Winter 2-26-2018

Opportunity Identification for New Product Planning: Ontological Semantic Patent Classification

Farshad Madani
Portland State University

Recommended Citation

Madani, Farshad, "Opportunity Identification for New Product Planning: Ontological Semantic Patent Classification" (2018). *Dissertations and Theses.* Paper 4232.

10.15760/etd.6116

Opportunity Identification for New Product Planning:

Ontological Semantic Patent Classification

by

Farshad Madani

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

Doctor of Philosophy
in
Technology Management

Dissertation Committee:
Charles Maria Weber, Chair
Antonie Jetter
Steven Bedrick
Robert Harmon

Portland State University
2018

Abstract

Intelligence tools have been developed and applied widely in many different areas in engineering, business and management. Many commercialized tools for *business intelligence* are available in the market. However, no practically useful tools for *technology intelligence* are available at this time, and very little academic research in *technology intelligence methods* has been conducted to date.

Patent databases are the most important data source for technology intelligence tools, but patents inherently contain unstructured data. Consequently, extracting text data from patent databases, converting that data to meaningful information and generating useful knowledge from this information become complex tasks. These tasks are currently being performed very ineffectively, inefficiently and unreliably by human experts. This deficiency is particularly vexing in product planning, where awareness of market needs and technological capabilities is critical for identifying opportunities for new products and services. Total nescience of the text of patents, as well as inadequate, unreliable and untimely knowledge derived from these patents, may consequently result in missed opportunities that could lead to severe competitive disadvantage and potentially catastrophic loss of revenue.

The research performed in this dissertation tries to correct the abovementioned deficiency with an approach called patent mining. The research is conducted at Finex, an iron casting company that produces traditional kitchen skillets. To 'mine' pertinent patents, experts in new product development at Finex modeled one ontology for the required *product features* and another for the attributes of requisite metallurgical *enabling technologies* from which new product opportunities for skillets are identified by applying natural language processing, information retrieval, and machine learning (classification) to the text of patents in the USPTO database.

Three main scenarios are examined in my research. Regular classification (RC) relies on keywords that are extracted directly from a group of USPTO patents. Ontological classification (OC) relies on keywords that result from an ontology developed by Finex experts, which is evaluated and improved by a panel of external experts. Ontological semantic classification (OSC) uses these ontological keywords and their synonyms, which are extracted from the WordNet database. For each scenario, I evaluate the performance of three classifiers: k-Nearest Neighbor (k-NN), random forest, and Support Vector Machine (SVM).

My research shows that OSC is the best scenario and SVM is the best classifier for identifying product planning opportunities, because this combination yields the highest score in metrics that are generally used to measure classification performance in machine learning (e.g., ROC-AUC and F–score). My method also significantly outperforms current practice, because I demonstrate in an experiment that neither the experts at Finex nor the panel of external experts are able to search for and judge relevant patents with any degree of effectiveness, efficiency or reliability.

This dissertation provides the rudiments of a theoretical foundation for patent mining, which has yielded a machine learning method that is deployed successfully in a new product planning setting (Finex). Further development of this method could make a significant contribution to management practice by identifying opportunities for new product development that have been missed by the approaches that have been deployed to date.

Dedication

To my wife for her love, patience, and encouragement

Acknowledgements

I would like to acknowledge my committee chair and PhD advisor, Dr. Charles Weber, for his effective leadership and professionalism in advising me on my thesis. Dr. Weber spent many hours patiently discussing the details of every step of the research. He supported me and my family through many difficult times, and gave us hope and encouragement. Dr. Weber showed me how to be a great PhD adviser, how to conduct scientific research, write a thesis, and, most importantly, how to deal sympathetically with people.

It is also with deep gratitude that I acknowledge the support of my committee members: Dr. Antonie Jetter, Dr. Robert Harmon, and Dr. Steven Bedrick, for their valuable time and efforts during the completion of my dissertation. Dr. Jetter added many valuable contributions to my research, suggesting I add ‘patent behavior’ and ‘patent search’ analysis. Dr. Harmon constructively took care of theoretical contributions in the research, and offered productive amendments. Dr. Bedrick directed the research at critical points, especially those involving patent retrieval measurement and machine learning analysis. I appreciate all of my committee members for their contributions to the research and my efforts.

I am grateful to the faculty of the Department of Engineering and Technology Management, particularly Dr. Tugrul Daim, Dr. Ron Khormaei, and Dr. Jisun Kim. Dr. Daim’s classes informed my thinking significantly and contributed a great deal to my PhD. Also, I learned from the professional opportunities provided by Dr. Khormaei.

Finally, I would like thank Dr. Kim who planted the seeds of the main idea of my dissertation in his ‘technology forecasting’ class.

I also want to thank my friend, Yasser Alizadeh, who was the first person who welcomed me to Portland and helped me get set up for living in Portland. I would like to extend my thanks to my friends, Kevin Williamson, Mike Cook, and Richard Sain, and their families, who supported my family with their love. Also, I would like to express my gratitude to my friend, Dr. Meysam Asgari, for his helps and guidelines.

Finally, I would like to express my deepest gratitude to my wife and daughter for their sacrifice and patience during my PhD.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. Research Problem

1.1.1. Technology Intelligence

Success in new product development (NPD) is driven by simultaneously "maximizing the fit with customer needs and minimizing the time to market" (Schilling & Hill, 1998). The firm has to develop product lines on a timescale that is in alignment with the product lines' market windows, and it has to do so cost effectively (Hull, 2004). For this purpose, it has to develop some technologies internally and combine them with technologies that it procures externally (R. G. Cooper, 1979)(R. Cooper, 1987). Technologies that yield critical capabilities tend to be developed internally, whereas complementary, less critical technologies are procured less expensively through the open market (Prahalad & Hamel, 2006).

An increased awareness of which technologies are available externally allows a firm to combine and integrate these technologies more effectively with each other and with those that it has developed internally (Droge, Jayaram, & Vickery, 2004), thereby enhancing its chances to develop a better product line (Zhao, Huo, Selen, & Yeung, 2011). A heightened awareness of external technologies can also serve as a guidepost for new R&D projects (Enkel, Gassmann, & Chesbrough, 2009). Decision makers within the firm, which tend to be engineers and managers that are involved in the NPD process, may be able to specify an R&D project that fills a particular gap in the development of a new product line more precisely, if they become aware of many available, accessible and obtainable complementary technologies (Kogut & Zander, 1992). (These engineers or managers may be a part a variety of different departments such as marketing, finance, manufacturing, supply chain, (Lawrence & Lorsch, 1967).) Due to the need to minimize time to market, this awareness needs to be raised during the product planning phase of the NPD process, when most opportunities for matching available technologies to market demand are identified (Urban, Hauser, & Dholakia, 1987).

In order to enhance their awareness of externally available, accessible and obtainable technologies, product development organizations apply information technologies that scan and monitor the environment of their firm (Brenner, 1996). This approach, called business intelligence consists three main components: market intelligence, competitive intelligence, and technology intelligence, which respectively gather information about customers, competitors and technological opportunities (Kerr, Mortara, Phaal, & Probert, 2006). Technology intelligence is the focus of this dissertation. It extracts and analyzes information from multiple sources such as websites, patent databases and citation indexes. It also incorporates human intelligence that is obtained from experts (Veugelers, Bury, & Viaene, 2010). (See Figure 1)

One of the key aspects of technology intelligence is extracting actionable knowledge from patent data, which serve as a reliable source for tracking technological changes in many industries (Shih, Liu, & Hsu, 2010). Patent data have been available in electronic format in the United State Patent and Trademark Office (USPTO) since 1976, enabling academicians to develop university-internal tools for patent analysis with which they conduct research in technology intelligence (Brockhoff, 1991; Porter, 2005; Yoon, 2008). However, patent analysis does not provide a complete solution for technology intelligence. Some innovations are not patentable; others are not patented even though they could be (Archibugi & Planta, 1996). In addition, the patent records vary significantly from industry to industry. For instance, telecommunications, information technology, pharmaceuticals, biotechnology, chemicals, and automotive are among the most patent-intensive industries (Economics and statistics administration, 2012) (Breitzman & Thomas, 2002). So it is not surprising that Microsoft, HP, AT&T, and Intel are among the recent top ten patent holders among US companies (Intellectual Property Owners Accosiation, 2013).

A patent is inherently designed to protect the rights of its inventors. Thus a patent must explicitly mention the technologies that underlie the inventions in textual and visual form. This makes patents a good source of textual and visual information about a particular technology. Scanning a patent database will therefore provide textual and visual information about a variety of technologies, giving product planners a better overview of the set of complementary technologies that are potentially available.

For patent databases to be helpful in decision making, the information that they provide must be accurate, presented in a comprehensible format and delivered in a timely manner. This can only be done if the users of patent databases have access to capabilities in keyword extraction, pattern recognition and pattern analysis. These crucial aspects of modern text mining have thus become an integral component of decision making, both at the strategic and tactical levels (Bose, 2009).

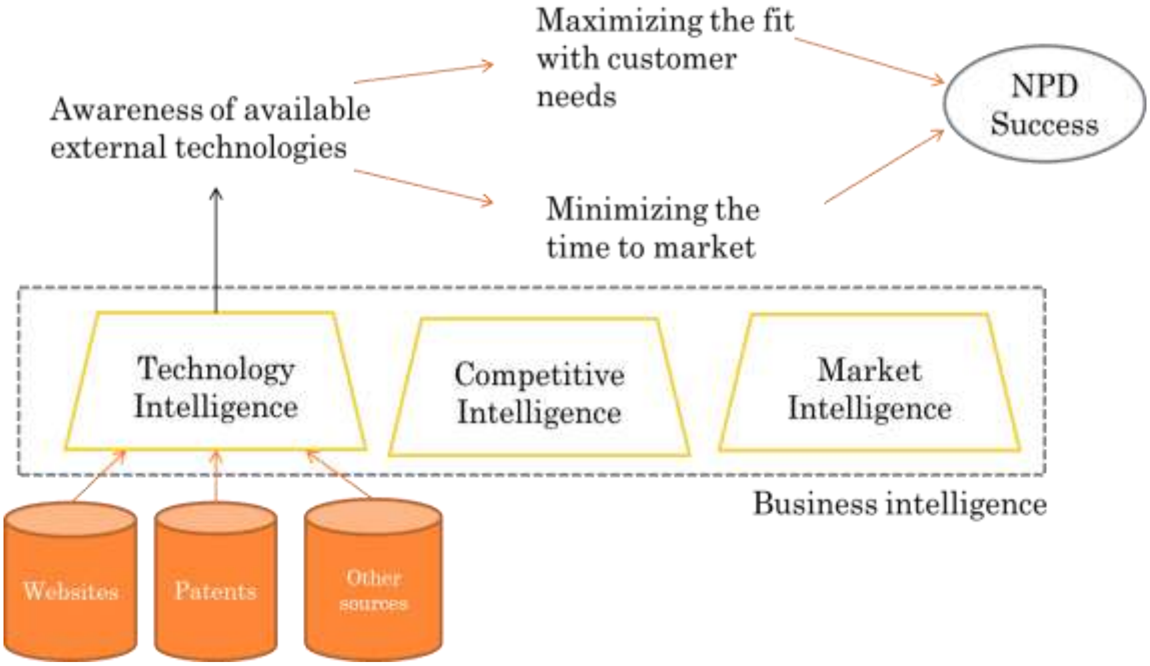


Figure 1- The role of technology intelligence in the success of NPD

Figure 1 summarizes the arguments made in this section. It shows that an awareness of externally available technologies drives the two most critical success factors of NPD, which are maximizing the fit with customer needs and minimizing time to market (Schilling & Hill, 1998). Technology Intelligence is an approach that provides this awareness by analyzing external sources of information such as websites, patent databases, conference papers and articles in academic journals. Patent analysis is a particularly useful aspect of technology intelligence, because patent databases are reliable sources for tracking technological changes in many industries and because patent data are freely available in electronic format.

After extensive marketing analysis, market needs tend to manifest themselves in a set of required product features that make a product attractive to the consumer (Souder, 1988). For successful product planning to take place, this set of features has to be matched up against a set of technologies that could enable the features in the product (Souder, 1988). The question is, how can this best be done efficiently, effectively and comprehensively through technology intelligence and, more specifically, through patent analysis?

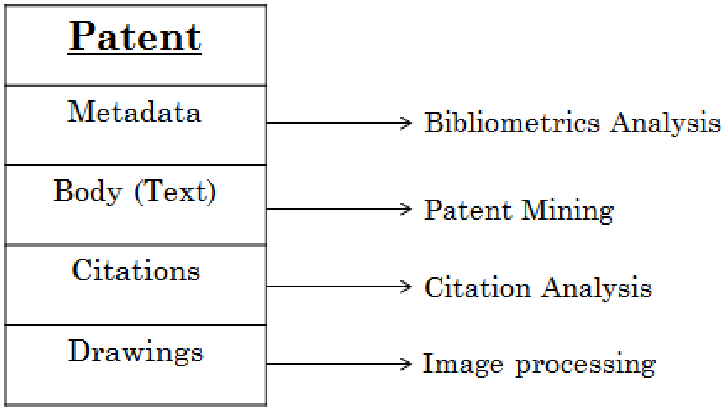


Figure 2- Known approaches to patent analysis

1.1.2. Patent Analysis

Figure 2 shows that patents have four main components: metadata, body text, drawings and citations. Metadata provides general information of including title, assignee, patent number, abstract, etc. The body text presents the technical information, including the background of the invention, a brief summary of the invention, a brief description of upcoming drawings, a detailed description of the invention, and a claim set. The drawings provide a visual representation of the key components of the inventions. Finally, the citations refer to prior art.

