TEXT CLASSIFICATION OF STUDENT PREDICATE USE FOR CONCEPTUAL CHANGE ASSESSMENT

By

Brian A. Landrón-Rivera

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

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS

December, 2016

Approved by:

Aidsa I. Santiago-Román, Ph.D Co-Chair, Graduate Committee	Date
José Fernando Vega Riveros, Ph.D Member, Graduate Committee	Date
Nayda G. Santiago-Santiago, Ph.D Chair, Graduate Committee	Date
Hilton Alers Valentín, Ph.D Representative of Graduate Studies	Date
José Colom Ustariz, Ph.D Chairperson of the Department	Date

ABSTRACT

In the field of educational engineering concepts have been placed into ontological categories that reflect the nature of those concepts. Furthermore learner misconceptions occur when students assign a concept to an incorrect ontological category within their individual mental models. Predicate tests used to estimate the categorization of student conceptions have proven to be successful for conceptual change assessment but they have not been automated with the use of modern computing systems. The main goal of this research is to show how predicate test automation is possible by applying knowledge discovery in databases theory to a previously annotated dataset to achieve. The secondary goal is to find which data mining techniques can be used to extract a feature set that yields high quality text classification results. This thesis documents how the predicate test can be automated with knowledge discovery in databases techniques using data from engineering students enrolled in a U.S. midwestern public institution.

RESUMEN

En el campo de la ingeniería de la educación los conceptos transmitidos o adquiridos durante el aprendizaje se han colocado en ciertas categorías ontológicas que van de acuerdo con la naturaleza de cada concepto. Además la concepción errónea de conceptos se refiere al fenómeno donde se adquieren conceptos nuevos y la mente los asigna a una categoría que no va en acorde con la categorización correcta establecida por los expertos de la ingeniería de la educación. Para el avalúo de concepciones erróneas se han utilizan las pruebas de predicado para estimar la categorización de las concepciones de los estudiantes pero estas técnicas no han sido implementadas utilizando sistemas de computación modernos. El propósito principal de esta investigación es mostrar que las pruebas de predicado se pueden automatizar aplicando la teoría de descubrimiento de conocimiento en bases de datos a un conjunto de datos que contengan anotaciones previas para lograr aprendizaje supervisado. La meta secundaria será determinar las técnicas de minería de datos que podrían extraer un conjunto de características que produzcan resultados de clasificación de alta calidad. Esta tesis documenta cómo se puede automatizar el cómputo de las pruebas de predicado utilizando técnicas de descubrimiento de conocimiento en bases de datos con datos de estudiantes de ingeniería matriculados en una intitución pública del medio oeste de E.E.U.U.

To my son Ethan Allen, my daughter Charlotte Zoé, and my wife Cristal.

ACKNOWLEDGMENTS

I want to specially thank my chair Nayda Santiago for guiding me in the art of high quality research ever since my days as an undergraduate. Thank you. I deeply appreciate your ever present support.

I also would like to thank my co-chair Aidsa Santiago for all her contributions to my understanding of conceptual change theory and Fernando Vega for all his contributions to my understanding of ontologies and machine learning concepts.

Finally I would like to thank all collaborators involved in the work supported by the National Science Foundation under grant no. EEC-0550169 "Developing Ontological Schema Training Methods to Help Students Develop Scientifically Accurate Mental Models of Engineering Concepts", including but not limited to Dazhi Yang, Ruth Streveler, and Ronald Miller.

TABLE OF CONTENTS

LIST OF TABLES

Table page

LIST OF FIGURES

Figure page

LIST OF ABBREVIATIONS

SVM	Support Vector Machine.
ITS	Intelligent Tutoring Systems.
KDD	Knowledge Discovery in Databases.
EDM	Educational Data Mining.
BSM	Bayesian Student Model.
PTAP	Predicate Test Automation Pipeline.
WEKA	Waikato Environment for Knowledge Analysis.
SMOTE	Synthetic Minority Oversampling Technique.
SMO	Sequential Minimal Optimization.
TD-IDF	Term Document Inverse Document Frequency.
TP	True Positive.
FP	False Positive.
TN	True Negative.
FN	False Negative.

CHAPTER 1
INTRODUCTION

The U.S. Department of Education has long established that students must meet the requirements associated with their current K-12 educational standards in order to be competent in their subsequent grades. According to Pellegrino, Chudowski, and Glaser [9] students tend to develop knowledge gaps due to misconceptions. These misconceptions are the result of incorrect delivery of instruction and assessment techniques [10]. When instruction is not accompanied by cognitive assessment, these misconceptions are hard to detect. Thus, they must be addressed with real-time assessment in order to avoid learning difficulties in new topics [9].

To optimize transfer of knowledge and to improve academic achievement educational systems must ensure that misconceptions are corrected in a way that students can correctly apply newly acquired knowledge to new contexts [11]. As soon as misconceptions about any given domain are identified and conceptual change occurs students are able to improve their perception and critical thinking about a concept [11][9]. Helping students apply newly acquired knowledge results in an efficient transfer of knowledge form teachers to students [11].

According to Chi learning can occur under the following circumstances regarding the topic under study: (1) the learner has no prior knowledge, (2) the learner has some correct prior knowledge, and (3) the learner has some incorrect prior knowledge [12][13]. When learners have this incorrect prior knowledge they encounter difficulty understanding science concepts because they tend to incorrectly categorize their ideas [1][6][14][15] (e.g. considering electric current to be a substance that

can be contained in a battery when in fact it is an equilibrium seeking process). Thus the third circumstance for learning must include conceptual change for knowledge acquisition to occur.

Chi et. al. [1][6][14][15][12][16] have made substantial contributions to modern conceptual change theory. Over the past two decades the focus of their research has closely aligned with a process called the predicate test. Predicate tests consist of extracting student predication used to describe a concept and comparing it to correct predication used by experts. By examining the correctness of student verbal predication found in their description of answers to multiple-choice question answers teachers can identify each student's mental categorization of concepts [1][6][14][15][12][16].

The predicate test is a state-of-the-art approach for misconception assessment. It has the potential to become a powerful tool for all types of classrooms if it becomes readily accessible to teachers. Its caveat is revealed once you consider the fact that the predicate test heavily depends on educational engineering expertise to discern which conceptual category students are assigning to acquired knowledge to the correct category.

During this research it was possible to automate the predicate test process using knowledge discovery in databases (KDD) theory by performing text mining using text classification. Our proposed approach is named the Predicate Test Automation Pipeline (PTAP) and is based on determining if a student's mental categorization of concepts can be predicted using text classification. The PTAP was designed to determine if a student's mental model about a certain concept is aligned with the emergent process category or the sequential process category.

The KDD process was applied to an expert annotated dataset gathered from a midwestern U.S. public institution. This dataset consists of student's labeled textual descriptions of answers to multiple-choice science questions about dynamics and heat

transfer. The data was collected as part of a research project where the predicate test is manually performed by educational engineering experts. Among the results of that research is an expert annotated dataset where the verbs or phrases used by students in their explanations are identified as ontological attributes. Furthermore those ontological attributes were categorized as belonging to the *emergent* process category, *direct sequential* process category, and a mixed category that represents a student categorization of a concept into both the *emergent* and the *sequential* categories simultaneously (a data point having multiple labels). [10]

The raw dataset consists of 680 textual descriptions of student answers to multiple-choice questions about dynamics and heat transfer. Not all of the documents found in this raw data were labeled by experts. This dataset was manually preprocessed by constructing a file in Attribute-Relation File Format (ARFF), which is the document format used by WEKA [17]. The result was a dataset consists of 41 instances of the *emergent* class, 99 instances of the *sequential* class, and 50 instances of the *mixed* class.

Textual explanations of a concept and their corresponding expert labels were the only features extracted from the raw dataset. We chose to use the *sequential* and *emergent* process categories as our classification labels because, as previously stated, students also tend to confuse the inherent nature of *sequential* and *emergent* processes [12].The feature space was created using feature extraction and selection algorithms provided by WEKA [17].

We used this dataset to perform a multi-class, single label classification with Support Vector Machine (SVM) as our classification algorithm. Four classification models were built based on varying feature selection techniques and evaluated them using 10-fold cross-validation. The selection of our best performing classification model was based on the resulting measures of precision, recall, accuracy, f-measure, and kappa coefficient.

With the implementation of the PTAP we have shown that the predicate test proposed by Chi can be automated using KDD theory.

1.1 Motivation

Chi's predicate test is an effective yet time consuming task, which consists of expert categorization of student textual descriptions of concepts, is a time consuming, expert dependent process that is not yet an acceptable real-time learner assessment solution needed in modern classrooms as described by [9].

According to Chi misconception detection via predicate tests and their interpretation are key players in learner assessment [12]. As previously mentioned a state-of-the-art technique performed by educational engineering experts is known as the predicate test is an affective yet time consuming task that requires a trained expert. In order to visualize mental models that help pinpoint student misconceptions experts have to manually inspect and annotate student predication.

The predicate test automation process can provide precise assessment and significantly reduce the time it takes to deliver individualized assessment feedback to students within a learning environment, wether it be in an educational research setting or an actual classroom. The presence of real time assessment within a learning environment for each individual student can be essential in the detection of the knowledge gaps present in his/her understanding of concepts [9]. Furthermore modern educational engineering theory states that real-time assessment performed on each individual student can greatly improve transfer of knowledge from teacher to student [9].

In short the PTAP solves the issue of waiting for experts to manually inspect and annotate student predication in order to obtain real-time assessment, which helps detect knowledge gaps in learner mental models and positively impacts the transfer of knowledge from teacher to students.

1.2 Problem Statement

Predicate tests require the participation of educational engineering experts to isolate the student predication relevant to understanding a learner's mental model of a certain science concept under study. The main goal of this research was reduce the time it takes for teachers to deliver real-time assessment through predicate tests by building a classification model that could successfully learn to categorize student textual descriptions of concepts into the *emergent* and *sequential* process ontological categories. Developing this process, named the PTAP, was considered as a task of KDD that makes use of text mining and text classification. The PTAP can serve as an essential building block of a complex e-learning system with the capability of providing individual, real-time assessment to students based on their correct or incorrect mental categorization of concepts.

1.3 Scope

The research presented in this document describes the use of multi-class, single label classification using text categorization of student predicate use. Due to the nature of conceptual change assessment via predicate tests our text classification task could have been aligned to multi-class, multi-label classification where student predication can contain components of more than one class [16]. The scope of this work is limited to multi-class classification or hard classification.

In addition this research proposed and developed a methodology for automatizing predicate tests assuming there is an existing pre-labeled dataset that has already been gathered. In other words this research is not involved with the gathering of student data and assumes the data is already available for preprocessing, transformation, classifier training and evaluation.

1.4 Contributions

The main contribution brought forth by this research is a novel approach that combines Chi's predicate tests with KDD theory to reduce the time it takes to

deliver predicate test results while helping teachers deliver real-time, individualized misconception assessment to students studying science concepts. We have shown that it is possible to successfully classify learner's textual descriptions of a science concepts as belonging to the emergent and *sequential* concept categories described by Chi. This in turn reveals information about each student's mental categorization of such concepts, thus revealing knowledge gaps in their understanding of that concept.

The approach documented in this thesis for predicate test automation can help to ease the difficulties associated with learning new topics due to accumulated misconceptions within actual learning environments. This is why our research directly impacts the efficiency of transfer of knowledge from teachers to students, which affects academic achievement in all educational systems [9][11].

Although the PTAP can quickly become useful for speeding up misconception research it can further be used as an essential component a system that automatically collects and classifies student data to provide for a fully automated (including data gathering) predicate tests.

