Early Alert of At-Risk Students: An Ontology-Driven Framework

Elias S. Lopez
University of New Mexico - Main Campus

Candidate

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

, Chairperson

Early Alert of At-Risk Students: An Ontology-Driven Framework

by

Elias S. Lopez

A.S., Wyoming Technical Institute, 2009
B.S., University of New Mexico, 2014

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Engineering

The University of New Mexico

Albuquerque, New Mexico

May, 2018

Dedication

To my parents, Elias G. Lopez and Margarita Simental, for their support, encouragement, and for giving the opportunity to attain this valuable education. Without both of you, none of this would've been possible.

"Un pueblo ignorante es un instrumento ciego de su propia destrucción" – Simon Bolivar

Acknowledgments

I would like to thank my committee members, Prof. Gregory L. Heileman, Prof. Don Hush, Dr. Terry Babbit and Dr. Ahmad Slim. Also, I would like to express special gratitude towards my advisor Dr. Gregory Heileman as he took a chance on a Mechanical Engineer and turned him into a Computer Engineer. Lastly, I will like to thank all of the IDIers, especially Jarred Kozlick and Ahmad Slim, with whom I have worked with and became friends with.

Early Alert of At-Risk Students: An Ontology-Driven Framework

by

Elias S. Lopez

A.S., Wyoming Technical Institute, 2009

B.S., University of New Mexico, 2014

M.S., Computer Engineering, University of New Mexico, 2018

Abstract

As higher education continues to adapt to the constantly shifting conditions that society places on institutions, the enigma of student attrition continues to trouble universities. Early alerts for students who are at-risk academically have been introduced as a method for solving student attrition at these institutions. Early alert systems are designed to provide students who are academically at-risk a prompt indication so that they may correct their performance and make progress towards successful semester completion. Many early alert systems have been introduced and implemented at various institutions with varying levels of success. Currently, early alert systems employ different techniques for identifying students that may be at-risk. These techniques range from using machine learning algorithms for predicting students that may become at-risk to more manual methods where the professors are responsible for assigning at-risk tags to students in order to notify the student.

This thesis will introduce an ontology-driven framework for early alert reporting of students at-risk. To be more precise we will determine early alerts for students

who are at-risk with an ontology-driven framework employing situational awareness. Ontology-driven frameworks allow us to formalize situations in a way that is similar to the human interpretation of situational awareness. The ontology presented will be constructed using OWL the Web Ontology Language. The use of this language will facilitate the description and reasoning of the situation as it is a commonly supported programming language with computable semantics. In this piece we will consider factors such as advisor notes, learning management system interaction, as well as other factors related to student attrition to assign at-risk tags to students who may be at academic risk.

Contents

List of Figures

List of Figures

List of Tables

Glossary

LMS	Learning Management System
EA	Early-Alert
SWRL	Semantic Web Rule Language
SAO	Situational Awareness Ontology
NLP	Natural Language Processing
RDF	Resource Description Framework
XML	eXtensible Markup Language
W3C	World Wide Web Consortium
CSV	Comma Separated Values
NLTK	The Natural Language ToolKit

Chapter 1

Introduction

Many things have changed in higher education but student attrition continues to be a problem and therefore a topic that continues to be researched. Student attrition, as defined by the first author credited for researching this topic in 1993, Tinto, is "a longitudinal process of interactions between the individual and the academic and social systems of the college during which a person's experiences in those systems ... continually modify his goals and institutional commitments in ways which lead to persistence and/or to varying forms of dropout [27]." Student attrition is attributed to many factors along the journey of the student through higher education at an institution as Tinto describes it. The interest of young students is rapidly changing therefore many students often disenroll from their perspective college in order to switch career paths. According to a recent study from the National Center for Education Statistics, the average 6-year graduation rate in the United States is 57% [14]. Many students simply transfer to other institutions that are either more prestigious or that are more in-line with their current interest. According to Tinto more than half of the students who disenroll from their first institution often do not come back to higher education[27]. Most of the students who decide to not come back often leave their first institution due to academic or financial reasons. There is a need for a system that pinpoints at risk students so that the institution may give them the guidance needed to realize their higher education aspirations.

Student attrition proves to be a much more complex and serious problem at Hispanic Serving Institutions (HSI), like the University of New Mexico. Currently in the United States hispanic graduation rates are much lower than the graduation rates of caucasian americans [13]. Nationally Hispanics graduate at a rate of 51% in six years while caucasian students graduate at a rate of 59% in the same time frame. This case proves to be much more severe at HSIs as these nationally assigned institutions graduate less then half of their Hispanic population. This problem exits at the University of New Mexico as currently the six-year graduation rate for hispanics is 45%, and 48% for the entire student population [1].

The need to graduate more of the students is urgent as the consequences for students opting to drop out of higher education are quite drastic and affect a wide range of stakeholders not just the individual student electing to disenroll. The stakeholders for a perspective student's academic success are himself, society, our economy and the institution. On average according to U.S. Census Bureau data, a young adult with a college degree working full-time in a year earns close to 40% more than a young adult of the same age without a college degree [2]. Also, students who opt to end their college career without obtaining a degree often accumulate large debts which they tend to have a hard time repaying as their income is drastically lower than a college graduate. The implications these students impose on our society and economy are quite large as well. Society makes a direct investment in allowing students access to education. The origin of this investment comes from taxes paid by the majority of the population. In 2010 it was reported that states spend $1.3 billion every year on students who elect to drop out within their first year of attending higher education institutions [20]. The federal government also spends an additional $300 million on the same set of students. A report in 2011 by the American Institutes for Research estimates the amount of lost revenue for students that opt to drop out college is \$3.8 billion for the entire population for one year [22]. The same report calculates the amount of lost revenue for these students is \$566 million in federal taxation revenue and \$164 million in state taxation revenue nationwide. The loses for one cohort of students are immense and they place a burden not only on themselves but also on our society. Institutions are left with aftereffects when a student decides to leave one of the institution's programs, they also face many challenges after students leave. Many costs are incurred by an institution when a student enrolls at an institution as the institution has to dedicate time and resources (faculty, infrastructure, etc.) to a student [21]. To make matters worse, when a student decides to leave the institution many of those costs that were initially incurred by the institution are not discontinued but the tuition compensation for that student is absent. The institution has to ensure that even though there are empty seats in the classroom, all the faculty, staff and facility expenses are met. Student attrition affects every stakeholder that is associated with academic success, as students, our economy and institutions are linked to an individual's academic achievement. Institutions need to continue to research methods that help to retain students and develop procedures to graduate more students.

Tinto commenced research on attrition in the mid 1980s and this phenomenon continues to be a relevant form of research. There are various causes for student attrition at postsecondary institutions and many systems have been designed to help address student attrition. Some of the factors that cause a student to disenroll from a given institution are simply enrolling at another institution, but more often than desired many students disenroll due to factors institutions control. There are many causes for attrition, they can range from under preparation before college to lack of class attendance [26]. To help alleviate this problem many systems have been designed to examine targeted advising, lengthy course sequences, early alerts, students at risk, etc. [10, 23, 11, 24, 19]. Those systems have helped decision makers make progress in the battle against student attrition but this multifaceted problem is yet to be solved.

In this document we introduce a new approach aimed at helping institutional administrators identify students at risk in order to help them diminish student attrition. We introduce the situational awareness framework to the student scenario. Situational awareness has been applied to many scenarios which range from military applications to cybersecurity. Situational awareness will facilitate pinpointing students that might be having academic, financial, or other problems that might hinder their academic motivation and/or lead them towards disenrolling from the institution. Currently many of the early alert systems that have been designed depend solely on professor participation and submission of their records [11, 24, 19]. Situational awareness will alleviate the load on professors, meaning professors will not be relied on so heavily for reporting students in their class, by introducing techniques that will find students with academic difficulties such as absences, poor test scores, etc.

Next in this passage we will give a brief background into all of the artifacts that will come together to shape our situational awareness student scenario system. First we explain some of the causes of student attrition, the methods we use to measure them and the limitations our system has in measuring them accurately. Consequently, we expand on the topic of situational awareness and how we can form a uniform ontology to describe our student scenario. We explain both of these techniques and how they have been applied to other settings and why it is important to apply it to this scenario. After we have introduced all of these ideologies we explain what a threat model is and how it can help us identify the anomalies of individual student academic difficulties in order to pinpoint the corresponding students. Finally, we illustrate how our proposed situational awareness system functions detecting at risk students.

Chapter 2

Background

Student attrition in higher education is due to a collection of issues in the journey of a student through an institution. More unambiguously, as Tinto describes it, a longitudinal process involving a series of complex socio-psychological interactions between the student and the institutional environment. According to this theory, the student comes to college with many characteristics such as family personal attributes, background and experiences. Each of those characteristics is believed to influence not only college performance, but also initial levels of goals and institutional commitment. These characteristics and commitments, in turn, interact with various structural and normative features of the particular college or university and lead to varying levels of integration into the academic and social systems of the institution. According to Tinto, “Other things being equal, the higher the degree of integration of the individual into the college systems, the greater will be his commitment to the specific institution and to the goal of college completion”.

