DEVELOPING TRANSDISCIPLINARY METRICS USING DATA MINING TECHNIQUES

Christopher M. Adams, B.S. Math (UNT) & M. Eng (TTU)

Doctoral Dissertation

In

Mechanical Engineering

Submitted to the College of Engineering at
Texas Tech University in
Partial Fulfillment of
The Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

Approved

Dr. Derrick Tate (Chairman of Committee)

Dr. Atila Ertas

Dr. Hong Chao Zhang

Dr. Roman Taraban

Dr. Eunseog Youn

December 2009

Acknowledgements

I would like to dedicate this Phd to my wife, kids, advisor Derrick Tate, and dissertation committee, management at Raytheon for allowing this research to take place, and my mother and father. I have spent the last 3 years of my life dedicated to this research. Over this time, my wife and kids have seen me mostly behind the computer performing research and data analysis to provide the successful results documented in this paper. I would like to thank them for their encouragement and for bearing with me while I spent night after night and weekend after weekend away from them. I owe the world to their dedication in seeing me through this journey in our life. In addition, I would like to thank Dr. Derrick Tate for providing his technical advisement during my research and for the constructive criticism and informed decision making that helped me along the way.

Furthermore, I would like to dedicate this research to my dissertation committee, Dr. Atila Ertas, Dr. Hong Chao Zhang, Dr. Roman Taraban, and Dr. Eunseog Youn. Their meetings with me and technical insight helped me tremendously along the way to completing my research. In addition, I would like to thank Raytheon for giving me this opportunity. In particular I would like to thank Randy Case and John Zanoff for selecting me for this program. Additionally, I would like to thank John Zanoff for his encouragement during our long weekends in Lubbock. I would also like to thank Roddy Metroka for helping me during busy times at work to understand my priorities and organize them to complete this difficult task.

Finally, I would like to thank my mother, father, mother-in-law, and father in-law. My mother for helping me with my kids and keep my sanity during busy times and my father for working with me to develop software used to perform information extraction of patent data from the world wide web used as part of this research. Finally, my mother in-law, and father in-law for babysitting while I was at work and performing research.

Table of Contents

Abstract

Measuring the level of new product success based on its impact in a given market is very important. Is it possible to create a set of product measures that can be used to assess market impact based on product historical economic factors, functionality, and physical attributes? The approach of this research is to assess whether the transdisciplinarity of new products has a measurable effect on product success. As part of this dissertation, multidisciplinary theories in design and innovation—such as TRIZ, innovation management theory, functional basis of design, and economics—were reviewed to create a set of transdisciplinary metrics. First, natural language processing of patent data is used to quantify the number of product functions and physical components to estimate product degree of ideality. In addition, a prediction model is created using neural network regression techniques to predict the level of invention of a new design. Next, a semantic functional basis of design is created to measure a product's level of functional synthesis. This metric is constructed by using natural language processing and latent semantic analysis to generate a functional basis for ten disciplinary areas of research. Finally, a set of novel transdisciplinary metrics were developed. This set of metrics can be used to quantify the transdisciplinarity of functional and physical terms based on their use in ten disciplinary areas. A neural network prediction model is trained using these transdisciplinary metrics to predict the market impact of a product based on patent citation measures. These metrics are tested using machine learning to train prediction models, validate the models, and test model prediction results using test data

sets. The new method for predicting a product's level of invention received a correlation coefficient of .98 for training data and .90 for test data representing a high accuracy in prediction results for the model. In addition, the new set of transdisciplinary metrics received a correlation coefficient of greater than .50 based on test results and validates the contribution of this research. The research contribution is a method for creating a set of transdisciplinary metrics and the application of these metrics in a machine learning model to predict the success of new designs.

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Statement of Research

This research demonstrates how the integration of metrics from engineering design processes such as TRIZ [1], Innovation Management [2, 3], Functional Basis of Design [4, 5], and economics can be used to form the basis of a new set of transdisciplinary metrics. The current research includes machine learning techniques and approaches such as artificial neural networks [6-8] to construct predictive models using patent data from the National Bureau of Economic Research (NBER) and United States Patent Office (USPTO) patent databases. This includes creating a set of independent variables consisting of a set of transdisciplinary metrics used in a machine learning model employed to predict the market success of new designs. To complete this task a set of patent data was mined to extract functional and physical descriptions used as part of a set of transdisciplinary metrics. This includes measuring the transdisciplinarity of functional and physical terms mined from text and using 22 transdisciplinary measures to characterize the transdisciplinary (inter-disciplinary, mono-disciplinary, bi-disciplinary, etc) composition of a new invention. These metrics are used in a predictive model to predict the innovation potentialmarket successof new inventions.

The objective of this research is to develop a method to measure the degree by which new designs integrate knowledge across multiple disciplines. This objective is fulfilled

by first creating a predictive model using patent citation measures to estimate the level of invention of a new design. Next, a method is created to develop a semantic functional basis of design for multiple disciplinary research areas. The semantic functional basis of design is used to measure the use of functionality and physical attributes in new inventions. The ability to measure the use of functionality and physical attributes enables the creation of set of transdisciplinary knowledge integration measures. The research topic investigated is the possibility to generate transdisciplinary knowledge integration measures based on the use of terms that span across many disciplines. Next is an overview of the motivation of this research.

1.2 Motivation

The use of data mining and machine learning techniques facilitates the discovery of new knowledge from existing data. The NBER and USPTO online patent databases include a large amount of design data that provides insight into how integrating technologies from multiple disciplines can increase the market acceptance of a design. Can this information be used to develop transdisciplinary metrics for new designs that measure how increasing the use of transdisciplinary design, process, and science approaches may improve the probability that a new design will succeed? The next paragraphs will discuss the current motivation for this research.

The defense industry is a good example of an industry interested in system integration. The systems integration paradigm relies on the integration of technologies from many product categories to enable the execution of a capability or broader system

function. Increasing the ability to measure the success of product innovation will benefit the defense and commercial industries greatly by increasing the breadth of and impact of the technologies employed. Developing design metrics to quantify the transdisciplinarity of a new design helps programs increase their ability to integrate new technologies and measure the success of a design based on the levels of integration across diverse fields and different facets of a company. Using patent data to understand how the integration of diverse technologies has increased the level of innovation provides a useful model for industry to predict he success of future product developments.

Currently, Microsoft is performing research on the use of learning Bayesian networks to find surprising events in large amounts of time-series data [9]. Example surprising events include sporting events, townhall meetings and other community events that affect trafflic flow and congestion. Furthermore, the National Science Foundation has a current initiative titled "Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation [10]." This initiative considers the use of data mining and knowledge discovery to discover knowledge from distributed data across the World Wide Web. All of these current areas of research within academic and industrial fields contribute to the motivation for the research discussed in this dissertation. Next, is an overview of the main objective of this research including a discussion of establishing a set of transdisciplinary metrics for measuring the probability of success for a new design.

1.3 Main Objective

The main objective of this research is to develop a set transdisciplinary metrics used in a predictive model to predict the success of new designs. Invention is the first instance of a new design. The transdisciplinary metrics established as part of this research are based on functionality and physical attributes and are used to build a machine learning model used to predict the success of new inventions. For a new idea such as a generated patent, the measure of acceptance is based on the number of citations or the number of grant dollars received. For new designs, measures such as market success based on patent citations and breadth of impact are used to measure the success of new designs.

This research reviewed common design metrics used from existing transdisciplinary [11, 12] design theories that are proposed to improve a design or idea. Transdisciplinary design processes reviewed include the Theory of Inventive Problem Solving (TRIZ) [1, 13-15], innovation management [1-3, 16-19], functional basis of design [4, 5], and economics. This set of transdisciplinary design theories are reviewed to establish a set of transdiscplinary metrics used in a machine learning model to predict the success of new inventions.

The relationships in the predictive models were developed using the techniques of data mining, text mining, natural language processing, latent semantic analysis, and machine learning by utilizing mathematical operations such as regression, classification, supervised learning, and statistical analysis. The next section includes a discussion of the research question for this dissertation.

1.4 Research Question

This research answers the question of how can new designs be evaluated based on their use of transdisciplinary design, process and science approaches? In addition, is there a link between the use of concepts from many disciplines and design breadth of impact on future inventions? Finally, can transdisciplinary knowledge integration measures be employed to measure new design success and economic impact in the marketplace? The methods used in this dissertation include an approach to extract physical and functional information from patent descriptions. In addition, a set of transdisciplinary metrics are constructed to predict the success of new products and designs using machine learning techniques. A set of independent variables were constructed using patent data from the NBER and USPTO patent databases. A set of dependent variables were selected from previously defined NBER patent data fields. The dependent variables selected include forward importance, citations received and generality. The independent and dependent variables were used to form the basis of a set of predictive models. The predictive models set the basis for prediction using the transdisciplinary metrics to evaluate the success of new products and designs. Finally, the transdisciplinary metrics are validated using a predictive model to evaluate product success using a subset of the patent data to train the model and the remainder of the patent data to verify the model and validate the metrics. The next section will discuss the significance of this research.

1.5 Significance

The significance of this research is centered on the need to form a transdisciplinary basis for measuring the success for new products and designs. Currently, a set of metrics that unifies engineering design concepts from many disciplines does not exist. In addition, data mining and machine-learning techniques have not been significantly employed to generate a predictive model that predicts design success based on the use of cross-disciplinary knowledge. Two areas of current research that are exploring the use of data mining and machine-learning techniques and approaches to generate new knowledge are the activities conducted by the National Science Foundation and Microsoft.

The National Science Foundation currently has an initiative titled "Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation [10]." Performing research in the area of data mining of NBER patent databases [20] will augment current research performed by the National Science Foundation. In addition, Microsoft is performing research on the use of learning Bayesian networks to find surprising events in large amounts of time series data [9]. Microsoft's research includes creating a directed graph of events that affect the flow of traffic. Nodes on the graph include sporting events, holidays, weather, time of day and other factors that may result in traffic congestion. The significant contribution of this research is to employ data mining and machine-learning techniques and approaches to form predictive models that can be used to evaluate the success of new designs. In addition, establishing a definition and means to quantitatively measure the transdisciplinary metrics and validate the measure of existing engineering design metrics will provide a contribution to the transdisciplinary field of research.

1.6 Scope of Work

The next section will provide an overview of the work included in this doctoral dissertation. It includes a summary of using computer aided methods to estimate TRIZ metrics and TRIZ concepts like contradictions, using natural language processing and latent semantic analysis to develop a semantic functional basis of design, and building prediction models using a set of novel transdisciplinary metrics as inputs to an artificial neural network model.

1.6.1 Computer-Aided TRIZ Metric Estimation

Degree of Ideality is defined in TRIZ as "The benefit to cost ratio of the system or the ratio of its functionality to the sum of various costs associated with the building and functioning of the system" [1]. In addition, a design's level of invention is defined based on the type of design conflict resolved for a new invention and the number of disciplines used in resolving the conflict [2].

Table 1-1 includes the criteria for the five levels of invention from TRIZ. This table also includes the percent share of US patents that are estimated to be level 1 – level 5.

Table 1-1 TRIZ Five Levels of Invention

Level	Criteria	Share
Level 1	Apparent solution: A component intended for the task is used. No system conflicts are resolved.	32%
Level 2	Small improvement: Existing system slightly modified. System conflicts are resolved by the transfer of a solution from a similar system.	45%
Level 3	Invention inside paradigm: System conflicts are resolved by radically changing or elimination at least one principal system component. Solution resides within one engineering discipline.	19%
Level 4	Invention outside paradigm: System conflicts are resolved. A new system is developed using interdisciplinary approaches	4%
Level 5	Discovery: Resolving system conflicts results in a pioneering invention. Often based on recently discovered phenomenon.	<0.3%

An approach for calculating a patent's degree of ideality and level of invention from patent data can be created. These measures can be used to identify example designs that can be used as reference points during early phases of the design process to support design functional modeling and concept generation [3, 4]. This research will discuss a computer-aided approach for extracting design functional and physical information from patent data. This approach will be used to generate hierarchical and non-hierarchical functional and physical models that are utilized to estimate TRIZ metrics. First, an overview of the use of natural language processing of patent data to extract design information from patents is provided. Second dissertation describes the use of patent design information to estimate the degree of ideality for each patent. Next this research provides a discussion of how patent citation measures [5], such as originality, number of

backward patent citations made, number of forward patent citations received and the mean forward and backward citation lag can be used as training data to classify patents into the five levels of invention using machine learning techniques. Finally, section a discussion of how TRIZ metrics such as degree of ideality and level of invention can be used to support design concept generation and functional modeling during early phases of the design process.

1.6.2 Developing a Semantic Functional Basis of Design

The next section includes an overview of using natural language processing techniques to verify the functional basis of design for mechanical engineering developed by Stone and Wood. In addition, a method to generate a functional basis of design for other disciplines is discussed. This semantic functional basis of design is used to calculate transdisciplinary metrics.

1.6.2.1 Verification of Functional Basis of Mechanical Design

In order to verify the functional basis of mechanical design proposed by Stone and Wood, the frequencies with which mechanical functions appear in patents is analyzed. First, the functional terms listed in the Class, Secondary and Tertiary lists are analyzed to determine with what frequency they appear within the patent descriptions. Stone and Wood claim that the functional basis of mechanical designs in [4, 5] represent a set of mechanical design functions that can be used to evaluate new designs in the design process. Upon analysis of the term frequencies it was discovered that Class functions from functional basis of design do not appear as frequently in patent descriptions. The

function terms Branch, Channel, Signal and Provision, which reside in the Class (primary) category of mechanical functions, did not appear frequently in the analyzed patents. This suggests that a semantic functional basis of design must be developed to capture all potential functional terms used by designs. This is performed by using natural language processing to extract design functionality from text. Furthermore, a semantic functional basis of design is developed based on using verbs and objects that appear frequently in patent descriptions.

1.6.3 Developing transdisciplinary Knowledge Integration Measures

The functional basis of design, developed by Stone and Wood, [4, 5] provides a taxonomy of commonly used functions employed by mechanical designs. The functional basis of design provides a list of functions that can be used by designers to represent the operations performed by a mechanical device or artifact. Created as part of the engineering design process, the functional architecture represents the purpose of a design including how functions are used to meet customer requirements. The functional architecture of a design represents functions and sub-functions implemented by a system's design parameters and physical components. Individual functions are expressed as action verbs and objects that represent a system, device, or sub-functions performed by one system component on another system component.

The starting point of this activity is the use of the concepts from functional basis of design to generate a list of physical design terms and functional terms. This list of functional and physical terms is extracted from USPTO patent documents to develop a list of basic terms used by different disciplines. The list of functions and physical

components are used in assessing the transdisciplinarity of new designs and products. This measurement is based on the use of a set of proposed transdisciplinary metrics. This measurement uses the frequency of terms that appear in a new invention. Transdisciplinary metrics are used to assess the success that a new design or product can expect in terms of breadth of impact on other products and economic impact. The importance of this method is to understand how the difference between using "abstract and case specific knowledge" [12] influences design and product success. Higher values of design transdisciplinarity indicate a higher degree of design abstraction or "generality," and a lower value of design transdisciplinarity shows a more case-specific product or design.

1.6.4 Overview of Dissertation Sections

The second section of this dissertation provides a background of the current research consisting of a literature review of current engineering methodologies that are used in this work to create a set of transdisciplinary knowledge integration measures. These engineering methodologies include transdisciplinary and interdisciplinary research, the theory of inventive problem solving (TRIZ), and the functional basis of design. In addition, natural language processing (NLP) software, latent semantic analysis (LSA), and current implementations of text mining to extract design functional descriptions from patent text, and current research in the field of machine learning will be discussed. Chapter 3 of this dissertation presents the theoretical approach taken to establish transdisciplinary knowledge-integration measures. In addition, included in chapter 4 of this dissertation is a description of the software architecture developed and the software

implemented to extract the functional and physical design terms from patent text descriptions, perform latent semantic analysis and principal component analysis (PCA) to create a list of functional and physical design terms for each discipline, and extract first- and second-generation patent citations to measure the forward importance of a patent. This includes providing an overview of software implemented in the Visual Basic programming environment for the analysis. Chapter 5 presents a method to test the transdisciplinary metrics by using the metrics in a neural-network machine-learning model to predict invention importance and breadth of impact. Finally areas for future research are discussed, including the further use of transdisciplinary knowledge integration measurements—calculated from patent descriptions—to predict probability of success for new designs. Figure 1-1 provides a flow chart that shows the structure of this dissertation. It includes the major chapters of the dissertation as well as supporting appendices.

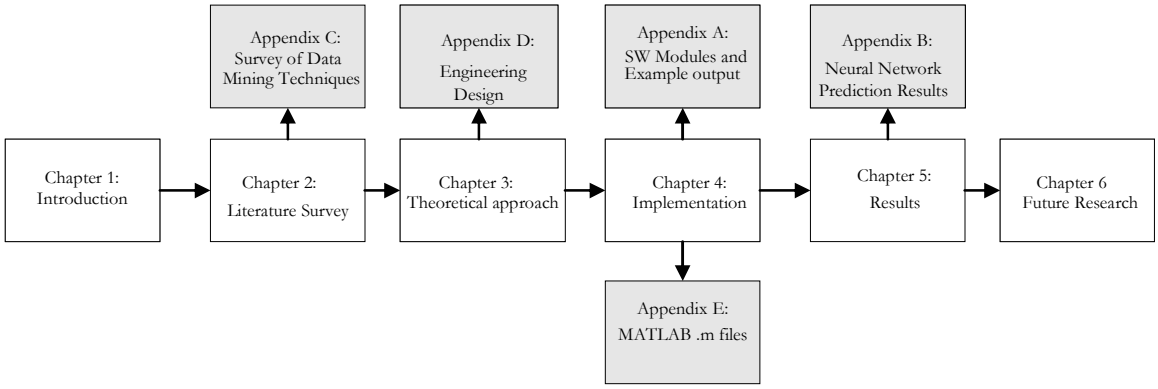


Figure 1-1 Flow Chart Showing Dissertation Structure

CHAPTER 2

LITERATURE SURVEY

2.1 Transdisciplinary Research

Transdisciplinary design process and science provides a novel method for generating products that have depth and breadth of impact on future engineering outcomes. Tanik, Ertas and Maxwell in [21] discuss a novel research model for transdisciplinary Design Process and Science education. In addition, Gumus in [22] developed a transdisciplinary life-cycle management process based on the methods of Axiomatic Design [23] and Complexity Theory [24] developed by Suh. In addition, Chapter 1 of the *Handbook of Transdisciplinary Research* [12] cites the following aspirations of transdisciplinary research:

a) "to grasp the relevant complexity of a problem

b) to take into account the diversity of life-world and scientific perceptions of problems

c) to link abstract and case-specific knowledge, and

d) develop knowledge and practices that promote what is perceived to be the common good"

Transdisciplinary metrics are constructed in this dissertation to measure the amount of abstract and case-specific knowledge in new designs. This is performed to assist in

developing designs that include a high level of generality, creativity, and have a large impact on future designs.

Other related transdisciplinary research includes methods for measuring transfer of knowledge between disciplines. In [25], Bordons et al. investigate cross-disciplinary knowledge transfer between authors. Bordons et al.'s, paper analyzes how authors collaborate across different disciplines. The analysis performed in their dissertation is conducted by reviewing bibliometric citations to understand the flow of knowledge from one discipline to another. The research in this dissertation is similar in motivation to the research by Bordons et al. This dissertation proposed a means to quantify knowledge integration across multiple disciplines by developing a set of transdisciplinary metrics to evaluate the mix of disciplinary knowledge residing within new inventions. Breschi et al. also discuss the subject of knowledge flows between patents in [26]. They discuss links that exist between citations made by patents to other patent documents and the linkage that exists between citations made and citations received by patents. They infer that this linkage is correlated with knowledge flows between inventors and that patent citation data can be used to predict the economic and market success of new inventions.

Furthermore, Brusoni et al. in [27] investigate the integration of knowledge between scientific publications and patent data. This includes study of knowledge flows between different firms and different technological sectors. Brusoni et al.'s paper also includes an overview of different knowledge generation processes. The paper discusses measures for quantifying the breadth of a company's knowledge base. In addition, it proposes a "Relative Specialization Index (RSI)" estimated based on the number of citations that

occur from different scientific fields. This is very similar to the measures of originality and generality discussed by Jaffe and Trajtenberg in [20]. Originality is based on the number of citations made to different USPTO patent classification categories and how these citations are distributed as a percentage of the overall number of citations made to patents in the past. In comparison, generality is a similar measure based on the number of citations received in the future and whether a patent receives citations across a number of different technological categories. The research in this dissertation incorporates generality from the work of Jaffe and Trajtenberg into a method for quantifying the breadth of an inventions' impact. Machine-learning models are developed as part of this research to predict the breadth of impact that a new invention will have, given the mixture of functional and physical design information used across diverse disciplines. Finally, Ziedonis et al. in [28] investigate the economic value of patents based on the number of citations they receive. Their research infers there is a link between the numbers of citations received by patents and the economic and market success of a new invention. This implies that patent citations can be used as a success measure for new inventions that is related to a patents' economic value. This research demonstrates that a predictive model can be created to predict the economic value of a new invention in terms of its expected citations based on the functionality and physical components that are incorporated among multiple disciplines.

2.2 Engineering Design Process

Fey and Rivin define the concept of degree of ideality as the following: Degree of ideality: "The benefit-to-cost ratio of the system or the ratio of its functionality, to the

sum of various costs associated with the building and functioning of the system." Degree of ideality can be explained by the following qualitative formula:

$$\text{Degree of Ideality} = \frac{\text{Functionality}}{\text{Costs+Problems}}$$

In addition, the idea of level of invention for a newly generated patent, design or idea is defined by Fey and Rivin [1] to be the following: Level of Invention: "Altshuller suggested dividing all inventions into five novelty levels[1]:

Level 1 A component intended for the task is used. No system conflicts are resolved.

Level 2 Existing system is slightly modified. System conflicts are resolved by the transfer of a solution from a similar system.

Level 3 System conflicts are resolved by radically changing or eliminating at least one principal system's component. Solution resides within one engineering discipline.

Level 4 System conflicts are resolved and a new system is developed using interdisciplinary approaches.

Level 5 Resolving system conflicts results in a pioneering invention (often based on a recently discovered phenomenon.)"

Furthermore, Utterback and Suarez [19, 29] discuss the concept of Dominant Design and how it is a key aspect of generating sustained innovation. Chen, Li, Huang and Roco [30] discuss the creation of a patent citation network and use of the network to understand knowledge transfer across technical fields.

Additionally, Stone and Wood [5] developed a set of functions and functional flows and further discussed this development in [4]. Functional basis of design developed by Stone and Wood defines functions and functional flows that will be useful as part of this proposal research. As an extension to the research put forth by Stone and Wood, the research discussed in this proposal will suggest using the functions defined in [5] to form a typical set of functions employed by designs. Mining functions and functional flows from patent data will enable the measurement of the degree of ideality [1, 5] for new designs.

Functional basis of design provides a taxonomy of commonly used functions and functional flows employed by mechanical designs. The number of functions and functional flows employed by a design also provides a useful metric for design evaluation. [4, 5] Stone and Wood [5] developed a set of functions and functional flows and further discussed this development in [4]. The further development of a semantic functional basis of design defines functions and functional flows that will be useful for disciplines other than mechanical designs. Pahl and Beitz were among the originators of creating a functional basis of design[31]. They developed a "generally valid list of functions" consisting of the functions *change*, *vary*, *connect*, *channel*, and *store* that are typically applied to the conversion of energy, materials and signals identified as the basic objects residing in a verb-object pair. In addition, Altshuller, the founder of the theory of inventive problem solving (TRIZ) [13], claimed that all mechanical designs can be described using 30 basic functions [13]. The National Institute of Standards and Technology (NIST) also developed a general taxonomy of functions to use as the basis of

mechanical designs. This list consists of 100 functional flows and 130 functions that make up the NIST taxonomy [32]. In addition, Kirschman and Fadel also discuss the classification of functions for mechanical design into a taxonomy [33]. This article will discuss natural language processing of patent data to extract design information in the form of action verbs, objects, and subjects that describe a design in the form of a list of design functions and physical components.

Table 2-1 Functional Basis of Design List of Mechanical Functions includes a taxonomy developed by Stone and Wood. This dissertation will discuss natural language processing of patent data to verify the claims of Stone and Wood and discuss automating the process of creating a semantic functional basis for multiple design classes.

Table 2-1 Functional Basis of Design List of Mechanical Functions

CLASS	Secondary	Tertiary
Branch	Separate	
		Divide
		Extract
		Remove
	Distribute	
Channel	Import	
	Export	
	Transfer	
		Transport
		Transmit
	Guide	
		Translate
		Rotate
		Allow DOF
Connect	Couple	
		Join
		Link
	Mix	
Control Magnitude	Actuate	
	Regulate	
		Increase
		Decrease
	Change	
		Increment
		Decrement
		Shape
		Condition
	Stop	
		Prevent
		Inhibit

CLASS	Secondary	Tertiary
Convert	Convert	
Provision	Store	
		Contain
		Collect
	Supply	
Signal	Sense	
		Detect
		Measure
	Indicate	
		Track
		Display
	Process	
Support	Stabilize	
	Secure	
	Position	

Currently, a set of metrics that unifies design concepts from many disciplines does not exist. Jaffe and Trajtenberg discuss the idea of "basicness," which means that a new idea is "basic" if it has a "diffused", substantial impact across a number of different fields as well as a significant impact on a single field [20]. In comparison, Rowlands introduced the idea of journal diffusion factors as a way to "measure" the breadth of a journal's knowledge across literature [34] This was also stated by Frandsen, et al. as the transdisciplinary reception of a journal [35].

In addition, Jaffe and Trajtenberg also stated that it is necessary to develop "backward-looking" and "forward-looking" measures that will give insight into the "basicness" of an idea [20]. Backwards-looking measures give insight into the source and the history of the research associated with an idea and forward-looking measures give insight into the impact the idea has on future ideas [20]. Therefore, it is possible to develop a set of transdisciplinary metrics to use in a machine learning model that provides a method to predict he probability of success for a new design/idea by looking at the "backward-looking" measures that have historically caused an idea to succeed.

2.3 Data Mining and Machine Learning

Knowledge discovery and data mining [6, 36-40] are techniques used to generate new knowledge during from existing data. The five-step knowledge discovery and data-mining process is discussed by Kawasaki, Ho and Granat in [39]. The first step in the data-mining process includes understanding the domain in which data resides and

understanding problems that will be solved by mining the data. Step two in the process is to collect and preprocess data. This step is the most work intensive and costly part of the data mining process because data may be retrieved in a number of formats such as text files, database files, multimedia files, etc. The third step in the data mining process includes extracting patterns/models from the data to discover new knowledge. The fourth step of the process includes interpreting and evaluating the newly discovered knowledge. The final step in the process is to use the mined data in a pragmatic way.

Data mining can be very useful in the engineering design process to help improve one's understanding of innovation. For example, data mining can provide designers and inventors with a means to review historical market trends that lead to the development of a new product or service. Data mining can also be useful in discovering new technological and market knowledge by uncovering surprising technological events [9] that result in new products, resolve design contradictions [1], or result in a dominant design [19, 29, 41]. It can also be useful to determine when technological discontinuities [2] may occur for a product or process resulting in a new technological innovation.

Machine learning is defined as "extracting patterns relevant to predictive attributes using one or more" data mining algorithms [38, 42, 43]. In addition, machine learning is focused on the development of data mining algorithms that enable computers to learn from data [8]. This research will explore many different data mining and machine learning techniques in creating a set of transdisciplinary metrics to estimate the potential new designs have for innovative impact.

Horvitz in [9] discusses mining of traffic congestion data and the use of the data to predict surprising events. Horvitz defines surprising events as the following:

"To identify surprises, we compare the output of the marginal models with the real-time states to identify rare flows and congestion. We mark these situations as situations that would likely be surprising to users." [9]

Basically, a surprising event is an unusual event that does not follow a model of normal activities.

Zhu and Porter [44] discuss the concept of mining information from text. Ho and Granat in [39] discuss the overall process of mining events through the use of data mining. Daim et al. discuss the use of mining patent bibliometric data in [45]. In addition, Cascini, et al.. discuss implementing Natural Language Processing of patents in [46-48], in order to automate the extraction of patent functional descriptions and design components from a patented design. The complete set of functions extracted from a patented design can be used to represent a patented design.

A list of a design's functions and components can be used for analysis of patents to determine the novel approach taken by the patented design that formed the basis for the patent. This information can then be used to develop transdisciplinary metrics related to functional synthesis. The following machine learning and data mining techniques and approaches were reviewed as part of this research: genetic algorithms and genetic programming [8, 49], support vector machines [7, 50-52], artificial neural networks [6-8], decision trees [8, 53-55], clustering [7, 8], and learning Bayesian networks [7-9, 56-58].

Heckerman discusses learning Bayesian networks in more detail in [56, 57]. An overview of Bayesian networks including a discussion of conditional independency and directed acyclic graphs is included in [42]. A further discussion of these techniques is in included in Appendix C.

Kusiak in [59, 60] discusses a process for innovation and the link between the use of data mining of patent data to define metrics for creativity and innovation. In his paper he discusses data mining techniques such as clustering analysis, genetic programming and decision trees. Kusiak also discusses methods from TRIZ for resolving design conflicts and contradictions.

Another form of data mining, also known as text mining, consists of the use of natural language processing to perform part of speech tagging of textual information to extract relevant data from design descriptions. Test mining is also referred to as information retrieval, information extraction, or knowledge management. [61] A number of natural language processing software packages have been developed that help in the process of extracting relevant information from textual descriptions. One such software package is MontyLingua [62]. MontyLingua [62] is a software developed to perform natural language processing of text. MontyLingua was implemented using the Python programming language. MontyLingua is claimed to be “an end-to-end natural language processor with common sense” [63]. MontyLingua provides different tools to process English text that range from semantic processing of meanings from text to summarizing textual paragraphs and sentence summarization in the form of verb, subject, object, object phrases. MontyLingua is stated to contain “common sense” due to the incorporation of

rule-based part-of-speech tagging methods originally developed by Eric Brill [64, 65]. Eric Brill's rule based process is incorporated into MontyLingua's part-of-speech (POS) tagger, MontyTagger.