1.5 Outline

The following chapter details Chi's misconception theory and the reasons as to why Chi's misconception theory was chosen for this research. It also details the definition of text classification and modern tools used for text mining. The next chapter describes our proposed solution for performing text classification on students answers. Our research objectives and methodology follow. The chapter that follows presents our experimental results and results discussion. The last chapter of this document states our conclusions and future work.

CHAPTER 2
LITERATURE REVIEW

This chapter focuses on documenting our literature review on which this research was built on. Section 2.1 discusses conceptual change research. Sections 2.2 and 2.3 detail Knowledge Discovery in Databases theory and Educational Data Mining theory respectively. The following section, Section 2.4 discusses text classification. Section 2.5 mentions related work in the field of student modeling for learner assessment.

2.1 Conceptual Change in Science Education

Science education research focuses on how learners acquire knowledge about science and how that knowledge is applied. Within the domain of conceptual change research two broad perspectives have emerged to describe the nature of knowledge structure coherence, misconceptions, and conceptual change. These are known as the *knowledge-as-theory* perspectives and the *knowledge-as-elements* perspectives [18]. These perspectives state that science domain knowledge acquired by learners can be broadly described, respectively, as unified frameworks with coherent theoretical structure or as independent collections of elements [18].

This research is aligned with the *knowledge-as-theory* perspectives, specifically the perspective documented by Chi et. al. in [1][6][14][15][12][16]. Chi et. al.'s knowledge structure theory describes how science concepts can be ontologically categorized by learners and how incorrect categorization of concepts gives way to misconceptions and the need for conceptual change.

This remainder of this section describes the role of ontologies in misconceptions and provides a formal definition of conceptual change theory. It also includes a discussion about the difficulties involved in conceptual change and its assessment. The last section discusses predicate tests and their use in conceptual change assessment.

2.1.1 Ontologies and Conceptual Change in Science Education

Conceptual change in the past two decades has become aligned with the notion that novice and expert understanding of concepts is based on ontological categories. When a learner incorrectly categorises a concept the he or she is said to have a misconception of that concept. This theory has been prominently documented by Chi, Slotta, et. al. in [1][6][14][15][12][16].

The definition of conceptual change established by Chi, Slotta, et.al. is based on a combination of accepted positions within conceptual change literature and is an attempt to systematically diagnose and assess conceptual change. Three types of conceptual change have been defined by Chi: belief revision, mental model transformation and categorical shift [12]. These three types of conceptual change are all based on the assumptions that entities in the world can be ontologically categorized, that the nature of physics science concepts dictates their categorization into *constraint-based interaction* concepts, and that students hold naïve preconceptions aligned with *substance-based* descriptions of concepts. [10]

A contradicting theory documeted by Gupta et. al. in recent years [19][20][21][22] has surfaced which proposes that associating physics science concepts to a single ontological category can be detrimental in the development of expertise. Gupta et. al. claim that novices do cross ontological boundaries when describing physics science concepts. Also they claim that resources which have been associated with a specific ontological category can be utilized to help teach concepts from different ontological categories as well [21]. Gupta et. al. argue that students do not have

rigid ontological commitments, rather they can switch from matter based understanding to process based understanding of concepts, although most novice categorization of concepts tends to be substance based [21]. Gupta et. al.'s evidence is based on the inspection of expert literature and novice description of concepts. Although their findings are theoretically coherent, their research does not explain how it is acceptable for experts to use *substance-based* descriptions for *constraint-based interaction* concepts. In other words, the fact that expert literature contains *substance-based* conceptions or examples of *constraint-based interaction* concepts proves that metaphors can be used to describe *constraint-based interaction* concepts with *substance-based* descriptions. In addition, this does not prove that concepts which belong to the *constraint-based interaction* category can also belong to the *substance* category [6]. For these reasons we subscribe to Chi's definition of conceptual change.

The following section of this document is dedicated describing the role of ontologies in Chi's conceptual change theory. Following that section it is noted why some types of conceptual change are apparently difficult. The next section details how the predicate test can be used to determine if and when conceptual change must take place.

2.1.1.1 Three Suppositions of Conceptual Change

As stated before, this work focuses on describing Chi's theory of conceptual change, which is based on three assumptions. The first is an epistemological assumption about the natural categorization of entities in the world. The second is a metaphysical assumption that describes how most science concepts belong to the process or constraint-based interactions category. The third is a psychological assumption related student's naïve preconceptions [6]. This section details those three assumptions.

2.1.1.1.1 Epistemological Proposition of Conceptual Change. The epistemological proposition of Chi's theory of conceptual change states that entities in the world naturally belong to certain major ontological categories, these being matter, processes, and mental states [6] (refer to Figure 2–1 for an example of this categorization of entities). The theory is loosely aligned with the exact names of those three major categories and states that more than three can exist, but focuses on the matter and processes categories or their equivalent descriptions [12].

Ontological attributes are those properties that an ontological category can possess due to being associated to that category. The following example aims to explain this notion: considering a shoe as an artifact from the matter category it can be said that the shoe must have defining attributes like a sole, most frequently has characteristic attributes like laces, and can potentially be worn, which is an ontological attribute [6].

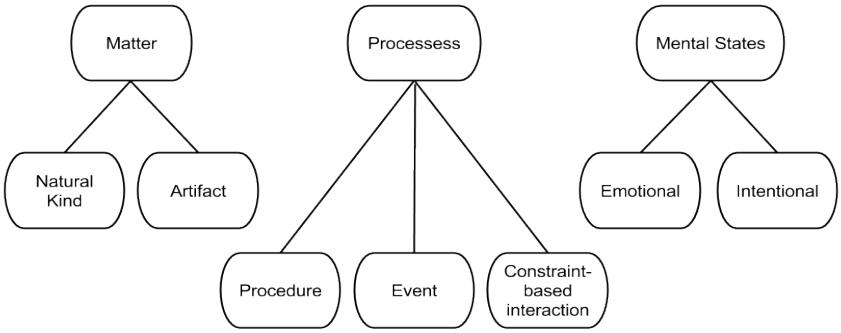


Figure 2–1: Three major ontological trees from Chi et. al.'s epistemological proposition of conceptual change theory. Adapted from [6].

Ontological trees are considered distinct if they possess mutually exclusive ontological attributes [6][12]. In other words, ontological attributes from a given ontological tree cannot be applied to categories that belong to a different tree. For example, entities categorized as matter have ontological attributes such as "storable" and "having color, volume, or mass", which are attributes that cannot be assigned to a process category. In a similar way, ontological attributes such as "resulting in" or "occurring over time" can only be assigned to entities that form part of the

processes tree. Categories within the same ontological tree can also be ontologically distinct when their respective categories cannot be shared between them [6][12].

Ontological categorization is the basis of Chi's theory of conceptual change since conceptual change happens when students shift their categorization of concepts from one distinct ontological category to another [6]. This category shift by students can occur across major ontological trees or within them [6][12].

2.1.1.1.2 Metaphysical Proposition of Conceptual Change. The second proposition of Chi's conceptual change theory is about the nature of science concepts. It exposes how the constraint-based interaction category, a subcategory of the processes ontological tree, can be used to describe many science concepts [6][12]. Constraint based interactions are unpredictable and emergent with no definite beginning or end. This category encompasses concepts such as heat, electric current, and light, which are processes that cannot be assigned to any subcategories of the matter ontological tree [6][12]. For example, electric current exists when a charged particle moves through an electric field. Thus the electric current process is considered a constraint-based interaction which emerges from the interaction of various components that belong to the matter ontological tree such as particles, wires, and batteries [6]. Since constraint-based interactions involve components from the matter category this can be confusing for students [6]. This assumption can apply to concepts outside of the physical sciences [6].

2.1.1.1.3 Psychological Proposition of Conceptual Change. The third proposition of Chi's conceptual change theory states that students hold naïve knowledge or preconceptions about science concepts [6][12]. It explains the nature of some science misconceptions through a psychological point of view. In short, students' naïve conceptions of physical science concepts tend to consist of assigning process concepts to the matter category.

These preconceptions can exist at the proposition and mental model levels. Naïve knowledge at the proposition level is simple to remove, is referred to as non-robust misconceptions, and is corrected with conceptual reorganization. In other words, those preconceptions can be easier to correct or remove [6][12]. Some naïve knowledge can be resistant to creative types of pedagogy approaches and is referred to as robust misconceptions [6][12]. Diagnosing the presence or absence of preconceptions in individual students reveals information about the incorrect category to which those preconceptions have been assigned by students.

2.1.2 Conceptual Change Difficulty

In their research Chi et. al. [6] documented how science concepts can be represented with ontology trees. This theory is based on the assumption that entities in general are part of ontological categories like matter and processes. According to those authors [6], science concepts can be categorized by students to create individual mental models of representation and the incorrect categorization of concepts leads to certain types of misconceptions. Identifying this mis-categorization reveals how conceptual change can be assessed by comparing the ontological categorization of concepts by students with domain expert categorization of concepts. The difficulty lies within student's own lack of awareness of misconceptions, student's naive misconceptions, and the incompatibility between student and expert categorizations [6][12].

When students learn new science concepts they automatically try to assign those new concepts to an ontological category within their understanding. This becomes a burden when the category for new concepts does not exist within the student's mental model [6][12].

It has been stated by Chi et. al. in [14] that when students are learning science they have two problems: learning many things at once which are missing from their current understanding and holding naïve preconceptions. Those authors also

established that naïve preconceptions have two properties: they are often incorrect and they often impede acquiring deep knowledge of concepts.

Some mis-categorizations of concepts can seem to be easier to repair than others depending on each student's prior conceptions. Furthermore it has been documented by Chi et. al. in [6][14] that misclassification of a concept into a hierarchically related category is not considered a robust misconception and its repair is not considered conceptual change rather conceptual reorganization.

Chi and Roscoe have stated in [14] that conceptual shift is not difficult. It becomes difficult when students lack awareness for its need and they have limited knowledge of the categories to which their ontologically incorrect concept should be correctly assigned.

2.1.3 Predicate Tests and Conceptual Change Assessment

Contributions by Chi and her colleagues Slotta, deLeuw, Santiago, et.al. have documented the use of the predicate test as the means to assess conceptual change, i.e. concept re-categorization at the ontological tree level [1][15][16]. The predicate test theory states that verbs used by students and experts to describe concepts correspond to ontological attributes of those concepts [10]. Decades of research document how novices use *matter-based* predicates to describe *substance-based* concepts and *constraint-based interaction* concepts with the same frequency [1][6][14][15][12][16]. Since it is usual for novice predication to contain ontological attributes of the wrong category predicate tests consist of analyzing verbal predicates used by students and contrasting them to the predicate use of experts.

It is also noted that expert predicate use contains a high frequency of *substance-based* predicates when describing *substance-based* concepts. In addition, the *predicate use profiles* reveal high frequency of *constraint-based interaction* predicate use for *constraint-based interaction* concept descriptions [1][15][16]. This is aligned with psychological proposition of conceptual change, which considers that novices hold

naïve preconceptions causing the alignment of their mental models with *substance-based* predication for *constraint-based interaction* physics concepts [6]. In contrast, the proposition also states that experts consistently use *constraint-based interaction* predicates for *constraint-based interaction* descriptions of physics concepts.

Slotta et. al. developed expert taxonomies based on expert explanations of *substances* and *processes*. The taxonomies describe predicates in the form of single or multiple word phrases or ideas that are explicitly associated to certain ontological attributes and the ontological categories to which the associated ontological attributes belong to [1]. These taxonomies are used as the main criteria for predicate tests. Refer to Table 2.1.3 for a snippet of the substance and process predicate taxonomies.