Although there are many factors of student retention in which an institution cannot help, there are various characteristics of a student that an institution can influence in order to guide the student to achieve the ultimate goal of higher education. It was reported in the Chronicle of Higher Education that approximately 9 out 10 students that come into higher education needing at least one remedial course intend to use academic assistance programs at their respective institutions [15]. Also, according to Kuh, approximately half of those students do use those academic assistance programs by the end of their first year. This is a factor an institution can assist a student with. If the institution is able to pinpoint a student in need of academic assistance they may guide that student to the resource they need. As Kuh states, "we (institutions) can do better ... by engaging them in purposeful activities that enhance their learning and personal development. The real question is whether we have the will to increase the odds that more students will get ready, get in, and get through [15]."

2.1 Early Alert Systems

Throughout the last thirty years a number of early alert systems have been implemented at various institutions with the aim to help solve the student attrition problem plaguing colleges. The goal of these systems is to pinpoint students at risk and get them the help they may need before they decide to quit higher education. All of the systems that have been implemented focus on several attributes that cause student attrition but the majority of them regarding early alerts focus on excessive absenteeism as it is often leads to a student leaving his goal of higher education. At Morehead State University (MSU) an early alert system was implemented in 2003 to track student course attendance [11]. This system was designed to only track excessive absenteeism as the institution approximated that "more than 20% of freshman students who fail courses do so as a result of excessive absenteeism during the fist 4-6 weeks of the semester." At this university a web-based system for reporting attendance was provided to the instructors as well as a reporting system to notify students and advisors of the absences. Although a longitudinal study was not provided, this report concluded that by contacting students with excessive absenteeism they demonstrated to the students the importance of attending class.

At other institutions similar early alert systems were implemented but they tracked more attributes that may contribute to a student's scholarly success. At the University of North Texas (UNT) and at Columbia College (CC) a system was introduced that allowed course instructors to identify and report a student whom they had identified as at risk for one or more academic reasons [24, 19]. Similar to MSU, at UNT a web based system was implemented to track students at risk. The main difference being that professors at UNT were able to identify students at risk for various reasons, as listed in below.

1. poor class attendance;
2. poor performance on quizzes/exams;
3. poor performance on writing assignments;
4. does not participate in class;
5. difficulty completing assignments;
6. difficulty with reading;
7. difficulty with math;
8. sudden decline in academic performance;
9. concerns about their major;
10. college adjustment issues;
11. financial problems;

12. physical health concerns;
13. mental health concerns;
14. alcohol or substance use concerns;
15. roommate difficulty;
16. disruptive behavior;
17. absent from work; and
18. student needs veterans assistance.

If faculty at UNT identified a student as at risk then a central clearing house was notified which would then notify the student and, if appropriate, connect him to the resources that he may need. In the first term that this system was implemented, 255 students had been identified as at risk in 108 courses with over 56% of the students referred for poor class attendance. At UNT they found that when a student was contacted successfully by both the central clearing house and the professor, the term grade point average for that student was on average one point higher than a student whom they failed to contact. With an early alert system that resembles the system implemented at UNT, at Columbia College early alerts were integrated in the mid 1980s. Although this system was not technology enabled, it allowed this large institution to identify and intervene in manner similar to UNTs.

Pfleging [19], unlike UNT, provided a study of the effectiveness of such system at CC. This institution attracts a student body that is mainly formed of students who are often categorized as high-risk. They are categorized as high-risk because many of these students are often ill-prepared for college, employed full-time, head of households, etc. At CC professors were only able to report students as at risk once a semester. Also, professors at CC were only able to tag students as at risk students for five reasons, as listed below.

Reason for Alert	Number Alerts	Percent of Total
Failed Test	42	34.7
Incomplete Assignments	14	11.6
Lacking Participation	12	9.9
Not Attending	51	42.2
Lacking Basic Skills	2	1.6
Total	121	100

Table 2.1: Reasons for Alerts in Surveyed Classes, Fall Semester 2001

1. not attending class;
2. incomplete assignments;
3. failed test;
4. not participating;
5. lacking basic skills;

Once a student is tagged as at risk a letter informing the student of his status in the course is mailed to him/her. The letter that is sent to the student contains various suggestions for follow-up actions and it also attempts to connect the student to services that can guide him/her to success in the given course. For the fall 2001 semester the early alerts given are summarized as follows in table 2.1.

As seen in table 2.1 over forty percent of the students tagged as at risk were for excessive absenteeism, much like UNT and MSU. For gauging the effectiveness of this early alert program students were surveyed and academic performance statistics were gathered. 370 students were surveyed and out of those students 14 indicated they had received an early alert letter. Academic performance for students that were not considered at risk were compared to those that were considered at risk and summarized in the table 2.2.

Actual Academic Performance	EA Students	Non-EA Students
% students with grades of A, B or C	13.1	68.5
% students with grades of D, F or I	19.5	11.8
% students with grade of W	67.4	19.7

Table 2.2: Actual Academic Performance of Early Alert (EA) and non-EA Students in the Surveyed Classes Fall 2001

All of the early alert systems which have been summarized in this section focus on professors reporting those students who are at risk. Recently machine learning has enabled another method for reporting students who may be at risk. These systems focus on predicting which students are likely to be at risk based on a limited number of student attributes. The method introduced by Jayaprakash [12] focuses on detecting, as early as possible in the semester, students that might be academically challenged by a course based on student attributes. As Jayaprakash describes it, "this task was re-expressed as a binary classification process with the purpose of discriminating between students (a) in good standing or (b) academically at risk" The classification model is trained by labelled data samples with a target value. The system Jayaprakash proposes is illustrated by figure 2.1.

In conclusion, currently no method that has been introduced in literature is the definitive solution for reporting early alerts. Although some of the first three methods introduced are technology enabled, all require manual reporting from professors and do not offer any automation for reporting early alerts. The last method introduced attempts to predict which student might be at risk based on student parameters, but like all predictive models perfect accuracy is not guaranteed.

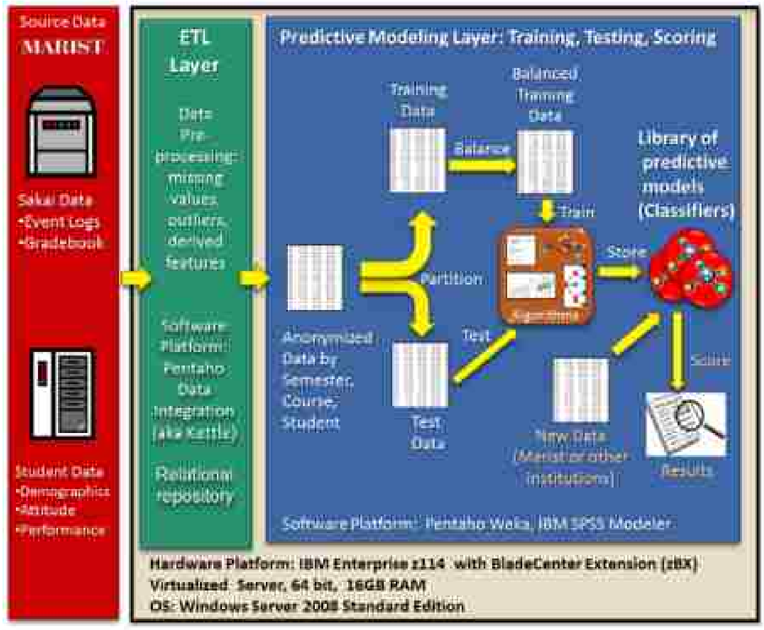

Figure 2.1: OAAI predictive modelling architecture

2.2 Learning Management System Interaction

Recently data scientists in higher education have been analyzing the relationship between student success and the amount of time a student spends using a Learning Management Systems (LMS) [28]. The guiding question of this analysis was: "is total time spent in a Blackboard Learn course a good predictor of final grade" In this work it is understood that LMS interaction does not imply that a student is learning better because of interaction with the system. In the research of Whitmer [28, 29] all that he is attempting to decipher is if there is a relationship between interaction with the LMS system and a student's final grade. In his first study [28] Whitmer investigated the relationship between the time spent in course portal (in the LMS system) of a course and the final grade of that course. His first finding was that there was a small relationship between time spent on the LMS course portal and the final grade of that course, as shown in figure 2.2. In this reports he also

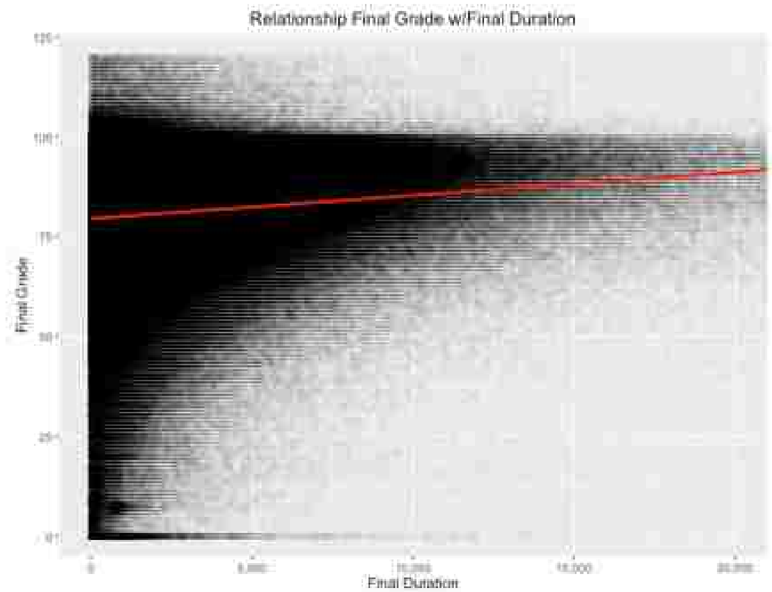

Figure 2.2: Relationship of Final grade to LMS Interaction

investigated if this relationship is higher in some courses more than in others. As Whitmer explained in his report the relationship between final course grade and time spend in the LMS system tends to be higher in some courses. Out of 34,519 courses that were sampled, 7,648 appear to have a higher relationship between course grades and time spend on the LMS course portal.

