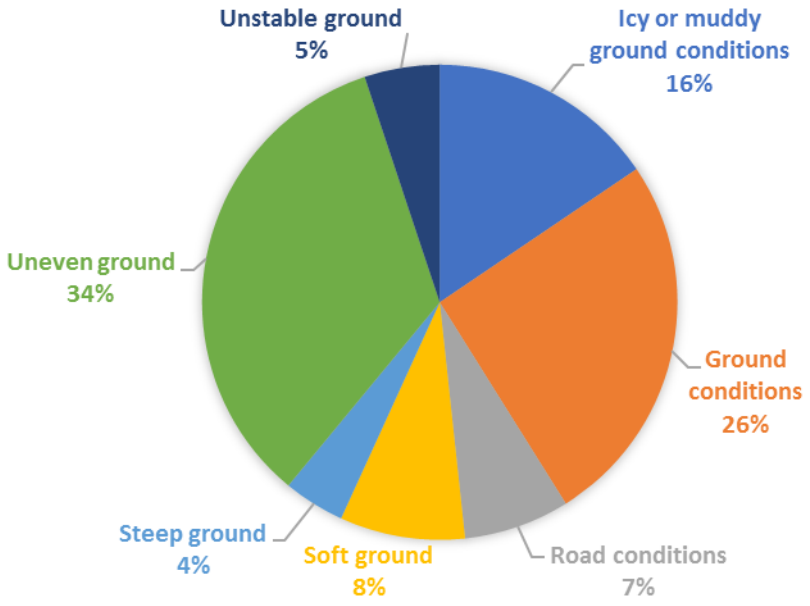


Figure 4-22 Ground related hazards

Heavy equipment in pipeline construction projects is the main execution component. Construction activities such as excavation, backfilling, pipe stringing, lifting and moving, and pipe bending, all require heavy equipment for execution. Hazards related to equipment are diverse and contribute to critical incidents at construction sites. Equipment in moving status is a source of hazards at construction sites. Hazards such as “moving equipment,” “collisions,” and “equipment tipping” are examples of

hazards that can lead to incidents on site. “Moving equipment” is the most common hazard identified in JHA forms and related to equipment sub concept (see Figure 4-23). Moving equipment is a hazard that is strongly associated with activities such as excavation and backfilling.

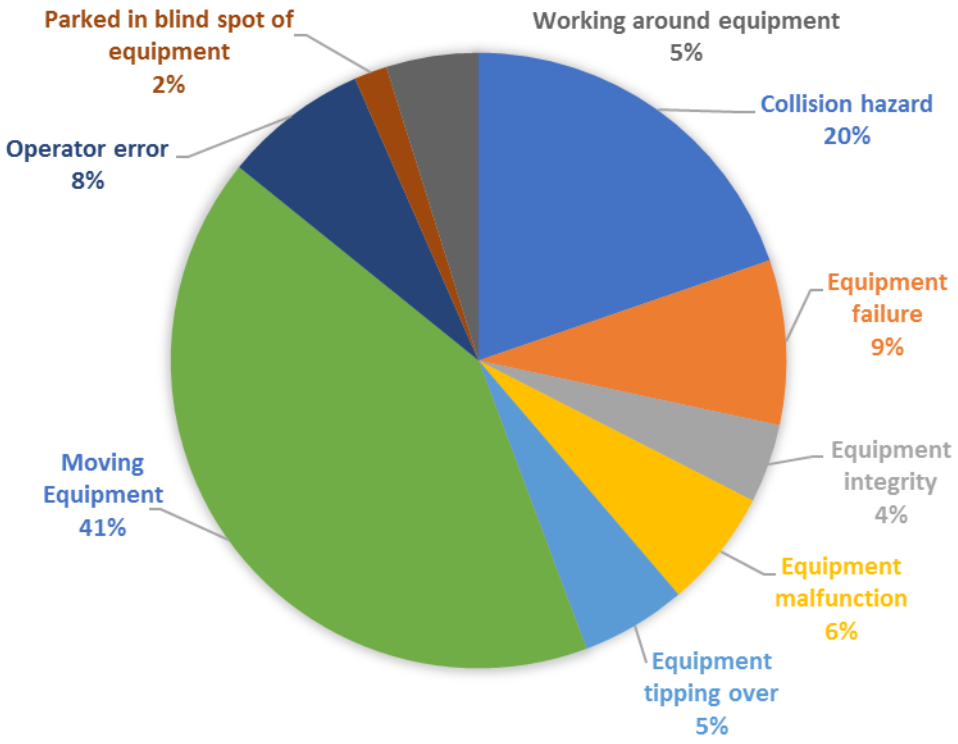


Figure 4-23 Equipment related hazards

Handling pipelines material during construction is linked to many hazards. Hazards related to “pipes failing,” rolling,” “shifting” and “moving pipe” are very common during pipe-handling operations. Other hazards are related to loose material falling on exposed live pipes during excavation. Also, line-striking during excavation is frequently identified in JHA forms (see Figure 4-24).

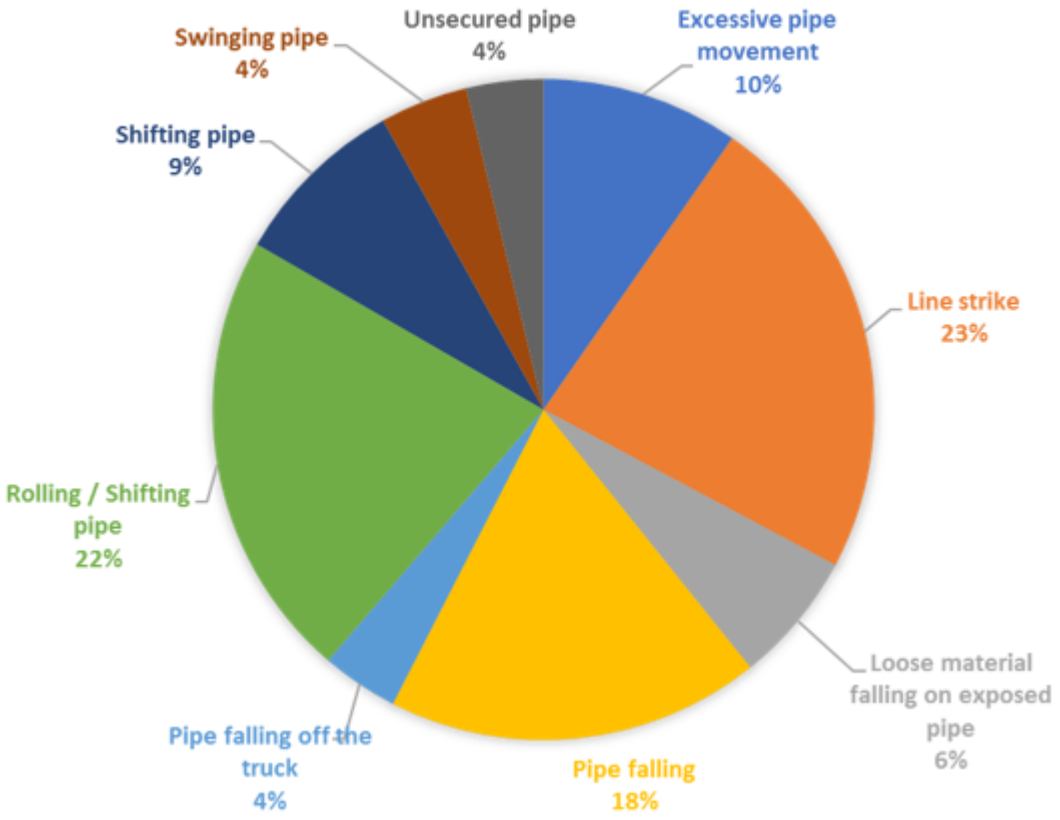


Figure 4-24 Pipe related hazards

Weather conditions are one of the main hazards that have a direct impact on construction workers, especially in Alberta in Canada. Pipeline construction projects are very sensitive to weather condition. Pipeline construction activities are executed in uncontrolled, remote, and open locations. Weather conditions, specifically wintry weather, have a significant impact on construction progress and safety. Heavy equipment can be affected during the snow time in Alberta. Ground condition-related risks are escalated due to snow accumulation, icy and muddy conditions. Most of the hazards related to weather were indicated in general term “weather conditions” to

describe hazards related to weather. Other terms were used to describe more specific conditions such as freeze up and wind direction (see Figure 4-25)

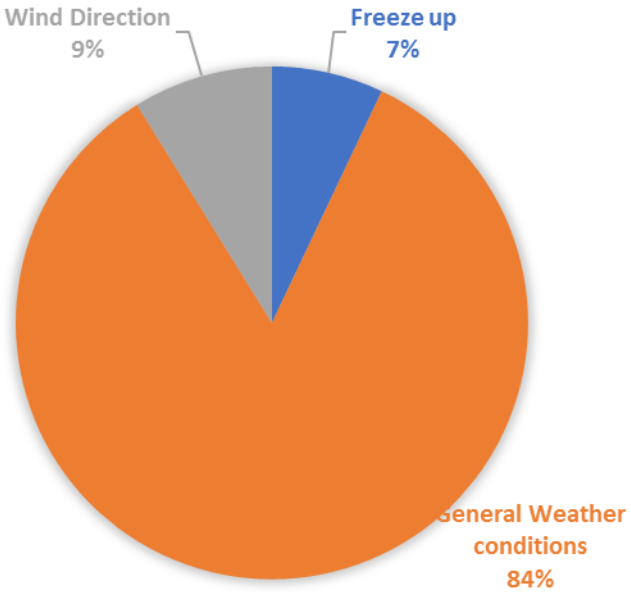


Figure 4-25 Weather related hazards

4.4.3 Construction activity completing stage:

After finishing a construction activity, cleaning up and organizing the site is crucial for site safety. Hazards associated with this stage are related to all site cleaning and organizing either on a daily basis or when construction activities are finished. Hazards determined in this stage are connected to properly disposing of construction waste. Construction waste has a highly negative impact on wildlife and the surrounding environment (see Figure 4-26). A pipeline project progresses from one segment to another segment in a repetitive manner over distances that may cover hundreds of kilometers. Hence, Continuous construction site cleaning is very crucial to avoid causing pollution and affecting wildlife.

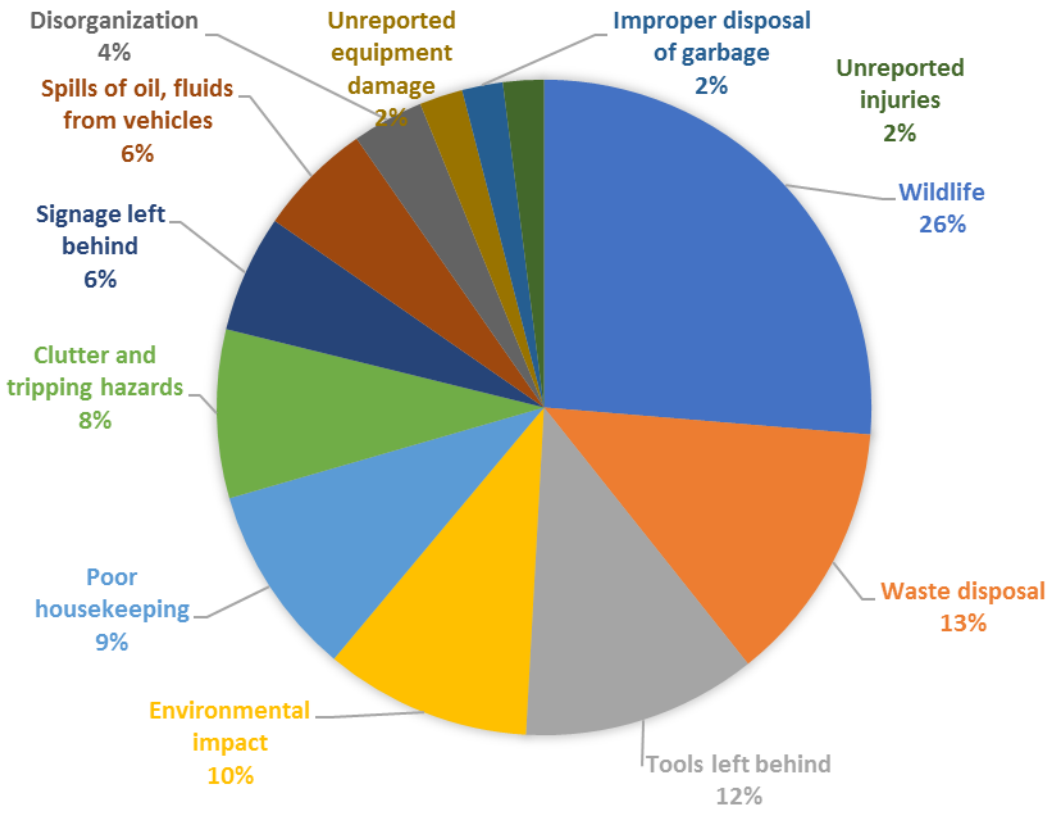


Figure 4-26 Completing and cleaning up hazards

Another completing hazard is related to leaving a construction site without reporting injuries occurred during the construction time. Another example is not reporting equipment damage or leaving tools behind at work site without proper storage. Figure 4-27 shows different hazards related completing stage and their occurrence per class.

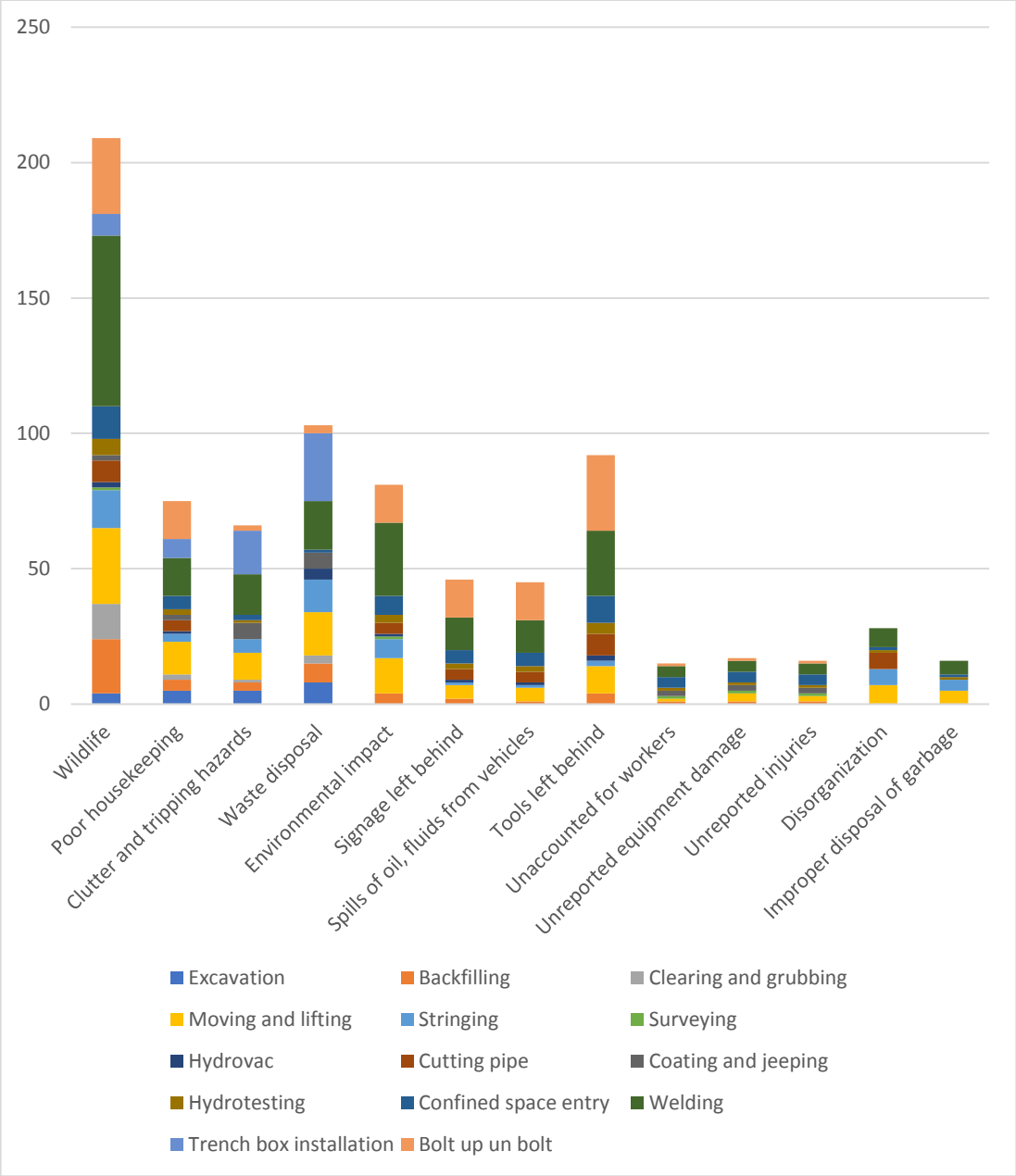


Figure 4-27 Completing hazards per activity classes

4.4.4 Hazards relationships development:

After extracting all hazards associated with different constitution activities, a hazard co-occurrence analysis is conducted to extract the semantic relationship between hazard concepts. Hazard co-occurrence analysis is the process of associating each

hazard with its co-occurred hazards in the JHA forms. Linking hazards with co-occurred hazards can benefit the hazard identification process. This means that identifying one hazard for construction activity could lead to identifying several hazards that had co-occurred in the previous JHA document classes. To identify these links, the method used was the same explained in Section 4.3.3. Example of hazard concept similarities are shown in Table 4-6 and full table of similarity matrix for excution hazards are shown in the appendices. For example, the hazard "arc flash" in welding activities co-occurred with several hazards such as "open flame," "kickback of grinders," and "electrical shock" (see Figure 4-28). Table 4-6

