2018-12-07

Predictive Analysis and Recommendation for Managing Risk and Avoiding Hazard in Chemical and Oil & Gas Industrial Infrastructures

Polat, Serhan

UNIVERSITY OF CALGARY

Predictive Analysis and Recommendation for Managing Risk and Avoiding Hazard in Chemical and Oil & Gas Industrial Infrastructures

by

Serhan Polat

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

DECEMBER, 2018

Abstract

Chemical processing industrial infrastructures such as oil & gas plants are operated with the risk of hazardous events which may lead to casualties, economic and/or environmental consequences. Fortunately, a variety of devices and mechanisms are already available or rapidly emerging to capture data which may be used to develop techniques that may assist in issuing timely hazard alerts. This would help to avoid or prevent the hazard and hence save lives, the environment and the economy. Thus, the aim of this thesis is to develop an approach capable of analyzing the reports data captured after operations of infrastructure which can be used to guide domain experts in handling various causes and consequences of hazards. Available data may be publicly available or may exist in private repositories of processing companies. The latter data may not be accessible outside the company premises. However, the data available for this thesis has been crawled from publicly available data which exists as reports in various formats varying from plain text, semi-structured to structured. The crawled reports have been preprocessed using natural language processing techniques. Domain ontology has been used to guide the whole processes of clustering, and classification and a multiagent system have been integrated into the developed approach. Utilizing a multiagent system in the process allows for multiple perspectives to be incorporated into the process. These aspects are represented by independent agents who collaborate and negotiate to reach a consensus. The developed approach has been successfully applied to some publicly available gas and oil infrastructure hazard related data. The reported results may be used to issue recommendations to use certain safeguards to reduce the risk level in the processes.

Acknowledgments

Since I put the first step on this road leading to MSc in Computer Science I was highly confident that I will hit the target successfully. This is mainly attributed to two professors from the University of Calgary, namely Prof. Jon George Rokne and Prof. Reda Alhajj who I heartedly thank for all their support and encouragement. Prof. Rokne shared his long experience unconditionally and helped me shape my research vision and contributions. He has been very friendly and always available for consultation and discussion of research related matters. The material I learned from the courses of Prof. Reda Alhajj helped me a lot in my research methodology. He brought to my attention how various computing techniques could be applied in different domains including oil and gas industry. I would like to thank the Department of Computer Science for the facilities they provided to support my research. I want to thank Ayeesha AbdulWahed with whom I worked for a long time and from whom I learned a lot of things as well.

I also want to thank my family, my grandfather and grandmother, my dad, mom and uncle and aunt for their support and encouragement. Indeed, my grandfather and grandmother both deserve most of the credit because they envisioned my today's status since the day I was born. They drew my roadmap and worked hard to prepare and motivate me to head towards the target with firm steps. I am happy that I have been able to satisfy their expectations and turn their hopes into reality.

Table of Contents

List of Figures and Illustrations

List of Tables

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
PSID	Process Safety Incident Database
PHA	Process Hazard Analysis
SIL	Safety Integrity Level
MARS	Major Accident Reporting System
MAHB	Major Accident Hazard Bureau
CSB	Chemical Safety and Hazard Investigation Board
HAZOP	Hazard and Operability Study
RR	Risk Ranking
P&ID	Piping and Instrumentation Diagram
FMEA	Failure Mode and Effect Analysis
CBR	Case Based Reasoning
FCA	Formal Concept Analysis
HACCP	Hazard Analysis and Critical Control Points
NLP	Natural Language Processing
MAS	Multiagent System
KNN	K-nearest Neighbors Algorithm
SVM	Support Vector Machines
DT	Decision Trees
NB	Naïve Bayes
CSS	Cascading Style Sheets
CSV	Comma-separated Values
JADE	Java Agent Development Framework
ACL	Agent Communication Language
AID	Agent Identifier
AMS	Agent Management System
DF	Directory Facilitator
RMA	Remote Monitoring Agent
CNP	Contract Net Protocol

Chapter 1 Introduction

Over time, humans have gradually shifted from leading a simple life to a more luxury style of living made possible by the rapid developments in technology. New inventions have been quickly accepted and adapted in daily life from wheels to engines to computers and mobile devices, etc. As a result, a huge variety of systems exist to serve humanity in one way or the other. Each system has its own complications and risks which are generally ignored until they are faced with adverse outcomes. Some adverse events have minor effects and may be tolerated. However, others can lead to disastrous outcomes and hence should be avoided, if ever possible. Indeed, the fact that risk and associated accidents may occur is often ignored because avoidance does not come at no cost. The cost associated with risk avoidance and its frequency dictates how the decision makers proceed. And unfortunately, a hazard-free environment is likely not possible in real life. Thus, it is important to detect faulty items and fix them with the hope that the associated risks and consequences on individuals, society that could lead failure in a processing plant and economy can be avoided.

This chapter defines the problem tackled in this thesis and the motivation to developed the proposed solution. An overview of the proposed approach is also described together with the contributions. The last section of the chapter presents the organization of the rest of this document.

1.1 The Need to Handle Risk

A hazard can have various definitions and can be explained as a risk or danger. But, it is generally a potential damage, or harm of any source. Yet, this also raises another question, namely the need to clarify what a risk would mean. Risk may be defined as the probability or chance that a failure may occur whether major or minor. Almost every system in real life is subject to risk during its lifetime because it is almost impossible to predict the behavior of a given system. There is no guarantee that a system will proceed well from its start to termination without any possible failure.

In general, a failure once happens it may be associated with damage or causalities. Damage may affect an infrastructure, a system, the environment, the economy, a building, a region, a city, a country, etc. Both humans and animals may be listed under causalities which may encountered by taking risk. For instance, a person may be harmed or experience to an adverse health effect if exposed to a hazard. It may also apply to situations with property or equipment loss, or may lead to harmful effects on the environment [1]. Further, the uncontrolled release of radiation or a toxic chemical may have immediate short-term safety consequences, more protracted health impact, and much longer-term environmental impact. Events such as Chernobyl, may cause immediate deaths, and in the longer term, may lead to death from cancer. It may have a lasting environmental impact leading to birth defects, impacts on wildlife, etc. These are some consequences of risk which, in other words, may be expressed as a probability or likelihood of developing a disease or getting injured. Finally, hazard refers to agent's responsible (e.g., radiation and toxic chemical release) [2].

Some of the vital systems and processing infrastructures that have visible impact on humanity are oil and gas platforms, pipelines, refineries, etc. (hereafter platforms will be used to mean all oil and gas related infrastructures). There are serious risks associated with operating these platforms. Indeed, it cannot be ignored that major accidents mostly come with lots of irreversible consequences such as fatalities, damages and many other adverse events that happen to affect the environment, economy, society, etc. Most of the crucial incidents are associated with major explosions and release of dangerous chemicals that may lead to lots of damage to properties or even serious causalities, including unrecoverable injuries or death of some people. Affected people are generally employees present at the site of an incident, and in some cases, the events may even include people nearby who were just unlucky to be hit by the explosions or they were exposed to released chemicals or to fire.

To illustrate how dangerous and deadly the risk associated with certain accidents can be, it is worth mentioning the case of an accident described in DuPont Corporation Toxic Chemical Releases which caused an operator's death:

> "On January 23, there was a release of highly toxic phosgene, exposing a veteran operator at the DuPont facility in Belle, West Virginia and resulting in his death one day later. DuPont officials told the Chemical Safety Board that a braided steel hose connected to a one-ton capacity phosgene tank suddenly ruptured, releasing phosgene into the air. An operator who was exposed to the chemical was transported to the hospital, where he died the following day. The phosgene release followed two other accidents at the same plant in the same week, including an ongoing release of chloromethane from the plant's F3455 unit, which went undetected for several days, and a release from a spent sulfuric acid unit. The plant announced over the weekend that it would be shutting down a number of process units immediately for safety checks. The Chemical Safety Board is also investigating a November 2010 accident at the Dupont facility outside Buffalo, NY, that fatally injured one worker." [3].

Another severe incident is the one related to Arkema Inc. Chemical Plant Fire accident described as follows:

> "On August 29, 2017, flooding from Hurricane Harvey disabled the refrigeration system at the Arkema plant in Crosby, TX, which manufactures organic peroxides. The following day people within a 1.5-mile radius were evacuated. As the trailers increased in temperature the peroxides spontaneously combusted on August 31. Officials ignited the remaining trailers, on Sunday, September 3, 2017. The evacuation zone was lifted on September 4, 2017." [4].

This is where safety analysis defined as an analysis for creating a hazard-free productive workplace environment for all operators and employees starts to play an important role [5]. This requires controlling all processes associated with risks and hazards. A process is defined as an operation or a series of operations which are expected to cause a physical or chemical change in a substance or mixture of substances [6]. There are many different regulations and rules for process safety management and all the processes depend on capturing data related to critical instruments or items which may lead to hazard. Data may be captured as an unstructured report file and this brings up lots of problems with itself which is difficult to process. This is why semantic analysis needs to be implemented to assist professionals with data analysis requirements. There are many different ways to help those users, such as text mining and classification, data mining and its methods, etc. The aim of this work was to create an application that will give accurate results and could also be flexible with the data provided in the way of thinking like professionals related to this field.

1.2 Process Safety Incident Databases and Challenges

To illustrate what has been learned from major adverse events such as in Seveso, Flixborough, and Bhopal, professional analysts reported some details related to accidents by adapting the process safety management legislation. As noted by [7], “In the United States, regulations such as the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) and the Emergency Planning and Community Right-To-Know Act (EPCRA) require reporting when any facility releases more than a specified amount of a hazardous substance to the National Response Center”. The National Response Center keeps a large database which contains details about the reported incidents [8]. Similarly, in Europe, chemical processing industries are required to report major accidents or near misses to the Major Accident Reporting System (MARS) operated by the Major Accident Hazard Bureau (MAHB) [9]. The Chemical Safety and Hazard Investigation Board (CSB) has also made available to the public results of the investigation of accidents in the chemical and processing industries [10]. For example, the Process Safety Incident Database (PSID) allows collecting, tracking, and sharing of process safety incidents and experience among participating companies [11]. The Japan Science and Technology Agency (JST) maintains a database of accidents and failures since 2001 [12]. As of 2009, the database stores about 1160 records of incidents, including 333 records related to chemical substances, some of which are available in English.

1.3 Risk and Safety in Oil and Gas Platforms: Problem Definition and the Proposed Approach

There are many different public, private and member-based databases which are accessible through companies or by individuals who do safety analysis and process safety management. However, due to the different regulations and different kinds of rules applied when reporting incidents, it can be challenging to generalize the rules and gather all data from different reports required to extract the information they need. There are different kinds of algorithms proposed for different categories of problems to handle and analyze process safety incident data for events available from the Web. However, these algorithms have not realized the vital need for collecting data and turning it into valuable knowledge for effective decision making. This thesis focuses on finding effective and intelligent ways to help and ease the work of process safety engineers and analysts. This will be achieved by developing an approach to do safety analysis which integrates data mining and machine learning techniques for data extraction and analysis. Multiple agents have also been integrated in the process to model the situation in a more natural way. The rest of this section briefly describes the various components of the proposed approach.

1.3.1 Information Retrieval (IR).

One of the problems encountered in the existing literature is that every one of the databases mentioned above has a different information retrieval mechanism. This causes difficulties for engineers and/or analysts when they want to identify specific data related to a given problem, then parse the causes, consequences, injuries and environmental effects of the report and use them to do safety analysis, and finally apply the rules of process safety. The main source of this confusion is the fact that the documentation for the reports differ for each of the available databases.

To overcome this problem, a data crawler tool has been created that gathers and extracts all the attributes and properties of incidents from each of the reports. This way, professionals can reduce the time they need for looking at the reports in their effort to extract all the necessary information.

After building the database with all the extracted information, a user can perform a search with the help of keywords.

For the data collection, different kinds of sources are handled in different ways. The ExxonMobil Torrance Refinery incident will be used here to demonstrate the process. It has been taken from one of the U.S Chemical Safety Board documents and will be used to explain how they report the study node. The statements as taken from CSB reads as follows:

> "As a result of this incident, a near miss event occurred in the modified hydrofluoric acid (MHF) alkylation unit when explosion debris nearly hit tanks in close proximity to the ESP, each containing hydrofluoric acid (HF), water, hydrocarbons, and a chemical additive intended to reduce the amount of HF vaporized during a loss of containment event. HF is a highly toxic chemical that can seriously injure or cause death at a concentration of 30 parts per million (ppm). ExxonMobil resisted CSB requests for safety information pertaining to the potential release of HF in the event the tanks were struck by explosion debris. ExxonMobil continues to refuse to provide the CSB with information detailing safeguards to prevent or mitigate a release of HF."

This is an example that indicates event type, release major occurrence, release initiating event, and if there is an explosion. The above statement describes explosion major occurrence and/or explosion initiating event, or the safeguards applied (if applicable).

Multiple other attributes like the above mentioned ones can be extracted from these reports. Since the Chemical Safety Board explains everything in a very detailed way, however, only the key root causes, and their consequences, safeguards are extracted. This data can be found in the summary page, and hence there is no need for processing the complete files.

After having detailed discussion of the problem with some process safety/chemical engineers, it was decided that there was a lack of data collection similar to the data they had in their database. They also indicated that they needed to investigate as many incident reports as possible develop safety recommendation based on reported incidents. This will allow them to collect them all together and learn from the mistakes that happened earlier.

1.3.2 Multiagent Clustering Methodology.

Collecting all the extracted data and extracting knowledge which is assumed to be valuable for decision making is only one aspect of the problem. Unfortunately, this process does not help much when it comes to extracting knowledge from the unstructured text that has been collected. A classification problem arises here because it is important to classify all documents gathered from different databases. This is indeed a further challenge. There is a huge number of accident investigations recorded on many distinct oil and gas facilities. All these need to be considered in a comprehensive analysis. The more recorded data we consider, the better vision we will acquire, and hence the more solid guidance can be derived by safety professionals and communicated to decision makers in companies and government agencies.

As mentioned earlier, we need to identify the root causes and all consequences that may arise due to component failure whether immediately or gradually as a propagative effect. On most oil and gas platforms, process safety engineers and analysts are already doing this manually to determine the type of accident which occurred (this might sometimes include more than one type). However, this process is time consuming for them and it is highly subject to errors. To overcome this and to speed up the process with possible errors minimized, a multiagent based clustering algorithm has been proposed for the data scraped from documents. A domain specific ontology has been proposed who the target is to classify every study node in order to determine the risk level accordingly. A multiagent based clustering algorithm employs domain specific ontologies finds some appropriate clusters for data and document classification. Firstly, ontologies need to be built, yet the background knowledge was not sufficient. The research about the domain knowledge was therefore conducted with the help of engineers and professionals from the related field. Then multiagent ontologies were created to achieve the target clustering. This will be explained in detail later in Chapter 3. After syntactical and semantical extractions, the multiagent based clustering approach has been applied using the domain ontologies.

1.3.3 Supervised Learning and Recommendations.

As part of the solution developed to cope with risk and safety problems on oil and gas facilities, the other proposed algorithm underlines the recommendation system for safeguards. This system has been employed to reduce the risk level in an automated way by using keywords and by learning from the manually entered data. An automated process is preferred to the option of handling the situation manually and figuring out the outcome for each one of the data instances. The three terms study node, risk level, and safeguards are explained in Chapter 2 with associated detailed description of each. After determining which safeguard(s) should be applied to the investigated study node, some experiments have been conducted to realize the reduction of risk level. One other problem here is a categorization of safeguards.

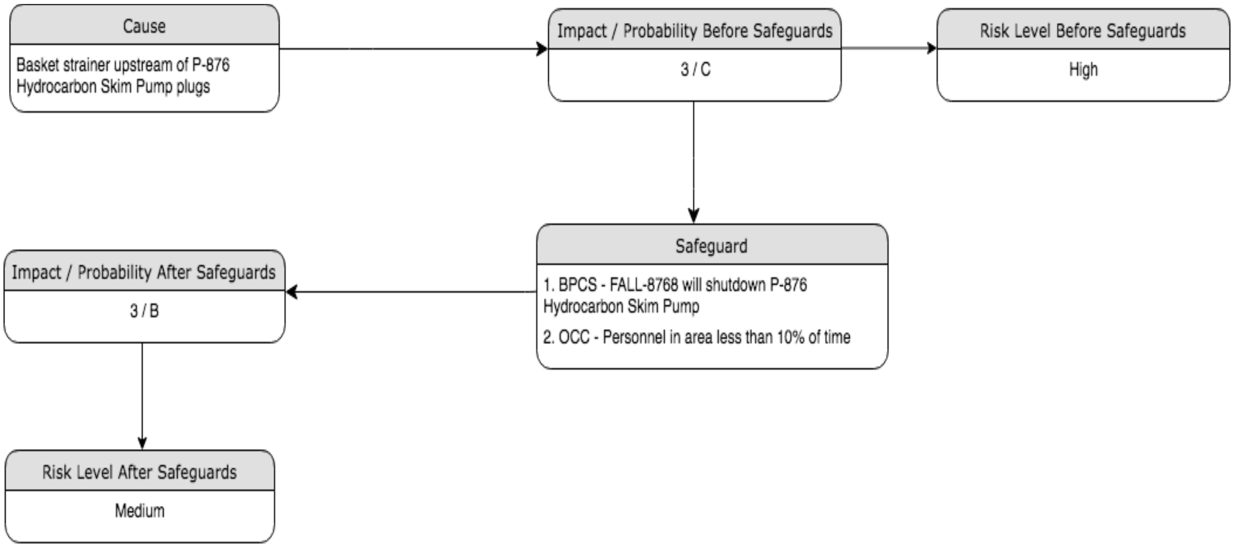
Cause
Basket strainer upstream of P-876 Hydrocarbon Skim Pump plugs
Impact / Probability Before Safeguards
3 / C
Risk Level Before Safeguards
High
Safeguard
1. BPCS - FALL-8768 will shutdown P-876 Hydrocarbon Skim Pump
2. OCC - Personnel in area less than 10% of time
Impact / Probability After Safeguards
3 / B
Risk Level After Safeguards
Medium

Figure 1.1 Reduction of the Risk Level with Applied Safeguards

In the available reports and online databases, descriptions about how safeguards are applied mostly in an unstructured text format. Nevertheless, this will not summarize the type of safeguards to be applied together with the given causes and consequences relatively. Categorization is needed here to determine risk level change with safeguards. The process is illustrated in the block diagram shown in Figure 1.1. This problem has been solved by text analysis using keyword index search, where keywords are normally pre-defined by professionals and engineers.