Whitmer followed up on the report that was described above by investigating how successful students use the LMS. In this report they included data from over 900 institutions and 70,000 courses that used the gradebook feature. The first finding of this report is how it is mainly used by the entire population in the study, this can be seen in figure 2.3.

The findings of this study were interesting in regard to predicting student success from LMS data. Whitmer found that the most important LMS data features for predicting student success were:

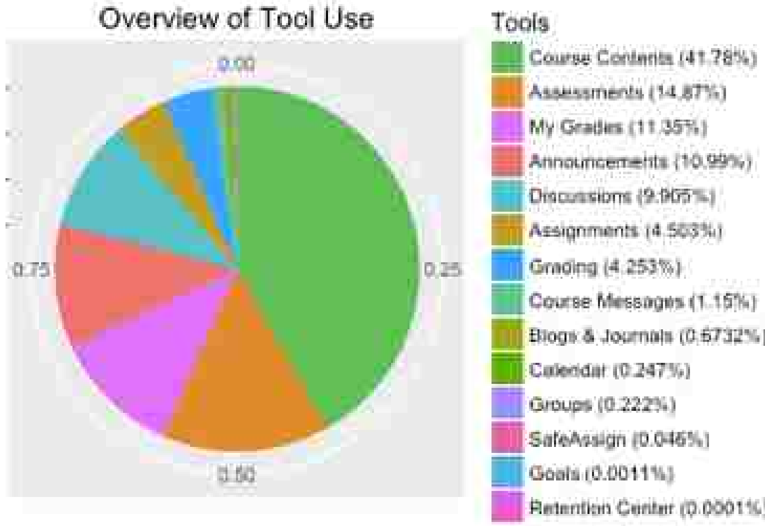

Figure 2.3: Student Interaction with LMS

- MyGrades
- Course Content
- Assessments
- Assignments
- Discussion

To support this conclusion please see figures 2.4, 2.5, 2.6 and 2.7. The student population is categorized into three groups according to their course performance: 80% or higher in the course, 60% to 79% and below 60%. Also, the x-axis variable in these figures is quantized into the following groups: high interaction (75%-100%), medium interaction (50%-74%), low interaction (25%-49%), really low interaction (0-24%) and no interaction.

Although Whitmer's reports did not provide a solution to solving student attrition, they did provide an indicator of student engagement that should be valuable if

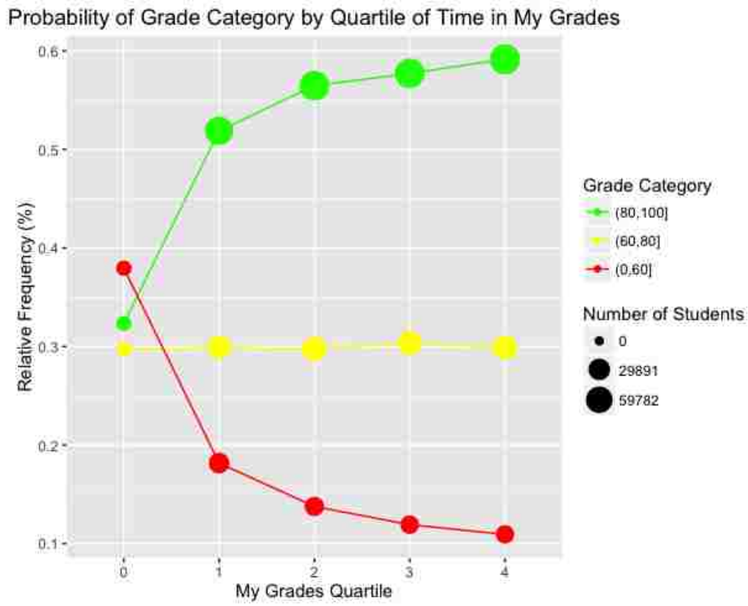

Figure 2.4: MyGrades to Course Final Grade Relationship

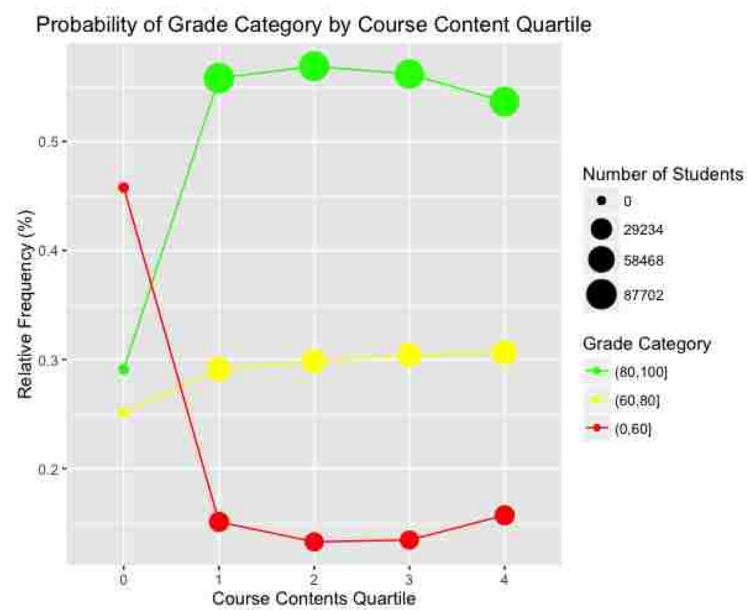

Figure 2.5: Course Contents to Course Final Grade Relationship

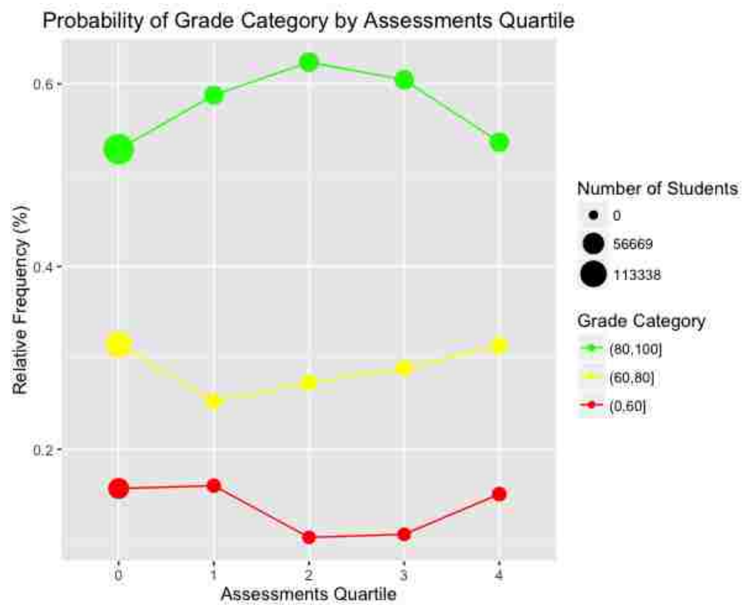

Figure 2.6: Assessments to Course Final Grade Relationship

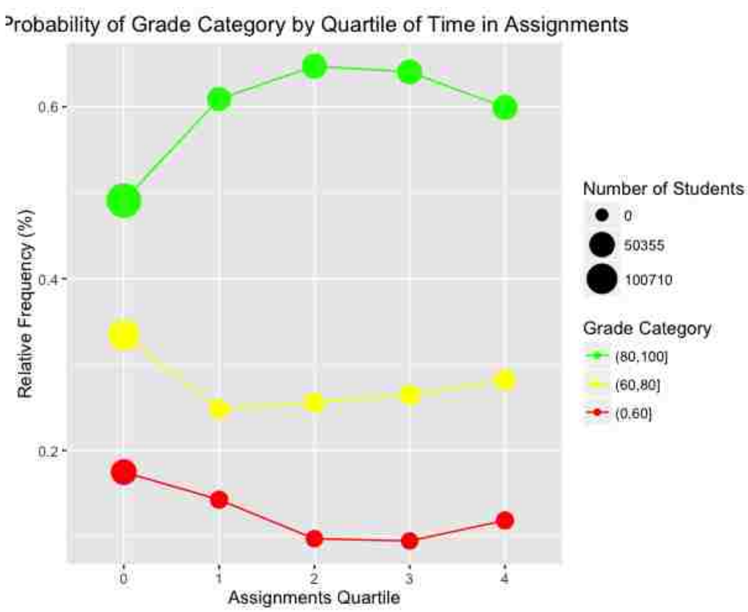

Figure 2.7: Assignments to Course Final Grade Relationship

implemented in an early alert system.