There are currently multiple different approaches for conducting part of speech tagging of text. The first is by using statistical based stochastic processes to determine the part of speech of words in a text using probabilities. The second is based on using rule based methods to perform part of speech tagging. In general, stochastic part of speech tagging methods use a Hidden Markov Model (HMM) process where probabilities are set by first assuming a fixed set of tags that represent the part of speech for each word in a text. The tags are also set by assuming that each word in a text is based on a fixed vocabulary that forms the English language. Part of speech tagging is then performed by looking at the process used to generate the text, which may consist of looking at neighboring words that surround the text and determining the probability that a word in the text is a noun, verb, adjective, adverb, etc. based on an HMM developed using a set corpus of documents as the basis for the stochastic model.

Typically the Brown Corpus is used as the basis of the part of speech tagging HMM. The Brown Corpus was used by Charniak at Brown University to develop the part of speech tagging model discussed in [66-68]. The part of speech tagger used as part of the MontyLingua engine is a form of the Brill Tagger developed by Eric Brill also known as the MontyTagger [62-65]. The Brill Tagger is a rule based part of speech tagger that determines the part of speech for words in a text by reviewing the part of speech of words that surround the word to be tagged. The rule based part of speech

tagger developed by Brill has seen a prediction rate of between 96.7 and 97.2% when applied to the Brown Corpus [64]. This can be compared to the HMM part of speech tagger developed by Charniak which has a prediction rate of 96.45% when used on the Brown Corpus [66-68]. The error rates from the two part of speech taggers, whether statistical or rule based are close to the same. The approach taken as part of this research uses the rule based approach for part of speech tagging using the open source NLP engine MontyLingua. The MontyLingua part of speech tagger was selected due to its use of open source software, ease of integration with other software, and more efficient code that does not rely on a large database of statistics to tag the part of speech for each word in a text document. In addition, the results of the MontyLingua software to extract SAO instances were compared to the results obtained by Invention Machine Corporation using their patented SAO extraction method [69]. Figure 2-1 includes the SAO extraction for the following source sentence as described in USPTO patent #6167370 assigned to Invention Machine Corporation:

> The present invention shields a noise of an external magnetic field with the slider and improves a recording performance because the slider is isolated magnetically.

Subject	Action	Object
Present Invention	Shield	Noise of External Magnetic Field
Present Invention	Improve	Recording Performance
	Isolate	Slider

Figure 2-1 Invention Machine SAO Extraction Example

C:\SAO\SAO6167370.TXT

Verb	Subject	Object 1	Object 2	Object 3
shield	present invention	noise	of external magnetic field	with slider
improve	recording performance	because slider		
isolate	slider			

Figure 2-2 SAO Extraction From MontyLingua

Figure 2-2 provides the SAO extraction from MontyLingua using the same source sentence. As can be seen from the two tables, both SAO extraction methods provide the same results. The Invention Machine Corporation patented SAO extraction method is also a rule based part of speech tagger.

Part of Speech tagging has also recently been performed using support vector machine methods. Use of support vector machines to perform part of speech tagging was recently demonstrated by Gimenez and Marquez in [70]. This method uses the widely implemented SVMlight software to perform part of speech tagging developed by Joachims [51, 52]. The accuracy of the support vector machine part of speech tagger is competitive with the rule based and HMM taggers with a total accuracy of 97.16% [70]. In addition, Marquez and Rodriquez developed a decision tree based part of speech tagger [71] that also has performs well when compared to the HMM and Brill rule based taggers.

In addition, Chu and Shu develop a method using natural language analysis to identify biomimetic functions and other cross-domain terminology that can be used in the

concept generation phase of a design effort in [72-77]. Shu's et al. efforts help identify unique functional terms from biology text that will help designers create new ideas when developing new systems. Shu et al.s, process is based on using WordNet [78] as a keyword database to identify synonymous terms that reside across multiple domains. Yang and Cutkosky in [79-82] also discuss using text mining and data mining to develop engineering design thesauri based on keywords found design text that can be used during the concept generation process. Their method uses tools such as singular value decomposition to reduce the dimensionality of term frequency and document frequency matrices so that the most popular words found in specific documents can be used later in the design process. Futhermore, Li, et al.. in [83] discuss developing computer-aided tools to perform semantic processing of design text using techniques such as latent semantic indexing, ontology engineering and natural language processing.

Horvitz in [9] implements learning Bayesian networks to model traffic congestion in metropolitan areas. This leads to the prediction of surprising events. Where surprising events include sporting events, holidays, weather and other circumstances that lead to traffic bottle necks . The research discussed in this proposal covers the prediction of surprising events in the area of transdisciplinary design and process science and will review the use of learning Bayesian networks, neural networks and support vector machines to predict the occurrence of these events based on trends found in patent data. In addition, Horvitz in [9] demonstrates the use of machine learning techniques and discusses training a learning Bayesian network by using approximately 75% of data from a metropolitan traffic database and then verifying the model using the remaining 25% of

the data from the database. This approach will be taken to validate and verify the predictive model used to develop a set of transdisciplinary metrics.

Trappey, et al. discuss using a neural network trained using back propagation in [84] to classify patents based on the international patent classification (IPC). In addition, Trappey, et al.. Include an overview of several machine learning methods such as Naïve Bayes, K-Nearest Neighbor, and the genetic algorithm. As part of their analysis, it was shown that the neural network back propagation model performed best when classifying patents into different IPC categories. There neural network achieved a test data correlation coefficient of 0.90 which is a very good fit for the classification model. In addition, Fattori, et al.., [54] discuss using decision trees to classify patent current awareness bulletins based on their information content. Furthermore, Matthews in [85] discusses building a machine-learning model using a Bayesian Belief Network so that engineers can rapidly explore the design space of a new project. The model uses a new information content metric developed by Matthews that is based on exemplary design solutions of automobiles. In addition, Loh, et al.. [55] discuss using multiple machine-learning methods to classify patents into different categories based on TRIZ inventive principles [86, 87] used in the patents. In Loh et al.'s. paper, several machine learning methods are used such as k-nearest neighbor, decision tree, support vector machine (SVM), and Naïve Bayes to classify patents. The software package used to train the classification models is WEKA [88].

WEKA is used in this research along with Matlab [89] to train regression and classification models to predict the success of new designs based on a set of

transdiscplinary metrics. Regression methods are used for data classification when continuous dependent variables consisting of numeric data are used as target data for a machine learning model. Classification methods use nominal target data as dependent variables in the machine learning model. WEKA can be used to train either regression or classification models. [88] WEKA includes a number of different classification and regression methods. Figure 2-3 shows a screen shot of the WEKA interface that provides a list of Bayesian and functional machine learning models that can be used for either numeric or nominal classification of training and target data.

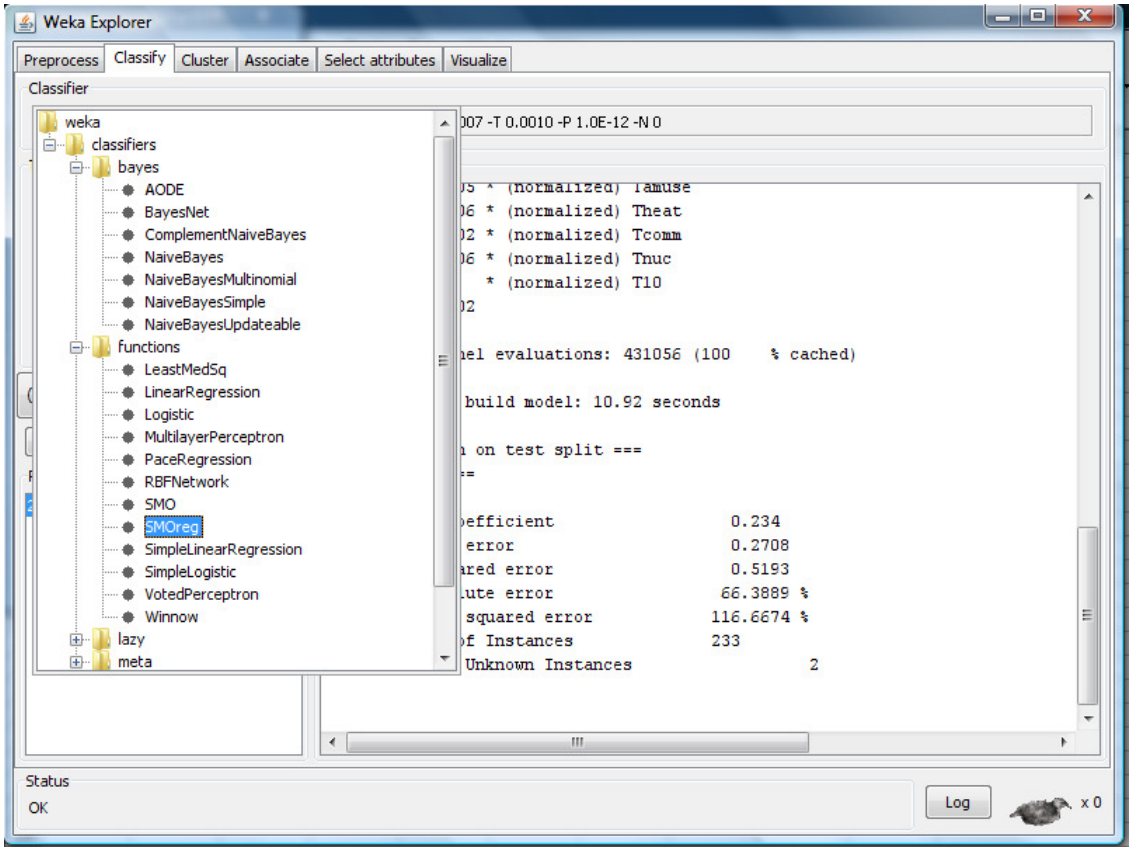


Figure 2-3 WEKA Screenshot

A number of articles have been written that discuss mining patent data to extract relevant design information, its use in the concept generation phase of product development, and the creation of models of design innovation. Tan Runhua, et al. discuss patent text mining in [90, 91] as part of an effort to pursue computer aided innovation similar to the work performed in this dissertation. Next, Chapter 3 includes an overview of the theoretical contribution put forth as part of this research.

2.4 Latent Semantic Analysis

Latent semantic analysis is used widely by web-based search engines such as Google to index webpage text to enable information retrieval of web content. [82] Latent semantic analysis is used to characterize a set of documents, or corpus, by calculating the frequency with which each term in a document appears and then determining the relevance of the term to a specific set of documents. This is accomplished by using the term frequency –inverse document frequency (tf-idf) method to index a set of documents [82, 92]. Latent semantic analysis consists of first creating an m x n matrix that represents the term frequency in the matrix rows and the document in the matrix columns. This generates a tf-idf index matrix represented in matrix form by the following equations [82, 92]:

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{\mathrm{n}_{\mathrm{i,j}}}{\sum_k \mathrm{n}_{\mathrm{k,j}}}$$

Where $n_{i,j}$ is the number of appearances of term t_i in document d_j and the denominator represents the total number of appearances of all terms in the document. Inverse document frequency is described by the equation below:

$$\mathrm{idf}_{\mathrm{i,j}} = \log \frac{|\mathrm{D}|}{|\{\mathrm{d}_j : \mathrm{t}_i \in \mathrm{d}_j\}|}$$

Where |D| is the total number of documents included in the corpus under consideration and the denominator in the log equation consists of the number of documents in which the term t_i occurs.. This process gives rare terms more weight in the tf-idf matrix. Thus, the cross product of the term frequency and inverse document frequency yields the tf-idf matrix shown by the equation below:

$$\text{tf-idf}_{\mathrm{i,j}} = \mathrm{tf}_{\mathrm{i,j}} \times \mathrm{idf}_{i,j}$$

Before creating the tf-idf matrix a number of text preparation techniques must be employed. Two of the techniques employed as part of this research include first, removing stop words from the text—frequently used words like *and*, *be*, *because*, etc.—and then using Porter's stemming algorithm to remove the letters at the end of commonly used words [93].

Once the tf-idf matrix is generated, the next step in latent semantic analysis consists of reducing the number of dimensions in the matrix. To reduce the number of matrix dimensions techniques such as singular value decomposition, or principal component

analysis [94], and clustering techniques can be used to generate a smaller number of uncorrelated variables in the matrix from the total set of terms [40, 95, 96]. These approaches group common terms usually according to the synonymy of the words [97-99]. Principal component analysis was chosen as the method to reduce dimensionality of the matrix. Specifically Hotelling's T^2, which provides the variance of terms in the matrix, was selected to rank terms based on their relationship to a specific discipline.

CHAPTER 3

THEORETICAL APPROACH FOR PREDICTING INVENTION SUCCESS

This chapter includes an overview of three theoretical approaches developed as part of the contribution of this dissertation. The first theoretical approach consists of using computer aided methods to estimate metrics from the TRIZ. Degree of ideality from TRIZ is estimated by using natural language processing to extract functions and physical attributes from patent text. In addition, level of invention is estimated by building a machine learning model using artificial neural networks based on patent citation measures. Next a theoretical approach for building a semantic functional basis of design is discussed. Finally, an approach is discussed for creating a set of transdiscplinary metrics. These transdisciplinary metrics are used in a machine learning model to prediction the success of a new invention.

3.1 TRIZ Degree of Ideality and Level of Invention Estimation

This section will discuss the theoretical approach to estimate the value of two TRIZ metrics, degree of ideality and level of invention.

3.1.1 TRIZ Degree of Ideality Estimation

Patent component names and numbers can be constructed in a hierarchical list by reviewing action verbs such as "comprise, form, have, include, mount, etc" that indicate that certain components are subcomponents of other components [47] Figure 3-1 provides an example hierarchical component list assembled by extracting the component names and verbs using MontyLingua.

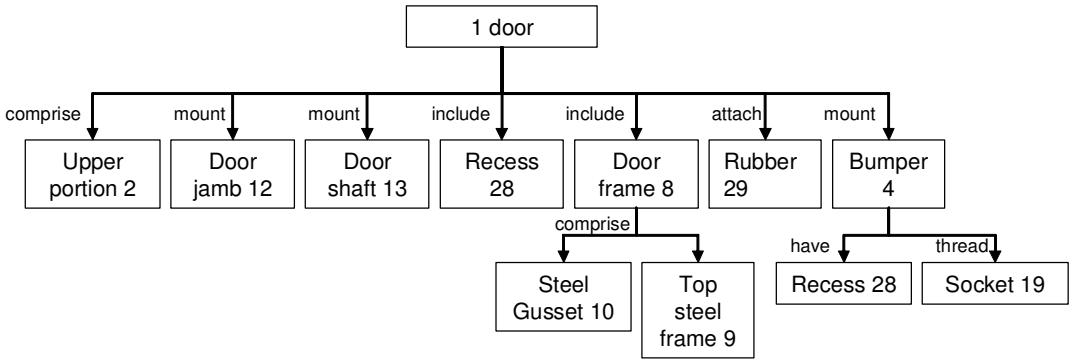


Figure 3-1 Component Hierarchy for Patent #3,858,357

Once the hierarchical list of components is generated, a list of functions performed by the design can be created for each hierarchical level. This can then be used to generate patent functional models to review component and function relationships that increase a design's benefit to cost ratio or degree of ideality.

The next section will discuss using computers to estimate the level of invention for a patented design.

3.1.2 TRIZ Level of Invention Estimation

A patent's level of invention can be estimated by using patent citation analysis, patent generality and originality measures [20] as independent variables in a supervised learning model to classify patents into the five TRIZ levels of invention. The steps involved in using software to aid in the estimation of patent level of invention consist of first estimating the level of invention manually for a set of patents to use as training data for a supervised learning model. Next, data from the National Bureau of Economics Research (NBER) patent database is used as a training data set consisting of data fields such as number of citations made and number of citations received. In addition, the measures of patent generality and originality from the NBER database is used in the training set. The originality measure is calculated based on the number of patents cited by the patent under analysis that are from different patent classes. The measure of originality is calculated using the following equation [20]:

$$o_i = 1 - \sum_{k=1}^{n_i} \left(\frac{b_{ik}}{b_i}\right)^2$$

Where i is the patent under consideration, b is the number of patents cited, and k indicates the subclass of the CITED patent as indicated in the NBER database. For example if one patent cites 3 patents and 2 of the patents are from subclass X and 1 patent is from subclass Y, then the originality measure is 1- $((2/3)^2 + (1/3)^2) = 0.44$. A patent's generality is measured in a similar way, but considers forward patent citations to different patents from different subclasses. The measure of generality is calculated using the following equation [20]:

$$g_i = 1 - \sum_{k=1}^{n_i} \left(\frac{f_{ik}}{f_i}\right)^2$$

Where f is the number of patents citing patent i, and k indicates the subclass of the cited patent as indicated in the NBER database. Finally, citation information such as the mean forward citation lag and mean backward citation lag is also used as part of the network training data. This data is used to determine the breadth of influence a patented design has on future inventions.

Table 3-1 includes an example of the training data set used to classify patents by level of invention. It includes the independent variables as well as dependent variable, level of invention. This training data can be used with a number of different machine learning techniques to perform data classification. The machine learning technique used to perform the classification in this example is the artificial neural network back propagation algorithm supplied in MATLAB Neural Network Toolkit. (Other machine learning techniques that can be used include support vector machines and Naïve Bayes Networks.) An artificial neural network is used to train a classification model using an expanded set of training data, similar to the example training data shown in Table 3-1

Table 3-1 is used to estimate the level of invention for a large number of patents. The patents were initially selected using the number of citations received as an indicator of patent level of invention. Intuitively, it is expected that patents that receive a large number of citations will be inventive and likely to support many other inventions. In addition, patents with a high forward to backward citation ratio were selected for review to identify whether these patents have a high level of invention. In addition, patents were

only selected from the patents that fall in the category of mechanical designs. Mechanical designs were selected to keep the neural network from artificially selecting patents from fields such as biotechnology that draw from many diverse disciplines. It was found that using patents from many diverse disciplines skews the number of level five patent estimates.

Table 3-1 Example Training Data

patent	cmade	creceive	general	original	fwdaplag	bckgtlag	LOI
4387297	12	233	0.34	0.72	12.52	6.67	2
4575330	18	216	0.80	0.69	8.05	9.72	4
4251798	14	181	0.24	0.64	14.09	5.71	2
4409470	30	178	0.30	0.52	11.02	8.50	2
5040715	18	160	0.49	0.54	5.34	11.28	2
4835834	19	151	0.66	0.77	6.68	21.21	2
4277837	12	144	0.77	0.71	9.42	5.58	2
4361060	31	142	0.71	0.26	9.77	10.03	3
4506387	9	128	0.79	0.81	10.45	8.11	4
4834306	10	127	0.64	0.66	5.32	21.40	2
4369361	13	126	0.25	0.26	12.53	7.38	2
4520817	16	121	0.28	0.34	11.05	17.13	2
4130095	3	121	0.69	0.44	8.64	1.67	3
4944443	16	118	0.66	0.68	6.02	22.63	2
4807222	6	112	0.65	0.67	7.45	3.83	4
4728020	13	109	0.35	0.26	9.08	4.92	2
4127322	3	109	0.79	0.50	15.47	9.67	3
4236880	13	108	0.62	0.66	11.87	13.15	3
4636346	24	104	0.64	0.74	7.77	8.50	3
4303904	5	87	0.80	0.80	13.40	10.80	5
4162397	4	53	0.79	0.44	6.68	7.75	3
4714144	17	16	0.58	0.44	5.19	5.53	1
5265694	3	16	0.12	0.44	1.94	0.67	2
4074996	3	16	0.12	0.44	11.88	11.33	2
4656994	6	16	0.12	0.44	11.88	33.83	2
4717094	15	16	0.23	0.44	7.19	31.73	2
4646904	3	16	0.23	0.44	7.94	2.00	1
4051924	3	16	0.23	0.44	12.88	7.33	1
4385609	7	16	0.30	0.44	3.88	6.29	2

The first patent dataset initially used to train the neural network model epresents a larger number of mechanical and electrical patents than biotechnology and chemical patents. From reviewing the neural network results it was found that biotechnology and chemical patents typically have higher originality and generality scores than other patents. Based on this result it was determined to use only mechanical designs to train the neural network. Future research will investigate a hypothesis that patents that rely on knowledge across many disciplines will result in designs with higher levels of invention. This can be overcome by normalizing the data set by giving biotechnology patents higher generality, originality and citations made and received values based on the values for these measures across many other disciplines.

Patents with a low number of citations made and high number of citations received may be based on a new technological discovery that initiates the evolution of novel technological trends. Patents that receive a large number of citations support the evolutionary development of new technologies since a large number of inventions result from this novel concepts. Using this training data to estimate level of invention, information can also be used to understand trends in design evolution and innovation. This will aid in understanding whether TRIZ level of invention is correlated with other innovation metrics such as the emergence of a dominant design [19] and the evolution of technological discontinuities [3]. The next section includes the theoretical approach for developing a semantic functional basis of design.

3.2 Semantic Functional Basis of Design

The next section includes an overview of using information extraction techniques to calculate the frequency of functional basis of design terms. The frequency of functional basis of design terms is used to create a semantic functional basis of design that can be used early in the design process for concept generation.

3.2.1 Functional Basis of Design Term Extraction and Frequency

To determine the frequency with which action verbs and objects appear in mechanical design patents, an algorithm was written in Visual Basic to count the number of appearances of each functional basis term. Once the frequency of functional terms is determined from the list of action verbs, the frequency of the functional terms from Functional Basis of Design is analyzed. This is constructed by generating a Pareto list of function terms extracted from the patent descriptions and then mapping the term frequencies to the functional terms from Functional Basis of design. Table 3-2 includes an example functional term frequency report that provides a short list of the function terms extracted from the patent descriptions. The table also includes the frequency with which the function terms appear across 1,000 mechanical design patent descriptions. This table does not include all of the functional terms extracted from the patents, but the report indicates that the frequency is based on a total of 39032 occurrences of action verbs extracted from the patent descriptions.

Table 3-2 Patent Description Functional Term Pareto Report

VERB	FREQUENCY	REPORT						
Verbs	Extracted	from	file:	C:\SAOALL\SAOALL_5.TXT				
Filtered	Verbs	contained	in	FILTERED	VERB	OCCURENCES		REPORT:
C:\PATENTREPORTS\VERBFREQ5.TXT								
11/20/2008	8:58:04	PM						
VERB:	FREQUENCY:	PERCENT	OF	39032				
provide	1703	4.36%						
include	767	1.97%						
illustrate	632	1.62%						
extend	589	1.51%						
connect	552	1.41%						
make	528	1.35%						
apply	509	1.30%						
comprise	429	1.10%						
position	395	1.01%						
produce	389	1.00%						
control	359	0.92%						
indicate	327	0.84%						
pass	307	0.79%						
generate	294	0.75%						

3.2.2 Data Representation and Analysis

Once the verb list is created by extracting verbs from the SAO phrases using MontyLingua, unwanted terms are filtered, and functional term frequencies are counted, it is necessary to group the functional terms into functional classes. To classify functional terms into a hierarchy of functions by class (primary) functions, secondary functions, and tertiary functions as performed by Stone and Wood in [4, 5], the functional basis for mechanical systems is used as a template. The functional basis for mechanical designs is provided in Table 3-3. This table is augmented with the term frequencies for each of the functional terms to the right of the function text. The term frequency analysis result is provided in Table 3-3.

Table 3-3 Mechanical Design Functional Basis Term Frequencies

CLASS		Secondary		Tertiary	
Term	**Frequency**	**Term**	**Frequency**	**Term**	**Frequency**
Branch	0.01%	Separate	0.20%		
				Divide	0.06%
				Extract	0.04%
				Remove	0.71%
		Distribute	0.05%		
Channel	0.04%	Import	0.00%		
		Export	0.00%		
		Transfer	0.18%		
				Transport	0.04%
				Transmit	0.35%
		Guide	0.08%		
				Translate	0.05%
				Rotate	0.73%
				Allow DOF	0.00%
Connect	1.41%	Couple	0.33%		
				Join	0.09%
				Link	0.03%
		Mix	0.11%		
Control	0.92%	Actuate	0.16%		
Magnitude		Regulate	0.06%		
				Increase	0.44%
				Decrease	0.10%
		Change	0.24%		
				Increment	0.01%
				Decrement	0.00%
				Shape	0.15%
				Condition	0.03%
		Stop	0.18%		
				Prevent	0.55%
				Inhibit	0.04%

CLASS		Secondary		Tertiary	
Term	**Frequency**	**Term**	**Frequency**	**Term**	**Frequency**
Convert	0.12%	Convert	0.12%		
Provision	0.00%	Store	0.25%		
				Contain	0.51%
				Collect	0.05%
		Supply	0.30%		
Signal	0.05%	Sense	0.16%		
				Detect	0.15%
				Measure	0.15%
		Indicate	0.84%		
				Track	0.10%
				Display	0.04%
		Process	0.13%		
Support	0.70%	Stabilize	0.05%		
		Secure	0.66%		
		Position	1.01%		

The next section will discuss the theoretical approach for creating a set of transdisciplinary metrics to predict the success of new inventions.

3.3 Transdisciplinary Knowledge Integration Measures

A novel approach was developed as part of this dissertation research to establish a set of measures used to evaluate designs based on the number of functions and physical components employed from many different disciplines. The approach includes using natural language processing techniques to extract functional and physical terms used by different disciplines. In addition, latent semantic analysis techniques such as tf-idf text indexing and Hotelling's T^2 are employed to generate a list of key design terms for each discipline. Once key-term lists are

generated, the lists are used to search new design text to identify the use of the functional and physical design terms. This creates a method to generate different measurements of the level of transdisciplinary knowledge integration for a new design used as part of a predictive model for understanding the breadth of impact a new invention will have on future inventions.

The transdisciplinary knowledge integration metrics developed as part of this research consist of first measuring the frequency of functional and physical terms used within one discipline for a set of n disciplines, then measuring the use of functional and physical terms used within two disciplines for a set of n disciplines, then within three disciplines for a set of n disciplines until the interaction of functionality and components across multiple disciplines is completely characterized for a total of n disciplines under study. This provides n measurements that can be used to characterize the transdisciplinarity of a design or process. The first transdisciplinary knowledge integration measure is labeled by the term T_1. T_1 is constructed by calculating the frequency of functional and physical terms that occur within one discipline for each of the n disciplines under study and then measure the integration of knowledge from multiple disciplines employed within a new invention

The next transdisciplinary measure T^2, measures the use of functions and physical components that span two disciplines within a set of n disciplines. This measure calculates the number of functional or component terms that reside within two disciplines out of n disciplines. The measure T^2 is a second measure to T_1 and considers the transdisciplinary knowledge residing within two disciplines for a set of n disciplines where T_1 considers knowledge that resides only within one discipline.

The measure for T^2 can then be repeated to create a T_p measure that represents the terms that span across p disciplines, where p is less than n. The T_p measure is calculated in the same

way as T^2, but takes into account p out of a total of n disciplines. T_n is another measure that considers only functional and physical term transdisciplinary knowledge and does not focus on knowledge that resides with a specific discipline. It follows that as p is greater than half of n, the measure is a stronger term transdisciplinary measure than disciplinary measure and when p is less than half of n the measure is thought of as a more disciplinary measure than term measure. Therefore, the lower the transdisciplinary index of the measure the better the measure estimates the level of transdisciplinary knowledge integration of a new invention.

Another set of measures were created to measure the frequency of physical and function terms from each of the n disciplines. These measures are labeled T_k, where k represents the discipline understudy, and T_k is the frequency of the terms from a given discipline. This allows for the measurement of the number of functions and components that reside within only one discipline. In addition, an interdisciplinary metric and transdisciplinary metric were created to measure the interdisciplinarity and transdisciplinarity of the overall invention. Key functional and physical terms from each discipline are used to measure T_1 through T_n for a large set of patent documents. The transdisciplinary measures are then used with other patent measures such as number of citations made to other patents, patent backward importance measure, and patent originality measure from the NBER database [20] to train a machine learning model to predict the breadth of impact a patent has on later inventions based on its measures of transdisciplinary knowledge integration. Figure 3-2 includes a flow diagram of the inputs and target variables used in the machine learning model. The inputs and targets of the machine learning model are explained further in this section. The machine learning model [100, 101] selected to predict a patents breadth of impact is the neural network back propagation method. Once the machine learning model is developed it is tested against another set of patent data to validate this approach

to predict future impact of a design based on the use of transdisciplinary knowledge integration measurements.