Table 2–1: Snippet of substance and constraint-based interaction predicate taxonomies proposed by Slotta et. al. in [1].

Substance Predicates	Process Predicates
block	movement process
move	excitation
consume	equilibrium seeking
quantify	systemwide
accumulate	simultaneous
equivalent amounts	transfer

The following describes the events that lead up to and occur after predicate tests take place. Given a carefully crafted multiple-choice question about a certain science topic (e.g., dynamics, heat transfer) the students are required to explain their selected answer. The textual description of their answer is then examined to isolate predicates and the ontological attributes they contain. Then the isolated words, phrases, and sentences are classified as belonging to a certain ontological category that may or may not be the category to which the concept being described belongs

to. This last step is considered to be the actual predicate test. In other words the predicate test does not include crafting multiple-choice questions whose answers can best reveal mental models of whoever answers them, administering the questions, or isolating predicates from textual descriptions the selected multiple-choice question answer. What it does include is the analysis of determining which of Chi's ontological categories for science concepts the student's predication corresponds to.

The resulting predicate test analysis is used to demonstrate the robustness of student's incorrect ontological categorizations by determining whether conceptual change should take place at the ontological tree level and to what degree [1][15][16].

2.1.3.1 Predicate Test and Concept Incommensurability

The predicate test is a comparison of student and expert predication as a protocol analysis, which helps to draw inferences about the difference in ontological commitments between expert and novice language of the science domain [1][15][16]. This is used to estimate the need for conceptual change in a student as well as the degree of incommensurability between novice and expert predication. The incommensurability between concepts or ontological categories is defined by Chi and Roscoe in [14] as "irresolvable differences in concepts, propositions, and explanations of theories". They further explain that concepts can be considered incommensurate if the one can replace the other, be differentiated from each other, or coalesced to further understand a single concept [14]. This theory of incommensurability helps detect if robust conceptual change is needed or is taking place by using the results of a predicate test to estimate incommensurability [14].

The predicate test is performed by analyzing student predication of a certain concept. Experts determine which ontological categories are present in a given student predication by analyzing if it contains attributes from the *substance* category, the *constraint-based instructions* category, or any of their subcategories. The student's conception of that given concept can then be considered commensurate or

incommensurate with respect to the correct predication established by experts. In early discussions of this literature it is assumed that ontological categories which facilitate science thinking are reliably incommensurate [14].

As stated in the previous section, conceptual reorganization differs from misclassification of concepts into lateral or hierarchically distinct categories, which is considered a robust misconception and its repair is considered robust conceptual change [6][14]. Estimating the incommensurability between expert and novice predication of physics science concepts distinguishes if conceptual reorganization or robust conceptual change must take place or is already happening [1][14].

Chi and her collegues were able to relate misconceptions to the identification of ontological boundaries between two explanations [1][6][10][14][15][16]. This does not explain the actual change of conception in student mental models but allows for a systematic approach for the assessment of conceptual change and pinpoints when it must take place [1][6][10][14][15][16]. The predicate test stands as proof that naïve conceptions of the physics sciences are based on the *matter* category, experts predicate use is consistent with ontological attributes from a category that is equivalent to the *constraint-based interactions* category, that novice predicate use is not [1][6][14][15][16], and the degree to which conceptual change must take place is directly related the concept incommensurability between student and expert perceptions of a science topic under study.

2.2 Knowledge Discovery in Databases

In essence KDD refers to the process of obtaining useful knowledge from large datasets. The KDD process consists of the sequential iteration of selecting data of interest from a large dataset, preprocessing and transforming the data into a format that is appropriate for mining, and extracting models or patterns from the transformed data. Figure 2–2 illustrates the major steps involved in the KDD process.

The following bullet list further describes each major step in the KDD process:

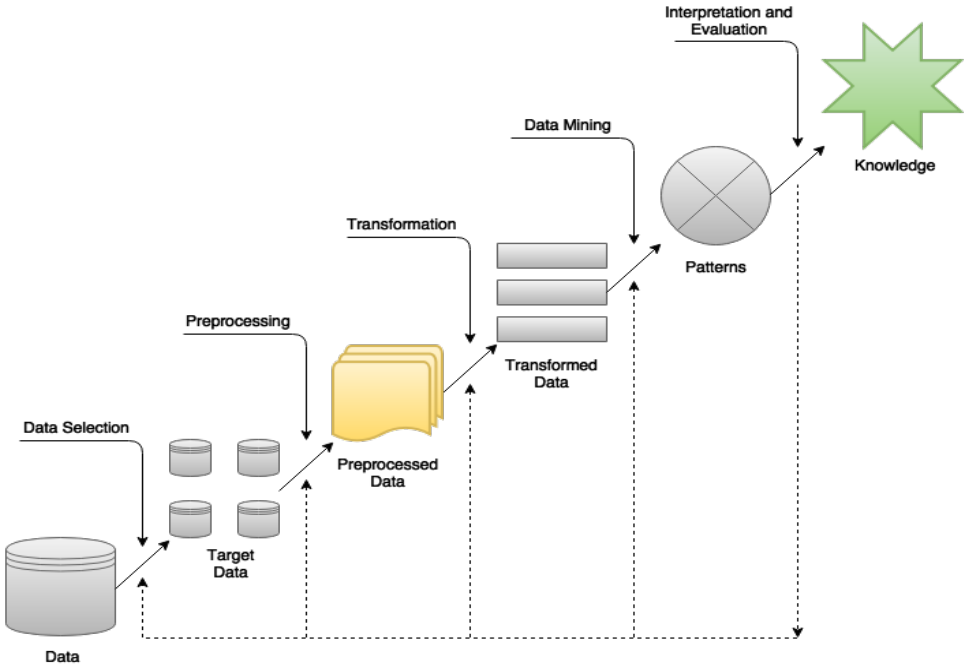


Figure 2–2: Steps involved in the KDD process.

1. **Data Selection**: a relevant dataset that could potentially be useful for knowledge acquisition is extracted from an existing data warehouse.
2. **Data Preprocessing**: selected data is cleaned to minimize noise introduced by missing, irrelevant, or erroneous data points. This step enhances data reliability.
3. **Data Transformation**: this step aims to represent the preprocessed data in a format that is appropriate for mining. Transformation can include feature selection, feature extraction, feature weighting, dimensionality reduction, etc.
4. **Data Mining**: this step is where knowledge is discovered in the form of associations, patterns, anomalies, and other significant data structures using machine learning algorithms.
5. **Evaluation**: discovered knowledge is interpreted and validated using charts, graphs, and other data illustration tools.

The steps involved in the KDD process can iteratively repeated when the selected preprocessing and transformation techniques do not yield satisfactory results.

Although steps 3, 4, and 5 of the KDD process are sometimes collectively referred to as *data mining* our research is aligned with the notion that *data mining* refers to the specific task of extracting patterns from data and KDD refers to the overall process from data selection to results evaluation.

2.2.1 Data Mining

According to knowledge discovery in databases (KDD) theory *data mining* is an essential step of the KDD pipeline. Data mining is usually accomplished using machine learning techniques, which include and are not limited to:

1. **Classification**. This mining technique is based on the use of previously labeled data, known as a training set, to inductively construct a model for each known label based on the features of each class. The process of building a model for each class results in a set of classification rules, which can be applied to classify future data or to gain insight into previously existing data.
2. **Prediction**. As its name states this mining technique is used to predict value distributions of attributes or to predict missing values of attributes belonging to a set of objects. It is based on the determining a set of attributes that represent *independent variables* to be used for predicting a relevant attributes of interest considered as *dependent variables*. For example, a theatre's attendance for an upcoming day of the week can be predicted by analyzing the theater's previous attendance distribution of that same week day.
3. **Clustering**. Clustering is used to identify collections or clusters of data objects considered similar with respect to one another. The term similar is relative and can be defined with distance measurements or any type of function that describes the difference between two data points. High quality clustering results are considered as such when *intra-cluster similarity* is high and *inter-cluster similarity* is low. In other words objects belonging to the

same cluster are considered to be highly similar and objects belonging to different clusters should be easily differentiable. For example, employees from a certain company can be clustered according to their area of expertise or college degree major.

4. **Association Rule Discovery**. The aim of this technique is to determine which patterns, associations, or correlations occur with most frequency within a given data repository. The representation of discovered associations is denoted as $X \rightarrow Y$ where X and Y are collections of one or more items known as item sets, and the presence of an item set X in a database implies the presence of Y. For example, an association rule can describe how a department store customer that buys a movie is likely to buy popcorn as well in the same transaction.

   Discovered associations are evaluated with measures of support and confidence. Support refers to the fraction of data that contain both item sets X and Y. Confidence measures how often item set Y is implied by item set X. The process of discovering association rules consists of two major steps known as frequent item set generation and association rule generation. Frequent item set generation determines which item sets have support count above a minimum threshold and association rule generation determines which frequent item sets have the highest confidence score.

2.3 Educational Data Mining

Educational data mining (EDM) focuses on applying knowledge discovery techniques to large datasets generated in educational contexts [2][23]. It is an emerging field of research that combines three major domains: education, computer science, and statistics [2]. The Venn diagram found in Figure 2–3 illustrates the interdisciplinary nature of EDM, where the intersection of the main domains comprise the subdomains that are closely related to EDM.

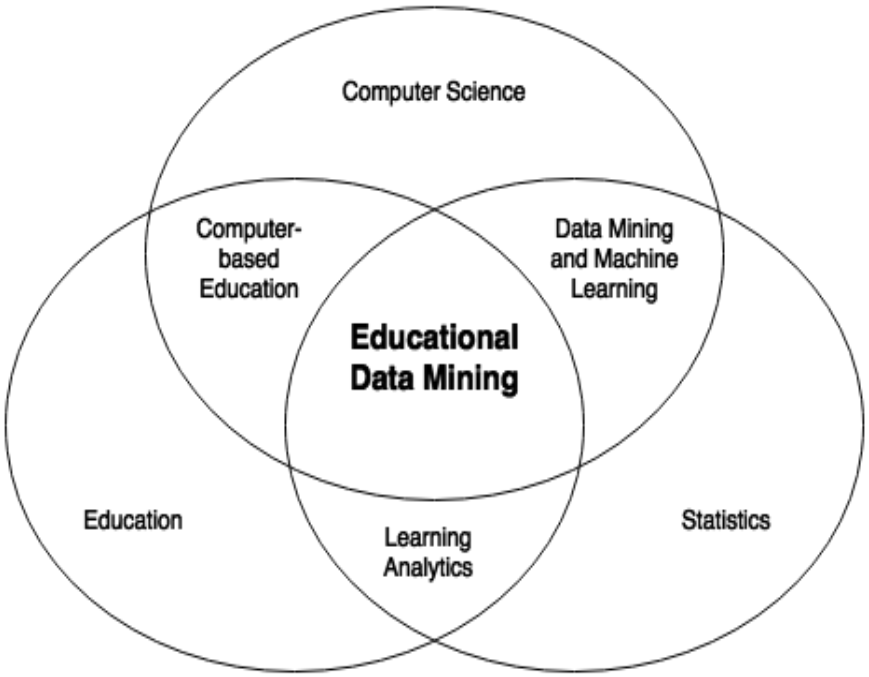


Figure 2–3: Venn diagram illustrating the multiple disciplines associated with educational data mining. Adapted from [2].