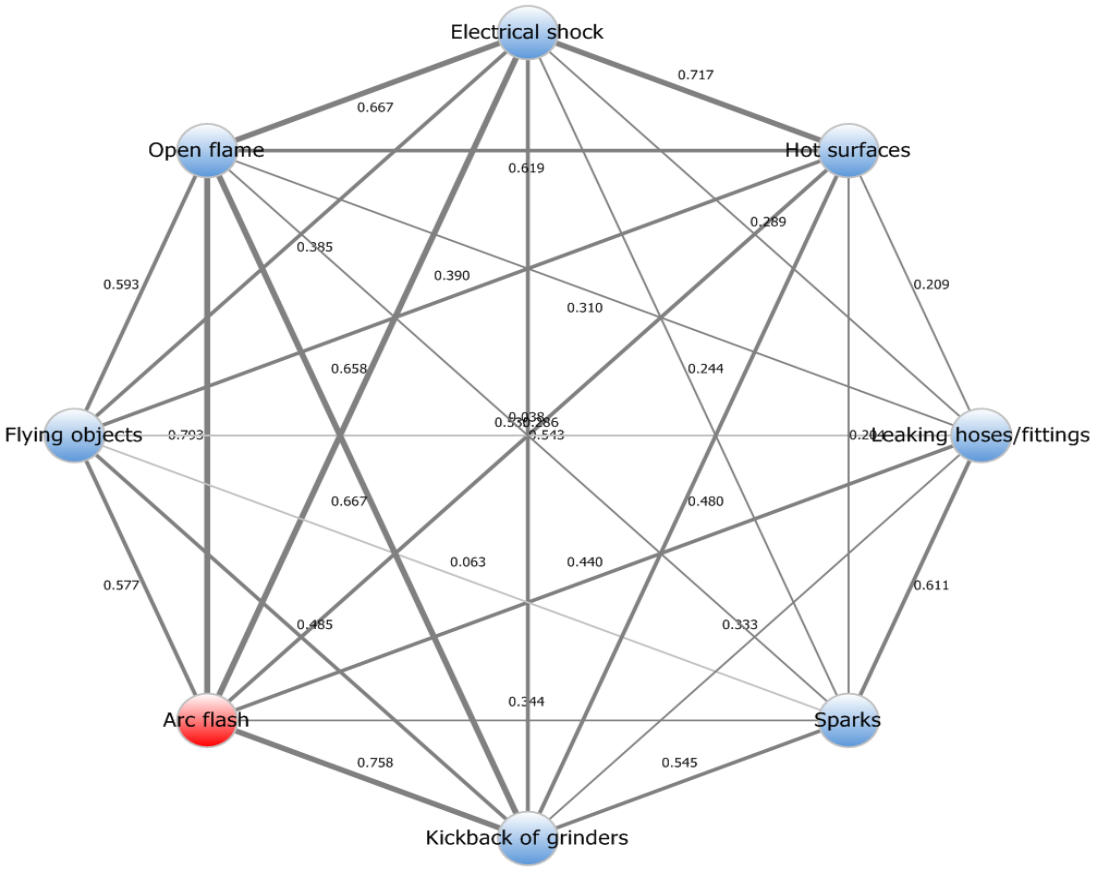


Figure 4-28 "Arc flash" hazard and co-occurred hazards

Table 4-6 Similarity matrix developed for hazards concepts

	Access and egress	Air quality	Arc flash	Awkward body positioning	Blind spots	Boredom	Cave-in hazard	Chain movement	Chainsaw kickback	Chemical exposure	Collision hazard	Communication failure	Compressed gasses	Confined space	Congested work area	Contact with buried utilities
Access and egress	1.00															
Air quality	0.08	1.00														
Arc flash	0.02	0.00	1.00													
Awkward body positioning	0.05	0.03	0.29	1.00												
Blind spots	0.08	0.00	0.03	0.00	1.00											
Boredom	0.00	0.24	0.00	0.00	0.00	1.00										
Cave-in hazard	0.14	0.00	0.00	0.03	0.12	0.19	1.00									
Chain movement	0.02	0.00	0.00	0.00	0.06	0.00	0.00	1.00								
Chainsaw kickback	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.50	1.00							
Chemical exposure	0.02	0.00	0.00	0.00	0.06	0.08	0.09	0.00	0.00	1.00						
Collision hazard	0.00	0.00	0.00	0.00	0.03	0.45	0.17	0.00	0.00	0.14	1.00					
Communication failure	0.10	0.00	0.00	0.00	0.06	0.00	0.00	0.11	0.07	0.09	0.03	1.00				
Compressed gasses	0.04	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.07	0.00	0.00	1.00			
Confined space	0.36	0.11	0.02	0.09	0.00	0.00	0.04	0.00	0.00	0.12	0.06	0.03	0.00	1.00		
Congested work area	0.17	0.06	0.06	0.09	0.06	0.13	0.10	0.01	0.01	0.02	0.15	0.01	0.01	0.10	1.00	
Contact with buried utilities	0.05	0.00	0.02	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00

4.5 Knowledge schema

Looking back to the starting point of this research, we had a large number of JHA forms from a previous pipeline construction project. The forms were scattered, and unused despite the value of knowledge inside them. The forms are the output of significant efforts of construction professionals in previous projects. Two main stages of knowledge mining were introduced in this research to reuse the knowledge embedded in the JHA forms.

In the first stage, documents were grouped in classes using clustering and classification models. Organizing JHA documents in construction activity classes is important for further content analysis. In the second stage, knowledge mining was done by integrating text mining and the qualitative approach. Text mining helps to speed up the process of knowledge concept extraction. In addition, text mining aids in discovering hidden relationships between text terms. However, some knowledge taxonomies cannot be obtained by merely applying text mining. For example, the classification of hazards by activity stage is based on studying and surveying hazard concepts qualitatively.

Hazard knowledge schema is the final output of processing previous JHA forms using different levels of extraction. The knowledge schema consists of knowledge concepts and their hierarchal and horizontal relationships. Hazard knowledge schema presents hazard concepts associated with each construction activity. Moreover, hazard taxonomies based on construction activity execution stages are also presented. Each construction hazard has horizontal relationships (co-occurrences) with other hazards'

concepts (see Figure 4-29). Knowledge schema is the base element for digitizing hazard knowledge using information technology tools to enable automatic retrieval and communication of knowledge.

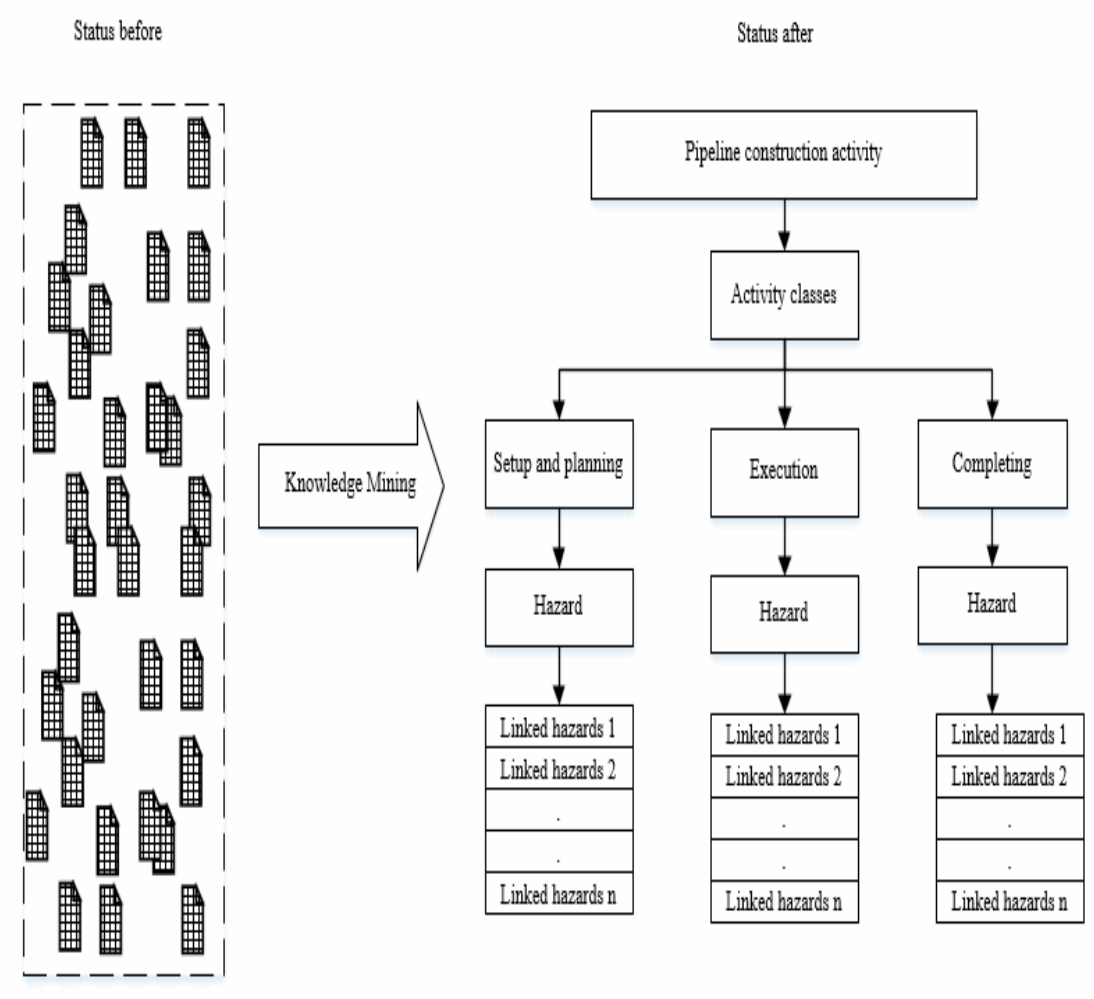


Figure 4-29 Hazards knowledge schema

4.6 Conclusion

JHA documents were analyzed to extract hazard concepts and taxonomies. The text mining and qualitative approaches were used together to build a hazard dictionary. Hazard concepts for each construction activity were extracted and explored. Hazards associated with specific pipeline construction components such as heavy equipment,

pipe material and ground condition, were extracted and analyzed. Text hierarchal clustering and co-occurrence analysis was used to explore underlying patterns in the text. A similarity matrix for hazard concepts was derived to represent relationships between concepts. Hazard knowledge schema was produced based on the analysis and reflects hazard concepts, taxonomies and relationships between the concepts.

CHAPTER 5

Knowledge modeling using ontology

5.1 Background

Knowledge modeling aims to represent real existing domain knowledge to enable knowledge storing, communication and sharing. Enabling effective knowledge input will improve the output performance for construction management processes such as planning, and execution, monitoring and control and completing.

Ontology is a formal information model that explains and describes knowledge of a specific domain that can be communicated and shared by people and computer applications (Batresa, et al., 2014). Ontology consists of hierarchies (taxonomies) of concepts, concepts relationships and the axioms that describe their behaviour (El-Diraby, 2013). It also includes the machine-interpretable meaning of concepts and relationships between them, (Noy, et al., 2001). Ontology can be built to fulfill one or more of the following requirements:

- Share structured information that has a common understanding between people and program software.
- Enable reuse and transfer of specific domain knowledge
- Perform analysis of domain knowledge.

Ontology has an important advantage over other data modeling due to its knowledge domain philosophy that enables the use of information by multiple program agents. In addition, ontologies support a knowledge reuse strategy and allow the extensibility and evolution of a knowledge domain (Verstichel, et al., 2011). Ontology is crucial for the semantic web and recognized as the backbone of the semantic Web. Semantic is a form of communication that communicates sufficient and effective meaning to produce actions. Semantic technology represents the new trend of transforming knowledge and information from being machine-readable to machine-understandable, (Taye, 2010).

Process automation and advancement in information communications have attracted researchers to ontology-based information modeling. Most construction management and technical processes need knowledge as an input. Currently, Construction knowledge is scattered and not formalized, organized and stored properly for future reuse.

An ontology was built to represent highway construction domains to enable transferring and sharing knowledge among all project parties, (El-Diraby & Kashif, 2005). Another knowledge-based system was built using an ontology to assess risk factors that may lead to cost overruns in international projects, (Yildiz, et al., 2014). A social network system using ontology modeling was introduced to enable the sharing of safety information among construction teams, (Le, et al., 2014). Ontology is to support the construction quality inspection process by modeling regulation constraints associated with different construction tasks, (Zhong, et al., 2012).

JHA knowledge stored on previous JHA documents was represented by ontology to support hazard identification process (Wang & Boukamp, 2011). Chi, et al., (2014) created an ontology model that include safety documents resources such as safety regulation and standards. Moreover, ontology modeling was used to model safety knowledge and enable integration with building information modeling (BIM) to automate safety planning using BIM environment (Zhang, et al., 2015).

However, the hierarchical structure used to build previous job-hazard knowledge models consisted of activity, action steps, hazards, and controls. It was based on the explicit structure founded in the JHA documents and not built on a deeper knowledge domain analysis. Knowledge domain analysis is crucial to extract knowledge schema that is required for knowledge modeling (Gasevic, et al., 2009). Extracting embedded knowledge concepts and semantic relationships between hazard entities, can leverage the structure of the knowledge domain and improve the performance of the ontology model. Semantic relationships between concepts and entities is essential for effective retrieval process of hazards and their control measures.

5.2 Ontology modeling

5.2.1 What is ontology

The word ontology comes from the Greek word, “ontos,” which means “being.” It is defined as the study of categories of things that belong to a domain, (Sowa, 2000). Ontology has many definitions in the computing field and in the area of artificial intelligence. One of the widely used definitions is that “ontology is a specification of a conceptualization.” (Gruber, 1993).

A conceptualization is an abstract and simplified view of a specific domain. The specification is a formal and declarative representation of knowledge concepts. Since ontology represents knowledge that will be used by a program agent, concepts, classes, constraints, and relationships must be written using formal language that is machine-readable.

5.2.2 Why ontology?

Ontologies provide a beneficial and valuable feature for intelligence systems and knowledge representation. These features relate to domain vocabularies, domain taxonomy, knowledge sharing, and reuse.

Vocabularies represent the names that describe the concepts in an ontology. Vocabularies in ontology are machine-readable and are different from other forms of vocabulary content such as catalogs, glossaries, and thesaurus. The catalog is a finite list of text terms and their clear interpretation and such terms refer exactly to their identifiers. In the glossary, each term refers to meaning described by natural language that may often be interpreted in another way by different people. A thesaurus offers more semantic relationships among meanings of terms, in the form of synonyms. However, a thesaurus does not explicitly show the hierarchies of the terms. In ontology, logical statement is used to describe the meaning of terms and how they are or are not related to each other. Thus, an ontology provides a common ground of understanding for both human and machines.

Taxonomy (also called concept hierarchy) is a hierarchal classification of entities or concepts in a domain of knowledge. A good taxonomy is one that has its concepts

presented in clear and easy structure to interpret and remember (Gasevic, et al., 2009). Taxonomy and vocabulary are presented using specifications written in formal language to produce a conceptual framework for knowledge analysis, sharing reuse, and retrieval

Knowledge sharing and reuse are among the most important objectives of building an ontology. Computer application plays a significant role in the current time in processing operations that are required in our daily life. Enabling the application to access and reuse knowledge represented in ontologies is a Major goal. to accomplish this goal, compatibility and mutual understanding of terminologies between agent application and knowledgebase systems are essential.

However, this is not the case in real life for several reasons. One reason is that there are different languages for representing ontologies. Another reason is that many technical groups and communities competing to produce different approaches and technologies in building ontologies. In addition, building multiple ontologies for the same domain will cause confusion and problems of reuse. Ontology is designed to evolve over the time and this will require maintenance and updating.

5.2.3 Ontology application areas

Ontology has different areas of application depending on why the ontology is being developed. Ontology is used for collaboration purposes in a specific domain of knowledge such as project management, health, and construction. Ontology provides a united knowledge structure, so it can be used for further knowledge development and sharing. Knowledge sharing and reuse can also be performed using an intelligent

application agent. The existence of a united knowledge schema enables an efficient knowledge exchange between application and the information model (ontology).