Figure 2 illustrates that each of the known approaches to patent analysis focuses on a main component of the patent. Bibliometric analysis looks at metadata. Methods that analyze the main body (or text) of the patent are commonly referred to as patent mining. Image processing techniques find commonalities between drawings “to verify the originality of an invention” (Hanbury, Bhatti, Lupu, & Mörzinger, 2011). Finally, citation analysis identifies the relations between patents by applying network analysis and cluster analysis to the patent’s list of references (or citations).

Historically, patent analysis has consisted of capturing metadata, which provides an analysis of the relationships between individuals, organizations and institutions that comprise an industry or a national/regional innovation system (M Acosta & Coronado, 2003; Melin & Danell, 2000; Naoki Shibata, Kajikawa, & Sakata, 2010). In addition, patent analysis uses citation indexes to identify early and emerging technologies (Karvonen & Kässi, 2013; N. Shibata, Kajikawa, Takeda, Sakata, & Matsushima, 2009; Naoki Shibata, Kajikawa, & Sakata, 2011), to quantify the impact of a particular technology (C Lee, Cho, Seol, & Park, 2012; Madani & Zwick, 2017) and to characterize how knowledge flows, i.e. how technology diffuses through organizations and socio-technical systems (Chang, Lai, & Chang, 2009; Montobbio & Sterzi, 2011; Tijssen, 2001). However, patent analysis that is based on metadata and citations exclusively is highly longitudinal; it does not provide current information (B. Yoon & Park, 2004). For example, a patent that has been granted most likely documents an invention that is at least five years old (G Cascini, Russo, & Zini, 2007). Citation analysis consequently does not allow organizations to respond to rapid shifts in the environment in a timely manner.

1.1.3. Patent Mining

Fortunately, recent advances in text mining have enabled scholars to extract textual information from the content of a patent, not just from its metadata and its citations (Tseng, Lin, & Lin, 2007) (Russo, 2011). This approach, known as patent mining, allows researchers to obtain technical information from patents, which has greatly improved the accuracy of patent analysis (B. Yoon & Park, 2004) (Tseng et al., 2007) (Fattori, Pedrazzi, & Turra, 2003). As a result, patent mining has become very popular since its advent in the late 1990s, and the number of academic publications pertaining to patent mining has been growing exponentially since 2005 (Madani, 2014). Patent mining has been used for different applications such as strategic technology planning (H. Park, Kim, Choi, & Yoon, 2013), technology monitoring (Gerken, 2012), technology roadmapping (S. Choi, Kim, Yoon, Kim, & Lee, 2013), technology trend analysis (S. Choi, Yoon, Kim, & Kim, 2011; Changyong Lee, Jeon, & Park, 2011a; J. Yoon, Choi, & Kim, 2011) and technology acquisition (Jeon, Lee, & Park, 2011).

Current approaches to patent mining consist of broad searches that track the changes of a specific technology (J. Choi & Hwang, 2014b; Changyong Lee, Jeon, & Park, 2011b; Changyong Lee, Park, Kim, & Park, 2011; Ruffaldi, Sani, & Bergamasco, 2010; Scopel, GREGOLIN, & FARIA, 2013; J. Yoon, Choi, & Kim, 2010) or try to find opportunities within a specific industry (S. Lee, Yoon, Lee, & Park, 2009; Thorleuchter &

Van den Poel, 2014). These practices are not well suited for market-pull approaches to new product development planning, which tend to derive the features of a product from data that pertains to a specific market. Instead, the ideal method of patent mining for new product development planning would identify technological opportunities, i.e. matching patented technologies that can be incorporated into products to meet specifically identified, desirable product features.

Unfortunately, an approach to patent mining that matches available patented technologies to product features has not been developed to date. As a consequence, patent mining still relies noticeably on experts to identify and manage the right keywords (Russo & Montecchi, 2011), despite all the advances in text mining. These experts are generally expensive, and there may even be a shortage of experts in specific knowledge domains. Furthermore, different experts may introduce their respective biases into the search process, which could lead to faulty or ambiguous conclusions.

1.1.4. Ontological Semantic Analysis

Ontological semantic analysis (Nirenburg & Raskin, 2004), an approach that integrates ontology design with semantic analysis, reduces reliance on experts and makes patent mining more objective. In ontology design, the nature of a concept is represented as a hierarchical structure of terms (Chandrasekaran, Josephson, & Benjamins, 1999; Gavrilova, Farzan, & Brusilovsky, 2005). Semantic analysis allows searching for all synonyms of these terms (M. L. Murphy, 2003). In ontological semantic analysis, ontology design and semantic analysis are executed in sequence.

Ontological semantic analysis has been applied successfully in identifying patent infringement (H. Park, Yoon, & Kim, 2012) and developing a product design process (A.J.C. Trappey, Trappey, Wu, Liaw, & Zhang, 2013). However, significant challenges to broadly based implementation of this approach remain. For example, it has been observed (Russo, 2014) that two inventors may express the same concept at different levels of detail and that inventors with different backgrounds may use different expressions or different syntax to explain the same concept. Furthermore, an inventor may choose to patent some aspects of his/her invention but keep others as a trade secret. He/she can then withhold critical information from or purposely introduce ambiguous language into the disclosure and the claim. In all cases, keyword searches on the same topic may yield different results (Russo, 2014). A researcher who is looking into the database may thus miss crucial patent information by applying an incomplete or incorrect set of keywords.

An incomplete set of patent data could result in very adverse consequences for a firm that engages in new product development (Quinn, 2017). First and foremost, the firm may not pursue the fastest and most effective approach to developing its product, thereby missing the product's market window. It may also invest in developing the wrong technology, not acquire the best technology or not develop the optimal strategic alliances. Alternatively, the firm may develop technology that already exists, which would lead to a wasteful duplication of effort. Even worse, the firm could be sued because it may have inadvertently encroached on someone else's intellectual property. Finally, patent information may hide the potentially best approach to addressing product features. For example, a firm may decide to develop a product without a very important function, simply because the firm is unaware that the technology for doing so is available at the time.

A novel approach to patent mining that is faster, more concrete and accurate would reduce all these risks and thus substantially benefit the new product development efforts of many firms. It would provide firms with a timely and accurate source of patent information that is organized by functionality. A firm that has access to a patent mining toolkit that has all these capabilities will therefore come much closer to developing the right product for the right market segment at the right time at a much lower cost. Novel approaches to patent mining consequently constitute an area of research that is worth pursuing.

A review of the literature on patent mining and its applications in NPD (Section 2.3 and 2.4) reveals the primary research gap that has motivated this dissertation: *No significant patent mining research that matches required product features with enabling technologies has been conducted to date.* Therefore, no currently available patent mining method can identify opportunities for new product planning.

1.2. Purpose of Dissertation Research

The purpose of this dissertation is to close the primary research gap. This entails developing a patent mining method, which makes R&D engineers and managers who are involved in NPD planning more aware of external technologies that generate opportunities for their specific NPD effort. To achieve the stated purpose of this dissertation, I will conduct an exploratory empirical study that analyzes keywords extracted from U.S. patents, which are provided *electronically* by United States Patent and Trademark Office (USPTO). Figure 3 indicates that the USPTO is one of the two largest patent databases in the world. It has been growing steadily and exponentially since 1990 (USPTO, 2015). Over 615,243 patents were filed within 2014, suggesting that the USPTO comprises a very rich source of technological knowledge. According to a USPTO report ("General Patent Statistics Reports," 2016), 46% to 50% of patents filed between 2001 and 2014 belong to foreign applicants. This shows that many companies, governments and individuals from across the world file their inventions in the USPTO to protect their intellectual property when they want to introduce their products to the highly important US market.

The method to be developed will search the USPTO database for patents that pertain to specific technologies. These technologies potentially *enable product features* that the product needs to exhibit, in order to meet market needs. The essential management question being addressed in this research is: *How can R&D engineers/managers that are engaged in product planning find patents that will provide new technological opportunities?* To be of use to practicing engineers and managers, these patents must be found within a timeframe that allows NPD teams to effectively exploit the markets for the products they are developing.

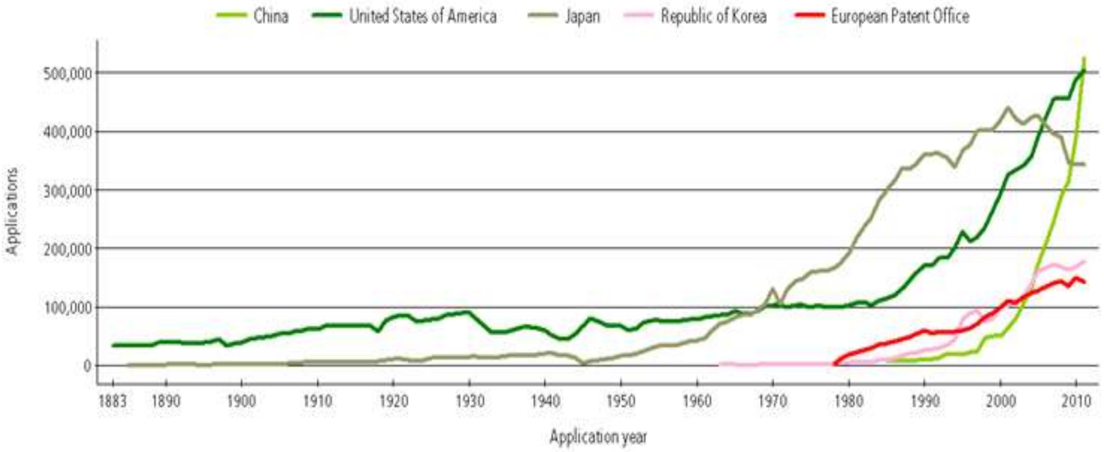


Figure 3- Trend of patent publication in five major international offices

1.3. Gaps in the Literature, Research Questions and Hypotheses

The following steps are required to close the primary research gap. First, you need to generate a conceptual model of the enabling technologies and product features under consideration. This is typically achieved by using ontologies. Second, you need to look for all possible keywords that address the conceptual model and its synonyms. This mandates a semantic analysis of the keywords generated by the ontologies. Finally, you need to determine whether a patent is related to these conceptual models or not. This is achieved by deploying a classification algorithm. The literature review in chapter 2 indicates that none of these steps have been attempted to date. Each of these steps consequently constitutes a sub-gap of the primary research gap, which will be addressed in this dissertation.

Ultimately the goal of the patent mining method to be developed in this dissertation is accurate classification, which the use of ontologies and semantic analysis may assist. For this purpose, I consider a *baseline* scenario in which patents are classified without ontologies or semantic analysis. I call this *regular classification*. I also define the term *ontological classification* as an approach where classification utilizes keywords that are generated by ontologies. I define the term *ontological semantic classification* as an approach where classification relies on ontological semantic analysis—it utilizes keywords that are generated by ontologies, as well as their synonyms, which are generated by semantic analysis.

Under these circumstances, the following research questions must be asked.

Research Question 1 (RQ1): Does applying ontologies to model product feature(s) and technological attribute(s) (ontological classification) lead to a better patent classification than the baseline scenario?

Research Question 2 (RQ2): Does applying ontological semantic analysis to model product feature(s) and technological attribute(s) lead to a better patent classification than the baseline scenario?

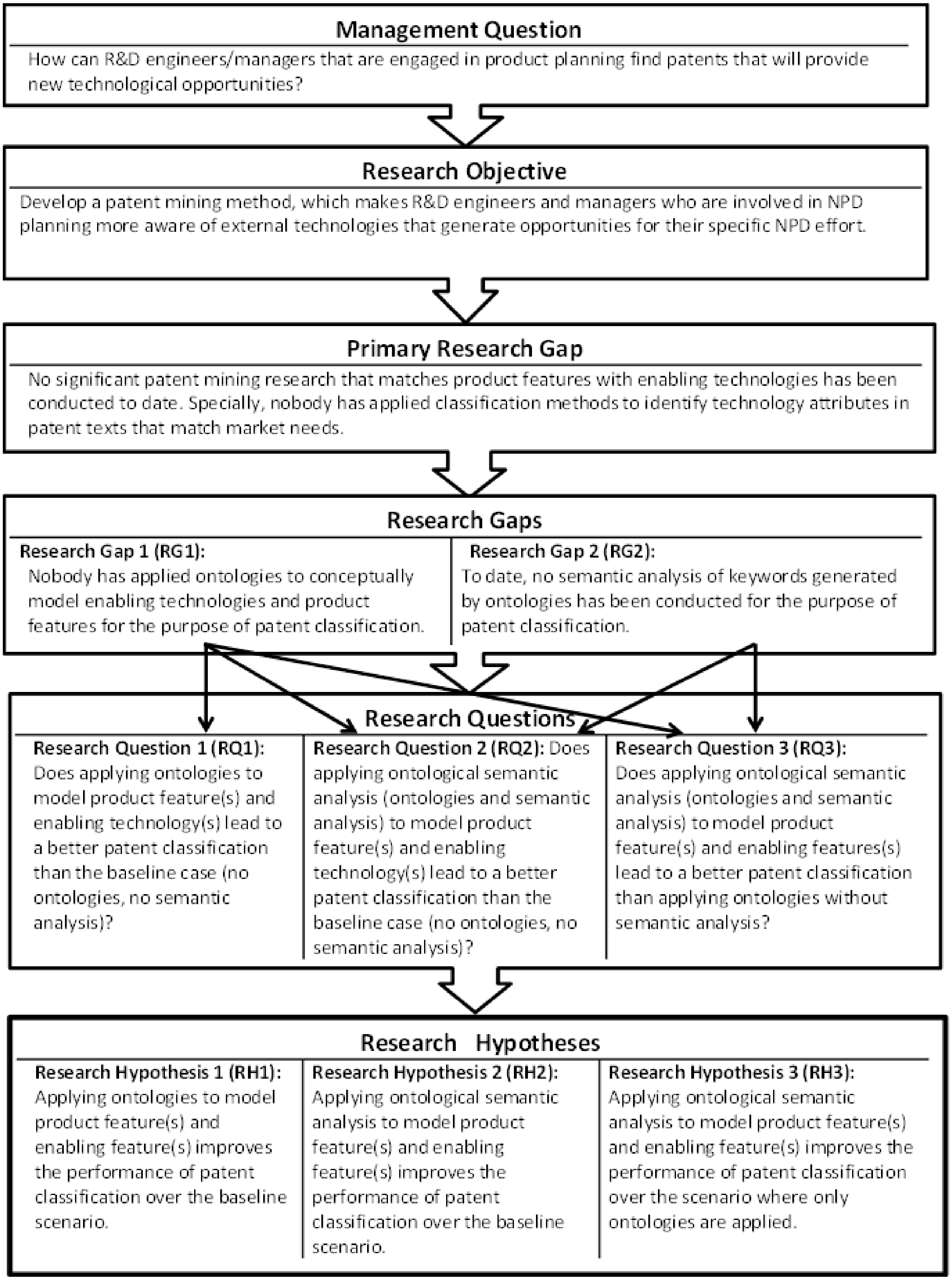
Research Question 3 (RQ3): Does applying ontological semantic analysis to model product feature(s) and technological attribute(s) lead to a better patent classification than applying ontologies without semantic analysis (ontological classification)?