2.3 Situational Awareness, Ontology and Threat Model

Situational awareness was originally introduced to the military domain in order to provide information preeminence through information fusion. The reason for information preeminence is to guide military directors towards superior decisions in a military scenario. Through information fusion a computational system can help the decision maker cope with a military environment that has millions of sensors and that is rapidly changing. The most widely accepted definition for situational awareness is, "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future [8]." From this definition we can derive that in order to achieve situational awareness we do not only require the perception of the entire set of data that is dynamically changing, but we also require the reasoning of the data to give the decision maker a general view that is easier to comprehend. Situational Awareness combines ontologies, semantic web technologies and rule-based reasoning to give the decision maker a scenario that is uncomplicated to comprehend. Although to date this approach has not been applied to higher education, it has been applied to financial markets, transportation management, energy management, environmental control, disaster management and many other domains.

The most fundamental component to achieve computational situational awareness is the ontology of the scenario. As Pai et al. [17] describes it, "ontologies concentrate on defining classes/subclasses and characterizing the relationships among them and their instances." Even though we form the constraints for the ontology used, the data is what governs the overall structure of the ontology. There are many methods in which an ontology can be implemented onto the domain the user is constrained to. The simplest method is to have a single layer ontology on which every element of the data is mapped into a single ontology. Other domains call for a more advanced implementation of the ontology such as a multi-layer ontology merging paradigm. In this method data is divided into sub ontologies that may or may not be connected. The sub ontologies are united by top level ontologies that fuse data from both sources in order to achieve situational awareness.

After the ontology approach has been derived there is a need for a technology such as semantic web to give the machine the ability to decipher the meaning of words. The Semantic Web as formally introduced by Berners-Lee [4] is, “an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation.” The Semantic Web supports a set of tools and standards that give machines the ability to detect meaning in information on the web.

The last step towards situational awareness is to derive knowledge from our annotated data using rule-based reasoning. The threat model for the defined scenario is usually captured using rules which can be applied to the ontology to extract features that might indicate a threat in the given framework. Many programing languages have been introduced for rule-based reasoning with the most adopted being Semantic Web Rule Language (SWRL).

As previously stated, situational awareness has been widely adopted for military scenarios. Most recently, Pai et al. [17] introduced a multi-layer situational awareness model to achieve high-level data fusion and inference in a military situation. In this multi-layered approach many existing and publically known military ontologies were integrated to achieve situational awareness, as seen in figure 2.8. First, the Military Scenario Ontology (MSO) is were military scenarios (units, equipment, tactical

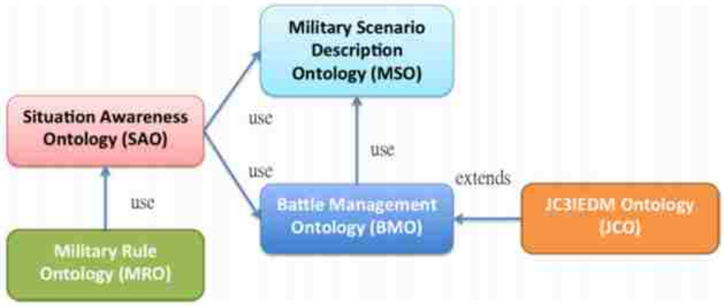

Figure 2.8: Ontology Integration Model

graphics, environment) are captured. The Battle Management Ontology (BMO) is where orders, reports and requests are managed. The BMO ontology is able to simulate behaviors in order to make predictions and run simulations. The JC3IEDM ontology is also introduced to this system in order to enable interoperability of systems. As explained in the previous paragraph a rule-based reasoning model is needed to achieve situational awareness. In this multi-layered approach the rules are separated into their own ontology, specifically the Military Rule Ontology (MRO). Lastly, what connects all of the ontologies previously mentioned is the Situation Awareness Ontology (SAO). In this ontology knowledge associated with all of the ontologies is gathered and presented to the end user in order to provide high-level reasoning.

In recent literature the applications of situational awareness have expanded beyond military applications. Applications in transportation management have also begun to form a growing research community. Barrachina et al. [3], recently revealed a situational awareness model introduced to transportation management. Much like the problem this document is attempting to address, in transportation management a system that enhances the efficiency of the scenario is needed. To be more specific, this work attempts to enhance traffic safety by implementing a single layer ontology approach to transportation. In this approach the system ingests data from the envi-

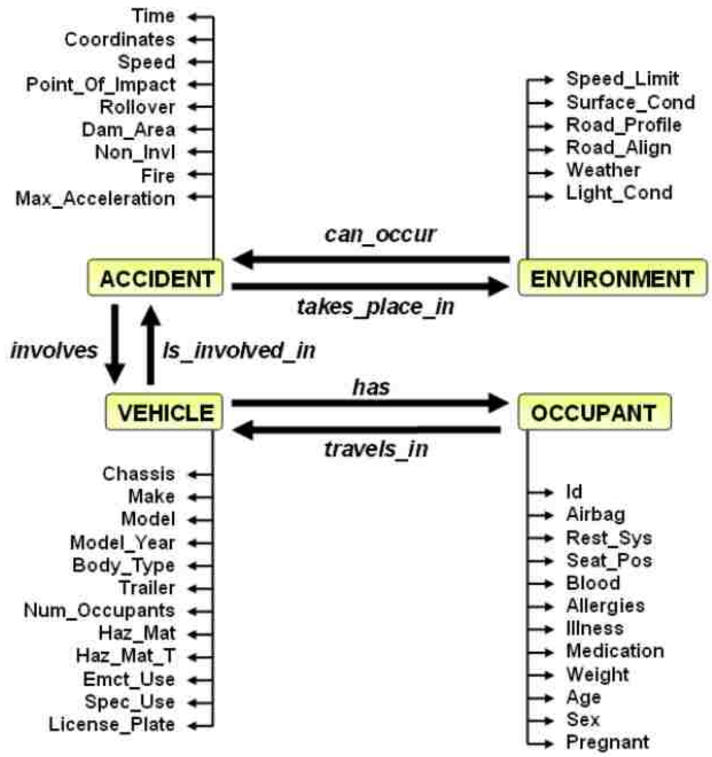

Figure 2.9: VEACON Ontology Model

ronment, a vehicle, occupancy of that vehicle, and if an accident has occurred into the ontology. The data captured by the ontology is shown on figure 2.9. Like the previous situational awareness scenario described, this system leverages a rule-based reasoner to report accidents that may have occurred.

2.4 Ontology Mapping

In the semantic web field, the importance of ontology-driven data conceptualization has been established and many ontologies have been developed for specific domain knowledge. Currently, this field is devoting more effort towards sharing information between ontologies in order to avoid data duplicity. Data sharing will not only solve data duplicity problems but it can also allow us to study the effects one ontology can have on another. The effects one ontology can have on another have recently become an area where a lot of research efforts have been dedicated, since none of the solutions presented in the literature appear to be acceptable. In the literature this problem is usually titled ontology mapping, as the approach taken is to map one ontology to another ontology. Once the the first ontology is mapped to the main situation ontology changes can be inflicted onto the scenario. The task of mapping one ontology to another might appear trivial at first glance, but this problem tends to get very complex. It gets complex specially when the same meaning in different ontologies can have different label names or a certain label can be in both ontologies but have different meanings. This problem only gets more complicated when ontologies with different taxonomy structures are introduced. Currently, the literature mainly presents approaches that are very limited as they can only map ontologies one-to-one [6, 7]. This is not a concrete solution as there are vast amounts of ontologies that can share information if a viable method is introduced.

Tang et al. [25] introduce the problem of ontology mapping and attempt to solve the problem as a directional problem. In this publication the example shown in figure 2.10 is given to aid in understanding the complexity of the problem. As seen in the figure, one ontology represents a section of the institutional structure at Washington University while the other represents a section of the structure at Cornell University. The goal of mapping these ontologies is to show which courses may be equivalent based on the institutional structure. As shown in this example by the dashed lines among these institutions there are equivalencies although the name of the entities at each institutions are different. In order to solve this complex problem, the author suggest a directional decision-based method for mapping similarities between ontologies. This problem is a directional problem as we want to know what changes one ontology has on the other. Also, in this method a decision problem is desired

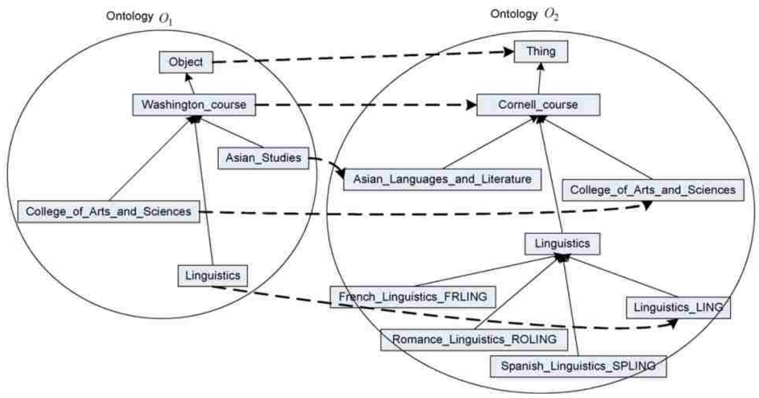

Figure 2.10: Example of two heterogeneous ontologies and their mappings.

as the need to discover optimal mappings with minimal risks is desired so Bayesian decision theory is employed. Tang et al. [25] proposes

> "an approach called Risk Minimization based Ontology Mapping (RiMOM) to conduct ontology mapping by running several passes of processing: first multi-strategy execution in which each decision find the mapping independently; and then strategy combination in which the mappings output by the independent decisions are combined; thirdly mapping discovery in which some mechanisms are used to discover the mapping based on the combined results. Mapping process can take place iteratively until no new mappings are discovered. In each iteration, user interaction can be used to refine the obtained mappings."