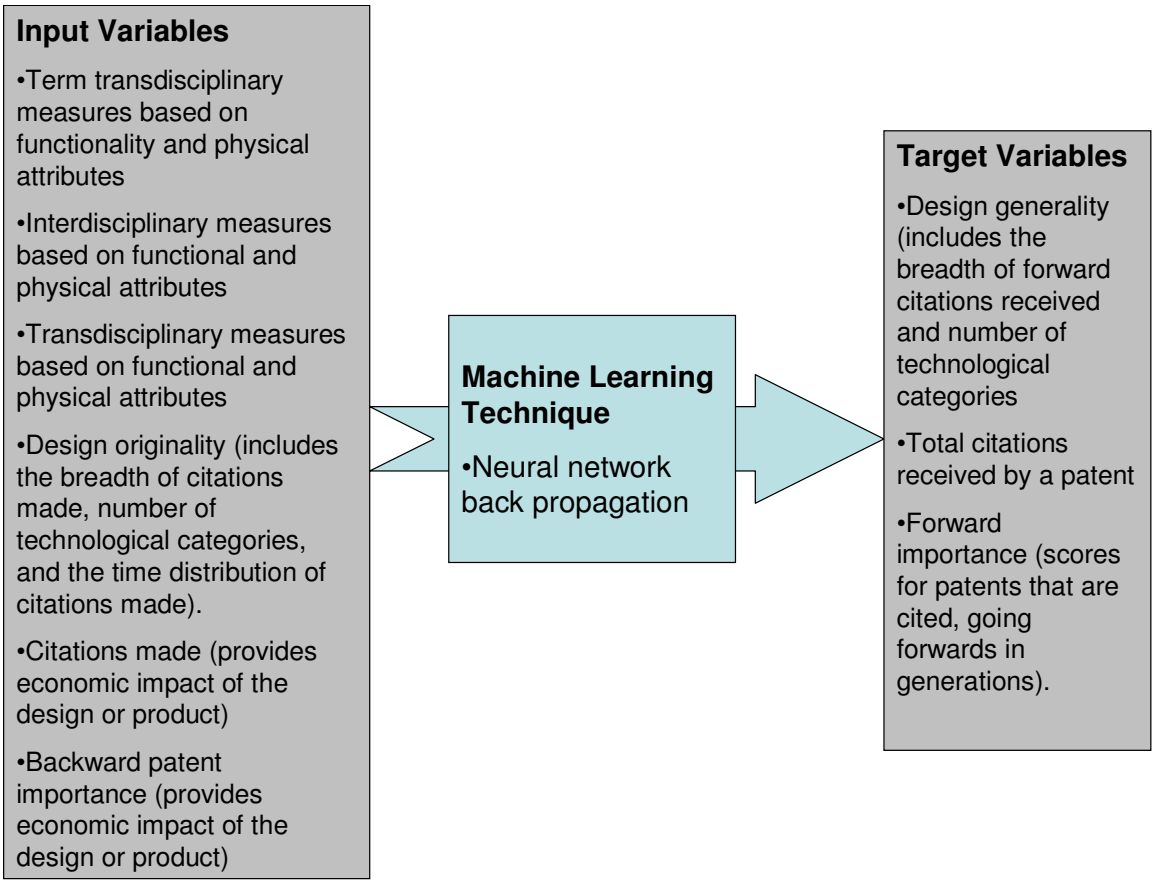


Figure 3-2 Input and target variable flow diagram

USPTO patent documents provide a good representation of a design that includes the functions performed by the design as well as the patented design's components. USPTO patent documents must follow a set of rules that define how a patent document is constructed [102]. This includes identifying sections for a patent document such as patent title, abstract, claims, and patent description. Each of the patent document sections includes useful information that can be used to extract the functional representation for each patented invention. To extract patent functional descriptions from patent text it is necessary to employ natural language processing

(NLP), information extraction and information retrieval techniques. This is conducted by extracting the action verbs and objects applied in patents that represent the functions employed by the patented design.

Natural language processors are used to extract the Subject Action and Object (SAO) from each sentence in a patent's textual description. An SAO is defined as subject, action verb and object of a sentence included in a body of text. As part of this work the NLP software MontyLingua [62] implemented in Python was selected to extract all of the SAO instances that reside in each sentence included a patent's textual description. Once the SAO instances are extracted Latent Semantic Analysis [82, 92, 103] is used to generate a term frequency – inverse document frequency matrix that serves as an index of terms for a corpus of patent documents that resides in a given discipline.

The USPTO currently uses a method for identifying patents that includes 400 unique patent classes. [20] This consists of 36 subcategories of patents and 6 higher level categories of patents. The six categories of patents include chemical, computers and communications, drugs and medical, electrical and electronics, mechanical and other. To establish a list of key functional and physical design terms used by different disciplines, patent text was selected for 100 patent documents from the 1970s, 100 documents from the 1980s and 100 documents form the 1990s that fell within the different categories and subcategories defined in the NBER database.

3.3.1 Transdisciplinary Term Measure Calculation

In order to measure the transdisciplinary integration level of a new design based on the text used to describe it, it is necessary to first gather information from historical data to measure the frequency of which terms appear within different documents classified for a given discipline and across documents that represent many different

disciplines. To solve this task, latent semantic analysis is employed to generate a list of terms that appear frequently within a set of patent documents that fit within ten different disciplines. Latent semantic analysis consists of first creating an m x n matrix that represents the documents under analysis. This consists of creating a matrix where terms that appear in the documents are the rows in the matrix and the patent documents under consideration make up the columns in the matrix.

After constructing the tf-idf matrix we must determine which terms contribute the most to the variance of the matrix. This is accomplished by using principal component analysis to reduce the number of matrix dimensions and create a number of uncorrelated variables that represent many related terms instead of using a matrix with a large number of variables that represent many correlated terms. Another method that can be used to determine what terms contribute the most to matrix variability is the use of Hotelling's T-Square (T^2) method. Hotelling's T^2 method identifies terms that contribute the most to tf-idf matrix variability. Hotelling's T^2 score identifies terms that are farthest away from the matrix mean. Hotelling's T^2 is given by the following equation:

$$T^2 = n(x-\mu)' W^{-1}(x-\mu)$$

Where n is the number of points (the number of terms in the tf-idf matrix), x is a column vector of n elements (x is a vector representing terms that reside in each document), μ represents the mean tf-idf frequency in the tf-idf matrix, and W is a n x n sample covariance matrix (covariance matrix of the tf-idf matrix).

After determining which terms are the most distant from the mean, Hotelling's T^2 is used as a measure to rank the terms that appear in each discipline. Once Hotelling's T^2

function is employed using Matlab to analyze the IDF matrix for each patent, then the T^2 estimate for each term is exported into a .csv format. Each term is then ranked by its associated T^2 measure to create an ordered list of terms by importance. A ranked list of terms for each discipline is then created by sorting the list of terms extracted from the latent semantic analysis in descending order based on its associated T^2 score. Ten disciplines were studied as part of data analysis included in this dissertation based on the work by Jaffe in [20]. The ten disciplines studied in the analysis include chemical, computers and communications, electrical, measuring and testing, nuclear and x-rays, mechanical, heating, amusement devices, biotechnology and other. A sample of the term list for each discipline is provided in Table 3-4.

Table 3-4 Example list of disciplinary terms

Amusement Devices	Heating	Computers & Communication	Nuclear & X-rays	Measuring & Testing	Biotechnology	Chemical	Electrical	Mechanical	Other
speric	comprtment	sfp	mqw	zero-quantum	phenoxyacetyl	bmf	misfet	sub-convey	ti-zr
putter	ptc	pmgr	radon	loop-gap	replicat	urea-sulfur	igfet	airbag	sub-bodi
jackpot	setback	svd	piezo-actu	counterpropag	cftr	dmc	non-barri	helicoid	seizur
lotto	burning-out	mbc	pyro-opt	fid	hpv	photometri	vapor-grown	coiler	weigher
turtl	precalcin	atm	micro-dr	piezoresist	amyloid	polyanhydroaspart	epitaxial	fan-fold	non-stick
honey-gath	heat-convei	itinerari	slit-rai	electrodeform	cryoprotect	hydrocrack	hi-c	signpost	after-touch
outwardli	fin	vlan	photo-luminesc	corioli	angiogenin	cba	loco	sunshad	truss
puzzl	slump	ethernet	electrosprai	microbridg	extracapillari	peroxyacid	non-singl	extruderhead	superthin
hockei	circulating-air	palm-top	probe-carri	liquid-level	integu	contrast-enhanc	spin-on-glass	tilted-up	percuss
moebiu	photoflash	workspac	markabl	motion-encod	cea	photochrom	polyoxid	treadmil	oxide-bas
playfield	fluid-distribut	unipost	near-ir	grase	hind	catam	light-activ	force-resist	parapet
jigsaw	overtemperatur	multicast	secular	flow-encod	transgen	micro-fibers-gener	quantum-wel	uat	scrapper
goaltend	fireplac	snoop	light-detect	lfzp	a-factor	fluoroaliphat	lpd	underscor	cryopanel
three-piec	swirler	timelin	shear-forc	bondpad	unstain	ga-treat	teo	electro-conduct	aluminid
pinbal	air-suppli	vci	fov	magnetoelast	amebocyt	nematogen	ig-fet	unton	cymbal
audiotex	high-calori	subscrib	focussing-error	dient	telomeras	electrograin	tiw	tread	cementiti
racket	blowoff	simm	thermosprai	phase-encod	stromal	azeotrope-lik	split-gat	prepressur	intumesc
doll	windbox	hash	fclum	unsoak	cdna	photocatalyt	anti-punchthrough	sac	earth-lik
hitter	checkerwork	bsr	megavoltag	borehold	dt-a	light-respons	graphoil	ni-ti	dockboard
headfram	spear	keyword	sub-area	bragg-typ	tissue-deriv	expuls	in-process	control-pressur	tex
footbal	lement	edo	ontain	mansfield	biomass	bronst	tiwn	arch-shap	plasma-deposit
poker	vac	over-eras	ftr	proofmass	keratinocyt	photoharden	punchthrough	glasslik	volut
prize-win	heatpip	broker	constrast	rlg	spermatozoon	hydroconvers	ode	doser	congel
dealer	raft	vsync	lessend	lightpip	cmcase	steel-mad	p-conduct	lockr	roll
allei	desalt	parser	visabl	bjt	aav	diboride-bas	conductivity-impart	derailleur	self-set
pressureless	high-tens	pen-bas	cho	coastdown	plasmid	radiation-block	phospho-sil	reinforcing-fram	propylene-ethylen
prize-award	stagnat	raid	lwir	gradient-coil	limulu	raffin	macro	pacemak	shoeboard
basketbal	firebox	full-writ	effect-induc	rason	carolina	projection-typ	virtual-ground	skateboard	splint
sub-cub	kiln-typ	telesketch	nsom	sfm	ionen	photobleach	pre-stag	fixer	hrj
gondola	otari	directori	quistor	comparitor	synthas	tle	rapd	aramid-epoxi	colour-form
coin	pulverized-coal-fir	pram	field-sensit	two-port	mab	tert-butyl	defect-caus	reaction-pl	awn
card-edg	tramp	midi	mid-infrar	y-ax	nrrl	clau	mesfet	heat-solubl	self-flux
skittl	after-burn	cd-i	q-piezoelectr	memor	myeloma	light-modul	requenc	eec	rosari
pre-chosen	silo	mccam	grin	rephas	hemicellulas	octafin	isoplanar	sidepl	asm
keno-gam	atc	dct	tubehead	exor	plasmin	tact	non-ldd	smif	load-sens

The list of terms included in Table 3-4 does not represent the complete list of disciplinary terms. The analysis generated the following number of terms for each discipline:

- Amusement Devices = 5936 terms
- Heating = 5166 terms
- Computers and Communications = 8937 terms
- Nuclear and X-rays = 7185 terms
- Measuring and testing = 6512 terms
- Biotechnology = 12979 terms
- Chemical = 9857 terms
- Electrical = 4269 terms
- Mechanical = 6268 terms
- Other = 8229 terms

To develop a set of transdisciplinary knowledge integration metrics, term transdisciplinarity is measured using Hotelling's T^2 to rank the terms extracted from a set of patents representing a given discipline. First, terms listed in the tf-idf matrix for each discipline are ranked in descending order using the T^2 score. The data analysis considered in this paper uses ten disciplines.

The next step in the process is to create a matrix of terms for each discipline where the rows in the matrix represent terms in the tf-idf matrix for one of the n disciplines used for the measurement. The columns in the matrix represent the disciplines under study. The data in the matrix represents the T^2 score for each term where the first column in the matrix

includes the discipline from which the term list was extracted and the T^2 score terms ranked in descending order. The data in the other columns includes the T^2 score for terms that are found within the other disciplines under study. Table 3-5 includes an example matrix for one of the disciplines used for the data analysis discussed in this dissertation.

Table 3-5 Hotelling's T-Squared Matrix for Each Discipline

BIO	TSQ	BIO in Chem	Bio in Elec	Bio in Mech	Chem	TSQ	Chem in Bio	Chem in Elec	Chem in Mech	Elec	TSQ	Elec in Bio	Elec in Chem	Elec in Mech	Mech	TSQ	Mech in Bio	Mech in Chem	Mech in Elec
polynucleotid	11679	2777			bmf	8460				diamond	3907		5425	4316	bobbin	5845			
bibul	11377	103			urea-sulfur	8395				film	3895	788	3453	1621	grind	5681	806		175
trident	10844				modul	8281	549	2791	1732	sub-lay	3811		366		ply	5577		58	
cassett	10680	9		760	dmc	8000				pre-charg	3782			15	screen	5457	1015	2054	748
zone	10623	2269	2802	1027	silver	7960	582	134	40	storag	3749	1031	1177	1339	code	5436	1540	49	67
compart	10431	4969		964	label	7722	3903	68	890	polyimid	3696	34	2003	511	exercis	5322	24		
cellulas	10404	9			crystal	7688		1299	2237	emiss	3633	4552	262	2024	turbin	5283	29	2692	
electrod	10016	6460	2308	3203	light-sensit	7559	57			sidewal	3623	870	494	844	transfer	5250	8053	2369	1937
phenoxyacetyl	9745				transport	7406	5235	499	1833	pyroelectr	3601				bag	5091		3283	
replicat	9298				toner	7268			3371	oxynitrid	3584		2		fuel	4988	386	2807	
insolubil	9289	111			hollow	7183	4599	757	454	fire	3566	97	533	965	neck	4966	593	1459	126
unsubstitut	9280	2769			grid	7147	33	144	498	tantalum	3563	404	394	3	abras	4966	338	291	222
transesterif	9253	113			filter	7097	2709	524	1066	insul	3552		3515	570	sub-convey	4958			
cftr	9186				amorph	7088	794	3292	916	segment	3522	2223	1705	848	suspens	4892	1769	445	232
hpv	9118				align	7045	619	1004	411	dielectr	3509	357	4484	3236	mold	4883	314	578	1530
amyloid	8830				photoconduct	7015	15	294	2726	logic	3498	570	189	2306	valv	4875	5654	5640	2013
phase	8828	1896	848	2548	solar	6986		3226		isol	3489	473	130	106	assist	4765	196	159	2904
bind	8799			1904	membran	6890	3739		2921	c-mo	3471				scroll	4756			
microlaboratori	8652	5482			fasten	6873			1101	inhibit	3469	1713		103	layer	4698	4651	5664	2191
matrix	8634	2629	1080	205	fibrou	6820	3730		97	silicid	3450	14	289		gear	4687	806	155	
ascorb	8483	138			photometri	6763				protect	3426	2647	390	1840	polish	4673	34	2	2315
analogu	8294				emitt	6751	4113	3423	298	emitt	3423	4113	6751	298	optic	4622	1283	5409	491
cryoprotect	8170				polyanhydroaspart	6735				gate	3421	106	1361	4207	crucibl	4608		4814	833
transfer	8053	2369	1937	5250	whiten	6731	139		30	groov	3412	2228	876	2646	wafer	4606	120	2051	2117
cholesterol	7980	50			platen	6706	43		594	mount	3409	331	2333	3103	panel	4596	918	2515	361
starch	7925	3295		8	core	6661	1197	98	4432	float	3368	72	50	151	belt	4596	776	5876	5
angiogenin	7880				vessel	6569	6810	231	2821	cmi	3359				fill	4531		1556	
cluster	7843		75		reactor	6563	6355	1559	197	transduc	3301	254	117	219	launcher	4528			
cancer	7749	831			slide	6509	2374	501	1541	antifus	3296				core	4432	1197	6661	98
ferment	7684	83		180	electrod	6460	10016	2308	3203	amorph	3292	794	7088	916	imag	4354	3652	3576	1133
extracapillari	7668				tubular	6437	537	6	1643	impur	3266		306	4	conveyor	4342	1082	2538	3
load	7493	2917	1033	420	sputter	6402		1441	6	charge-sens	3252				diamond	4316		5425	3907
wire	7468	2542	1865	2614	boiler	6265				trench	3251		86		limb	4313			
integu	7461				hydrocrack	6166				solar	3226		6986		slave	4312			
wall	7452	3950	2377	2977	jet	6164	1726	1	3669	carbid	3223	128	3277	3052	card	4269	5439		333
nucleic	7352	524	114		cathod	6145	230	615	134	misfet	3193				trim	4261			
micropor	7304	2678			corrug	6067			950	analog	3189	3037	218	145	color	4231	3534	4718	170
surface-act	7222	620			cba	6022				intermetal	3170			98	gate	4207	106	1361	3421
acceptor	7205	686	247	48	pulp	5996			233	depress	3167	173	365	1398	cover	4181	1089	1050	2438
cea	7205				honeycomb	5984			858	memori	3152	703	1152	2649	member	4143	5101	3321	2805
hind	7161				chelat	5973	3493	1159		light-emit	3152	59	56	258	circuit	4141	966	5780	1328

Next, the matrices are used to measure the transdisciplinary level for each term. If a term appears in only one of the disciplinary term lists then it contains no transdisciplinary knowledge and receives a transdisciplinary term measure of 0%. If a term appears in 2 to n disciplines, then it receives a percent transdisciplinary measure that is estimated based on the T^2 score for the term. The transdisciplinary measure for a specific term is estimated using the following equation:

$$t_d = 1 - \sum_{d=1}^{n} \left(\frac{T_d^2}{\sum_{d=1}^{n} T_d^2} \right)^2$$

Where t_d is the term transdisciplinary measure, d is the discipline that includes the term and T_d^2 is Hotelling's T^2 score extracted for a specific term from discipline d.

The term transdisciplinary measure can then be used to evaluate terms extracted from the patent list. Terms with a transdisciplinary measure of 0 are considered disciplinary terms and are included in the list of terms associated with a specific discipline. The lists are ranked in order first using td and then the term T^2 score to perform the ranking.

3.3.2 Transdisciplinary Knowledge Integration Measurement

To measure levels of transdisciplinary knowledge integration, the first step consists of extracting the text from the design description document for the design under study. Stop words must be filtered from the text, then Porter's stemmer algorithm should be applied, and then the frequency of each term in the document is calculated. Terms in the design document are then compared to the list of terms extracted from patent data and a set of transdisciplinary

measures are calculated to measure the transdisciplinarity of a given patent. First, measure of the interdisciplinarity, T_I of a patent is calculated using terms extracted from the document with t_d = 0%. The following equation provides the method to calculate T_I using the term frequency of terms where t_d =0:

$$T_I = 1 - \sum_{d=1}^{n} \left(\frac{\left(\sum_{t=1}^{m} tf_t \right)_d}{\sum_{d=1}^{n} \left(\sum_{t=1}^{m} tf_t \right)_d} \right)^2$$

Where n = the number of disciplines under study. t is a term with td =0%, m = the number of terms in discipline d with t_d = 0%, d = the number of disciplines used to measure transdisciplinarity and *tf* is the term frequency for a given term t.

Next, T_d is constructed to measure the contribution of terms with t_d > 0% that reside within all of the disciplines under study. This set of measures is given by the following equation:

$$T_d = 1 - \frac{\sum_{r=1}^{l} (tf)_r (t_d)_r}{\left(\sum_{r=1}^{l} tf_r \right)}$$

In this equation d represents the number of disciplines under study, r represents the terms with t_d > 0% that reside in all disciplines d, l = the total number of terms that reside in all disciplines d, and *tf* = the frequency of which the term r appears within all disciplines.

Furthermore, a set of measures are generated that estimate the contribution of terms with t_d = 0% that reside within each of the disciplines under study. This set of measures is given by the following equation:

$$T_x = \sum_{t=1}^{m} (tf)_t$$

In this equation x represents the discipline under study, t represents the terms with t_d = 0%, m = the total number of terms with td = 0%, and *tf* = the frequency of which the term t appears within discipline x.

Another set of measures are estimated that constructs the disciplinary composition of the invention. This set of measures reviews the percentage of the invention that is mono-disciplinary, bi-disciplinary, tri-disciplinary, etc . These measures are estimated by summing the term frequencies for terms that appear in one discipline, in two disciplines, in three disciplines, etc, then dividing by the total frequency of terms that reside in a discipline until all of the d disciplines are characterized. Figure 3-3 includes 22 transdisciplinary metrics, T_I, T_d, T_{mech}, T_{elec}, T_{chem}, T_{bio}, T_{other}, T_{amuse}, T_{heat}, T_{comm}, T_{nuc}, T_{meas}, T_1, T_2, T_3, T_4, T_5, T_6, T_7, T_8, T_9, $T_{10.}$. This set of transdisciplinary metrics is used to represent the cross-disciplinary interaction at all levels of transdisciplinary knowledge integration. It includes interdisciplinary interaction, mono-disciplinary interaction and multiple levels of transdisciplinary knowledge integration.

Monodisciplinary				3.59%
Bidisciplinary				1.20%
Tridisciplinary				9.08%
T4				1.20%
T5				4.93%
T6				8.09%
T7				15.20%
T8				4.50%
T9				11.68%
T10				40.53%
Tmech			1.96%	
Telec			3.91%	
Tchem			0.00%	
TBIO			1.96%	
Tother			0.00%	
Tamuse			1.96%	
Theat			64.72%	
Tcomm			7.86%	
Tnuc			17.64%	
Tmeas			0.00%	
Transdisciplinary		16.94%		
Interdisciplinary	54.12%			

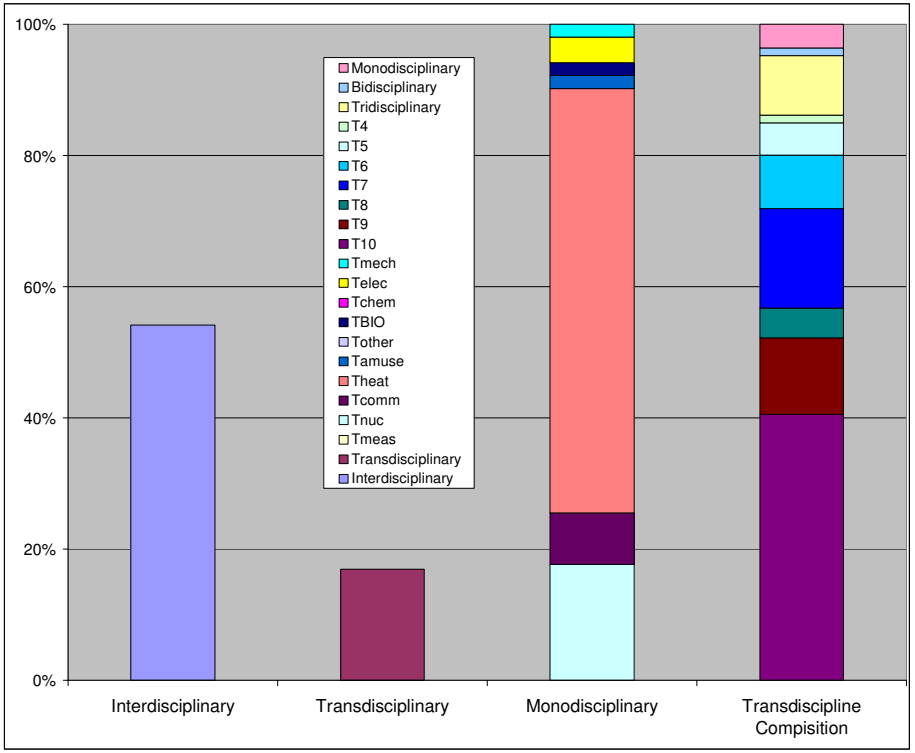


Figure 3-3 Graphical Representation of Transdisciplinary Knowledge Integration Metrics

The next section will discuss the software implemented to extract functional and physical descriptions from patent text. This includes an overview of the patent software toolkit used to extract patent information, perform natural language processing by incorporating MontyLingua, and perform latent semantic analysis of patent text.

CHAPTER 4

IMPLEMENTATION

4.1 Implementation of TRIZ Degree of Ideality and Level of Invention Estimation

TRIZ degree of ideality estimates were performed using natural language processing and machine learning techniques. The next section will discuss how TRIZ metric estimation was implemented in software.

4.1.1 TRIZ Degree of Ideality Estimation

USPTO patent documents provide a good representation of a design that includes the functions performed by the design as well as the patented design's components. USPTO patent documents must follow a structured set of rules that define how a patent document is constructed. [102] Patent document rules describe different sections that must be included in the text of a patent including the patent title, abstract, claims, and patent description. Each of the patent document sections includes useful information. The patent description section can be used to build functional hierarchical models for a patented invention. To extract patent functional and physical component information from patent textual descriptions, it is necessary to employ Natural Language Processing (NLP) techniques [47]. Each patent in the USPTO patent database includes drawings that depict a numbered list of all of the components of the patented design. A list of patent physical components may be used to create hierarchical functional and physical models of a patented design. This is

accomplished by first extracting Subject Action Object (SAO) phrases from patent text including the title, abstract, claims, and description to retrieve each component name and number residing in the patent as well as the action verb and object that reflects the function performed by the physical components.

A number of open source Natural Language Processors were reviewed to perform the extraction of Subject Action and Objects (SAO) from the sentences in patent textual descriptions [62, 63, 66-68]. As part of this research the NLP software MontyLingua [62] implemented in Python was selected to perform the SAO extraction. Specialized software, the Patent SW Toolkit, was generated in the Visual Basic programming environment to extract patent information from the USPTO.gov website. The patent software toolkit graphical user interfaces are provided in Appendix A of this dissertation. This includes example screenshots of the software as well as example outputs that result form analyzing the patent documents. This software is used in the data processing phase to extract patent textual descriptions in HTML format. In addition, the Patent SW Toolkit is used to convert patent textual descriptions previously extracted in HTML format into tagged XML. Tagged XML is created by the Patent SW Toolkit to label and segregate the patent text so that the MontyLingua NLP software [62] can be implemented on different sections. An example of the XML report is provided below:


```
<?xml version="1.0" ?>
- <patentdata>
  <patentnumber>3858357</patentnumber>
```

<Patent_Title>BUMPER MOUNTING FOR FLEXIBLE TRAFFIC DOOR</Patent_Title>

<Abstract>A door having intermediate the ends thereof …

…said bumpers.</Abstract>

<Inventors>McGuire; Winston B. (Plattsburgh, NY)</Inventors>

<Assignee>W. B. McGuire Co., Inc. (Hudson, NY)</Assignee>

<Filed_date>June 5, 1973</Filed_date>

<Current_US_Class>49/460 ; 16/86R; 160/354</Current_US_Class>

<Field_of_Search>49/460,9,34 160/354 16/86R,86A,86B,1,111R,DIG.2 293/62 85/45 29/526 42/74</Field_of_Search>

- <Citations_made>

- <citations>

<citation>3091818 June 1963 Clark</citation>

</citations>

</Citations_made>

<Claims>I claim: 1. A door having intermediate the ends thereof a pair ...Description</Claims>

<Description>The present invention relates to doors particularly flexible traffic doors… </Description>

Once XML reports are generated, the MontyLingua NLP software is used to extract the SAO instances from the different sections of the patent text. Next, an algorithm is employed in the Patent SW Toolkit to extract component names by the component numbers that appear in the SAO instances that are extracted by MontyLingua from patent textual descriptions. MontyLingua provides an SAO report that includes the verb, subject, and objects that appear in each line of the patent text. It does not include subordinate clauses. As part of the subject and object extraction, each component from the patent can be extracted by looking for the component number and name in the subject and object in each SAO instance. This can be used to generate a list of components for each patent.

Once the component list is generated, list of functions performed by the components can be generated by extracting functions that appear in each SAO that includes a component name and number in the subject or object. The following provides an example of an SAO extraction from MontyLingua:

['comprise', 'door 1', 'clear plastics upper portion 2']

['form', 'which portion 2 and 3 are', 'at bumper 4']

['have', 'their close position', 'overlap 6']

['close', '', 'by magnet assembly 7']

['include', 'door 1', 'door frame 8']

['comprise', 'door frame 8', 'top steel frame 9', 'steel gusset 10']

['bolt', 'which', 'to upper and lower portion 2 and 3', 'by bolt']

['5', 'bolt']

['mount', 'door 1', 'on door jamb 12', 'by door shaft 13', 'to which shaft 13']

['bolt', 'frame 11', 'by bolt 14']

['reinforce', '', 'rubber seal 15']

['extend', 'rubber seal 15', 'between longitudinal edge', 'of side steel frame 11']

The terms shown in single quotes represent the [‘verb’, ‘subject’, ’object’, ‘object’] extracted from each sentence using MontyLingua.

The process of estimating a patent’s degree of ideality is conducted by first using software to automatically generate functional and physical models. Patent functional and physical models are generated with the use of patent text mining and information extraction techniques. This is accomplished by generating a list of components for each patent and modeling the functions performed by each patent component. A patent component is a physical feature, part of the system, or combination of physical parts that are used to satisfy a function fulfilled by the patented invention. Functional modeling is performed by extracting subject action object relationships for each component with the use of natural language processing techniques to perform part of speech tagging [64-66] of patent text. Patent

functions and components are used to estimate patent degree of ideality by dividing the number of functions by the number of components to determine a patent's benefit to cost ratio at each functional and physical hierarchical level.

The next section includes the approach estimating the level of invention of new designs.

4.1.2 TRIZ Level of Invention Estimation

To estimate the level of invention of new designs it was necessary to build a training data set based on inventions with know level of invention estimates. A matrix data set of 43 patents was used to train the neural network back propagation model using Matlab to estimate the level of invention for a set of 48,986 patents. 23% of the patents in the data represent level one inventions, 51% of the data represents level two, 14% level three, 7% level four and 5% for level five. To train the neural network the data set was broken up into a training set that consists of 80% of the data, a validation set representing 10% of the data and another 10% of the data to test the network performance. The data set is split into different data sets automatically. The training data set is used to train the neural network model, the validation set is used in Matlab to validate training and the test data set is used to test the results of the model. The network was built using 100 hidden neurons to train the inputs to the targets. The inputs to the network represent the number of citations made, citations received, patent generality, patent originality, mean forward citation time lag and mean backward citation time lag for each patent. The target data represents the estimated level of invention for each patent. The level of invention data was prepared by reviewing the data set manually to make an initial estimate of level of invention for each patent.

4.2 Implementation of Semantic Functional Basis of Design

A semantic functional basis of design was constructed by using natural language processing and latent semantic analysis to create a list of disciplinary terms for ten different disciplines. The ten disciplines used to create a semantic functional basis of design were chosen based on the patent categories in the NBER patent database. The disciplines include amusement devices, communications and computers, heating, mechanical, electrical, chemical, biotechnology, nuclear & X-rays, measuring and testing, and other. The next section includes an overview of the software implemented to complete this task. First NLP is performed on patent text using MontyLinqua to create a list of SAO instances. Next LSA is applied to the list of SAOs using the patent SW Toolkit to measure term frequencies and document frequencies for all terms. Finally, Matlab is used to rank terms based on the variance of the terms to the mean of the document frequency matrix.