The application of ontologies in text mining has had positive impact on text clustering and classification (Bloehdorn, Cimiano, & Hotho, 2006)(Jing, Zhou, Ng, & Huang, 2006). Also, it is reported that semantic analysis has had positive impact on the performance of different text mining applications such as sentiment analysis (Nasukawa

& Yi, 2003), intellectual property management (W. M. Wang & Cheung, 2011a), patent matching between international classification systems (Y.-L. Chen & Chiu, 2013), and technology monitoring (Gerken, 2012).

Given the potential impact of ontologies and semantic analysis in text mining, the following hypotheses respectively address the research questions:

- **Hypothesis 1 (HP1):** Applying ontologies to model product feature(s) and technological attribute(s) improves the performance of patent classification over the baseline scenario.
- **Hypothesis 2 (HP2):** Applying ontological semantic analysis to model product feature(s) and technological attribute(s) improves the performance of patent classification over the baseline scenario.
- **Hypothesis 3 (HP3):** Applying ontological semantic analysis to model product feature(s) and technological attribute(s) improves the performance of patent classification over the scenario where only ontologies are applied.



Management Question

How can R&D engineers/managers that are engaged in product planning find patents that will provide new technological opportunities?

Research Objective

Develop a patent mining method, which makes R&D engineers and managers who are involved in NPD planning more aware of external technologies that generate opportunities for their specific NPD effort.

Primary Research Gap

No significant patent mining research that matches product features with enabling technologies has been conducted to date. Specially, nobody has applied classification methods to identify technology attributes in patent texts that match market needs.

Research Gaps

Research Gap 1 (RG1): Nobody has applied ontologies to conceptually model enabling technologies and product features for the purpose of patent classification.

Research Gap 2 (RG2): To date, no semantic analysis of keywords generated by ontologies has been conducted for the purpose of patent classification.

Research Questions

Research Question 1 (RQ1): Does applying ontologies to model product feature(s) and enabling technology(s) lead to a better patent classification than the baseline case (no ontologies, no semantic analysis)?

Research Question 2 (RQ2): Does applying ontological semantic analysis (ontologies and semantic analysis) to model product feature(s) and enabling technology(s) lead to a better patent classification than the baseline case (no ontologies, no semantic analysis)?

Research Question 3 (RQ3): Does applying ontological semantic analysis (ontologies and semantic analysis) to model product feature(s) and enabling features(s) lead to a better patent classification than applying ontologies without semantic analysis?

Research Hypotheses

Research Hypothesis 1 (RH1): Applying ontologies to model product feature(s) and enabling feature(s) improves the performance of patent classification over the baseline scenario.

Research Hypothesis 2 (RH2): Applying ontological semantic analysis to model product feature(s) and enabling feature(s) improves the performance of patent classification over the baseline scenario.

Research Hypothesis 3 (RH3): Applying ontological semantic analysis to model product feature(s) and enabling feature(s) improves the performance of patent classification over the scenario where only ontologies are applied.

Figure 4- Management question, research objective, research gaps, research questions and hypotheses

Figure 4 outlines the line of reasoning that underlies my proposed dissertation research. It begins with the management question from which the research objective is derived. The literature review in chapter 2 reveals the primary research gap, which I decompose into two critical sub-gaps. The sub-gaps give rise to the dissertation's research questions. The hypotheses respectively follow the research questions.

1.4 Research Scope

The exploratory empirical study proposed for my dissertation focuses on patent mining and its application to the product planning process within new product development. The later stages of new product development including design, production, and post-production are beyond the scope of this dissertation, since many studies in regard to the application of patent mining for design activities in NPD process have been conducted (section 2.4.2, (Fu, Murphy, et al., 2013; Fu, Chan, Schunn, Cagan, & Kotovsky, 2013a, 2013b; Yan Liang & Liu, 2013; Yan Liang, Liu, Kwong, & Lee, 2012; Yanhong Liang & Tan, 2007; Yanhong Liang, Tan, & Ma, 2008; J. Murphy, Fu, Otto, Yang, et al., 2014; A.J.C. Trappey et al., 2013; P.-A. Verhaegen, D'hondt, Vandevenne, Dewulf, & Duflou, 2011; Paul-Armand Verhaegen, D'hondt, Vandevenne, Dewulf, & Duflou, 2011)) and patent mining is not applicable in production and post-production activities (section 2.4.2). The research will also not cover patent metadata and patent citations because information pertaining to product features and attributes of technologies is contained in the main body of the patent text. Finally, the proposed study analyzes keywords extracted from U.S. patents, which are provided *electronically* by United

States Patent and Trademark Office (USPTO). The research will not consider patents that are not published electronically.

1.5 Overview of Research Method—Ontological Semantic Classification

Figure 5 highlights my research method—ontological semantic classification (OSC). In this approach, natural language processing converts every patent extracted from the USPTO database into a set of keywords. Information retrieval subsequently filters out keywords that are common and keeps keywords that are likely to possess high discriminatory power. In parallel, ontological semantic analysis of data generated through interviews with experts generates keywords that are synonyms of the ontologies of interest. The final classification step identifies patents that possess the desired enabling technologies and those that meet prescribed product features. Technological opportunities are discerned by comparing those two sets of classified patents.

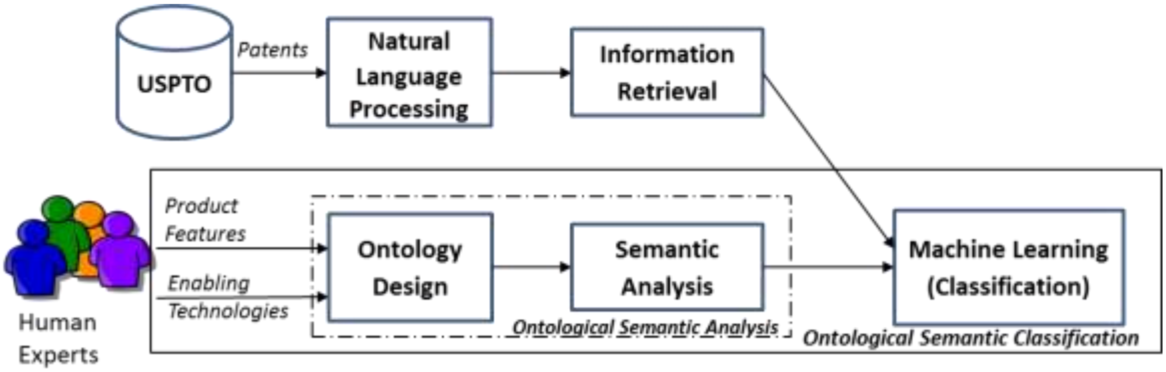

USPTO
Patents
Natural Language Processing
Information Retrieval
Product Features
Enabling Technologies
Human Experts
Ontology Design
Semantic Analysis
Machine Learning (Classification)
Ontological Semantic Analysis
Ontological Semantic Classification

Figure 5- Overview of Ontological Semantic Classification

Separate ontological semantic classifications will be performed on the same patent data set to identify enabling technologies and product features. Patents that are identified in both categories are treated as an opportunity for NPD planning. Applying semantic analysis allows me to consider all possible synonyms for the keywords coming from the ontologies. Applying an appropriate classification method permits me to quickly model the pattern of keywords which address the ontologies. Three classification methods will be evaluated: k-Nearest Neighbor (kNN), Support Vector Machine (SVM) and random forest (Tan, Steinbach, & Kumar, 2006). The output variables of the three classifiers will act as performance criteria. To assess the performance of the classifiers, three measures are often utilized based on a confusion matrix, which is utilized to define the *precision*, *recall* and *F* measures (Tan et al., 2006, p. 297).

Scenario	Ontology	Semantic analysis	Classification	Related Hypothesis		
				HP1	HP2	HP3
I (baseline)			✓	✓	✓	
II	✓		✓	✓		✓
III	✓	✓	✓		✓	✓

Table 1- The three Scenarios of the Research

To assess the performance of ontological semantic classification, I will examine the impact of ontology and semantic analysis in the classification process. To do so, three scenarios, shown in Table 1, are considered. In scenario I, the baseline scenario, the patent would be classified without the application of ontology or semantics. In scenario II, the patents will be classified only by considering the vocabulary presented in the ontology. In scenario III, the patents will be classified by considering the synonyms of the vocabulary presented in the ontology. In each scenario, three classifiers including k-NN, SVM, and random forest will be applied to see how each performs in that scenario. Addressing the research questions and testing the hypotheses from section 1.3 consists of comparing the results of the three scenarios to each other. To examine the hypotheses introduced in Figure 4, the scenarios are pairwise compared.

The study is conducted at Finex, an iron casing company located in Portland, Oregon, which produces traditional iron kitchen skillets. Experts from that company design two ontologies (one for product features and one for enabling technologies), which serve as a basis for ontological classification and ontological semantic classification. The ontologies are validated by a panel of outside experts, which consists of two professors in materials science, two PhD students in materials science, and two industry professionals in related industries. Cross-validation occurs in an iterative process, in which the parameters of the classifiers are tuned to identify the best classifier.

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

The management question that motivates this dissertation is: "*How can R&D engineers/managers that are engaged in product planning find patents that will provide new technological opportunities*?"

In the review of the academic literature that follows, I look at the prior research that has been done, and based on this prior research I identify gaps in knowledge that warrant further scientific study. From these gaps, I shall generate research questions for my dissertation. The major contributions of this dissertation will close the gaps in knowledge that I identify in this chapter, and address the research questions that they generate.

As an introduction to this discussion, I briefly review (in section 2.2) the field of text mining and its constituent disciplines, upon which the field of patent mining depends. The following issues, which are addressed in section 2.3 and 2.4, are of particular interest to practicing technology managers:

1. What are the most recent patent mining methods and their capabilities in patent analysis? (Section 2.3)
2. What are the potential gaps in the application of patent mining in NPD process? (Section 2.4)

I discuss the abovementioned issues in sections 2.3 and 2.4, respectively, identifying the literature streams in which these issues are debated.

2.2. Text Mining – A Brief Overview

Text mining is a variation of data mining that tries to identify valid, novel, potentially useful, and ultimately understandable hidden patterns in large textual databases (Hotho, Nürnberger, & Paaß, 2005). Text mining is an interdisciplinary field which is built upon natural language processing (NLP), information retrieval (IR), and machine learning (ML), which are all deeply rooted in statistics (Gupta & Lehal, 2009). Natural language processing supplies materials (keywords); information retrieval extracts information from the keywords; machine learning recognizes and studies the pattern of keywords based on the information provided by information retrieval (see Figure 6). These three fields of study are briefly introduced in this section.

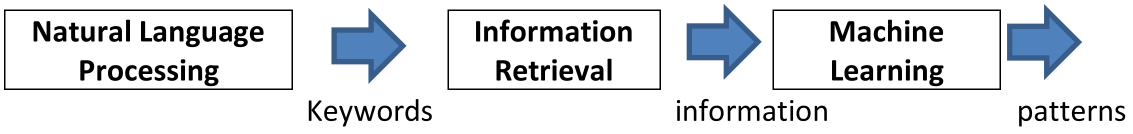


Figure 6- Text mining components

2.2.1. Natural Language Processing

Natural language processing (NLP) is an area of computer science that explores how computers can understand and analyze natural language text or speech (Chowdhury, 2005). The main applications of NLP in text mining are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering (Gupta & Lehal, 2009). NLP consists of the following steps (Nadkarni, Ohno-Machado, & Chapman, 2011), which are executed in the sequential order shown in Figure 7:

- *tokenization*: breaks a stream of text to and identifies individual tokens (words, punctuations). Common words, called *stop words* (e.g., ‘is’, ’at’, ‘which’), are filtered out to improve the performance of NLP.