1.4 The Developed System at a Glance

The main idea of this thesis is to combine the components described above into a unified approach capable of aiding professionals in the oil and gas industry, with safety related problems and in particular those working on platforms, pipes, etc. The target is to reduce their workload of manually handling and processing a large amount of data to assess potential hazardous cases. Instead, we developed an automated system which collects, integrates and analyzes data available online to produce appropriate recommendations which may help in avoiding hazard as much as possible. It is crucial and vital role to help these fellows by learning from past incidents and giving them better recommendations based on the knowledge derived from the data collected from the Web in addition to locally available data. Collecting data from the Web is essential to enrich the data to be used in the analysis because local data may not cover all possible cases. Thus, it is highly beneficial and rewarding to benefit from the experience shared by others. Even it is important to reduce the time for giving recommendations to reduce risk level.

The system was developed in Java programming language with the help of external libraries. Some tools have been used for accuracy checking and for evaluating the results as well. Other tools were used to understand how the related HAZOP domain works, how it could be interpreted, and how it could be used as a guide for the development of the application leading to the working approach described in this thesis. With the help of the application for hazard and operability, the analysis procedure was accelerated. Clustering and classification methods have been implemented to make safety analysis in a semi-automatic way with a multiagent system to proceed faster and shorten the period for the whole process.

1.5 Contributions

The main contributions of the work described in this thesis may be enumerated as follows:

1. A crawler for collecting data from the web related to various aspects of risk and safety in the oil and gas HAZOP data
2. Integration of data from various sources so that it can be processed by an algorithm for more effective knowledge discovery. We process the collected data as a whole rather than processing data in pieces.

3. Clustering and classification of the unified data to highlight interesting discoveries which may better guide the decision-making process for facilities so that risk of hazardous events is reduced or eliminated, if possible.
4. A recommendation system for properly handling various components and instruments to better avoid associated hazard.

1.6 Outline of the Thesis

This thesis has been structured into five chapters. Following this introductory chapter, the second chapter covers the background and related work necessary to understand the content of this thesis. The methodology is described in Chapter 3 where all components of the proposed solution are presented in detail. Some experiments have been conducted to validate the proposed methodology. The data used in the experiments, the test environment, and the reported results have been all included in Chapter 4. Conclusions and future research directions are discussed in Chapter 5.

Chapter 2 Background and Related Work

The work presented in this thesis may be described as the development of a system which integrates data mining and machine learning techniques into a recommendation system directly applicable to risk and safety analysis in the gas and oil industry. It may also be tuned and adapted to serve other field, including health, homeland security, military, etc.

This chapter is dedicated to cover the necessary background and related work in order that the remainder of the thesis will be easier understood. First, the basics of data mining and machine learning techniques are covered. Then there is a discussion on how process safety analysis has been handled and then the terms used by professionals will be explained. This will lead to a better understanding of the data, what has been done, and how data could be manipulated by text analysis techniques for information extraction and knowledge discovery. The aim of this preliminary discussion is also to get an idea of how text mining would work on domain specific unstructured text. Finally, this chapter provides an overview of other works related to our approach described in this thesis. Studying these approaches and identifying their shortcomings has motivated for the approach described in this thesis. The conducted experiments and the results reported in Chapter 4 clearly highlight the advantage of our approach over other approaches described in the literature which tackles the same problem.

2.1 Background

2.1.1 Clustering, Classification and Multiagent System.

Clustering, classification and agents may be roughly described as learning techniques. Each technique has its own distinct approach to incorporate learning. That is each of the three has its own flavor and interpretation of learning. Clustering is an unsupervised learning approach and classification is a supervised learning approach. Agents are in general assumed to be autonomous self learning entities who try to simulate the behavior of some living bodies.

Clustering.

Clustering, e.g., [13, 14], may be considered as the process of distributing a set of objects or data instances into groups based on the similarity of their individual characteristics such that all similar objects fall in the same group which is called as cluster.

The number of clusters and the similarity measure used to decide on the destination cluster for each object are the two key distinctions of a clustering technique. The choice of which similarity measure to use in the process is highly problem specific. In general, characteristics of the objects or data to be clustered dictate whether a Euclidian or a Non-Euclidian based measure is to be employed. Roughly speaking, the former works best for numeric values while the latter is more applicable to non-numeric values. Even within each of the two categories there are various measures which may be used based on computation power available.

Most clustering techniques, like k-means, require explicitly specifying the number of clusters as an input. Others like DBScan expect some parameters like number of neighbors needed for an object to be considered as the core for a cluster, and a minimum degree of similarity between two objects to be considered neighbors so they join the same cluster. The values of these two parameters may be adjusted to produce different number of clusters. A third category takes on the similarity measure as input and produce a hierarchy of all possible clustering alternatives from one object per cluster to all objects in the same cluster. A specific result may be obtained by limiting the hierarchy at a certain level depending on how much details are desired for the output.

Classification.

Classification, e.g., [15], is a supervised learning approach where the number and characteristics of target classes are expected to be known as input. Labelled data is used to build a model capable of deciding on the destination class of a new object based on how its characteristics match one of the predefined classes.

For classification the available data is generally split into two disjoint subsets, one subset is used for training or model building and the other subset is used for testing the accuracy of the model in classifying unknown instances. Decision trees, support vector machines, Bayesian and neural networks, are some of the classification techniques that have been widely utilized for many applications.

Multiagent System.

An agent is defined as an entity capable of completing a certain task. It may be intelligent and possess capability to learn. Learning may be simple or advanced, and this determines the complexity of an agent model. An agent may learn only simple tasks or may be capable of learning from the history, from the environment, and from other agents. A good introduction to agents is provided in [16]. An agent may build its knowledge by experience. However, it is hard for a single agent to cover a complete knowledge domain. For instance, for a designer to build a house on his/her own, he/she is expected to be knowledgeable in plumbing, electrical wiring, heating/cooling, etc. Having one person to know all these skills is not possible in this era. Thus, multiple experts come together to complete a house. The same applies to agents who in practice simulate humans or animals and hence are expected to accomplish similar outcomes. Thus, having multiple agents involved in a process better simulates a real-life scenario. Agents negotiate to resolve conflicts and cooperate to achieve a common target. Their combined expertise and effort aim to have each agent concentrating only on its role within the group.

2.1.2 Process Safety.

Hazard and Operability Study.

A hazard and operability study (HAZOP) is a technique that allows engineers to see the overall system design for an individual facility and it studies how the system operates to highlight or pinpoint to potential unwanted hazards and/or operability issues on the incident/event. And it is a method performed by a team specialist to review a system architecture as part of their effort to figure out unforeseen problems. In general, it covers a list of records that generally summarize accidents. It is not concerned with solving issues, but rather concentrates on identifying them. In other words, it is a collection of possible breakdowns of the review process from the overall system design reports. The more comprehensive the list is, the better will be the knowledge acquired by professionals who will be responsible for fixing faulty parts.

HAZOP documents contain many terms. Here are some of the terms that have been used for the study:

Node: Nodes refer to components that should be investigated. A node could for example be a pipe or an equipment. There might be more than one issue associated with the same node, and this might be linked to different causes and consequences.

Table 2.1 Some of the Specific and General Parameters

Flow	Pressure
Temperature	Time
Composition	Safety
Phase	Level
Reaction	Agitation
Draining / Venting	Testing
Sources of Ignition	Accessibility / Visibility
Concerns and Comments	Sampling
Contamination	Maintenance
Siting	DCS failure

Parameter: Parameters are the identifying keywords of the processes. They may encapsulate explanations of a problem and indicate what has really happened. They can be specific or general and related to a certain issue. Some specific parameters are listed in Table 2.1.

Intention: Intention reflects the expectation of how the system should react/behave for the provided nodes.

Deviation: Deviation is a combined usage of parameters and guidewords together.

The usage of guidewords and parameters together might be more descriptive. For instance, “more or less” could be used for “higher or less” when the parameter is pressure or temperature. It is possible to have more guidewords associated with parameters, and their reasonable combinations will be used.

Table 2.2 Some Guidewords and Their Descriptions

Guideword	**Meaning**
No	Negation of the design intent
More	Quantitative increase
Less	Quantitative decrease
As Well As	Qualitative increase
Part Of	Qualitative decrease
Reverse	Logical opposite of the intent
Other Than	Complete substitution
General	Inclusive

Guidewords.

Guidewords are sets of words used with parameters to identify possible deviations in each node in the process. Table 2.2 shows some guidewords and their meanings.

Causes.

Causes are possible reasons for a deviation. They may be seen as kind of triggers based on which some action, like deviation, might occur. They could be due to some external or environmental factors, they may be initiated by equipment failures, or they may even result from human factors. For each of the causes, some special keywords, like equipment number, are mentioned rather than giving just a simple description. For example, “Draining of a Vessel” as a chosen cause is not sufficiently specific, and may be attributed to several initiating events such as:

- “LV-101 malfunctions open.”
- “Operator inadvertently drains vessel V-201 during normal operation.”

Table 2.3 Cause Categories

Cause Category	**Description**
Human Factor	When the event occurs from a human's activity or inactivity, depending on the situation and what is expected.
Equipment Failure	A failure caused with an instrument or equipment malfunction happened naturally without direct human failure.
Environmental or External Effect	All natural disasters and/or actions that are out of the facility's control may be classified in this category.
Unknown	If the cause is unknown or cannot be determined, it will be considered as unknown.

There are generalized cause categories that are determined by professionals. In this thesis, they are used relatively to the clustering process. Here is an illustrating example from the data collected for the causes; "During the restarting of the Acetylene hydrogenation reactors (after a shutdown caused by instrumentation malfunction) strong vibrations interested the drain valve in the boiler's candle. These vibrations caused the unscrewing of many flanges bolts of the structure allowing gas leakage." To see the outcome of the analysis for this study node, the category of this cause must be assigned first. Categories have been defined after careful consideration. Finally, some identified cause categories are listed in Table 2.3.

Table 2.4 Consequence Categories

Consequence Category	**Description**
Health & Safety	Mostly the effects of consequences impacted to the actual human life, as in fatalities or injuries and/or disabilities.
Environmental Impact	Chemical/toxic material's impacts to the environment can be defined in this category.
Economic Impact	Financial losses because of the incident, such as replacement of an equipment, or renovation of the total facility.

Consequences.

Consequences reflect what would happen as a result of an event caused by a deviation. It is the result of an action, such as the release of a toxic chemical. According to the HAZOP document format that has been used in this thesis, all Consequence descriptions must reference equipment by tag numbers and a semi-colon (;) can be used as an abbreviation in record keeping. The meaning or interpretation of this is "leading to" or "resulting in".

Here is an example which illustrates consequences, "Overpressure of V-110 Inlet Separator and inlet 300 ANSI piping; leaks or rupture; loss of containment of sour gas and HC liquid; fire or explosion; health and safety impact." From this example, it can be easily seen that there may be more than one consequence for a given deviation. Possible consequence categories for the clustering and classification tasks are given in Table 2.4.

Recommendations.

Recommendations are valuable for eliciting consequences resulting from deviations. They provide some information which is deemed necessary for preventing or reducing the outcome of the specific consequence. Recommendations may also help to identify more detailed discussions on the deviations linked to the consequences.

Here it is worth mentioning that HAZOP study is normally made in order to identifying a specific node and related deviations rather than fixing the original problem. This is where an automated solution to the identified problem could be useful.

Risk Assessment.

Risk Assessment (RA) is the process which quantifies, and calculation of the risk associated with some given severities. It also covers the estimation of event consequences and likelihoods of deviation. In a conservative set up, the ripple effect of a given cause and its associated consequence(s) may be considered as part of the process to avoid future surprises due to unforeseen side effects.

Risk Matrix.

The risk matrix indicates of the relationship between the frequency of events (likelihood) and the level of the consequence (severity) in a matrix format. Each entry in the matrix reflects an association between the frequency of an event and its level of severity.

Initiating Event Likelihood.

The likelihood of an event is the representation of an event's causes frequency. This can be recorded if it already exists, otherwise it should be determined by process safety engineers.

Severity.

Severity values are passed according to the company's given risk matrix. As in the Likelihood case, there can be different rules when it comes to determining these values. They might change from one company to another.

Table 2.5 An Example Risk Assessment Matrix

S L	1	2	3	4	5
A	I	I	II	II	III
B	I	I	II	III	IV
C	I	II	III	IV	IV
D	II	III	III	IV	IV
E	III	III	IV	IV	IV

A risk level is assigned to each of the given severity values and its corresponding likelihood. Table 2.5 provides an illustrating example of a risk level matrix which includes risk ranking and their associated priorities. Table 2.6 includes a list of risk ranking for risk assessment.

Table 2.6 Risk Ranking for Risk Assessment

RANK	MEANING
IV	High, unacceptable risk. Action required immediately
III	Reduce risk to the lowest level possible
II	Operation can be executed after necessary actions are taken
I	Acceptable risk, no action required

Safeguard.

Equipment repairs, and preventive systems maintenance are aimed at the reduction of hazard or for stopping the current situation from getting worse. These remedial actions are taken based on some safeguard recommendation. An equipment, or a human factor may be counted as a safeguard.

Safeguards generally assist to reduce the likelihood (initiating event frequency) and/or severity values for potential hazards and eventually, this will lead to a reduction of the overall risk level. Some safeguard categories together with their corresponding recommendations are listed in Table 2.7. Safeguards may be divided into the following three parts:

- Making sure it is possible to prevent the distribution or emission of a level 1 severity release of hazardous or flammable material
- Determining and providing sooner precaution warning for the likelihood of a release of hazardous or flammable material
- Systems that may highlight consequence(s) associated with the release of hazardous or flammable material

Table 2.7 Safeguard/Recommendation Categories

Safeguard/ Recommendation Categories	**Description**
MEC	Mechanical - Active Mechanical Devices, e.g., Clean Service PSVs, PVRVs, PSEs
MEC-P	Mechanical (Passive Mechanical Device, e.g., dykes, berms, fire blast walls, restriction orifices, mechanical stops etc.)
BPCS	Alarm with Operator Action through BPCS (primary panel system)
BPCS-T	AutomatedShutdown/Trip/Permissive through BPCS that brings the process to a safe state (primary panel system)
BPCS-C	Automated Controller through BPCS that brings the process to its normal operating state (primary panel system)
SIS	Safety Instrumented System
Other LS	Local Alarm - An alarm safeguard that doesn't have status back to the primary panel system (e.g., LEL alarm on a build with horn and blinking lights. No communication back to Primary Panel System). Or Alarm with field panel, but no communication back to Primary Panel System.
Other LS-T	Local Automated Shutdown/Trip/Permissive - An action safeguard that doesn't have status back to the primary panel system
Round	Operator Rounds (e.g. a "normal operation" routine action, e.g., monitoring level in chemical totes, monthly observation of sand builds up in pig receivers, etc.)

Other LS	Other logic solver, non-shutdown action - An alarm safeguard that links back to the primary panel system, but whose safeguarding function will still operate as designed if the communications link to the primary panel system fails (e.g., SCADA alarm for high pressure)
Other LS-T	Other logic solver, shutdown action - An action safeguard which links back to the primary panel system, but whose safeguarding function will still operate as designed if the communications link to the primary panel system fails (SCADA shutdown for flow path permissive)
Other LS-C	Other logic solver, controller action - An action safeguard that links back to the primary panel system, but whose safeguarding function will still operate as designed if the communications link to the primary panel system fails (e.g., BMS excess oxygen control system)
PRO	Operating Procedure (a "by request" action that is dependent on operator action, e.g., procedure for operator to open manual valve prior to start-up of a pump)
OCC	Occupancy Modifier
PM	Preventative Maintenance (includes inspections, integrity, maintenance procedures, maintenance based chemical injection, such as corrosion inhibitor etc.)
Other	Anything that doesn’t fit the above categories (e.g., Inherent design, PPE, EHT, other modifiers)
Design Review	Recommendations only: A design review and potential modification to the design is required

Hence, HAZOP documents differ from one company to another, and even between professionals working for the same company. There is, unfortunately, no generalized standards when it comes to documenting adverse events. Even within each company or for individual uses different software might be used to enter the data. This leads to different identifications for the same event which makes it hard to understand the documentation for someone who is not familiar with the topic. Thus, the first step towards more successful handling of HAZOP documents is to enforce some standards which will allow all parties to speak the same language and hence come to a common consensus faster. In other words, all data should be gathered in common concise format, should be self-descriptive and as complete as possible, and should be equally applicable for each of the nodes of an incident.