The core purpose of EDM is to facilitate the understanding of how students learn, further enhance educational managerial issues, and aid in resolving educational research issues [2][23]. Further descriptions of EDM objectives can be specified in terms of its users or stakeholders. Refer to Table 2–2 for the aforementioned descriptions.

Table 2–2: EDM stakeholders and their respective objectives [2].

Stakeholders/Users	Objectives
Teachers	Gain insight on how students learn in order to improve the pedagogical approaches, teaching performance, etc.
Students	Improve learning performance, provide assessment feedback, gain insight on individual student situations, etc.
Administrators	Find optimal ways to organize institutional resources, etc.
Researchers	Determine which data mining approaches are best suited for specific educational tasks, and measure the learning effectiveness of educational tasks when employing different data mining methods.

The process of knowledge discovery within educational data is same as the traditional KDD process with the single constraint, which dictates that data generated in educational contexts must be used as input to the first step (data warehousing) of the KDD process [2].

The following are KDD tasks that have been proven to be relevant to EDM [2]:

- Prediction
- Clustering
- Outlier Detection
- Relationship Mining
- Social Network Analysis
- Process Mining
- Text Mining
- Distillation of Data for Human Judgement
- Discovery with Models
- Knowledge Tracing
- Nonnegative Matrix Factorization

During this research we utilized *text mining* or *text data mining* techniques to achieve our goal of predicate test automation. Furthermore, our text mining tasks were carried out using text classification.

2.4 Text Classification

Text classification is a modern topic of information processing that finds uses in spam filtering, knowledge-base creation, information retrieval, etc. [10][24][17][25][26]. There are two types of classification known as supervised and unsupervised classification.

According to classification theory a set of document-category tuples (d_j, c_i) can result from the cross product $D \times C$ of all available documents D to be classified into known or unknown categories C [10][24][17][26][27].

The process of supervised text classification involves three major steps for determining a category for any given document or vice versa. These steps are known as preprocessing, model creation or classifier training, and classification into previously defined categories [10][24][17][25][26].

In supervised text classification for each (d_j, c_i) tuple the semantic similarity or distance between d_j and c_i is estimated and thresholded to decide which known category c_i each document d_j belongs to [24][17][26][27]. Boolean and weighted values can be assigned to each $(d_j, c_i) \in D \times C$ as well [24][17][26][27].

Conversely unsupervised text classification involved preprocessing of data and assigns unknown categories to documents without undergoing classifier training. The aim is to cluster related documents based on their content and assign a category to each document cluster.

During this research we focused solely on supervised learning to achieve our classification tasks. For this reason further discussion of text classification is assumed to describe supervised text classification.

Types of classification tasks include [24][17][26]:

- **Binary** - determine if a new document does or does not belong to a certain category. Is this an article about mathematics?
- **Multi-class** - assign a label to a document from a previously defined set of labels. Which sport if this article about?
- **Multi-label** - assign zero or more categories to a new document. Which are the best career choices for this high school graduate?
- **Ranking** - assign categories with ranks to a new document. Rank the best career choices for this high school graduate according to the student's probability of success.

Preprocessing is the first step of text classification, which includes everything from parsing to feature extraction, weighting, and extraction. Once a dataset has

been transformed classification algorithms such as Support Vector Machines (SVM) and Naïve Bayes (NB) inductively construct a model using a subset of the preprocessed dataset known as the training set [24][17][26][27]. The model is then used to perform classification on unknown data points.

Aside from using a larger training set compared to the test set no definitive ratio of training data to testing data that is considered a standard [17].

2.4.1 Classifier Evaluation

Measures used to evaluate classifiers include precision, accuracy, recall, and f-measure, and kappa coefficient (κ). These measures vary with different combinations of data set size, training-test set ratios, preprocessing, feature selection, and classification algorithms [24][17][26][27]. The following lists the equations for each of the aforementioned measures:

$$Accuracy = \frac{TP + TN}{N} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F - measure = 2\frac{Recall * Precision}{Recall + Precision} \tag{2.4}$$

$$\kappa = \frac{ObservedAccuracy - ExpectedAccuracy}{1 - ExpectedAccuracy} \tag{2.5}$$

$$ErrorRate = \frac{FP + FN}{N} \tag{2.6}$$

Accuracy refers to the ratio of instances that are correctly categorized. Precision, also known as positive predictive value, describes how well a positive result actually predicts the presence of the positive category. Recall, also known as sensitivity or true positive rate is the probability that a datapoint is classified as positive given that it actually is positive. F-measure is a ratio that combines measures of precision and recall into a single measure.

Kappa or inter-rater agreement is a metric used to compare a classifier's accuracy with random chance and measures the agreement between the classifications and the true classes. The kappa statistic reveals information about a classifier's performance and is used to compare the performance between two or more classifiers. [3][4][5]

Tables 2–3, 2–4, and 2–5 show guidelines for interpreting kappa measures proposed by Cicchetti et. al. [3] Landis et. al. [4], and Fleiss et. al [5] respectively.

Table 2–3: Guidelines by Chicchetti et. al. for interpreting kappa measures [3]

Rating	κ
Poor	$\kappa < 0.40$
Fair	$0.40 < \kappa < 0.59$
Good	$0.60 < \kappa < 0.74$
Excellent	$0.75 < \kappa < 1.0$

Table 2–4: Guidelines by Landis et. al. for interpreting kappa measures [4]

Rating	κ
Poor	$\kappa < 0.20$
Fair	$0.21 < \kappa < 0.40$
Good	$0.41 < \kappa < 0.60$
Very Good	$0.61 < \kappa < 0.80$
Excellent	$0.81 < \kappa < 1.0$

Classification results are presented in the form of a confusion matrices where a two-by-two contingency table is constructed for each binary classification problem of N documents. The cells in this table contain the following information:

Table 2–5: Guidelines by Fleiss et. al. for interpreting kappa measures [5]

Rating	κ
Poor	$\kappa < 0.39$
Good	$0.40 < \kappa < 0.74$
Excellent	$0.75 < \kappa < 1.0$

- True Positives (TP) - the amount of positive instances which have been correctly classified as such. For example, healthy people determined to be healthy by a classifier.
- False Positives (FP) - the amount of negative instances that have been incorrectly classified as positive. For example, when a classifier incorrectly determines that sick people are healthy.
- True Negatives (TN) - the amount of negative instances that have been correctly classified as such. For example, sick people that have been determined to be sick by a classifier.
- False Negatives (FN) - the amount of positive instances that have been incorrectly classified as negative. For example, when a classifier incorrectly determines that healthy people are sick.

where

$$N = TP + FP + TN + FN \tag{2.7}$$

2.5 Related Work

Misconception assessment using Machine Learning techniques has found its place in the construction of student models within Intelligent Tutoring Systems (ITS). In addition student models play a major role in ITS' personalization strategies. This is due to the fact that ITS incorporate adaptive learning techniques based on each individual student's knowledge. [28]

This work focuses on student misconception detection at the ontological level to develop student models that describe current ontological categorization of concepts. In other words we have focused on categorizing misconceptions. We are not aware of any work that uses our approach to model erroneous conceptions in students.

The the rest of this section mentions research work that has a close resemblance to our work, although they do not address the problem of categorizing misconceptions. Instead they mainly focus on the presence of misconceptions in students or describing the general knowledge possessed by students. In addition, they focus on determining which learning materials are needed to correct student misconceptions or to advance to subsequent lessons respectively.

In the research documented by Liu in [29] the proposed system is already aware of the possible misconceptions that can arise while learning statistics topics. The system focuses on evaluating students to determine if the possess any of these previously determined misconceptions and providing feedback to make them aware of their existence.

In order to advance to subsequent lessons Wang used a pre-test and a two-tier tests to identify knowledge gaps and dynamically select additional learning materials a student may need [30].

Ehimwenma et. al. in [31] propose the use of a multi-agent system to determine which concepts have not been learned by students. Pre-assessment strategies are employed to model current knowledge in students, then knowledge gains and gaps are identified to determine which learning materials are recommended for learners to able to advance to further lessons.

Bayesian Student Models (BSMs) have been used by Millán et. al. in [32] to develop student models. They are based on using knowledge and evidence variables, as well as their correlation, to construct Bayesian Networks. In other words the network structure or nodes are elements that represent whether or not students

have knowledge about a specific domain, and the answer to questions about that domain. In addition the edges between these two types of nodes is used to describe their correlation for a particular student.

We have observed that the previously mentioned research can model student knowledge and identify knowledge gaps, they are not focused on misconception categorization.

CHAPTER 3
CONCEPTUAL FRAMEWORK

It has been clearly stated by Chi's that misconception assessment is possible through the use of predicate tests performed on learner data [1][6][14][15][12][16]. The data for predicate tests usually consists of student answers to multiple choice selections. Students are also asked to describe their answers and explain why they chose the answers wherever possible. Since specific problems and questions are chosen by domain experts each item reveals details about individual student mental models.

The predicate test is a process where each student response is analyzed at the individual sentence or proposition level to determine the predicates used and how they were used to describe their answer [1][16]. These authors have documented how student predicate use is compared to expert a priori categorizations of *substance-based* or *constraint-based interactions* verbal predicate use. This in turn reveals whether a student used any combination of *substance-based* or *constraint-based interaction* predication in their answer's description.

Details of how conceptual change assessment can be interpreted as a KDD task using text classification are discussed in Section 3.1. Section 3.2 discusses how text is represented as word vectors in order to be used for text classification. The following section, Section 3.3 details Support Vector Machines theory.

3.1 Text Classification of Student Predicate Use

Consider a student's textual explanations of multiple choice science questions as the set of documents D to be categorized into the set of ontological categories C

used by Chi. According to text classification theory [24][17][26][27] the mentioned documents could be classified into each category using text classification.

Considering the KDD process, which includes a classification step, the complete automation of predicate tests has to account for the following:

- Gathering of textual descriptions from students
- Preprocessing and transforming the gathered data
- Performing feature selection, extraction, and weighting techniques to the transformed data
- Building a prediction model using the resulting feature set
- Classify unknown data points into the *substance* category, the *constraint-based interaction* category, or any combination of those categories, including their respective sub sets.
- Results evaluation

What is needed to implement the proposed automated solution for conceptual change assessment is an available dataset that has been manually annotated to partition into training and testing datasets. If this were not the case and supervised learning was to be employed then clustering would have to be performed in order to determine relevant labels and features for use during classifier training [33][34].

The classification to be performed consists of assigning student explanations about science topics to one or a combination of the major categories of concepts involved in learning science. This is the approach taken by Chi in [1][6][14][15][12][16]. Their research has identified the three main categories used in science concept discussions as [12] substances, sequential processes, and emergent processes. These categories can be used as our text classification labels.

3.2 Representing Text as Word Vectors

Classifier accuracy is directly related to the way data is represented and fed to learning algorithms [7][8][35][36]. Before *text classification* can be applied to text

documents they have to be transformed into word vectors [37]. After transforming the documents the features are extraction and selection are performed using the resulting word vectors. This remaining data is used to train learning algorithms that will eventually perform classification tasks.