- *stemming (morphology decomposition)*: reduces words to their roots. For example, "argue", "argued", "argues", "arguing", and "argument" are reduced to the stem "argu”.

- *Part-of-Speech* (POS): classifies and tags words to lexical categories such as nouns, pronouns, adjectives, verbs, adverbs, etc.

- *parsing*: converts a sentence into a tree whose nodes hold POS tags, but the rest of the tree would tell how the words are exactly joined together to make a sentence.

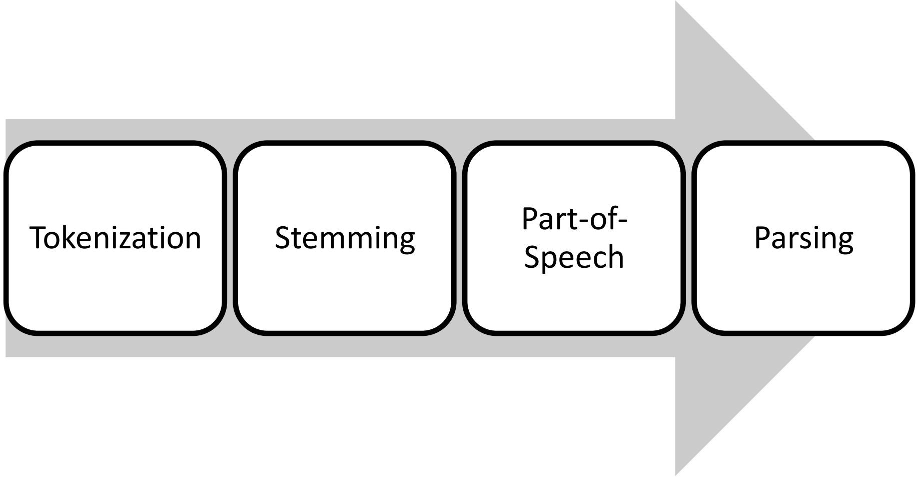


Figure 7- General steps of natural (text) language processing (NLP) (Nadkarni et al., 2011)

2.2.2. Information Retrieval

Information Retrieval (IR) is the area of study which deals with the representation of, storage of, organization of, and access to information of textual sources such as electronic documents and Web pages (Baeza Yates & Ribeiro-Neto, 1999). The process of IR, shown in Figure 8, contains three main sub-processes: 1) representing the content of the document; 2) representing the user’s information need; and 3) comparing of the two representations (Hiemstra, 2009). Most IR systems assign a numeric score to every document and rank them by this score. The three most common research topics in IR are the vector space model, the probabilistic models, and the inference network model (Singhal, 2001). In the vector space model (VSM) (Salton, Wong, & Yang., 1975), a document is represented by a vector of *terms* which are typically words and phrases. The VSM model measures the *similarity* between the query vector and the document vector by applying the cosine of the angle between the two vectors. In the probabilistic models, documents are ranked by decreasing probability of relevance to a particular query (Salton & Michael, 1986). Many probabilistic models have been proposed, each based on a different probability estimation technique (Singhal, 2001). In inference network models, there are two main components: a document network and a query network (Turtle & Croft, 1991). The document network represents the document collection, and the query network represents the user’s information need. The nodes of document network and of the query network are linked by representation concepts and query concepts.

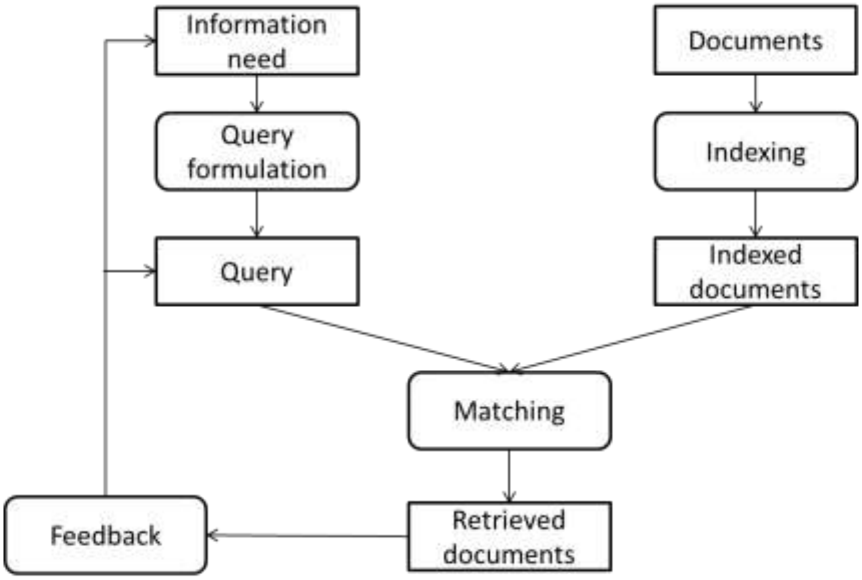


Figure 8- Information retrieval process (Hiemstra, 2009)

2.2.2.1. Information Retrieval Evaluation

Information retrieval systems are developed to serve users, so it is very important to observe and evaluate how users behave and how information retrieval systems perform. Most of the early user studies recognized that the *knowledge* of the subject matter and the level of general search *experience* are the main success factors in information retrieval experiments (Harman, 2011), and *human error* is the main factor in search failures (Lancaster, 1968).

During information retrieval experiment, data collection contains logging *time* spent for different tasks (e.g. query design, document opening, document judgment, etc.) and *facts* (like number of keywords and queries used by each participant or completion time) (Petrelli, 2008). In addition to the quantitative data mentioned before, data collection can contain users' opinions which are qualitatively analyzed through a questionnaire or an interview (Petrelli, 2008).

Efficiency, *effectiveness*, and *user satisfaction* are main measures which are studied in information retrieval experiments (Petrelli, 2008). Efficiency is measured often based upon *time* spent for an information retrieval experiment (Dunlop, 2000). Effectiveness is measured based on the average of *precision* and of *recall*. Users' relevance judgement is, also, another source to measure effectiveness (Dunlop, 2000).

Query formulation is often considered as the main success factor in information retrieval experiment (Petrelli, 2008). Query formulation not only depends on the experience and the knowledge of a user, but also it is affected by user interfaces provided in an information retrieval system (Petrelli, 2008). To assess the success of a user in query formulation, the number of queries used, the number of different terms used, and the average length of queries can be considered as efficiency indicators (Belkin et al., 2003).

The criteria introduced in this section will be applied to evaluate a patent retrieval experiment which is called *patent search* and is introduced in section 3.4.3. The Patent search is designed in order to 1) collect patents judged by experts, and 2) evaluate the performance of experts in patent retrieval. The criteria are customized and introduced in section 4.2.

2.2.3. Machine Learning

Machine learning (ML) is a paradigm to construct and study of algorithms, in order to learn automatically without human intervention. The goal is to do in the future based on what has been experienced in the past. For example, *data mining* is a type of machine learning where patterns are discovered within large volumes of data ("big data") (Zhang & Tsai, 2003). Machine learning methods are grouped to three main categories: 1) supervised, 2) unsupervised and 3) semi-supervised (Mohri, Rostamizadeh, & Talwalkar, 2012) (see Figure 9).

Supervised ML methods, including *regression* and *classification*, use training data with specific labels. For example, 'spam' versus 'not spam' are labels used for spam email filtering. The training process continues until the model achieves a desired level of accuracy with respect to the training data. *Classification* methods assign a category to each item. For example, a document can be classified to different categories such as politics, business, sport, etc. *Regression* predicts a real value for each item. Prediction of stock values or variations in economic variables are examples of regression applications (Mohri et al., 2012).

The only known *unsupervised* ML method is *clustering*, in which data are grouped after patterns are recognized. A clustering algorithm prepares a *deductive structure*[1] in the input data, which, unlike supervised methods, does not have any label. Clustering is often utilized to analyze large data sets to reduce redundancy or organize data by similarity. One of the main applications of clustering is *dimensionality reduction* where many random variables are reduced to fewer. Document frequency, mean TF-IDF (term frequency-inverse document frequency), term frequency variance are the most common dimension reduction techniques used for text clustering (Tang, Shepherd, Milios, & Heywood, 2005). These techniques are categorized as *feature selection*

[1] The deductive structure is a system of thought in which conclusions are justified by means of previously assumed or proved statements.

methods, which select a subset of relevant features from a dataset by removing irrelevant or redundant features (L. Liu, Kang, Yu, & Wang, 2005). The second group of dimensional reduction methods is *feature extraction* or *feature transformation,* which transform the data in a high-dimensional space to a space of fewer dimensions (Tang et al., 2005). Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) are examples of feature extraction methods, which are widely used to cluster text documents (Shafiei et al., 2007).

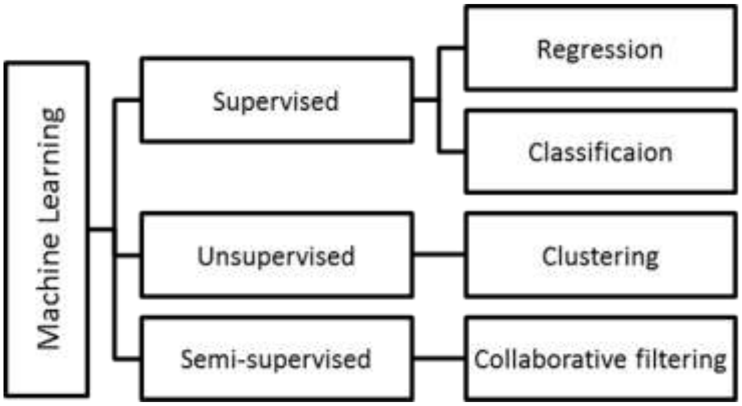


Figure 9- Machine Learning Methods (Mohri et al., 2012)

Semi-supervised ML methods classify large amounts of unlabeled data, under specific assumptions, by applying a training algorithm to a small set of labeled data. Classification occurs by binning the unlabeled data according to the outcome of the training process (Zhu, 2005). Speech recognition, Webpage classification, and genome sequencing are the samples of semi-supervised ML applications (Chapelle, Schölkopf, & Zien, 2006, p. 4).

2.3. Patent Retrieval

Identifying patents for ideation and innovation is one of the most common application of patent retrieval (Bonino, Ciaramella, & Corno, 2010). The main emphasis in most of this kind of tasks is to find all relevant patents, particularly when missing relevant patents is unacceptable (Joho, Azzopardi, & Vanderbauwhede, 2010). As a result, professional patent searchers, like product planners, prefer more functionality in patent search in comparison to occasional users who often require an easy to use interface and simpler commands (Bonino et al., 2010). Specially, after the emergence of the internet, professional users' needs have significantly been changed in patent search (Newton, 2000). While fast response time is among of users' top expectations (McDonald-Maier, 2009), professional users, like product planners, have to deploy iterative search strategies to ensure they have found as many as of the relevant patents as possible (Bonino et al., 2010)(Atkinson & H., 2008). Joho *et al* did a patent search survey and reported that a typical search task takes 12 hours to complete and ranges from a minimum of 3 hours to a maximum of 40 hours (Joho et al., 2010). Also, they reported that such a search task includes roughly 15 queries and a judgment on 100 patents, with each query taking 5 minutes to formulate, while each document takes 5 minutes to judge (Joho et al., 2010).

2.3.1. Patent Retrieval Methods

Query formulation is the key component in every patent retrieval method. Therefore, several methods are developed in order to select, remove, or add keywords in query design. The main patent retrieval methods: 1) keyword-based methods, 2) semantic-based methods, and 3) interactive methods (Fafalios & Tzitzikas, 2017). In *keyword-based methods*, a patent searcher uses keywords that match exactly a target corpus. The weakness of keyword-based method is that they cannot catch patents with similar ideas, but different vocabulary. In *semantic-based methods*, a patent researcher expands queries so that he/she can find relevant patents to the meaning that he/she is looking for. Because neither keyword-based methods nor semantic-based methods show satisfactory performance, interactive methods are proposed (Fafalios & Tzitzikas, 2017). In *interactive methods*, a patent searcher interactively designs queries with reasonable efforts.

Keyword based methods are developed based on three factors:

1) *Targeted Data:* In this approach, to expand or reduce queries, relevant terms are selected based on their position in the body of a patent such as title, abstract, first sentence of the claims, and description (Braschler, Harman, Pianta, & CLEF., 2010; Magdy, Leveling, & Jones, 2010; Mahdabi, Keikha, Gerani, Landoni, & Crestani, 2011; Verberne & D'hondt, 2010).

2) *Term selection/removing approaches:* There are two main approaches in query design: 1) position based approaches, and 2) pattern based approaches. In *position based approaches*, scholars develop queries from different sections of patents. Mahdabi et al. reported that they observed better patent retrieval performance when they constructed their queries based on terms in the *description* section (Mahdabi, Keikha, Gerani, Landoni, & Crestani, 2011). In *pattern based approaches*,

3) *Terms weight calculation:* In addition to frequency-based weighting approaches like calculating tf-idf, some authors developed position-based weighting methods (Magdy, Leveling, & Jones, 2010). They manually assigned different weights to terms according to their positions in a patent.