4.3 Implementation of Transdisciplinary Knowledge Integration Measures in Software

This section includes an overview of the software architecture employed to perform NLP and LSA to extract textual information from patents to measure multiple levels of transdisciplinary knowledge integration of a new design or invention. In addition, the software architecture discussed in this section is employed to generate a list of physical and functional terms used to represent a given discipline. The list of physical and functional terms for different disciplines can be used in the concept generation phase by designers to help increase the success of their design concept by promoting the use of cross-disciplinary approaches, synthesize functions, create function-based representations, and develop function structures for new designs and inventions [104-108].

In this section a summary of the methods used to support the claims of this dissertation is provided. This includes an overview of data collection and preprocessing of patent data from the USPTO.gov website. In addition, the section discusses the conversion of USPTO.gov patent data from html format to XML format to facilitate extraction of patent functional descriptions. Furthermore, this section also includes an overview of NLP software modules employed by the Patent SW Toolset to extract action verbs, the subjects performing the action, and the objects of the action representing a system function from patent texts. Finally, the implementation of LSA is discussed to calculate term frequency metrics and to build a list of key physical and functional terms that represent a set of ten unique disciplines. Figure 4-1 includes an overview of the Patent SW Toolkit functional architecture.

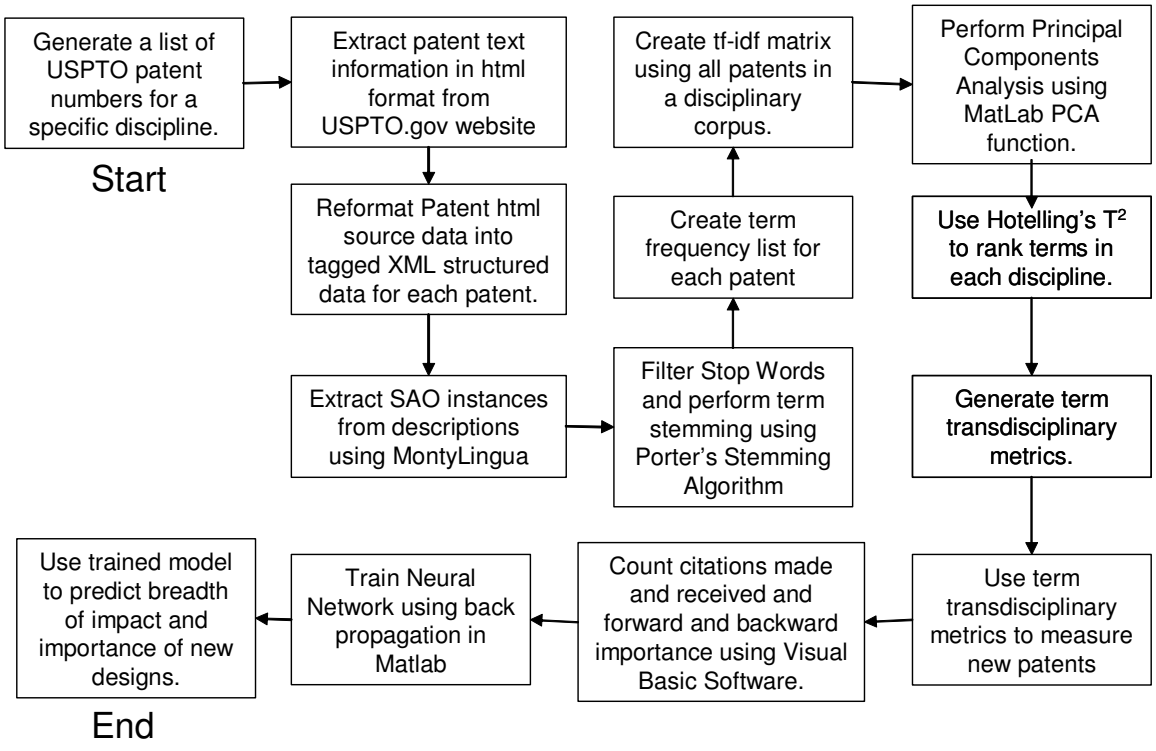


Figure 4-1 Patent SW Toolkit Functional Architecture

A graphical user interface was constructed to review each step in the patent data analysis process performed by the Patent SW Toolkit. The graphical user interface allows the user to select a list of patents by patent number to download raw patent data in HTML format. In addition, the interface enables the user to create XML and text files from the html data. Once the XML data is generated, the user interface enables the user to select the patent data files that will be analyzed by the MontyLingua NLP engine [62] to extract Subject Action Object instances and save them in list format in the form of text files. Once SAO text files are created, the user can then calculate function term frequencies from the list of SAOs in each patent. An example of the Patent SW Toolkit graphical user interface is provided in Figure 4-2.

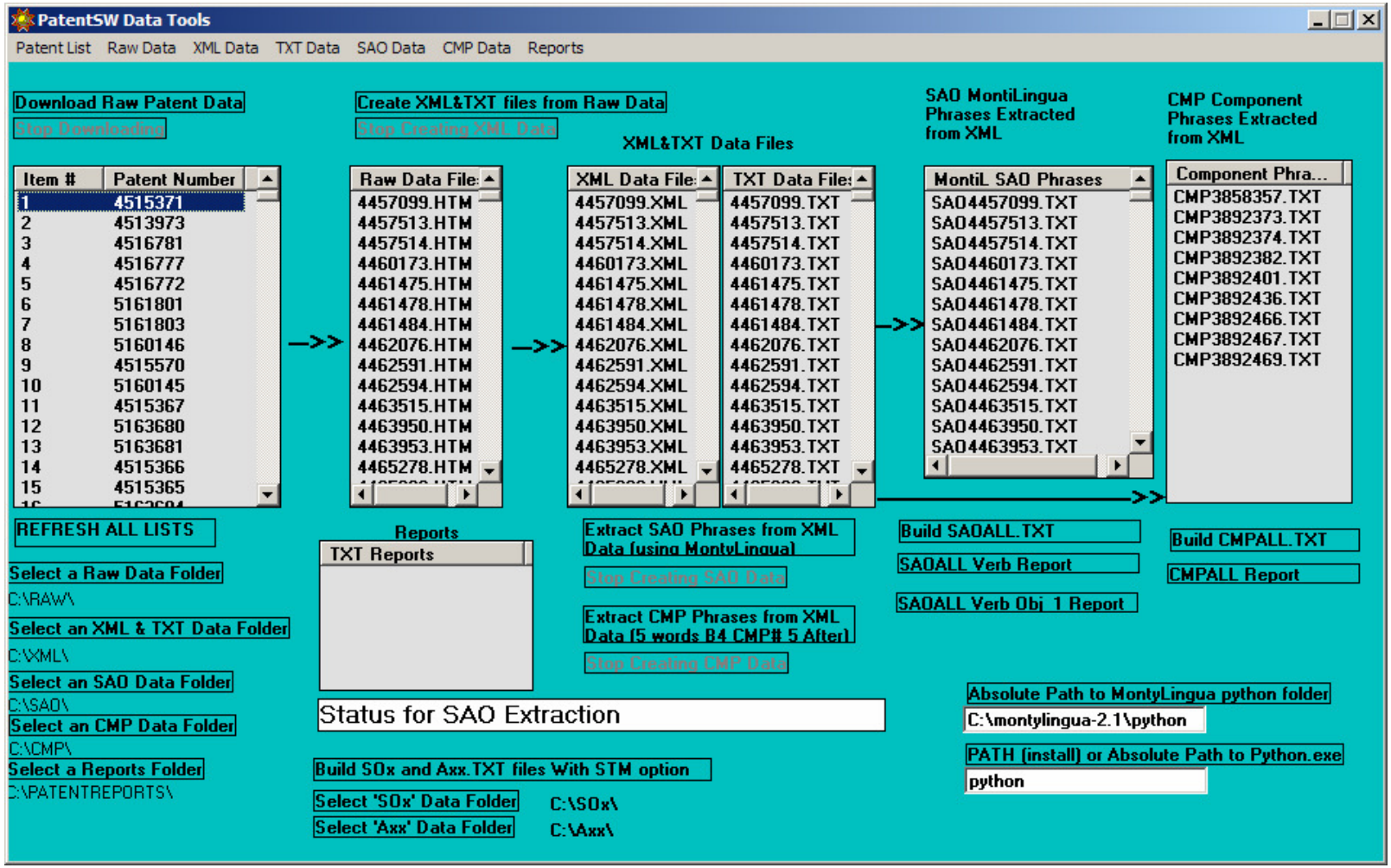


Figure 4-2 Patent SW Toolkit Graphical User Interface

Figure 4-2 includes a screen capture of the report windows provided by the Patent SW Toolkit used to review XML data files generated by the tool, patent files in text format, Subject Action Objects extracted using MontyLingua in list format, and a list of the function term frequencies extracted from the patent text.

4.3.1 Natural Language Processing of Patent Data

Natural language processing of patent data is accomplished using the MontyLingua NLP software written in Python [62]. The MontyLingua NLP software provides functionality to read a body of text. MontyLingua provides several types of text summarization and information extraction algorithms. The MontyLingua algorithm used to extract subject-action-objects from a sentence is the MontyLingua JIST function, which extracts the verb, subject, first object and second object from each sentence. Monty Lingua JIST function is applied to each sentence in each patent description. Only 300 patents were chosen from each discipline for analysis at this time because this data set will provide a good statistical sample set that can be expanded to all patents. Once the verb, subject, object-1, object-2 is extracted from the patent descriptions, it is possible to separate the action verbs, subjects, and objects from each SAO component phrase. The action verbs are the part of the sentence that represents the functions employed by the mechanical design described by each patent, and the subjects and objects represent the physical design components. Once the list of verbs is extracted, the list of verbs is then reduced by removing unwanted stop words from the list. In addition, term stemming is applied using Porter's stemming algorithm. [109] The following represents a subset of the stop words that are filtered from each set of patent terms:

a, about, above, across, after, afterwards, again,… .. be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, ….can, cannot, cant, co, con,… except, few, fifteen, fifty, fill, find, fire, first, five, for, former, formerly, forty, found, four, from,… put, rather, re, same, see, seem, seemed, seeming, seems, serious, several, she, should, show,… yet, you, your, yours, yourself, yourselves…

The next section will discuss software used to perform Latent Semantic Analysis using principal components analysis to develop a list of key disciplinary terms.

4.3.2 Latent Semantic Analysis Software

Software was created to generate a tf-idf matrix output in comma-separated value (csv) file format in order to easily import the software into Matlab to analyze the tf-idf matrix using principal component analysis. Figure 4-3 presents the tf-idf software graphical user interface. The tf-idf software allows the user to download patents in HTML format from the web. In addition, the software filters HTML tags and other non-value added fields from the text. The software then enables the user to construct a list of stop words to filter from the text and applies Porter's stemming algorithm. [109] Finally, functionality is provided for constructing the term frequency matrix and then calculating the inverse document frequency matrix. Once the tf-idf matrix is constructed, the Matlab principal component analysis function is used to analyze the matrix. The results are exported in .csv format to use in constructing the disciplinary term lists. The disciplinary term lists are then used to measure new inventions based on the functional and physical terms included in the description of the invention.

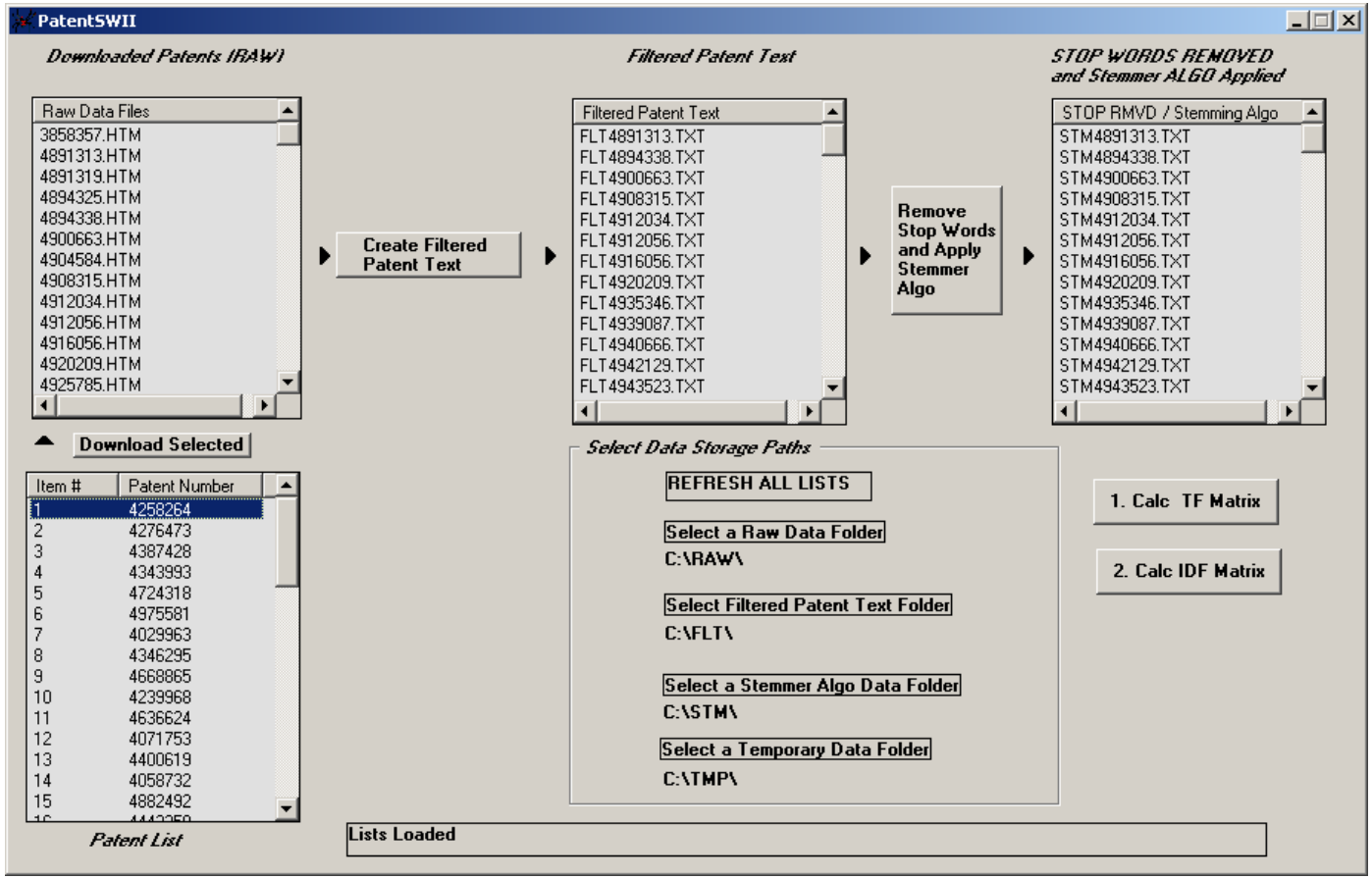


Figure 4-3 tf-idf software graphical user interface

Next, transdisciplinary metrics are employed to train a neural network model that is used to predict the breadth of impact and the importance of a new invention. The next section includes results of employing machine learning models to predict design success.

CHAPTER 5

RESULTS

5.1 Results of TRIZ Degree of Ideality and Level of Invention Estimation

The next section includes the results of estimating the degree of ideality for a set of patents. Two example patents are provided for discussion of the degree of ideality estimation results.

5.1.1 TRIZ Degree of Ideality Estimation

For patent number 3,858,357 the degree of ideality was estimated to quantify the patented design's value. First, a list of components was generated using the method discussed in the previous section. Next, a list of functions is assembled by identifying function terms that reveal actions performed by components that represent subjects in an SAO phrase or that act on components that represent objects in the same SAO phrase. If no subject or objects exist, then the SAO phrase is not considered in the list of components. Using this approach the following list of components was extracted for patent 3,858,357:

1 Doors, 2 upper portion, 3 lower portion, 4 bumpers, 5 bolts, 6 overlap, 7 magnet assembly, 8 door frame, 9 top steel frame, 10 steel gusset, 11 side steel frame, 12 door jamb, 13 door shaft, 14 bolts, 15 rubber seal, 16 header door jamb, 17 hinge, 18 bolts, 19 threaded

socket, 20 bolts, 21 threaded sockets, 22 V-cam follower, 24 recesses, 25 compression spring, 26 washer, 27 collar, 28 recess, 29 rubber

This list provides approximately 29 components for this design that represents a mount for a rubber bumper to be attached to a door. The sole function of the bumper mount is to allow the door to enable impact on the bumper. This function can be extracted from the patent text by listing all SAO phrases that include the door component name and then filtering out SAOs that do not include typical action verb terms. Once all action verbs are filtered by parsing the text and removing text other than the first term listed in single quotes it is possible to see that the only door function that remains is ['allow', 'door', 'for impact', 'on door', 'by goods'] which represents the component ‘door’ which allows for impact on door. (Note: Multiple objects are provided by Montylingua as part of the SAO phrase.) Other subfunctions can be found at lower subcomponent levels after SAO filtering. Functions found at subcomponent levels include thread bumper, form bumper, extend side steel frame, bolt door jamb, mount door shaft, attach V-cam follower, extend bolt, receive bolt, extend side steel frame, rise hinge, accommodate bolt, and compress spring. This represents a total of ~12 sub functions and one primary component function. By counting the subcomponents and subfunctions of the design, it is possible to estimate the degree of ideality of this system by taking the ratio of patent functions to components. This yields a ration of 13 functions to 29 components and a degree of ideality for this patented design of 45%.

This metric can be used to quantify designs that provide a high level of value within a specific patent technical category. It can be used as part of the concept generation process to review designs that have a high degree of ideality and use these designs as a benchmark early in the design process. The next section will discuss using computers to estimate the level of invention for a patented design.

5.1.2 TRIZ Level of Invention Estimation

Figure 5-1 provides the regression performance of the training data to the neural network.

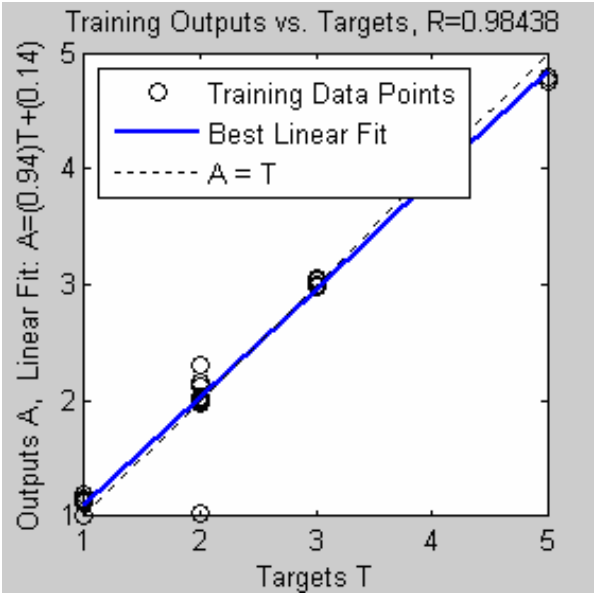


Figure 5-1 Neural Network Performance Data

The figure portrays how the outputs of the neural network model, shown on the Y axis in the Figure 2 fit the target level of invention data shown on the X axis. The network fits the data well with a regression coefficient of 98.4%. This shows the network performs well

based on the training data input against the level of invention target data. The neural network is then saved to the workspace to use to estimate the level of invention for a new set of patents. This new set of patents was selected using the NBER patent database by first gathering patents from the category of mechanical designs. These patents were then narrowed to patents that have a minimum of one citation made, one citation received, a measure of patent generality and patent originality in the NBER database. The input data was then run through the neural network to classify 48,986 patents into the five levels of invention. Figure 5-2 provides a summary of the initial network training results from applying the network to 48,986 patents. In addition, statistics from classical TRIZ research are included for comparison to the level of invention estimates.

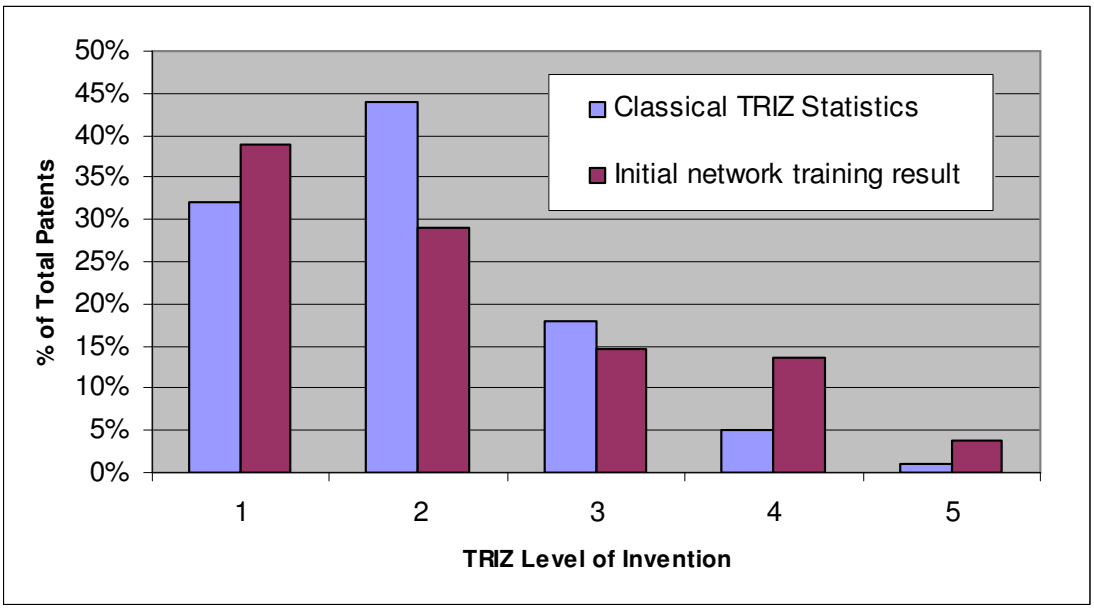


Figure 5-2 TRIZ Level of Invention Statistics

The initial network results show more patents classified as level one than the classical TRIZ research. In addition, the estimate shows fewer patents classified as level two, close to the same number of patents for level three and far more patents estimated as level four and five. The neural network classification results fit more closely with the distribution of levels of invention described in [18]. However, the percentage of level one invention is much higher and the percentage of level three through five inventions is much lower. The number of level three through five inventions in the distribution of level of invention data is <4% for the top three level of invention classifications. The level five patents predicted by the network were reviewed to determine estimate accuracy. One of the level five patents predicted by the network includes patent number 4,863,655 and is titled "Biodegradable packaging material and the method of preparation thereof". This patent fits the criteria of a level four to five invention given that it represents the use of a new scientific discovery. The discovery in this patent is the use of materials for packaging materials that will degrade in the environment to prevent future environmental pollution. The neural network predicted this patent as a level five invention since it has a high citation received to citation made ratio and has a large mean forward citation time lag. This reflects that the patent has an effect over a wide range of future inventions.

Another patent identified by the neural network as a level five invention is patent number 5,232,243 titled "Occupant sensing apparatus." This patent describes a novel method for determining when an occupant has entered a vehicle and then adjusting the seat restraint based on the size of the occupant. This involves using a material containing an electrical

characteristic that adjusts based on the size of the individual. This patent is closer to a level four invention than a level five invention because it lies outside of the existing paradigm of mechanical design by using a material to sense an occupant. The use of the material that contains an electrical characteristic identifies that a tool from science is used. This meets the definition of a level four invention. This patent is not a level five invention since it does not result in a new scientific discovery. The next section will discuss the approach taken to develop a semantic functional basis of design.

5.2 Developing a Semantic Functional Basis of Design

Some of the function terms in the secondary and tertiary categories appeared much more frequently than those in the class (primary) category of functions. Therefore, it is proposed by this research that the class (primary) function term "Branch" be replaced by the tertiary term "remove" since remove is the 19th most used verb in the list of extracted verbs. In addition, it is very unlikely that the term "Branch" will be used as a major function by mechanical designs since it appears in less than 0.01% of patented designs. The analysis also suggests that class function terms provision, channel and signal be replaced with the terms supply, transmit and indicate, respectively. This is suggested since the later terms supply, transmit, and indicate are used more widely by patented mechanical designs and provide a better semantic functional basis that can be used to extract functionality from textual design descriptions. Finally, the analysis also suggests that a new class of function be added, rotate. Rotate is used much more frequently than other terms under the transmit class and represents a different function than transmitting an object.

Table 5-1 and Table 5-2 include proposed changes as a result of analyzing the frequency of occurrence of functional terms within patent descriptions for mechanical designs.

Table 5-1 Proposed Changes to Functional Basis of Mechanical Designs

CLASS		Secondary		Tertiary	
Term	**Frequency**	**Term**	**Frequency**	**Term**	**Frequency**
Remove	0.71%	Separate	0.20%		
				Divide	0.06%
				Extract	0.04%
				Branch	0.01%
		Distribute	0.05%		
Transmit	0.35%	Import	0.00%		
		Export	0.00%		
		Transfer	0.18%		
				Transport	0.04%
				Channel	0.04%
Rotate	0.73%				
				Translate	0.05%
				Guide	0.08%
				allow	0.37%
Connect	1.41%	Couple	0.33%		
				Join	0.09%
				Link	0.03%
		Mix	0.11%		
Control	0.92%	Actuate	0.16%		
Magnitude		Regulate	0.06%		
				Increase	0.44%
				Decrease	0.10%
		Change	0.24%		
				Increment	0.01%
				Decrement	0.00%
				Shape	0.15%
				Condition	0.03%
		Stop	0.18%		
				Prevent	0.55%
				Inhibit	0.04%

Table 5-2 Proposed Changes to Functional Basis of Mechanical Designs

CLASS		Secondary		Tertiary	
Term	**Frequency**	**Term**	**Frequency**	**Term**	**Frequency**
Convert	0.12%	Convert	0.12%		
Supply	0.30%	Store	0.25%		
				Contain	0.51%
				Collect	0.05%
				Provision	0.00%
Indicate	0.84%	Sense	0.16%		
				Detect	0.15%
				Measure	0.15%
				Signal	0.05%
				Track	0.10%
				Display	0.04%
		Process	0.13%		
Position	1.01%	Stabilize	0.05%		
		Secure	0.66%		
		Support	0.70%		

The analysis suggests that the terms representing the class (primary) set of functional basis terms should be adjusted to represent more highly used functional terms to form a semantic functional basis. As a result of the analysis it was found that many other functional terms occurred more frequently than terms listed in the functional basis of mechanical designs. Therefore, it may be necessary to revisit the list even further to determine if the secondary and tertiary list of functions and corresponding functions provided by Stone and Wood should be augmented. Some of the highly used functional terms not listed as part of functional basis of design Class, secondary, tertiary or correspondent terms include the terms

extend, drive, receive, and cut. These functional terms should be considered as potential correspondents within a semantic functional basis of design and possibly considered as secondary or tertiary functions. The next section includes analysis performed to create a functional basis for the "coded data generation or conversion" class of patents. This process can be used on other classes of patents to develop a functional basis.

5.2.1 Developing Semantic Functional Basis for all design Classes

A process has been developed to aid the selection of class, secondary and tertiary terms for a patent class with the intent of generating a semantic functional basis for other patent classes. The coded data generation or conversion patent class from the USPTO was used as an example class to demonstrate the process. The first step in the process of generating a semantic functional basis for patent classes is to develop a matrix X that consists of the term frequency of a patent in a specific class as a column vector, and the occurrence of each functional term in the Pahl & Beitz [31], Hundal [110] and TRIZ [1, 13] functional basis as either a class, secondary, or tertiary functional term.

Table 5-3 Matrix for Functional Basis Classification

	Matrix X				Output Y
Function	**Pahl & Beitz**	**Hundal**	**TRIZ**	**Term Frequency**	**FB**
Branch	0	1	1	0.01%	1
Channel	1	1	0	0.04%	1
Connect	1	1	0	1.41%	1
Control	0	0	1	0.92%	1
Convert	0	0	0	0.12%	1
Provision	0	0	0	0.00%	1
Signal	0	0	0	0.05%	1
Support	0	0	0	0.70%	1
Separate	0	2	1	0.20%	2
Distribute	0	0	0	0.05%	2
Import	0	0	0	0.00%	2
Export	0	0	0	0.00%	2
Transfer	0	0	1	0.18%	2
Guide	0	0	0	0.08%	2
Couple	0	0	0	0.33%	2
Mix	0	2	0	0.11%	2
Actuate	0	0	0	0.16%	2
Regulate	0	0	0	0.06%	2
Change	1	2	1	0.24%	2
Stop	0	2	0	0.18%	2
Convert	0	2	0	0.12%	2
Store	1	1	0	0.25%	2
Supply	0	1	0	0.30%	2
Sense	0	2	0	0.16%	2
Indicate	0	0	0	0.84%	2
Process	0	2	0	0.13%	2
Stabilize	0	0	1	0.05%	2
Secure	0	0	0	0.66%	2
Position	0	0	0	1.01%	2
Divide	0	2	0	0.06%	3
Extract	0	0	0	0.04%	3
Remove	0	0	0	0.71%	3
Transport	0	2	0	0.04%	3
Transmit	0	2	0	0.35%	3
Translate	0	0	0	0.05%	3
Rotate	0	0	0	0.73%	3
Allow DOF	0	0	0	0.00%	3
Join	0	0	0	0.09%	3
Link	0	0	0	0.03%	3

This is used to set up a matrix of inputs as shown in Table 5-3. The numbers from 0 to 3 represent the occurrence of function as either not used (0), a class function (1), a secondary function (2), or a Tertiary function (3). The Matrix X is used to create multiple linear regression coefficients used to classify each new functional term in the coded data generation or conversion patent class as either a class, secondary or tertiary function. Table 5-4 includes results of the functional basis classification. Functional terms that have a high frequency of occurrence and predicted with a functional basis (FB) value of 0, 1, or 2 may be selected in the class (primary) group for functional basis classification. Functional terms that appear with a two will be classified as secondary functions and Functional terms that have a two or three will be classified as tertiary functions. This method is proposed as process to aid the synthesis of a semantic functional basis of design for other patent classes.