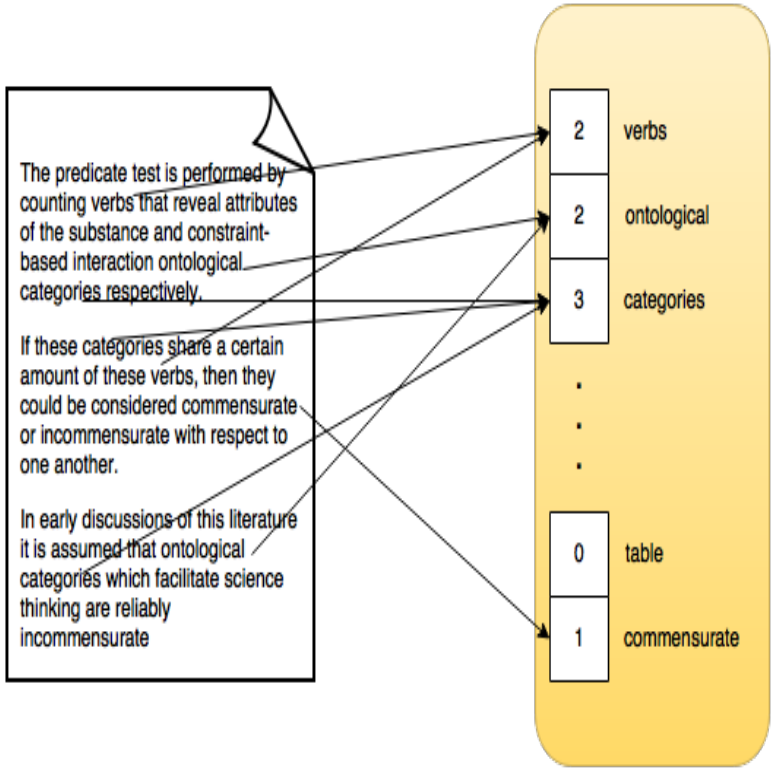


Figure 3–1: *Text as Feature Vector* [7].

To transform a document to vector representation each word in a document is tokenized, punctuation is removed, case is ignored, stop words are removed, etc., and a feature vector for each document is constructed. The constructed feature vector's components contain information about word frequency. In formal terms each word (w_i) in a text document is considered a feature or vector component with value represented by $TF(w_i, d_i)$, where $TF(w_i, d_i)$ denotes the number of times the word w_i appears in document d_i [7][8][35][36].

This is known as basic attribute-value representation of text, which results in a feature vector whose length is the size of the vocabulary being processed. Common vocabulary size is 10,000-100,000. In other words the task of text classification is

usually very high dimensional with near independence of features. Figure 3–1 depicts an example document and its transformation to feature vector form.

The resulting vector representation is likely to be high dimensional in nature, which can potentially enhance the complexity of the problem to be solved, cause classifier overfitting, etc. That is the main reason feature reduction techniques are called for [7][8][35][36].

3.3 Support Vector Machines

Support vector machine (SVM) algorithms are based on the structural risk minimization principle [7]. They have been proven to achieve good results when classifying high dimensional data sets with many relevant features [35].

SVMs have been widely applied to machine learning to search for maximal-margin hyperplanes that separate positive and negative data points form each other with the least true error [8]. This type of classifiers are trained with a pre labeled dataset [7][8][35][36].

The use of SVM has been determined to be very effective in text classification [7][8]. Joachims in [7] has stated this is due to:

- High-dimensional input space of textual data
- Textual data usually contains very few irrelevant features
- Representation of text data results in sparse vectors
- The majority of text classification problems are linearly separable

The following formally details SVM theory for machine learning:

Let T be a set of labeled training samples defined by:

$$T = (x_1, y_1), ..., (x_m, y_m); x \in R^n, y_i \in \{-1, 1\}, i = 1, ..., m$$

T is said to be linearly separable if there exists a vector $x \in R^n$ and a scalar b that can be used to construct hyperplanes, which conform to:

$$\begin{aligned} wx_i - b \geq 1 \quad & \text{if } y_i = 1, \\ wx_i - b \leq -1 \quad & \text{if } y_i = -1, \\ & i = 1, ..., m \end{aligned} \tag{3.1}$$

Equation 3.1 can be summarized as:

$$y_i(wx_i - b) \geq 1 \quad y_i \in \{-1, 1\}, i = 1, ..., m \tag{3.2}$$

The distance between hyperplanes is known as margin. The optimal separating hyperplane is expressed as:

$$w^* x_i - b^* = 0 \tag{3.3}$$

The aim is to find the maximum-margin hyperplanes that divide data points x_i when $y_i = 1$ from data points x_i when $y_i = -1$. When T is considered to be linearly separable we can select parallel hyperplanes with the largest possible distance between them. Figure 3–2 shows a visual representation of hyperplanes separating two distinct classes. The classifier on the left would yield better results since there is a more distinct gap between classes.

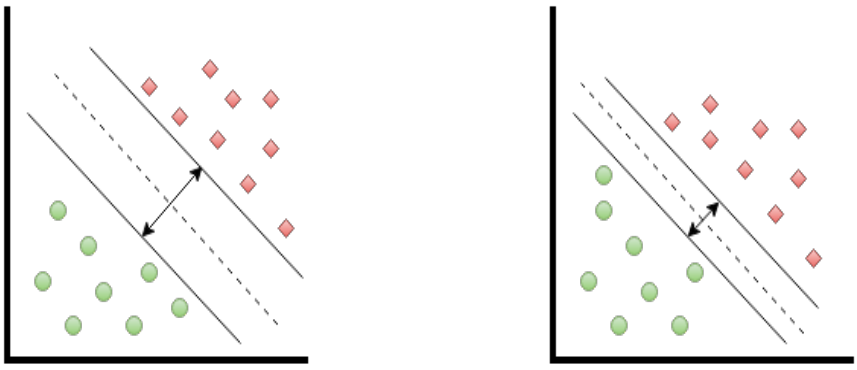


Figure 3–2: SVM Hyperplanes [8].

Finding vector w^* and scalar b^* that minimize $||w||$ subject to $y_i(wx_i + b) \geq 1 \quad y_i \in \{-1, 1\}, i = 1, ..., m$ determines a linear support vector classifier or generalization function as:

$$g(x) = sgn(wx_i - b) \tag{3.4}$$

where x_i that lie nearest to the maximal-margin hyperplane are known as the support vectors.

If T is not a linearly separable dataset kernel functions are used to map the data from one domain to another where data can become linearly separable. Kernel equations can be *linear*, *quadratic*, *Gaussian*, etc. Linear kernel equations take the form:

$$K(x_i, x) = x_i \cdot x \tag{3.5}$$

The support vector classifier or generalization function for nonlinear classification is then expressed as:

$$g(x, \alpha) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) - b\right) \tag{3.6}$$

subject to

$$\forall_i : 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

where x_i with $\alpha_i \neq 0$ are the support vectors, n is the number of support vectors, and C is the cost of classification error.

The optimal decision boundary or maximal-margin hyperplanes are determined by computing α_i and b that solve the Lagrange maximization problem:

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) y_i y_j \tag{3.7}$$

subject to

$$\forall_i : 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

CHAPTER 4
METHODOLOGY

This chapter presents a detailed explanation of the solution use to implement our predicate test automation problem. Section 4.1 mentions the tools used to implement the PTAP. Section 4.2 details the PTAP architecture, including the specific techniques used to implement each component. Section 4.3 describes how those techniques were employed to perform our experiments and Section 4.4 lists the research questions our experiments aim to answer. Finally Section 4.5 describes the evaluation metrics used to measure our results.

4.1 Text Mining Tools

The main tool we used to accomplish our text classification tasks is known as the Waikato Environment for Knowledge Analysis (WEKA) toolkit [17]. WEKA was developed in the University of Waikato in the upper North Island of New Zealand. It is a workbench for data mining and machine learning that supports tasks like classification, clustering, and visualization [17].

We also made use of TextWrangler, a common text editor for OS X, for creating our dataset in ARFF.

4.2 PTAP System Architecture

The PTAP system architecture is best described in context of the KDD process. An alternate version of Figure 2–2 from Chapter 2, which describes the KDD process, will be used to describe our proposed architecture. Figure 4–1 illustrates the PTAP system architecture and the following subsections describe it in detail.

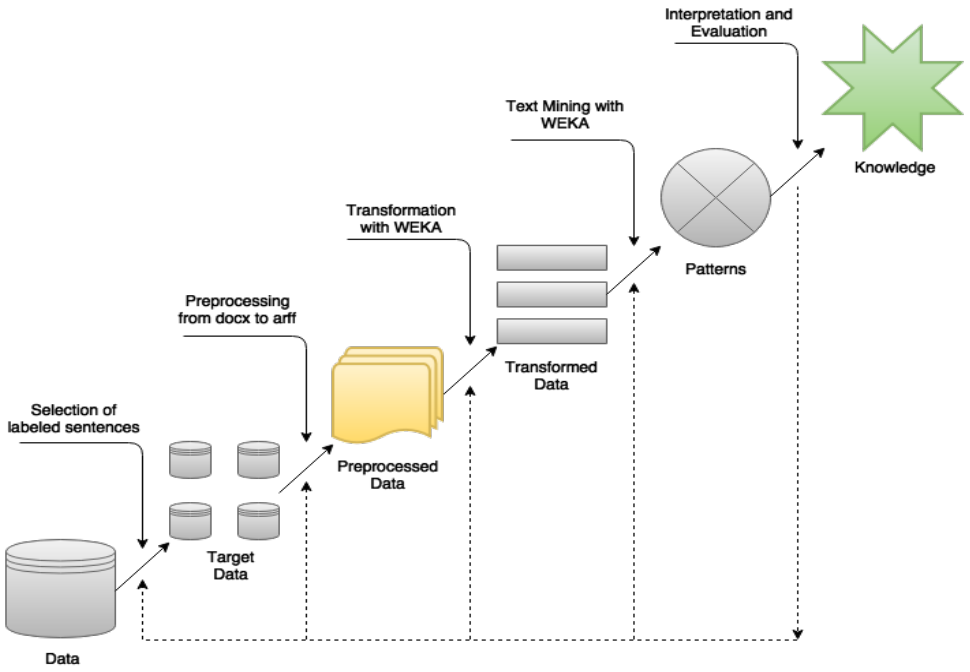


Figure 4–1: PTAP Architecture

4.2.1 Data Gathering

Our dataset was gathered from engineering students enrolled in a midwestern public institution. It consists of textual explanations to multiple-choice science questions [16]. The data was collected as part of a research project where the predicate test was manually performed by educational engineering experts. Among the results of that research is an expert annotated dataset where the verbs or phrases used by students in their explanations are identified as ontological attributes. Furthermore those ontological attributes were categorized into the *emergent process*, *direct process*, and *emergent and sequential process* or *mixed* categories by educational engineering experts as well. [10]

The dataset consists of 680 textual descriptions of students' answers to multiple-choice questions about dynamics and heat transfer. Unfortunately the entire dataset was not labeled by experts. [10]

The sentences that are being considered for classifier training have not been tampered with by experts. This suggests that any computing system can train a

classifier with expert labeled predicate test data and compute predicate test results on student textual descriptions of science concepts that have no expert labels. Using that approach to compute predicate test results could be considered a fully automated predicate test. [10]

The bullet list found below shows textual descriptions written by students to describe their selected answer to multiple-choice questions about diffusion and heat transfer. It was extracted from the original dataset under consideration. According to Chi's predicate test theory each sentence contains *sequential* and *emergent* phrases, which reveal information of each student's mental categorizations of concepts. This phrases were identified by experts and highlighted in yellow if they were *sequential* phrases and highlighted in green if they were determined to be *emergent* in nature. [10]

- Stirring the water causes more molecular motion and allows more salt become evenly dispersed throughout the container
- The air cannot escape and helium can (as stated above), therefore due to random movement of helium, some helium is likely to escape in the process
- The rates will be the same because the non stirred glass will eventually reach equilibrium through diffusion and have the same temperature as the stirred glass.
- The random motion of the dye molecules causes them to collide and move into the beaker with just the water
- Thermal excitation makes molecules move faster, therefore there is an increase of the molecules random motion , therefore the concentration of the dye reaches equilibrium with the water quicker in a heated beaker

In the mentioned research the manually annotated dataset was used to detect if a student's verbal description of concepts is aligned with *sequential* or *emergent process* conceptions. Our available dataset shows the presence of more than one

conceptual schema used by students to describe concepts. This fact points to the use of multi-class, multi-label classification, but the multi-class, single label approach was implemented to show that a binary SVM classifier can be used for misconception assessment automation. [10]

4.2.2 Data Preprocessing

Our preprocessing step was manually completed by constructing a file in ARFF format, which is the document format used by WEKA. Each labeled sentence was extracted from the original dataset to yield a total of 41 instances of the *emergent* class, 99 instances of the *sequential* class, and 50 instances of the *mixed*.