Ontology is useful for information integration from several different resources. Computer applications may require different information from several resources as a response to a query. If the application can access ontologies that have this information, integration will be easier and more automatic.

A united knowledge structure of ontology can help people who are seeking to educate themselves more about a domain. Moreover, the ontology provides the expert domain a medium through which they can share their understanding of domain concepts and schema. Ontology is also used for intelligent modeling as a pre-developed knowledge base model. In information modeling, ontologies are used as reusable building blocks in the system model.

5.2.4 Ontology component

Ontology is built on a classification of a hierarchal structure of a group of concepts. Ontology adds a richer network of semantic relationships such as functions, constraints, and inference rules and axioms (Jakus, et al., 2013). Ontology typically includes concepts, classes, objects, individuals, entities, attributes, and properties. **Error! Reference source not found.** Figure 5-1 illustrates an example of ontology structure for animal knowledge domain.

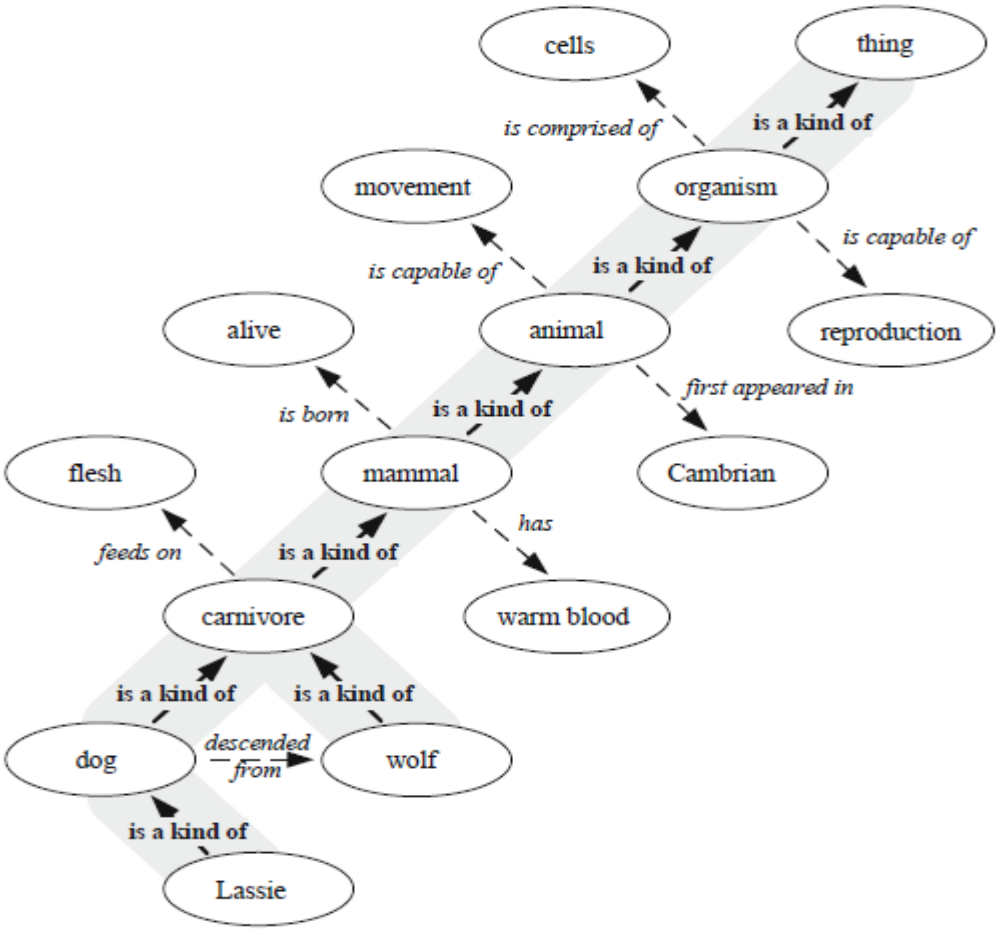


Figure 5-1 An example of ontology structure, (Jakus, et al., 2013)

5.2.5 The semantic web

The semantic Web is new generation of existing Web. It represents the efforts that contribute to enabling machine to read information from the Web. Semantic web is not new and independent from the current World Wide Web (WWW), but it is an extension to current one, (Berners-Lee, et al., 2001). Semantic technology is aiming to make The Web more understandable by machines (Heflin & Hendler, 2001). It is about building efficient infrastructure, so intelligent application agents can access the Web and performing complex operations required by the end users.

The Semantic Web aims to improve information integration by establishing a common ground of understanding of knowledge structure by both human and the computer, (Taye, 2010). The WWW is huge and consists of numerous information blocks that are not easy to integrate. Most of WWW information formats are represented in natural language. Computers cannot interpret the meaning of neutral languages and humans cannot process and analyze a large amount of information. Using machines to interpret information will enable processing and analyzing WWW content (Noy, et al., 2001). The Semantic Web comprises two major components: the common language format for integrating information from diverse sources and how to relates information to real world entities. Web ontology languages are referred to as Semantic Web Languages.

Semantic web structure is shown in Figure 5-2. It consists of several layers: URI is the uniform system identifier that is required for locating any resources on the WWW. Unicode is the universal standard computer representation for characters. Extensible markup language (XML), a resource description framework (RDF), and ontology vocabulary (OWL ontology language) will be explained in the next sections. Logic and proof blocks are about checking the logic and consistency of the structure of the ontology using reasoner. Trust is considered the final layer of the semantic Web and is related to the trustworthiness of knowledge on the WWW to assure the quality (Taye, 2010).

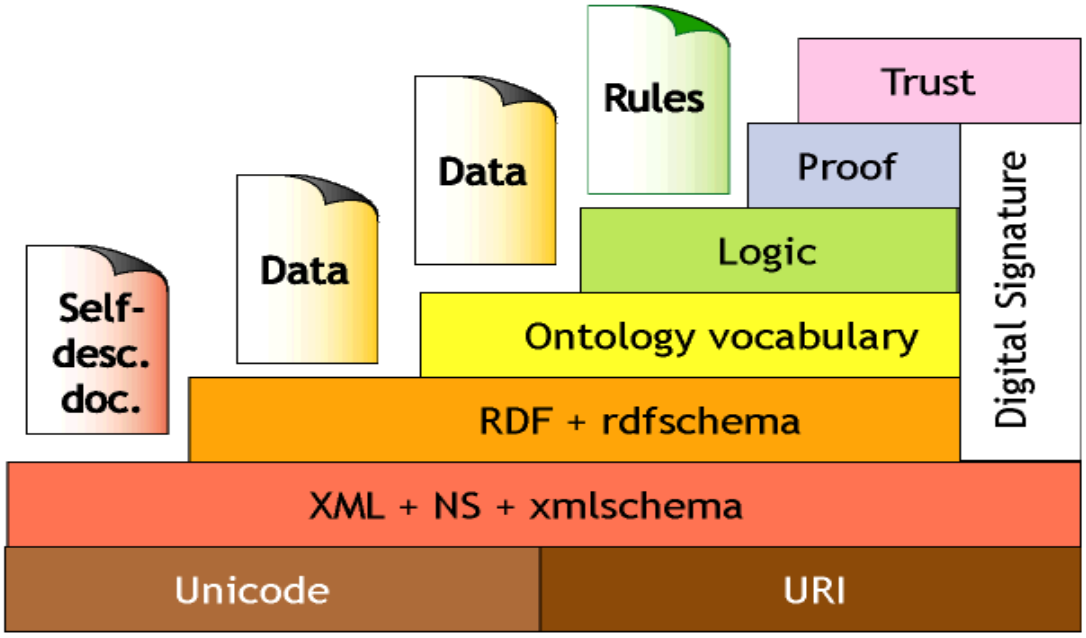


Figure 5-2 Semantic web structure, (Berners-Lee, 2000)

5.2.6 XML and XML schema

XML is a markup language used to produce machine-readable documents. It is used to mark up unstructured documents without using fixed tag vocabulary and to make the tag for a document content understandable by a human. Tags can be logically customized by people to make XML documents self-describing.

XML is a data format-based language. XML is giving structure for data for documents and access process for programs and application to perform required operation such as updating and structure documents. To use XML to exchange data among applications, it is necessary to have an agreed-upon dictionary of vocabulary. This can be achieved by using XML schema that specifies the method of structuring XML documents (Gasevic, et al., 2009) .

5.2.7 RDF and RDF schema

RDF and an RDF schema model are the tools to describe data in documents in a semantic way. This semantic relationship enables the machine to be able to interpret domain knowledge. RDF uses statements called triples to describe the relationship between data objects. Triples use a (subject-predicate-object) format where the subject and object can be any form of information object in the knowledge domain. The predicate is the relationship or properties that describe how the subject relates to the object. RDF statements provide the feature of representing triples in a graphical semantic network. RDFS (RDF Schema) introduces XML-based vocabulary to enable the creation of taxonomies and to create classes, their relationships and their properties (Brickley & Guha, 2014).

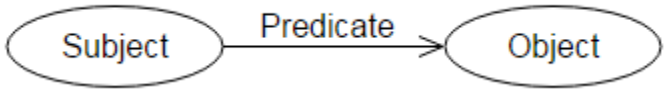


Figure 5-3 RDF graph concept representation (Cyganiak, et al., 2014)

5.2.8 OWL ontology language

OWL language greatly facilitates the machine interpretability of the information content of the WWW. OWL introduces additional vocabulary to the formal semantics, (McGuinness & Harmelen, 2004). OWL ontology language is on the top of RDF layer and aims to describe the terminology used in Web documents.

The main feature in OWL ontology language is an extremely rich vocabulary for describing links among classes, individuals and properties (Gasevic, et al., 2009). An

extension and revision for OWL ontology language (Now known as OWL 1) is OWL 2, (W3C OWL Working Group, 2012).

5.2.9 SPARQL query

SPARQL is a group of specifications that introduces languages and protocol for querying RDF content format from the Web or RDF store (W3C SPARQL Working Group, 2013). It is used to extract information from an RDF graph and subgraph. In addition, it is used build a new graph based on a new query structure based on clients' requests.

A SPARQL query can be presented in one of four forms: SELECT, CONSTRUCT, ASK, and DESCRIBE.

SELECT: Return all, or a subset of queried information, and variable values that match the query pattern.

CONSTRUCT: Return the RDF graph that is produced by the substitution of the variable into a triple template.

ASK: Return a Boolean to show if a query pattern matches or not.

DESCRIBE: Return RDF triples that contain resources which match the query pattern.

5.3 Safety ontology development methodology

Ontology development required sufficient understanding of the targeted knowledge domain. Construction Safety knowledge domain comprises many concepts which is related to different knowledge area in construction safety. Hazard knowledge

embedded in JHA forms is one the most critical knowledge because of its direct relation to construction activities. Developing an ontology comprises main phases (Gasevic, et al., 2009) as follows:

Specification: which is relating to identifying the objective and scope of targeted ontology.

Conceptualization: is related to structuring knowledge in an organized and semi formal way. The output of this phase is classes, concept, relations, properties, attributes and constraints.

Implementation: is related to using ontology platform to implement the ontology map that produced on conceptualization map.

The objective of safety ontology is to build hazard knowledge base for pipeline construction project to assist in performing future hazard analysis processes. Moreover, targeted ontology can be used as an education tool for learning about hazards associated with pipeline construction activities.

Pipeline construction project is the range of the safety ontology. The ontology consists of classes of construction activities, activities stages, and hazard instances associated with each activity. Ontology map was developed to reflect knowledge structure extracted from JHA forms (see Figure 5-4).

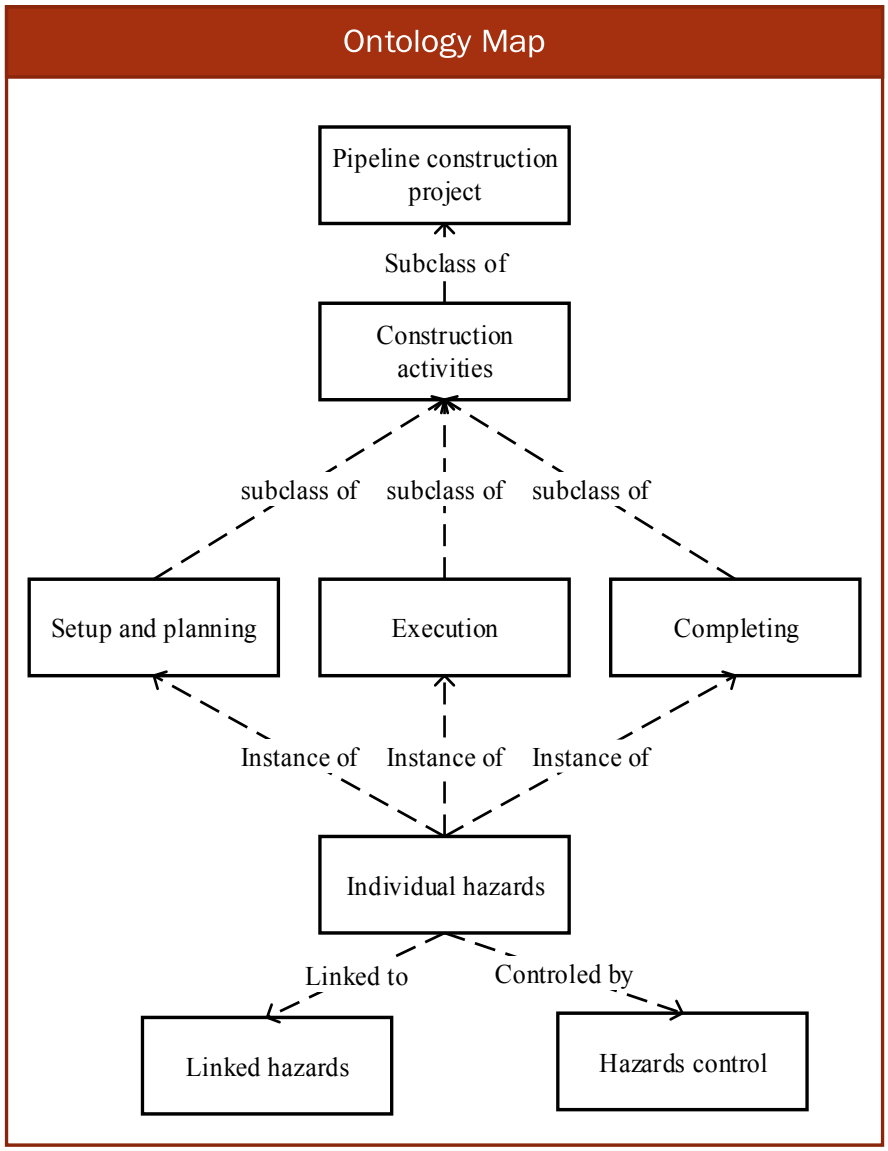


Figure 5-4 Hazards ontology map

5.4 Safety ontology implementation

To develop the ontology, an ontology editor, Protégé was used. Protégé is an open source platforms environment developed by the Stanford Center for Biomedical Informatics at the Stanford University (Musen, 2015). The platform enables an environment for building ontologies to represent knowledge base systems (see Figure 5-5).

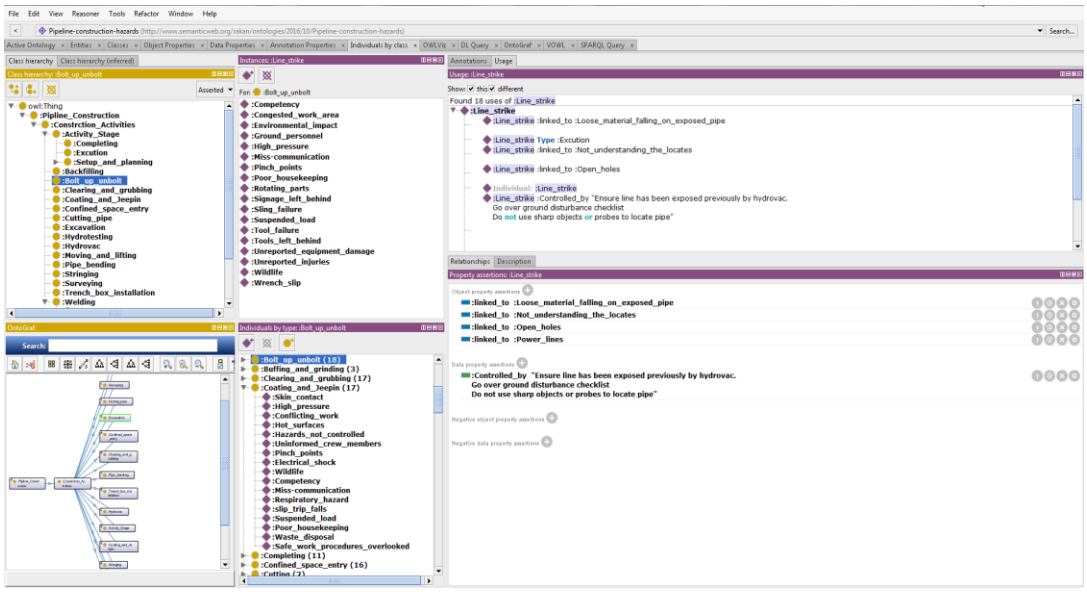


Figure 5-5 Protégé environment

5.4.1 Ontology main concepts

The main concepts of safety ontology are construction activities and construction stages of activity execution (see Figure 5-6). Pipeline Construction activities such as excavation, backfilling, stringing and stringing are all presented as subclasses of pipeline construction project, (see Figure 5-7).