The RiMOM approach is illustrated in figure 2.11.

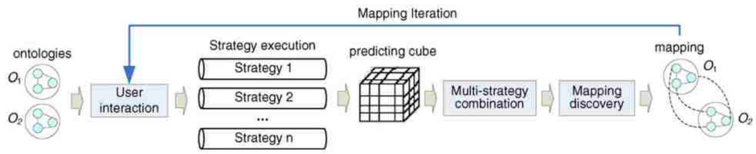

Figure 2.11: Mapping process in RiMOM.

2.5 Natural Language Processing

If the labels exemplifying the data in the ontology are descriptive and they represent the data object captured, then the data can be mapped according to the method described in the previous section. In order to complete the ontology mapping method, we need to introduce Natural Language Processing (NLP) to the algorithm. NLP is a topic of research that attempts to close the gap between computers and natural language. NLP attempts to give computers the ability to understand and manipulate natural written and spoken language. NLP gives computers the ability to give structure to text so that tasks such as summarization, translation, named entity recognition, relationship extraction, sentiment analysis, etc. can be completed computationally.

Chapter 3

Design and Implementation

3.1 Ontology

The first step toward the implementation of this early alert situational awareness model is to define the situational awareness ontology. As previously described, ontologies are specification of concepts and relationships that may exist in a given setting with those concepts. In order for a reasoner to reason about situations that might arise in this model a formal ontology is to be incorporated that captures the entire situation regarding a student at a given institution. In order to conclude with an ontology that allows for this type of reasoning there are various approaches we can take. We can attempt to list all of the classes and relationships among those classes and attempt to build our ontology. Another approach is to use a previously built university ontology. Much research has gone into to university ontologies and this research has one goal in mind, to represent a university structure as naturally as possible.

In this thesis, as a starting point, we will use the university ontology presented by Guo et al. [9]. This ontology represents the university setting in a very natural fashion. In order to use this ontology we will have to adapt this ontology to some of the scenarios we presented in the previous chapter.

First, we will start by describing the classes of the Bench University Ontology [9] and also describing the connections between the classes in this ontology. In an ontology a class is linked to subclasses and other classes with object properties. In table 3.1 a listing of ontology classes and subclasses is listed.

Next, we will list the object properties that connect the classes between each other and to its subclasses. In table 3.2 the object properties are listed along with the domain class and the range class. In ontologies the domain class points to the range class, signifying a belongs to relationship.

Lastly, in the provided ontology we must define data properties. Data properties spell out the specific features that populate the class or the object property. For example, our Undergraduate Student class has data property name, which is where you add the name of the student as a string. Data properties are provide in table 3.3.

As previously mentioned, we still need to add various classes, object properties and data properties to this ontology in order to achieve the task established at the beginning of this thesis. The proposed goal of this effort is to assign alert tags to undergraduate students that are in risk of academic failure. In table 3.4 the classes that need to be added to the university ontology employed for this investigation are listed. Similarly, object properties and data properties that need to be added to this ontology are listed in figures 3.5 and 3.6 respectively. A complete representation of the ontology can be found in figure 3.1 as well.

Figure 3.1: Early Alert Ontology

3.2 Resource Description Framework

Resource Description Framework (RDF) has been widely adopted as the the general purpose language for representing information on the web. The main goal of RDF is to provide a model which describes resources in a machine readable and standard format. RDF is a specifically structured eXtensible Markup Language (XML) file according to specified standards proposed by World Wide Web Consortium (W3C) [18]. In this framework data is structured in triples, with each element in the triple representing a specific purpose. An RDF triple is composed of a resource, a property type, and a value. The resource is any object that is uniquely identifiable. The property type represents the type of connection between a resource and its value. Finally, the value is the property that captures the description of the data, this can be a number, a string, etc. We will use the following sentence to illustrate an RDF triple:

Student John Doe has student identification number: 999999.

Figure 3.2: Example of an RDF Triple.

As shown in the figure 3.2 the resource is the student John Doe, the property type is "has" and finally the value is the student identification number. Once the ontology has been created the need for the situational awareness model to ingest data arises. Due to the capabilities of RDF, the data that is to be ingested by our ontology will need to be defined as RDF. There are various ways of completing this task but for this project we have opted to use the cellfie plugin within protege to map data into the ontology defined in the previous section. The reason cellfie will be used to convert our data to the ontology-driven RDF is because it is well documented and is easy to use. Cellfie gives us a user interface which allows us to convert data from the origin Comma Separated Values (CSV) file, pictured in figure 3.3 to RDF seen in code snippet below.

```
<owl:NamedIndividual rdf:about="http://swat.cse.
   lehigh.edu/onto/univ-bench.owl#Delmer">
 <rdf:type rdf:resource="http://swat.cse.lehigh.edu/
    onto/univ-bench.owl#UndergraduateStudent"/>
 <univ-bench:is_taking rdf:resource="http://swat.cse.
    lehigh.edu/onto/univ-bench.owl#CALCII-Delmer"/>
 <univ-bench:OWLDataProperty_000000000000000000041
    rdf:datatype="http://www.w3.org/2001/XMLSchema#
    decimal">2.153</univ-
```

```
        bench:OWLDataProperty_00000000000000000041>
    <univ-bench:has_attendance rdf:datatype="http://www.
        w3.org/2001/XMLSchema#integer">10</univ-
        bench:has_attendance>
    <univ-bench:has_lms_logins rdf:datatype="http://www.
        w3.org/2001/XMLSchema#string">[&quot;15-DEC-2017&
        quot;, &quot;14-DEC-2017&quot;, &quot;17-DEC-2017&
        quot;, &quot;11-DEC-2017&quot;, &quot; 9-DEC-2017&
        quot;, &quot;18-DEC-2017&quot;]</univ-
        bench:has_lms_logins>
    <univ-bench:lms_logins_amount rdf:datatype="http://
        www.w3.org/2001/XMLSchema#integer">6</univ-
        bench:lms_logins_amount>
    <univ-bench:name rdf:datatype="http://www.w3.org
        /2001/XMLSchema#string">Delmer</univ-bench:name>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="http://swat.cse.lehigh.
    edu/onto/univ-bench.owl#Emerson">
    <rdf:type rdf:resource="http://swat.cse.lehigh.edu/
        onto/univ-bench.owl#UndergraduateStudent"/>
    <univ-bench:is_taking rdf:resource="http://swat.cse.
        lehigh.edu/onto/univ-bench.owl#CALCII-Emerson"/>
    <univ-bench:OWLDataProperty_00000000000000000041
        rdf:datatype="http://www.w3.org/2001/XMLSchema#
        decimal">3.319</univ-
        bench:OWLDataProperty_00000000000000000041>
    <univ-bench:has_attendance rdf:datatype="http://www.
        w3.org/2001/XMLSchema#integer">8</univ-
```

	A	B	C	D	E	F	G	H
1	Student	Student_Cou	Attandence	GPA	LMS_Login	LMS_Login_c	Course	Classes
2	Ethelyn	CALCII-Ethely	6	3.951	["14-DEC-201	10	Calc II	10
3	Marco	CALCII-Marco	7	3.448	["18-DEC-201	11	Calc II	10
4	Orie	CALCII-Orie	8	3.074	[" 8-DEC-201	17	Calc II	10
5	Myron	CALCII-Myroi	8	3.175	["10-DEC-201	4	Calc II	10
6	Guillermo	CALCII-Guille	7	3.177	[" 6-DEC-201	5	Calc II	10
7	Timothy	CALCII-Timot	10	3.531	["12-DEC-201	19	Calc II	10
8	Emerson	CALCII-Emers	8	3.319	["15-DEC-201	1	Calc II	10
9	Shanel	CALCII-Shane	7	3.974	[" 5-DEC-201	1	Calc II	10

Figure 3.3: Example of CSV file.

```
      bench:has_attendance>
    <univ-bench:has_lms_logins rdf:datatype="http://www.
      w3.org/2001/XMLSchema#string">[&quot;15-DEC-2017&
      quot;]</univ-bench:has_lms_logins>
    <univ-bench:lms_logins_amount rdf:datatype="http://
      www.w3.org/2001/XMLSchema#integer">1</univ-
      bench:lms_logins_amount>
    <univ-bench:name rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">Emerson</univ-bench:name>
</owl:NamedIndividual>
```

3.3 Advisor Notes

In order to computationally infer any knowledge about advisor notes we first need to provide the ability to understand natural written language. As explained in sections 2.4 and 2.5 we can allow an ontology to infer information about natural language by using NLP algorithms and ontology mapping. For this investigation, we will examine using natural language techniques that are to be completed before the information is passed to the ontology so that reasoning only occurs once. If we attempt to map information from one ontology to the other, as described in section 2.4, the risk of losing information presents itself. The method of pre-analyzing advisor notes also imposes difficulties as well, as not only do we have to pass the most amount of information to the ontology but we have to pass all of this information pre-analyzed and classified with the correct gauges.