4.2.3 Data Transformation

This step is where our feature space is created. Feature extraction was performed using the WEKA tool. The following describes the WEKA filters used to complete our feature extraction tasks.

Since our experiment is aligned with multi-class classification instances that belong to both the *emergent* and *sequential* classes were eliminated from the original dataset. This was done with a filter known as Remove with Values, which removes class instances with a specified value [38]. In our dataset the samples that belong to both the *emergent* and *sequential* classes were labeled as *mixed*. [10]

The second filter applied to our dataset is known as the Synthetic Minority Over-sampling Technique (SMOTE) [39]. This algorithm considers minority class instances and their k nearest neighbors to generates synthetic data. Synthetic samples are generated in three steps .The first calculates the difference between a minority class feature vector and its chosen nearest neighbor. Then the difference is multiplied by a random number between 0 and 1 and the result is added to the feature vector under consideration [39]. [10]

Our dataset contains 41 examples of the *emergent process* class and 99 examples of the *sequential process* class, leading to class imbalance. We used SMOTE

to correct this considerable bias. *Emergent process* class examples were randomly oversampled by 100% with 5 nearest neighbors to obtain a total of 82 instances from the *emergent process* class [10]. Although the synthetic data points double the amount of *emergent process* class instances they are reliable due to the fact that the original data point are very similar and if additional natural data points were available, they would not be much different in comparison to the ones generated by SMOTE.

Another filter that applied to our dataset for feature extraction is known as StringToWordVector. This filter tokenizes each string attribute in our data samples into a set of attributes consisting of each word in the string and information about word occurrence [10][40] [41].

Considering the first sentence found in the sample dataset from Section 4.2.1 its vector representation would have a component for each word in the entire dataset and information about word occurrence for each word present in that same sentence [10].

In addition feature selection is also supported by the StringToWordVector filter by means of Inverse Document Term Frequency. The following parameters were configured:

- IDFTransform - sets whether if the word frequencies in a document should be transformed into:

$$TD - IDF(t_i, d_j, D) = TF(t_i, d_j)\dot{I}DF(t_i, D); i = 1, ..., m, j = 1, ..., n \quad (4.1)$$

where $TF(t_i, d_j)$ is the number of times term t_i appears in document d_j, D is the data corpus under consideration, and

$$IDF(t_i, D) = \log \frac{N}{|\{d_j \in D : t_i \in d_j\}|} \quad (4.2)$$

where $N = |D|$ (total number of documents) and $\{d_j \in D : t_i \in d_j\} =$ number of documents where term t_i is present

- normalizeDocLength - sets whether if the word frequencies for a document (instance) should be normalized or not
- tokenizer - selects the tokenizing algorithm to use on the strings
- wordsToKeep - the number of words per class to attempt to keep

Feature weighting was also performed with WEKA's AttributeSelection filter. The following parameters were configured for the AttributeSelection filter:

- evaluator - determines how attribute subsets are evaluated
- search - determines the search method used to find attributes based on an information gain threshold

4.2.4 Text Mining

Our text mining tasks were performed using WEKA's implementation of SVM known as SMO, which uses a linear kernel function. SMO is sequential minimal optimization algorithm for training a support vector binary classifier. It was trained and evaluated using 10-fold cross validation.

4.2.5 Interpretation and Evaluation of Results

Our aim was to achieve approximately 75% accuracy, precision, and recall, and 80% F-measure. Further details about our chosen evaluation metrics are found on Section 4.5. Our results discussion is found in Section 5.2 and the answers to our experimental questions are found in Section 5.2.1.

4.3 Experiments

The following is a list of four combinations of data transformation techniques that characterize our main experiments. We chose to document these four approaches due to the relevancy of the classification results obtained. All classification models were build with WEKA's SVM implementation known as SMO with 10-fold cross validation for test/training set partitions.

1. WEKA's SMOTE filter was applied to duplicate *emergent* class instances. Then WEKA's StringToWordVector filter was applied with the following configuration:
    - IDFTransform - false
    - normalizeDocLength - No normalization
    - tokenizer - WorkTokenize with delimiters .,;:'"()?!
    - wordsToKeep - 10000
2. WEKA's SMOTE filter was applied to duplicate *emergent* class instances. Then WEKA's StringToWordVector filter was applied with the following configuration:
    - IDFTransform - true
    - normalizeDocLength - Normalize all data
    - tokenizer - WorkTokenize with delimiters .,;:'"()?!
    - wordsToKeep - 10000
3. WEKA's SMOTE filter was applied to duplicate *emergent* class instances. Then WEKA's StringToWordVector filter was applied with the following configuration:
    - IDFTransform - true
    - normalizeDocLength - Normalize all data
    - tokenizer - NGramTokenizer with delimiters .,;:'"()?! considering phrases of one to three words
    - wordsToKeep - 10000
4. WEKA's SMOTE filter was applied to duplicate *emergent* class instances. Then WEKA's StringToWordVector filter was applied with the following configuration:
    - IDFTransform - true
    - normalizeDocLength - Normalize all data

- tokenizer - NGramTokenizer with delimiters .,;:"'()?! considering phrases of one to three words
- wordsToKeep - 10000

Finally WEKA's AttributeSelection filter was applied with the following parameters:

- evaluator - InfoGainAttributeEval
- search - Ranker with a threshold of 0.

4.4 Experimental Questions

With the experiments described in Section 4.3 we sought to answer the following research questions:

1. Is is possible to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16] with SVM?
2. Is it possible to build a successful classification model using complete student descriptions about concepts as the only source of input for our feature space creation?
3. Which if the ontological categories proposed by Chi in [16] can be predicted successfully?
4. Which combination of feature extraction, selection, and weighting techniques is best suited for classifying learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16]?
5. Can Chi's predicate test be automated with text mining techniques?

4.5 Evaluation Metrics

Classification models are commonly evaluated using measures from information retrieval. Commonly metrics used to evaluate classification results are known as accuracy, precision, recall and F-measure, kappa coefficient, and error rate. Classification results are usually presented in the form on confusion matrices. Refer to Section 2.4.1 for details about classification measures and confusion matrices.

Unbalanced datasets that are measured with accuracy and error rate usually lead to mis-interpretation of results. For example if a dataset has a high frequency of positive instances and a trivial classification model makes positive predictions only, then accuracy measure would be high and error rate would be low, but these results are useless. The same occurs when there is a high frequency of negative data points and a classifier predicts all instances to be negative. This is the reason why recall, precision, and F-measure are preferred instead of accuracy and error rate. Another relevant measure taken into consideration is known as Kappa Coefficient (κ).

To evaluate our experiments we chose a data corpus partitioning strategy known as k-fold cross-validation. The strategy partitions the data set into k-partitions and each partition is used as a training set with all the remaining partitions used as the test set. The results of each constructed model was averaged to determine the final classifier results.

For the aforementioned experiments we chose to use 10-fold cross-validation to partition our dataset. We also chose to describe relevant results in terms of accuracy, precision, recall, F-measure, and κ.

CHAPTER 5
EXPERIMENT RESULTS AND DISCUSSION

This first section of this Chapter presents the results we obtained from each constructed prediction model discussed in Section 4.3.

Section 5.2 of this Chapter contains a detailed discussion of our experimental results.

5.1 Experimental Results

Our four experiments yielded an average accuracy measure of 79.5%. This was computed using the average accuracy between each class in each model and then averaging each model's average accuracy.

We also obtained an average kappa coefficient of 0.5501 using all four experiment results.

Table shows the averaged measures of precision, recall, and F-measure.

Table 5–1: Averaged Experiment Results Using 10-Fold C-V

Label	Avg. Precision	Avg. Recall	Avg. F-measure
Emergent	0.845	0.62.2	0.709
Sequential	0.776	0.904	0.834

Table 5–2 shows the measured and expected accuracy of each constructed model using 10-fold cross-validation, as well as each model's kappa coefficient.

The rest of the tables found in this chapter (tables 5–3 to 5–10) show measures of precision, recall, and F-measure, as well as confusion matrices resulting from each of our experiments respectively.

Table 5–2: Measured and Expected Accuracy and Kappa Coefficient for Each Model Using 10-Fold C-V

Experiment	Accuracy	Expected Accuracy	Kappa Coeff.
Experiment 1	76.24%	50.69%	0.5181
Experiment 2	77.34%	50.06%	0.5366
Experiment 3	81.77%	51.94%	0.6206
Experiment 4	82.86%	63.90%	0.5251

Table 5–3: Experiment 1 Support Vector Machine Results Using 10-Fold C-V

Label	Precision	Recall	F-measure
Emergent	0.753	0.707	0.730
Sequential	0.769	0.808	0.788
Average	0.762	0.762	0.762

Table 5–4: Experiment 1 Confusion Matrix

	Emergent	Sequential
Emergent	58	24
Sequential	19	80

Table 5–5: Experiment 2 Support Vector Machine Results Using 10-Fold C-V

Label	Precision	Recall	F-measure
Emergent	0.797	0.671	0.728
Sequential	0.759	0.859	0.806
Average	0.776	0.773	0.771

Table 5–6: Experiment 2 Confusion Matrix

	Emergent	Sequential
Emergent	55	27
Sequential	14	85

Table 5–7: Experiment 3 Support Vector Machine Results Using 10-Fold C-V

Label	Precision	Recall	F-measure
Emergent	0.962	0.622	0.756
Sequential	0.758	0.980	0.855
Average	0.850	0.818	0.810

Table 5–8: Experiment 3 Confusion Matrix

	Emergent	Sequential
Emergent	51	31
Sequential	2	97

Table 5–9: Experiment 4 Support Vector Machine Results Using 10-Fold C-V

Label	Precision	Recall	F-measure
Emergent	0.870	0.488	0.625
Sequential	0.821	0.970	0.889
Average	0.835	0.829	0.812

Table 5–10: Experiment 4 Confusion Matrix

	Emergent	Sequential
Emergent	20	21
Sequential	3	96

5.2 Results Discussion

The scope of this research focuses an a novel approach to achieve predicate test automation with text classification techniques in order to avoid learning difficulties in new topics due to accumulated misconceptions. According to the results presented in Section 5.1 the proposed can PTAP successfully predict samples from the *emergent* and *sequential* classes. This section details our analysis of the results found in Section 5.1.

Our first classification model derived from experiment 1 has an accuracy score of 76.24%. This is our lowest obtained accuracy score. In terms of precision and recall 70.70% of *emergent* class samples were correctly classified as such, where each data point labeled as *emergent* has a 75.30% chance of actually belonging to the *emergent* class. Furthermore this model scored a recall of 80.80% for *sequential* class samples with 76.90% precision.

This first model's kappa score is 0.5181 which is considered fair or good according to the kappa coefficient interpretation guidelines proposed by [4][5][3] (refer to Section 2.4.1).