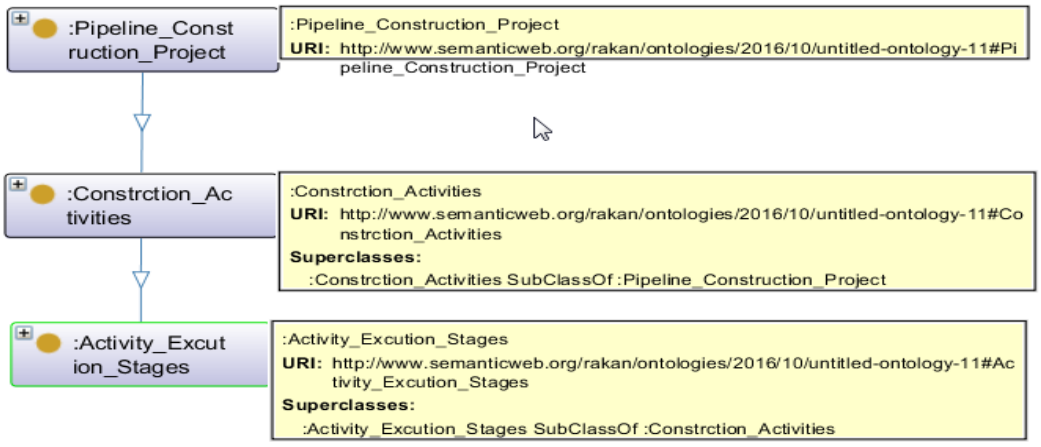


Figure 5-6 High-level ontology concepts

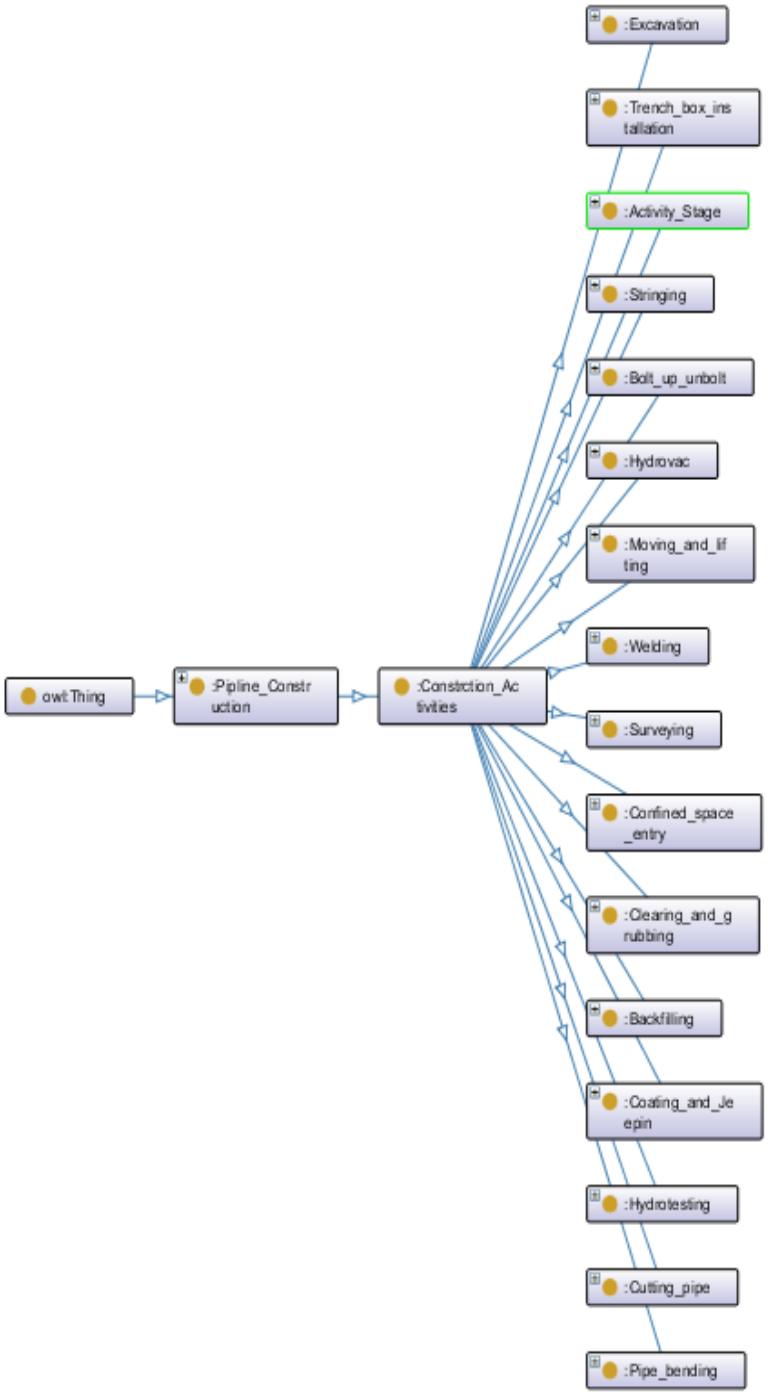


Figure 5-7 Construction activity subclasses

Some construction activities have been broken into activity steps to show more details about the execution process. This typically depends on the complexity of activity execution and the level of risk associated with the activity. Breaking activity into steps can enable the exploration of more embedded hazards and consequently the assigning of control measures. For example, pipe welding activity in JHA was broken down into steps including lining up the pipe, spacing, preheating, buffing and grinding, and welding (hot pass, fill and cap) (see Figure 5-8).

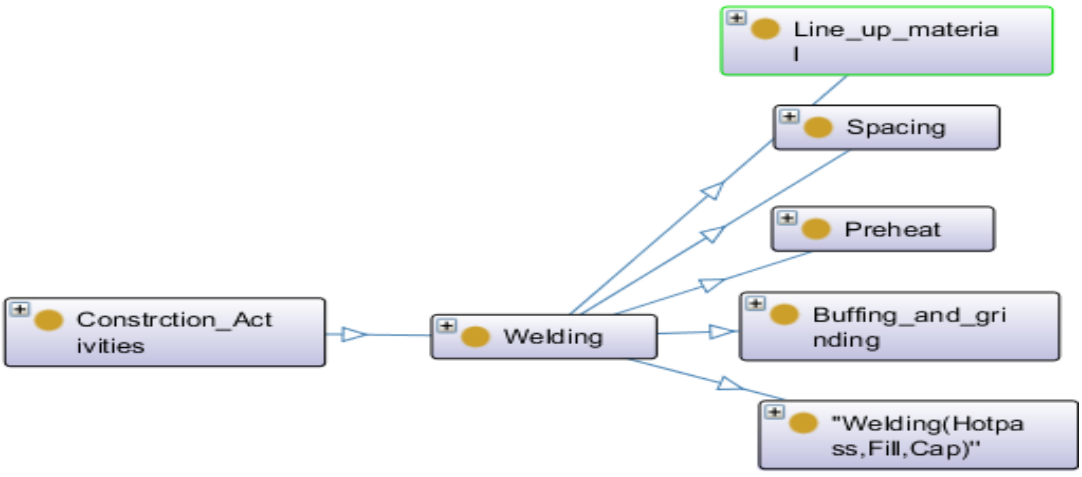


Figure 5-8 Welding activity steps

Construction activity execution passes through three stages which are setup, execution and completing. For example, hazard related to oil spills that have an impact on environment and wildlife is mostly identified to be investigating during completing stage of construction activity. Another example, the hazard of moving equipment on a construction site is identified in the execution stage of construction activities such as

backfilling and excavation. These stages were represented using a subclass of construction activities class (see Figure 5-9).

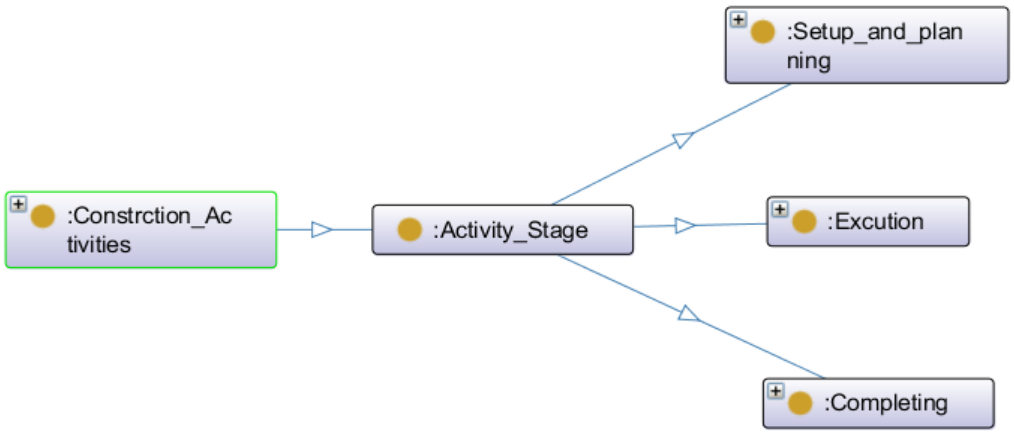


Figure 5-9 Construction activity stages

5.4.2 Ontology instances

Hazards are the instances of construction activity classes and activity stage classes. Hazards' instances are inserted for each construction activity class, an example of backfilling activity is shown in Figure 5-10. Construction activity classes could include shared hazard instances. For example, activities such as excavation, backfilling and stringing are sharing the same hazards instances as "uneven ground" and "equipment damage or failure".

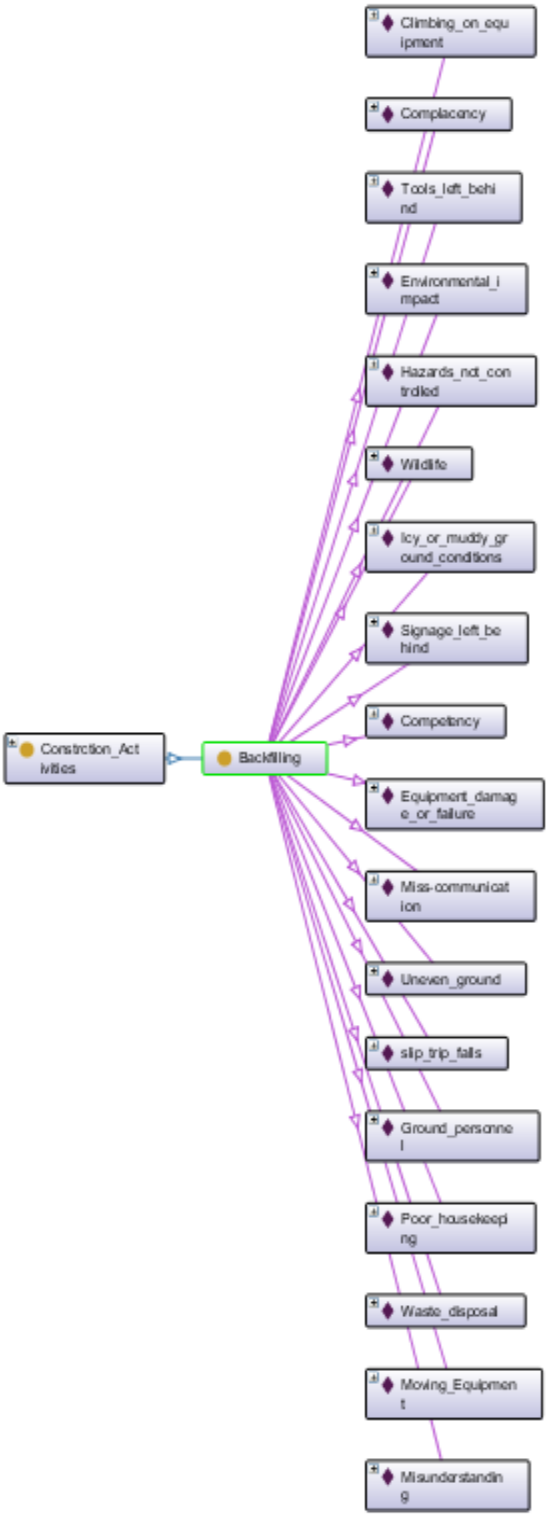


Figure 5-10 Backfilling hazard instances

5.4.3 Data and object properties

Each hazard has its own control measures that are required for eliminating or mitigating the hazard. A control measure is represented by the data property "controlled by." The data property is a linking of instances to an XML schema data type or RDF literal, both of which can be referred to generally as the data value. In Protégée, there are several built-in data properties such as integer, float, and string. The data property "controlled by" was assigned as a string. Examples of hazards controls are shown in Figure 5-11.

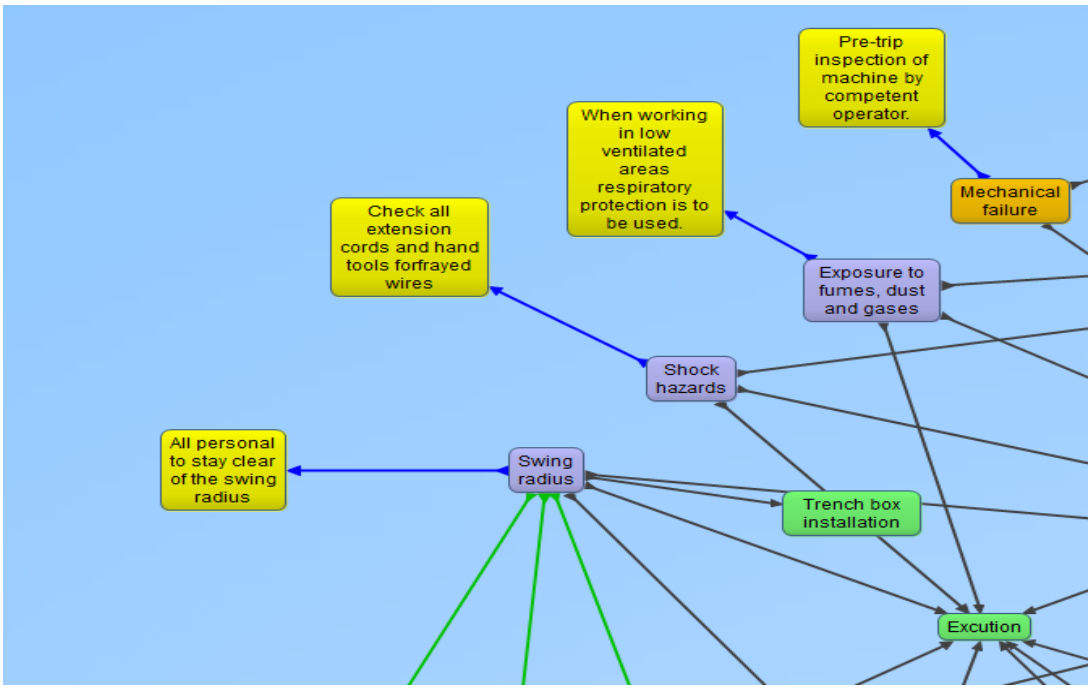


Figure 5-11 Data property "Controlled by" for hazards controls

Hazard instances are linked together to reflect the semantic relationship produced by measuring similarity occurrences. Hazard instances are linked using the OWL object property named "linked to." The object property's domain is construction activities class. Because the hazard similarity matrix is symmetric, the property is also

symmetric. This means that if Hazard A is linked to Hazard B, Hazard B is also linked to Hazard A (see Figure 5-12).

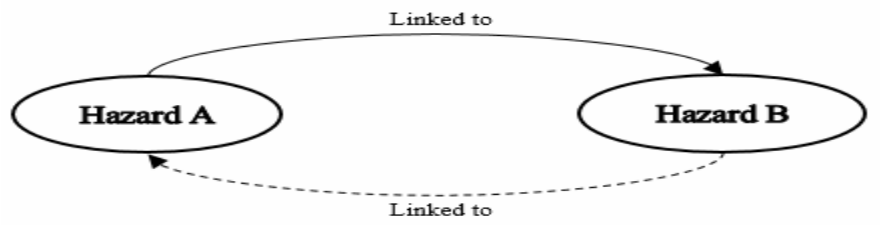


Figure 5-12 Symmetric object property

5.5 Ontology validation

Ontology modeling has three dimensions that describe ontology development and usage: the philosophical dimension, artificial intelligence (AI), and linguistic, (El-Diraby, 2014). The validation of present ontology will adopt the same validation philosophy introduced by (El-Diraby, 2014) as follows:

5.5.1 Philosophical dimensions

The philosophical dimension is related to how knowledge concepts of hazard ontology are constructed and based on what theory the concepts are related to each other. Research output ontology uses the output of intensive text mining applications to represent extracted knowledge schema from JHA documents that are real-world knowledge domain sources. The main structure of hazard knowledge (activity, activity step, hazard, hazard control) was adopted by the previous research to build the ontologies to support the JHA process, (Chi, et al., 2014; Zhang, et al., 2015). However, the proposed knowledge structure in this research introduces new sub-

concepts and new semantic relationships among entities based on an analysis of JHA documents. In addition, discussions were held with research partners during research developemnt regarding the knowledge concepts and their validity in the pipline projects domain.