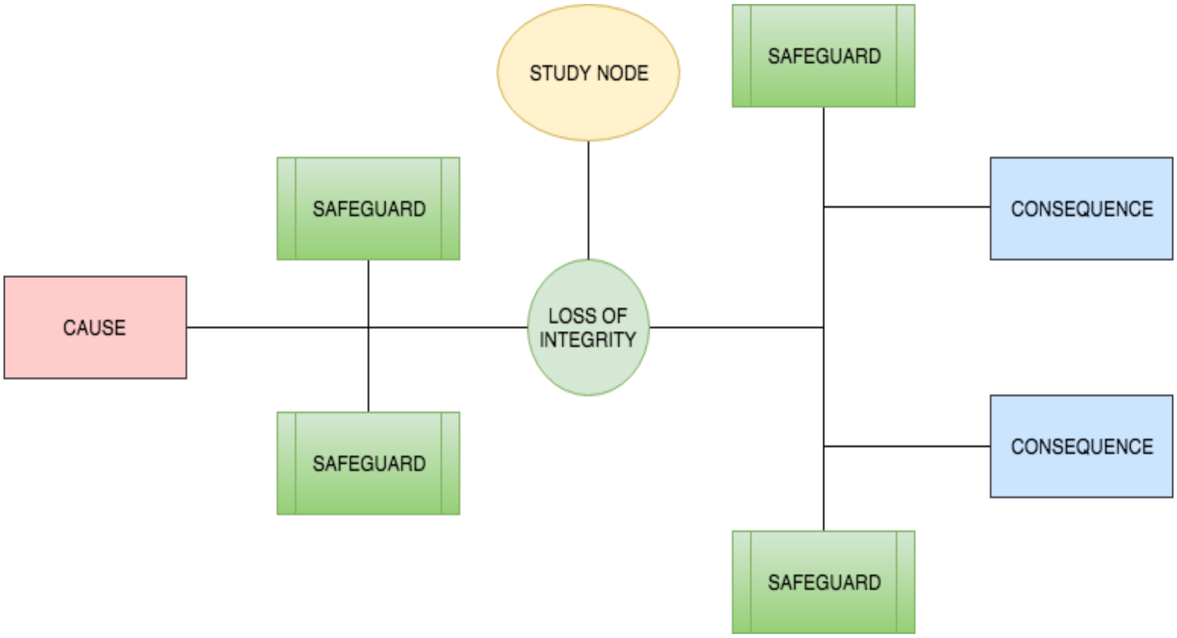

Figure 2.1 A Simple Bowtie Representation of a HAZOP Analysis

Each event might have a cause, but for some events this cause might not be clear. There are also documented consequences which includes recommendations as well. Initiating event likelihood and severity might not be documented; and there could be times when a professional need to determine them. This is also applicable to safeguards as well. This is the main reason why safeguards and risk level determined by severities and likelihoods are not always included in the data collected from various sources. A graphical schema with the bowtie HAZOP is shown in Figure 2.1, where the flow of control is clearly realized.

Process Hazard Analysis (PHA) covers the evaluation process for potential hazards. Hazard and Operability Analysis can be considered as one of them. Process safety engineers usually use a software to conduct their research and analysis for creating a PHA file. Their attributes and instances will be discussed in detail in the data collection section.

Finally, it is worth keeping in mind that this was just a brief introduction to domain knowledge. There are many other definitions, terms, and factors that are included in HAZOP analysis. Some of these are P&ID, Quantitative Risk Assessment, Re-HAZOP, Revalidation, Safety Integrity Level (SIL), Single Point Failure, Worst Credible Consequence, Bowtie, etc. In this thesis, we will use the concepts (features) relevant for the conducted study.

2.2 Related Work

There exist in the literature many evolutionary research efforts and improvements on process safety incident analysis. Some of the ideas and thoughts described in the literature have been adopted into the research approach described in this thesis. The remainder of this section covers related research about process hazard evaluation.

Thapa [17] wrote a report where he tried to highlight the importance of reducing risk to avoid hazard. He tried to describe best practices which should be followed by the industry in order to manage risk. Rodhi et al. [18] discussed risk factors and risk assessment techniques which can help development of programs. The authors realized the complexity of handling risk factors and emphasized the need for machine learning based techniques to cope with the problem. Mearns and Flin [19] reported on studies which covered a review of the psychology of offshore workers on UK and Norwegian installations. The author concentrated on the need to deal with environmental and socio-organizational factors which may affect risk perception and attitudes to safety, and ultimately risk-taking behavior and accident involvement.

The literature includes details related to a number of methodologies that discuss the automation of process hazard analysis. Some existing process hazard analysis (PHA) techniques are: Hazard and Operability (HAZOP) analysis and Failure Mode and Effect Analysis (FMEA), Checklist, What If?, What If?/Checklist, etc. Automated and computerized algorithms underlying these methodologies will be discussed in the sequel.

Daramola et al. [20] developed KROSA. It is a framework which has a modular architecture that builds on the integration of Natural Language Processing (NLP), Case Based Reasoning (CBR), and ontology. The software system developed by Daramola et al. [20] works as follows. It takes a source document and performs a semantic case-based risk analysis (study node recommendation, knowledge retrieval, retention, and adaption) using domain ontologies with the help of a Hazard and Operability (HAZOP) ontology, and Failure Mode and an Effect Analysis (FMEA) ontology.

Daramola et al. also described the ontology library and natural language processing for their semi-automatic domain ontologies.

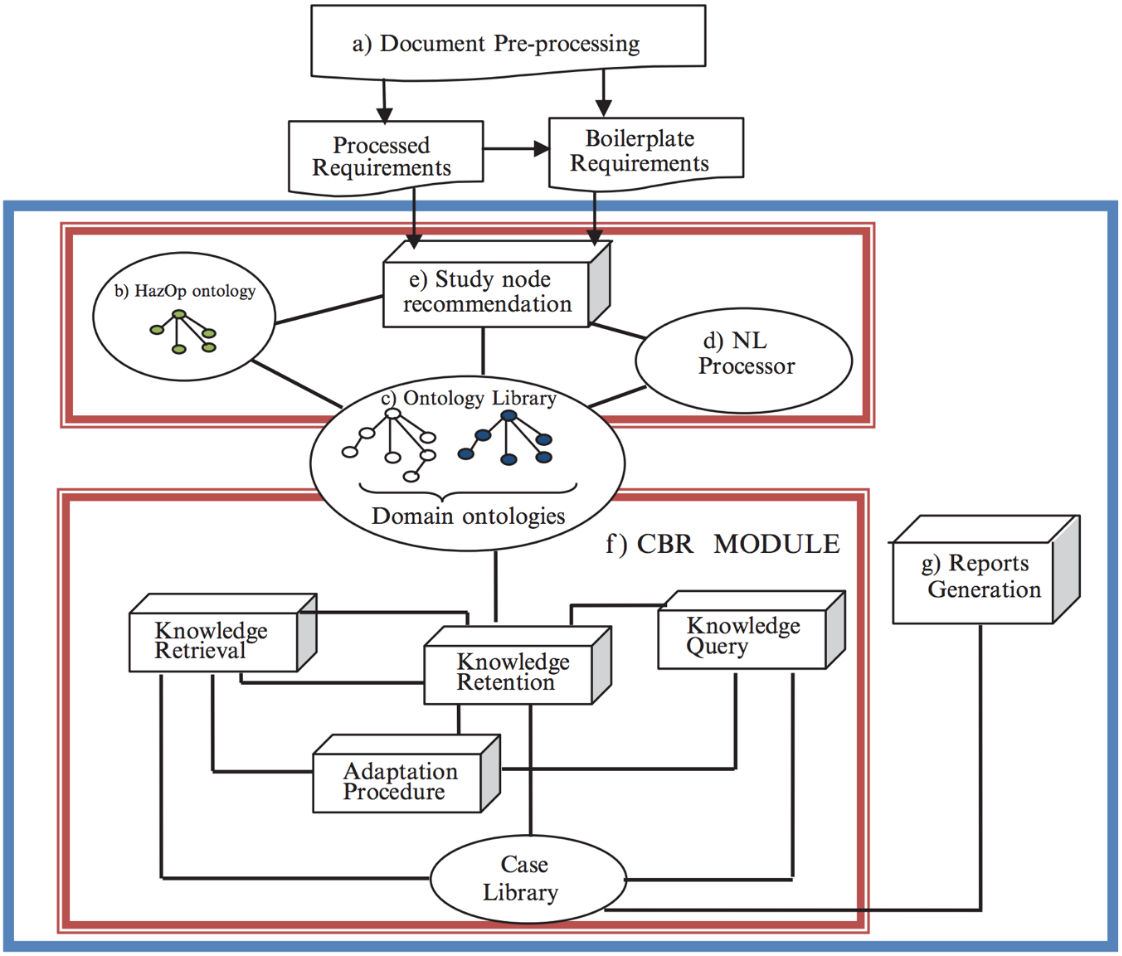

Figure 2.2 Conceptual Framework for Semantic Case-Based Safety Analysis (Daramola et al.)

As depicted in the block diagram shown in Figure 2.2, the process which generates the semi-automatic domain ontology of Daramola et al. [20] works as follows. It first takes a pre-processed document (PD). Second, it does term extraction (TE). Third, concept identification (CI) is accomplished with the verification of professionals. Finally, the ontology is created (OC) automatically based on results relationship mapping (RM). This is one of the few research efforts reported on this subject. Some other domain related articles helped to understand the HAZOP structure model.

Narapan, Unchalee and Thongchai [21] described a systematic formulation for HAZOP analysis based on a structural model. Digraph technique is explained for computer-based HAZOP analysis. Here, they want to generate the cause and effect relationship with a HAZOP based structural model. Their proposed methodology contains the following five stages.

- Selection of the unit from the library and defining parameters
- HAZOP digraph model (HDM) for the selected unit
- Formulation of HAZOP structural model (HSM) from HDM
- Implementation of Cause and Effect Matrix (CEM) prototype
- Interpretation of CEM into HAZOP analysis

There are many other HAZOP analysis approaches which are based on Fault Tree Analysis which is a very common approach. Guo and Kang [22] developed one of the main methods that use Fault Tree Analysis to identify potential hazardous events happening in chemical plants. They employed generic HAZOP analysis to determine causes and effects. Then, they showed how hazard flows from causes to consequences within the hazard scenario model. They presented a dynamic fault tree analysis simplifying quantitative calculation with a first binary decision diagram. Finally, a Markov chain approach was applied to subtrees. Afterward, they compared the occurrence probability of the top-level event with the occurrence probability of each of the other events. To summarize, at the end conventional HAZOP analysis was extended with a dynamic fault tree approach.

Rossing, et al. [23] proposed a methodology which is based on a new functional model rather than conventional HAZOP analysis. They used Multilevel flow modeling (MFM) to improve on the traditional model. Further, Rossing, et al. [23] state the following: “MFM combines the means-end dimension with the whole-part dimension, to describe functions of the process under study and enable modeling at different abstraction levels.” They also mentioned that this functional method will enable cause and consequence analysis. In addition, it will help in identifying potential hazard with a computer-based reasoning tool.

Cui et al. [24] developed a framework which combines Hazard and Operability Study (HAZOP), Layer of Protection Analysis (LOPA), Safety Requirements Specification (SRS), and Safety Integrity Level (SIL) systems to reduce the amount of incomplete PHA studies. They integrated all these methods into their software system, called HASILT, to handle new cases by employing case-based reasoning (CBR) based on old cases.

The integration framework works as depicted in the flowchart shown in Figure 2.3. Advanced HAZOP study is converted into LOPA worksheets. It exports LOPA into SRS and specifies it. It performs SIL validation with SRS design. It checks if SILs meet the requirements of SRSs.

Zhang, et al. [25] developed an approach for building information modeling and performing ontology-based job hazard analysis. They integrated ontologies to automate a procedure for job hazard analysis. They divide the procedure into four classes: Task, Activity, Job-Step and Potential-Hazard.

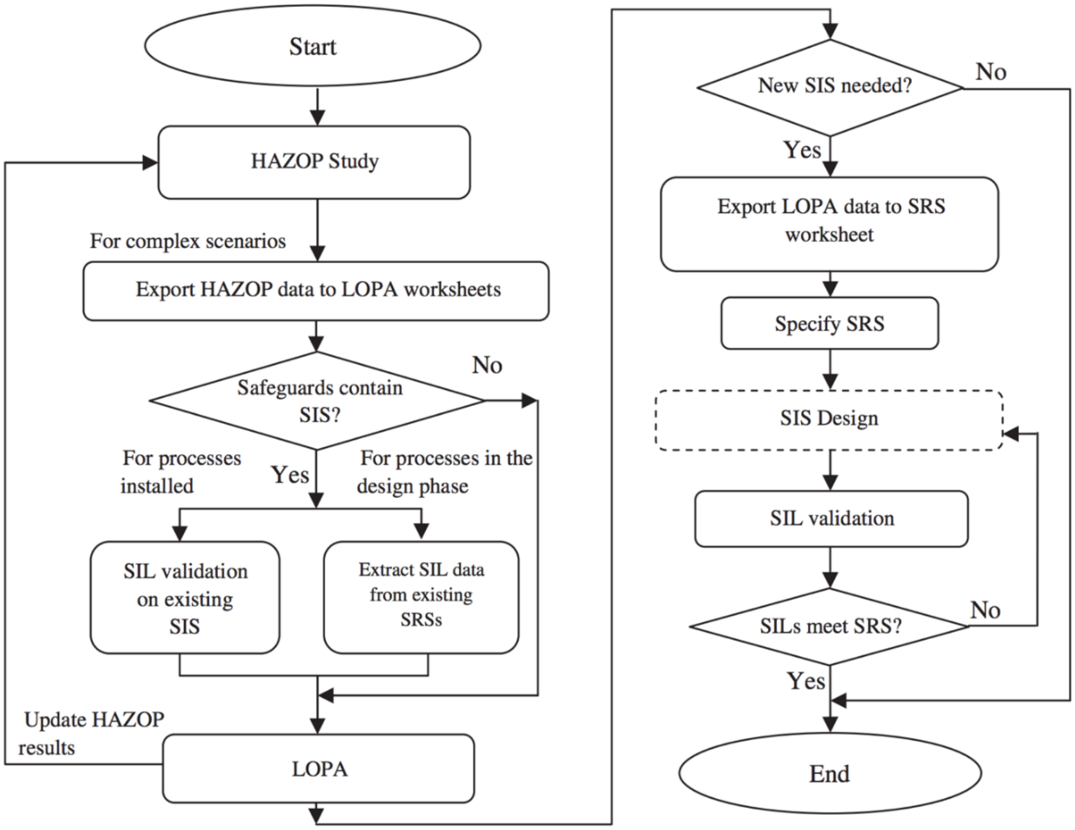

Figure 2.3 Integration Framework (Cui et al.)

OWL based ontologies and XML based recommendation procedures were integrated into the process to accomplish automated job hazard analysis. They manually selected tasks and activities from the ontology and used them in the procedure to get potential hazard and recommendations. The block diagram of the framework developed by Zhang, et al. [25] is shown in Figure 2.4.

Batres, et al. [26] discussed ontology usage to improve acquiring incident information on the study node. They mentioned that ISO 15926 ontology provides relationships such as oil and gas facility plants' operations, process behavior, plant equipment, chemical processes, batch recipes, and engineering diagrams.

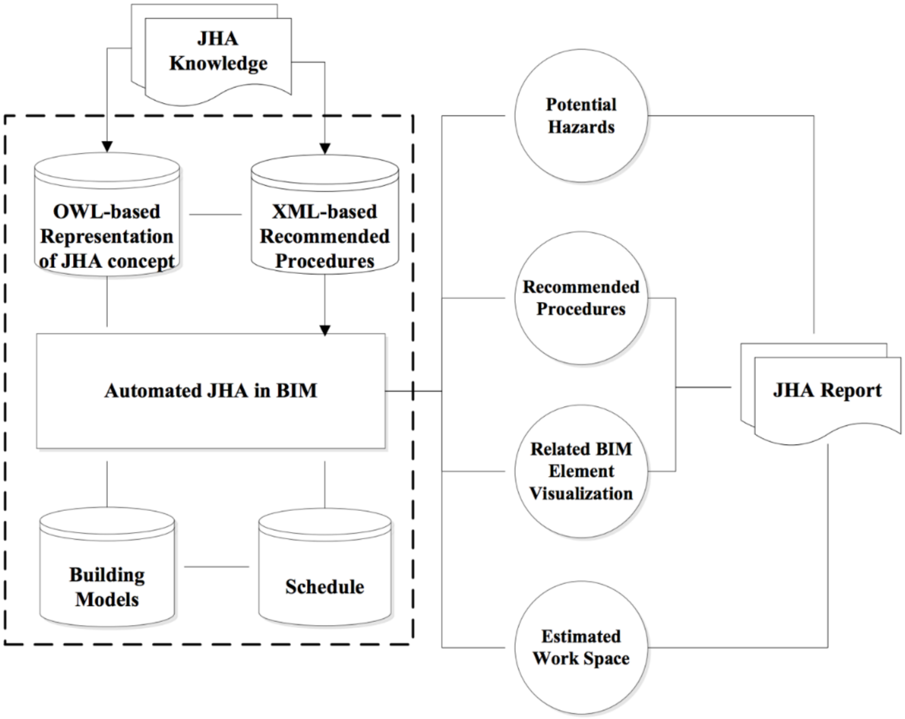

Figure 2.4 The Proposed Framework for Implementing Automated JHA (Zhang et al.)

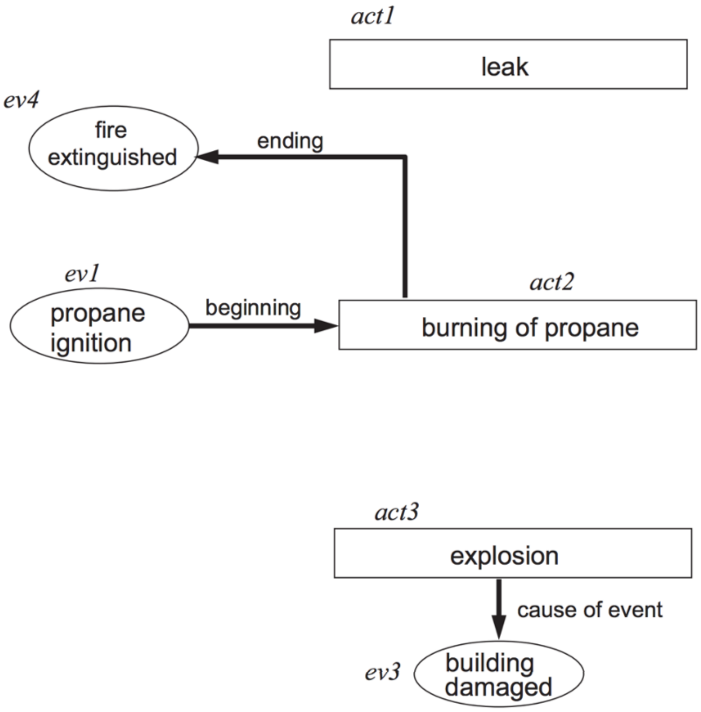

Figure 2.5 Main Activities and Associated Events with ARC (Batres et al.)