The classifier's expected accuracy of 50.69% was improved by 25.55% and its kappa coefficient and confusion matrix are indicative of good classifier performance. Although this first model yielded our poorest results the aforementioned reasons justified our use of these results as a benchmark with which to evaluate all consequent experiments.

Results obtained from experiment two show a negligible 1.10% increase in accuracy and a negligible 0.0185 increase in its computed kappa coefficient. Experiment two's average F-measure was chosen to represent its precision-recall relationship. Compared to experiment one's average F-measure experiment two differed by a negligible .9%. For these reasons we determined that the results from experiment one and two are the same in terms of statistical significance.

Experiment three's accuracy was considered as the best accuracy measure with a score of 81.76%. Its F-measure is also the highest with a score of 81%, which is a significantly larger score when compared to the experiments one and two.

Experiment three also has a kappa coefficient measure of 0.6206, which is considered a very good kappa score according to [4][5][3]. This kappa score is considerably higher compared to the other three experiments. When comparing its expected accuracy of 51.94% vs. its measured accuracy of 81.76% experiment three yields an impressive improvement of 29.82% compared to random chance. The classification model constructed as part of experiment three also resulted in the lowest amount of incorrectly classified instances in comparison to the other experiments.

These improvements in F-measure and kappa coefficient for experiment three may have manifested as a result of using the NGramTokenizer to represent student predication with bags of words that closely resemble the phrases which are part of Chi's conceptual schemas.

Results from our fourth experiment varied from our first three due to the use of the AttributeSelection filter applied for information gain. This fourth classification

model yielded the highest accuracy of 82.86%. This could have led us to believe that experiment four yielded the best performance, but upon careful inspection of the statistical significance of those results we concluded otherwise.

Experiment four's results are biased due to the unbalance distribution of data points used as input for the cross-validation process. The unbalanced dataset reflects the use of AttributeSelection filter applied for information gain, which resulted in the removal of *emergent* class data points generated using SMOTE (refer to Section 4.3). In other words the test data used to evaluate this classifier was biased by having twice the amount of *sequential* class samples than *emergent* class samples.

Furthermore the expected accuracy for the fourth model we constructed is 63.90%. This expected accuracy is the highest amongst all other experiment's expected accuracy, meaning that there is a higher probability that if classification is left to random chance the model constructed for experiment four could yield a better performance compared to the other experiments.

It is worth noting that upon considering both the kappa coefficient (refer to Table 5–2) and confusion matrix (refer to Table 5–10) obtained from experiment four it could be possible to use this model for accurate predictions of *sequential* class samples only. This is because experiment four yielded the worst results for the *emergent*.

After analyzing the overall accuracy, precision, recall, F-measure, and kappa coefficient of each classification model we concluded that our best performing classification model resulted from experiment 3. This is because its F-measure is the highest with a score of 81%, its kappa coefficient measure is also the highest with a score of 0.6206, it showed the most significant improvement over random chance, and it resulted in the least amount of incorrectly classified instances when compared to the other experiments.

5.2.1 Experimental Questions Answered

With the results obtained from the experiments described in Section 4.3 we answer the following research questions:

1. Is is possible to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16] with SVM?

Considering the results shown in tables 5–3, 5–5, 5–7 and 5–9, yes it is possible to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16] with SVM. This answer is based on our best performing model, which clarifies unknown data point with an acceptable mean absolute error of 18.23%.

2. Is it possible to build a successful classification model using complete student descriptions about concepts as the only source of input for feature space creation?

This question is also answered with the same results that answer experimental question number one. According to these results a successful classification model can be built using complete student descriptions about concepts as the only source of input for feature space creation. This answer is based on our best performing model, which clarifies unknown data point with an acceptable mean absolute error of 18.23%.

3. Which combination of feature extraction, selection, and weighting techniques is best suited for classifying learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16]?

Our best performing classifier was based on manual feature extraction and bag of words theory (3-Gram Tokenizer) for feature selection. Our results indicate that the feature selection technique known as bag of words used to build the classification model in experiment three yielded the best results. It is with noting that combining

feature selection with bag of words and feature weighting in the form of information gain negatively impacted our classification results.

4. Which of the two ontological categories chosen our classification labels and proposed by Chi in [16] can be predicted successfully using text classification?

It is possible to successfully predict both of the proposed labels (*emergent* and *sequential*) using text classification. The answer to this question is based on the results obtained from our best performing model (experiment three). Refer to Tables 5–7 and 5–8 for details about precision, recall, F-measure, and kappa coefficient for experiment three.

5. Can Chi's predicate test be automated with text mining techniques?

We have shown that Chi's predicate test can be automated by applying KDD theory to a expert annotated dataset that assigns student textual descriptions of science concepts into their corresponding ontological category.

CHAPTER 6
IMPLICATIONS AND FUTURE WORK

6.1 Implications

This research is based on Chi's conceptual change theory. Chi's research arguments that learners assign newly acquired knowledge about a science concept to an ontological category that best describes the nature of the such concept. Misconceptions occur when learners assign an incorrect category to concepts under study. In addition Chi has developed a process known as the predicate test, which aids in the assessment of misconceptions by inspecting student textual descriptions of science concepts and categorizing the descriptions as belonging to an *emergent* process, a *sequential* process, or both. This process is known to be a time consuming task that can only be performed by trained experts from the educational engineering domain.

We sought to automate Chi's predicate test by applying the KDD process to existing predicate test results. We used this previously annotated dataset to train a SVM classifier in order to show that the *emergent* and *sequential* process categories can be successfully learned.

The main purpose of our research was to explore the possibility of reducing the time it takes to deliver predicate test results while helping teachers in the delivery real-time, individualized misconception assessment to students studying science concepts.

Our dataset was collected from engineering students enrolled in a U.S. midwestern public institution. Student explanations to multiple choice question answers were considered documents for classification belonging to two categories: *sequential* and

emergent. Furthermore we experimented with bags of words for feature selection and information gain for feature weighting to determine which is best suited for predicting the *emergent* and *sequential* ontological categories. We built our classification models with WEKA's SVM implementation known as SMO and evaluated our classification models based on measures of accuracy, precision, recall, F-measure, and kappa coefficients.

According to our classification results we have concluded that it is possible to classify learner descriptions about concepts into the *emergent* and *sequential* ontological categories proposed by Chi [16]. We also found it is possible to build a classification model of acceptable performance using complete student descriptions about concepts as the only source of input for our feature space creation. Furthermore, we have determined that our best performing classification model resulted form experiment number three, which is based on using bags of words theory for feature selection without the use of feature weighting. We have arrived at this conclusion because experiment number three has the highest score of F-measure of 81%, its kappa coefficient measure is also the highest with a score of 0.6206, it showed the most significant improvement over random chance, and it resulted in the least amount of incorrectly classified instances when compared to the other three experiments.

A positive contribution to the science of educational engineering has resulted from the combination of Chi's predicate tests with KDD theory. With the intent to automatize Chi's predicate tests we have documented a methodology (PTAP) for exploring the possibility of having teachers deliver real-time, individualized misconception assessment to students. This is because the PTAP can reveal insight about a learner's mental categorization of concepts aiding teachers to pinpoint learner misconceptions and knowledge gaps.

Although the PTAP can quickly become useful for speeding up misconception research it can further be used as an essential component in a system that automatically collects and classifies student data to provide a fully automated (including data gathering) predicate tests.

The PTAP has the potential to serve as on of many basic building blocks used in the development of modern e-learning systems due to its positive impact on the efficiency of transfer of knowledge from teachers to students, which affects academic achievement in all educational systems.

6.2 Future Work

Issues for future work can be described in terms of classification algorithm of choice, the supervised vs. unsupervised learning paradigm, and dataset alteration.

First of all we have not yet explored the possibility of improving our results by using alternate classification algorithms. For example Naïve Bayes (Multinomial Naïve Bayes), Multinomial Logistic Regression (Maximum Entropy) [42], or Neural Networks [43]. The results of exploring the performance of these classifiers as part of our research's future work would be necessary to identify which is better suited for the task of predicate test automation.

Another topic to be considered for future work is centered around the absence of an expert annotated dataset for classifier training. The research would focus on prototyping a solution for predicate test automation using clustering to gather similar student descriptions and assign a category to them. The resulting clusters can be inspected in order to document wether the clustered categories are useful or relevant within the context of conceptual change assessment. This clustering approach could also yield a training set that can be used in the construction of a classification model based on supervised learning [25][40][44].

The discussion found in Section 4.2.1 describes the available dataset as lacking in size and consisting in entire textual description, the phrases experts identified

as features for categorization, and the category they actually belonged to. Our classification results can be improved by increasing the size of our dataset with an even distribution of samples for each class. In addition since we only explored the use of student textual descriptions a the only input for feature space creation we are aligned with the notion that using the annotated phrases as the input for feature space creation may improve classifier performance. It is also possible that the combination of textual descriptions and the phrases that experts used to identify student categorization of concepts can both be used to improve performance.

Considering the social, political, intellectual, and economic context of a learner can also open the door to another variant of future work for this research. That would lead to the exploration of the unmeasured relationships between the social, political, intellectual, and economic stories present in a learner's life at a given point in time, and how that learner is achieving the understanding of science concepts and the retention of that newly acquired knowledge.

Future work for this research encompasses applying the PTAP to the domain of mathematics as well.

APPENDICES

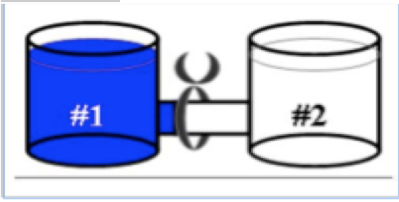
APPENDIX A
RAW DATASET EXAMPLES

A7. S1. Q2 (Paragraph)

Dogs are well known to have a tremendous sense of smell. They can find objects and people by their scent. Explain how a dog can find you by your scent.

Molecules of your scent are released into the air where due to thermal motion, the mix with the air randomly and eventually some of them reach the dog's nose.

Suppose you have 2 beakers connected by a short tube with a clamp. Beaker #1 contains a highly concentrated solution of darkly colored blue dye (and water), and Beaker #2 contains no dye, only water. We will refer to Beaker 1 as dye and Beaker 2 as water. Thus there is a high concentration difference between the two beakers. At first the tube is clamped shut and nothing can flow between the two beakers.



A7. S3. Q6 (Paragraph)

When the clamp is removed in Scenario 2, what causes the blue colored solution to appear to flow from Beaker #1 to Beaker #2?

The random motion of the dye molecules causes them to collide and move into the beaker with just the water. (code 1)

Figure A–1: Raw dataset example

Using your knowledge about diffusion, explain how cells receive the oxygen they need from the bloodstream and lose harmful carbon dioxide.

The random (MR) motion of the CO2 molecules will cause them to travel through the membrane where there is a lower concentration of CO2, while the O2 molecules will do the same thing, but in the opposite direction. (code 2 – MR, CE, ER, AS)

A7. S6. Q17 (Multiple Choice)

Pyrex glass is almost impermeable to all gases except Helium (He). If a mixture of air and He (50% of each) is fed through a long circular tube, what would you expect the composition to look like at the exit of the tube?

3.The exit composition would contain more air because ...

A7. S6. Q18 (Paragraph)

Please complete and explain your choice in the above question.

the He was able to diffuse through the sides of the glass tube before it reached the end of the tube. (code 0)

Figure A–2: Raw dataset example

REFERENCE LIST

[1] James D Slotta, Michelene TH Chi, and Elana Joram. Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cognition and instruction*, 13(3):373–400, 1995.