5.5.2 Lexical /linguistics dimension

This dimension is about the validation of terms' sources used in building an ontology. A hazard words missing? built based on extracting terms from the JHA analysis collected from different pipeline construction projects. Terms were analyzed to identify frequencies and co-occurrences to extract semantic relationships in each class of construction activity (see Chapter 4). In other words, the ontology is built based on an analysis of lexical terms extracted from JHA documents and reflects the current linguistic pattern used by construction professionals. To more generalize the ontology, other sources of JHA documents are needed for further lexical extraction and representation.

The present hazard ontology represents the first step toward formalization of hazard taxonomies extracted from previous The developed ontology is and should not be expected to be final and fully comprehensive; rather, it requires continuous hazard updates.

5.5.3 AI dimension

The AI dimension of ontology is related to the formal part of ontology development. AI's role is to emulate human reasoning in knowledge cognition. This can be achieved

by coding knowledge using formal languages. Validating this part is considered straightforward and uses testing queries to recall specific ontology content.

1- Reasoning

Reasoning is the process of checking the quality and consistency of an ontology. Reasoning means using facts of explicit knowledge modeled in the knowledge base to drive a conclusion or make an inference (Jakus, et al., 2013). In the Semantic Web and ontology, reasoning is applied by using an inference engine.

An example of a reasoning tool that is used along with Protégé for designing safety ontology is the Hemi T reasoner. The reasoner algorithm is written using ontology language based on hypertableau calculus. The reasoner works automatically during the development process. In case any inconsistency happens during the building of the ontology, reasoner advice and locate the errors immediately.

2- Validating ontology using SPARQL queries

The competency question was developed to validate the AI part of the ontology. Competency questions are transformed to SPARQL queries to retrieve ontology content. Ontology content consists of RDF triples that have a subject-predicate-object format. A tool application, Gruff 6.4.3, developed by Feanzinc Inc (Franz Inc., 2017), is used to construct visual queries and query results for competency questions (see Table 5-1). For example, Competency question No.6 is about retrieving welding activity construction hazards and proposed control measures as shown Figure 5-13.

Table 5-1 Competency questions to validate hazard ontology

S. N	Competency Questions	Gruff Query Codes
1	Pipeline construction activities?	select ?node_variable_1 where { ?node_variable_1 rdfs:subClassOf <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Constrction_Activities> . } limit 32
2	Hazards associated with completing stage?	select ?node_variable_1 where { ?node_variable_1 rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Completing> . } limit 32
3	Hazards associated with excavation activities and their control?	select ?node_variable_1 ?node_variable_3 where { ?node_variable_1 rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Excavation> ; <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_3 . } limit 32
4	Hazards associated with pipeline stringing activity, including controls and linked hazards?	select ?node_variable_7 ?node_variable_8 ?node_variable_9 ?node_variable_10 ?node_variable_11 where { ?node_variable_10 <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_11 ; rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Stringing> . ?node_variable_7 <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_8 ; rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Stringing> ; <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#linked_to> ?node_variable_9 . }

5	Hazards associated with stringing activities in execution stage and their control?	select ?node_variable_1 ?node_variable_3 where { ?node_variable_1 rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Excution> , <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Stringing> ; <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_3 . } limit 32
6	Retrieve welding activity steps and associated hazards, their controls' measures?	select ?node_variable_1 ?node_variable_2 ?node_variable_3 ?node_variable_5 ?node_variable_6 where { ?node_variable_3 <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_6 ; rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Welding> . ?node_variable_1 rdfs:subClassOf <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Welding> . ?node_variable_2 <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_5 ; rdf:type ?node_variable_1 . }
7	From excavation activity retraveling strike hazard, its control measure and its linked hazard?	select ?node_variable_34 ?node_variable_36 ?node_variable_37 where { ?node_variable_34 rdf:type <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Excavation> . <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Line_strike> <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Controlled_by> ?node_variable_36 ; <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#linked_to> ?node_variable_37 . filter (?node_variable_34 = <http://www.semanticweb.org/rakan/ontologies/2016/10/untitled-ontology-11#Line_strike>) }

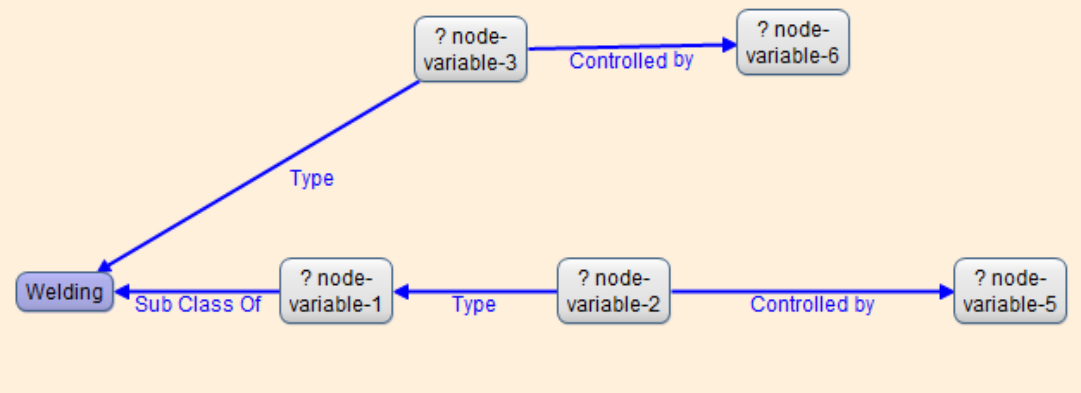
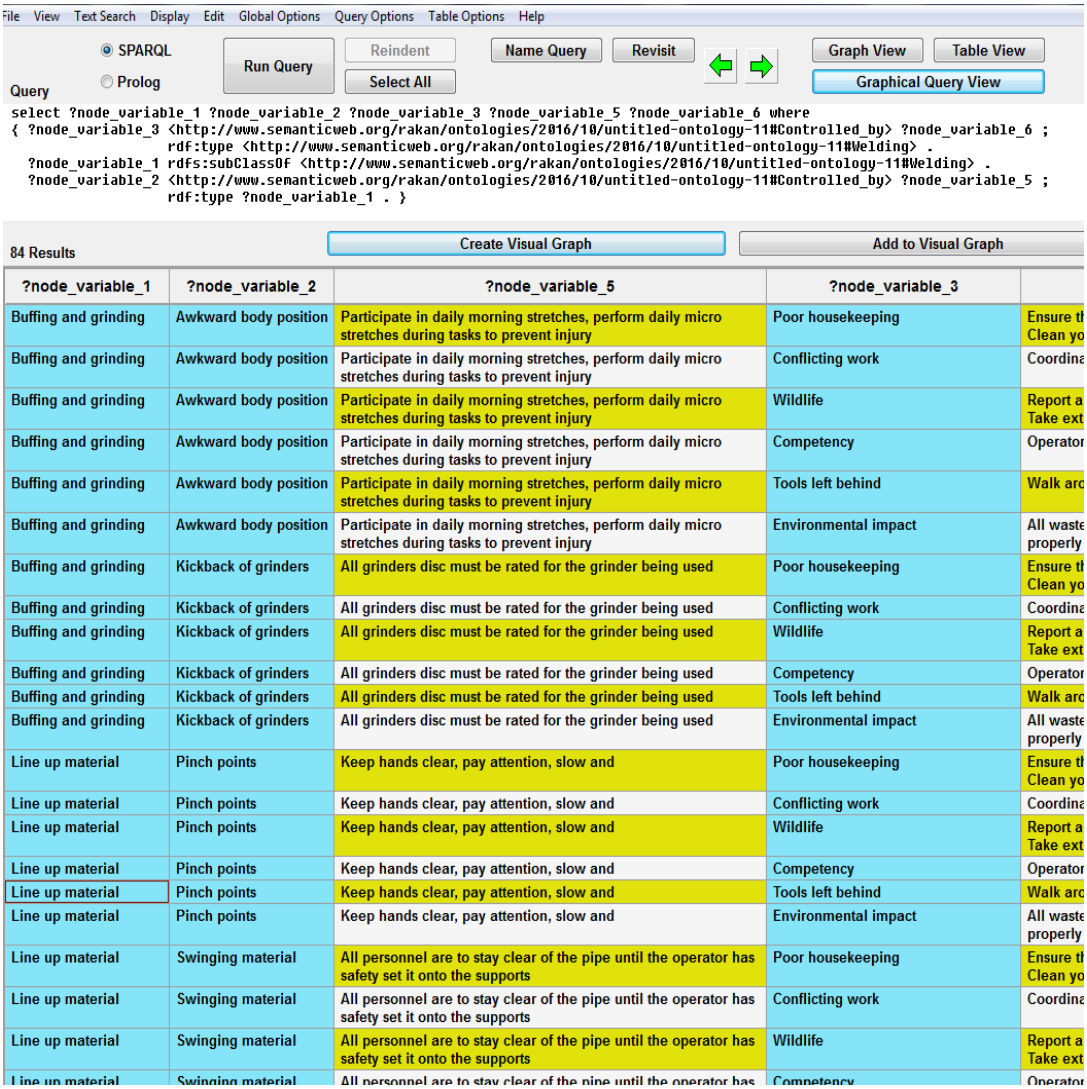


Figure 5-13 Competency question No. 6 visual query design



?node_variable_1	?node_variable_2	?node_variable_5	?node_variable_3	
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Poor housekeeping	Ensure th Clean yo
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Conflicting work	Coordina
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Wildlife	Report a Take ext
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Competency	Operator
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Tools left behind	Walk arc
Buffing and grinding	Awkward body position	Participate in daily morning stretches, perform daily micro stretches during tasks to prevent injury	Environmental impact	All waste properly
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Poor housekeeping	Ensure th Clean yo
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Conflicting work	Coordina
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Wildlife	Report a Take ext
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Competency	Operator
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Tools left behind	Walk arc
Buffing and grinding	Kickback of grinders	All grinders disc must be rated for the grinder being used	Environmental impact	All waste properly
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Poor housekeeping	Ensure th Clean yo
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Conflicting work	Coordina
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Wildlife	Report a Take ext
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Competency	Operator
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Tools left behind	Walk arc
Line up material	Pinch points	Keep hands clear, pay attention, slow and	Environmental impact	All waste properly
Line up material	Swinging material	All personnel are to stay clear of the pipe until the operator has safety set it onto the supports	Poor housekeeping	Ensure th Clean yo
Line up material	Swinging material	All personnel are to stay clear of the pipe until the operator has safety set it onto the supports	Conflicting work	Coordina
Line up material	Swinging material	All personnel are to stay clear of the pipe until the operator has safety set it onto the supports	Wildlife	Report a Take ext
Line up material	Swinging material	All personnel are to stay clear of the pipe until the operator has	Competency	Operator

Figure 5-14 Competency question No. 6 table results

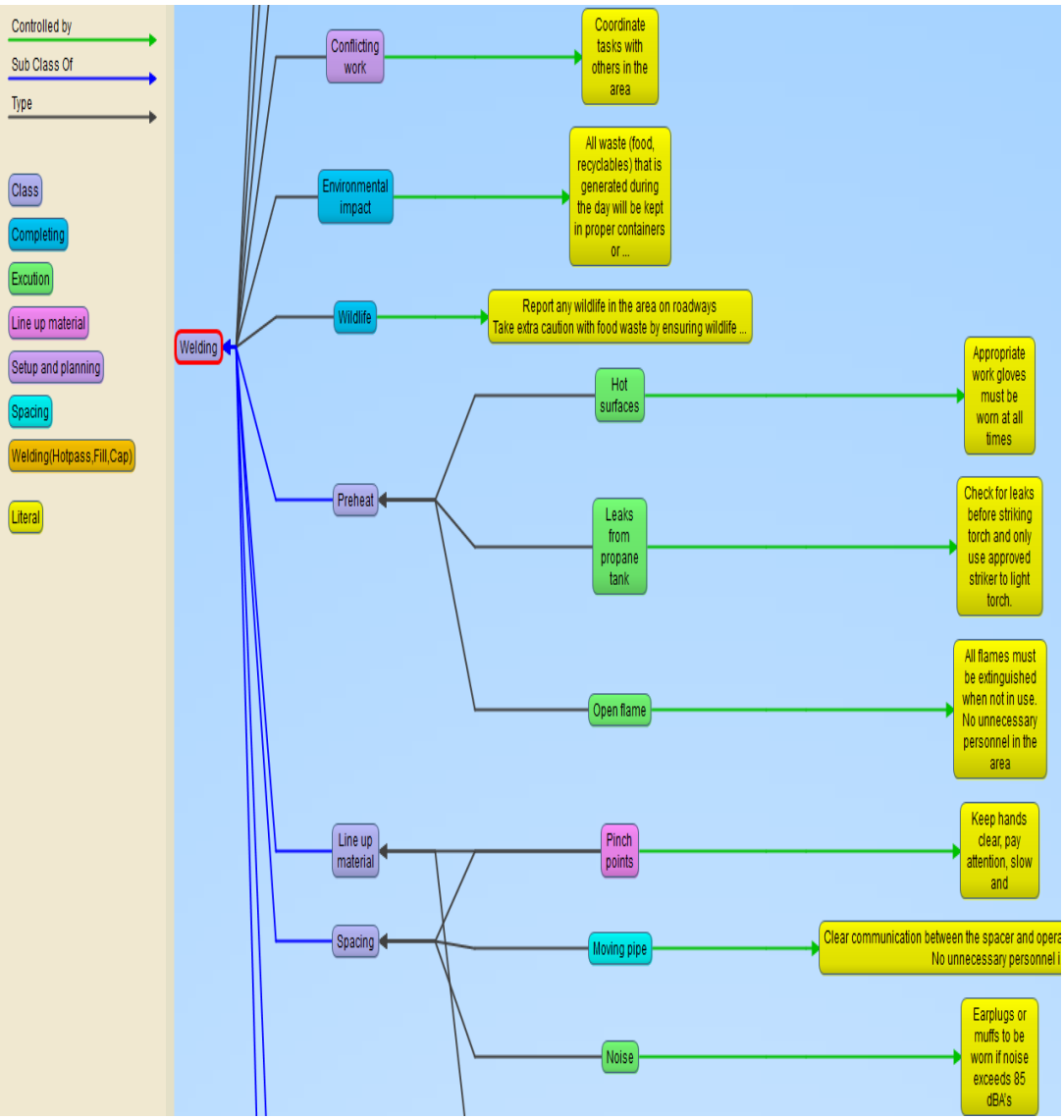


Figure 5-15 Competency question No. 6 visual graph results

Query results can be obtained using a table or graphical view as indicated in in Figure 5-14 and Figure 5-15. Other example of competency question is competency question No. 4 which is retrieving all hazards associated with pipeline stringing activity, their control measures and linked hazards (see Figure 5-16 and Figure 5-17). Competency question No. 7 is for retrieving linked hazards to a specific hazard "line strike" that belongs to the excavation activity class. The visual query graph, visual graph results and answer table are shown in Figure 5-18, Figure 5-19 and Table 5-2.