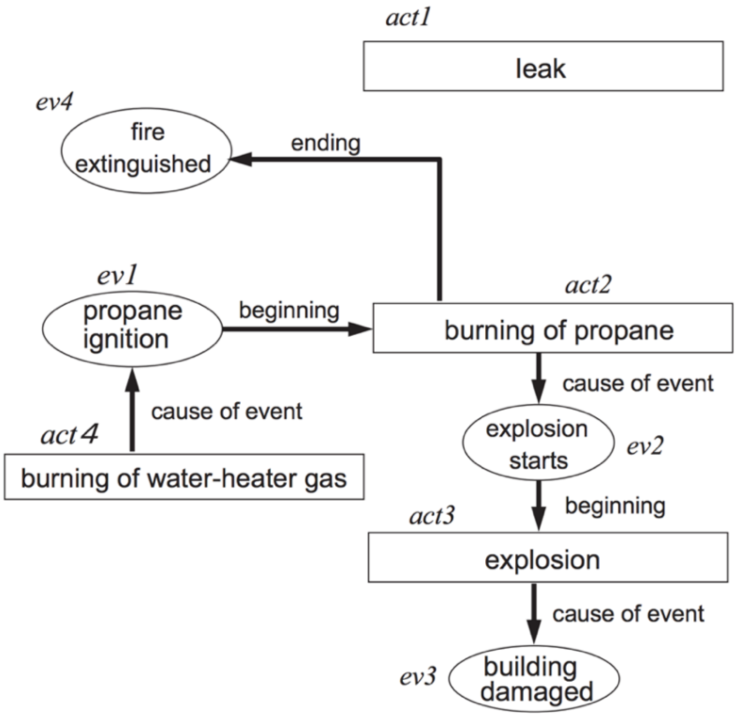

Figure 2.6 Completing the Causality with Additional Activities and Events (Batres et al.)

According to ISO 15926, each item (ontology component) has two classes; “possible_individual” and “abstract_object”. Batres, et al. [26] discussed an illustrating example in their article. It is an explosion and fire accident for incident identification. They also explained how the ontology works as depicted in Figure 2.5 and Figure 2.6. They used a graphical tool, called ARC, to show problem scenarios and schemas with an editor to help users.

The illustrating example given by Batres, et al. [26] has been articulated as follows: “An explosion occurred at a bottling plant injuring six workers. The incident occurred due to leak on a 1100-gallon tank containing propane, which is tough to have been ignited by a water heater. The fire was extinguished in forty-five minutes. Nearby buildings within half a mile were damaged by the blast.”

Batres, et al. [26] aimed to enhance knowledge retrieval using specifically ISO 15926 as an upper ontology which is more comprehensive. The idea underlying the ontology covered here is one of the approaches that influenced our semantic analysis. However, the process has been integrated in a completely different way.

There are classifications in the ISO 15926 ontology, e.g., “Activity class” is for an event which has a beginning and an end time. An event is a class for activities that are happening instantly. Physical objects represent equipment material or activities which have beginning and end time. Casual relations are for “cause_of_event”, they represent events caused by some activities which might lead to another event as well. The beginning relation will be the start time of an activity. Ending relation refers to the end of an activity. Participation relations are specific to objects that are affected by the considered activities. The mereological relation is an association which represents a part-whole relation for an object. Containment relation is a sub-relation representing other objects. Location relation is another mereological relation for the location of an object. And lastly, topological relation reflects the connectivity between objects.

Some of the achievements presented here are not related just to automated HAZOP analysis, but they also help to understand the meaning of the data. Another tool relevant for hazard and risk analysis is a knowledge-based expert system developed by Rahman et al. [27]. The approach may be considered as a fault propagation algorithm. It enables users to update knowledge with the provided GUI. The dynamic knowledge-based method helps, in addition to fault propagation algorithm, to identify all causes and consequences of each study node to all downstream equipment for process performance.

For the last part of the related work section, the PHASuite tools by Zhao et al. [28] will be covered. The problem divided into two parts. The first part is a knowledge engineering framework and the second part is dedicated to software development and a case study. They explain how the knowledge engineering framework could be created. The operation and equipment levels are represented in the two parts of this framework. Colored petri nets representation of information sharing is handled with ontology-based information. Knowledge management is handled with case-based reasoning.

All the solutions described in the literature and briefly covered above somehow help users either for making data on adverse events manageable and configurable or for knowledge acquisition. In some cases, it is possible to achieve both together. Ontologies are used in most of the cases and they play an important role in the conversion of data into recognizable information claimed into appropriate classes. They all provide case studies to show how their proposed algorithms work; they also highlight their efficiencies.

The approach described in this thesis will tackle the creation of domain ontologies in a different manner. A multiagent mechanism will be used to define ontologies in Chapter 3. Case Based Reasoning (CBR) is an approach that has been used widely for safety analysis. It is a way of understanding a new problem by re-using old experience [29, 30]. Similarity measures have been evaluated for different types of cases. Another approach used for assisting in generating domain ontologies is Formal Concept Analysis (FCA). It is a mathematical approach used for heterogeneous data analysis [31, 32]. Consequently, it is a good way for the integration of ontologies. Fu et al. [33] described in detail how to use FCA approach for domain ontologies. They clarified well how it can help to develop better ontology.

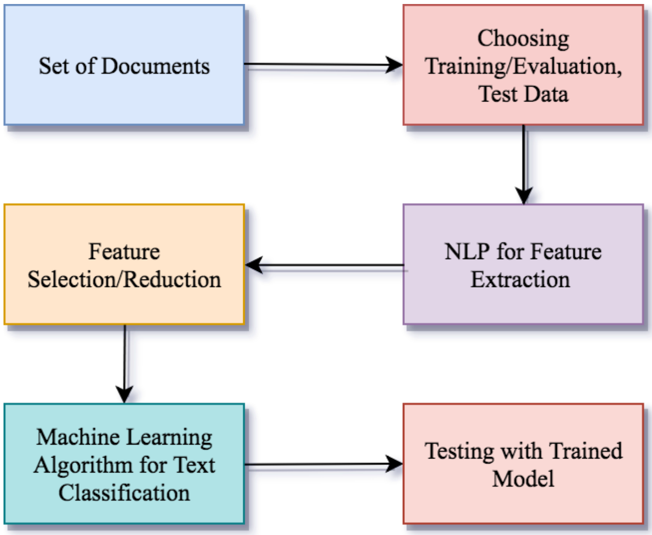
Set of Documents
Choosing Training/Evaluation, Test Data
NLP for Feature Extraction
Feature Selection/Reduction
Machine Learning Algorithm for Text Classification
Testing with Trained Model

Figure 2.7 Text Classification Procedure

Concerning the classification of unstructured documents, there are some methodologies which have been developed for different purposes. These methodologies to cluster and create templates for car insurance documents in a semi-automatic way [34]. They depend on an ontology to create a semantic network for each sentence of the document. Then they cluster them in the old-fashioned statistical way. Finally, the research described in [35] used agents for the analysis of unstructured text. The authors used two main agents, namely rule and instance, for the analysis. They also have a controller agent for the termination.

The work described in [36] used agents for clustering method. By integrating agents into k-means and KNN clustering. They checked the efficiency of their algorithms. The algorithm was started with individuals as separate clusters [36]. They also explained how JADE's framework has been used. The work described in [37] discussed knowledge representation for dynamically updated data. Clustering happens in a changing environment with accurate findings of the vector space model and by getting their cosine similarity. They also used trial methods to determine system effectiveness with new data.

When it comes to supervised text classification, there exist many proposed solutions which are capable of handling a variety of domains, such as spam detection, emotion analysis, label categorization, prediction, among others. As depicted in Figure 2.7, the procedure works as follows. It starts by data gathering and preprocessing. Second, it concentrates on extracting features using natural language processing, and then it selects the necessary features. The process continues by applying a suitable learning method considering the given features. They learn the classes from the training data. They then use the classification engine on the test (unclassified) data to get the actual class for each instance.

The most commonly used classification techniques include Support Vector Machines (SVM), Naïve Bayes Classifier, K-Nearest Neighbor, Latent Semantic Indexing, Decision Trees (ID3, C4.5), etc. One of the relevant classification techniques which integrated a multi-agent system is described in [38]. It is dedicated to scientific data with KNN learning method.

Naïve Bayes MAS has been applied on news articles and RCV1-v2 datasets, e.g., [39, 40]. The work described in [41] used agents to build a recommendation system by collecting information about web pages visited earlier. It suggests new pages based on historical data analysis.

Chapter 3 The Methodology and the Proposed Solution

Ad hoc handling of situations which may lead to disaster is not appropriate in general and should be avoided. A systematic approach should be adopted in help preventing disasters or at least to provide some guidance for taking some precautions which will help in reducing the often unpleasant consequences of a hazardous event. A methodology has been developed in this thesis aiming to accomplish this goal. The rest of this chapter covers the problem definition and various aspects of the methodology proposed, from data collection to knowledge discovery leading to recommendations.

3.1 Problem Definition

Most of the tragic and devastating disasters related to oil and gas facilities have some underlying reason(s) which could be an explosion, a dangerous chemical release after adverse events. This is when Process Hazard Analysis (PHA) is needed for the determination of the root causes, and the establishment of recommendations done by professionals using a combination of guidewords and parameters. HAZOP analysis is one type of PHA and will be used here for domain specific research. HAZOP basically covers the usage of causes and consequences for a study node and suggesting possible deviations based on existing safeguards and recommendations. All terms discussed here have been explained in Chapter 2.

The HAZOP process is handled by process safety engineers and related professionals from the same field. They use several software systems to help documenting an incident report based on the rules of HAZOP. Yet, these tools only enable entering the data manually. They are not capable of easing or reducing the worth of engineers' with interpreting the data in any way.

A careful investigation of existing automated systems for HAZOP analysis by process safety engineers revealed the conclusion that these systems are not well satisfying the expectations of domain experts who care the most for having the best possible systems. Safety analysts focus more because they are the first to be responsible by any hazardous event. Thus, they need a system characterized by fast information retrieval and informative summarization of incidents for appropriate HAZOP analysis leading to timely recommendations.

Rather than having a simple word processing tool which aids data collection for given attributes of incidents, it is necessary to have concise interpretation of data, make it (re)usable, and add meaning to each incident. Unfortunately, existing tools are either semi-automatic or only concentrate on some parts of the data, i.e., they mostly focus only on specific sections of HAZOP analysis. This is unacceptable because it is not guaranteed to have the target information only in the analyzed sections. Hence comprehensive data analysis should be the target of any system to be accepted and willingly used by safety engineers and domain experts.

The documented data is mostly unstructured text and cannot be manipulated by computers in a straightforward manner. This raises the need to employ data mining and machine learning techniques for effective and informative analysis of unstructured text. Intelligent data analysis will reveal valuable knowledge which is mostly implicitly present in data and cannot be retrieved by traditional information retrieval and query processing mechanisms. The outcome will be used by safety engineers and other domain experts who will otherwise find it almost impossible to extract same knowledge by manually processing the same data. This has been addressed in this thesis. A new approach has been proposed to enhance (re)usability of data, to speed up the process, and to extract the knowledge needed for informative and timely decision making by professionals, analysts, and engineers.

Information retrieval is another problem faced by domain experts because many different resources are available online. These documents are either in different formats or reflect incompatible process hazard analysis rules because every organization may have its own rules. For the given problem statement, semantic analysis is needed for unstructured text to assign specific meaning to data instances. Further, clustering and classification algorithms have been used to determine target categories by clustering, and then analyze the whole process safety incident report files to classify them into their appropriate categories. In addition, the same reports were supposed to be converted into HAZOP structures. Domain ontologies have been used in the identification process since domain knowledge was not adequately available and the process could not be done fast and automatically. In addition to the utilized data mining techniques, a multiagent system has been employed in the approach not only to ease the problem-solving process, but also to speed up the processing of data.

This chapter covers data collection and preprocessing, the proposed methodology, the techniques and algorithms applied to the collected data. The following anonymous illustrating example gives an idea about the issues that have been encountered. This specific incident is described as follows:

> “In the electrical network of the installation, there were works in execution. Then the safety fuse wire fused. Thereby the shutoff valve (safety closed) which was mounted in the output of a gasometer jacket towards the collector over the piston-compressor, closed. The gasometer jacket also in the future will be filled with electrolytic hydrogen.

> The hydrogen control valve will be closed for reached peak load in the storage vessel. In the condenser, aspiration pipe formed an under pressure. In the gas, aspiration line was installed a water trap, whose drain had a plastic tube stopped that in turn was submerged in the bottom of a plastic tube filled with water. For the under pressure, the water was aspirated out of the plastic tube, but deposits remained in the lower parts. Due to an under-pressure protection, the mean pressure condenser failed. Now the back-flowing air arrived at the high-pressure condenser. The hydrogen formed together with the air an explosive gas which auto-ignited. The plant keepers, who were in the near building, were seriously injured by an iron pieces breaking into the building. The built with brick building assembly, two pressure vessels, pipes and other parts of the plant were largely damaged."

Here, the incident description reveals a human failure (cause) as the main reason for the hazardous event. The consequence reported two injuries, who were present on site. In the full report, the applied safeguards can be found. One other issue here is knowledge retrieval, hydrogen-control-condenser keywords would return this result. However, there are some other incidents not related to the given guidewords and parameters. Therefore, it is not necessary that querying some keywords would return the correct results.

Conducting such a complex analysis just from a text report was another challenge. Many data sources have complicated implementation and they do not address the causes, the consequences or the recommendations. Hence, domain knowledge was inadequate. Hence, ontologies were used to overcome these problems.

3.2 Data Collection and Pre-processing

As mentioned earlier, there are distinct resources for the process safety incident data to be used in the analysis. Data collection is handled with a scraper. For public HAZOP databases, a data crawler, Scrapy [42], was used. This open source framework helped to crawl data from online databases by employing Spiders. Unfortunately, since each process incident database website has a different format and belongs to a different organization, their corresponding rules are mostly distinct. Thus, dedicated Spider classes have been created for each website. The extracted data feed is in JSON format which is then converted into CSV format. For the selectors, CSS or Scrapy selectors are available for XPATH. The following structured public databases were used for data scraping;

- ProcessNet [43]
- The Center for Chemical Process Safety (CCPS) [44]
- ARIA - Analyse, Recherche et Informations sur les Accidents [45]
- eMARS - Major Accident Reporting System [46]
- NTSB – Aviation Accident Database & Synopses [47]
- JST (Japan Science and Technology Agency) Failure Knowledge Database [48]

Since the structured data was gathered with the help of a data scraper, it does not come ready to use as in a structured way. For text columns, there are some HTML tags that should be clean. Also, not all data sources clearly and explicitly specify names of causes, consequences, and the given recommendations in the same way. This case imposes the need for the cleaning procedure to be employed in order to filter the data from all unexpected and undesired content.

Unfortunately, not all databases come often in structured form and available for direct download. There are some process incident databases which provide full-text reports to describe cases to be utilized in a conducted study. In these cases, identifying the root causes and other properties of HAZOP is not very feasible without extra effort. Here, the role of ontology becomes important to guide the classification of these documents into the expected format. The following databases include full-text reports that have been used for the study described in this thesis:

- U.S. Chemical Safety Board (CSB) [49]
- FACTS - Failure and Accident Technical Information System [50]
- IRTAD - International Road Traffic and Accident Database [51]

In this thesis, only important attributes have been taken into consideration rather than focusing on all HAZOP properties. These properties are causes, consequences, safeguards, recommendations, severity, and likelihood values for risk level(s). Semantic analysis and suitable data mining techniques have been applied to the data to give recommendations.

3.3 A Multiagent System for Ontology Development

One of the most important usages of Multiagent Systems in this thesis is to achieve the different goals and handle the coordination between them. A multiagent system will be used especially for the ontologies since they all describe specific domains. This allows ontologies to create their own agents so that they can represent accurately what they aim for. Multiagent Systems would also provide a good environment for parallel computing where different procedures are handled by dedicated agents. This will lead to better efficiency. The other benefit of multiagent systems is they provide the ability to add a new domain ontology to the system without modification of others.

For the work discussed in this thesis, each of the subtasks will be handled by the corresponding agent(s). Some of these agents are aiming for locating and retrieving the information and some for the classification of unstructured data. Ontology generation and actions performed by agents will be described in the sequel.