A method developed to create a functional synthesis transdisciplinary metric includes using natural language processing and latent semantic analysis to develop a list of physical and functional terms for different disciplines. This list of terms can be used to measure the number of functions synthesized in a new invention. The list of functional and physical terms was developed using the method discussed in the previous sections. Table 3-4 contains an example list of disciplinary terms for 10 different disciplines. This list of terms can then be used to measure the term frequency for each of the functional terms that reside in each of these disciplinary lists. This can then be used to measure the overall disciplinary function term frequency of a new design and measure the level of functional synthesis. The functional

synthesis transdisciplinary metric can also be used to measure the success of new designs by predicting the breadth of impact and market success based on this measure.

Table 5-4 Exemplary Functional Basis Classification

Function:	Frequency:	FB
connect	4.42%	0
change	0.53%	1
store	0.92%	1
control	0.91%	2
channel	0.04%	2
indicate	1.35%	2
transfer	0.31%	2
separate	0.11%	2
supply	1.11%	2
stabilize	0.04%	2
branch	0.01%	2
convert	1.01%	2
couple	0.99%	2
transmit	0.67%	2
contain	0.55%	2
increase	0.52%	2
actuate	0.37%	2
inhibit	0.32%	2
measure	0.32%	2
detect	0.29%	2
divide	0.24%	2
signal	0.27%	2
sense	0.22%	2

Function:	Frequency:	FB
position	0.23%	2
prevent	0.23%	2
track	0.19%	2
display	0.14%	2
decrease	0.16%	2
translate	0.14%	2
stop	0.11%	2
rotate	0.13%	2
process	0.10%	2
remove	0.13%	2
condition	0.11%	2
mix	0.02%	3
support	0.05%	3
shape	0.04%	3
regulate	0.04%	3
secure	0.04%	3
distribute	0.04%	3
transport	0.00%	3
link	0.02%	3
increment	0.02%	3
extract	0.02%	3
join	0.02%	3
collect	0.01%	3

5.3 Developing a Prediction Model Using Machine Learning

A set of thirteen hundred patent documents was selected at random to test the transdisciplinary knowledge integration measures. A machine-learning model was developed to test the transdisciplinary measures to identify whether the measures serve as good predictors of design success. The machine-learning model constructed to test the transdisciplinary knowledge integration measures is an artificial neural network back-propagation model. To build the machine-learning model, a set of training data was gathered using data from the NBER patent database and USPTO patent database. NBER patent data includes variables defined by Jaffe and Trajtenberg in [20], such as citations made to other patents, the level of originality of a patent, the level of backward patent importance calculated using first and second generation patent citations and the set of transdiscplinary knowledge integration measures. These variables represent independent variables used to predict the dependent variables: number of citations received, forward patent importance, and the level of patent generality, also defined by Jaffe and Trajtenberg in [20]. These three dependent variables are used as part of the training data to build prediction models to test the capabilities of the transdisciplinary metrics. Table 5-5 includes training data used to train the artificial neural network model.

Table 5-5 Artificial Neural Network Training Data

Independent Variables																									Dependent Variables		
cmade	IMPORTB	original	TI	Td	Tmech	Telec	Tchem	TBIO	Tother	Tamuse	Theat	Tcomm	Tnuc	Tmeas	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	general	creceive	IMPORTF
3	9.5	0.44	0.54	0.17	0.02	0.04	0.00	0.02	0.00	0.02	0.65	0.08	0.18	0.00	0.04	0.01	0.09	0.01	0.05	0.08	0.15	0.05	0.12	0.41	0.69	12	85.5
4	7.5	0.63	0.50	0.16	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.07	0.01	0.07	0.19	0.63	0.00	3	18.0
10	19.5	0.69	0.13	0.17	0.00	0.00	0.02	0.93	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.08	0.01	0.03	0.04	0.02	0.06	0.10	0.61	0.67	3	9.5
4	10	0.38	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.01	0.00	0.02	0.02	0.02	0.14	0.14	0.14	0.50	0.26	14	110.0
18	63.5	0.73	0.00	0.16	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.04	0.05	0.11	0.22	0.55	0.44	3	25.0
11	54	0.22	0.50	0.16	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.01	0.02	0.02	0.01	0.01	0.05	0.10	0.15	0.20	0.44	0.79	9	26.5
6	24	0.78	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.02	0.02	0.01	0.08	0.17	0.10	0.59	0.67	3	6.5
9	27.5	0.81	0.52	0.16	0.00	0.00	0.21	0.64	0.00	0.00	0.14	0.00	0.00	0.00	0.01	0.01	0.00	0.08	0.02	0.04	0.05	0.19	0.11	0.49	0.63	4	16.0
11	28.5	0.63	0.00	0.16	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.17	0.01	0.01	0.03	0.08	0.08	0.04	0.55	0.50	2	16.5
6	26	0.28	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.10	0.14	0.06	0.08	0.56	0.00	1	3.0
8	18	0.32	0.00	0.14	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.01	0.00	0.03	0.14	0.14	0.14	0.50	0.24	7	25.0
9	25.5	0.67	0.62	0.18	0.00	0.57	0.10	0.00	0.19	0.00	0.10	0.05	0.00	0.00	0.02	0.00	0.04	0.01	0.02	0.03	0.02	0.11	0.18	0.58	0.68	22	214.5
8	39.5	0.38	0.60	0.17	0.00	0.00	0.11	0.06	0.06	0.06	0.00	0.06	0.06	0.00	0.03	0.01	0.02	0.03	0.01	0.09	0.03	0.22	0.13	0.43	0.54	42	289.5
9	20.5	0.24	0.00	0.17	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.03	0.02	0.06	0.02	0.15	0.13	0.52	0.31	11	99.5
3	11.5	0.44	0.66	0.18	0.10	0.40	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.03	0.04	0.04	0.01	0.05	0.04	0.07	0.05	0.19	0.47	0.68	18	154.5
13	43	0.49	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.05	0.05	0.11	0.11	0.10	0.56	0.58	89	206.0
10	34	0.59	0.63	0.17	0.00	0.00	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.02	0.02	0.01	0.01	0.03	0.05	0.05	0.13	0.08	0.61	0.47	78	645.0
6	19	0.50	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	0.03	0.04	0.01	0.01	0.05	0.05	0.06	0.20	0.53	0.31	11	325.0
6	18	0.28	0.76	0.17	0.03	0.08	0.08	0.05	0.03	0.00	0.30	0.35	0.00	0.00	0.04	0.02	0.07	0.03	0.03	0.06	0.05	0.04	0.20	0.47	0.12	16	64.0
5	18	0.56	0.71	0.17	0.00	0.00	0.18	0.00	0.18	0.09	0.45	0.00	0.00	0.00	0.02	0.03	0.06	0.02	0.04	0.05	0.01	0.10	0.21	0.47	0.67	6	23.5
10	34.5	0.66	0.80	0.17	0.11	0.05	0.00	0.37	0.05	0.11	0.11	0.00	0.11	0.00	0.03	0.01	0.02	0.03	0.03	0.04	0.03	0.11	0.18	0.53	0.56	23	322.5
6	14	0.50	0.45	0.17	0.00	0.73	0.00	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.03	0.01	0.03	0.02	0.03	0.08	0.09	0.08	0.03	0.61	0.78	12	158.5
6	24	0.28	0.54	0.16	0.50	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.03	0.10	0.10	0.19	0.54	0.45	11	476.5
11	34	0.59	0.00	0.15	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.04	0.00	0.02	0.04	0.10	0.15	0.10	0.51	0.78	8	59.5
3	8.5	0.44	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.09	0.14	0.22	0.49	0.00	2	19.5
9	30	0.53	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.02	0.00	0.00	0.00	0.00	0.08	0.03	0.09	0.16	0.62	0.79	24	184.5
11	31.5	0.22	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.12	0.71	0.41	8	56.0
15	33	0.38	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.01	0.01	0.16	0.03	0.17	0.08	0.48	0.00	8	108.0
14	44	0.67	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.09	0.08	0.11	0.10	0.57	0.64	5	18.5
4	8.5	0.50	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.17	0.09	0.16	0.49	0.00	1	27.0
14	33	0.49	0.22	0.15	0.00	0.00	0.00	0.13	0.00	0.87	0.00	0.00	0.00	0.00	0.03	0.03	0.01	0.00	0.01	0.04	0.03	0.07	0.09	0.69	0.63	4	51.5
11	18.5	0.67	0.48	0.15	0.00	0.00	0.60	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.01	0.02	0.03	0.02	0.01	0.01	0.07	0.10	0.06	0.67	0.47	40	372.0
9	37.5	0.37	0.00	0.15	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.04	0.02	0.03	0.23	0.08	0.55	0.43	22	246.0
13	34	0.46	0.32	0.17	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.20	0.00	0.00	0.01	0.05	0.01	0.02	0.03	0.07	0.05	0.10	0.11	0.55	0.67	3	6.0
18	54	0.69	0.00	0.18	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.12	0.02	0.01	0.05	0.06	0.06	0.10	0.54	0.61	6	67.5
9	36.5	0.72	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.03	0.01	0.03	0.09	0.10	0.71	0.00	1	8.0
6	12.5	0.38	0.00	0.15	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.04	0.07	0.07	0.78	0.50	2	45.0

The machine learning technique used to train the model in this example is the artificial neural network back-propagation algorithm supplied in Matlab Neural Network Toolkit. The neural network is used to train a model using an expanded set of training data, similar to the example training data shown in Table 5-5 Artificial Neural Network Training Data. To train the artificial neural network, the data set was broken up into a training set that consists of 90% of the data, a validation set representing 5% of the data and another 5% of the data to test the network performance. The network was built using 50 hidden neurons to train the inputs to the targets to improve the training and test data correlation coefficients. The inputs to the network represent the number of citations made, patent backward importance, patent originality, and 22 transdisciplinary measures for each patent. The target data represents the level of generality, the number of citations received, and the forward importance for new inventions.

The measure of originality is calculated using the following equation [20]:

$$o_i = 1 - \sum_{k=1}^{n_i} \left(\frac{b_{ik}}{b_i}\right)^2$$

where i is the patent under consideration, b is the number of patents cited and k indicates the subclass of the CITED patent as indicated in the NBER database.

The measure of generality is calculated using the following equation [20]:

$$g_i = 1 - \sum_{k=1}^{n_i} \left(\frac{f_{ik}}{f_i}\right)^2$$

where f is the number of patents citing patent i and k indicates the subclass of the cited patent as indicated in the NBER database.

The measure of forward and backward patent importance is measured using the following equations:

$$IF_i = Cr_i + \lambda \sum_{j=1}^{Cr_i} Cr_{i+1,j}$$

$$IB_i = Cm_i + \lambda \sum_{j=1}^{Cm_i} Cm_{i+1,j}$$

Where IF is the forward patent importance, Cr is the number of citations received by other patents, IB is the backward patent importance, Cm is the number of citations made to other patents by the patent, i is number of first generation citations, j is the number of second generation citations and $\lambda = 0.5$ is a discount factor provided by Jaffe and Trajtenberg that is used to reduce the contribution of 2nd generation patents in the measurement. [20]

Figure 5-3 presents neural network performance data for predicting the level of generality, number of citations received and forward patent importance. In addition, more detailed results are provided in Appendix B. The predictive model trained to predict design generality fit the test data set with a correlation coefficient of 0.51. In addition, the citations received model fit the test data with a correlation coefficient of 0.33. The forward importance prediction model fit the test data with a correlation coefficient of 0.37. This verifies that the transdisciplinary measures have the potential to predict the impact an invention will have on future inventions based on the functionality employed across multiple disciplines.

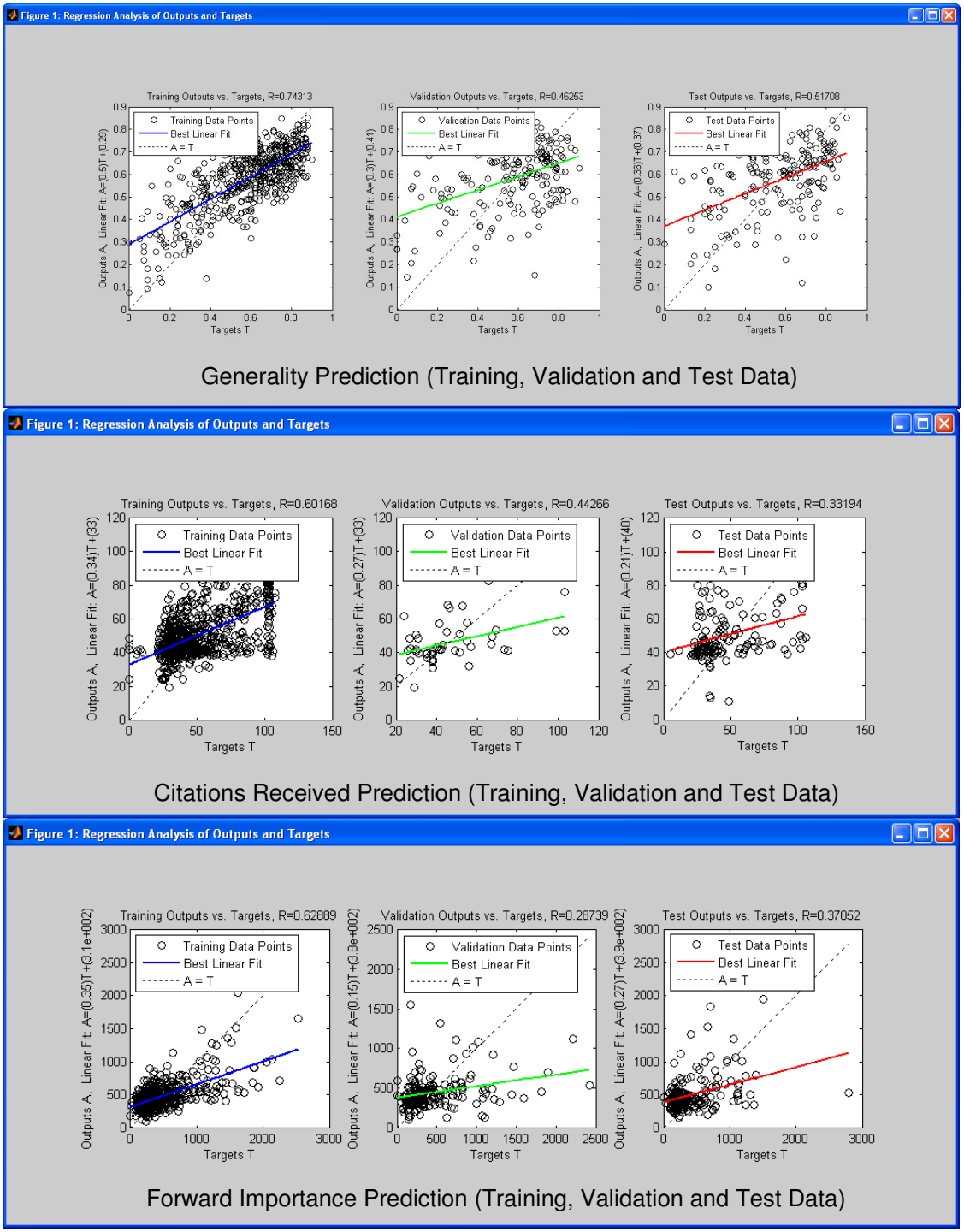


Figure 5-3 Final Neural Network Prediction Results

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

6.1 TRIZ Metric Estimation

Future research will investigate methods for improving the performance of the machine learning model. One method to improve the neural network performance includes increasing the training data set by manually estimating the level of invention for more patents or getting standard/published data from TRIZ researchers and increasing the number of independent variables used in the training data set. Another approach for increasing neural network performance is to base the neural network on functional and physical terms using latent semantic analysis to identify key design terms that represent each of the patents within the field of mechanical design. This method could provide a predictive measure that indicates the level of invention for more recent patents that have yet received many citations. Other areas to increase network performance may include using term frequency metrics based on the use of natural language processing and latent semantic analysis. Latent semantic analysis could be used to identify key terms that when used result in a higher level of invention. Furthermore, functions used in the estimation of degree of ideality could be separated in to useful and harmful functions to increase the accuracy of the degree of ideality estimation. Finally, machine learning models could be created to determine the design position on TRIZ laws/trends of technology evolutions.

6.1.1 Use of TRIZ Metrics to Support Design Innovation

TRIZ metrics such as degree of ideality and level of invention can be used early in the design process to support concept generation, functional modeling, and functional synthesis. A wide range of patents that span across multiple disciplines can be analyzed to see how new design components evolve over time to perform existing system functions or create new functions by integrating cross domain knowledge. Patent citation networks are used to identify patents that initiate technological discontinuities by using a new set of components

to perform common functions. These trends of evolution can be examined using metrics such as degree of ideality and level of invention. In addition, new metrics can be created by employing text mining techniques to extract a set of key words that represent a given discipline. These key words identify how components from multiple disciplinary fields can be integrated to develop new technologies to perform existing system functions. [17] Data mining techniques can be used to generate TRIZ metrics for a large number of patents. Machine learning techniques can then be used to train neural network, or other machine learning, models to uncover evolutionary trends that reside in patent data. These trends can be used in the concept generation process to review design trends that lead to innovative design concepts. As part of future research, the use of machine learning techniques and approaches will be used to train prediction models that can help in the innovation process by predicting future success of new designs.

6.2 Semantic Functional Basis of Design

This dissertation includes an approach to classify functional terms to aid in the process of creating a semantic functional basis for multiple disciplines. A semantic functional basis of design can be used in the concept generation phase to develop a metric for design success based on determining the integration of functionality across many domains. For future research the authors suggest that it is possible to classify patents into clusters using a semantic functional basis of design centered on Stone and Woods function and flow definitions. This classification can then be used as a metric during the concept generation phase to understand design concepts from other domains that result in successful design

problem solutions. In addition, WordNet can be used to identify hypernyms and hyponyms for functional concepts. Finally, Wordnet could be used to create different levels of functional basis based on the relationships between different functional concepts.

6.3 Transdisciplinary Knowledge Integration Measures

The results of this research indicate that it is possible to employ a set of transdisciplinary metrics to serve as inputs for predicting the success of new inventions. This dissertation has put forth a methodology for use in generating a set of transdisciplinary knowledge integration measures based on the use of natural language processing and latent semantic analysis techniques. In addition, a machine-learning model has been developed using this methodology that will help in the evaluation of new inventions.

As part of future research the authors suggest that hierarchical relationships of knowledge between disciplines be studied to understand the difference between lower- and higher-level knowledge transfer. Furthermore, a new set of transdisciplinary metrics should be developed that includes the integration of physical design components with functions to create a set of technology-based term constructs that can be used to measure the transdisciplinarity and interdisciplinarity of new designs based on existing technological systems. Technological systems would then be used to measure invention transdisciplinarity instead of just the physical and functional items that make up these systems. In addition, other machine learning techniques can be explored further to improve the accuracy of this method.

Further research will also investigate using an artificial neural network to predict innovative potential. Approaches for developing an innovation potential model include the following:

First, use functions and objects separately by creating a list of functional and physical transdisciplinary measures. This is accomplished by separating subjects and objects from action verbs. Another method includes using Google PageRank as a measure of importance. In addition, machine learning techniques could be used to train prediction models containing other engineering design theory metrics. This includes using dominant design, technological discontinuities as targets in a machine learning model. Finally, a series of patents could be modeled. Functions and solutions for a series of patents are determined and then compared along the series to see changes. This method could also be used to identify how TRIZ contradictions are solved. All of these areas should be explored as part of future research to develop a method to predict innovation potential.

REFERENCES

[1] V. Fey and E. Rivin, *Innovation on demand: New product development using triz*: Cambridge University Press, 2005.

[2] A. Afuah, *Innovation management strategies, implementations and profits*, 1998.

[3] C. M. Christensen, *The innovator's solution: Creating and sustaining successful growth*, 2003.

[4] J. Hirtz, R. B. Stone, D. A. McAdams, S. Szykman, and K. L. Wood, "A functional basis for engineering design: Reconciling and evolving previous efforts," *Research in Engineering Design*, vol. 13, pp. 65-82, 2002.

[5] R. B. Stone and K. L. Wood, "Development of a functional basis for design," *Journal of Mechanical Design*, vol. 122, pp. 359-370, 2000.

[6] K. Kaikhah and S. Doddameti, "Knowledge discovery using neural networks," *Lecture Notes in Artificial Intelligence*, vol. 3029, pp. 20-28, 2004.

[7] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, 2nd Edition ed: Elsevier Academic Press, 2003.

[8] T. M. Mitchell, *Machine learning*: McGraw-Hill, 1997.

[9] E. Horvitz, J. Apacible, R. Sarin, and L. Liao, "Prediction, expectation, and surprise: Methods, designs and study of a deployed traffic forecasting service," 2007.

[10] T. Finin, J. Gama, R. Grossman, D. Lambert, H. Liu, K. Liu, O. Nasraoui, L. Singh, J. Srivastava, and W. Wang, "National science foundation symposium on next generation of data mining and cyber-enabled discovery for innovation (ngdm'07): Final report," 2007.

[11] D. Tate, A. Ertas, M. M. Tank, and T. T. Maxwell, "Foundations for a transdisciplinary approach to engineering systems research based on design & process," *The ATLAS Module Series On Transdisciplinary Science & Engineering*, vol. 2, pp. 1-37, 2006.

[12] G. H. Hadorn, H. Hoffmann-Riem, S. Biber-Klemm, W. Grossenbacher-Mansuy, D. Joye, C. Pohl, U. Wiesmann, and E. Zemp, "*Idea of the handbook*," in *Handbook of transdisciplinary research*, H. H.-R. Gertrude Hirsch Hadorn, Susette Biber-Klemm, Walter Grossenbacher-Mansuy, Dominique Joye, Christian Pohl, Urs Wiesmann, and Elisabeth Zemp, Ed.: Springer, 2007.

[13] G. Altshuller, *Creativity as an exact science*: CRC Press, 1994.

[14] M. G. Moehrle, "What is triz? From conceptual basics to a framework for research," *Creativity and Innovation Management*, vol. 14, pp. 3, 2005.

[15] G. Cascini and P. Rissone, "Plastics design: Integrating triz creativity and semantic knowledge portals," *Journal of Engineering Design*, vol. 15, pp. 405-424, 2004.

[16] R. Chun, "Innovation and reputation: An ethical character perspective," *Creativity and innovation management*, vol. 15, pp. 63-73, 2006.

[17] C. Herstatt and E. v. Hippel, "Developing new product concepts via the lead user method: A case study in a "Low tech" Field"," *Journal of Product Innovation Management*, vol. 9, pp. 213-221, 1991.

[18] E. v. Hippel, " Democratizing innovation," 2005.

[19] F. F. Suarez and J. M. Utterback, "Dominant designs and the survival of firms," *Strategic Management Journal*, vol. 16, pp. 425-430, 1995.

[20] A. B. Jaffe and M. Trajtenberg, *Patents, citations & innovations a window on the knowledge economy*: MIT Press, 2002.
[21] T. T. M. Murat M. Tanik Atila Ertas, " Transdisciplinary engineering education and research model," *Transactions of the Society for Design and Process Science*, vol. 4, pp. 1-11, 2000.
[22] B. Gumus, "The transdisciplinary product development lifecycle model," *Journal of Engineering Design*, vol. 19, pp. 185, 2008.
[23] N. P. Suh, *Axiomatic design: Advances and applications*: New York:Oxford University Press, 2001.
[24] N. P. Suh, *Complexity theory and applications*: Oxford University Press, 2005.
[25] M. Bordons, F. Morillo, and I. Gomez, "Analysis of cross-disciplinary research through bibliometric tools," in *Handbook of quantitative science and technology research*: Kluwer Academic Publishers, 2004.
[26] S. Breschi and F. Lissoni, "Knowledge networks from patent data: Methodological issues and research targets," in *Handbook of quantitative science and technology research*: Kluwer Academic Publishers, pp. 613-643, 2004.
[27] S. Brusoni and A. Geuna, "Specialisation and integration: Combining patents and publications data to map the 'structure' of specialised knowledge," in *Handbook of quantitative science and technology research*: Kluwer Academic Publishers, pp. 733-758, 2004.
[28] B. N. Sampat and A. A. Ziedonis, "Patent citations and the economic value of patents: A preliminary assessment," in *Handbook of quantitative science and technology research*: Kluwer Academic Pulishers, pp. 277-298, 2004.
[29] J. M. Utterback and F. F. Suarez, "Patterns of industrial evolution, dominant designs and firms' survival," *Research on Technological Innovation, Management and Policy*, vol. 5, pp. 47-87, 1993.
[30] X. Li, H. Chen, Z. Huang, and M. C. Roco, "Patent citation network in nanotechnology (1976–2004)," *Journal of Nanoparticle Research*, vol. 9, 2007.
[31] G. Pahl and W. Beitz, *Engineering design a systematic approach*: Springer-Verlag, 2007.
[32] S. Szykman, J. W. Racz, and R. D. Sriram, "The representation of function in computer-based design," in *Proceedings of the 1999 ASME Design Engineering Technical Conferences (11th International Conference on Design Theory and Methodology),*, Las Vegas, NV, 1999a.
[33] C. Kirschman and G. Fadel, "Classifying functions for mechanical design," *Journal of Mechanical Design*, vol. 120, pp. 475-482, 1998.
[34] I. Rowlands, "Journal diffusion factors: A new approach to measuring research influence," *Aslib Proceedings*, vol. 54, pp. 77-84, 2002.
[35] T. F. Frandsen, R. Rousseau, and I. Rowlands, "Diffusion factors," *Journal of Documentation*, vol. 62, pp. 58-72, 2006.
[36] M. Last, P. S. Szczepaniak, Z. Volkovich, and A. Kandel, *Advances in web intelligence and data mining*: Springer, 2006.
[37] A. L. Porter, "How "Tech mining" Can enhance r&d management," *Research Technology Management*, vol. 50, pp. 15, 2007.
[38] S. Sumathi and S. N. Sivanandam, *Introduction to data mining and its applications*, vol. 29: Springer.
[39] T. B. Ho, S. Kawasaki, and J. Granat, *Knowledge acquisition by machine learning and data mining*, vol. 59, 2007.

[40] Y.-H. Tseng, H.-J. Lin, and Y.-I. Lin, *Text mining techniques for patent analysis*, vol. 43, 2007.
[41] R. Srinivasan, G. L. Lilien, and A. Rangaswamy, "The emergence of dominant designs," *Journal of Marketing*, vol. 70, pp. 1-17, 2006.
[42] N. R. Pal and L. Jain, *Advanced techniques in knowledge discovery and data mining*: Springer, 2004.
[43] D. J. C. MacKay, *Information theory, inference, and learning algorithms*: Cambridge University Press, 2005.
[44] D. Zhu and A. L. Porter, "Automated extraction and visualization of information for technological intelligence and forecasting," *Technological Forecasting and Social Change*, vol. 69, pp. 495-506, 2002.
[45] G. Rueda, T. U. Daim, H. Martin, and P. Gerdsri, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," *Technological Forecasting and Social Change*, vol. 73, pp. 981-1012, 2006.
[46] G. Cascini and D. Russo, "Computer-aided analysis of patents and search for triz contradictions," *International Journal of Product Development*
vol. 4, pp. 52-67, 2007.
[47] G. Cascini, A. Fantechi, and E. Spinicci, *Natural language processing of patents and technical documentation*. Berlin Heidelberg, 2004.
[48] G. Cascini, D. Lucchhesi, and R. P., "Automatic patents functional analysis through semantic processing.," in *XII ADM International Conference*, 2001.
[49] J. R. Koza and R. Poli, *Search methodologies introductory tutorials in optimization and decision support techniques* Springer US, 2005.
[50] H. Yu, J. Yang, J. Han, and X. Li, "Making svms scalable to large data sets using hierarchical clustering indexing," *Data Mining and Knowledge Discovery*, vol. 11, pp. 295-321, 2005.
[51] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the 21st International Conference on Machine Learning*. Banff, Canada, 2004.
[52] T. Joachims, *Making large-scale svm learning practical*: MIT Press, Cambridge, USA, 1998.
[53] C. Cardie, "Using decision trees to improve case-based learning," in *Proceedings of the Tenth International Conference on Machine Learning*, Proceedings of the Tenth International Conference on Machine Learning, pp. 25-32.
[54] E. Doganavsargil and M. Fattori, "Decision tree analysis as a tool to optimise patent current awareness bulletins," *World Patent Information*, vol. 30, pp. 212-219, 2008.
[55] L. H. Tong, H. Cong, and S. Lixiang, "Automatic classification of patent documents for triz users," *World Patent Information*, vol. 28, pp. 6-13, 2006.
[56] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197-243, 1995.
[57] D. Heckerman, "Bayesian networks for data mining," *Data Mining and Knowledge Discovery*, vol. 1, pp. 79-119, 1997.
[58] K. P. Murphy, "The bayes net toolbox for matlab," in *Computing Science and Statistics*, 2001.
[59] A. Kusiak, "Innovation science: A primer," *Int. J. Computer Application in Technology*, vol. 28, pp. 140-149, 2007.