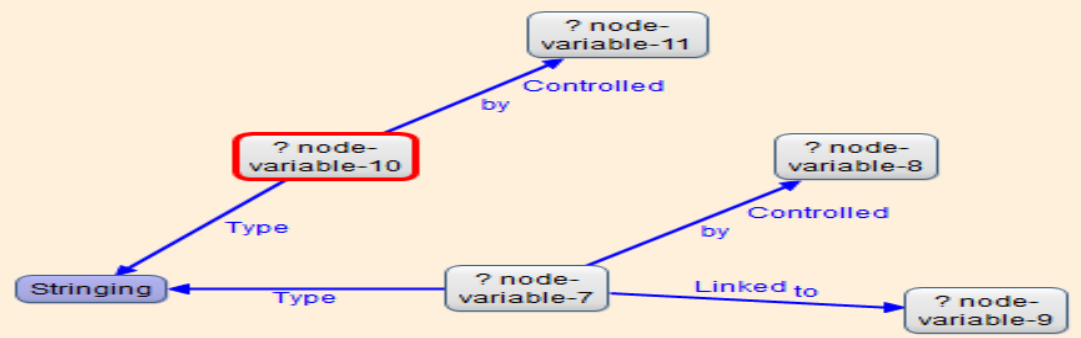


Figure 5-16 Competency question No. 4 visual query design

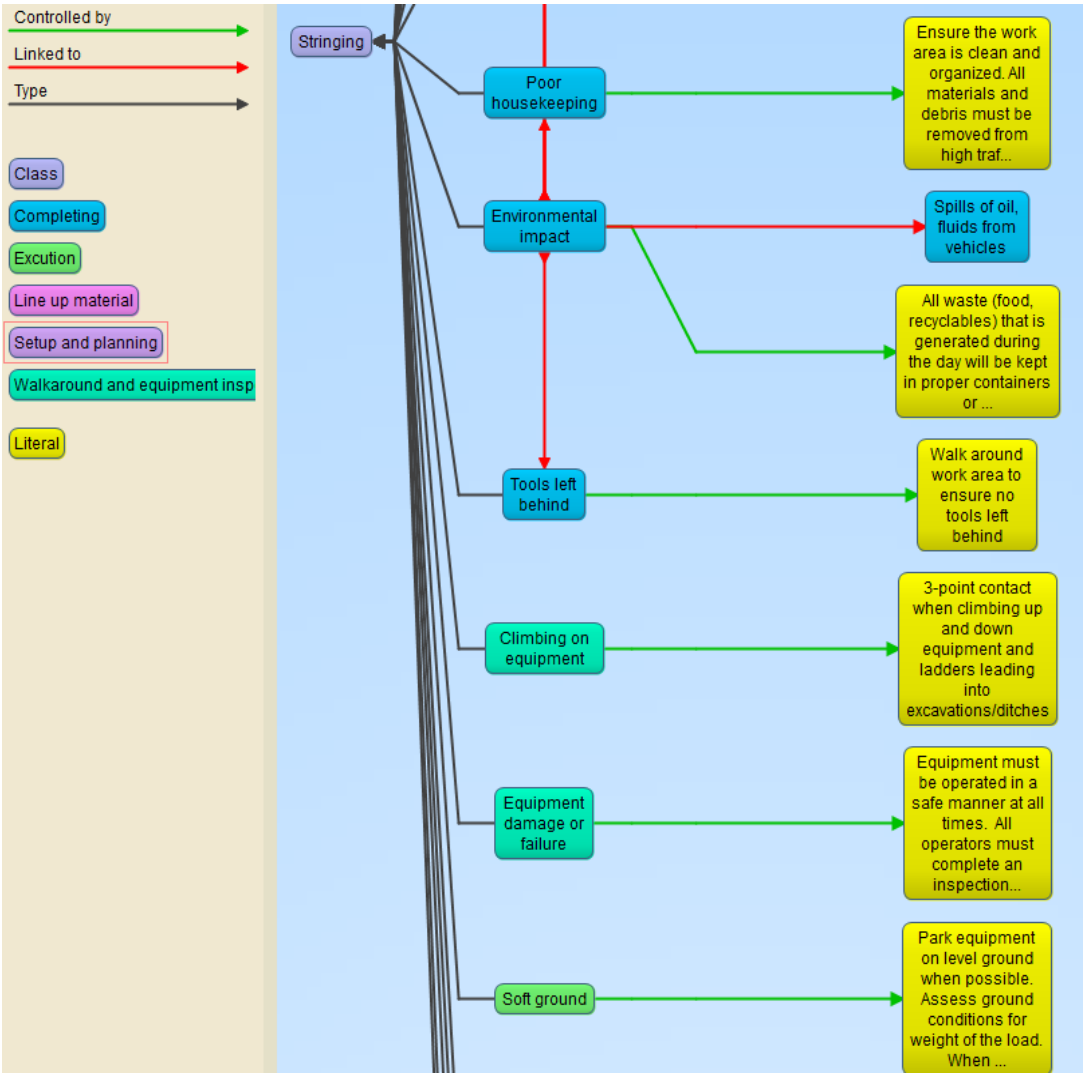


Figure 5-17 Competency question No. 4 visual graph results

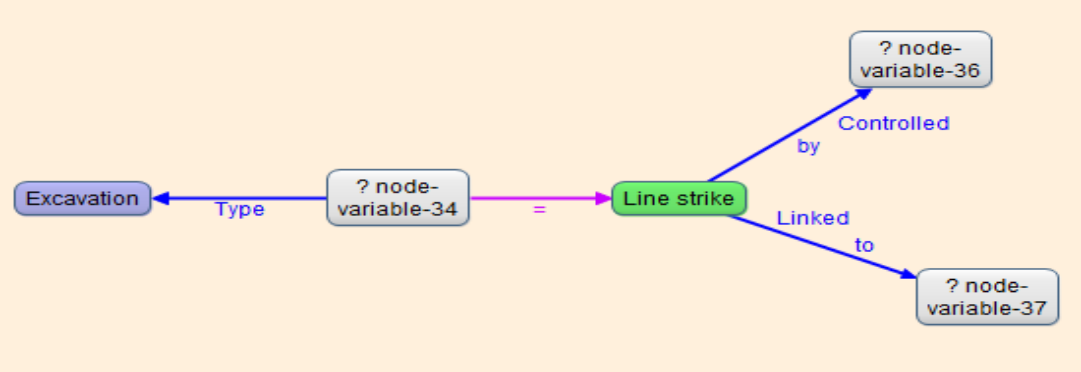


Figure 5-18 Competency question No. 7 visual query design

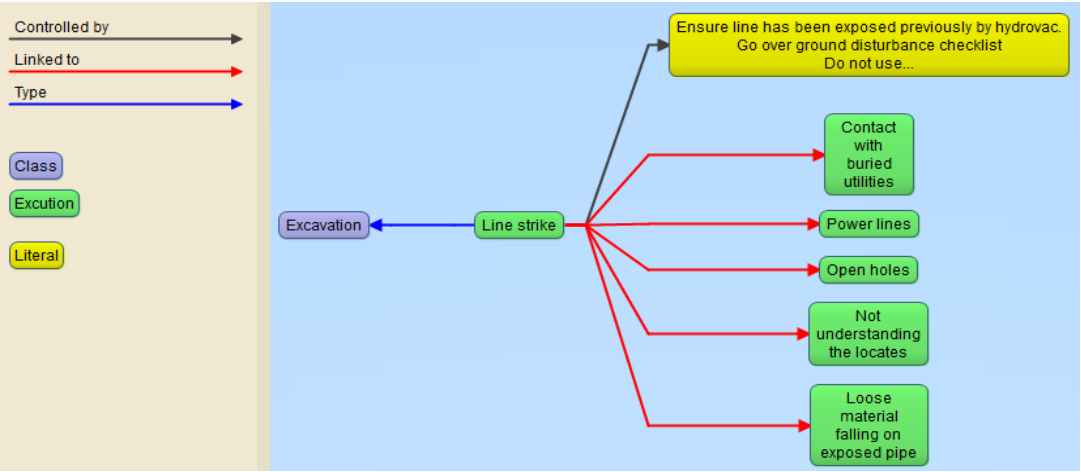


Figure 5-19 Competency question No. 7 visual graph results

Table 5-2 Competency question No. 7 answer table

Hazard	Hazard Controls	Linked Hazards
Line strike	1- Ensure line has been exposed previously by hydrovac. 2- Go over ground disturbance checklist 3- Do not use sharp objects or probes to locate pipe	Contact with buried utilities
		Loose material falling on exposed pipe
		Not understanding the locates
		Open holes
		Power lines

The hazard ontology performed well in answering the competency questions. This validation is important for the ontology model; it confirmed that the ontology is performing according to its design objectives. Present ontology is built based on knowledge schema that was extracted by performing knowledge content analysis for

JHA knowledge domain. Other ontologies which were developed in related previous reserch were not based on JHA domain analysis and they were validated based in AI dimention (Wang & Boukamp, 2011; Zhang, et al., 2015). Present ontology has more reliability in represnting hazard concepts , taxonomies and sematic relationships of JHA knowledge domain and hence is more effectiv in supporting future JHA process.

5.6 Conclusion

In this chapter, knowledge modeling using ontology was explored. Ontology-related components were presented and described. Extracted knowledge schema produced from Chapter 4 was implemented and formalized in classes, subclasses, instances, and properties' semantic relations.

The ontology was validated and was very successful in responding to the designed test queries. Knowledge modeling is a key step toward the digitization of hazard knowledge to promote hazards communication during the JHA process prior to commencing construction activities. In addition, the hazard knowledge ontology model can be used to benefit education and training operations for less experienced and new construction personnel.

CHAPTER 6

Research conclusion

6.1 Research summary

The research introduced new methods for the multi-level extracting of knowledge from past Job hazard analysis (JHA) documents. Extracted knowledge was transformed to knowledge schema that can be used to support the JHA process for oil and gas pipeline projects. JHA documents are a critical source of knowledge and contain a mix of explicit and tacit knowledge. The research started with exploring incident data related to different pipeline construction projects in the province of Alberta to highlight direct causes and root causes of incidents. While "failure to identify hazards" was the most frequent direct incident cause, a lack of knowledge was the most common root cause of incidents. This reflects the pivotal role of the hazard identification process in eliminating and mitigating incidents at construction sites.

A survey of recent literature indicates that many researchers have introduced different methods to support the JHA process to improve safety performances in construction projects. Recent research centralized around one idea: how to communicate and retrieve hazard knowledge during JHA in the planning stage to decrease incidents and increase safety performance in projects. However, most of the recent safety research is related to building projects. Nonbuilding projects such as pipeline construction lack sufficient safety research.

In Chapter 3, a method of categorization of JHA documents was introduced. The method includes two stages of processing. The first uses the clustering technique to group and label unknown documents. The second uses labeled documents from Stage One as a training set of data to build a classification model.

The clustering technique performed well in grouping documents in similar groups. The clusters purity measure was up to 80.9 %. Then, labeled documents produced from the clustering process were used as training data input for classification algorithms. JHA documents were classified using different algorithms. The best performance was for the K-NN algorithm using cosine similarity measures and has 93% accuracy in classifying the test documents group.

In Chapter 4, an analysis of the content of JHA documents was introduced to extract hazards associated with each construction activity. Text mining was used to build a hazards dictionary. Utilizing tokenization of collocation was very efficient in extracting hazard concepts related to different pipeline construction activities. A co-occurrence analysis was used to extract semantic relationships between hazards. The output of the overall analysis was formal knowledge schema and hazard dictionary for pipeline construction projects.

In Chapter 5, ontology modeling was used to model a knowledge structure obtained from hazard knowledge analysis. The model aims to formally represent the knowledge concepts, taxonomies, and relationships. This representation enables knowledge retrieval to support the JHA process in future projects. The ontology was validated and successfully responded to designed competency questions.

6.2 Research contribution

The research was developed in collaboration with an industry partner. The research contribution is presented in two categories, industrial and academic contribution, as follows:

6.2.1 Industrial contribution

- Currently, JHA document organization is an existing problem in the industry. Text clustering and classification algorithms is used in solving this problem.
- The research developed a new methodology for extracting JHA knowledge that was buried in previous JHA documents.
- The research developed a knowledge model using semantic technology to support and promote hazards communication during the JHA planning stage.
- The research contributes to the construction safety related to oil and gas pipeline projects which has not been studied extensively.

6.2.2 Academic contribution

- The research explores the potential of using text mining and machine learning in document categorization.
- Text mining and its associated machine learning algorithms were used to extract concepts and semantic relationships from JHA documents. Text mining decreases the time and effort needed for extraction and at the same time increases the level of objectivity in the analysis process.

- The research utilized semantic technology to solve the problem of communicating and sharing hazard knowledge.
- The overall research output represents a step forward toward the digitization of construction hazard knowledge for more consistent and systematic hazard identification process in the planning phase.

6.3 Limitation and recommendation for future research

The research methodology was applied using JHA documents from a research industry partner. The documents belong to several previous pipeline construction projects. However, having JHA documents from projects done by different construction organizations can add new knowledge to the extracted hazard knowledge dictionary, taxonomies and semantic relationships. Number of documents is essential to extract sufficient hazard knowledge. Increasing the number of input documents will add credit to the extracted knowledge.

This research can be extended to different areas in several dimensions as follows:

- The extracted knowledge can be used to build an education and training model to increase the level of hazard knowledge for workers with less experience or for newly employed construction professionals.
- Analyze incident records and reports and transform them to safety lessons learned that can be inserted as additional knowledge concepts to current knowledge schema. This will broaden the safety knowledge and benefit the JHA process communication.

- Digitizing the JHA process will enable integration with other construction management disciplines, especially planning and scheduling. Hazards and associated knowledge can be presented as a new dimension of pipeline construction activity in the construction schedule.
- The research methodology can be used to extract and represent knowledge related to different area of construction management such as contract administration, project planning, and control.

References

Abdelhamid, T. S. & Everett, J. G., 2000. IDENTIFYING ROOT CAUSES OF CONSTRUCTION ACCIDENTS. *JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT,* 126(1), pp. 52-60.

Agarwal, S. & Yu., H., 2009. *FigSum: Automatically generating structured text summaries for figures in biomedical literature.* s.l., In AMIA Annual Symposium Proceedings.

Aggarwal, C. C., 2015. *Data Mining: The Textbook.* New York: Springer International Publishing Switzerland.

AL Qady, M. & Kandil, A., 2014. Automatic clustering of construction project documents based on textual similarity. *Automation in Construction,* Volume 42, p. 36–49.

Al Qady, M. & Kandil, A., 2015. Automatic Classification of Project Documents on the Basis of Text Content. *Journal of Computing in Civil Engineering,* 29(3), pp. 1-11.

Alberta Construction Safety Association, 2013. *Leadership for Safety Excellence.* Edmonton: Construction Safety Association.

Anon., 2013. Document Clustering on Various Similarity Measures. *International Journal of Advanced Research in Computer Science and Software Engineering,* 3(13), pp. 1269-1273.

Association of Workers' Compensation Boards of Canada, 2015. *Fatality Statistics.* [Online]
Available at: http://awcbc.org/?page_id=14#fatalities
[Accessed 4 February 2017].

Batresa, R., Fujiharaa, S., Shimadab, Y. & Fuchinoc, T., 2014. The use of ontologies for enhancing the use of accident information. *Process Safety and Environmental Protection,* Volume 92, pp. 119-130.

Berners-Lee, T., 2000. *Semantic Web - XML2000.* [Online]
Available at: https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html
[Accessed 13 July 2017].

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American,* pp. 34-43.

Bharti, K. K., 2014. A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science,* Volume 5, pp. 156-169.

Brickley, D. & Guha, R., 2014. *RDF Schema 1.1.* [Online]
Available at: https://www.w3.org/TR/rdf-schema/
[Accessed 26 June 2017].

Buetow, T. et al., 2003. A Spatio Temporal Visualizer for Law Enforcement. *Intelligence and Security Informatics,* p. 181–194.

Buzydlowski, J. W., 2015. Co-occurrence analysis as a framework for data mining. *Journal of Technology Research,* Volume 6, pp. 1-19.

Caldas, C. H., Soibelman, L. & Han, J., 2002. Automated Classification of Construction Project Documents. *J. Comput. Civ. Eng.,* 16(4), pp. 234-243.

Canadian Standards Association, 2012. *Z662-11 Oil and gas pipeline systems.* Mississauga,: Canadian Standards Association.

Cao, N. & Cui, W., 2016. *Introduction to Text Visualization.* s.l.:Atlantis Press.

Card, S., Mackinlay, J. & Shneiderman, B., 1999. *Readings in Information Visualization: Using Vision to Think.* Los Altos: Morgan Kaufmann.

Carter, G. & Smith, S. D., 2006. Safety Hazard Identification on Construction Projects. *J. Constr. Eng. Manage,* 132(2), pp. 197-205.

Carter, G. & Smith, S. D., 2006. Safety Hazard Identification on Construction Projects. *Journal of Construction Engineering and Management,* 132(2), pp. 197-205.

Chi, N.-W., Lin, K.-Y. & Hsieh, S.-H., 2014. Using ontology-based text classification to assist Job Hazard Analysis. *Advanced Engineering Informatics,* Volume 28, pp. 381-394.

Cho, I., Dou, W., Wang, D. X. & Ribarsky, E. S. a. W., 2016. VAiRoma: A Visual Analytics System for Making Sense of Places,Times, and Events in Roman History. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS,* 22(1), pp. 2010-219.

Chua, K. D. & Goh, M. Y., 2002. Application of Case Based Reasoning In Construction Safety Planning. *J. Constr. Eng. Manage.,* Volume 138, pp. 1169-1180.

Cui, W. et al., 2010. Context preserving dynamic word cloud. *IEEE Symposium on Pacific Visualization,* p. 121–128.

Cyganiak, R., Wood, D. & Lanthaler, M., 2014. *RDF 1.1 Concepts and Abstract Syntax.* [Online]
Available at: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/
[Accessed 26 June 2017].

Dash, M. & Liu, H., 1997. Feature Selection for Classification. *Intelligent Data Analysis,* pp. 131-156.

El-Diraby, T. E., 2013. Domain Ontology for Construction Knowledge. *J. Constr. Eng. Manage.,* Volume 139, pp. 768-784.

El-Diraby, T. E., 2014. Validating ontologies in informatics systems: approaches and lessons learned for AEC. *Journal of Information Technology in Construction (ITcon),* Volume 19, pp. 474-493.

El-Diraby, T. E. & Kashif, K. F., 2005. Distributed Ontology Architecture for Knowledge Management in Highway Construction. *J. Constr. Eng. Manage.,* Volume 131, pp. 591-603.