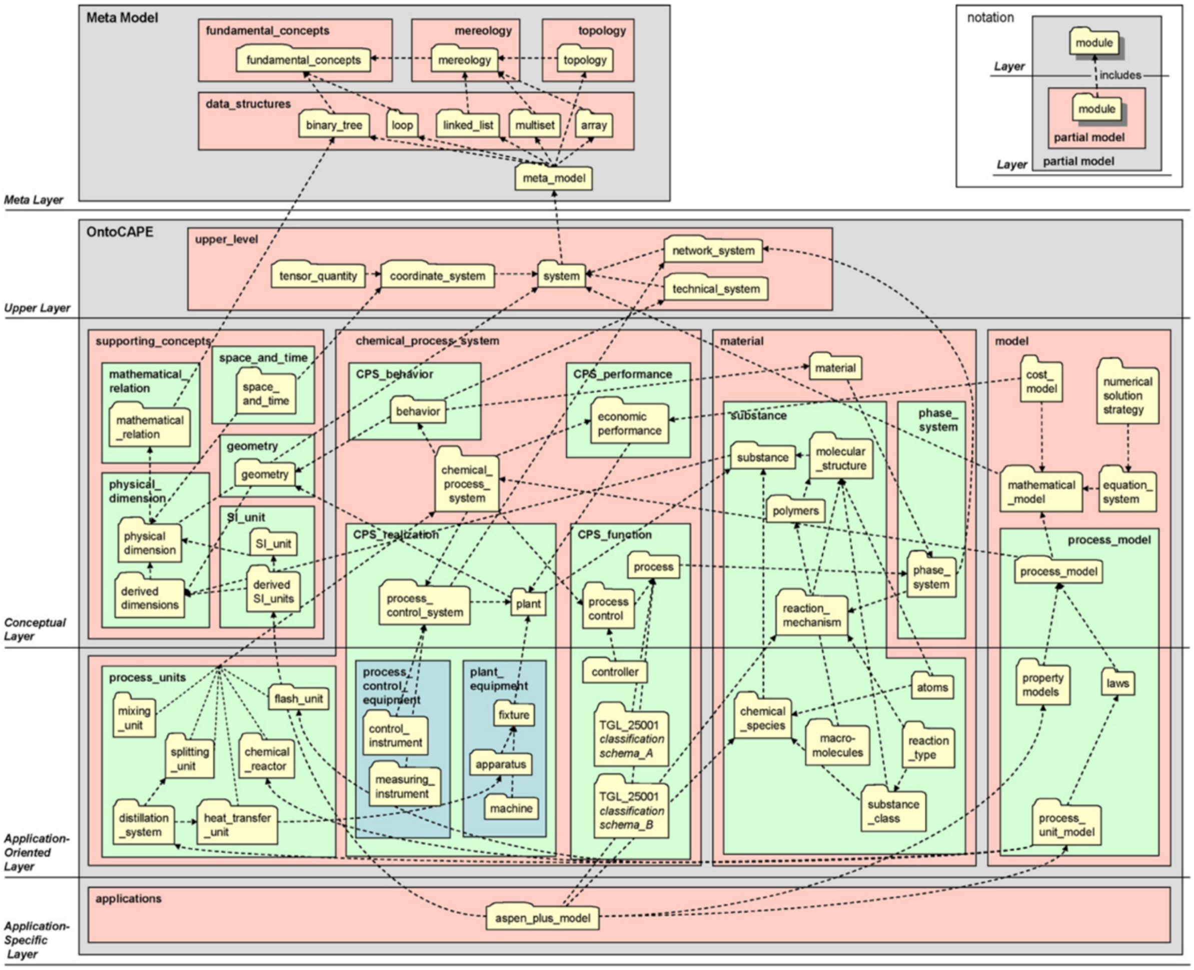

Figure 3.1 Model Structure of OntoCAPE (Morbach et al.)

3.3.1 Ontology Development.

As stated in [52], “An ontology is an explicit specification of a conceptualization”. For the work described in this thesis, we realized the need for a domain ontology for the representation of deviations, causes, consequences, equipment, and their relations.

A multiagent based ontology system has been developed to serve this purpose. The terminology and rules used here such as classes, instances, and relations have been taken from OntoCAPE ontology developed by Morbach et al. [53]. As shown in Figure 3.1, OntoCAPE is a reusable ontology for computer aided process engineering. It is one of the earliest and most comprehensive ontologies in the chemical process engineering field.

OntoCAPE is an ontology available under an open source GNU General Public License [54, 55]. OntoCAPE has been developed with the Web Ontology Language (OWL), and Protégé [56] was used as an ontology editor. They verified the outcome with the reasoner RacerPro [57].

3.4 System Development Using Java Agent DEvelopment Framework (JADE)

Multiagent-based ontologies have been implemented using JADE [58], following the rules and hierarchies of OntoCAPE. Protégé editor have been used to see the representation and structures of OWL files. As shown in Figure 3.1, OntoCAPE is divided into two parts, one is the meta-model and the other is the core ontology. The meta-model (uppermost layer) includes root terms, generic concepts, and design principles. Fundamental concepts, mereology, and topologies are subclasses of the meta-model layer. And the core ontology contains different types of ontologies that are required for a certain application.

All classes, descriptions of relations, and instance descriptions have been taken from OntoCAPE terminology. The conceptual layer has chemical engineering classes and entity relations like operations, equipment/instruments, events, properties, etc. For instance, distillation systems and chemical reactors are subclass ontologies of the chemical process system ontology. Chemical process system is the conceptual layer and chemical reactors and distillation systems belong to the application-oriented layer. Distillation systems ontology will have all the information related to the distillation equipment and their procedures. The chemical reactors ontology will introduce all chemical reactors and their properties and components. The application-specific layer will have all relations and classes related to the specific application under development. Individual agents and a community of agents have been created using OntoCAPE. The visualization tool has been used to extract existing ontology rules and to convey them into the multiagent system.

3.4.1 Multiagent Systems and JADE Key Points.

A multiagent system (MAS) is a distributed architecture developed as a sub-branch of distributed artificial intelligence. MAS is a combination of agents and their interactions in an environment [59]. Agents may be intelligent to an arbitrary level and may be learning agent. They interaction and cooperate when one agent cannot handle a specific case by itself [59]. They try to resolve conflicts whenever they arise. Here are some characteristics of multiagent systems [59];

- Agents have individual duties and limited information with limited capacities. They may be set as static or dynamic. Static agents do not learn and improve while dynamic agents try to adapt to the environment by learning from the environment, from encountered cases, from other agents, and from the history.
- There is no global control, though in some settings, specific agents are assigned coordination role.
- Data is decentralized
- Computation is asynchronous

Researchers and developers in academia and industry realized the need for using multiagent systems when the systems get more and more complex. The interest in adapting multiagent systems is considerably increasing to the level that they have been integrated into solutions developed to serve a wide variety of domains and disciplines.

There is some motivation around multiagent systems development. This could be attributed to several reasons. There is in general resource limitations, or lack of capacity in case of having a single centralized agent or system that might be subject to failure. Instead, a multiagent system is more attractive because in case of a distributed system the failure of one or even more nodes or agents will allow the rest of the system to continue to survive with some limitations. A multiagent system could also enable interoperability of existing systems. Building an agent wrapper could be a good example of the cooperation of legacy systems [60]. Another motivation to adapt a multiagent system is its effectiveness for scenarios when a problem can be more naturally solved with autonomous interacting component-agents. For example, in meeting scheduling, a scheduling agent who manages the calendar of one participant can be considered as autonomous entity who can interact with other similar agents who manage calendars of other participants. All agents may interact to find a good time slot for a future meeting to be attended by majority of participants in case it is not possible to find a time slot which fits all of them [61, 62]. A multiagent system may be ideal for crawling data from multiple sources where one agent may be dedicated to handle one of the available sources. Then agents may negotiate how to integrate the collected data by minimizing overlap and eliminating noise [63]. A multiagent system could well fit a healthcare system where agents may negotiate to determine reasonable diagnosis for a particular patient. They can also be dapted for manufacturing kind of concurrent engineering issues where distributed expertise could lead to more convenient solutions for certain design and manufacturing problems [64].

Adapting a multiagent system may improve the overall performance of a complex system. This is achieved by having better efficiency and concurrency of computation by keeping the communication minimal, such as transmission of high-level information rather than low-level data. A multiagent system is reliable and extensible by its nature of being distributed, and hence the workload of each agent is reduced. Further, limiting the capabilities might help in fault detection and recovery since each agent can be modified to handle specific need. Finally, a multiagent system may be considered robust for tolerating uncertainty.

Maintainability is another key attractive characteristic of a multiagent system since the existence of multiple agents makes it easy to handle them individually and as a group. Responsiveness is almost always achievable in case of a multiagent system. For instance, an issue encountered in one or more agents would not affect the whole system. Specific agents can be chosen and organized for a special problem and this enables system's flexibility. Also, reusability of agents provides the opportunity to combine some well tested agents to solve a new different problem [60]. Using multiagent systems instead of standard text mining and semantic analysis will not just only reduce the overall time of the process, but also prevents the extraction of duplicated information due to negotiation and conflict resolution.

3.4.2 Agent Communication.

A multiagent system was created using Java Agent Development Framework (JADE). JADE is a framework which simplifies the implementation of multiagent systems through a middle-ware that complies with FIPA specifications. It uses a set of graphical tools that support the debugging and deployment phases. JADE comes with a remote GUI and can work in any environment without causing problems. JADE has been implemented fully in the JAVA language and provides all libraries for development.

When two agents communicate with one another there is an information medium between them. Each agent has its own ways to represent this information. To illustrate the process, consider the following example. Assume one report includes this statement: "Four people were injured by the explosion. All were covered with a wet burnt powder and were drenched under emergency showers. They suffered superficial burns to the hands and face and spent one night in a local hospital. They suffered no side-effects. The total cost of the material losses has been evaluated at about 0.648 M Euros. The extents of the material losses were: both front and side pressure relief windows in the process area blown out; one roof vent just above the reactor slightly lifted; the flexible extractor hose above the reactor burnt."

In this example, we can see that four people were injured, and in addition there is also a financial impact on the installation, mainly burning situation of the extractor hose. Rather than parsing this string each time with JADE agents, relevant information can be conveniently represented inside an agent as Java objects.

Consider the following scenario, when agent A sends meaningful information to agent B, both agents should interpret the internal meaning of the communicated information. Agent B should also perform a control check to see if it is appropriate according to its rules. For instance, in the case above, it was important to know that the burnt hose affects the installation. With JADE, all proper conversions and checking mechanisms are provided. The block diagram shown in Figure 3.2 shows how it is easier for a developer to accomplish the task [60].

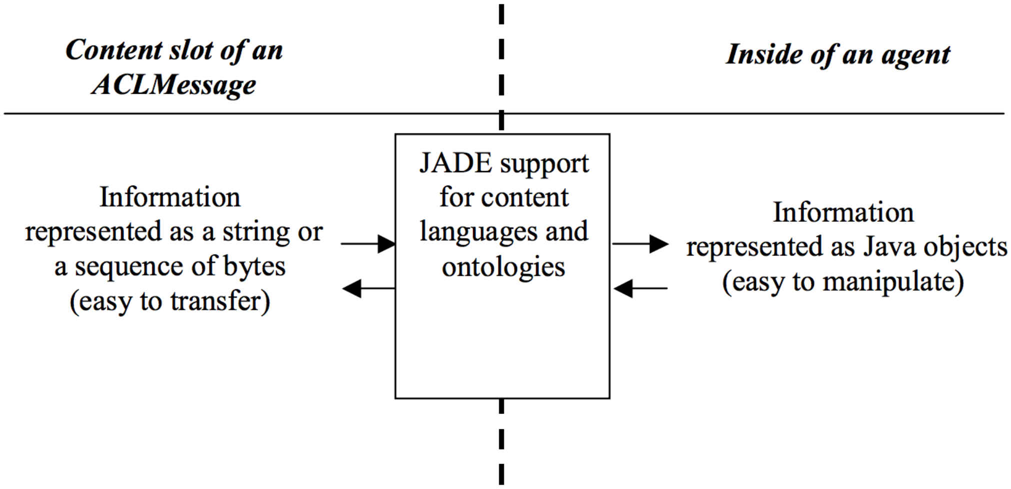
Content slot of an ACLMessage
Inside of an agent
Information represented as a string or a sequence of bytes (easy to transfer)
JADE support for content languages and ontologies
Information represented as Java objects (easy to manipulate)

Figure 3.2 The Conversion Performed by JADE Support for Content Languages and Ontologies (Caire et al.)

JADE comes with appropriate classes and packages which allow developers to use JADE to generate ontologies. The conversion and check methods described above, are carried out with an object from ContentManager class which exists in jade.content package. The ContentManager class provides all the methods necessary to transform Java objects into strings (or sequences of bytes) and to insert them in content slot of ACL messages and vice versa [65]. Further, ContentManager provides a convenient interface to access the conversion functionality. Actually, it just delegates conversion and check operations to an ontology (i.e., an instance of the Ontology class included in the jade.content.onto package) and a content language codec (i.e., an instance of the Codec interface included in the jade.content.lang package). More specifically, the ontology validates the information to be converted from the semantic point of view and codec performs the translation into strings (or sequences of bytes) according to syntactic rules of related content language [65].

3.4.3 Ontology Structure Based on JADE and FIPA Specifications.

Even though JADE provides the most common properties of checking and conversion by FIPA standards, an ontology should be defined for the process to be accomplished. This means defining a vocabulary and semantics for the content of the messages exchanged between agents. Using strings is the most basic way for messages. Using Serialized Java objects is another way of transmitting messages, but there is a drawback of not having human readable messages. Predefined classes permit transferring object definitions by encoding/decoding messages in standard FIPA format.

For message content, there are setter and getter methods which are classified as follows: for strings getContent() and setContent(), for Java objects getContentObjescts() and setContentObjrcts(), and for ontology objects extractContent() and fillContent() are available.

ACL messages should have the proper semantics. To be able to communicate with FIPA standards, it is necessary to be bounded with some predicates and terms. Predicates, also knowns as facts are expressions which may be either true or false about something. Terms represent the identification of entities (abstract or real) about which agents are discussing. Entities with a complex structure are defined as concepts. Agent actions are tasks performed by an agent based on special concepts.

Primitives are types of entities such as strings or integers. An indication of groups of other entities is named aggregates. Identifying Referential Expressions (IRE), represent expressions of entities for a given predicate as true. Variables are expressions which were not recognized earlier.

Shown in Figure 3.3 is the Plant module mereological class diagram taken from OntoCAPE. It describes classes, subclasses, and their aggregations. The plant has four subclasses, namely Fixture, Equipment, TransportCahnnel, and Instrumentation. The taxonomy of both ontologies used for this work have been modified and fitted for suitable entity matching and relationship identification.

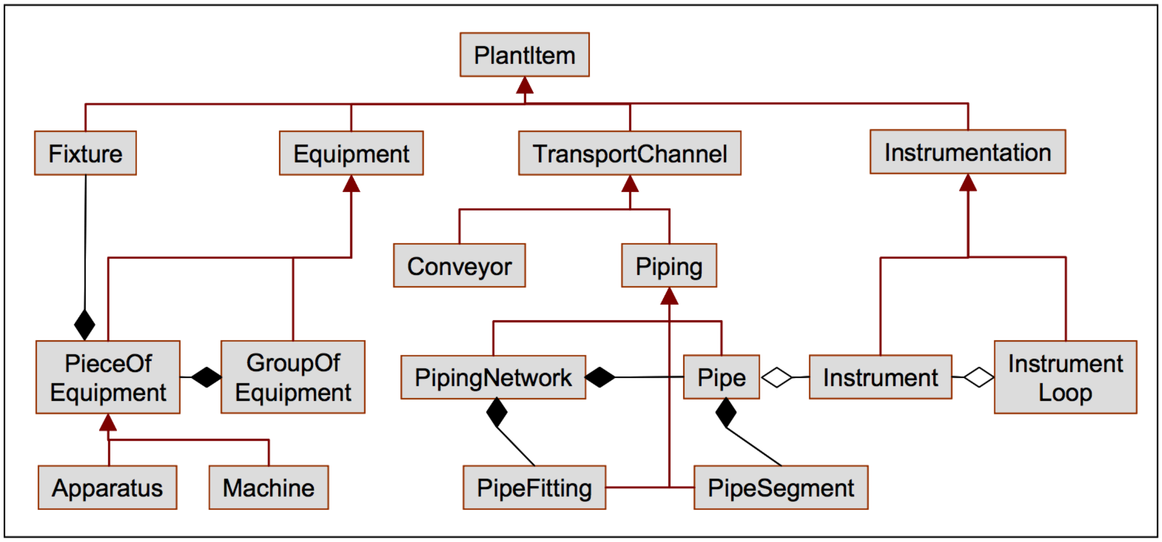

Figure 3.3 Class Diagram of Mereological Considerations of the Module Plant (Marquardt et al.)

For example, a valve is used to control the flow of fluids, and temperature sensors are used for measurements. A valve is an instrument. The connectivity of Nozzles is shown in Figure 3.4.

PipeSegmentEnd and Instrumentation of a plant may be described as follows. PieceOfEquipment, Instrument, and Pipe are subclasses of the plant, while plantItem is a subsystem of the system module. By considering connections between pieces of equipment, piping network and loops can be created. PipeSegmentEnd has two pipe sides which can be connected to equipment, piping, or instrumentation. A nozzle can be connected to ends of a pipe segment or instrumentation connection to build a valid connection. There is also the direct connection of PipeSegmentEnd to InstrumentConnection, or vice versa. A PieceOfEquipment might have one-to-many relationship with nozzles. Pipes are considered to have two ends. But, forking of pipes situation is also possible. Finally, Instrumentation can have one-to-many relationship with IntrumentationConnection [53].

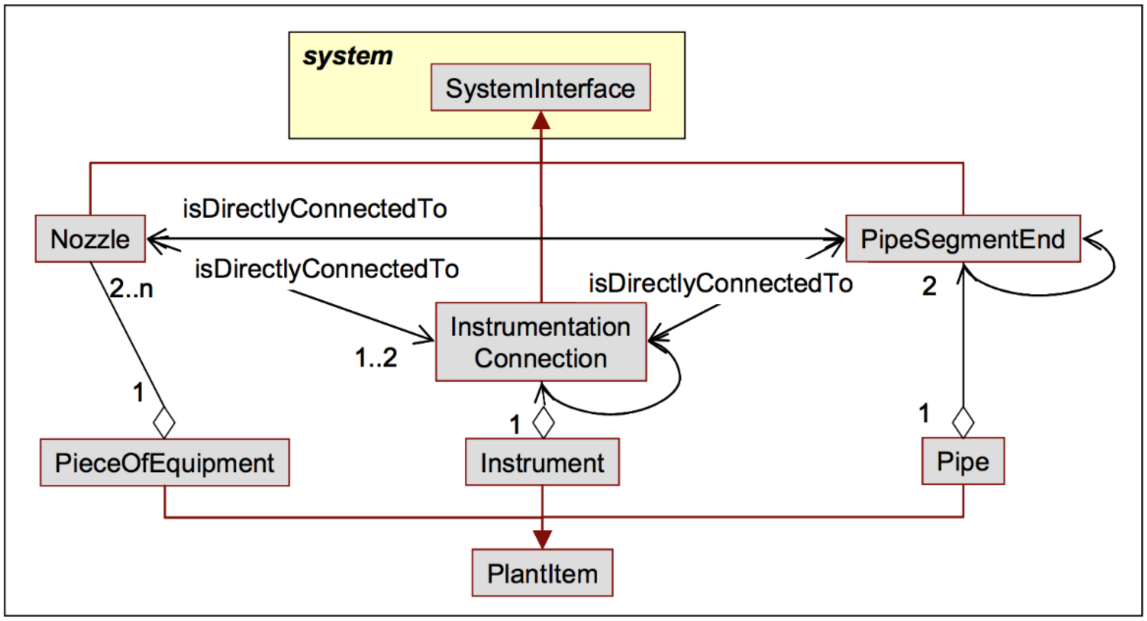

Figure 3.4 Class Diagram for Topological Consideration of the Module Plant (Marquardt et al.)