[60] B. Agard and A. Kusiak, "Data-mining-based methodology for the design of product families," *International Journal of PRoduction Research*, vol. 42, pp. 2955-2969, 2004.
[61] U. R. Kulkarni, S. Ravindran, and R. Freeze, "A knowledge management success model: Theoretical development and empirical validation," *Journal of management information systems*, vol. 23, pp. 309, 2006.
[62] *Montylingua website, (web.Media.Mit.Edu/~hugo/montylingua/)*, 2004.
[63] H. Liu, "Conceptnet: A practical commonsense reasoning toolkit," *BT Technology Journal*, vol. 22, pp. 211-226, 2004.
[64] E. Brill, "A simple rule-based part of speech tagger," in *In proceedings of the Third Conference on Applied Natural Language Processing*, 1992.
[65] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Computational Linguistics*, vol. 21, pp. 543-565, 1995.
[66] E. Charniak, "Statistical techniques for natural language parsing," in *AI magazine*, vol. 18, 1997, pp. 33-43.
[67] E. Charniak, C. Hendrickson, N. Jacobson, and M. Perkowitz, "Equations for part of speech tagging," in *In Proceedings of the Eleventh National Conference on Artificial Intelligence*, Menlo Park, 1993, pp. 784-789.
[68] E. Charniak, G. Carroll, J. Adcock, A. Cassandra, Y. Gotoh, J. Katz, M. Littman, and J. McCann, "Taggers for parsers," *Artificial Intelligence*, vol. 85, 1995.
[69] V. M. Tsourikov, L. S. Batchilo, and I. V. Sovpel, "Document semantic analysis/selection with knowledge creativity ", USPTO, Ed., G06F 1727 ed: Invention Machine Corporation, 2000.
[70] J. Giménez and L. Márquez, "Svmtool: A general pos tagger generator based on support vector machines," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, 2004.
[71] L. Màrquez and H. Rodríguez, "Part-of-speech tagging using decision trees " in *Machine learning*, *Lecture notes in computer science*: Springer Berlin/Heidelberg, 2006.
[72] T. L. L. Shu., H. Hansen, and L. Alting., "Biomimetics applied to centering in microassembly," *Annals of the CIRP*, vol. 52, pp. 101-104, 2003.
[73] L. H. Shu, N. H. Hansen, A. Gegeckaite, J. Moon, and C. Chan, "Case study in bio-mimetic design: Handling and assembly of microparts," in *Proc. ASME IDETC/CIE, Paper No. DETC2006-99398*, Philadelphia, PA., 2006.
[74] I. Chiu and L. H. Shu, "Natural language analysis for bio-mimetic design," in *ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Salt Lake City, Utah USA, 2004.
[75] I. Chiu and L. H. Shu, "Bridging cross-domain terminology for bio-mimetic design," in *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Long Beach, California, USA, 2005.
[76] I. Chiu and L. H. Shu, "Bio-mimetic design through natural language analysis to facilitate cross-domain information retrieval," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 21, pp. 45-59, 2007.
[77] V. Vakili and L. H. Shu, "Towards biomimetic concept generation," in *ASME 2001 Design Engineering Technical Conferences Design Theory and Methodology*. Pittsburgh, Pennsylvania, 2001.
[78] *Wordnet 2.0 (n.D.) available at: Http://www.Cogsci.Princeton.Edu/~wn/*,

[79] M. C. Yang and D. J. Epstein, "A study of prototypes, design activity, and design outcome," *Design Studies*, vol. 26, pp. 649-669, 2005.
[80] M. C. Yang and M. R. Cutkosky, "Automated indexing of design concepts for information management," in *Proc. Int. Conf. Engineering Design*, Tampere, Finland, 1997.
[81] M. C. Yang, W. H. Wood, and M. R. Cutkosky, "Data mining for thesaurus generation in informal design information retrieval," in *Proc. 1998 Int. Congr. Civil Engineering*, Boston, 1998.
[82] M. C. Yang, W. H. W. III, and M. R. Cutkosky, "Design information retrieval: A thesauri-based approach for reuse of informal design information," *Engineering with Computers*, vol. 21, pp. 177-192, 2005.
[83] Z. Li, M. Liu, D. C. Anderson, and K. Ramani, "Semantics-based design knowledge annotation and retrieval," in *ASME 2005 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. Long Beach, California USA, 2005.
[84] F.-C. H. Amy J.C. Trappey, Charles V. Trappey and Chia-I. Lin, "Development of a patent document classification and search platform using a back-propagation network," *Expert Systems with Applications*, vol. 31, pp. 755-765, 2006.
[85] P. C. Matthews, "Bayesian networks for engineering design," in *WCE*, 2007, pp. 284-289.
[86] G. Altshuller, "Triz keys to technical innovation ." in *40 principles*, R. S. Shulyak L., Ed. Worcester, MA.: Technical Innovation Center, 1997.
[87] *40 invention principles with examples available from: Http://www.Oxfordcreativity.Co.Uk/.*
[88] *Weka. Available from: Http://www.Cs.Waikato.Ac.Nz/ml/weka/.*
[89] *Matlab availble at: Http://www.Mathworks.Com*,
[90] L. Yanhong and T. Runhua, "A text-mining-based patent analysis in product innovative process," *Trends in Computer Aided Innovation*, vol. 250, pp. 89-96, 2007.
[91] Z. Peng and T. Runhua, "Using the cai software (inventiontool 3.0) to solve complexity problem," *Trends in Computer Aided Innovation*, vol. 250, pp. 115-123, 2007.
[92] Salton&McGill, *Introduction to modern information retrieval*. New York, USA: McGraw-Hill, 1983.
[93] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130–137, 1980.
[94] J. Shlens, "A tutorial on principal component analysis." La Jolla, Ca: Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 2005.
[95] A. Selamat and S. Omatu, "Web page feature selection and classification using neural networks," *Information Sciences*, vol. 158, pp. 69-88, 2004.
[96] C.-P. Wei, C. C. Yang, and C.-M. Lin, "A latent semantic indexing-based approach to multilingual document clustering," *Decision Support Systems*, vol. 45, pp. 606-620, 2008.
[97] I. C. a. L. H. Shiu, "Natural language analysis for biomimetic design," in *ASME 2004 International Design Engineering Technical Conference and Computers and Information in Engineering Conference*. Salt Lake City, Utah, USA, 2004.
[98] I. C. a. L. H. Shiu, "Bridging cross-domain terminology for biomimetic design," in *ASME 2005 International Design Engineering Technical Conference and Computers and Information in Engineering Conference*. Long Beach, Ca, USA, 2005.
[99] I. C. a. L. H. Shiu, "Bimimetic design through natural language anaysis to facilitate cross-domain information retrieval," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 21, pp. 45-59, 2007.
[100] S. T. a. K. Koutroumbas, *Pattern recognition*, 2nd ed: Elsevier Academic Press, 2003.

[101] T. M. Mitchell, *Machine learning*: McGraw-Hill, 1997.
[102] USPTO, "Patent rules: Title 37 - code of federal regulations - patents, trademarks, and copyrights," vol. Tilte 37, 2000.
[103] C.-P. W. Christopher C. Yang, Chia-Min Lin, "A latent semantic indexing-based approach to multilingual document clustering," *Decision Support Systems*, vol. 45, pp. 606-620, 2008.
[104] A. Chakrabarti and T. P. Bligh, "An approach to functional synthesis of mechanical design concepts: Theory, applications, and emerging research issues," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 10, pp. 313-331, 1996.
[105] P. S. a. M. I. Campbell, "A study on the grammatical construction of function structures," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 19, pp. 139-160, 2005.
[106] J. S. G. a. U. Kannengiesser, "A function-behavior-structure ontology of processes," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 21, pp. 379-391, 2007.
[107] C. R. B. Michanel Van Wie, Matt R. Bohm, Daniel A. Mcadams, and Robert B. Stone, "A model of function-based representations," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 19, pp. 89-111, 2005.
[108] H. D. William H. Wood, and Clive L. Dym, "Integrating functional synthesis
" *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 19, pp. 183-200, 2005.
[109] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, pp. 130-137, 1980.
[110] M. Hundal, "A systematic method for developing function structures, solutions and concept variants," *Mech Mach Theory*, vol. 25, pp. 243-256, 1990.
[111] K. M. Nelson and H. J. Nelson, "Technology flexibility: Conceptualization, validation, and measurement," in *Proceedings of The Thirtieth Annual Hawwaii International Conference on System Sciences*, 1997.
[112] G. M. Seddon, "The measurement and properties of divergent thinking ability as a single compound entity," *Journal of Educational Measurement*, vol. 20, pp. 393-402, 1983.
[113] M. Ozman, "Breadth and depth of main technology fields: An empirical investigation using patent data," in *Science and Technology Policies Research Center Working Paper Series*, 2007.

APPENDIX A

SOFTWARE MODULES AND EXAMPLE OUTPUT

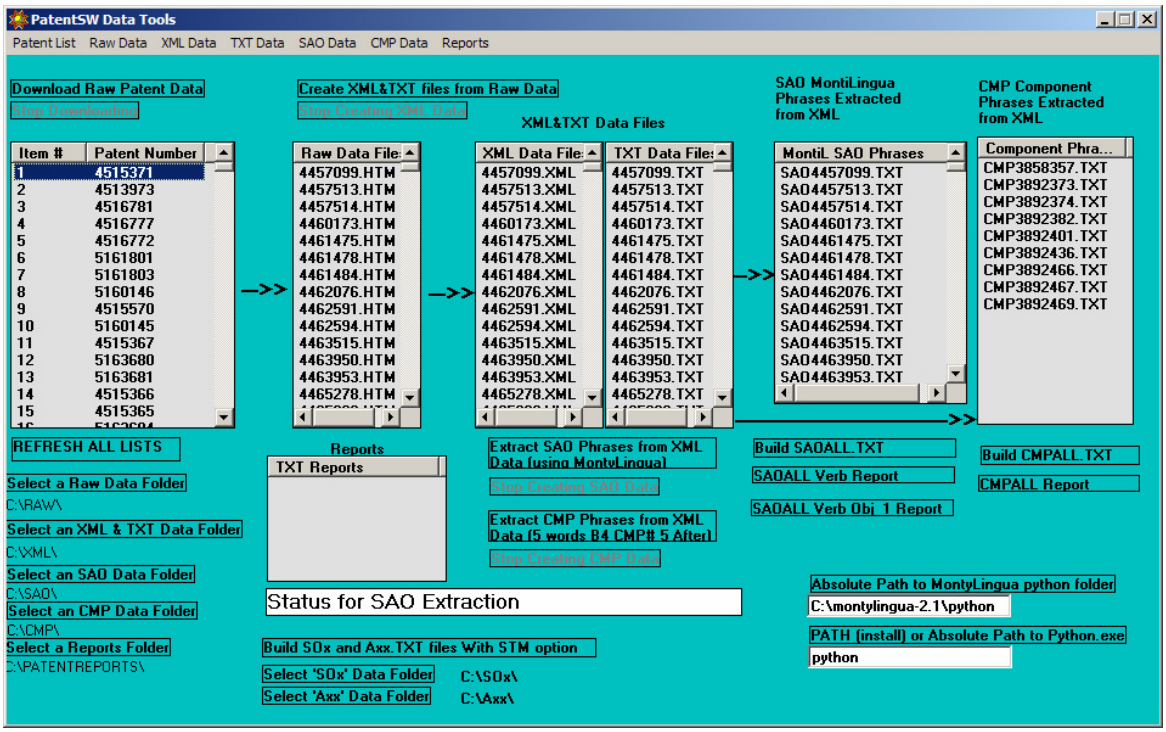


Figure A-1 Patent Information Extraction and NLP SW

Example Patent XML Report Generated using Patent SW for Patent # 5178393


```
<?xml version="1.0" ?>
- <patentdata>
  <patentnumber>5178393</patentnumber>
  <Patent_Title>Method and apparatus for measuring golf driving
    distance</Patent_Title>
  <Abstract>A method and apparatus for the measurement of the driving distance of
    a golf swing by measuring the initial velocity of a golf ball tethered by a string
    to a rotating axle.</Abstract>
  <Inventors>Dennesen; James J. (Haverhill, MA)</Inventors>
  <Assignee>Dennco, Inc. (Salem, NH)</Assignee>
  <Filed_date>November 4, 1991</Filed_date>
  <Current_US_Class>473/140</Current_US_Class>
    <Field_of_Search>273/185C,185D,184B,184R,184A,185R,185A,185B,197R,197A
    ,2R,2B</Field_of_Search>
- <Citations_made>
- <citations>
  <citation>3815922 June 1974 Brainard</citation>
  <citation>4429880 February 1984 Chen et al.</citation>
  <citation>4660835 April 1987 Lorurto</citation>
  <citation>4971326 November 1990 Montone</citation>
  <citation>5035432 July 1991 Lew</citation>
    </citations>
    </Citations_made>
```

<Claims>**What is claimed is: 1. Apparatus for estimating the length of a golf drive comprising: a golf ball tethered to a rotatable axle at a preselected distance from the axle; circuitry for calculating the estimated length of the drive from the initial velocity of the ball; and display apparatus to display said calculated estimated length. 2. Apparatus of claim 1 wherein the initial velocity is calculated from the angular velocity of the axle and the preselected distance of the ball from the axle. Description**</Claims>

<Description>**BACKGROUND OF THE INVENTION This invention relates to a method and apparatus for measuring the length of a golf drive within a very limited space. Their are currently no inexpensive and accurate means of measuring the driving range of a golf swing. By combining a compact design based on rotary motion with a very sensitive sensing scheme for the initial velocity of the ball, the invention disclosed herein allows for instant readouts of driving distance without requiring a large open space. SUMMARY OF THE INVENTION According to the invention, the apparatus for estimating the length of a golf drive includes a golf ball tethered to a rotatable axle at a preselected distance from the axle. A sensor is provided to measure the angular of velocity of the axle and circuitry is provided for calculating the estimated length of the drive from the initial velocity of the ball. Display apparatus displays the calculated estimated length. In a preferred embodiment, the initial velocity is calculated from the angular velocity of the axle and the preselected distance of the ball from the axle by means of a digital circuit. BRIEF DESCRIPTION OF THE DRAWINGS FIG. 1 is a perspective view of the apparatus for measuring driving distance. FIGS. 2A and 2B show a circuit diagram for the processor that calculates the driving distance. DESCRIPTION OF THE PREFERRED EMBODIMENT With reference first to FIG. 1, a golf ball 1 is tethered by a string 2 to a rotatable axle 3. The axle 3 also carries a slotted disc 4 which cooperates with an optical source and receiver unit 5 for determining the angular velocity of the axle 3. As the slotted disc 4 rotates it alternately blocks and passes light creating electrical pulses**

which are conveyed by wires to the processing unit that performs the calculations as will be described herein below. The apparatus of the present invention includes a base plate 6 which may be affixed to the ground by means of spikes 7. The output of the optical unit 5 serves as the input to the circuit diagram shown in FIG. 2. As will be appreciated by those skilled in the art, the circuit of FIG. 2 will calculate the distance that the ball would have traveled if untethered and can be programmed to use calibration constants based on field measurements. The estimated distance is displayed in liquid crystal display 8. In operation, a golfer selects a club and strikes the ball which will then rotate on the axle 3. The particular club selected can serve as one of the calibration factors in computing the estimate of the distance. After striking the ball, the display 8 will show the estimate for the distance the ball would have traveled if untethered and struck with that particular club. Thereafter, the apparatus is reset for an additional swing. The calculation of the distance estimate is based on simple Newtonian mechanics as is well known to those in the ballistics arts. * * * * *</Description>

</patentdata>

C:\SA0\SA05178393.TXT

Verb	Subject	Object 1	Object 2	Object 3	Obje
relate	invention	to method and ...			
measure	length	of golf drive			
be	inexpensive an...				
measure	drive range	of golf swing			
combine	compact design				
base	compact design	on rotary motion			
allow	invention	for instant read...			
drive	distance				
require	large open space				
estimate	length	of golf drive			
include	golf drive	golf ball			
tether	golf ball	to rotatable axle	at preselected ...	from axle	
measure	sensor				
provide	axle and circuitry				
calculate	estimate length	of drive	from initial veloc...	of ball	
display	Display apparatus	calculate estim...			
calculate	initial velocity	from angular vel...	of axle		
mean	of digital circuit				
be	1	perspective view	of apparatus		
drive	distance				
calculate	that				
drive	distance				
tether	golf ball 1	by string 2	to rotatable axle 3		
carry	axle 3	slot disc 4	which		
cooperate	which	with optical sou...			
determine	angular velocity	of axle 3			
rotate	slot disc 4	it			
block	it				
pass	light				
create	light	electrical pulse	which		
convey	which	by wire	to processing unit		
perform	that	calculation			
describe					
include	present invention	base plate 6	which		
affix	which	to ground			
mean	of spike 7				
serve	optical unit 5	as input	to circuit diagram		
show	circuit diagram	in FIG			
appreciate					
calculate	2	distance	that ball		
travel	ball	if untethered			
use	calibration cons...				
base	calibration cons...	on field measur...			
display	estimate distance	in liquid crystal ...			
select	golfer	club			
strike	ball	which			
rotate	which	on axle 3			
select	particular club				

Figure A-2 Example Verb, Subject, Object Report for Patent # 5178393

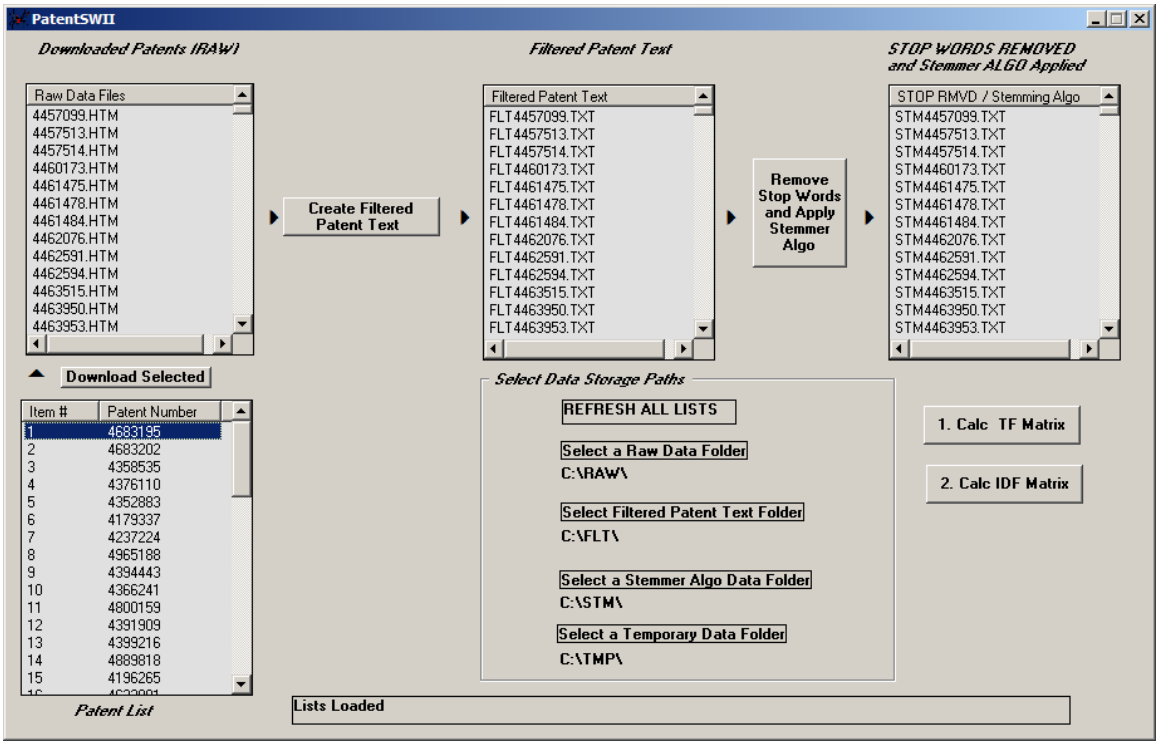


Figure A-3 Latent Semantic Analysis SW

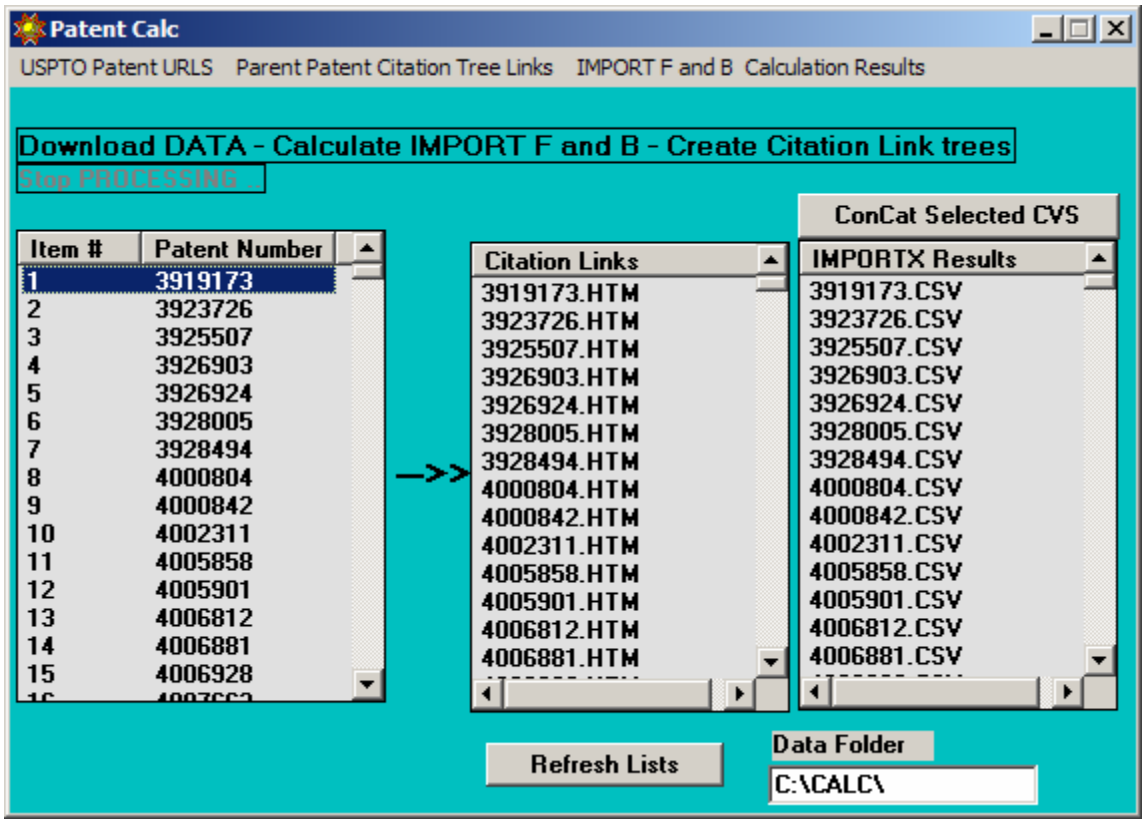


Figure A-4 Patent Citation and Importance Calculation Source Code and Example output

Example Citation Tree Output file

PARENT PATENT: 3919173

Moisture curable polyurethane systems

CITED PATENTS [2nd GEN CITED COUNT]

2830037 [0]
3192186 [0]
3351573 [0]
3352830 [0]
3425973 [0]
3463748 [0]
3479325 [0]
3549569 [0]
3554962 [0]
3652508 [0]
3663514 [0]

IMPORTB = NCITED + discountFactor * sumNCITED
11 = 11 + 0.5 * 0

CITING PATENTS [2nd GEN CITING COUNT] - CITING PATENT TITLE

7,435,464 [0] - Articles comprising aqueous dispersions of polyureaurethanes
6,846,849 [0] - Saccharide-based resin for the preparation of foam
6,822,042 [0] - Saccharide-based resin for the preparation of composite products

6,720,385 [2] - Polyurethane latexes, processes for preparing them and polymers prepared therewith

6,680,356 [2] - Coating composition having improved early hardness and water resistance

6,677,425 [0] - Clear coating composition having improved early hardness and water resistance

6,610,228 [0] - Dry process for bonding silica-rich plant materials

6,514,572 [0] - Polyurethane films prepared by electrodeposition from polyurethane dispersions

6,472,493 [5] - Clear coating composition having improved early hardness and water resistance

6,451,908 [3] - Polyurethane films prepared from polyurethane dispersions

6,433,059 [0] - Method for preparing binder materials containing diisocyanates

6,368,714 [2] - Moisture-activated adhesive compositions

6,180,713 [0] - One-can moisture-curing urethane compositions

6,087,440 [2] - Continuous process for preparing a polyurethane latex

5,959,027 [9] - Continuous process for preparing a polyurethane latex

5,574,114 [6] - Mixture of isocyanate-terminated polyurethane prepolymers

5,559,196 [4] - Mixture of isocyanate-terminated polyurethane prepolymers

5,558,941 [2] - Article including an adhesively bonded moisture cured material and a method of making the same

5,536,805 [8] - Mixture of isocyanate-terminated polyurethane prepolymers having good adhesion

5,506,328 [16] - Low VOC, moisture curable, two-component coating compositions based on organic polyisocyanates

5,436,302 [2] - Mixture of isocyanate-terminated polyurethane prepolymers

5,418,310 [10] - Mixture of isocyanate-terminated polyurethane prepolymers having good adhesion

5,147,897 [3] - Method for producing non-yellowing polyurethane urea foam

5,064,871 [3] - Latent catalysts comprising bismuth carboxylates and zirconium carboxylates

RE33,175 [0] - Method for making decorative emblems

4,889,748 [12] - Display device

4,812,356 [8] - Coating composition for flexible substrates and the use thereof, and a method for the production of a protective coating

4,710,560 [12] - Polyurethane coating composition

4,562,289 [2] - Homogeneous storage stable cyanamide solutions in polyols and a process for their production

4,513,112 [8] - High build, ambient cure coating compositions

4,395,530 [7] - Catalyst initiated prepolymer systems

4,345,058 [10] - Urethane prepolymer repair system

4,211,847 [3] - Polyurethane foams and foam forming compositions containing amine scavengers

4,199,489 [12] - Moisture curing polyurethane topcoat paint displaying geometric metamerism

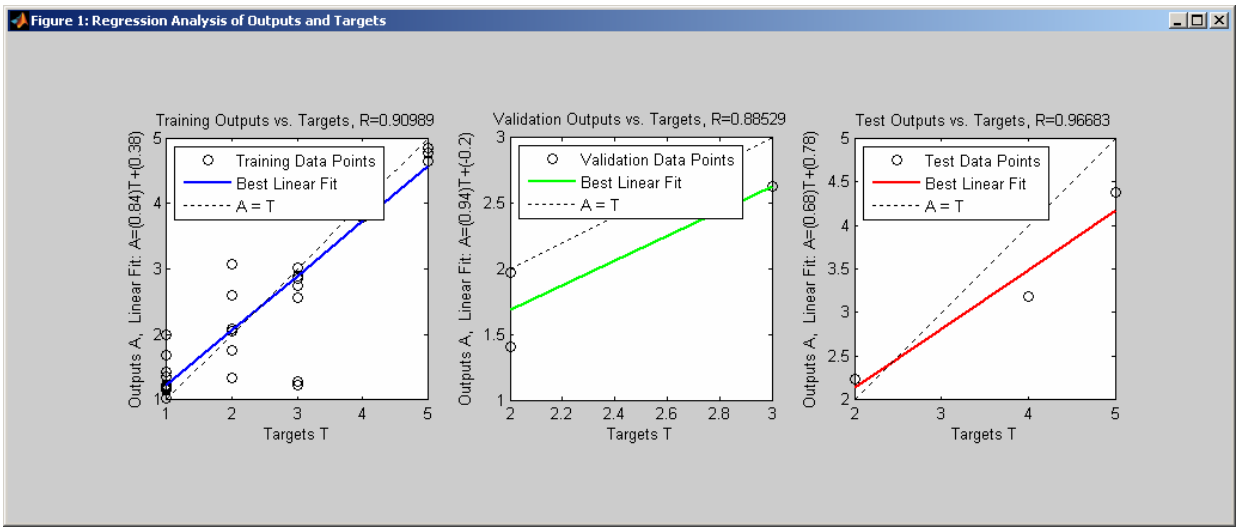
4,100,010 [30] - Method for making decorative emblems

IMPORTF = NCITING + discountFactor * sumNCITING

126.5 = 35 + 0.5 * 183

APPENDIX B

NEURAL NETWORK PREDICTION RESULTS



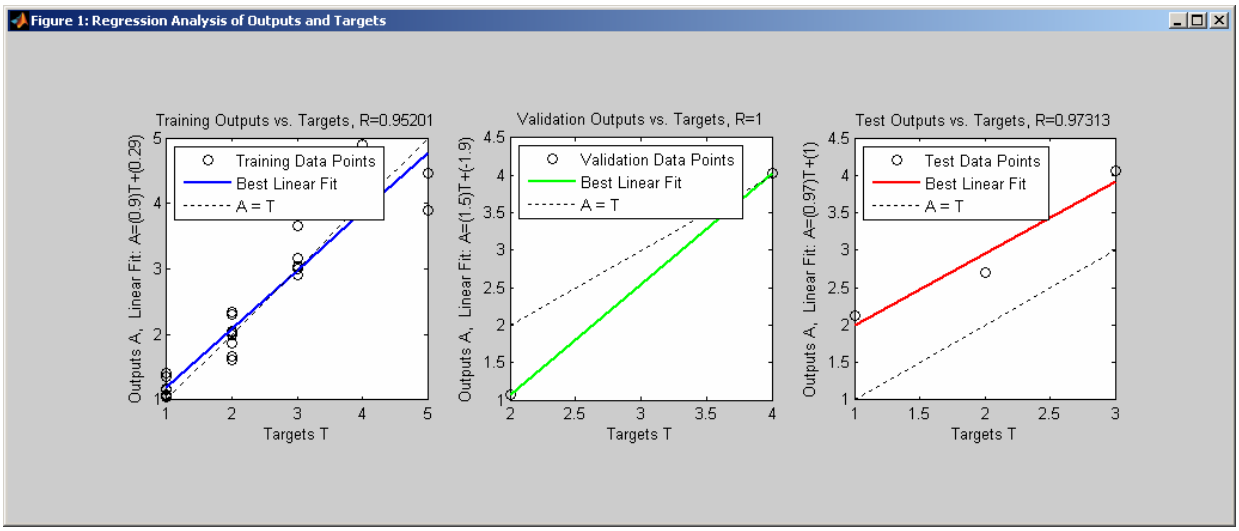


Figure B-1 Predicting Level of Invention Using Neural Network back propagation and 100 hidden neurons