Fan, H. & Li, H., 2013. Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. *Automation in Construction,* Volume 34, p. 85–91.

Fargnoli, M., 2011. Knowledge Management integration Integration in Occupational Health and Safety systems Systems in the construction industry. *Int. J. Product Development,* 14(1), pp. 165-185.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook.* s.l.:Cambridge University Press.

Franz Inc., 2017. *Home.* [Online]
Available at: https://franz.com/#
[Accessed 17 MAy 2017].

Gajzler, M., 2010. TEXT AND DATA MINING TECHNIQUES IN ASPECT OF KNOWLEDGE ACQUISITION FOR DECISION SUPPORT SYSTEM IN CONSTRUCTION INDUSTRY. *Baltic Journal on Sustainability,* 16(2), p. 219–232 .

Gasevic, D., Djuric, D. & Devedzic, V., 2009. *Model Driven Engineering and Ontology Development.* Verlag Berlin Heidelberg: Springer.

Goh, Y. M. & Chua, D. K. H., 2010. Case-Based Reasoning Approach to Construction Safety Hazard Identification: Adaptation and Utilization. *Journal of Construction,* 136(2), pp. 170-178.

Gruber, T., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition,* 5(2), pp. 199-220.

Hadikusumo, W. H. B. & Rowlinson, S., 2004. Capturing Safety Knowledge Using Design-for-Safety-Process Tool. *Journal of Construction Engineering and Management,* 130(2), pp. 281-289.

Hallowell, M. R., 2012. Safety-Knowledge Management in American Construction Organizations. *J. Manage. Eng,* Volume 28, pp. 203-211.

Hallowell, M. R., 2015. Capturing Safety Knowledge Using Design-for-Safety-Process Tool. *Journal of Management in Engineering,* 28(2), pp. 203-211.

Han, J. & Kamber, M., 2006. *Data Mining Concepts and Techniques.* San Francisco: Diane Cerra.

Heflin, J. & Hendler, J., 2001. A portrait of the Semantic Web in action. *IEEE Intelligent Systems,* 16(2), pp. 54-59.

He, Q., 1999. Knowledge discovery through co-word analysis. *Johns Hopkins University Press,* 48(1), pp. 133-159.

Huang, A., 2008. Similarity Measures for Text Document Clustering. *NZCSRSC,* pp. 49-56.

Jakus, G., Milutinovic, V., Omerovic, S. & Tomazic, S., 2013. *Concepts, Ontologies, and Knowledge Representation.* New York: Springer.

John, M. et al., 2016. *Visual Analytics for Narrative Text: Visualizing Characters and their Relationships as Extracted from Novels.* s.l., Proceedings of the 7th International Conference on Information Visualization Theory and Applications.

Jun, S., Park, S.-S. & Jang, D.-S., 2014. Document clustering method using dimension reduction and support. *Expert Systems with Applications,* Volume 41, pp. 3204-3212.

K.Sruthi & Reddy, B., 2013. Document Clustering on Various Similarity Measures. *International Journal of Advanced Research in Computer Science and Software Engineering,* 3(8), pp. 1269-1273.

Kamardeen, I., 2013. *OHS Electronic Management Systems for Construction.* New York: Routledge.

Kantardzic, M., 2011. *Data Mining Concepts, Models, Methods, and Algorithms.* 2nd ed. New Jersey: John Wiley & Sons, Inc..

Kartam, N. A., 1997. INTEGRATING SAFETY AND HEALTH PERFORMANCE INTO CONSTRUCTION CPM. *JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT,* 123(2), pp. 121-126..

Kim, H. et al., 2015. Information Retrieval Framework for Hazard Identification in Construction. *J. Comput. Civ. Eng.,* 29(3), pp. 1-10.

Kotu, V. & Deshpande, B., 2015. *Predictive Analytics and Data Mining: Concepts and Practice with.* Waltham,: Elsevier Inc.

Le, Q. T., Lee, Y. D. & Park, C. S., 2014. A Social Network System for Sharing Construction Safety and Health Knowledge. *Automation in Construction,* Volume 46, pp. 30-37.

Lin, Y.-C. & Lee, H.-Y., 2012. Developing Pproject Communities of Practice-Based Knowledge Management System in Construction. *Automation in Construction,* Volume 22, p. 422–432.

Li, Y., Chung, S. M. & Holt, J. D., 2008. Text document clustering based on frequent word meaning sequences. *Data & Knowledge Engineering,* Volume 64, pp. 381-404.

LI, Z., LI, X. & JIANG, S., 2014. *Web-based Text Mining for Extracting Relationships among Policies of Building and Construction Industry.* s.l., ICCREM 2014.

Manning, C. D. & Schutze, H., 2001. *Foundation of Statistical Natural Language Processing.* Cambridge: MIT Press.

McGuinness, D. L. & Harmelen, F. v., 2004. *OWL Web Ontology Language.* [Online] Available at: https://www.w3.org/TR/2004/REC-owl-features-20040210/#s1 [Accessed 26 June 2017].

Melzner, J., Zhang, S., Teizer, J. & Bargstädt, H., 2013. A Case study on Automated Safety Compliance Checking to Assist Fall Protection Design and Planning in Building Information Models. *Construction Management and Economics,* 31(6), pp. 661-674.

Meziane, F. & Rezgui, Y., 2004. A document management methodology based on similarity contents. *Information Sciences,* Volume 158, p. 15–36.

Miner, G. D. et al., 2012. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Waltham: Elsevier Inc.

Moens, M.-F., 2006. *Information Extraction: Algorithms and Prospects in a Retrieval Context.* 1 ed. Dordrecht: Springer Netherlands.

Musen, M., 2015. The Protégé project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence,* 1(4), pp. 4-12.

Nagarajan, R. & Aruna, P., 2016. Construction of Keyword Extraction using Statistical Approaches and Document Clustering by Agglomerative method. *Journal of Engineering Research and Applications,* 6(1), pp. 73-78.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. & Wanapu, S., 2013. *Using of Jaccard Coefficient for Keywords Similarity.* Hong Kong, Proceedings of the International MultiConference of Engineers and Computer Scientists.

Noy, N. F. et al., 2001. Creating Semantic Web contents with Protégé-2000. *IEEE Intelligent Systems,* 16(2), pp. 60-71.

Occupational Health and Safety Code, 2009. *Occupational Health and Safety Code.* Edmonton : Alberta Queen's Printer.

Podgórski, D., 2010. The Use of Tacit Knowledge in Occupational The Use of Tacit Knowledge in Occupational. *International Journal of Occupational Safety and Ergonomics,* 16(3), pp. 283-310.

Provalis Research, 2015. *WORDSTAT User's Guide.* Montreal: Provalis Research.

Rezgui, Y., 2007. Knowledge Systems and Value creation. *Industrial Management & Data Systems.,* 107(02), pp. 166-182.

SEBASTIANI, F., 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys,* 34(1), pp. 1-47.

Singthongchai, J. & Niwattanakul, S., 2013. A Method for Measuring Keywords Similarity by Applying Jaccard's, N-Gram and Vector Space. *Lecture Notes on Information Theory,* Volume 1, pp. 159-164.

Sowa, J. F., 2000. *Knowledge Representation: Logical, Philosophical, and Computational. Foundations.* Pacific Grove: Brooks Cole.

Stemler, S., 2001. An overview of content analysis. *Practical Assessment,Research & Evaluation,* 7(17), pp. 1-10.

Tatume, C. B., 2011. Core Elements of Construction Engineering Knowledge for Project and Career Success. *Journal of Construction Engineering and Management,* Volume 137, pp. 745-750.

Taye, M. M., 2010. Understanding Semantic Web and Ontologies:Theory and Applications. *Journal of Computing,* 2(6), pp. 182-192.

Tixier, A. J.-P. & Hallowell, M. R., 2016. Application of machine learning to construction injury prediction;Balaji Rajagopalan;Dean Bowman. *Automation in Construction,* Volume 69, pp. 102-114.

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B. & Bowman, D., 2016. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction,* Volume 62, pp. 45-56.

United States Department of State, 2014. *Final Supplemental Environmental Impact Statement for the Keystone XL Project Executive Summary.* Washington: s.n.

Verstichel, S. et al., 2011. Efficient data integration in the railway domain through an ontology-based methodology. *Transportation Research Part C,* Volume 19, pp. 617-643.

W3C OWL Working Group, 2012. *OWL 2 Web Ontology Language.* [Online]
Available at: https://www.w3.org/TR/owl2-overview/
[Accessed 26 June 2017].

W3C SPARQL Working Group, 2013. *SPARQL 1.1 Overview.* [Online]
Available at: https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/#sec-intro
[Accessed 27 June 2017].

Wang, H.-H. & Boukamp, F., 2009. *Ontology-based Job Hazard Analysis Support.* Austin, American Society of Civil Engineers.

Wang, H.-H. & Boukamp, F., 2011. Ontology-Based Representation and Reasoning Framework for Supporting Job Hazard Analysis. *J. Comput. Civ. Eng.,* Volume 25, pp. 442-456.

Wang, H. H. & Boukamp, F., 2011. Ontology-Based Representation and Reasoning Framework for Supporting Job Hazard Analysis. *Journal of Computing in Civil Engineering,* 28(2), pp. 203-211.

Wegerif, R. & Mercer, N., 1997. Using Computer-based Text Analysis to Integrate Qualitative and Quantitative Methods in Research on Collaborative Learning. *Language and Education,* 11(4), pp. 271-286.

Weiss, S. M., Indurkhya, N. & Zhang, T., 2015. *Fundamentals of Predictive Text Mining.* 2nd ed. London: Springer-Verlag London.

Williams, P. T. & Gong, J., 2014. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction,* Volume 43, pp. 23-29.

YANG, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval,* Volume 1, pp. 69-90.

Yildiz, A. E. et al., 2014. A knowledge-based riskmapping tool for cost estimation of international construction projects. *Automation in Construction,* Volume 43, pp. 144-155.

Yusof, M. N. & Bakar, A. H. A., 2012. Knowledge management and growth performance in construction companies: a framework. *Procedia - Social and Behavioral Sciences,* Volume 62, pp. 128-134.

Zhang, J., Kwigizile, V. & Jun-Seok, 2016. Automated Hazardous Action Classification Using Natural Language Processing and Machine-Learning Techniques. *16th COTA International Conference of Transportation Professional ,* pp. 1579-1590.

Zhang, S., Boukamp, F. & Teizer, J., 2015. Ontology-based semantic modeling of construction safety knowledge:Towards automated safety planning for job hazard analysis (JHA). *Automation in Construction,* Volume 52, pp. 29-41.

Zhang, W., Yoshida, T., Tang, X. & Wang, Q., 2010. Text clustering using frequent itemsets. *Knowledge-Based Systems,* Volume 23, pp. 379-388.

Zhong, B. et al., 2012. Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking. *Automation in Construction,* Volume 28, p. 58–70.

Zhou, Z., Goh, Y. M. & Li, Q., 2015. Overview and analysis of safety management studies in the construction industry. *Safety Science,* Volume 72, pp. 337-350.

Appendices

Appendix I: Sample of JHA form

JOB HAZARD ASSESSMENT
CONTRACTORS / PIPELINE / SET

JOB: 7015175
DATE: _____
NAME: HYDROVAC WITHIN 1.5M

EMERGENCY CONTACTS:
Supervisor: _____ Contact #: _____
Emergency Response #: _____ Meeting Point or Location: _____

STEPS TO DO THE JOB	POTENTIAL HAZARDS	INHERENT RISK RATING	CONTROLS	RESIDUAL RISK RATING
1. Bring the excavator into position to begin digging	• Equipment failure		• Operator will be provided with an excavator that has been both mechanically inspected as well as having passed yearly equipment certification requirements. Certification stickers will be affixed in a conspicuous location • Operator is required to perform a documented daily equipment inspection at the start of his work shift.	
	• Operator error		• Excavators shall be operated only by designated persons, trainees under the direct supervision of a designated person and maintenance and test personnel, when it is necessary in the performance of their duties.	
	• Congested work area		• No unnecessary personnel/equipment in the work area. • Clear communication between ground personnel and operator	
2. Determining the area to be excavated	• Underground facilities		• The "Excavation checklist SWP700" must be completed prior to commencing any type of excavation. • Ensure one call has been done and owner has been notified 2days previous	
3. Digging out area	• Line hits		• Ensure line has been exposed previously by hydrovac. • Have an oiler on site at all times and pay attention to signals, must have an air horn • Do not use sharp objects or probes to locate pipe • Go over ground disturbance checklist	
	• Power lines		• Ensure power line markers erect. • Use a spotter • When working in these areas to ensure clearance • Follow safe work practices regarding power lines	
4. Hydrovac	• Flying debris, wet conditions, cold weather, toxic environments, open holes		• Ensure crossing agreements are in place and one call notification has been submitted. Ensure safework permit is in place (if required). • Wear proper PPE, along with Nomex coveralls and have a portable monitor with you at all times. • Ensure hydrovac holes are either fenced off or backfilled after line has been located. • Ensure MSDS, owner's name, depth and size of line are attached to the post to notify others of the	

JOB HAZARD ASSESSMENT
CONTRACTORS / PIPELINE / SET

JOB: 7015175
DATE: _____
NAME: HYDROVAC WITHIN 1.5M

Appendix II: Execution hazards similarity matrix

	Access and egress	Air quality	Arc flash	Awkward body positioning	Blind spots	Boredom	Cave-in hazard	Chain movement	Chainsaw kickback	Chemical exposure	Collision hazard	Communication failure	Compressed gasses	Confined space	Congested work area	Contact with buried utilities	Crush hazards	Crush Zones (feet under outriggers)	Damaged tools	Defective tools	Dehydration	Electrical shock	Equioment damage	Equipment failure	Equipment integrity	Equipment malfunction	Equipment tipping over	Ergonomics	Excavator becoming unstable during excavation	Excessive pipe movement
Access and egress	1.00																													
Air quality	0.08	1.00																												
Arc flash	0.02	0.00	1.00																											
Awkward body positioning	0.05	0.03	0.29	1.00																										
Blind spots	0.08	0.00	0.03	0.00	1.00																									
Boredom	0.00	0.24	0.00	0.00	0.00	1.00																								
Cave-in hazard	0.14	0.00	0.00	0.03	0.12	0.19	1.00																							
Chain movement	0.02	0.00	0.00	0.00	0.06	0.00	0.00	1.00																						
Chainsaw kickback	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.50	1.00																					
Chemical exposure	0.02	0.00	0.00	0.00	0.06	0.08	0.09	0.00	0.00	1.00																				
Collision hazard	0.00	0.00	0.00	0.00	0.03	0.45	0.17	0.00	0.00	0.14	1.00																			
Communication failure	0.10	0.00	0.00	0.00	0.06	0.00	0.00	0.11	0.07	0.09	0.03	1.00																		
Compressed gasses	0.04	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.07	0.00	0.00	1.00																	
Confined space	0.36	0.11	0.02	0.09	0.00	0.00	0.04	0.00	0.00	0.12	0.06	0.03	0.00	1.00																
Congested work area	0.17	0.06	0.06	0.09	0.06	0.13	0.10	0.01	0.01	0.02	0.15	0.01	0.01	0.10	1.00															
Contact with buried utilities	0.05	0.00	0.02	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00														
Crush hazards	0.02	0.29	0.03	0.00	0.12	0.57	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.10	0.05	1.00													
Crush Zones (feet under outriggers)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.05	0.00	0.00	1.00												
Damaged tools	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.03	0.00	0.00	0.00	1.00											
Defective tools	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.06	0.00	0.00	0.00	0.00	1.00										
Dehydration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.05	0.00	0.00	0.39	0.00	0.00	1.00									
Electrical shock	0.05	0.00	0.66	0.27	0.04	0.00	0.02	0.00	0.00	0.07	0.00	0.00	0.21	0.08	0.08	0.02	0.02	0.00	0.02	0.02	0.00	1.00								
Equioment damage	0.02	0.20	0.00	0.13	0.09	0.15	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.13	0.02	0.21	0.00	0.00	0.00	0.00	0.00	1.00							
Equipment failure	0.03	0.00	0.12	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.14	0.00	0.00	0.00	0.04	0.00	0.00	0.11	0.12	1.00						
Equipment integrity	0.00	0.35	0.00	0.00	0.00	0.60	0.26	0.00	0.00	0.12	0.19	0.00	0.00	0.00	0.08	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.19	0.00	1.00					
Equipment malfunction	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.27	0.05	0.08	0.04	0.06	0.00	0.00	0.50	0.00	0.00	0.39	0.04	0.00	0.00	0.00	1.00				
Equipment tipping over	0.04	0.00	0.00	0.00	0.04	0.00	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.04	0.38	0.00	0.03	0.00	0.00	0.05	0.00	0.03	0.00	0.00	0.03	1.00			
Ergonomics	0.10	0.00	0.18	0.21	0.08	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.04	0.15	0.02	0.02	0.00	0.00	0.00	0.00	0.14	0.02	0.09	0.00	0.00	0.02	1.00		
Excavator becoming unstable during excavation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.00	1.00	
Excessive pipe movement	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.05	0.00	0.00	1.00	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.50	0.03	0.00	0.00	1.00