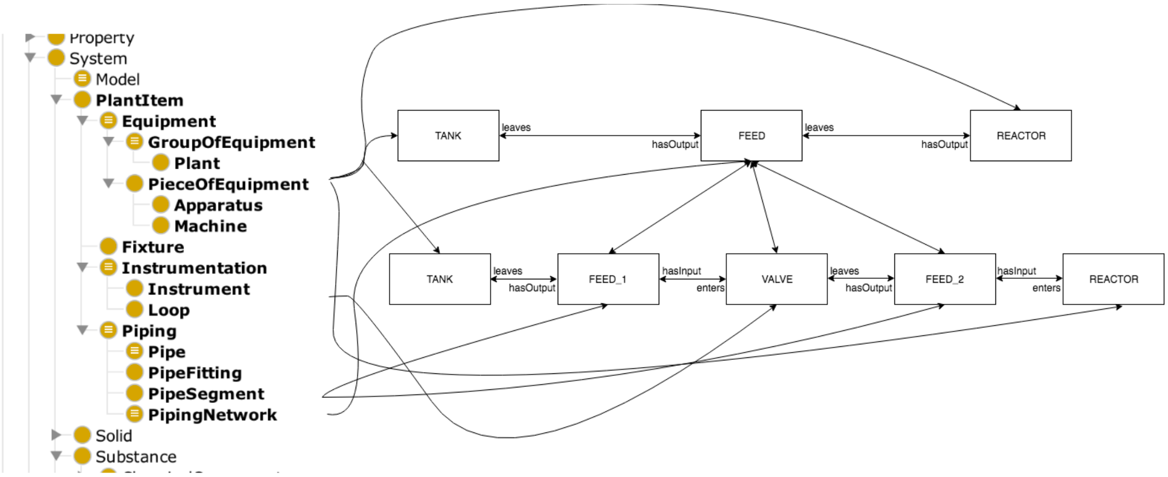

Figure 3.5 An Example Illustrating Elements and Their Relations from Plant Ontology

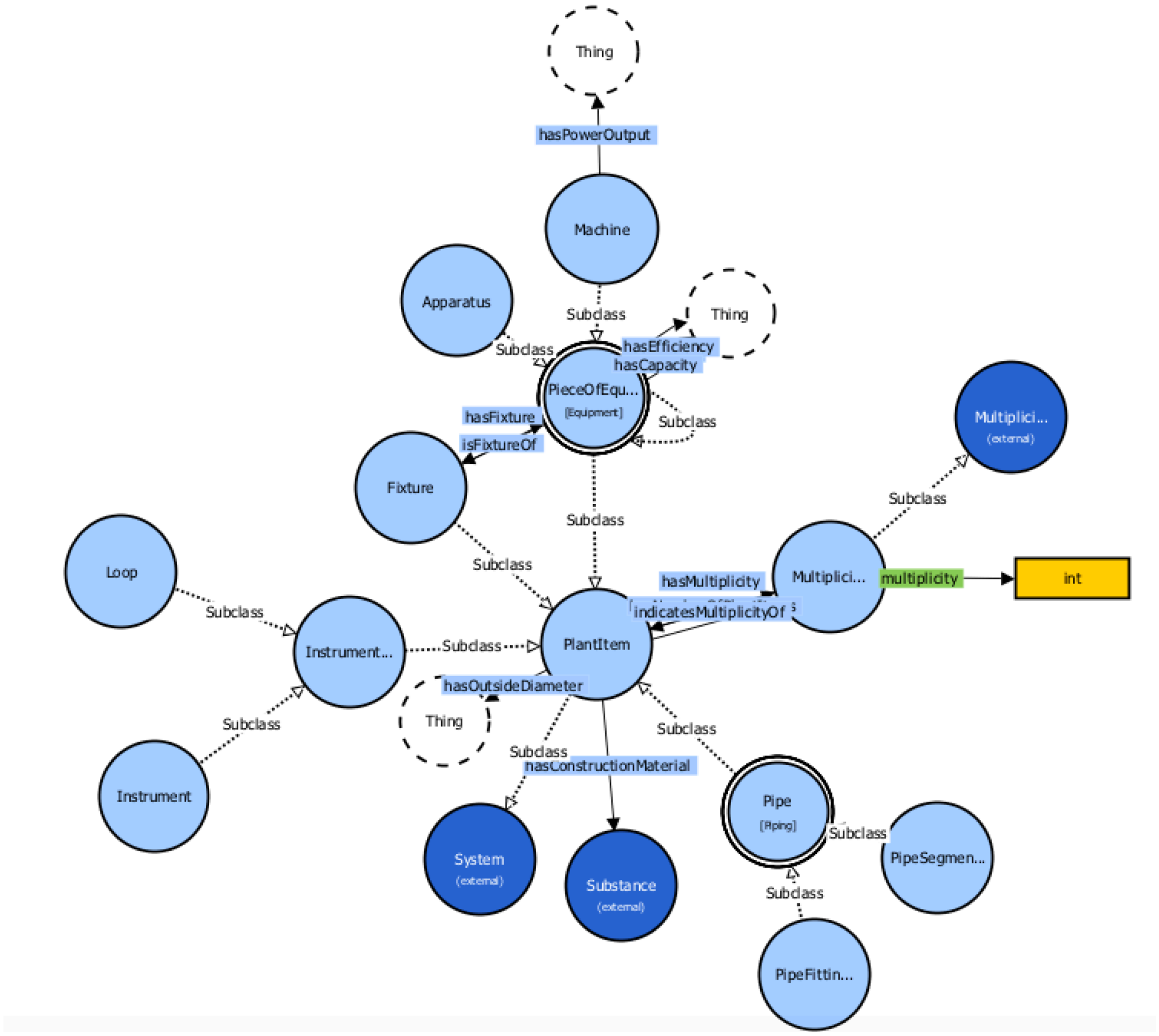

Figure 3.6 Visualization of Plant Module with Protégé Plugin VOWL

A simple representation of the connectivity of elements can be seen in Figure 3.5; this complies with the plant class in the ontology. A visualization of the plant module with Protégé plugin VOWL is shown in Figure 3.6.

3.5 OntologyBeanGenerator Plugin for JADE

Predicates, agent actions, and concepts have been declared using beangenerator plugin of Protégé software. Indeed, there are many hierarchical classes and their rules should be included. Protégé allows users to create ontology definition class together with predicates, agent actions and concept classes. This is facilitated automatically by a graphical user interface via the beangenerator plugin. It simplifies the implementation for the definitions of all Java classes.

The plugin has been installed and added to Protégé application. The OWL file is provided for OntoCAPE ontologies imported along with SimpleAbstractJadeOntology. Protégé does not support the definition of predicates, concepts and agent actions. It only provides the definitions of classes, while SimpleAbstractJadeOntology covers concepts, predicates, and agent actions. All ontology elements under classes are appropriately defined as predicates, agent actions and concepts, or agent identifier (AID). Concepts, agent actions, predicates, and AID have been chosen as subclasses of the class suitable for inter-agent communication. An illustrating example can be given as follow: hasConnector would be a predicate, in this case, Pipe would be a concept, and leaves are agent actions.

For this HAZOP study, ontologies are extended with Causes, Consequences, and Safeguards. However, the risk level is not included because it cannot be found all the time in the incident reports. Ontology rules are defined in the multiagent system. But first unstructured data should be classified according to causes and consequences to determine which category they belong to.

3.6 Clustering Algorithm Counting on Semantic Networks and Proximity

3.6.1 Ontology Inferred Semantic Network.

The proposed algorithm works in the following way. Unstructured data is considered, and each word is assigned to an agent. Agents communicate to decide whether they will meaningfully combine their words. The decision mechanism takes place under the control of rule agents who are defined to implement a descriptor which shows relations that are meaningful according to the provided domain. After acquiring morphological knowledge, syntactical analysis starts by accessing the syntax information in the ontology. This is accomplished by encapsulating and encoding ontologies as ACL messages with FIPA complaints. If different syntactical meanings have been identified, then they are assigned to different descriptors. All possible combinations of a sentence are generated based on its words. Then each meaningful grammatical sentence is taken into consideration as a semantic descriptor using the ontologies. The usage of ontologies here helps to define possible meanings of each word and eliminating the ones which are not needed by agents when they start the negotiation process.

All agents' communication is accomplished by considering FIPA Agent Communication Language, the same as in the earlier methodology. Agents agree on the correct sentence based on the syntax and semantics they have created for the semantic descriptor. The idea of semantic descriptor, which is a network representation of the sentence, has also been used in the works described in [34, 66, 67]. New sentences are checked with earlier descriptors which have been created by agents. Relations instantiated by agents are autonomously changed until agents reach a consensus. Each sentence is pre-processing with the processes of lemmatization, removing stop words, stemming, and morphological analysis. The semantic descriptor should be used for the whole document and/or to all causes or consequences of data.

3.6.2 Agents' Duty for the Clustering Process.

For the clustering process, the proximity of the semantic descriptor was used. Agents used here communicate through JADE by FIPA compliance rules. There are three main agents used for the analysis. The first one is ontology hierarchical rules representation agent. The second agent is for storing words in a sentence based on their descriptors. Agents communicate to extract semantic relations by exchanging information. The rule agent has a unique identifier, and its properties are defined as:

$$R = \{ \alpha_{id}, \Omega_{cl} (< R, I_{i \ldots n} >)\}$$

where α_{id} is the unique agent identifier, and Ω_{cl} is the class of attributes for ontological instances and relationships.

For the word agent, the following structure is considered:

$$W = \{\alpha_{id}, \Psi_{i \ldots n}, \lambda_{i \ldots k}\}$$

where Ψ is the set of words, and λ is for attributes of instance words. For each descriptor determined with the sentences as nodes, the similarities measured with the difference of each $I_{i \ldots n}$ instance relation are defined as concepts. They are used by negotiation agents to rank the results. At this stage, individually extracted causes and consequences have been handled separately for clustering leading to the categorization of causes and consequences.

$$F = \{a_{id}, N_d\}$$

JADE framework enables defining agents as Java threads in Agent Containers. It also provides Agent Management System (AMS) and Directory Facilitator (DF). Directory facilitator can be identified as yellow page service provider so that it can register their capabilities. And with the Agent Management System, automatic registration/deregistration has been handled to manage and control the lifecycle of an agent. Each attribute of the class has its attribute name and type. OntoCAPE provides all hierarchical knowledge representation needed. Finally, the third agent performs the process of creating the descriptors.

3.6.3 Hierarchical Agglomerative Clustering.

Autonomous agent methodology aids not just for communication and information sharing between agents, it also provides for clustering and classification methods with improved performance. Here, FIPA based ACL messages first take strings by communicating with the ontology schema, basically performing an action, and getting the morphological and topological meaning of each word in the sentence. Words are connected such that they will lead to a meaningful descriptor. In this case, one semantic descriptor is chosen for a sentence by extracting key words which are useful for the network. All stop words are discarded to get only meaningful descriptors, such as prepositions, conjunctions, etc.

After the preprocessing is completed, clustering starts by exchanging information for all semantic indexes identified either for causes or consequences. Indexed descriptors will be checked based on their similarity, whether they have the same relationship rules or belong to the same family, e.g., having same parents and same super classes. At this point, agglomerative clustering has been applied to clustering agents who take care of each word in each sentence.

3.6.4 Semantic Similarity Measures.

For a classical bottom-up hierarchical clustering, the similarity measure or distance is expected as input. It is determined by considering the similarity of semantic networks. For similarity measures on indexed semantic descriptors, we have used vector models created by relevance mapped by ontology using descriptors. In a descriptor graph, relationships have been checked using the approach described in [68]. Weight w has been chosen experimentally as 0.6. For each of the n concept nodes of a given mapped descriptor, the value of each component of vector $\vec{v}$ is computed as follows:

$$|\overrightarrow{v_n}| = \begin{cases} 1, & if\ concept\ is\ directly\ connected\ to\ node \\ w, & if\ concept\ has\ a\ relevance\ to\ the\ node \\ 0, & if\ concept\ is\ not\ relevant \end{cases}$$

The similarity matrix is created by considering each descriptor as a taxonomy concept. Dot product is applied on two vectors (cosine similarity). Having n different causes or consequences in a paragraph of sentences would lead to a nxn matrix, and since it is a symmetrical matrix, only one-half (e.g., the upper triangular part) will be needed for the process. Based on these combinations, similarity merge operation is considered monotonically as in single linkage maximum similarity.

$$sim\ (\vec{v}_i, \vec{v}_n) = \frac{\vec{v}_i \cdot \vec{v}_n}{\parallel \vec{v}_i \parallel \parallel \vec{v}_n \parallel}$$

Concepts which have high similarity are clustered together. This procedure continues in a bottom-up manner until a single cluster in obtained. The whole clustering process is initiated by agents. Each clustering agent receives a Request and initially forms a cluster by itself. Then. agents start sending Requests to other agents to build up the matrix. Each Request is followed by a response which is either Agree if it is accepted or Refuse if it is rejected. Alternatively, a Failure message is passed to the inter-communication agent in case a Request remains unanswered. If both agents agree on a Request, they form a cluster based on their distance similarity. Negotiation between agents has been handled by defined protocols based on Contract Net Protocol (CNP), and with Call for Proposal. The communication agent also organizes actions of agents and helps to find a way for handling not understood proposals.

After the communication agent gets the information from the cluster agent confirming the termination of the clustering process, it decides on a prespecified level of similarity to cut the hierarchy depending on the number of clusters relatively desired for causes and consequences.

In the tuples defined above, R stands for Ontology agents, while W and F are data agents. JADE provides Remote Monitoring Agent (RMA), Directory Facilitator (DF), and Agent Management System (AMS) for JADE Behavior as they are instantiated when the system is initiated. Overall the following agents have been used for this work:

- Ontology Agents
- Data Agents
- Protocol Agents
- Clustering Agents
- User Agent

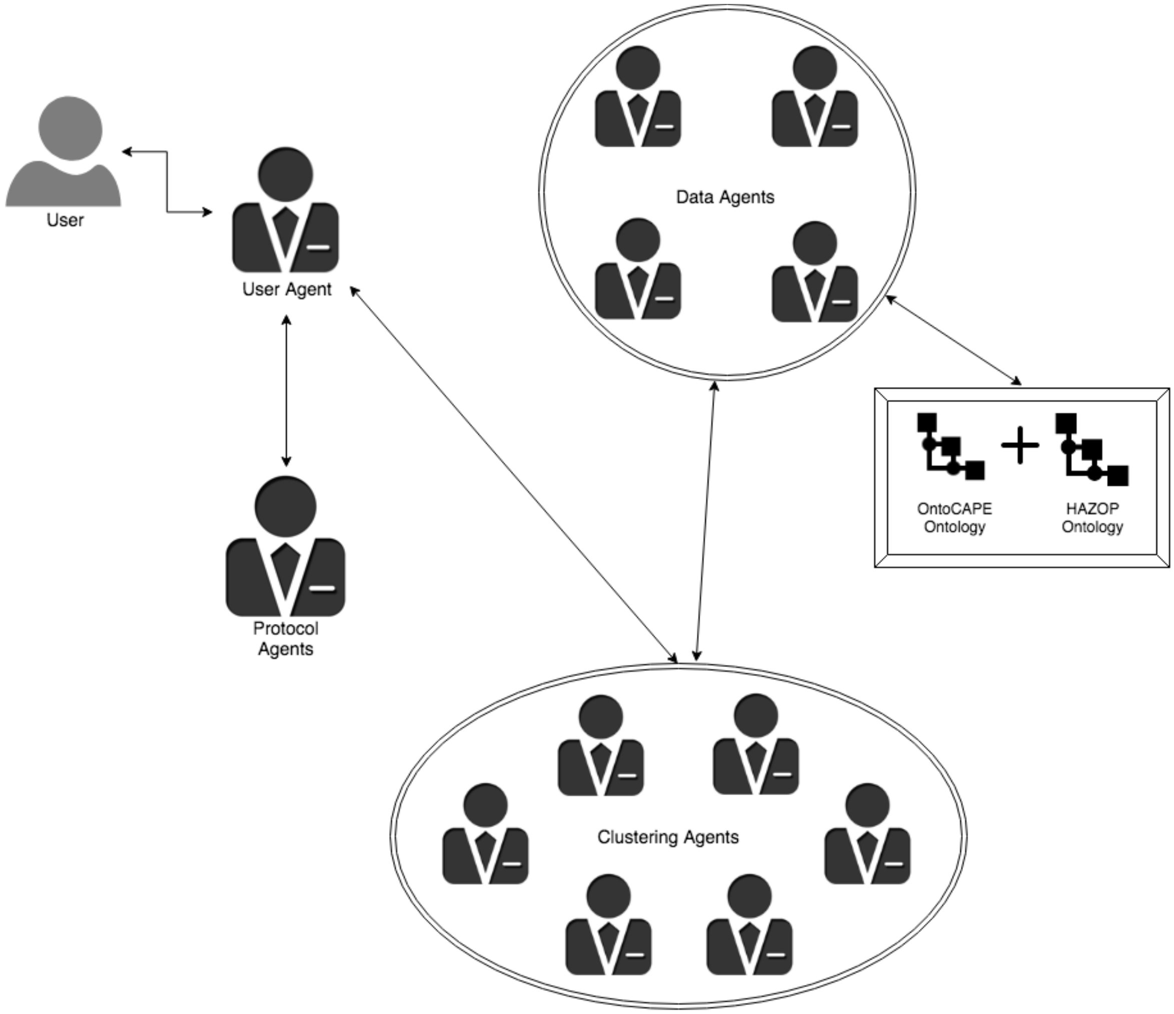

Figure 3.7 Proposed Multiagent System Architecture

3.7 Multiagent Classification and Analysis of Incident Data Risk Level Factor

The other HAZOP analysis concentrates on Safeguard(s) classification with the given causes, consequences, and dependency values of the risk factor level, and whether it has been reduced or not. There is serious need for figuring out which consequence will lead to which safeguard, and how would an identified safeguard be applied to reduce risk level for a more secure environment. This problem has been solved by automatic text classification of historical data which has been collected from earlier HAZOP studies.