Semantic based methods are developed as a remedy for the ineffectiveness of keyword based methods. Magdy reported that 12% of the relevant documents don't have common words according to the subject of the research (Magdy et al., 2010). Keyword based methods are ineffective due to vocabulary mismatch between the query and the relevant patents. Semantic based methods are categorized to *dictionary based* methods, and *corpus based* methods (Fafalios & Tzitzikas, 2017). In dictionary based methods, queries are expanded based on terms with similar meaning extracted either from existing generic lexical databases like WordNet or technical databases, or databases generated from patent-related data such as examiner's query log. In corpus based methods, semantically related terms are extracted from a corpus and used for query design.

Interactive methods are developed to increase the performance of patent search activities. Different approaches are applied to interact with and analyze the results of a patent search. The approaches include analyzing classification codes (e.g. IPC, CPC), keyword clustering, semantic enrichment, and others (Fafalios & Tzitzikas, 2017).

2.4. Patent Mining Evolution

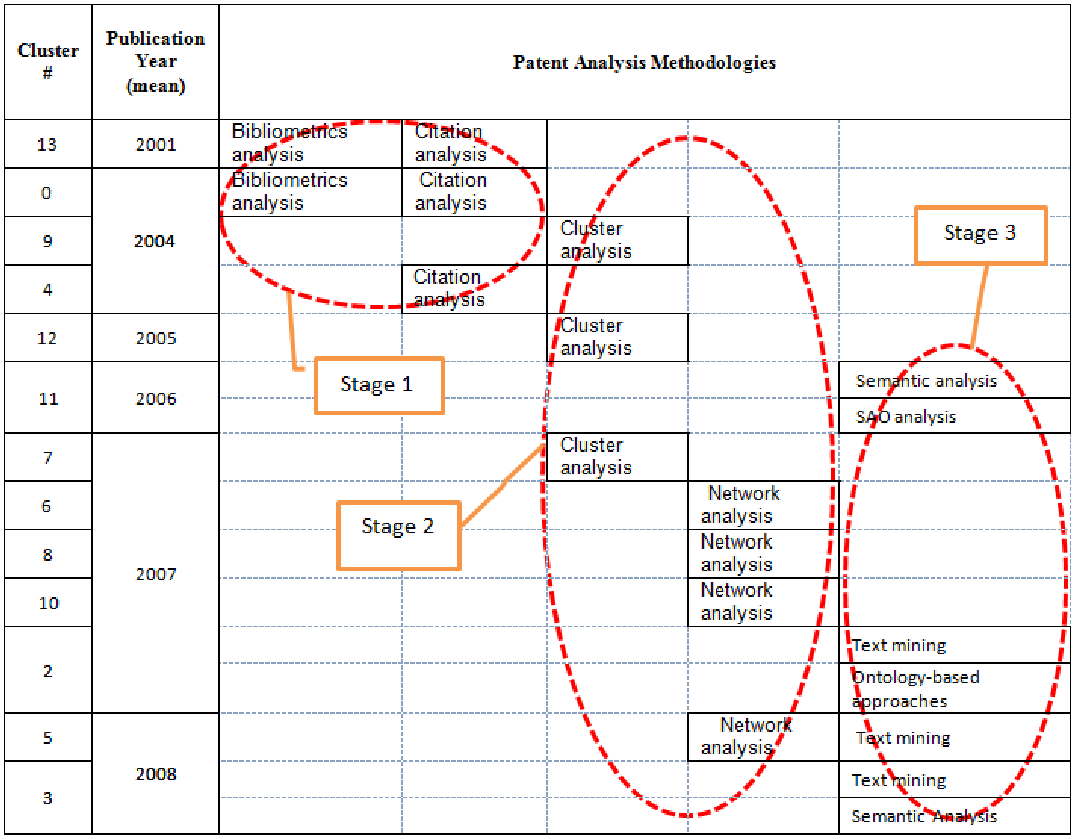
Due to advances in natural language processing, text mining methods and tools have become increasingly available in many different research areas where scholars try to extract useful information and textual patterns from technical documents, particularly patents. This includes technology management. Applying text mining methods to technical documents is named ‘tech mining’ or ‘technology mining’, and for patent analysis purposes, it is named ‘patent mining’. Porter, as one of the pioneers in technology mining, has defined ‘tech mining’ in his book (Porter & Cunningham, 2005, p. 19) as follows: *“the application of text mining tools to science and technology information, informed by understanding of technological innovation processes.”* Therefore, tech mining has two significant characteristics: 1) using ‘text mining tools’, 2) applying these tools to ‘technology management’.

The evolution of patent mining is not precisely understood, because the field has changed so rapidly over the past two decades (Madani & Weber, 2016). For example, it is not clear how scholars are applying methodologies discussed in section 2.3 to expand this research area. Few papers have been published in the field of patent mining; thus, the evolution of the field to its current state of the art is not completely understood. Abbas et al. (Abbas, Zhang, & Khan, 2014) have reviewed 22 published articles that pertain to patent analysis, and they have provided a general taxonomy of techniques that can be deployed in the field. Also, in an editorial note (Chiavetta & Porter, 2013), Porter and Chiavetta investigated six papers published in the proceedings of the first Global Tech Mining (GTM) Conference. They report four main analytics tools which are bibliometrics, data mining, network analysis and cluster analysis. In addition, they reveal eight application areas including emerging technologies and technology dynamics (trend analyses); technology forecasting, roadmapping and foresight; R&D management; engineering industries; science and technology (S&T) indicators; evolutionary economics; technology assessment and impact analysis; as well as science, technology and innovation policy studies. These two articles are based upon expert judgments about the very few papers that have been written in an attempt to explain the evolution of patent mining.

In an attempt to generate a comprehensive literature review, I have deployed a systematic methodology to investigate the majority of patent mining papers that have been published to date. This approach consists of three steps. First, I use bibliometric analysis to recognize the main papers, authors, universities, and journals. I subsequently apply cluster analysis on a keyword network that is extracted from the abstracts of the papers (Madani & Weber, 2016). Finally, CiteSpace (C. Chen, 2014), a free Java application for visualizing and analyzing citations and contents in scientific literature, is applied as the main analysis tool to identify and visualize emerging trends. CiteSpace has been developed by Chaomei Chen, whose research is 'information visualization' (C. Chen, 2004, 2006; C. Chen, Zhang, & Vogeley, 2009). CiteSpace enables me to identify co-citation clusters and trace what research trends have developed (C. Chen et al., 2009). The main techniques implemented in the software are spectral clustering and feature selection algorithms (C. Chen et al., 2009). Visualization of the results is the main characteristic of CiteSpace, which helps more analysts make sense of trends and evolutionary patterns (C. Chen, 2006). Information visualization in this software goes beyond merely visualizing graphical displays—it deploys cognitive, social, and collaborative activities to discover more effectively unstructured data (C. Chen, 2004). More information about the methodology applied in this research and the results of bibliometric analysis are available in my third independent study report (Madani & Weber, 2016). The results of the keyword network analysis are explained in the following separate sub-sections named 1) patent analysis evolution and 2) patent mining evolution.

2.4.1. The Evolution of Patent Analysis

To reveal the evolution of patent analysis methodologies, published articles are mapped with respect to the mean value of the publishing year of each cluster. The map, which is displayed in Figure 10, shows there are three main stages in this evolution. In the first stage, 'bibliometric analysis' and 'citation analysis' are the basic methods used by researchers to discover patterns and gaps in technologies. Different types of patent data are bibliometrically analyzed to examine different purposes particularly for national (Melin & Danell, 2000) or regional studies (Manuel Acosta, Coronado, & Angeles Martinez, 2012) (Montobbio & Sterzi, 2011) (Coronado & Acosta, 2005). For instance, 'forward patent citation' is used to examine the quality of university technology across European regions (Manuel Acosta et al., 2012), or, in another study, citations and co-inventors are represented as channels of knowledge flows from G-5 countries to Latin American countries (Montobbio & Sterzi, 2011).



Cluster #	Publication Year (mean)	Patent Analysis Methodologies				
13	2001	Bibliometrics analysis	Citation analysis			
0	2004	Bibliometrics analysis	Citation analysis			
9				Cluster analysis		
4			Citation analysis			
12	2005			Cluster analysis		
11	2006					Semantic analysis SAO analysis
7	2007			Cluster analysis		
6					Network analysis	
8					Network analysis	
10					Network analysis	
2						Text mining Ontology-based approaches
5	2008				Network analysis	Text mining
3						Text mining Semantic Analysis

- *Dotted area shows the emerging area of patent mining methods*

Figure 10- Map of the evolution of patent analysis methodologies (Madani & Weber, 2016)

In stage 2, clusters 7, 9, 12, and clusters 6, 8, 10, and 5 represent how patent analysis has evolved by applying cluster analysis and network analysis to provide more complicated analysis. Cluster analysis groups patents into similar categories whereas network analysis studies the structure of patent networks or citations networks.

Cluster analysis helps focus on more specific groups or classes of patents recognized in network analysis. This helps researchers conduct technology trend analysis and technology forecasting more efficiently.

Reviewing the key phrases of clusters 7, 9, and 12 suggests that the majority of the researchers utilize keywords extracted from patents to cluster patents based on their content. For instance, Trappey and colleagues clustered key phrases to group patents defining key innovations (C. V. Trappey, Trappey, & Wu, 2010), and to forecast RFID technologies (C. V. Trappey, Wu, Taghaboni-Dutta, & Trappey, 2011). In other research, Jun et al applied a matrix map and the K-medoids clustering method for vacant technology forecasting (Jun, Park, & Jang, 2012). In addition to keywords, citations are another facet of patents used for cluster analysis. For instance, Lee et al applied network analysis and cluster analysis on patent citations to explore technology evolution in electrical conducting polymer nanocomposite (P.-C. Lee, Su, & Wu, 2010a).

Reviewing the key phrases of clusters 6, 8, 10, and 5 suggests that network analysis has enabled scholars to discover the relation between patents and to interpret the content of patents more deeply and efficiently. In the majority of this type of research, patent citation is the most commonly deployed aspect of network analysis. It is often referred to as patent citation network analysis.

Patent citation network analysis is applied for different reasons. For instance, it is applied to analyze technology trends (P.-C. Lee, Su, & Wu, 2010b) (P.-C. Lee, Su, & Wu, 2010a) (J. Yoon et al., 2011), to detect emerging knowledge domains (Naoki Shibata, Kajikawa, Takeda, & Matsushima, 2008) (Kajikawa, Yoshikawa, Takeda, & Matsushima, 2008), to characterize the structure of research in a field of study (Kajikawa & Takeda, 2008), to explore technology diffusion (Chang et al., 2009), and to analyze other issues in the technology management field including technology identification (Shin & Park, 2007) and technology transfer (Y. Park & Lee, 2012). In addition to citations, other aspects of patents are used for network analysis. For instance, ‘co-inventors network’ is used to explore knowledge spillover in Latin America (Montobbio & Sterzi, 2011), or a ‘research grants network’ is utilized to analyze interdisciplinary research relationships (Yang, Park, & Heo, 2010).

In the third stage, where patent mining has emerged, reviewing key phrases of clusters 2, 3, 5 and 11 in Figure 10 discloses that text mining has enabled researchers to gain access to technical information in the patents’ content sections by extracting keywords and by extracting the latent knowledge contained in patents through the application of complementary methodologies, i.e. semantic analysis and ontology-based approaches.

Since patent mining is the focal point of this section, stage three is scrutinized and discussed independently in section 2.4.2.

2.4.2. The Evolution of Patent Mining

In the third wave of the evolution of patent analysis, scholars noticed that just relying on and analyzing citations and the other bibliographic aspects is not enough. There is a huge amount of knowledge and information in patent content section that had not been considered in prior analyses (G Cascini et al., 2007). As text mining methods progressed, scholars began developing content-based approaches (H. Park et al., 2013) by applying text mining methods to extract knowledge and information from patent content (Porter & Cunningham, 2005). This movement, known as ‘patent mining’, is progressing as scholars struggle to create synergies from applying text mining methods and other analytical methods such as network analysis and cluster analysis in conjunction, in order to develop more efficient patent mining methods. Reviewing clusters 2, 3, 5 and 11 in Figure 10 reveals how patent mining methods have been developed and applied over recent years. Figure 11 illustrates how patent mining has evolved over the last two decades, which will be discussed in detail in the next sections.