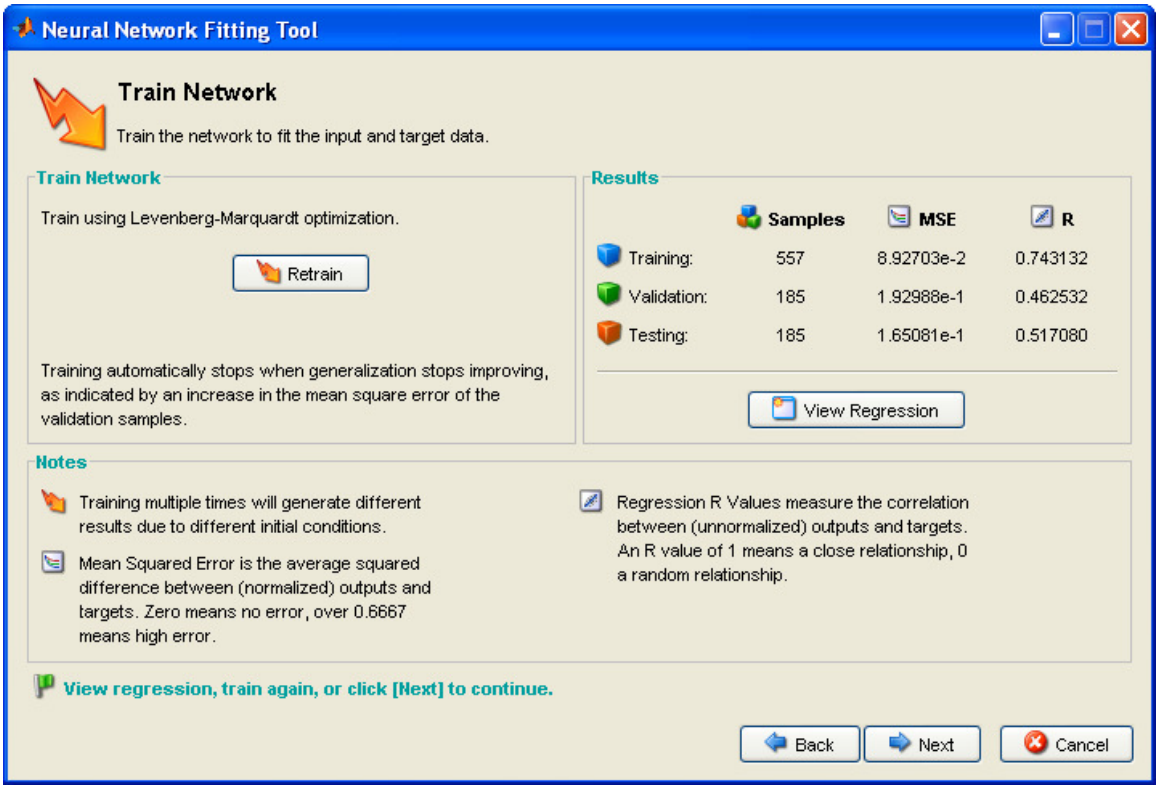
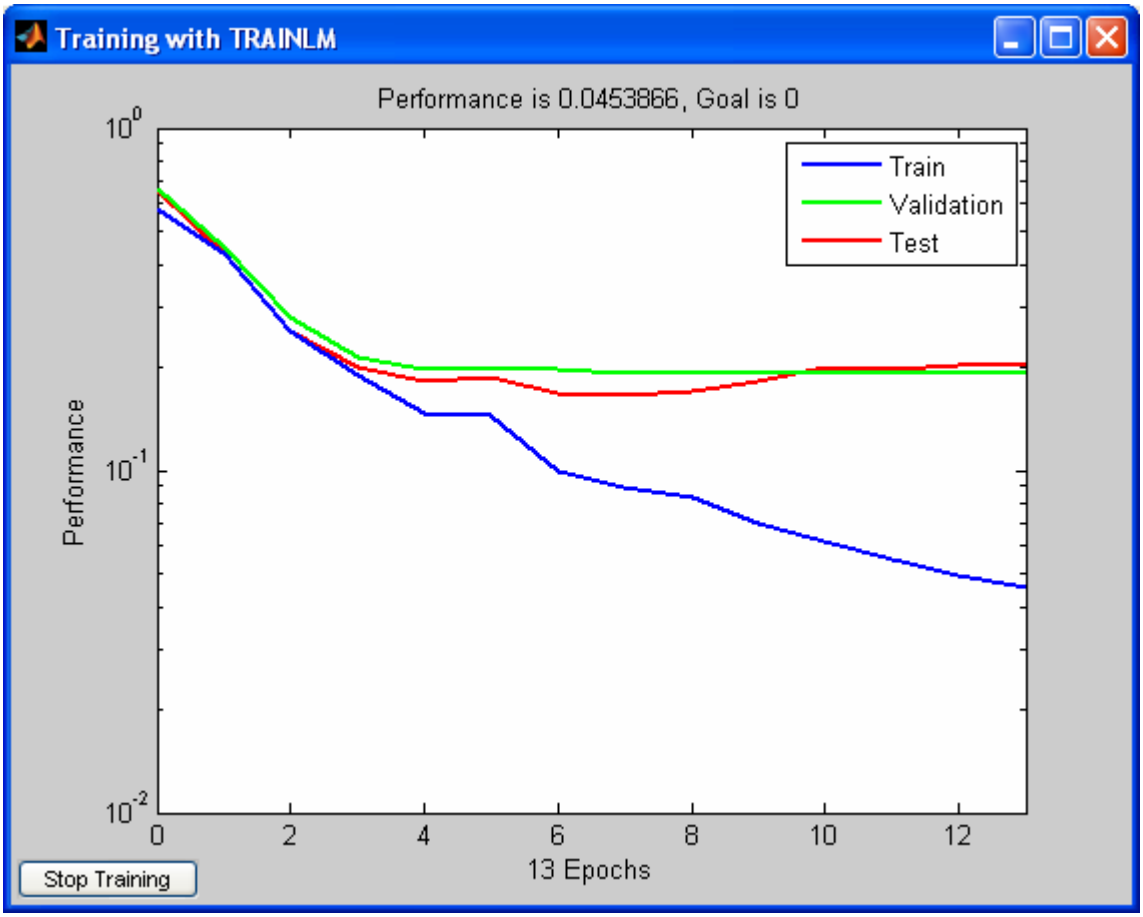


Figure B-2 Predicting Generality Using Neural Network back propagation and 20 hidden neurons



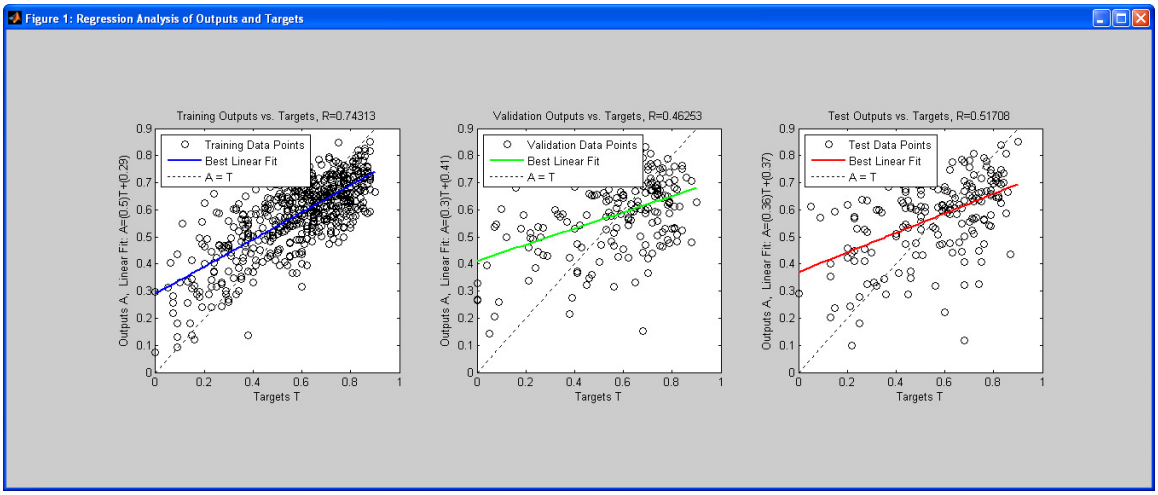


Figure B-3 Network Training Using Generality Data

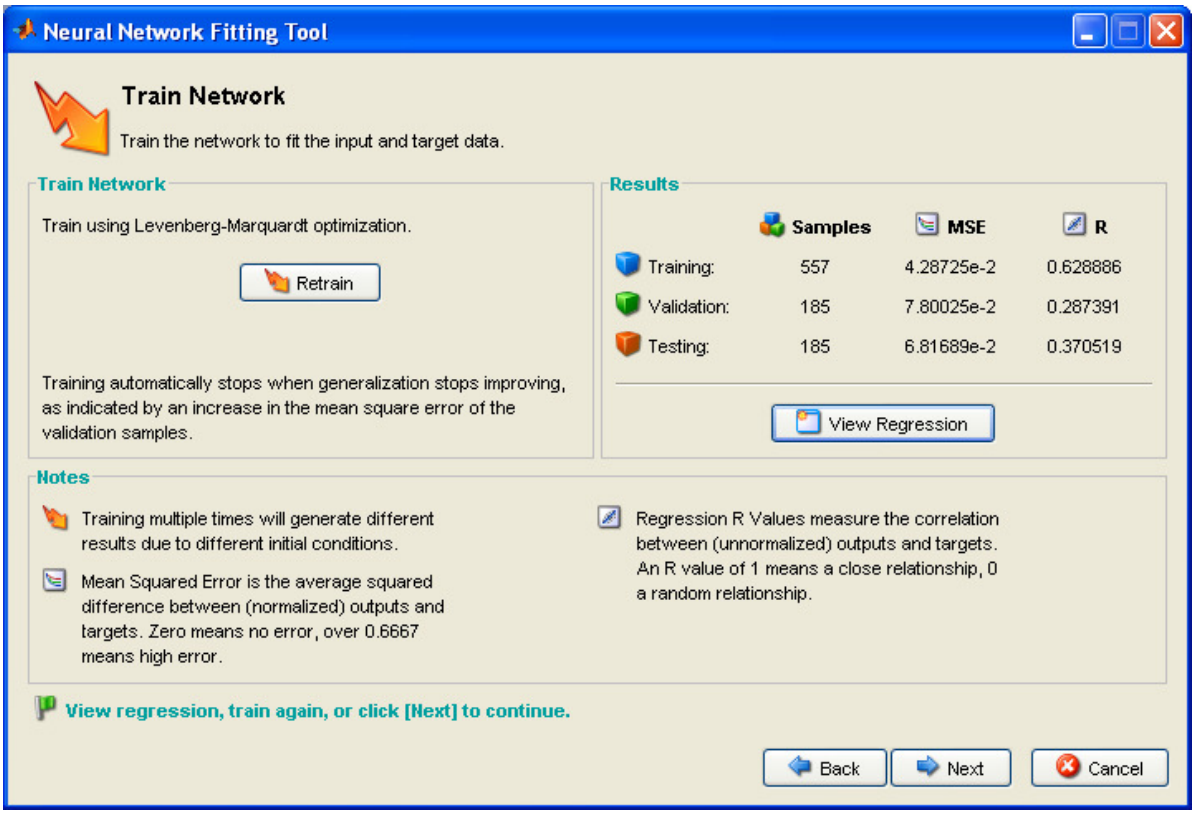
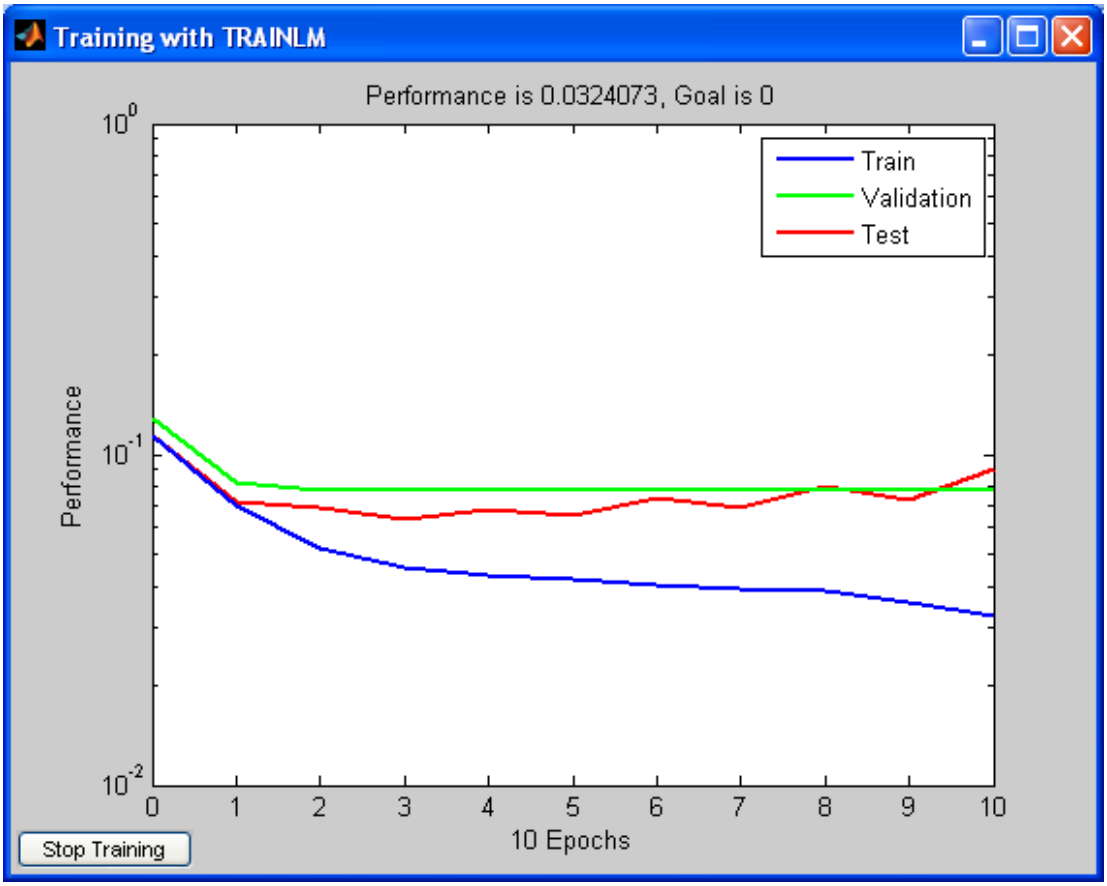


Figure B-4 Predicting Importance Using Neural Network back propagation and 20 hidden neurons



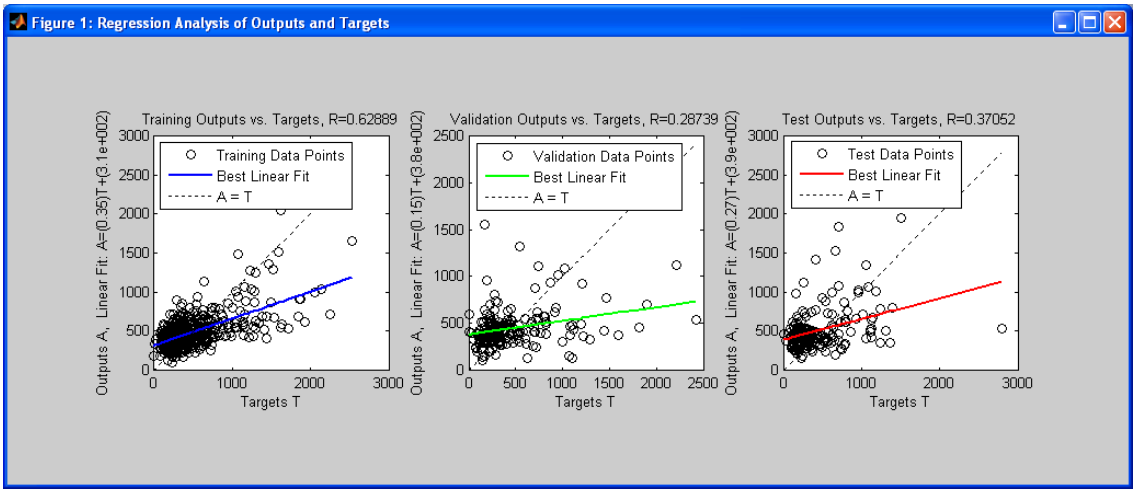


Figure B-5 Network Training Using Forward Importance Data

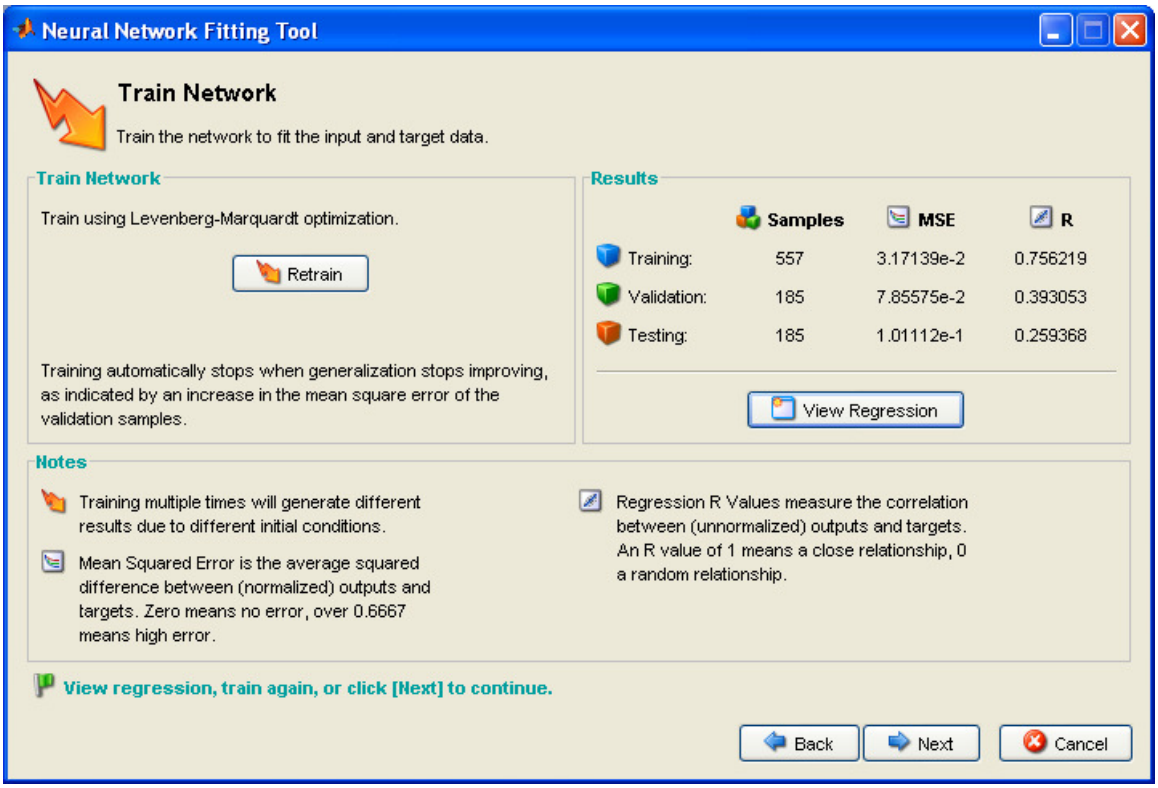


Figure B-6 Predicting Citations Received Using Neural Network back propagation and 20 hidden neurons

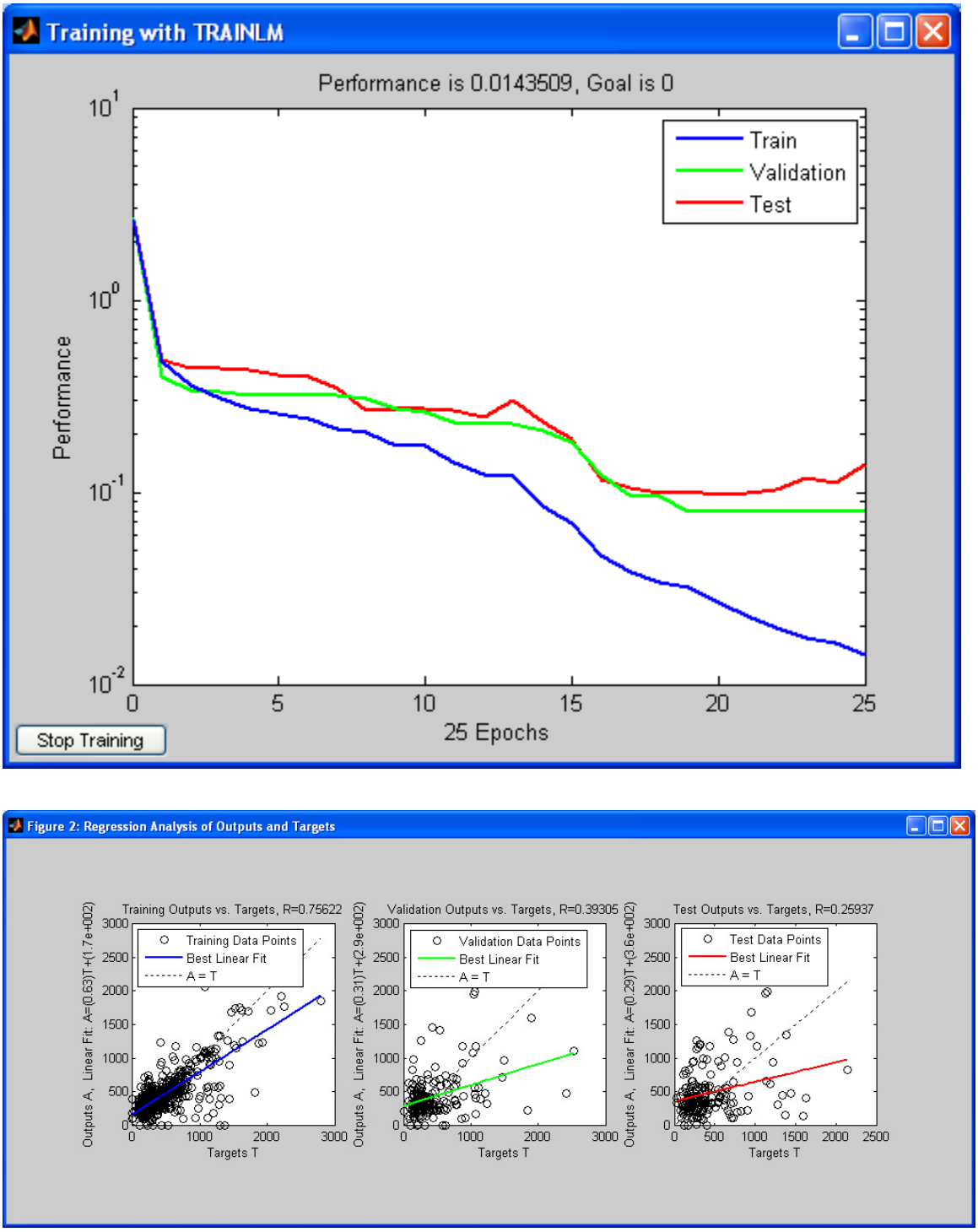


Figure B-7 Network Training Using Forward Citations Received Data

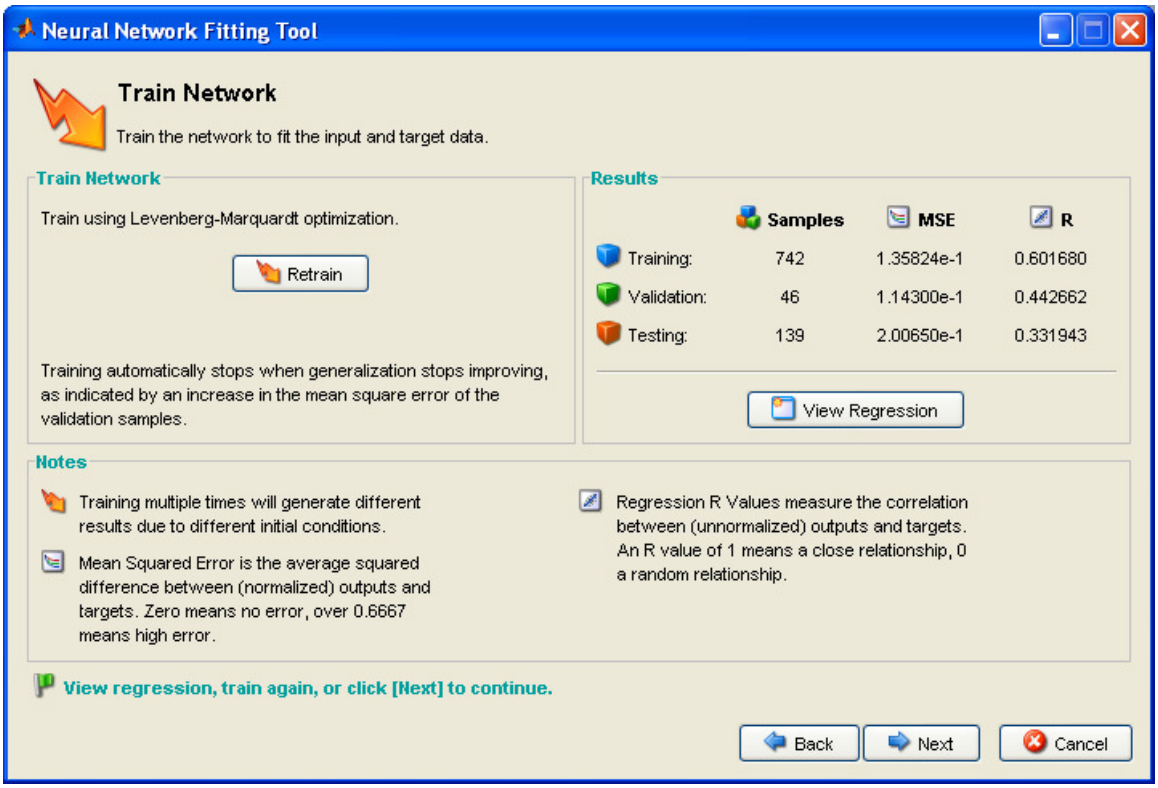


Figure B-8 Predicting Importance Using Neural Network back propagation and 20 hidden neurons

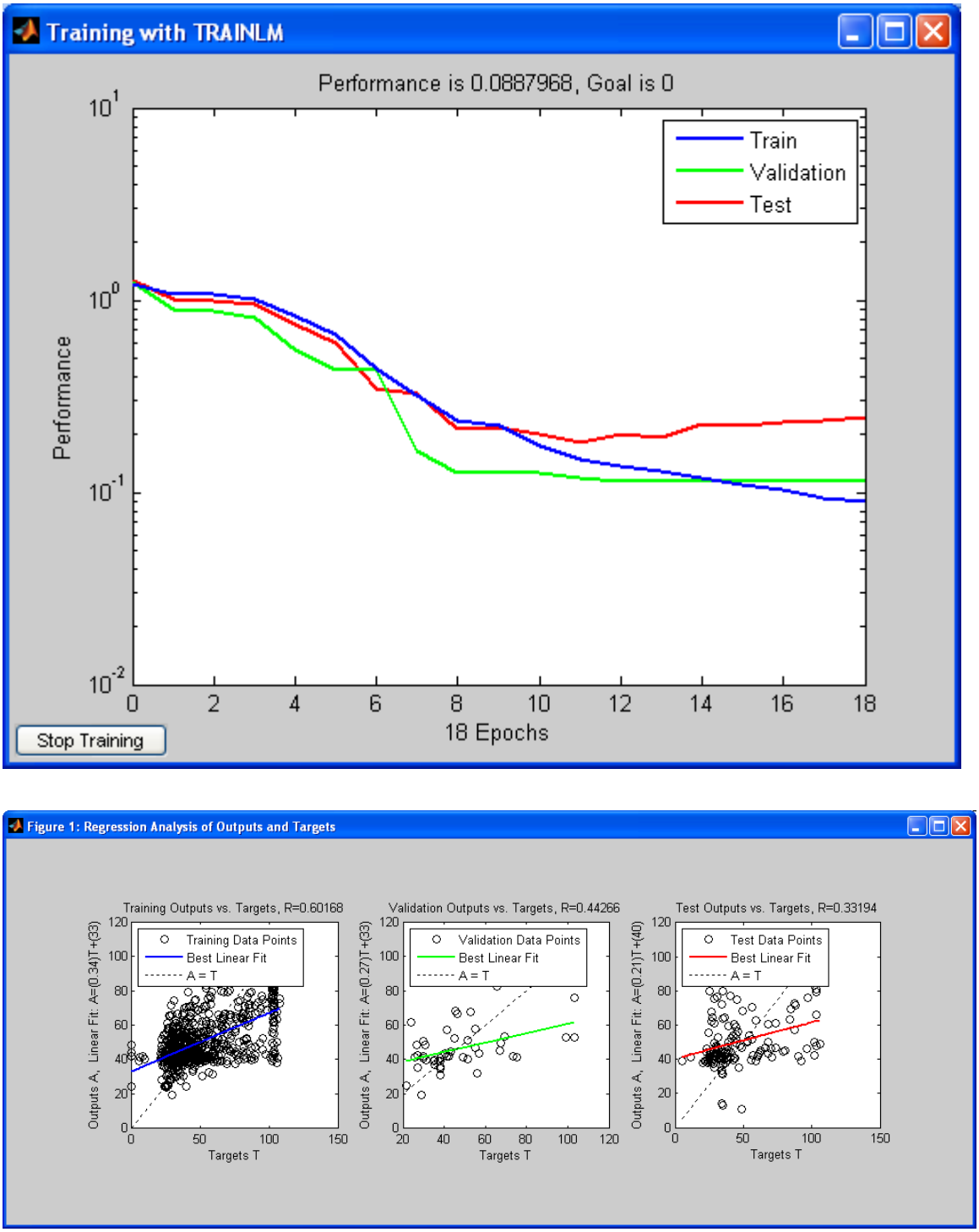


Figure B-9 Network Training Using Forward Citations Received Data

APPENDIX C

SURVEY OF DATA MINING TECHNIQUES

One of the tasks performed as part of this transdisciplinary research was reviewing different machine learning and data mining tools to determine the best approach/technique to apply to the problem. A number of data mining and machine learning techniques and approaches were reviewed to select the best method to apply to the problem. The initial methods reviewed include the use of genetic algorithms and genetic programming, support vector machines (SVM), artificial neural networks, decision trees, clustering, and learning Bayesian networks. The following includes a summary of each of the machine learning and data mining approaches.