3.7.1 Automated Text Classification Using Naïve Bayes Classifier.

For automatic text classification, supervised Naïve Bayes method has been used due to the availability of multi-class classification, and from the results reported in the literature, it has been confirmed as giving good results for automated text classification. The procedure has been handled by employing a multiagent model using JADE Agents. This setup is anticipated to work well for future classification of big and dynamic data flow.

Naïve Bayes method is one of the most commonly used and well-known learning methods for text classification based on maximum probability likelihood. It is effective for pattern recognition. Considering the natural language processing and feature selection tasks which have been done for clustering, the same data has been used to extract feature vectors. Two more agents have been added to the framework for the classification method, namely training agent and classifier agent. For a given document or piece of text d, a fixed set of safeguard classes which has $C = \{c_1, c_2, c_3, \dots, c_n\}$, a labeled training set $(d_1, c_1), \dots., (d_k, c_k)$ is specified. Further, the classification function $\gamma: d \rightarrow c$ is defined based on the Bayes' rule for the given d and c which is written as:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

where t_k represents words (terms/tokens) of document c, $P(c) = \frac{N_c}{N}$ is prior probability, where N_c is the number of documents in class c and N is the total number of documents and $P(t_k|c)$ is the conditional probability of term t_k in a document of a class . The best class has been chosen with maximum posterior, $c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$, where $P(c|d)$ can be substituted with conditionally independent probability of each token to get:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(t_1, t_2, t_3, \dots, t_n|c)\, P(c).$$

Furthermore, because of the independency between the probabilities, it is possible to have:

$$P(t_1, t_2, t_3, \dots, t_n \mid c) = P(t_1, c) \cdot P(t_2, c) \cdot P(t_3, c) \cdot \dots \cdot P(t_n, c).$$

Therefore [67]:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

The conditional probability $\hat{P}(t|c)$, which shows that frequency term t in the documents belongs to class c is calculated as:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T'_{ct} + 1)}$$

Laplace smoothing has been applied to the Multinomial Bayes Model above to avoid zeros in the absence of a word in a given class. Likelihood (probability of each word of the document given a particular class) has been multiplied with prior probability and the highest value has been selected for all the set of classes that have been evaluated. This has been done by maximizing the sum of their logarithms to prevent floating point underflow that a computer may cause due to the lack of enough memory.

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \left(\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c) \right)$$

3.7.2 Agent Identification.

User, data and protocol agents play role in the classification process as well. Furthermore, training agents wait for request response from data agents, and handle the preprocessing and feature selection. They check if the model has been trained or not. After training the model with the data with pre-defined Safeguard classes from the data agent, the protocol agent informs each classifier agent whose status in idle to perform the classification. After the classification process is completed, the protocol agent informs the user via the user agent. Finally, the test dataset will be sent to the classifier agent for testing and prediction in order to determine the overall accuracy.

Chapter 4 Experimental Analysis

This chapter covers all the experiments conducted to illustrate and demonstrate the various parts of the research completed as part of this thesis. Effectiveness of using a multiagent system with ontology, unsupervised and supervised learning methods is discussed. Accuracy is reported for the given classifier method. The features selected and where hierarchical clustering has its appropriate cut using semantic descriptors are explained in the following section. Different environments have been used by virtue of JADE's portability since it works in the Java Platform. Ontology conversion from OWL language to multiagent framework has been accomplished with Protégé [56] ontology editor.

The system created only works with structured Hazard and Operability data. It is not possible to have it directly working with all different kinds of process safety analysis data since every study has its own special structures for data, and hence their analysis differs. This chapter does not contain any comparative analysis because the literature does not include any study of data analysis in this specific field.

4.1 Dataset

The dataset which was used in this study was collected from the heterogeneous data resources mentioned in Section 3.2. A preprocessing and cleaning procedure were applied on data to fit into the HAZOP data structure. PHA-Pro is one of the applications that enable users to conduct and modify HAZOP study worksheets (see Table 4.1). The trials were made on 1000 different study nodes. Only important attributes have been considered for our experimental study. These are the ones which cover multiple consequences, causes, safeguards, severity and likelihood values for risk ranking and recommendation properties. Each of the attributes is described by some sentences which reflect the related topic in the given study node.

Table 4.1 A Snippet of an Example HAZOP Worksheet

Cause	**Consequence**	**Severity**	**Likelihood**	**Risk Level**	**Safeguard**	**Safeguard Category**	**Severity w/ Safeguard**	**Likelihood w/ Safeguard**	**Risk Level After Safeguard**
External fire on the vicinity of	Overpressure at D-100 o	4	B	High	PSV-10213 o	MEC	3	A	Medium
The manual valve on P-100 di	The liquid product line is	2	C	Medium	LIC-10204 on	BPCS	1	B	Low
Plugging of Demister pad in R	Reduced production; as	1	B	Low	PDAH-6120 a	BPCS	1	A	Low

4.2 Feature Sets

Text documents have been used as input for the clustering and classification algorithms that were used in this research. As a preprocessing step, unstructured text documents were converted into machine recognizable feature vectors. Vector space models were then used to translate the data into matrixes with number of features and number of documents, to be considered for this case as incident study nodes.

Automatic text classification reveals safeguards, causes, consequences and whether the suggested safeguards have diminished the power of risk level. Indeed, safeguards are the classes for supervised learning. Some heterogeneous sources include data about safeguards with their classes.

For other sources, some safeguards are covered in unstructured text format only. Some sources do not even include the safeguards at all. HAZOP related data has been the focus of the study described in this thesis. Unfortunately, only a limited amount of HAZOP data has been found that is publicly available. However, it is possible to expand this with private company HAZOP reports.

There are some regular expressions defined for safeguards with unstructured text to get their classes specified as a set of classes. For instance, if a document includes (alarm.*operator) then it would be categorized as BPCS, and (flame.*arrestor) would be classified as MEC-P. For clustering and classification, semantic descriptors are features which have been used for this study. The complete set of safeguard categories and their descriptions are listed in Table 4.2.

Table 4.2 RegEx Table for Manual Categorization

<table>
<tr><th>Safeguard Categories</th><th>Description</th></tr>
<tr><td>BPCS</td><td>operator.*alarm|alarm.*operator|alarm|shut.down|close|trip|shutdown|actuat|automat.*action|controller|control.*valve|control.*function|automat.*valve</td></tr>
<tr><td>ROUND</td><td>operator.*round|round.*operator|daily.*monitoring|monitoring.*daily|operator.*rundown|operator.*monitoring|daily|continuous.*monitoring</td></tr>
<tr><td>Other LS</td><td>PLC|BMS|compressor|(PLC|BMS|compressor|other).*(shut.down|close|trip|shutdown)|(shut.down|close|trip|shutdown).*compressor|hardwire</td></tr>
<tr><td>MEC</td><td>mech|mec|psv|mechanical.*trip|secondary.*containment|berm|mechanical.*stop|restriction.*orifice|flame.*arrestor|minimum.*stop</td></tr>
<tr><td>PRO</td><td>pro|operati.*procedure|procedure|shutdown.*procedure|SOP</td></tr>
<tr><td>SIS</td><td>safety.*function|SIL|SIF|SIS</td></tr>
</table>

<table>
<tr><td>PM</td><td>pm|preventative.*maintenance|pipeline|scheduled|integrity|procedure.*shutdown period|shutdown period.*procedure</td></tr>
<tr><td>OCC</td><td>(personnel|operator).*area.*less|occu</td></tr>
<tr><td>Other</td><td>check.*valve</td></tr>
</table>

Table 4.3 Incident Causes Values Addressed Clusters

Cause Category	Number of Documents (Incidents)
HF	647
EF	314
EEE	39

Table 4.4 Incident Consequences Values Addressed Clusters

Consequence Category	Number of Documents (Incidents)
HS	415
EE	456
EI	129

4.3 System Accuracy

The testing of the proposed methodology for clustering or classification has four main desired clusters or classes for causes, namely Human Factor (HF), Equipment Failure (EF), Environmental or External Effect (EEE), and Unknown. There are also three clusters or classes for consequences, namely Health & Safety (HS), Environmental Effect (EE), and Economical Impact (EI). The outcome will be compared with the corresponding ground truth that has been decided by domain experts who are trusted to be knowledgeable enough to produce almost perfect target classes. The outcome from hierarchical clustering has been cut appropriately to form three clusters for causes (excluding unknown) and three clusters for consequences which are listed in Table 4.3 and Table 4.4, respectively.

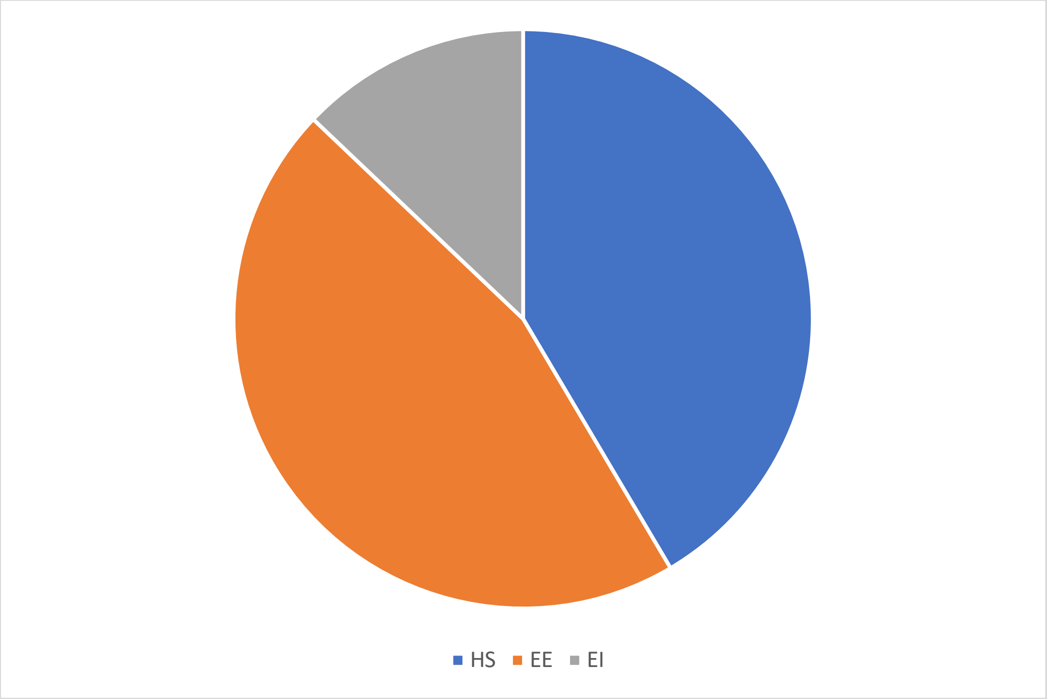

Figure 4.1 Incident Distribution by Their Causes

From the data shown in Table 4.3 and visualized in Figure 4.1, it is obvious that most recorded causes of hazard are attributed to human factors which are almost double the number of causes reported for equipment failure. This shows how it is important to take care of humans involved in oil and gas industry to make sure they are well trained and do not suffer from any stress or other factors which may negatively affect their performance or decision making. It is normal to have the number of external factors related causes low. This may reveal some stability in the environment surrounding oil and gas infrastructure. Hence, humans are the most important when trying to minimize causes of hazardous events. Equipment failure may be avoided by following a stricter regular maintenance program. Environmental effects cannot be avoided but their effect may be reduced by taking some precautions in case an unwanted environmental event is expected to occur. This is becoming more affordable with new technology which allows for early warning due to environmental phenomena. It might even be possible to predict ahead of time the risk level and expected degree of damage associated with a given environmental phenomenon. Because of this, necessary precautions must be put in place that minimizes the damage to oil and gas infrastructure, if failure occurs.

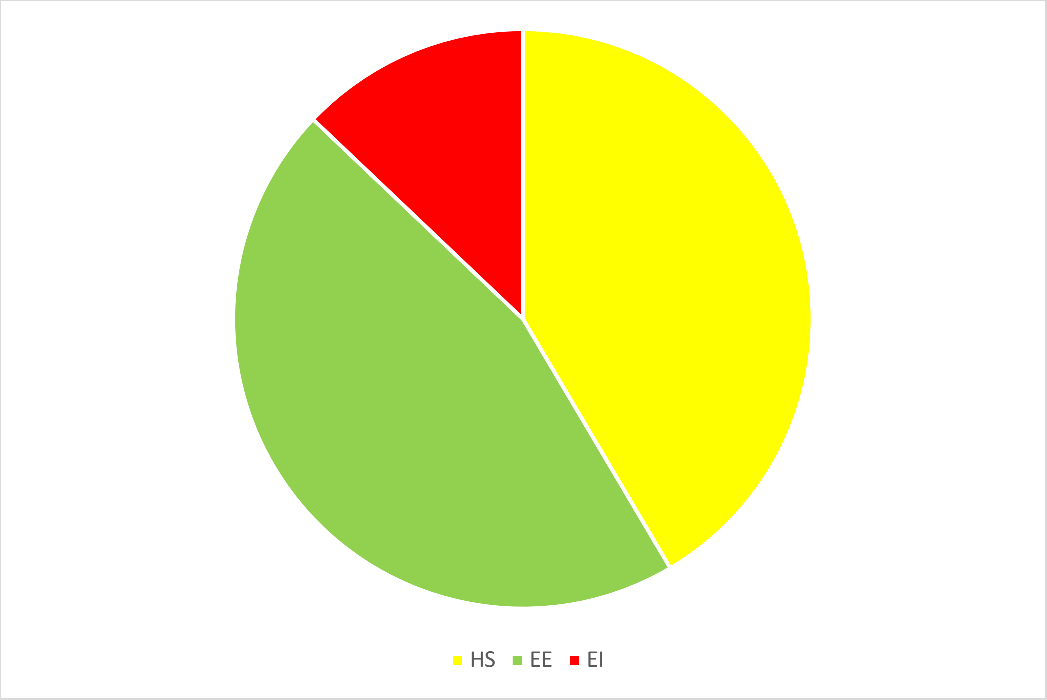

Figure 4.2 Incident Distribution by Their Consequences

The data reported in Table 4.4 and visualized in Figure 4.2 shows that humans are the most affected by hazardous events related to gas and oil industry. This is normal because in general humans are mostly present on site to operate various facilities. Thus, it is essential to avoid or at least minimize hazard as much as possible. Though the given data reflect environment damaging consequences almost at the same level of human related consequences, it is more important to save lives. It is also normal to have the economy the least affected by consequences because in most cases economical effect may stay local to the company running the affected facility. The identified nine safeguard classes or clusters have been specified based on ground truth data using Multinomial Naïve Bayes classifier. The confusion matrix for the mentioned safeguards' predicted actual classes can be seen in Table 4.5. While 80% of the data has been used for training, 20% of the data has been considered as a test dataset. A confusion matrix indicates the accuracy of actual and predicted classes for each incident in the given database. The plot shown in Figure 4.4 reports the number of correct classes and misclassified documents based on actual classes.

Table 4.5 Confusion Matrix for the Classification

P A	BPCS	ROUND	OTHERLS	MEC	PRO	SIS	PM	OCC	OTHER
BPCS	296	0	1	0	8	0	7	0	1
ROUND	0	97	3	2	5	2	6	4	0
OTHERLS	4	1	103	0	11	4	0	0	0
MEC	0	0	4	113	4	1	2	3	0
PRO	1	2	0	2	98	1	3	2	0
SIS	4	0	2	0	3	29	1	2	3
PM	1	2	1	0	1	4	49	2	1
OCC	0	1	0	3	4	1	0	69	0
OTHER	5	0	3	1	0	3	0	0	14

According to the chart shown in Figure 4.3, it can be clearly seen that most used safeguards are alarms with operator action or operating rounds with monthly monitoring. Some mechanical and logic solver safeguards are also highly in effect compared to safety instrument systems. Occupational and operating procedure safeguards have been classified correctly in a good manner. These results not just summarize the most common safeguards, but also enable engineers to select the correct safeguard for the corresponding consequence automatically.