In the NLP research field there are many methods of analyzing human written (natural language) data. These methods give machines the ability of deducing insight on to what information is contained in the written language. But in this composition we want to allow the ontology to infer the information therefore, we only want to pre-analyze the natural language data. There is an abundant number of tools, libraries and APIs that analyze natural language using different programming languages and different algorithms. For the purpose of analyzing advisor notes we will employ both the Nartural Laguage Toolkit (NLTK) and TextBlob which are both written in Python. TextBlob will allow us to rapidly assign sentiment scores to advisor notes. We will assign two types of scores to advisor notes, subjectivity and polarity. In this library the subjectivity score ranges from 0.0 to 1.0, where 0.0 is very objective and 1.0 is very subjective. This library also assigns polarity scores which range from -1.0 to 1.0 where -1.0 is very negative and 1.0 is very positive. It is important to mention that this library assigns these scores using a Naive Bayes analyzer. This Naive Bayes analyzer is trained using movie review data. To illustrate the functionality of this analyzer please refer to the following sentences:

"I disagreed with many of President Barack Obama's policies, but he was a good role model," – Rep. Ryan Costello

"Trump may hate Amazon, but there's little he could do to it." – Elizabeth Weise and David Jackson, USA TODAY

As one can read in both of the statements above they are more subjective than objective. That is, both of these statements seem to be open to interpretation based on someone's personal feelings, emotions, etc. To test the functionality of TextBlob, the sentiment analyzer, both of the statements were passed through the algorithm and we obtained subjectivity scores of 0.55 and 0.7 respectively. These scores align with the definition of subjectivity as one may also derive that the first statement is less subjective than the second statement. Lastly, one of the measures this sentiment analyzer provides, and the one we are mainly interested in, is the polarity of a statement. We want to know if an advisor is providing negative or positive notes to a student as this might include information that may require a student to receive an early alert. Opposite of the subjectivity score of the statements presented above, one can see that both of the previous statements are on the opposite side of the spectrum when it comes to polarity. The first statement discussed what a good role model the previous president was while the other statement discusses hate from the current president towards Amazon (an American e-commerce website). Both of the statements were also processed for polarity yielded the scores shown in table 3.7.

The last fragment of information we want to pre-process from the advisor notes before they are passed to the ontology is the topic of each sentence in the note. For this there are also many tools, algorithms and APIs but in this experimentation we will simply use the NLTK and the brown corpus [5]. A trained NLTK tagger has the ability to tag words in a sentence based on their usage within the sentence. Much like we all had to learn how to determine the grammar of a sentence (verbs, adjectives, etc.) with NLTK we give machines the ability to know how each word is used within a sentence. This can be a complicated task as a word can have different meanings depending on the usage within the sentence. To show the functionality of a tagger and how we will be using it to obtain the topic of each sentence please refer to the sentence below:

"A player always aspires to the top, and Madrid is always one of the highest ambitions for any footballer. Every player would love to get to the absolute top and

Real Madrid is it." – Luis Suarez

After analyzing this same sentence with a tagger we obtain the following:

```
[('A', 'AT'), ('player', 'NN'), ('always', 'RB'), ('aspires
    ', 'NN'), ('to', 'TO'), ('the', 'AT'), ('top', 'JJ'),
    (',', ','), ('and', 'CC'), ('Madrid', 'NNP'), ('is', 'BEZ
    '), ('always', 'RB'), ('one', 'CD'),('of', 'IN'), ('the',
    'AT'), ('highest', 'JJT'), ('ambitions', 'NN'), ('for', '
    IN'), ('any', 'DTI'), ('footballer', 'NN'), ('.', '.'), ('
    Every', 'AT'), ('player', 'NN'), ('would', 'MD'), ('love',
     'VB'), ('to', 'TO'),('get', 'VB'), ('to', 'TO'), ('the',
    'AT'), ('absolute', 'JJ'), ('top', 'JJ'), ('and', 'CC'),
    ('Real', 'NNP'), ('Madrid', 'NNP'), ('is', 'BEZ'), ('it',
    'PP'), ('.', '.')]
```

The tags next to each word in the sentence are described in appendix A. Once a sentence is tagged we will obtain the noun phrases and verb phrases from the tagged sentence. We will filter the noun phrases and verb phrases because as we have learned in linguistics the noun phrase is usually the topic or object in the sentence and the verb phrase is used to describe action within those objects.

For our use case we opted to build our own tagger for topic extraction because using some of the libraries and tools available ignored a number of occurrences that are important in our use case. For example, if an advisor had written a note similar to, "Student earning a C- in MATH 162 Calculus I" all of the taggers would ignore "C-" because they were not built to capture such university specific cases. This problem also repeated itself on course numbers like in the example presented above. The first step is to build our own tagger with NLTK and the brown corpus. To build our tagger we first define a general expression tagger as we can assign tags to text simply catching patterns in text such as “ing”. In the English language words that end with “ing” are usually present participle verbs. After the results are obtained from the general expression tagger they are passed to the NLTK UnigramTagger, which tags words based on a simple statistical tagging algorithm that has been trained with the brown corpus data. Lastly, the results from the UnigramTagger are passed to the NLTK BigramTagger which is also trained with the brown corpus data. The BigramTagger is a tagger that tags a word in a sentence based on its usage within a sentence.

The next step in topic extraction is to define our noun phrases. We will define noun phrases as seen in table 3.8 and we will use those noun phrases as rules to which we will compare the analyzed advisor notes.

3.4 Threat Model

After the ontology is defined and all of the data, including the pre-processed advisor notes, is passed to the ontology reasoning must occur in order to achieve situational awareness. Rule-based reasoning is how we will achieve this. In rule-based reasoning we define rules that will capture specific scenarios, which for the purpose of this thesis it will be assigning early alert flags to academically troubled students. The World Wide Web Consortium (W3C) proposed Semantic Web Rule Language (SWRL) as the standard rule language for the Semantic Web. SWRL is composed of a combination of the OWL DL and OWL Lite sublanguages along with Rule Markup Language (RuleML). To demonstrate the usability of SWRL please reference the following example: We have an ontology in which a student is enrolled in many classes and also this student obtains a failing grade in one class. When this situation occurs we simply want to add this student to list of students in academic probation.

```
UndergraduateStudent(?x) ^ is_enrolled_in(?x, ?course) ^
    earned_grade_in_course(?y, ?course) ^ swrlb:lessThan(?
    course, "C-") -> AcademicProbation(?x)
```

We will use SWRL to define the rules for our situation in order form alerts based on the scenarios presented in section 2.1. In section 2.1 it was established that the main reason a student will become academically in danger in a course is because of lack of attendance therefore an early alert tag will be assigned to this student. Also, as stated in the section previously referenced some students become fall at risk academically for failed exams and incomplete assignments although not as such an alarming rate as course absences. To capture these simple scenarios we will define our rules similarly to the rule described above. The first rule we want to define is the course attendance. We want to assign an early alert to any student that has been to class less than 50% of the total course time.

```
UndergraduateStudent(?x) ^ has_attendance(?x, ?atten) ^
    is_taking(?x, ?course) ^ is_course(?course, ?c) ^
    days_taught(?c, ?taught) ^ swrlb:integerDivide(?ratio, ?
    atten, ?taught) ^swrlb:lessThan(?ratio, "0.5"^^xsd:float)
    ^ early_alert(?x)
```

The rules implemented to capture students that have failed an exam and that have submitted incomplete assignments can be found on table 3.9.

Another factor that we wanted to capture in this thesis is LMS interaction, as it has recently come under scrutiny as its close relation to student failure has been quantified. As described in section 2.2 fewer log-ins and less interaction with the LMS portal tend to indicate that the student will have a lower final grade in the course. We want to capture and assign early alert flags to students that don?t interact with their LMS portal and we will define it as follows:

```
UndergraduateStudent(?x) ^ lms_logins_amount(?x, ?number) ^
    lms_logins_amount ^ swrlb:lessThan(?number, "10"^^xsd:int)
    ^ early_alert(?x)
```

In this rule we want to assign an early alert to any student that has logged on to their LMS course portal less than 10 times. As mentioned previously in this chapter, we also want to reason over advisor notes, with the help of NLP algorithms, in this framework. Once the advisor notes are passed to the ontology we can allow the ontology to assign early alerts to student in a couple of manners. In one rule we want to assign early alerts to students that have both a negative polarity score and a high subjectivity score as shown in the rule below.

```
UndergraduateStudent(?x) ^ advisor_note(?x, ?note) ^
    subjectivity_rating(?note, ?sub) ^ swrlb:greaterThan(?sub,
    0.4) ^ sentiment_rating(?note, ?sent) ^ swrlb:lessThan(?
    sent, -0.4) -> early_alrert(?x)
```

In the second rule we want to reason over the topics extracted from the advisor notes. This is also described in a previous section in this chapter. We will reason over advisor notes by building a dictionary of words or phrases that should be alarming if an advisor highlights them in the advising session notes. These phrases might include “not passing”, “failing course”, etc. We want to assign a student an alert if one of these phrases is extracted from the advisor notes. An example of a rule can be seen below.

```
UndergraduateStudent(?x) ^ advisor_note(?x, ?note) ^
    has_main_topic(?note, ?main) ^ swrlb:containsIgnoreCase(?
    main, "Non-Passing") -> -> early_alrert(?x)
```

The unique flexibility and the reason we chose rule driven situational awareness, is due to the modifiability it provides during implementation. In this type of situational awareness model we are able to manipulate rules that will capture radical use cases. This modifiability provides more flexibility than any other platform. In the setting we are implementing it is important to have this flexibility as we want to achieve a higher level situational awareness by combining rules. For example, if we have multiple students at an institution that have less than desirable attendance but not excessive absenteeism in any other implementation these students will “go under the radar” as their absenteeism will not provide an alarm as it is not excessive. But if we combine less than desirable absenteeism with both low LMS interaction and low assignment scores this student should invoke an early alert as this combination might contribute to a failing grade at the end of the term.

```
UndergraduateStudent(?x) ^ lms_logins_amount(?x, ?number) ^
    has_attendance(?x, ?atten) ^ is_taking(?x, ?course) ^
    is_course(?course, ?c) ^ days_taught(?c, ?taught) ^
    has_assigment(?c, ?assign) ^ max_score(?assign, ?max) ^
    has_assigment(?course, ?studentassign) ^ score(?
    studentassgin, ?stuscore) ^ lessThan(?atten, ?taught) ^
    swrlb:lessThan(?number, "5"^^xsd:int) ^ swrlb:lessThan(?
    stuscore, ?max) -> early_alert(?x)
```

Due to the flexibility of rule-based reasoning we wanted to assign early alert risk levels. In section 2.1 we learned that a student with low course attendance is much more at risk than a student that earns a low grade in a course. Also, we learned that LMS course portal interaction provides information regarding the studentś overall performance in the course therefore making this low interaction a high alert. It is to be advised that we applied risk levels in this ontology as a proof of concept and further research needs to be completed to assign risk levels appropriately.