1. Genetic Algorithms and Genetic programming: This machine learning technique is an optimization technique that is based on evolutionary concepts. Genetic algorithms and genetic programming use ideas such as natural selection, genetic mutation, and genetic combination to perform optimization. [49]

2. Support Vector Machines: This machine learning technique is a method used for classification and regression analysis. SVMs are also known as maximum margin classifiers. [50]

3. Artificial Neural Networks: This machine learning/data mining technique is based on biological neural networks and represents a non-linear predictive model that has the ability to learn by training the model with historical data. [43]

4. Decision Trees: This data mining method is used to create hierarchical trees that include nodes at each branch in the tree that represent a set of decisions. The decisions in the tree are used to construct rules for classifying the data in the database analyzed. [53]

5. Clustering: consists of classifying data elements into clustered groups by partitioning the data in a database into a subset of clusters. This is often performed using some measure of distance between the different data elements in the database. [38]

6. Learning Bayesian Network: A Bayesian network constructs a directed acyclic graphical model that is used to create probabilistic relationships between a set of variables. [57]

Once a machine learning or data mining technique/approach is selected to apply to the problem, a predictive model will be developed using the applicable method.

This section includes preliminary results from applying different data mining and machine learning techniques/approaches to NBER patent data. Figure C-1 demonstrates the use of clustering analysis to segregate different patent attributes into multiple clusters. The main patent attributes analyzed in Figure C-1 include the patent application year, number of citations received by each of the patents and the patent class for each patent. This analysis was conducted on a subset of the patent data associated with the communications patent category. The clustering analysis provides interesting insight into the different classes of patents in the category of communications and helps understand the correlation between the application year of a patent, the patent class and the number of citations

received. However, clustering analysis was not selected as the approach to apply to the problem due to the limited ability to predict the success for different patent parameters.

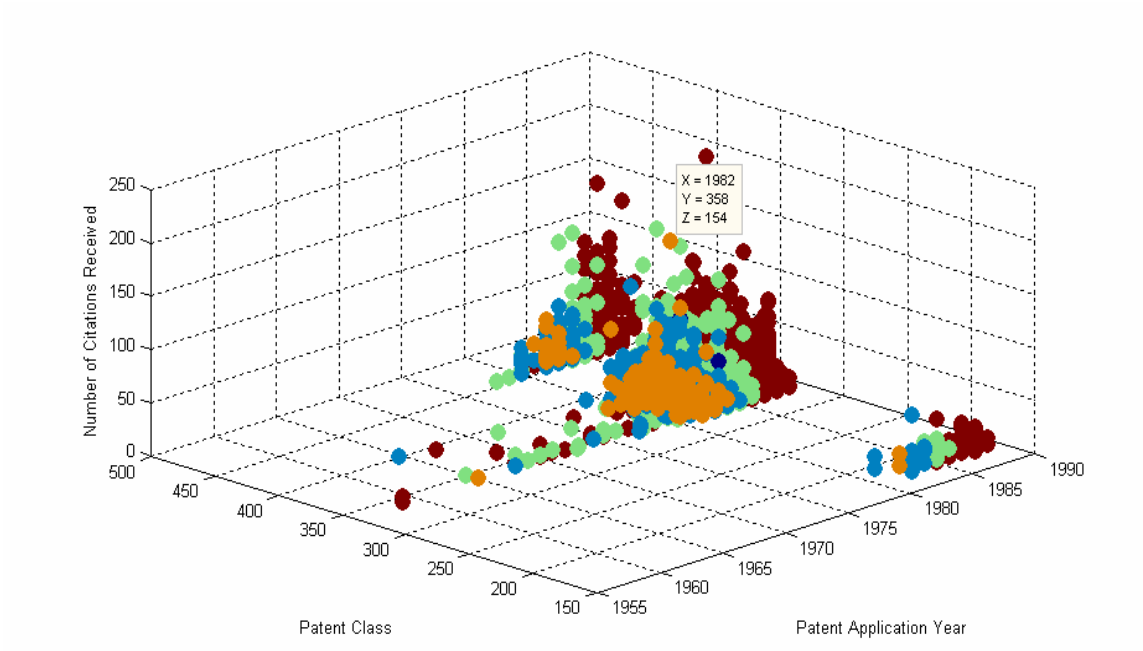


Figure C-1 Example Clustering Analysis Applied to Patent Data

The second data mining approach used to analyze patent data was the decision tree. Figure 3 2 provides an example decision tree applied to patents in the communications patent technological category. The decision tree includes branches that reflect the number of citations received for each patent and enables the data analyzer to make decisions when selecting patents to review based on the number of patent citations received. In addition, the decision tree is also thought of as a machine learning enabler since it can be used to learn from data by assigning a set of decisions to each of the branches in the tree by automatically arranging the data. Figure 3 2 shows the branches on the decision tree that provides patent numbers at the end of each branch on the tree based on the number of patent citations received.

The decision tree can become much more complicated by including data from the patent database to make decisions based on the number of citations made to other patents, the state in which the patent originated, the patents level of originality, and patent assignee code. A tree including more than three or four input variables is too complicated to show as an example in this Dissertation. Therefore, Figure C-2 is included as a representation of an example decision tree. Keep in mind that a much more complicated decision tree will have many more branches and decisions that would take hours to review and require the reviewer to prune branches on the tree in order to see higher level decisions. The decision tree could be used as a forward-looking measure to review what patented ideas have the biggest impact on future designs. In addition, the decision can be used to classify designs based on their level of success. The decision tree does provide a probability for each of the branches in the tree. However, the decision tree was not selected for use in the predictive models at this time. It will be considered in future work as a potential method to apply supervised learning to one or more data mining and machine learning techniques and approaches.

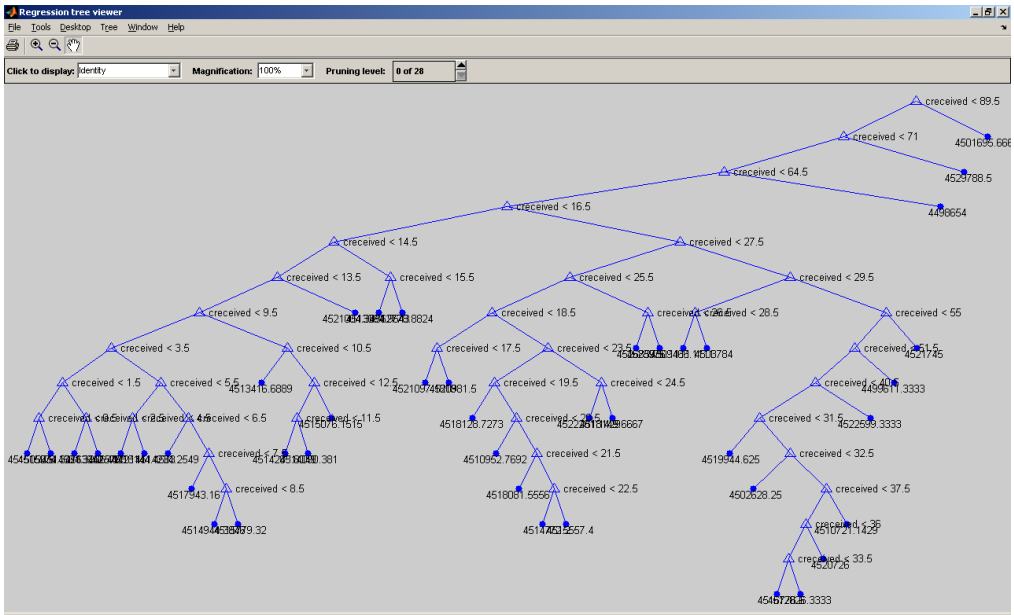


Figure C-2 Example Decision Tree Applied to Patent Data

Other data mining and machine learning techniques reviewed for use in the predictive models include genetic artificial neural networks and learning Bayesian networks. A preliminary learning Bayesian network was constructed using variables included in the NBER patent database. The initial set of original independent variables selected for the learning Bayesian network include the patent application year, the patent assignee code, the patent technological subcategory, and the state of the first inventor where the patent originated. In addition, a set of constructed independent variables were selected from the variables constructed by Jaffe and Trajtenberg [23]. The constructed variables include the number of patent citations received, the number of patent citations made, the mean forward citation lag, and the mean backward citation lag. Finally, three dependent variables were selected to include in the learning Bayesian network. This includes one constructed dependent variable, the level of generality, and two dependent variables that will be predicted as an outcome of the learning Bayesian network.

An initial Bayesian network was constructed in the form of a Directed Acyclic Graph (DAG). Figure C-4 demonstrates a Directed Acyclic Graph (DAG) that is used to construct conditional independence relationships between independent variables and dependent variables in the Bayesian network. The DAG included in Figure C-3 was initially constructed to be modeled using the Bayes Net Toolkit [40] developed by Kevin Murphy. This model was not selected for use to build the predictive models and will need to be explored as part of further research. The nodes in the DAG represent continuous and discrete random variables that are constructed using data from the NBER patent database. Construction of this DAG was initially started using the Bayes Net Toolkit, but it was determined that due to the size of the vectors for continuous and discrete random variables the Bayes Net Toolkit will require the use of “Conditional Gaussian Models” that will be much more time consuming to construct. The numbers assigned to each node in the graphical model included in

Figure C-3 are used in the Matlab code to reference the parent and child relationships between nodes in the graph. The DAG is used to generate conditional probability distributions between each node. Using Bayes Rule it is possible to write the joint probability distribution for the "surprise" metric, included and highlighted in blue in Figure 3 4, by the following equation:

$$P(12,10,3,9,5,6,4,2,1) = P(12 \mid 10,3,5,4,6,9) * P(10 \mid 9,6) * P(3) * P(5 \mid 3,4,1) * P(6 \mid 3,4,2,1) * P(9 \mid 5,6,1) * P(4 \mid 3,2) * P(2) * P(1)$$

A joint probability distribution can be constructed for each of the remaining nodes in the graph that represents the probability of occurrence for the node given the probability of occurrence for each of its parents. A number of other joint probability distributions could be constructed using the Bayes Net Toolkit to form predictive models for each of the transdisciplinary metrics.

The next chapter of this dissertation includes the results of this research. It provides detailed data analysis and documentation of the papers submitted to open literature used to validate the contribution of this dissertation. Four predictive models were constructed as part of the research in this dissertation using an artificial neural network. The first model discussed is used to measure the level of invention of a new product or design. In addition, a method is provided for estimating the degree of ideality of a new product using natural language processing. Furthermore, a semantic functional basis of design is constructed using natural language processing and latent semantic analysis. This semantic functional basis of design can be used to measure the use of functional and physical design terms in a new product or design. These terms can then be used to measure the level of design functional synthesis. Other predictive models discussed in this section are constructed using a set of transdisciplinary metrics to predict the level of generality of a new design or product.

The level of generality can be used to predict the breadth of impact a design or product will have on future inventions. Finally, two more predictive models were constructed that use the same set of transdisciplinary metrics used to predict generality to measure the level of forward importance and number of citations received of a new design.

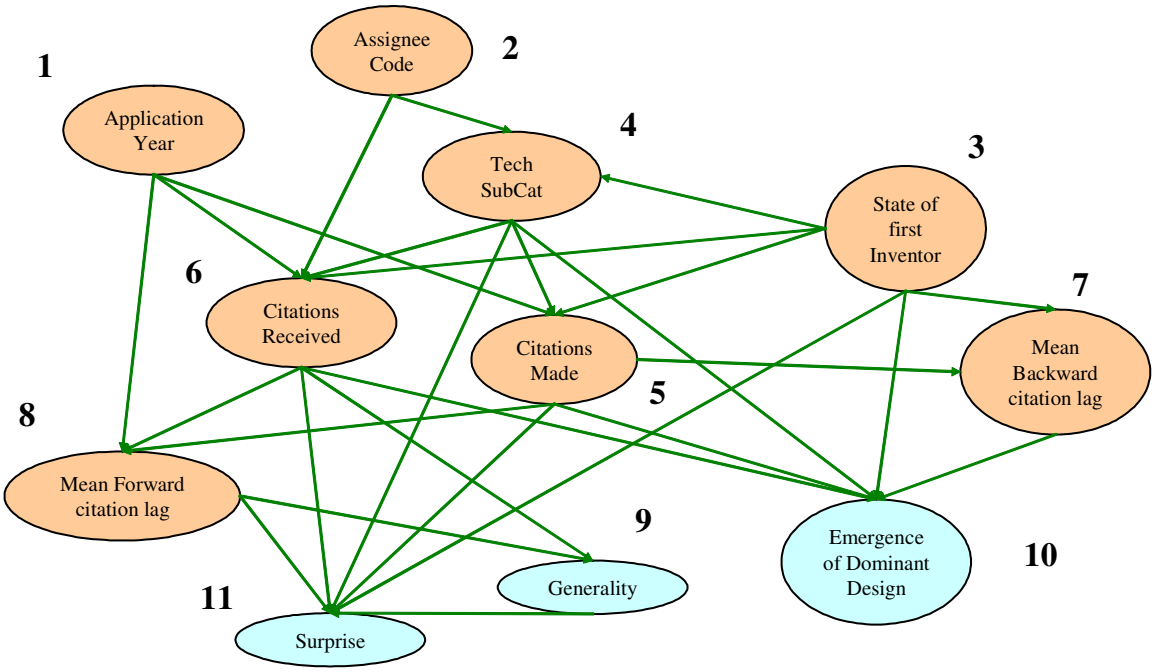


Figure C-3 Bayesian Network Directed Acyclic Graph

Predicting the innovation level of new designs helps inventors explore the possibilities of success when bringing a new product to market. The use of metrics from the field of TRIZ and innovation theory helps designers measure the future success of their products. This section provides a novel computer aided method to measure the emergence of a dominant design and birth of a technological discontinuity by developing a set of innovation metrics using machine learning techniques. A training data set developed using

patent citation data can be used to train a machine learning model employed to measure the growth and death of the number of firms over time within a patent citation tree that represents the evolution of new technological inventions. Patent citations, the number of firms involved in a technological market and the change in physical components used to perform common functions over time serve as independent variables in a machine learning model. The dependent variables in the model are used as predictors of the dominance a new design has in a given market by measuring the death of firms in that market and the technological discontinuity realized by a sudden explosion of firms. This machine learning model is built with variables extracted using natural language processing to measure the change in physical design components. In addition, unsupervised learning such as k-means clustering and supervised learning techniques like neural networks are used to build, train and test machine learning models.

Preliminary results suggest that the emergence of a dominant design was found to be closely correlated with the importance of a patent described by Jaffe and Trajtenberg [23]. What is a dominant design? A dominant design is "a specific path , along an industry's design hierarchy, which establishes dominance among competing design paths." [36] One of the major questions from reviewing patent citation data is whether the emergence of a dominant design can be correlated with the number of patent citations received from other patents. This is in someway correlated with the level of importance explained by Jaffe and Trajtenberg in [23]. Where level of importance, based on the number of forward patent citations, is quantified by the following equation [23]:

$$IMPORTFi = NCITINGi + \lambda$$

Where NCITINGi is the number of patents citing the originating patent and λ is an "arbitrary discount factor" used to reduce the weight given to citations that cite a patent that cites the originating patent. Or in other words, second generation patents. Another factor that plays into the emergence of a dominant design is the number of firms entering and exiting the market for a specific design [36]. Utterback and Suarez suggest that a firm has a higher probability of survival if it enters the market before the emergence of a dominant design. Therefore, a potential measure is created for a dominant design that calculates the number of firms entering the technological market (subcategory) and which are involved with the evolution of a design that has a high level of importance. Thus the emergence of a dominant design can be calculated by the equation below:

$$IMPORTFi + INVENTi *(TECHSUBn) = DOMDESi$$

Where INVENTi is the number of inventing firms in the design market and TECHSUBn represents the number of technological subcategories involved in the market. This metric can be enhanced further to include considerations for the number of firms entering and exiting a technology market. This enhancement helps forecast future emerging dominant designs.

An example dominant design based on mining patent data using exCITE patent citation software [41] demonstrates that a dominant design results in a high number of competing firms developing a similar product with a high number of citations received by the original dominant design. Figure C-4 includes an example of this using a multimedia device that represents the citation network as well as inventing firms growing exponentially upon the emergence of a dominant design. Also shown in Figure C-4 is the patent for an MP3 player design that represents a potential technological discontinuity that leads to the subsequent emergence of a dominant design. The technological discontinuity can be considered as a surprising event [15] in the design world that leads to the evolution of new product ideas. This element of "surprise" [34] can be further modeled as the basis of future research and form the foundation of generating other transdisciplinary metrics related to evolutionary design initiative. The next will include a summary of the conclusions obtained form this research and an overview of future research to be conducted.

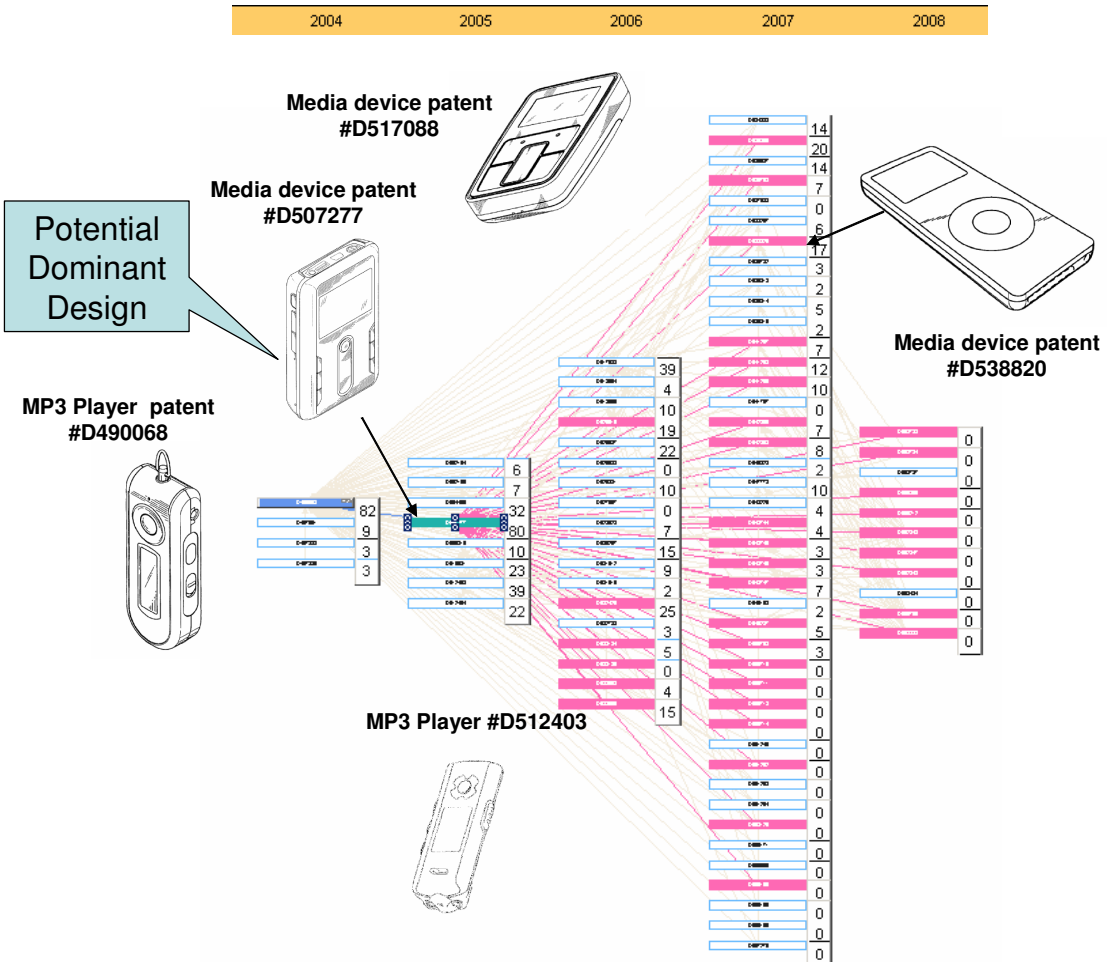


Figure C-4 Example Dominant Design Using Media Device Patents

The results of this research indicate that it is possible to employ a set of transdisciplinary metrics to measure the success of new inventions. This research has put forth a methodology to use in generating a set of transdisciplinary knowledge integration measures with the use of natural language processing and latent semantic analysis techniques. In addition, a machine learning model has been developed using this methodology that will help in the evaluation of new inventions. As part of future research

the author suggests that the hierarchical translation of knowledge between disciplines be studied to understand the benefit of exploring the difference between lower and higher level knowledge transfer. Furthermore, a new set of transdisciplinary metrics should be developed that includes the integration of physical design components and functions to create a set of technology based term constructs that can be used to measure the transdisciplinarity and interdisciplinarity of new designs based on existing technological systems. Technological systems are then employed to measure invention transdisciplinarity instead of just physical and functional items that make up these systems.

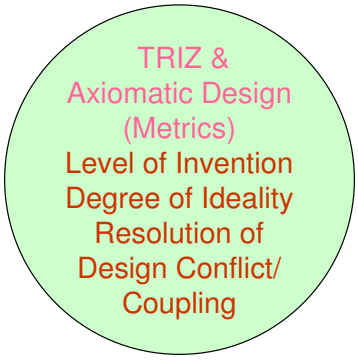
In addition, this research includes an approach to classify functional terms to aid in the process of creating a semantic functional basis for multiple design classes. A semantic functional basis of design can be used in the concept generation phase to develop a metric for design success based on determining the integration of functionality across many domains. For future research the authors suggest that it is possible to classify patents into clusters using a semantic functional basis of design centered on Stone and Woods function and flow definitions. This classification can then be used as a metric during the concept generation phase to understand design concepts from other domains that result in successful design problem solutions. In addition, this research created a method to classify patents into the five level of invention categories from TRIZ. Future research should be conducted to augment this approach using transdisciplinary term metrics, interdisciplinary and transdisciplinary metrics as independent variables in the level of invention classification models. Further research can also be conducted using other sections of the patent text such as patent abstracts,

and claims. This research can be used to further evaluate design solutions during the concept generation phase. Other design metrics should also be explored as part of future research including, but not limited to, using a semantic functional basis of design to extract TRIZ contradiction resolution instances from patent text, and determine the evolutionary potential of a design concept based on the use of functionality from many disciplinary areas. These areas will continue to be explored as part of further research. This will focus on the use of a semantic functional basis to extract function and flow information used to generate successful, creative and innovative design concepts.

APPENDIX D

ENGINEERING DESIGN PROCESS METRICS

The first Transdisciplinary design processes reviewed are TRIZ and Axiomatic Design [1, 23]. The metrics inspired by these two design processes include the five levels of invention, degree of ideality and resolution of contradictions from the TRIZ design process. Next, metrics extracted from the Axiomatic Design process include the resolution of design coupling. A definition of the metrics from TRIZ and Axiomatic Design is included in Figure D-1.

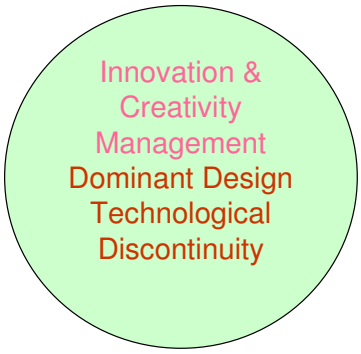


Metric	Measurement Definition
Level of Invention	Level 1 A component intended for the task is used. No system conflicts are resolved. Level 2 Existing system is slightly modified. System conflicts are resolved by the transfer of a solution from a similar system. Level 3 System conflicts are resolved by radically changing or eliminating at least one principal system's component. Solution resides within one engineering discipline. Level 4 System conflicts are resolved and a new system is developed using interdisciplinary approaches. Level 5 Resolving system conflicts results in a pioneering invention (often based on a recently discovered phenomenon.
Degree of Ideality	The benefit-to-cost ratio of the system or the ratio of its functionality, to the sum of various costs associated with the building and functioning of the system.
Resolution of Design Contradiction / Coupling	A physical contradiction is formulated by the pattern: To perform Action A1 the component must have property P, but to perform action A2 the component must have an opposite property P. Axiomatic design relates Design Contradiction to coupling of functional requirements and design parameters.

Figure D-1 Design Process Metrics from TRIZ and Axiomatic Design

Additionally, innovation management theory includes metrics such as the emergence of a dominant design[19] and technological discontinuities[2]. These are two widely

accepted metrics in the innovation community for measuring the success of a new design. A definition of metrics from Innovation and Creativity Management is included in Figure D-2.



Metric	Measurement Definition
Dominant Design	A specific path , along an industry's design hierarchy, which establishes dominance among competing design paths.
Technological Discontinuity	A competence destroying or competence enhancing technology that changes the current technological market.

Figure D-2 Design Process Metrics from Innovation & Creativity Management

Furthermore, functional basis of design provides a definition of commonly used functions and functional flows employed by designs. The number of functions and functional flows employed by a design also provides a useful metric for design evaluation. [4, 5] A definition of metrics from functional basis of design is included in Figure D-3.

Functional Basis of Design
Number of Functions Employed
Number of Functional Flows

Metric	Measurement Definition
Number of Functions Employed	A description of an operation to be performed by a device or artifact, expressed as the active verb of the sub-function.
Number of Functional Flows	Number of quantities that are input and output by functions. A flow is the recipient of a functions operation.

Figure D-3 Design Process Metrics from Functional Basis of Design

Finally, from the field of cognitive psychology metrics such as flexibility [111], originality [112], depth [113], and generality [20] of a design can be used to measure concept success. A definition of metrics from cognitive psychology is included in Figure D-4.

Cognitive Psychology
Flexibility
Originality
Depth
Generality

Metric	Measurement Definition
Flexibility	Degree of diversity in reactions shown under normal conditions and responsiveness to environmental pressures to change.
Originality	Concerned with giving ideas that are statistically unusual
Depth	The extent of which an idea/patent/design is specialized in its own field. $d_j = \frac{\sum_i d_{ij}}{n_j}$ d_j = depth of item j; d_{ij} = depth of item j in field i; n_j = number of total items in field i.
Generality	k is the index of item (patent) classes and N_i is the number of different class to which the citing items (patents) belong. NCITING is the number of patents citing the originating patent. $\text{General}_i = 1 - \sum_{k=1}^{N_i}\left(\frac{NCITING_{ik}}{NCITING_i}\right)$

Figure D-4 Design Process Metrics from Cognitive Psychology

APPENDIX E

MATLAB M FILES

```
%Read in Training Data to train neural network
Td=xlsread('TMfinalc', 'Sheet1', 'G2:BI1302');
general=xlsread('TMfinalc', 'Sheet1', 'CF2:CF1302');
creceive=xlsread('TMfinalc', 'Sheet1', 'CG2:CG1302');
ImportF=xlsread('TMfinalc', 'Sheet1', 'CH2:CH1302');
crecgenimport =xlsread('TMfinalc', 'Sheet1', 'CF2:CH1302');
crecnorm =xlsread('TMfinalc', 'Sheet1', 'CI2:CI1302');
successrate =xlsread('TMfinalc', 'Sheet1', 'CJ2:CJ1302');
factor =xlsread('TMfinalc', 'Sheet1', 'CK2:CK1302');
TdT = Td';
generalT = general';
creceiveT = creceive';
ImportFT = ImportF';
crecgenimportT = crecgenimport';
crecnormT = crecnorm';
successrateT = successrate';
factorT = factor';

%Perform principal components analysis
X=csvread('meas1549.csv', 1, 1);
[pc,SCORE,latent,tsquare] = princomp(X);
csvwrite('meas4CO1.csv', pc);
csvwrite('meas4SC.csv', SCORE);
csvwrite('meas4lat.csv', latent);
```

```
csvwrite('meas4tsq.csv', tsquare);

function [net,ps,ts] = fitwithnet(p,t)
%FITWITHNET Creates and trains a neural network to fit input/target data.
%
%  [NET,PS,TS] = FITWITHNET(P,T) takes:
%    P - RxQ matrix of Q R-element input samples
%    T - SxQ matrix of Q S-element associated target samples
%  arranged as columns, and returns these results:
%    NET - The trained neural network
%    PS - Settings for preprocessing network inputs with MAPMINMAX.
%    TS - Settings for postprocessing network outputs with MAPMINMAX.
%
%  For example, to create an network with this function:
%
%    load housing
%    [net,ps,ts] = fitwithnet(p,t);
%
%  To test the network on the original or new data:
%
%    pn = mapminmax('apply',p,ps); % Preprocess inputs
%    an = sim(net,pn); % Apply network
%    a = mapminmax('reverse',an,ts); % Postprocess outputs
%    e = t - a; % Compare targets and outputs
%
%  To reproduce the results you obtained in NFTOOL:
%
%    [net,ps,ts] = fitwithnet(TdT,generalT);
```

```
% Random Seed for Reproducing NFTool results
rand('seed',5.04820881E8)

% Normalize Inputs and Targets
[normInput,ps] = mapminmax(p);
[normTarget,ts] = mapminmax(t);

% Create Network
numInputs = size(p,1);
numHiddenNeurons = 20;  % Adjust as desired
numOutputs = size(t,1);
net = newff(minmax(normInput),[numHiddenNeurons,numOutputs]);

% Divide up Samples
testPercent = 0.20;  % Adjust as desired
validatePercent = 0.20;  % Adust as desired
[trainSamples,validateSamples,testSamples] =
dividevec(normInput,normTarget,testPercent,validatePercent);

% Train Network
[net,tr] = train(net,trainSamples.P,trainSamples.T,[],[],validateSamples,testSamples);

 % Simulate Network
[normTrainOutput,Pf,Af,E,trainPerf] = sim(net,trainSamples.P,[],[],trainSamples.T);
[normValidateOutput,Pf,Af,E,validatePerf] =
sim(net,validateSamples.P,[],[],validateSamples.T);
[normTestOutput,Pf,Af,E,testPerf] = sim(net,testSamples.P,[],[],testSamples.T);
```

```
% Reverse Normalize Outputs
trainOutput = mapminmax('reverse',normTrainOutput,ts);
validateOutput = mapminmax('reverse',normValidateOutput,ts);
testOutput = mapminmax('reverse',normTestOutput,ts);

% Plot Regression
figure
postreg(( trainOutput,validateOutput,testOutput ), ...
  t(:,trainSamples.indices),t(:,validateSamples.indices),t(:,testSamples.indices)));
```