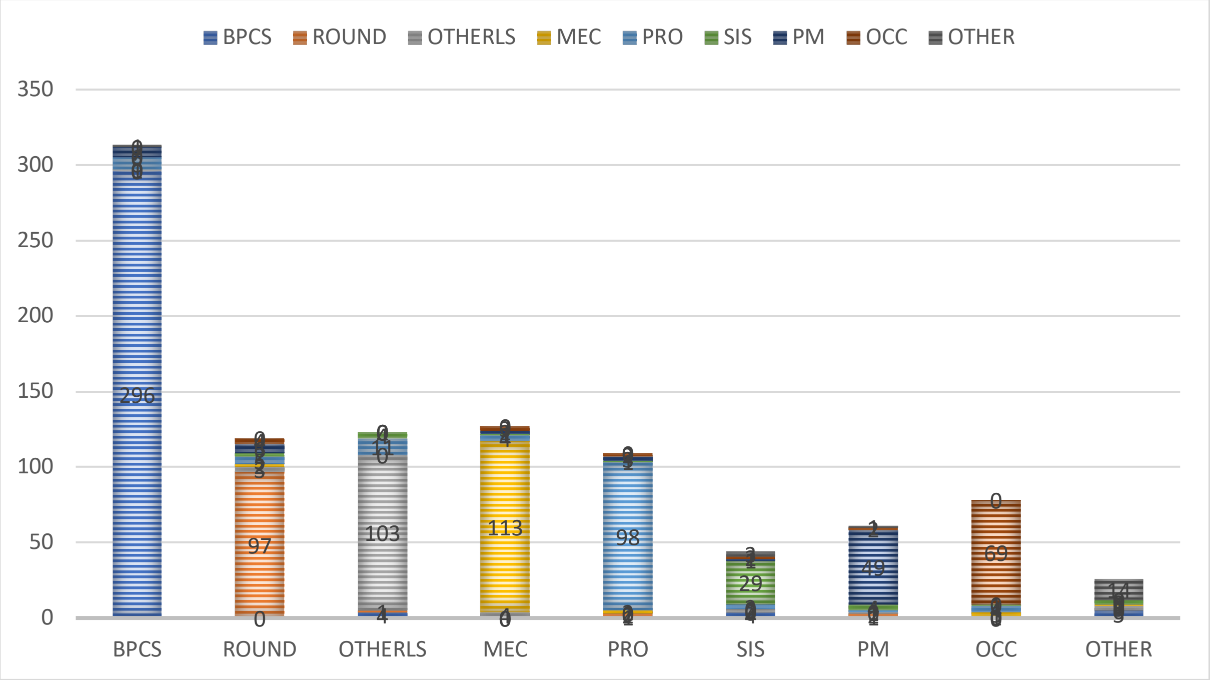

Figure 4.3 Chart of the Number of Correctly/Incorrectly Classified Safeguard Classes

Sensitivity, Precision, and Harmonic Mean of Precision and Recall (F1-Score) values obtained by the evaluation of the destination safeguard classes with the given confusion matrix are reported in Table 4.6 below.

Table 4.6 Evaluation Measures Refined by Each Safeguard Classes

SAFEGUARD	RECALL	PRECISION	F-MEASURE
BPCS	0.945	0.951	0.947
ROUND	0.815	0.941	0.873
OTHERLS	0.837	0.880	0.857
MEC	0.889	0.933	0.910
PRO	0.899	0.731	0.806
SIS	0.659	0.644	0.651
PM	0.803	0.720	0.759
OCC	0.884	0.841	0.861
OTHER	0.538	0.736	0.621

Looking at the values in Table 4.6, it can be easily seen that the values of recall are all above 0.5 (50%), which reflect a good performance of the model. Furthermore, regarding precision values, they are relatively high ranging from 0.644 to 0.95. This metric shows the actual number of positive values compared to all values which were assumed positive. Precisions values reported in Table 4.6 reflect a low false positive rate, which is a positive indicator of good performance of the model.

Explicitly speaking, the values of recall and precision are all high for BPCS, ROUND, OTHERLS, MEC, PRO and PM. These high values clearly illustrate the goodness of the classifier employed in the process. Its performance is indeed very good because it got almost everything classified correctly. It may be considered very sensitive and cautious, but this is very positive because it achieves to an accuracy percentage as expected. It returns results which will guide domain experts the best without tricking them with misclassification of data instances. Finally, it is noteworthy that these good results of recall and precision have produced an overall accuracy of 86,8% for the multinomial Naïve Bayes classifier with the given dataset.

Chapter 5 Conclusion and Future Work

Succinctly, the target of the research methodology which has been covered in this work is to help safety engineers to analyze process safety data in a more stable and accurate manner. To have the method more autonomous and self driven, a multiagent system has been integrated in the data analysis methodology. The developed method reported high accuracy by enabling communication between agents who were then able to produce more concise results.

5.1 Conclusion

Considerable development in automated text analysis informed in the recent literature. However, these developments have not been applied to process hazard safety analysis. Some existing works has outlined how it is possible to automate some duties in order to assist in the analysis of hazard and operability. However, they did not incorporate machine learning techniques for an advanced automated analysis. The approach proposed in this thesis has successfully achieved the aim of integrating automated text analysis by applying data mining techniques and a multiagent system. From the conducted experiments and the reported results into consideration, it can be confidently confirmed that this work reduced the amount of time required to complete a safety analysis with high accuracy.

In this work, data mining and text mining methods have been applied to analyze risk and safety related incident data using ontologies and multiagent systems. With the help of the developed approach, engineers will be capable of conducting the analysis faster by automatically categorizing incidents, and accordingly assigning appropriate safeguards to reduce risk level. Combination of semantic networks and their similarities, ease of interpreting and integrating ontology into a multiagent system could be mentioned as major key distinctions of this thesis.

5.2 Limitations of the Research

One of the limitations of the study described in this thesis is the lack of gold standard values to validate the obtained results. In other words, there was no ground truth for the categories of causes and consequences to validate clustering results. And the other issue that needs to be addressed is the limited amount of publicly available structured Hazard and Operability data that can be used for analysis.

Even though the results reported in this thesis have shown high accuracy on text analysis for HAZOP data, the system developed is not directly applicable to all kinds of process safety identification mentioned in the next section. Another problem is the insufficient available domain knowledge. It was not possible to question the reduction in risk level with the assigned safeguards since the values are determined by engineers.

5.3 Future Work

The approach described in this thesis could be extended in various directions. It can be also enhanced by integrating other techniques in the process. One possible enhancement to the developed approach could be by changing the system into a unique framework in such a manner that it could be applied to any kind of hazard identification analysis (HAZOP, FMEA, HACCP).

Another possible suggestion would be to expand the data covered beyond semi-structured HAZOP text. Instead, it would be more attractive to consider all kinds of incident reports, but this highly depends on availability of proper domain knowledge. Finally, techniques like Support Vector Machines and Convolutional Neural Networks can be used to analyze the given data, and then the methods could be compared for accuracy. It is also worth investigating how association rules mining [70] could help in the process by concentrating only on rules which incorporate incident causes leading to consequences.

Bibliography

[1] "Canadian Centre for Occupational Health and Safety," November 2018. [Online]. Available: www.ccohs.ca/oshanswers/hsprograms/hazard_risk.html.

[2] B. R. Gurjar and M. Mohan, "Environmental Risk Analysis: Problems and Perspectives in Different Countries," RISK: Health, Safety & Environment, vol. 13, no. 1, 2002.

[3] "U.S. Chemical Safety and Hazard Investigation Board," [Online]. Available: www.csb.gov/dupont-corporation-toxic-chemical-releases. [Accessed June 2017].

[4] "U.S. Chemical Safety and Hazard Investigation Board.," [Online]. Available: www. csb.gov/arkema-inc-chemical-plant-fire. [Accessed June 2017].

[5] A. F. Waly and W. Y. Thabet, "A virtual construction environment for preconstruction planning," Automation in Construction, vol. 12, no. 2, pp. 139-154, 2003.

[6] R. M. Felder and R. W. Rousseau, Elementary Principles of Chemical Processes, John Wiley & Sons, Inc., 2000.

[7] R. K. Mobley, Plant Engineer's Handbook, ButterworthHeinemann, 2001.

[8] S. Anand, N. Keren, M. Tretter, Y. Wang, M. S. Mannana and T. M. O'Connor, "Harnessing data mining to explore incident databases," Journal of Hazardous Materials, vol. 130, pp. 33-411, 2006.

[9] Z. Nivolianitou, M. Konstandinidou and C. Michalis, "Statistical analysis of major accidents in petrochemical industry notified to the major accident reporting system (MARS)," in Journal of Hazardous Materials, 2006, pp. A137: 1-7.

[10] R. Batres, Y. Shimada and T. Fuchino, "A Semantic Approach for Incidet Database Development," vol. 155, Institution of Chemical Engineers (IChemE), 2009.

[11] A. Sepeda, "Lessons learned from process incident databases and the process safety incident database (PSID) approach," Center for Chemical Process Safety, no. 130, pp. 1–2, 9–14, 2006.

[12] F. K. Database, "JST," 2009. [Online]. Available: www.shippai.jst.go.jp/en. [Accessed 2018].

[13] R. C. Dubes and A. K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.

[14] M. S. Chen, J. Han and P. S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, 1996.

[15] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1-13, 2007.

[16] D. Srinivasan and P. G. Balaji, "An Introduction to Multi-Agent Systems," in Innovations in Multi-Agent Systems and Applications - 1, L. C. Jain and D. Srinivasan, Eds., Springer, 2010, pp. 1-27.

[17] P. B. Thapa, "Oil Gas Offshore Safety Case (Risk Assessment)," Newfoundland, Canada, 2016.

[18] N. N. Rodhi, N. Anwar and P. A. Wiguna, "A Review on Risk Factors in the Project of Oil and Gas Industry," IPTEK: The Journal for Technology and Science, vol. 28, no. 3, December 2017.

[19] K. Mearns and R. Flin, "Risk perception and attitudes to safety by personnel in the offshore oil and gas industry: a review," Journal of Loss Prevention in the Process Industries, vol. 8, no. 5, pp. 299-305, 1995.

[20] O. Daramola, T. Stålhane, T. Moser and S. Biffl, "A conceptual framework for semantic case-based safety analysis," Emerging Technologies & Factory Automation (ETFA), 2011.

[21] B. Narapan, M. Unchalee and S. Thongchai, "A systematic formulation for HAZOP analysis based on structural model," Reliability Engineering and System Safety, vol. 121, pp. 152-163, 2014.

[22] L. Guo and J. Kang, "An extended HAZOP analysis approach with dynamic fault tree," Journal of Loss Prevention in the Process Industries, vol. 38, pp. 224-232, 2015.

[23] N. L. Rossing, M. Lind, N. Jensen and S. B. Jørgensen, "A functional HAZOP methodology," Computers & Chemical Engineering, vol. 34, no. 2, pp. 244-253, 2010.

[24] L. Cui, Y. Shu, Z. Wang, C. Zhao, T. Qui, W. Sun and Z. Wei, "HASILT: An intelligent software platform for HAZOP, LOPA, SRS and SIL verification," vol. 108, pp. 56-64, 2012.

[25] S. Zhang, J. Teizer and F. Boukamp, "Automated Ontology-based Job Hazard Analysis (JHA) in Building Information Modelling (BIM)," Proceedings of the CIB W099 International Conference, 2012.

[26] R. Batres, S. Fujihara, Y. Shimada and T. Fuchino, "The use of ontologies for enhancing the use of accident information," Process Safety and Environmental Protection, vol. 92, no. 2, pp. 119-130, 2014.

[27] S. Rahman, F. Khan, B. Veitch and P. Amyotte, "ExpHAZOP+: knowledge- based expert system to conduct automated HAZOP analysis," Journal of Loss Prevention in the Process Industries, vol. 22, no. 4, pp. 373-380, 2009.

[28] C. Zhao, M. Bhushan, V. Venkatasubramanian, "PHASuite: an automated HAZOP analysis tool for chemical processes," in Process Safety and Environmental Protection, vol. 86, no. 6, pp. 533-548, 2005.

[29] J. Koldoner, "An Introduction to Case-Based Reasoning, Artificial Intelligence," in Artificial Intelligence Review, vol. 6, 1992, pp. 3-34.

[30] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues. Methodological Variations, and System Approaches," Artificial Intelligence Communications, vol. 7, no. 1, pp. 39-59, 1994.

[31] B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Berlin: Springer, 1999.

[32] B. Ganter, G. Stumme and R. Wille, "Formal Concept Analysis: Foundations and Applications," Berlin, Heidelberg, Springer, 2005.

[33] G. Fu, "FCA based ontology development for data integration," Information Processing & Management, vol. 52, no. 5, 2016.

[34] I. Minakov, G. Rzevski, P. Skobelev and S. Volman, "Creating Contract Templates for Car Insurance Using Multi-agent Based Text Understanding and Clustering," in Holonic and Multi-Agent Systems for Manufacturing. HoloMAS 2007, Berlin, Heidelberg, Springer, 2007, pp. 361-370.

[35] N. O. Garanina, E. A. Sidorova and E. V. Bodin, "A Multi-agent Approach to Unstructured Data Analysis Based on Domain-specific Ontology," in CEUR Workshop Proceedings, Russia, 2013.

[36] S. Chaimontree, K. Atkinson and F. Coenen, "A Multi-Agent Based Approach to Clustering: Harnessing The Power of Agents," in Agents and Data Mining Interaction (ADMI), Berlin, Heidelberg, 2011.

[37] T. E. Potok, M. T. Elmore, J. W. Reed and F. T. Sheldon, "VIPAR: Advanced Information Agents Discovering Knowledge in an Open and Changing Environment," in 7th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, Florida, 2003.

[38] J. W. Reed and T. E. Potok, "A multi-agent system for analyzing massive scientific data," Software Engineering for Multi-Agent Systems (SELMAS), 2003.

[39] R. Ahmad, S. Ali and D. H. Kim, "A Multi-Agent system for documents classification," in International Conference on Open Source Systems and Technologies, Lahore, Pakistan, 2012.

[40] D. D. Lewis, Y. Yang, T. G. Rose and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," Journal of Machine Learning Research, vol. 5, pp. 361-397, 2004.

[41] I. Khan and H. C. Card, "Personal adaptive web agent: a tool for information filtering," in CCECE '97. Canadian Conference on Electrical and Computer Engineering. Engineering Innovation: Voyage of Discovery, Saint Johns, Newfoundland, Canada, 1997.

[42] "Scrapy," [Online]. Available: www.scrapy.org. [Accessed December 2016].

[43] "ProcessNet Eine Initiative von Dechema und VDI-GVC," [Online]. Available: www.processnet.org/process_net/en. [Accessed June 2017].

[44] "Center for Chemical Process Safety - An AIChE Technological Community," [Online]. Available: www.aiche.org/ccps. [Accessed July 2018].

[45] "ARIA (Analysis, Research and Information on Accidents)," [Online]. Available: www.aria.developpement-durable.gouv.fr. [Accessed June 2017].

[46] "EUROPA - eMARS Dashboard - European Commission," [Online]. Available: www.emars.jrc.ec.europa.eu/en. [Accessed June 2018].

[47] "National Transportation Safety Board," [Online]. Available: www.ntsb.gov/Pages/default.aspx. [Accessed June 2017].

[48] "Failure Mandalas," [Online]. Available: http://www.shippai.org/fkd/index.php.

[49] "U.S. Chemical Safety and Hazard Investigation Board," [Online]. Available: www.csb.gov. [Accessed June 2017].

[50] "Failure and Accidents Technical information System," [Online]. Available: http://www.factsonline.nl/. [Accessed June 2017].

[51] "International Transport Forum - Global Dialogue for Better Transport," [Online]. Available: www.itf-oecd.org/irtad-road-safety-database. [Accessed June 2017].

[52] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[53] J. Morbach, A. Wiesner and W. Marquardt, "OntoCAPE—A (re)usable ontology for computer-aided process engineering," Computers & Chemical Engineering, vol. 33, no. 10, pp. 1546-1556, 2009.

[54] "OntoCape - RWTH AACHEN UNIVERSITY Aachener Verfahrenstechnik," [Online]. Available: www.avt.rwth-aachen.de/cms/AVT/Forschung/Software/~ipts/OntoCape/lidx/1. [Accessed August 2017].

[55] "The GNU Operating System," [Online]. Available: www.gnu.org/licenses/gpl-2.0.html. [Accessed April 2018].

[56] M. A. Musen, "The Protégé project: A look back and a look forward," AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, vol. 1, no. 4, pp. 4-12, June 2015.

[57] V. Haarslev, K. Hidde, R. Möller and M. Wessel, "The RacerPro knowledge representation and reasoning system," Semantic Web Journal, vol. 3, no. 3, p. 267–277, 2012.

[58] G. Rimassa, A. Poggi, G. Caire and F. Bellifemine, "JADE: A software framework for developing multi-agent applications. Lessons learned," Information and Software Technology, vol. 50, no. 1-2, pp. 10-21, 2008.

[59] K. P. Sycara, "Multiagent systems," AI Magazine, vol. 19, no. 2, pp. 79-92, 1998.

[60] M. R. Genesereth and S. P. Ketchpel, "Software Agents," Communications of the ACM, vol. 37, no. 7, pp. 48-53, 1994.

[61] L. Garrido and K. Sycara, "Multiagent Meeting Scheduling: Preliminary Experimental Results," In Proceedings of the Second International Conference on Multiagent Systems, pp. 95-102, 1996.

[62] L. Dent, J. Boticario, J. McDermott, T. Mitchell and D. Zabowski, "A Personal Learning Apprentice," In Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 96-103, 1992.

[63] K. Sycara, K. Decker, A. Pannu and M. Williamson, "Distributed Intelligent Agents," IEEE Expert, vol. 11, no. 6, pp. 36-46, 1996.

[64] C. M. Lewis and K. Sycara, "Reaching Informed Agreement in Multispecialist Cooperation," Group Decision and Negotiation, vol. 2, no. 3, pp. 279-300, 1993.

[65] C. Giovanni and D. Cabanillas, "JADE Tutorial Apllication-Defined Content Languages and Ontologies," Telecom Italia S.p.A., 2010.

[66] G. Rzevski, P. Skobelev and I. Minakov, "Automated Text Analysis". United Kingdom Patent GB2412451A, 26 March 2004.

[67] G. Sokolov and V. Lanin, "One Approach to Document Semantic Indexing Based on Multi-Agent Paradigm," in SYRCoSE (Spring/Summer Young Researchers' Colloquium on Software Engineering), 2012.

[68] H. Liu and P. Wang, "Assessing Text Semantic Similarity Using Ontology," Journal of Software, vol. 9, no. 2, pp. 490-497, 2014.

[69] C. D. Manning, P. Raghavan and H. Schütze, An Introduction to Information Retrieval, Cambridge, England: Cambridge University Press, 2009, pp. 253-289.

[70] C. Silverstein, R. Motwani and S. Brin, "Beyond Market Baskets: Generalizing Association Rules to Correlations," in In Proceedings of ACM-SIGMOD International Conference on Management of Data, Tucson Arizona, 1997.