Class	Subclass of
Organization	
Institute	Organization
Program	Organization
Research Group	Organization
School	Organization
University	Organization
University Department	Organization
Person	
Chair	Person
Director	Person
Employee	Person
Student	Person
University Research Assistant	Person
University Teaching Assistant	Person
Administrative Staff Worker	Employee
Faculty Member	Employee
Clearical Staff Worker	Administrative Staff Worker
Systems Staff Worker	Administrative Staff Worker
Lecturer	Faculty Member
Post Doctorate	Faculty Member
Professor	Faculty Member
Assistant Professor	Professor
Associate Professor	Professor
Chair	Professor
Dean	Professor
Full Professor	Professor
Visiting Professor	Professor
Graduate Student	Student
Undergraduate Student	Student
Publication	
Article	Publication
Book	Publication
Manual	Publication
Published Specification	Publication
Software Program	Publication
Unnoficial Publication	Publication
Work	
Research Work	Work
Teaching Course	Work
Graduate Level Courses	Teaching Course
Schedule	

Table 3.1: Bench University Ontology Classes

Object property	Domain	Range
has a degree from	Person	University
has a doctoral degree from	Person	University
has a masters degree from	Person	University
has an undergraduate degree from	Person	University
has as a member	Oganization	Person
has as a research project	Research Group	Research Work
has an alumnus	University	Person
is a teaching assistant for	University Teaching Assistant	Teaching Course
is about	Publication	Research Work
is affiliated with	Oganization	Person
is affiliated with	Oganization	Organization
is being advised by	person	professor
is documented in	Software Program	Publication
is part of	Organization	Organization
is tenured	Professor	
is version	Software Program	
is taking		
lists as a course	Schedule	Teaching Course
member of	has as a member	
works for	member of	
is the head of	works for	
publishes	Organization	Publication
teaches	Faculty Member	Teaching Course
was written by	Publication	Person
was written on	Publication	

Table 3.2: Bench University Ontology Object Properties

Data Property	Domain
can be reached at	Person
is age	Person
is researching	
name	
office room No.	
telephone number	Person
title	Person

Table 3.3: Bench University Ontology Data Properties

Class	Subclass of
AdvisorNote	
Risk Level	
High Risk	Risk Level
Medium Risk	Risk Level
Low Risk	Risk Level
Student Course	
Student Course Work	Student Course

Table 3.4: Added Ontology Classes

Object property	Domain	Range
has advisor note	Undergraduate Student	Advisor Note
has assignment	Teaching Course	Student Course Work
has assignment	Student Course	Course Work
is assignment	Student Course Work	Course Work
is course	Student Course	Teaching Course

Table 3.5: Added Object Properties

Data Property	Domain
days taught	Teaching Course
has attendance	Student Course
has GPA	Person
has lms logins	Student Course
lms logins amount	has lms logins
has main subject	Advisor Note
has note	Advisor Note
has sentiment rating	Advisor Note
max score	Course Work
minimum rating sentence	Advisor Note
number of attendance	
score	Student Course Work
subjectivity rating	Advisor Note

Table 3.6: Added Data Properties

Statement by	Polarity Score
Rep. Ryan Costello	0.6
Elizabeth Weise and David Jackson	-0. 49375

Table 3.7: Statement Polarity Score

NLTK TAGS	Noun Phrase
NNP+NNP	NNP
NN+NN	NNI
NNI+NN	NNI
JJ+JJ	JJ
JJ+NN	NNI

Table 3.8: Noun phrases form NLTK tags

Rule	SWRL Code	Type of Alert
No engagement with LMS, course absences and low assigment or test scores	UndergraduateStudent(?x) ^lms_logins_amount(?x, ?number) ^has_attendance(?x, ?atten) ^is_taking(?x, ?course) ^is_course(?course, ?c) ^days_taught(?c, ?taught) ^has_assigment(?c, ?assign) ^max_score(?assign, ?max) ^has_assigment(?course, ?studentassign) ^score(?studentassgin, ?stuscore) ^swrlb:lessThan(?atten, ?taught) ^swrlb:lessThan(?number, 5) ^swrlb:lessThan(?stuscore, ?max) ->high-risk(?x)	High Risk
Low test or assigment scores and CAPS advisor notes	UndergraduateStudent(?x) ^is_taking(?x, ?course) ^is_course(?course, ?c) ^has_assigment(?c, ?assign) ^max_score(?assign, ?max) ^has_assigment(?course, ?studentassign) ^score(?studentassgin, ?stuscore) ^swrlb:lessThan(?stuscore, ?max) ->low-risk(?x)	Low Risk
Advisor notes with topic extraction	advisor_note(?x) ^has_main_subject(?x, ?main) ^swrlb:containsIgnoreCase(?main, ”Non-Passing”) ->low-risk(?x)	Low Risk
Advisor notes with low sentiments score and high subjectivity scores	advisor_note(?x) ^subjectivity_rating(?x, ?sub) ^swrlb:greaterThan(?sub, ”0.4”^^xsd:decimal) ^has_sentiment_rating(?x, ?sent) ^swrlb:lessThan(?sent, ”-0.4”^^xsd:decimal) ->medium-risk(?x)	Medium Risk
Low GPA	UndergraduateStudent(?x) ^has_GPA(?x, ?gpa) ^swrlb:lessThan(?gpa, 2.5) ->medium-risk(?x)	Medium Risk

Table 3.9: Rules Implemented to Capture Early Alerts

Chapter 4

Results

We have applied the situational awareness model presented in this thesis with the help of Protege [16]. In this implementation we have simulated the generic student data such as LMS interaction, course attendance, grades, etc. We have used advisor notes that were written by advisors at the University of New Mexico for their assigned students. It is to be noted that all of the advisor notes have been anonymized in order to keep student information confidential. Data was simulated for 100 students and was passed to the ontology.

To demonstrate the implementation presented in this thesis we will have three different examples. The first is students that have been categorized as high risk based on the rules presented in table 3.9. Figure 4.1 shows the categorization of students that had no engagement with LMS, excessive course absences, and low assignment or test scores as high risk.

Similarly, in figures 4.2 and 4.3 we demonstrate the students that were categorized based on the rules listed in table 3.9 to their respective groupings.

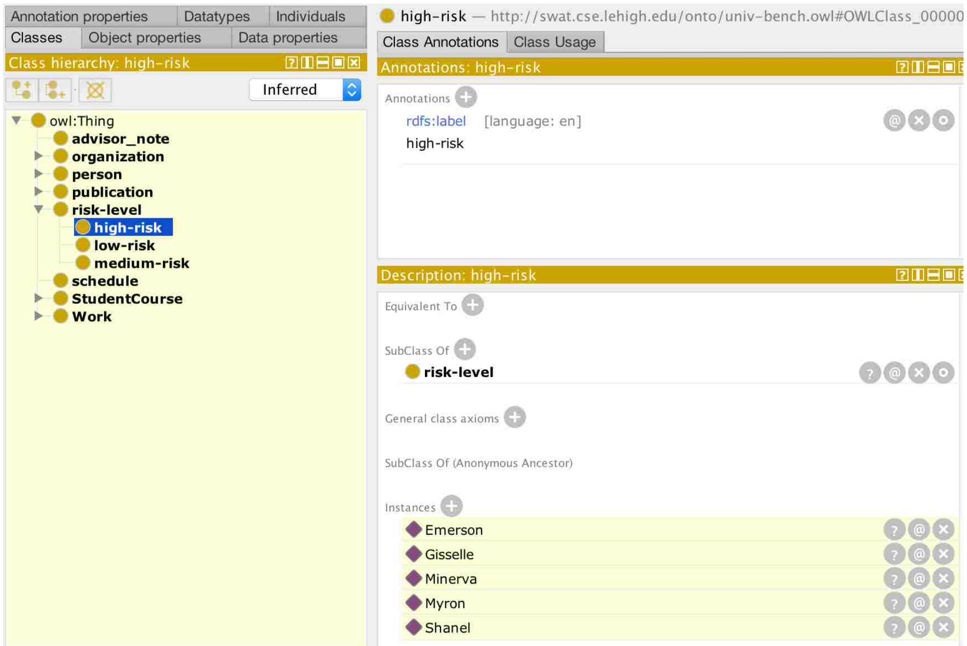

Figure 4.1: Example High Risk Output from Protege.