[2] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[3] Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.

[4] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[5] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.

[6] Michelene TH Chi, James D Slotta, and Nicholas De Leeuw. From things to processes: A theory of conceptual change for learning science concepts. *Learning and instruction*, 4(1):27–43, 1994.

[7] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[8] István Pilászy. Text categorization and support vector machines. In *The proceedings of the 6th international symposium of Hungarian researchers on computational intelligence*. Citeseer, 2005.

[9] James W Pellegrino, Naomi Chudowsky, Robert Glaser, et al. *Knowing what students know: The science and design of educational assessment*. National Academies Press, 2001.

[10] Brian A. Landrón-Rivera, Nayda G. Santiago, Aidsa I. Santiago-Román, and J. Fernando Vega Riveros. Binary classification for conceptual change assessment. 2015.

[11] Lorna Rosemary Sibbett. Ensuring each student reaches their potential:(2) transferability issues. 2010.

[12] Michelene TH Chi. Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. *International handbook of research on conceptual change*, pages 61–82, 2008.

[13] Brian P Coppola and Joseph S Krajcik. Discipline-centered post-secondary science education research: Distinctive targets, challenges and opportunities. *Journal of Research in Science Teaching*, 51(6):679–693, 2014.

[14] Michelene TH Chi and Rod D Roscoe. The processes and challenges of conceptual change. In *Reconsidering conceptual change: Issues in theory and practice*, pages 3–27. Springer, 2002.

[15] James D Slotta and MT Chi. How physics novices can overcome robust misconceptions through ontology training. *Manuscript submitted for publication*, 1999.

[16] Dazhi Yang, Aidsa Santiago Roman, Ruth A Streveler, Ronald L Miller, James Slotta, and Michelene Chi. Repairing student misconceptions using ontology training: A study with junior and senior undergraduate engineering students. In *Proceedings of the 2010 ASEE Annual Conference and Expo*, June 2010.

[17] Abdullah H Wahbeh and Mohammed Al-Kabi. Comparative assessment of the performance of three weka text classifiers applied to arabic text. 2012.

[18] Gökhan Özdemir and Douglas Burton Clark. An overview of conceptual change theories. *Eurasia Journal of Mathematics, Science & Technology Education*, 3(4):351–361, 2007.

[19] Ayush Gupta and Andrew Elby. Beyond epistemological deficits: dynamic explanations of engineering students difficulties with mathematical sense-making. *International Journal of Science Education*, 33(18):2463–2488, 2011.

[20] Ayush Gupta, Andrew Elby, and Luke D Conlin. How substance-based ontologies for gravity can be productive: A case study. *arXiv preprint arXiv:1305.1225*, 2013.

[21] Ayush Gupta, David Hammer, and Edward F Redish. Towards a dynamic model of learners' ontologies in physics. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 1*, pages 313–318. International Society of the Learning Sciences, 2008.

[22] Ayush Gupta, David Hammer, and Edward F Redish. The case for dynamic models of learners' ontologies in physics. *the journal of the learning sciences*, 19(3):285–321, 2010.

[23] RSJD Baker. Data mining for education. *International encyclopedia of education*, 7:112–118, 2010.

[24] Duan Li-guo, Di Peng, and Li Ai-ping. A new naive bayes text classification algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(2):947–952, 2014.

[25] Alex Marin, Roman Holenstein, Ruhi Sarikaya, and Mari Ostendorf. Learning phrase patterns for text classification using a knowledge graph and unlabeled data. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[26] Frans Coenen, Paul Leng, Robert Sanderson, and Yanbo J Wang. Statistical identification of key phrases for text classification. In *Machine Learning and*

Data Mining in Pattern Recognition, pages 838–853. Springer, 2007.

[27] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[28] Vladan Devedzic. Education and the semantic web. *International Journal of Artificial Intelligence in Education*, 14(2):165–191, 2004.

[29] Liu Tzu-Chien. Developing simulation-based computer assisted learning to correct students' statistical misconceptions based on cognitive conflict theory, using "correlation" as an example. *Journal of Educational Technology Society*, 13(2):180 – 192, 2010.

[30] Tzu-Hua Wang. Developing an assessment-centered e-learning system for improving student learning effectiveness. *Computers Education*, 73:189 – 203, 2014.

[31] Kennedy E Ehimwenma, Martin Beer, and Paul Crowther. Adaptive multi-agent system for learning gap identification through semantic communication and classified rules learning. In *7th International Conference on Computer Supported Education. In Doctoral Consortium (CSEDU)*, pages 33–38, 2015.

[32] Eva Millán, Guiomar Jiménez, María-Victoria Belmonte, and José-Luis Pérez-de-la Cruz. Learning bayesian networks for student modeling. In *Artificial Intelligence in Education*, pages 718–721. Springer, 2015.

[33] Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Text classification by labeling words. In *AAAI*, volume 4, pages 425–430, 2004.

[34] Deepak Kanojia and Mahak Motwani. Comparison of naive basian and k-nn classifier. *International Journal of Computer Applications*, 65(23), 2013.

[35] Xiaoyun Wu and V Vapnik. Support vector machines for text categorization. *Graduate School, vol. Ph. D. Buffalo, NY: State University of New York at Buffalo*, 2004.

[36] Atreya Basu, Christine Walters, and M Shepherd. Support vector machines for text categorization. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 7–pp. IEEE, 2003.

[37] Li Wang and Li Li. Automatic text classification based on hidden markov model and support vector machine. In *Proceedings of The Eighth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2013*, pages 217–224. Springer, 2013.

[38] Pooja Arora. A comparative study of instance reduction techniques. *International Journal of Advances in Engineering Sciences*, 3(3):7–13, 2013.

[39] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[40] M Jayakameswaraiah and S Ramakrishna. Development of data mining system to analyze cars using tknn clustering algorithm. *International Journal of Advanced Research in Computer Engineering Technology*, 3(7), 2014.

[41] V Umadevi. Sentiment analysis using weka. *International Journal of Advanced Research in Computer Engineering Technology*, 18(4), 2014.

[42] Hengshu Zhu, Huanhuan Cao, Enhong Chen, Hui Xiong, and Jilei Tian. Mobile app classification with enriched contextual information. 2013.

[43] Hugo Jair Escalante, Mauricio A García-Limón, Alicia Morales-Reyes, Mario Graff, Manuel Montes-y Gómez, and Eduardo F Morales. Term-weighting learning via genetic programming for text classification. *arXiv preprint arXiv:1410.0640*, 2014.

[44] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[45] Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference Information Society-IS*, pages 8–12, 2007.

[46] Ralf Krestel, René Witte, and Sabine Bergler. Predicate-argument extractor (pax). *New Challenges For NLP Frameworks Programme*, page 51, 2010.

[47] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 8–15. Association for Computational Linguistics, 2003.

[48] Burcin Acar Sesen. Internet as a source of misconception:" radiation and radioactivity". *Turkish Online Journal of Educational Technology*, 9(4), 2010.

[49] Norm G Lederman, Fouad Abd-El-Khalick, Randy L Bell, and Renèe S Schwartz. Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of research in science teaching*, 39(6):497–521, 2002.

[50] Michelene TH Chi. Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences*, 6(3):271–315, 1997.

[51] Tertia Jordaan. Misconceptions of the limit concept in a mathematics course for engineering students. 2009.

[52] Akiko Saito. Phylogeny, history, and ontogeny of human cognition, 2001.

[53] Miriam Reiner, James D Slotta, Michelene TH Chi, and Lauren B Resnick. Naive physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction*, 18(1):1–34, 2000.

[54] Karen Pine, David Messer, and Kate St. John. Children's misconceptions in primary science: a survey of teachers' views. *Research in Science & Technological Education*, 19(1):79–96, 2001.

[55] Reinders Duit and David F Treagust. Conceptual change: a powerful framework for improving science teaching and learning. *International journal of science education*, 25(6):671–688, 2003.

[56] James D Slotta and Michelene TH Chi. Helping students understand challenging topics in science through ontology training. *Cognition and Instruction*, 24(2):261–289, 2006.

[57] David Jonassen and Susan Land. *Theoretical foundations of learning environments*. Routledge, 2012.

[58] Tamer G Amin. Conceptual metaphor meets conceptual change. *Human Development*, 52(3):165–197, 2009.

[59] Julie C Libarkin and Josepha P Kurdziel. Ontology and the teaching of earth system science. *Journal of Geoscience Education*, 54(3):408, 2006.

[60] Robert Glaser. Components of a psychology of instruction: Toward a science of design. *Review of educational research*, 46(1):1–24, 1976.

[61] Michelene TH Chi, Rod D Roscoe, James D Slotta, Marguerite Roy, and Catherine C Chase. Misconceived causal explanations for emergent processes. *Cognitive science*, 36(1):1–61, 2012.

[62] Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367, 2008.

[63] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Feature selection for ordinal text classification. *Neural computation*, 26(3):557–591, 2014.

[64] Reshma Prasad and Mary Priya Sebastian. A survey on phrase structure learning methods for text classification. *International Journal*, 2014.

[65] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, and Jordan Pascual. Knn based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, 7(1):61–70, 2014.

[66] Vishwanath Bijalwan, Pinki Kumari, Jordan Pascual, and Vijay Bhaskar Semwal. Machine learning approach for text and document mining. *arXiv preprint arXiv:1406.1580*, 2014.

[67] Vasundhara Chakraborty, Victoria Chiu, and Miklos Vasarhelyi. Automatic classification of accounting literature. *International Journal of Accounting Information Systems*, 15(2):122–148, 2014.

[68] Robert Neumayer, Rudolf Mayer, and Kjetil Nørvåg. Combination of feature selection methods for text categorisation. In *Advances in Information Retrieval*, pages 763–766. Springer, 2011.

[69] Razieh Abbasi Ghalehtaki, Hassan Khotanlou, and Mansour Esmaeilpour. Evaluating preprocessing by turing machine in text categorization. In *Intelligent Systems (ICIS), 2014 Iranian Conference on*, pages 1–6. IEEE, 2014.

[70] Nivet Chirawichitchai. Emotion classification of thai text based using term weighting and machine learning techniques. In *Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on*, pages 91–96. IEEE, 2014.

[71] Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 104–111. IEEE, 2011.

[72] Osmar R Zaíane. Building a recommender agent for e-learning systems. In *Computers in Education, 2002. Proceedings. International Conference on*, pages 55–59. IEEE, 2002.

[73] Raymund Sison and Masamichi Shimura. Student modeling and machine learning. *International Journal of Artificial Intelligence in Education (IJAIED)*, 9:128–158, 1998.

[74] Tiffany Ya Tang and Gordon McCalla. Smart recommendation for an evolving e-learning system: Architecture and experiment. *International Journal on*

elearning, 4(1):105, 2005.

[75] Marie Bienkowski, Mingyu Feng, and Barbara Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, pages 1–57, 2012.

[76] Joseph Beck and Xiaolu Xiong. Limits to accuracy: how well can we do at student modeling? In *EDM*, pages 4–11. Citeseer, 2013.

[77] Oded Maimon and Lior Rokach. Introduction to knowledge discovery in databases. In *Data Mining and Knowledge Discovery Handbook*, pages 1–17. Springer, 2005.

[78] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[79] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57, 1992.

TEXT CLASSIFICATION OF STUDENT PREDICATE USE FOR CONCEPTUAL CHANGE ASSESSMENT

Brian A. Landrón-Rivera
Department of Electrical and Computer Engineering
Chair: Nayda G. Santiago
Degree: Master of Science
Graduation Date: December 2016