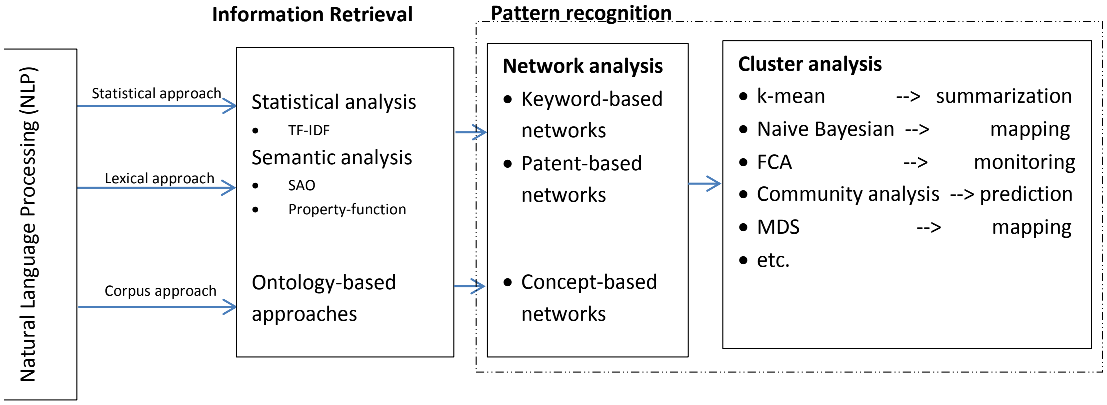


Figure 11- Patent mining evolution (Madani & Weber, 2016)

2.4.2.1. Information Retrieval

There are two ways to extract text content more accurately and more efficiently: 1) applying lexical approaches and 2) applying corpus approaches.[2] In a lexical approach, natural language processing capabilities including syntax tagging, word stemming, and stop-word elimination allow us to distinguish words in sentences based on their syntactic features. Since lexical approaches recognize semantic patterns, they can mine

[2] Basically, there are three main approaches to extract keywords from texts: 1) corpus approach, 2) lexical approach, and 3) statistical approach. In the corpus approach, it is often required to have a dictionary or predefined corpora developed by subject matter experts (SME's). In the lexical approach, natural language processing (NLP) is utilized to find out semantic relations among the keywords since it assumes the relation between keywords and semantic context of documents determines important keywords. In statistical approach deems term frequency is a proxy of the importance of a keyword. Inverse Document Frequency (TF-IDF) method is one of the first developed and broadly applied statistical method (A. J. C. Trappey et al., 2009).

patents more accurately and determine important keywords. Therefore, in comparison to traditional methods like TF-IDF, the scholars don't lose synonyms or polysemy and thereby extract all important keywords. In the corpus approach, experts provide a predefined collection of main concepts addressing the text content. These collections are made in either unstructured forms like dictionaries or in structured forms such as ontologies or morphologies. These collections create a more efficient keyword extraction and content analysis in later stages of analysis.

Natural language processing (NLP) is able to recognize 'lexical' and 'semantic' relation between words. This capability enables the scholars to follow two different strategies for pattern recognition: 1) semantic analysis, and 2) ontology-based analysis. In semantic analysis, important keywords and their relationships, and semantic patterns are recognized. In other words, semantic patterns are recognized based on the meaning of words and their roles in a sentence. The Subject-Action-Object (SAO) approach and its peer approach, the property-function approach, are two semantic analytical approaches developed to extract textual patterns for purposes of patent analysis. A SAO structure is composed of Subject (noun phrase), Action (verb phrase), and Object (noun phrase) that can be extracted by using natural language processing (NLP) of textual patent information (Gaetano Cascini, Fantechi, & Spinicci, 2004). For example, a sample SAO structure such as 'fire ignites oil' comprises the subject ('fire'), the action ('ignite'), and the object ('oil'). Similarly, in the property-function approach, the property is an adjective describing a specific character of a product and the function is a verb referring to an action of the product (J. Yoon & Kim, 2012). Numerous studies can be fulfilled by applying SAO or property-function approaches for different purposes such as technology roadmapping (S. Choi et al., 2013), technology trend identification (S. Choi et al., 2011) (J. Yoon et al., 2011), technology monitoring (Gerken, 2012), and strategic planning (H. Park et al., 2013). Regardless of SAO and property-function approaches, some scholars try to find syntactic patterns by applying heuristic algorithms. For instance, Wang and Cheung (W. M. Wang & Cheung, 2011b) have developed a method containing heuristics rules to detect simple syntactic patterns. This method enables users to search for patents related to a potentially new invention and to provide the relationship and patterns among a group of patents.

In ontology-based analysis, the concepts of patent content are modeled based on their properties, relationships, constraints and behavior (Noy & McGuinness, 2001). Therefore, to extract terms used for the same concept (Bermudez-Edo, Noguera, Hurtado-Torres, Hurtado, & Garrido, 2013), it is required to deploy NLP methods. Ontology-based analysis provides a framework for interacting with application systems, improving the communication model between humans and machines (Weng & Chang, 2008), and providing information with a knowledge domain (P.-C. Lee, Su, & Chan, 2010). Amy Trappey et al (A. J. C. Trappey, Trappey, & Wu, 2009) examined the performance of an ontology-based approach combined with TF-IDF approach in terms of compression ratio,[3] retention ratio,[4] and classification accuracy of the summarization results. The authors figured out the ontology based approach doesn't provide significant

[3] The compression ratio indicates the text reduction from the original document to the compressed summary.

[4] The retention ration indicates how much the information of the original document is available in the summarized document.

improvement in the compression ratio, but it does produce an 11% improvement for the retention ratio and a 14% improvement for the classification accuracy. In another research, Amy Trappey et al (Amy J. C. Trappey, Trappey, Chiang, & Huang, 2013) have developed a knowledge management approach and applied an ontology-based artificial neural network (ANN) to search and classify patent corpora. They combined term frequencies and the concept probabilities of key phrases as the ANN inputs. They produced significant improvement in classification accuracy.

2.4.2.2. Pattern recognition and analysis

After given keywords have been extracted and keyword vectors have been prepared, it is time to process keyword vector elements to provide information. In doing so, network analysis and cluster analysis are the two main methodologies that are applied. To process keyword vectors, it is necessary to take these steps: 1) calculate a similarity function and preparing a similarity matrix, 2) transform the similarity matrix into an adjacency matrix by applying a cut-off threshold value, 3) applying basic network analysis measures, and 4) applying cluster analysis algorithms.

In pattern analysis, it is not necessary to do all four of these steps. Some scholars only apply similarity calculations (step 1) to do an analysis. For instance, Jeon and colleagues (Jeon et al., 2011) only applied similarity calculations to search potential partners for collaboration purposes in open innovation, or Chen et al (Y.-L. Chen & Chiu, 2013) applied similarities for cross-language patent matching. Some scholars apply only basic network analysis, steps 1 to 3. For instance, Choi et al (S. Choi et al., 2013) applied basic network analysis measures to develop roadmaps. And finally, in the most advanced analysis, some scholars apply all of the four steps to create a cluster analysis. For example, Yoon and Kim (J. Yoon & Kim, 2011) applied a clustering method, *k* nearest neighbors (*k*-NNs), in their methodology to identify technological opportunities in the network.

2.4.2.2.1. Network analysis

Network analysis allows researchers to create a set of connected nodes with shared properties and to analyze them based on their network structure and their relationships. There are three types of networks applicable in patent mining: 1) patents-based networks, 2) keyword-based networks, and 3) concept-based networks. The nodes of the networks are patents, keywords, and concepts, respectively, and the relationships are created based on how similar the nodes are. In keyword-based networks, keywords are actors that are connected to a network and their relationships are specified, if a word-pair has been repeated in an extracted sentence or in a semantic pattern. Building a patent-based network requires two steps: 1) determine the similarity between the patents based upon their keyword vectors, and 2) convert the similarities to 0 or 1 by applying a pre-determined cut-off value. There are different approaches to determining similarity: 1) syntactical and 2) semantic. In a syntactical approach, a network is built based on the role of an extracted keyword in their related sentences. For instance, Choi et al (S. Choi et al., 2013) created a keyword-based network, named Product-Function-Technology (PFT) map, based on the SAO approach. They used network analysis to show how products and technologies are related to functions in order to develop a roadmap. Also, they utilized degree analysis to determine the technological trend, and used centrality measure to illustrate how a core function changes over time. Similarly, Choi et el applied this approach to technology trend identification (S. Choi et al., 2011)(J. Yoon et al., 2011). It is worth explaining that Choi et al didn't use a cut-off value in their research in order to obtain all possible keywords in the network. In a semantic approach, semantic similarity between patents is computed based on the similarity of the pairs of words. Tokenizing, stemming, tagging, and determining synonyms are the main steps to figuring out the words (J. Yoon & Kim, 2011). As mentioned above, the similarity matrix is converted to an adjacency matrix by applying a cut-off value to produce the relationships in the network. However, a cut-off value is a task-based and case-dependent variable; Lee et al (P.-C. Lee, Su, & Chan, 2010) suggested an empirical method to optimize the cut-off threshold value of similarity. In concept-based networks, concepts are extracted based on an ontology created by domain experts. Like patent-based networks, it is necessary to extract keywords addressing the concepts and to apply a similarity measurement in order to develop relationships. For instance, Amy Trappey et al (A. J. C. Trappey et al., 2009) used a combined ontology based and TF-IDF concept clustering approach to extract, cluster, and integrate the content of a patent to derive a summary and a cluster tree diagram of key terms.

2.4.2.2.2. Cluster analysis

Given similarity values, cluster analysis is the best methodology to figure out groups of patents, keywords, or concepts that are similar to each other but different from others in other groups. Cluster analysis contains various algorithms with significantly different views of what makes a cluster and how to efficiently catch a cluster. Over recent years, the scholars have tried to apply different clustering methods for various purposes in patent mining. For example, ‘k-mean algorithm’ for patent summarization (A. J. C. Trappey et al., 2009) (Amy J. C. Trappey & Trappey, 2008), ‘Naive Bayesian algorithm’ for patent mapping (W. M. Wang & Cheung, 2011a), ‘Formal Concept Analysis’ (FCA) for technology monitoring (Changyong Lee, Jeon, et al., 2011a), ‘community structure analysis’ for technology prediction (Naoki Shibata et al., 2008) (J. Choi & Hwang, 2014a), ‘Multi-Dimensional Scaling’ (MDS) for patent mapping (J. Yoon & Kim, 2011), etc.

2.4.3. Summary of Patent Mining Literature

Patent analysis has evolved over three main stages, as shown in Table 2. In the first stage, bibliometric analysis and citation analysis were the main methods applied for patent analysis. By applying network analysis and cluster analysis, more advanced bibliometric analysis and citation analysis emerged in the second stage. By advancements in natural language processing and text mining tools, patent mining methods are disclosed in the third stage.

Stage 1	Stage 2	Stage 3
Bibliometric Analysis & Citation Analysis	More advanced bibliometric analysis and citation analysis by applying network analysis and cluster analysis	Emerging patent mining methods

Table 2- Main stages of patent analysis

Network analysis and cluster analysis are still the main analysis methods in patent mining, as shown in Figure 11. This suggests the researchers intend to analyze the relation between patents by applying network analysis, and they intend to cluster specific groups of information by applying clustering analysis. As shown in Figure 11, network analysis and cluster analysis have been applied in technology management areas such as technology mapping, technology monitoring, etc., while there are some other areas such as technology acquisition, and new product planning that they need to *identify* specific technologies. Classification is the method that can remedy this need, as shown in Table 3. Therefore, I will apply classification methods in my research to identify pre-determined enabling technologies and product features for NPD planning purpose.

Type of Analysis	Network Analysis	Cluster Analysis	Classification
Purpose of analysis	Relation analysis	Grouping/clustering	Identification

Table 3- The purpose of three main analysis methods (network analysis, cluster analysis, and classification)

2.5. Application of Patent Mining in NPD Process

This section reviews the most common general NPD processes, citing several prestigious sources. The intent is to identify a consensus of what the main stages of the NPD process actually are (section 2.5.1). I subsequently discuss how patent mining research has been applied in each of these stages (section 2.5.2), which reveals potential gaps in the academic literature that pertain to applications of patent mining to the NPD process.

2.5.1. A review on NPD process

To review NPD process models, three prestigious sources are chosen and discussed in this section (Urban et al., 1987) (Ulrich & Eppinger, 2011) (Pahl, Beitz, Feldhusen, & Grote, 2007). Basically, NPD process activities can be classified into four main activities which are planning, design, production, and post-production. As illustrated in Table 4, Urban et al (Urban et al., 1987) have mainly concentrated on the planning aspects of NPD process, and they didn't get involved in 'design, 'production', and 'post-production' activities. Ulrich and Eppinger (Ulrich & Eppinger, 2011) have considered design and production, in addition to planning. Pahl et al (Pahl et al., 2007) have extended the NPD process and taken some 'post-production' activities such as recycling and energy recovery into consideration.

Source	Planning	Design	Production	Post-production
(Urban et al., 1987)	• Opportunity identification • Consumer management • Models of consumers	-----	-----	-----
(Ulrich & Eppinger, 2011)	Planning	• Concept development • System level Design • Detail design	• Testing and refinement • Production ramp-up	-----
(Pahl et al., 2007)	Product planning /task setting	Design/ development	Production/ assembly/ test	• Marketing/ consulting/ sales • Use/ consumption/ maintenance • Energy recovery • Recycling

Table 4- The NPD processes comparison

2.5.1.1. NPD Process by Urban et al.

Urban, et al. (Urban, Hauser, & Dholakia, 1987) suggest a five-stage product development process, which is shown in Figure 12. The product design process consists of two sub-processes, one managerial and one consumer. The managerial sub-process identifies the main categories of managerial decisions. The consumer sub-process illustrates how a company goes into the market to design the product. The main steps are described in turn.

- *Opportunity identification* is an effort that integrates technological opportunities and market demand. Customer needs and user solutions are the most important sources of ideas. Therefore, recognizing new technologies is a main function since it helps to find new opportunities to meet customer needs. In addition to internal sources such as R&D, engineering, production, and marketing, external sources including patents are very important to yield new ideas. To generate good technology-based ideas, companies should build their new product strategy based on both customer needs and technological advances.
- *Consumer management*: Emphasis in the beginning is the understanding of the consumers. Quantitative measurements are applied to clear consumers' behavior and insight to the market. Qualitative researches are employed to answer specific questions via surveys.
- *Models of consumers*: The models diagnose the market based on the awareness perception, preference, segmentation, availability, and choice factors. They provide the features of the product and direct the design process. Perception models explain what consumers consider to perceive the product, and preference

models illuminate how they evaluate the perceived dimensions. Segmentation determines the best strategy for allocating product(s) to different group of consumers. Choice models determine external events that impact consumers' purchase or use. The four models help managers understand the consumers.