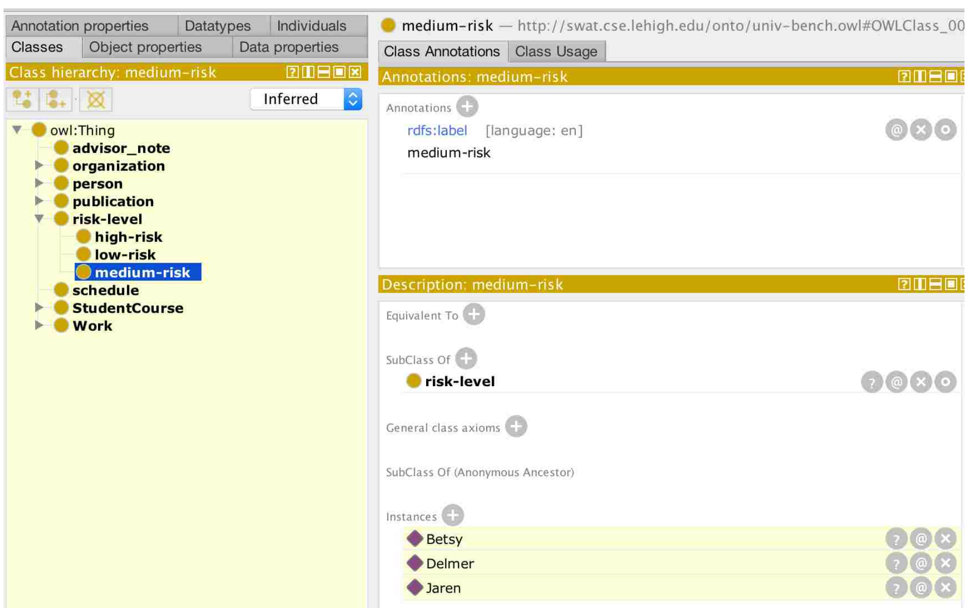

Figure 4.2: Example Medium Risk Output from Protege.

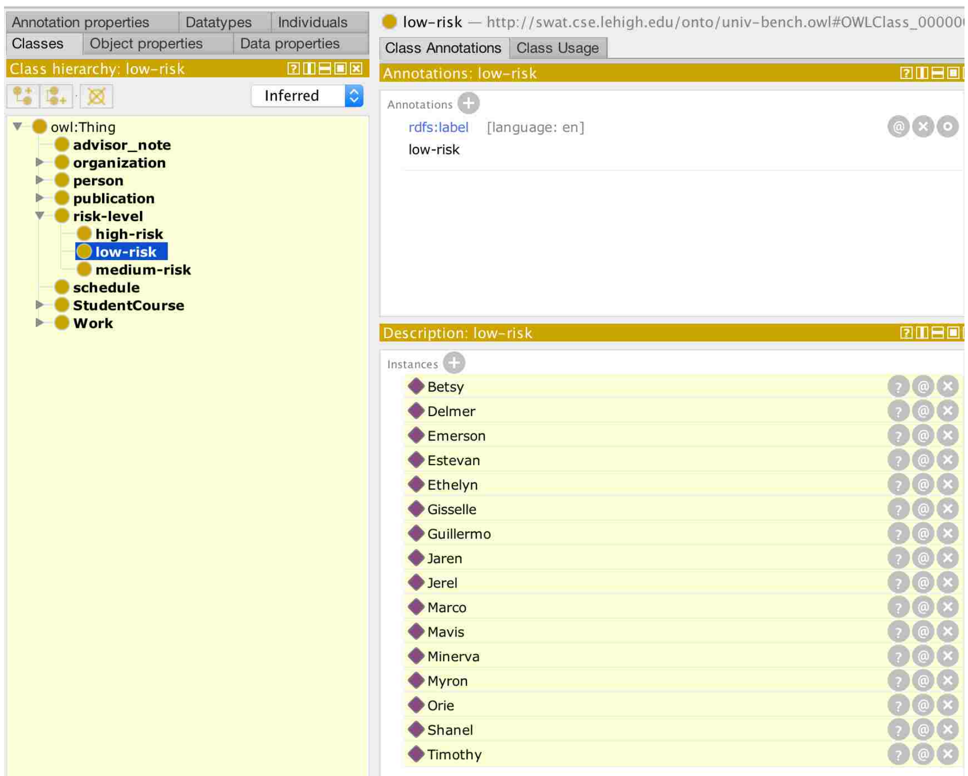

Figure 4.3: Example Low Risk Output from Protege.

Chapter 5

Discussion

The results obtained in the experimentation were promising although adding uncertainty to the rules would provide more assurance over the results obtained. As we learned in chapter 3, the Naive Bayes classifier is trained with movie review data which may not be sensitive to our use case. In order to achieve more accurate results training this classifier with data that is more sensitive to our use case is vital. In our experimentation we composed an algorithm to perform topic extraction on advisor notes but we noticed that in some cases we were losing information that might be important for early alerts. So the need to experiment with a semantic reader instead of topic extraction became important. In chapter 2 one can observe some of the predictive capabilities that using in-depth University LMS can provide. Unfortunately, during the time of experimentation this data was inaccessible therefore we were unable to experiment with some of the capabilities we learned about.

Appendices

A Part-Of-Speech Tagset

Appendix A

Part-Of-Speech Tagset

Table A.1: POS Tag Description

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	d?hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)

MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive to?	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are

VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have

VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
”	Quotation marks	”
“	Opening quotation marks	“
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	-;:...

References

[1] Ipeds: Integrated postsecondary education data system. Accessed: 2017-05-2.

[2] U.s. census bureau. Accessed: 2017-05-2.

[3] Javier Barrachina, Piedad Garrido, Manuel Fogue, Francisco J Martinez, Juan-Carlos Cano, Carlos T Calafate, and Pietro Manzoni. Veacon: A vehicular accident ontology designed to improve safety on the roads. *Journal of Network and Computer Applications*, 35(6):1891–1900, 2012.

[4] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):1–5, 2001.

[5] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

[6] Hong-Hai Do and Erhard Rahm. Coma: a system for flexible combination of schema matching approaches. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 610–621. VLDB Endowment, 2002.

[7] AnHai Doan, Pedro Domingos, and Alon Y Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *ACM Sigmod Record*, volume 30, pages 509–520. ACM, 2001.

[8] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[9] Yuanbo Guo, Zhengxiang Pan, and Jeff Heflin. Lubm: A benchmark for owl knowledge base systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3), 2005.

[10] Dana L Heisserer and Phil Parette. Advising at-risk students in college and university settings. *College student journal*, 36(1), 2002.

[11] William E Hudson Sr. Can an early alert excessive absenteeism warning system be effective in retaining freshman students? *Journal of College Student Retention: Research, Theory & Practice*, 7(3):217–226, 2005.

[12] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[13] Andrew P Kelly, Mark Schneider, and Kevin Carey. Rising to the challenge: Hispanic college graduation rates as a national priority. *American Enterprise Institute for Public Policy Research*, 2010.

[14] Laura G Knapp, Janice E Kelly-Reid, and Scott A Ginder. Enrollment in postsecondary institutions, fall 2009; graduation rates, 2003 & 2006 cohorts; and financial statistics, fiscal year 2009. first look. nces 2011-230. *National Center for Education Statistics*, 2011.

[15] George D Kuh. How to help students achieve. *Chronicle of Higher Education*, 53(41):B12–B13, 2007.

[16] Mark A Musen. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12, 2015.

[17] Fang-Ping Pai, Lee-Jang Yang, and Yeh-Ching Chung. Multi-layer ontology based information fusion for situation awareness. *Applied Intelligence*, pages 1–23, 2016.

[18] Jeff Z Pan. Resource description framework. In *Handbook on Ontologies*, pages 71–90. Springer, 2009.

[19] Elizabeth Pfleging. An evaluation of the early alert program at columbia college. 2002.

[20] Mark Schneider. Finishing the first lap: The cost of first year student attrition in america's four year colleges and universities. *American Institutes for Research*, 2010.

[21] Mark Schneider and Lu Yin. The hidden costs of community colleges. *American Institutes for Research*, 2011.

[22] Mark Schneider and Lu Yin. The high cost of low graduation rates: How much does dropping out of college really cost?. *American Institutes for Research*, 2011.

[23] Ahmad Slim, Jarred Kozlick, Gregory L Heileman, and Chaouki T Abdallah. The complexity of university curricula according to course cruciality. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2014 Eighth International Conference on*, pages 242–248. IEEE, 2014.

[24] Dale R Tampke. Developing, implementing, and assessing an early alert system. *Journal of College Student Retention: Research, Theory & Practice*, 14(4):523–532, 2013.

[25] Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. Using bayesian decision for ontology mapping. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(4):243–262, 2006.

[26] Patrick T Terenzini and Ernest T Pascarella. Toward the validation of tinto's model of college student attrition: A review of recent studies. *Research in Higher Education*, 12(3):271–282, 1980.

[27] Vincent Tinto. *Leaving college: Rethinking the causes and cures of student attrition.* The University of Chicago Press, 1993.

[28] John Whitner. How successful students use lms tools confirming our hunches, 2016.

[29] John Whitner. Research in progress: Learning analytics at scale for blackboard learn, 2016.