	Excessive pipe movement	Exhaust fumes	Explosion	Exposure to Dust	Exposure to fumes, dust and gases	Failure of fittings and tools	Falling trees	Fitting failure	Flame	Flying debris	Flying objects	Freeze up	Ground conditions	Ground personnel	Hammer / Lath	Hazards not controlled	Heavy clamps	Heavy Lifting	High pressure	Hose failure	Hot surfaces	Icy or muddy ground conditions	Improper fittings	Improper lift	Improper use of PPE	Improper use of skill saw and chainsaw	Inadequate lighting	Inadequate PPE	Inattention	Kickback of grinders	Leaking hoses/fittings	Leaks from propane tank	Limited visibility	Line of fire	Line strike	Load tipping	Local traffic	Loose material falling on exposed pipe
Excessive pipe movement	1.00																																				
Exhaust fumes	0.00	1.00																																			
Explosion	0.00	0.00	1.00																																		
Exposure to Dust	0.00	0.05	0.00	1.00																																	
Exposure to fumes, dust and gases	0.00	0.00	0.03	0.00	1.00																																
Failure of fittings and tools	0.00	0.00	0.00	0.00	0.00	1.00																															
Falling trees	0.00	0.00	0.00	0.00	0.00	0.00	1.00																														
Fitting failure	0.00	0.00	0.00	0.00	0.00	0.10	0.00	1.00																													
Flame	0.00	0.00	0.03	0.03	0.18	0.00	0.04	0.00	1.00																												
Flying debris	0.00	0.01	0.01	0.11	0.13	0.00	0.10	0.00	0.20	1.00																											
Flying objects	0.00	0.00	0.00	0.22	0.35	0.00	0.00	0.00	0.03	0.01	1.00																										
Freeze up	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00																									
Ground conditions	0.21	0.02	0.00	0.00	0.05	0.02	0.02	0.03	0.16	0.15	0.01	0.05	1.00																								
Ground personnel	0.17	0.03	0.10	0.06	0.02	0.07	0.02	0.03	0.01	0.15	0.07	0.03	0.32	1.00																							
Hammer / Lath	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.39	0.00	0.01	1.00																						
Hazards not controlled	0.00	0.10	0.00	0.00	0.13	0.05	0.00	0.05	0.13	0.05	0.00	0.08	0.18	0.04	0.00	1.00																					
Heavy clamps	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.20	0.06	0.00	0.00	0.05	0.00	0.00	0.25	1.00																				
Heavy Lifting	0.00	0.13	0.00	0.07	0.00	0.00	0.00	0.06	0.16	0.16	0.00	0.27	0.15	0.05	0.20	0.02	0.00	1.00																			
High pressure	0.00	0.00	0.00	0.24	0.03	0.15	0.00	0.09	0.16	0.23	0.11	0.00	0.16	0.18	0.02	0.05	0.00	0.23	1.00																		
Hose failure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.03	0.04	1.00																	
Hot surfaces	0.00	0.04	0.02	0.32	0.40	0.00	0.00	0.00	0.13	0.22	0.39	0.04	0.09	0.12	0.02	0.12	0.10	0.01	0.19	0.00	1.00																
Icy or muddy ground conditions	0.28	0.01	0.08	0.05	0.03	0.00	0.00	0.03	0.14	0.15	0.03	0.04	0.31	0.33	0.00	0.00	0.00	0.18	0.15	0.03	0.08	1.00															
Improper fittings	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.10	0.00	0.00	0.00	0.00	0.02	0.07	0.00	0.05	0.00	0.00	0.15	0.00	0.00	0.00	1.00														
Improper lift	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.01	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.05	0.00	1.00													
Improper use of PPE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00												
Improper use of skill saw and chainsaw	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.08	0.00	0.00	0.00	1.00											
Inadequate lighting	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.03	0.04	0.00	0.00	0.00	0.03	0.06	0.00	0.02	0.03	0.00	0.00	0.00	0.00	1.00										
Inadequate PPE	0.00	0.05	0.00	0.14	0.00	0.00	0.05	0.00	0.00	0.08	0.03	0.29	0.06	0.07	0.11	0.00	0.00	0.22	0.02	0.00	0.04	0.11	0.00	0.20	0.00	0.00	0.00	1.00									
Inattention	0.00	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.13	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.03	0.00	1.00								
Kickback of grinders	0.00	0.00	0.02	0.14	0.58	0.00	0.00	0.00	0.13	0.21	0.49	0.00	0.04	0.06	0.00	0.09	0.12	0.00	0.08	0.00	0.48	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.00	1.00							
Leaking hoses/fittings	0.00	0.00	0.04	0.00	0.58	0.00	0.00	0.00	0.24	0.14	0.04	0.00	0.04	0.00	0.00	0.17	0.36	0.00	0.00	0.00	0.21	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	1.00						
Leaks from propane tank	0.00	0.14	0.04	0.00	0.38	0.00	0.00	0.00	0.20	0.11	0.00	0.10	0.07	0.02	0.00	0.30	0.40	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.62	1.00					
Limited visibility	0.00	0.04	0.00	0.03	0.00	0.00	0.00	0.00	0.19	0.15	0.00	0.00	0.11	0.08	0.00	0.03	0.00	0.26	0.33	0.00	0.02	0.10	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	1.00				
Line of fire	0.14	0.08	0.01	0.02	0.21	0.01	0.02	0.04	0.07	0.17	0.08	0.07	0.31	0.22	0.01	0.16	0.05	0.12	0.09	0.03	0.26	0.23	0.01	0.04	0.00	0.06	0.04	0.09	0.01	0.18	0.13	0.13	0.04	1.00			
Line strike	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.03	0.07	0.12	0.00	0.14	0.13	0.08	0.00	0.02	0.00	0.11	0.01	0.00	0.00	0.06	0.00	0.10	0.00	0.00	0.03	0.19	0.00	0.00	0.00	0.00	0.02	0.06	1.00		
Load tipping	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	1.00	
Local traffic	0.00	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Loose material falling on exposed pipe	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.10	0.03	0.00	0.04	0.11	0.01	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.05	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.00	0.00	1.00

	Loose material falling on exposed pipe	Miscommunication	Missing information	Moving Equipment	Muscle strain	New Workers	Noise	Not understanding the locates	Open excavations	Open flame	Open holes	Operator error	Other workers in area	Overhead hazards	Overweight loads	Parked in blind spot of equipment	Pedestrian traffic	Pinch points	Pipe falling	Pipe falling off the truck	Poor communication	Poor lighting	Poor skids / skid piles	Poor visibility	Ppower lines	Propane leaks	Repetition	Respiratory hazard	Rigging failure	Road conditions	Rolling / Shifting pipe	Rotating parts	Rough terrain	Serious Injury	Sharp blade	Sharp edges	Shifting pipe	Shock hazards	Skin contact
Loose material falling on exposed pipe	1.00																																					
Miscommunication	0.00	1.00																																				
Missing information	0.00	0.04	1.00																																			
Moving Equipment	0.02	0.17	0.05	1.00																																		
Muscle strain	0.00	0.00	0.00	0.17	1.00																																	
New Workers	0.00	0.00	0.00	0.02	0.00	1.00																																
Noise	0.00	0.27	0.00	0.22	0.13	0.00	1.00																															
Not understanding the locates	0.85	0.00	0.00	0.03	0.00	0.00	0.00	1.00																														
Open excavations	0.03	0.04	0.19	0.08	0.00	0.10	0.03	0.07	1.00																													
Open flame	0.00	0.14	0.14	0.31	0.00	0.07	0.11	0.00	0.15	1.00																												
Open holes	0.27	0.04	0.13	0.10	0.00	0.07	0.05	0.31	0.25	0.12	1.00																											
Operator error	0.00	0.09	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.02	0.16	1.00																										
Other workers in area	0.00	0.21	0.10	0.16	0.00	0.04	0.03	0.00	0.17	0.15	0.11	0.03	1.00																									
Overhead hazards	0.00	0.14	0.00	0.20	0.00	0.00	0.18	0.00	0.04	0.45	0.06	0.02	0.15	1.00																								
Overweight loads	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.04	0.00	0.03	0.04	0.12	0.21	1.00																							
Parked in blind spot of equipment	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00																						
Pedestrian traffic	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.71	1.00																					
Pinch points	0.01	0.34	0.03	0.38	0.11	0.01	0.23	0.01	0.13	0.17	0.07	0.06	0.25	0.14	0.01	0.00	0.00	1.00																				
Pipe falling	0.00	0.18	0.00	0.33	0.19	0.00	0.25	0.00	0.00	0.18	0.00	0.00	0.09	0.17	0.00	0.00	0.00	0.22	1.00																			
Pipe falling off the truck	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.21	1.00																		
Poor communication	0.00	0.02	0.00	0.13	0.39	0.00	0.13	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.17	0.00	1.00																	
Poor lighting	0.00	0.08	0.31	0.08	0.00	0.13	0.03	0.00	0.29	0.16	0.21	0.00	0.32	0.02	0.05	0.00	0.00	0.09	0.00	0.00	0.00	1.00																
Poor skids / skid piles	0.00	0.14	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.12	0.16	0.00	0.00	0.00	1.00															
Poor visibility	0.00	0.03	0.03	0.16	0.65	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.17	0.16	0.00	0.46	0.02	0.00	1.00														
Ppower lines	0.21	0.09	0.00	0.15	0.15	0.00	0.16	0.21	0.11	0.02	0.28	0.11	0.13	0.13	0.10	0.00	0.00	0.18	0.12	0.00	0.16	0.01	0.07	0.15	1.00													
Propane leaks	0.00	0.03	0.22	0.20	0.26	0.12	0.10	0.00	0.19	0.29	0.15	0.00	0.10	0.05	0.00	0.00	0.00	0.11	0.16	0.00	0.22	0.22	0.00	0.19	0.10	1.00												
Repetition	0.00	0.00	0.00	0.15	0.70	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.17	0.00	0.52	0.00	0.00	0.85	0.13	0.22	1.00											
Respiratory hazard	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	1.00										
Rigging failure	0.00	0.12	0.00	0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.15	0.10	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.03	0.00	0.17	0.00	0.09	0.00	0.00	0.00	1.00									
Road conditions	0.00	0.00	0.00	0.14	0.52	0.00	0.15	0.00	0.02	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.15	0.16	0.00	0.48	0.00	0.00	0.67	0.11	0.21	0.77	0.00	0.03	1.00								
Rolling / Shifting pipe	0.00	0.14	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.12	0.16	0.00	0.00	0.00	1.00	0.00	0.07	0.00	0.00	0.00	0.17	0.00	1.00							
Rotating parts	0.00	0.09	0.05	0.07	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.16	0.03	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.03	0.00	0.05	0.03	0.00	0.00	0.05	0.19	0.00	0.00	1.00						
Rough terrain	0.48	0.00	0.00	0.09	0.26	0.00	0.00	0.55	0.05	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.10	0.00	0.00	0.20	0.25	0.00	0.22	0.00	0.00	0.15	0.00	0.00	1.00					
Serious Injury	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00					
Sharp blade	0.00	0.04	0.00	0.06	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.19	0.02	0.03	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.23	0.00	0.00	0.25	0.00	0.50	1.00			
Sharp edges	0.00	0.00	0.00	0.05	0.33	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.16	0.00	0.00	0.22	0.02	0.00	0.26	0.00	0.04	0.29	0.00	0.00	0.26	0.00	0.00	1.00		
Shifting pipe	0.00	0.06	0.24	0.11	0.00	0.00	0.01	0.00	0.13	0.16	0.10	0.12	0.08	0.02	0.00	0.00	0.00	0.06	0.02	0.00	0.00	0.14	0.00	0.00	0.01	0.18	0.00	0.00	0.14	0.00	0.00	0.15	0.00	0.00	0.18	0.00	1.00	
Shock hazards	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	
Skin contact	0.00	0.02	0.00	0.02	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.62	0.07	0.06	0.00	0.04	0.00	0.00	0.00	0.05	0.00	0.00	1.00

	Skin contact	Skin irritation	Sling failure	Slings slipping	Slippery conditions	Slips, Trips and Falls	Soft ground	Soil conditions	Sparks	Spills and leaks	Sprains and strains	Steep ground	Suspended load	Swing radius	Swinging load	Swinging pipe	Tight areas	Tool failure	Toxic Atmosphere	toxic environments	Unbalanced load	Uncontrolled intersections	Underground facilities	Uneven ground	Unplanned movement of load	Unsecured pipe	Unstable ground	Valve/fitting failure	Weather conditions	Wet conditions	Wind Direction	Worker fatigue	Working alone	Working around equipment	Working near water	Working without gloves	Wrench slip
Skin contact	1.00																																			
Skin irritation	0.23	1.00																																		
Sling failure	0.00	0.00	1.00																																	
Slings slipping	0.00	0.00	0.00	1.00																																
Slippery conditions	0.01	0.00	0.05	0.00	1.00																															
Slips, Trips and Falls	0.09	0.01	0.04	0.01	0.30	1.00																														
Soft ground	0.03	0.00	0.09	0.00	0.20	0.18	1.00																													
Soil conditions	0.04	0.08	0.00	0.00	0.00	0.09	0.03	1.00																												
Sparks	0.00	0.00	0.00	0.00	0.07	0.02	0.00	0.00	1.00																											
Spills and leaks	0.00	0.00	0.00	0.03	0.11	0.09	0.00	0.00	0.00	1.00																										
Sprains and strains	0.05	0.10	0.00	0.00	0.02	0.04	0.03	0.00	0.00	0.03	1.00																									
Steep ground	0.04	0.00	0.00	0.00	0.04	0.07	0.03	0.04	0.00	0.09	0.05	1.00																								
Suspended load	0.10	0.02	0.23	0.08	0.30	0.21	0.18	0.00	0.10	0.06	0.03	0.01	1.00																							
Swing radius	0.21	0.00	0.00	0.00	0.13	0.12	0.04	0.00	0.00	0.19	0.04	0.03	0.13	1.00																						
Swinging load	0.00	0.00	0.21	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	1.00																					
Swinging pipe	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	1.00																				
Tight areas	0.00	0.00	0.05	0.00	0.12	0.08	0.02	0.00	0.00	0.03	0.04	0.00	0.01	0.03	0.00	0.00	1.00																			
Tool failure	0.08	0.06	0.36	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.13	0.00	0.06	0.00	0.00	1.00																		
Toxic Atmosphere	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00																	
toxic environments	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00																
Unbalanced load	0.00	0.00	0.00	0.00	0.08	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00															
Uncontrolled intersections	0.00	0.00	0.00	0.38	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.05	0.27	0.00	0.00	0.00	0.00	1.00														
Underground facilities	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.00	1.00													
Uneven ground	0.03	0.01	0.11	0.07	0.20	0.29	0.06	0.01	0.10	0.06	0.05	0.09	0.24	0.05	0.05	0.00	0.14	0.06	0.00	0.00	0.07	0.12	0.00	1.00												
Unplanned movement of load	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00											
Unsecured pipe	0.00	0.00	0.00	0.00	0.08	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.07	0.00	1.00										
Unstable ground	0.03	0.03	0.00	0.00	0.13	0.13	0.02	0.03	0.02	0.02	0.13	0.02	0.13	0.07	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	1.00									
Valve/fitting failure	0.00	0.00	0.08	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00								
Weather conditions	0.06	0.00	0.05	0.02	0.19	0.15	0.05	0.03	0.08	0.05	0.02	0.17	0.14	0.07	0.00	0.02	0.04	0.03	0.00	0.16	0.00	0.01	0.15	0.29	0.00	0.00	0.07	0.03	1.00							
Wet conditions	0.00	0.00	0.00	0.00	0.03	0.02	0.03	0.00	0.00	0.06	0.06	0.08	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.57	0.00	0.00	0.54	0.03	0.00	0.00	0.00	0.00	0.16	1.00						
Wind Direction	0.04	0.00	0.00	0.00	0.11	0.07	0.05	0.00	0.00	0.16	0.00	0.00	0.07	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.03	0.00	0.12	0.00	1.00					
Worker fatigue	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00				
Working alone	0.00	0.00	0.00	0.00	0.12	0.02	0.00	0.00	0.00	0.07	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03	0.13	0.00	0.00	1.00			
Working around equipment	0.04	0.00	0.03	0.05	0.14	0.03	0.00	0.00	0.00	0.06	0.00	0.04	0.09	0.00	0.06	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.14	0.00	0.01	0.00	0.00	0.00	0.25	1.00		
Working near water	0.00	0.00	0.08	0.00	0.03	0.04	0.00	0.00	0.00	0.08	0.00	0.12	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.03	0.15	0.00	0.00	0.22	0.00	1.00	
Working without gloves	0.00	0.00	0.07	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	1.00
Wrench slip	0.04	0.08	0.35	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.12	0.00	0.07	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00