- *Prediction of market behavior*: The four models of consumer response are combined to predict market behavior.

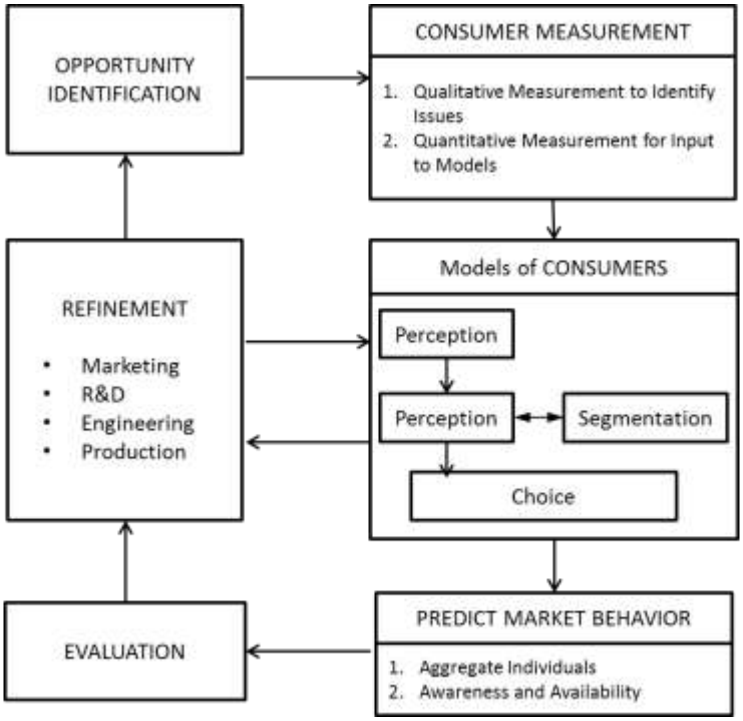


Figure 12- product design process (Urban et al., 1987, p. 26)

2.5.1.2. NPD Process by Ulrich and Eppinger

Ulrich and Eppinger introduce a generic product development process which is like a funnel that starts with a set of alternative product concepts and then narrows and specifies specifications of the product. The stages are shown in Figure 13 and described below:

1- **Planning:** Planning starts with 'opportunity identification' guided by corporate strategy and includes assessment of technology developments and market goals. The main output of this phase is the 'project mission statement', which specifies the target market, business goals, key assumptions, and constraints.

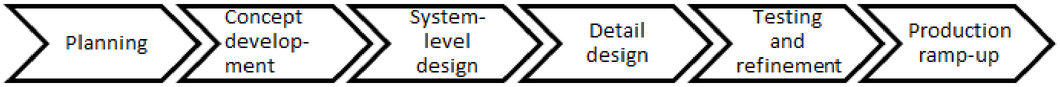


Figure 13- The product development process (Ulrich & Eppinger, 2011, p. 9)

2- **Concept development:** After identifying the needs of the target markets, several alternative product concepts are generated and evaluated to select one or more concepts for further development. The concept is a description of the form, function, and features of the product, and it is usually supported by more competitive and economic analysis.

3- **System level design:** The product architecture and its sub-systems and components are defined and preliminarily designed. The main outputs are a geometric layout, a functional specification of each sub-system, and a preliminary flow process diagram.

4- **Detail design:** All detailed technical aspects including specifications, materials, tolerances, standards, tooling, etc. are specified.

5- **Testing and refinement:** In order to examine the reliability and performance of the product, several prototypes are produced and tested from manufacturing and customer point of views.

6- **Production ramp-up:** To train the workforce and to solve the remaining manufacturing problems, products are manufactured and evaluated carefully. This phase gradually transits to ongoing production.

2.5.1.3. NPD Process by Pahl et al.

Pahl et al. mention a general new product development process which is called 'life cycle of a product' (see Figure 14). In contrast to the other NPD processes under study, the authors consider environmental and recycling phases. This explains why these authors apply the term 'life cycle' to the NPD process. Nonetheless, the process, like many others, contains general phases including product planning, design, production, marketing, and use/maintenance. Since my research relates to product planning, I will explain this phase in more detail.

The product planning process consists of a sequence of steps (Pahl et al., 2007, Chapter 3). In the first step, analyzing the situation, several aspects are considered and verified. The aspects are product life cycle, product-market matrix, company's competence, status of technology, and future developments. In life cycle analysis, product diversification is the focal point of decisions. It can lead to phased development and the sale of different products. The product-market matrix helps understand the status of existing products from the company and from competitors in the various markets. In company's competence assessment, the current market position is compared with competitors based on factors such as turnover, market share, market situation, etc. To determine the status of technology, the products of the company, related technologies, concepts, and products are reviewed in the literature and patents, and competitors' products. Future developments are estimated based on knowledge of future projects, expected customer behavior, technical trends, environmental requirements, and the results of fundamental research.

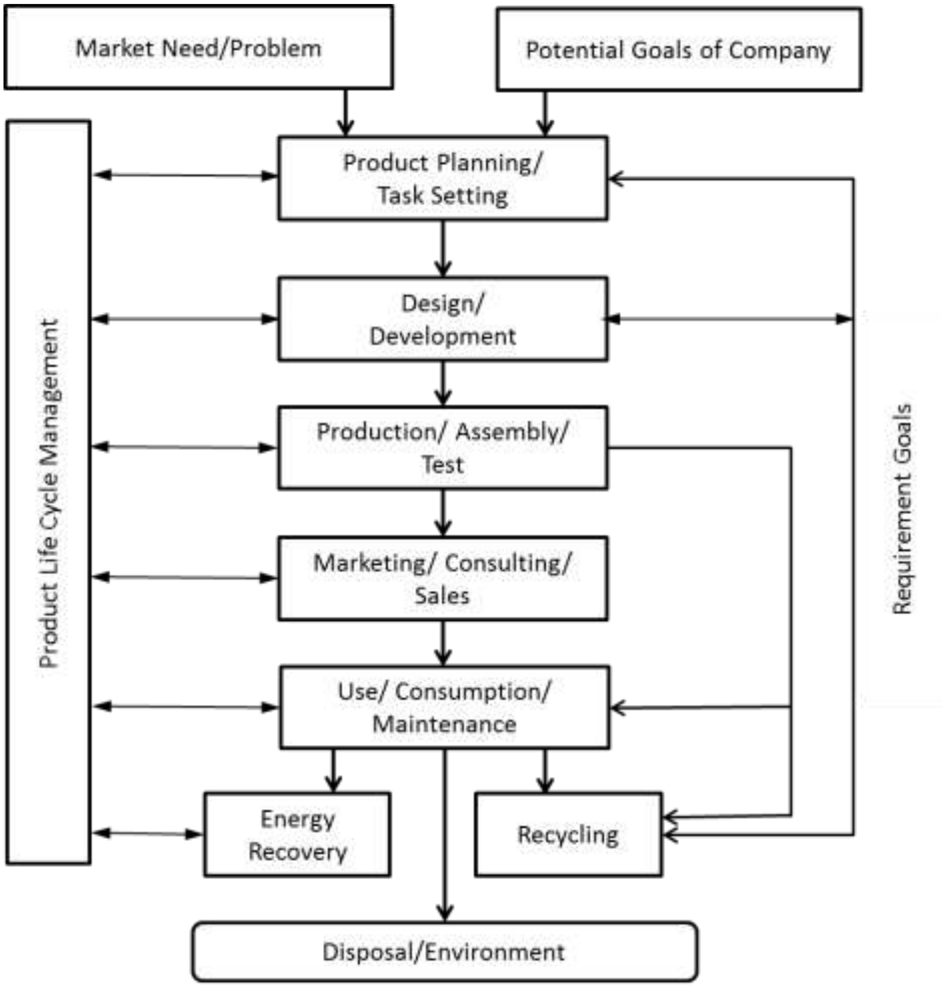

Market Need/Problem
Potential Goals of Company
Product Life Cycle Management
Product Planning/ Task Setting
Design/ Development
Production/ Assembly/ Test
Marketing/ Consulting/ Sales
Use/ Consumption/ Maintenance
Energy Recovery
Recycling
Disposal/Environment
Requirement Goals

Figure 14. Life cycle of a product (Pahl et al., 2007, p. 3)

In the second step, *formulating search strategies*, product planners reduce the number of search fields according the following criteria: customer needs, market trends and company aims. To identify strategic opportunities, different business strategies are adoptable. Introducing new products into the current markets or opening new markets with existing products are samples of business strategies that can be implied from a portfolio matrix. A promising gap determining the research field must be found by taking into account the company's goal, strength and market. To identify needs trends, customer behavior changes such as social developments, environmental awareness, transport problems, etc. should be conceived. Need-strength matrix is a tool commonly used to prepare the search of field proposal. Potential functions to carry degree of future client requirements are estimated. This analysis can result to R&D projects regard to new and future development components, assemblies, and products. At the end, goals and strength of the company must prioritize and select search fields.

In the third step, *finding product ideas*, the preferred search fields are explored for ideas by applying search methods such as brain storming, discursive methods, and **exploring patents**, which is one of conventional methods for information gathering to have access to state-of-the-art information. Depending on the degree of novelty, different product design strategies such as new product functions, new embodiments, and rearrangements can be employed.

In the fourth step, *selecting product ideas*, the generated ideas are subjected to a selection procedure whose criteria are linked to company's goals, company strength, market and other sources. High turnover, large market share, and functional advantages for the customers are the least criteria should be used at this step.

In the fifth step, *defining products*, the promising ideas are detailed more elaborately to characterize the requirements of the ideas. After elaboration, product ideas are evaluated again. Eventually, product proposals are prepared including preliminary requirements.

In the sixth step, *clarify and elaborate*, list of requirements are completed by considering external, internal, and structure requirements. Requirements related to disposal, recycling, and environmental impacts are more important in this step.

2.5.2. Patent Mining in NPD literature

To recognize current patent mining research applied in new product development, I looked at three main databases--IEEE, Web of Science (WOS), and Compendex[5]. After removing irrelevant and duplicated articles, 24 articles were identified and categorized into 9 groups based on their application (see Table 5). The papers are discussed based on their applications in main NPD activities introduced in Table 4. Patent mining cannot be applied in production and post-production activities, so only are planning and design activities considered.

As shown in Table 5, the papers are categorized into 10 groups based on their application of text mining. Table 6 also denotes the relationship between the papers and the main activities of the NPD process—technical, planning and design. Technical papers are papers that discuss the technical aspects of text mining rather than their applications in new product development. Therefore, only the papers pertaining to planning and design are discussed in the next sections.

[5] - The used query is: (("patent mining" or ("text mining" and "patent")) and ("product development" or "product design"))

Application	References	Technical	Planning	Design
Patent Classification	(Hong, Hua, & Hong, 2013)(J. Wang, Lu, & Loh, 2011)(Z. Li, Tate, Lane, & Adams, 2012)	✓		
Patent Mining (technical)	(Xu, 2009)(Bonaccorsi Andrea, 2007)	✓		
Patent Summarization	(Amy J.C. Trappey, Trappey, & Wu, 2009)(A. Trappey, Trappey, & S. Kao, 2006)	✓		
Design by Analogy	(Paul-Armand Verhaegen et al., 2011)(J. Murphy, Fu, Otto, Yang, et al., 2014)(Fu, Murphy, et al., 2013)(Fu, Chan, et al., 2013a)(Fu, Chan, et al., 2013b)			✓
Design Rationale	(Yan Liang et al., 2012)(Yan Liang & Liu, 2013)			✓
TRIZ	(Yanhong Liang & Tan, 2007)(Yanhong Liang et al., 2008)(Yan Liang & Liu, 2013)(Yanhong Liang & Tan, 2007)(A.J.C. Trappey et al., 2013)(P.-A. Verhaegen et al., 2011)			✓
Roadmapping	(S. Lee et al., 2006)		✓	
Technical parameter identification	(G. Cascini & Zini, 2011)			✓
Technology Identification	(Y.-R. Li, Tong, Hong, & Wang, 2006)		✓	
Tools Review	(Russo, 2011)(Lelescu et al., 2014)			✓

Table 5- Patent mining applications in new product development

2.5.1.1. Application of Patent Mining in the Design Stage

As illustrated in Table 5, the papers, which apply patent mining in the design stage, are classified into five specific subjects. Regarding to the similarity between the subjects, they can be classified into two main groups: 1) design methods and 2) special applications. Design methods include design by analogy, design by rationale, and TRIZ.

The second group discusses the special applications of patent mining in the design stage—a technical parameter identification and tools review.

2.5.2.1.1. Design Methods: Design-by-Analogy

Design by analogy is defined as "an ideation or problem-solving *method* based on analogies between products" (Paul-Armand Verhaegen et al., 2011). Design-by-analogy develops a set of ideas based on similar relationships from solutions to analogous problems (Fu, Murphy, et al., 2013). Design-by-Analogy (DbA) is an area that enables designers to identify and develop examples of related cases and scenarios, and connected experiences to solve a specific design problem (Moreno et al., 2014) (Linsey, Laux, Clauss, Wood, & Markman, 2007). Different analogical sources are applied in developing different DbA methods, such as answering direct questions to explore analogical categories in Synectics[6], looking for analogies in natural phenomena, developing analogous solutions from abstractions of functional models, and exploring analogous domains through semantic mapping (Moreno et al., 2014). Analogical search approaches and search engines are developed to identify potential analogies in digital sources like patent databases. Applying DbA can enhance the ability of automatic extraction analogies and systematic identification candidate products from patents, so the authors of the papers, shown in Table 5, have tried to develop DbA methods to facilitate analogies identification.

Verhaegen *et al*. have developed a method to extract product characteristics, called Product Aspects (PA), as a way to automatically and systematically identify

[6] Synectics is a creative problem solving method that stimulates thought processes of which the subject may be unaware (Gordon, 1961).

candidate products for design-by-analogy (Paul-Armand Verhaegen et al., 2011). The method is based on the analysis of the occurrence of words from specific WordNet categories. The TF-IDF is used to weigh the occurrence matrix, Principle Component Analysis (PCA) is deployed as the main method for dimensionality reduction, and VariMax is applied as a complementary method to make the result more interpretable. The Principle Components (PCs) gained by VariMax are called Product Aspects (PAs); they are ordered based on literal similarity and similarity under focus. By ordering PCs via literal similarity and plotting two products against literal similarity and similarity under focus, design-by-analogy candidates can be selected.

A group of scholars and practitioners have developed another DbA method and published the results in two papers (J. Murphy, Fu, Otto, Jensen, et al., 2014)(Fu, Murphy, et al., 2013). Their methodology is based on a five-step process which is developed based on the Vector Space Model as the basis of the analogy search method (Salton & Michael, 1986). In this method, a function vocabulary is created by indexing the extracted function via Zipf's law (Zipf, 1949) and subsequently a document vector is generated. The document vector is evaluated by applying *TF-IDF* and *cosine similarity*. Also, each document vector is normalized by patent *functional content measure* (*fcm*) to simplify the cosine similarity calculation. To rank the relevancy between patent vector and query vector, fcm and cos θ are linearly combined into a *total relevancy score* measure. The query is built by selecting primary and secondary functions, which address high level functionality from the functional model of the design problem. After the query completion, the metrics including cosine similarity, *fcm*, and *total relevancy score* are calculated. The top results are extracted and clustered by primary patent classification.

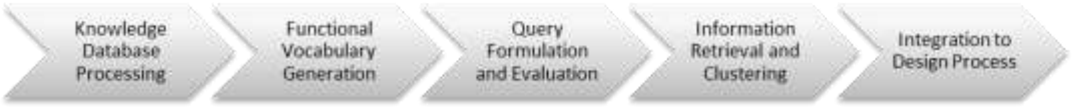

Knowledge Database Processing
Functional Vocabulary Generation
Query Formulation and Evaluation
Information Retrieval and Clustering
Integration to Design Process

Figure 15. Vector Space Model Process (Fu, Murphy, et al., 2013)(J. Murphy, Fu, Otto, Jensen, et al., 2014)

Fu et al. have evaluated the strengths and weaknesses of another algorithm (Fu, Cagan, Kotovsky, & Wood, 2013) and published the result in two papers (Fu, Chan, et al., 2013a) (Fu, Chan, et al., 2013b). This method is developed to organize the space of possible analogies based on structural forms in patent space. The algorithm discovers the best fitting-form for a given set of data from a space of 8 possible forms including partition, order, chain, ring, tree, hierarchy, grid, and cylinder. It instantiates the best form among the selected forms, which is called a structure. The computational methodology has three steps. First, a set of full-text patents are preprocessed by *Latent Semantic Analysis* (*LSA*). The result of the first step, similarity matrix shows the semantic pairwise relation between patents by assigning a numeric value between 0 and 1. In the second step, the similarity matrix is processed by a hierarchical Bayesian algorithm to discover structural form of the data. At the last step, LSA is used again to generate labels to describe the clusters of patents in the best structure.

2.5.2.1.2. Design Methods: Design Rationale

Many computerized systems have been developed since 1960s to assist engineers to model, design, and represent their ideas. Among such systems, Design

Rationale (DR) refers to decisions made over design process, and the reasons behind the decisions (Jarczyk, Loffler, & Shipmann, 1992). "*Design rationale includes all the background knowledge such as deliberating, reasoning, trade-off and decision-making in the design process of an artifact–information that can be valuable, even critical, to various people who deal with the artifact*" (Regli, Hu, Atwood, & Sun, 2014, p. 209). To represent the reasons and the decisions, different representation approaches are developed to facilitate reasoning or communication (J. Lee & Lai, 1991). Some scholars of Hong Kong polytechnic University have developed a DR representation model named ISAL (issue, solution, and aircraft layer) (Y. Liu, Liang, Kwong, & Lee, 2010). As the pioneers of DR application in patent analysis for new product development purposes, they have published two papers (Yan Liang et al., 2012) (Yan Liang & Liu, 2013), which are briefly explained below.

ISAL (Y. Liu et al., 2010) is a computational model in which consist of three layers including issue layer, solution layer, and artifact layer to represent DR, as shown in Figure 16. The issue layer refers to the reasons behind the design such as needs, problems and limitations of prior designs, or opportunities to develop a new product. The design solution layer describes thoughts, ideas, possible approaches and mechanisms used to address the reasons. The aircraft layer points to the main components and properties of the product, which can be a physical product like a printer or a soft product like codes of a software program.

In order to discover DR from design documents, Liang et al. (Yan Liang et al., 2012) have developed three algorithms for each layer of ISAL to 1) extract artifact information, 2) generate a summary of the issue, and 3) generate solution and reason pairs. For the issue layer, they have defined a semantic sentence graph to model sentence relationships through language patterns like terms and phrases that convey meaning. Based on this graph, they have improved the algorithm to extract issue sentences and to discover solution-reason sentences in the solution layer. For artifact information extraction, they have proposed two term relations, the positional term relation and the mutual term relation. These relations extend their document profile model to score the candidate terms and suggest terms with higher scores as artifact components. Finally, the authors have conducted some experimental studies to test the performance of the proposed algorithms using patents.

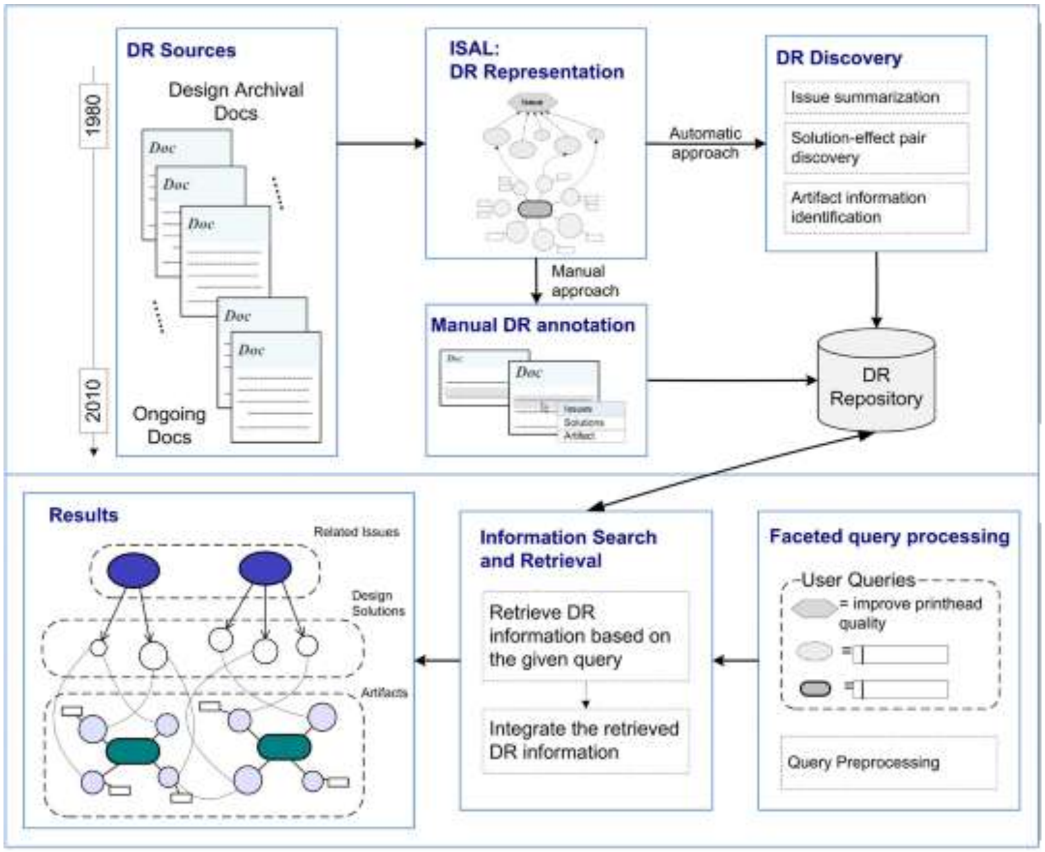


Figure 16- The ISAL-based framework of DR discovery, retrieval and management (Yan Liang et al., 2012)

Liang et al. have applied ISAL for another rationale-based patent analysis to analyze technical foci of corporations (Yan Liang & Liu, 2013). The proposed approach includes four stages: 1) pre-processing, 2) artifact feature connection measurement, 3) key feature grouping stage, and 4) patent categorization. In the first stage, the DR information of each patent is discovered and transformed into semantic form, via the ISAL model, in three layers, which are issue, solution, and aircraft. In the second stage, the relationship between invention features, attributes, and artifacts is measured based on a feature association-based approach. Highly connected features are clustered to categorize design foci from a corporation perspective. Each of the feature groups denotes a probable focus of design or a set of inventions that have been published previously. In the last stage, patents are assigned to classes based on their categories of invention. The patent grouping process facilitates investigating design issues and technology that has been designed or proposed for similar design foci.

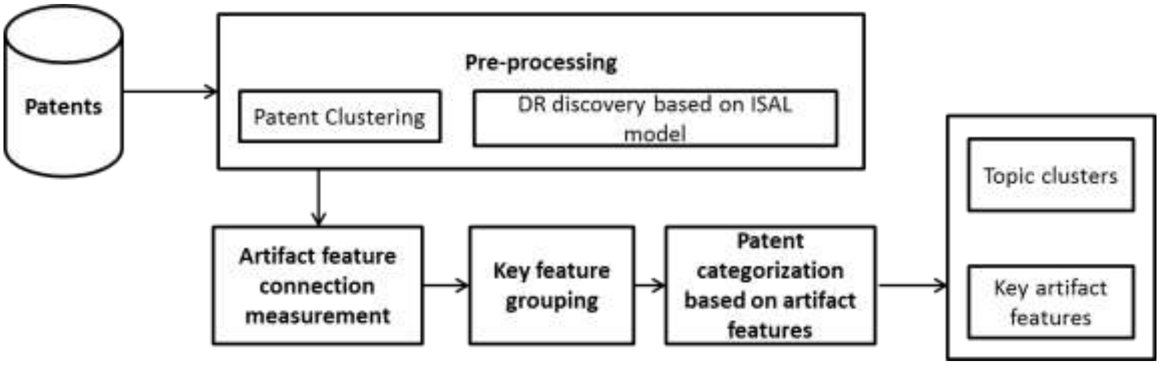

Patents
Pre-processing
Patent Clustering
DR discovery based on ISAL model
Artifact feature connection measurement
Key feature grouping
Patent categorization based on artifact features
Topic clusters
Key artifact features

Figure 17- The feature association-based method for the analysis of design focus (Yan Liang & Liu, 2013)

2.5.2.1.3. Design Methods: TRIZ

Theory of Inventive Problem Solving, TRIZ, is a philosophy, a method, and a problem-solving and analysis toolbox to generate innovative solutions (Mann, 2001). TRIZ is a systematic approach that helps to construct a problem definition and problem-solving process, works for any situation, and is effective across a broad spectrum of fields and problems types (Mann, 2001). TRIZ requires a mapping of a specific problem to an abstract problem; then the users can obtain abstract solutions and back to a specific solution (P.-A. Verhaegen, D'Hondt, Vandevenne, & Dewulf, 2010). After reviewing thousands of patents, Altshuller, the father of TRIZ, and his colleagues found only a limited number of inventive principles are used to solve or eliminate contradictions. They categorized the inventive principles into several retrieval forms including a contradiction table, 39 engineering parameters, 40 incentive principles, and 76 standard solutions (Yanhong & Runhua, 2007) (Yanhong Liang et al., 2008).

A product design problem can be considered as one or several contradictions and inventive principles (Yanhong & Runhua, 2007). According to TRIZ, to solve a design contradiction in product development, the product is developed by referring to the analogous inventions even among dissimilar problems. Since patents are a rich source of technological information for TRIZ applications, very few scholars have developed automatic tools to extract contradictions and inventive principles from patents to assist innovators to solve a design problem.

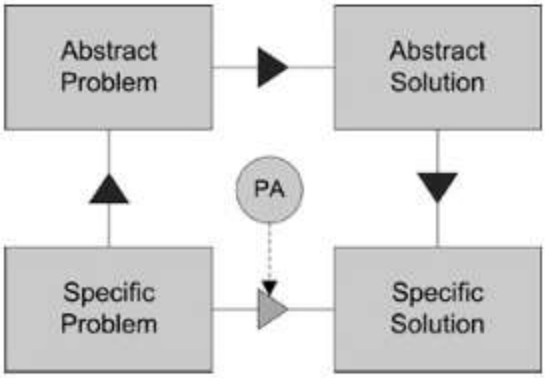


Figure 18- The TRIZ process (P.-A. Verhaegen et al., 2010)

Patent databases are important technical sources for product development for which knowledge and information to inspire designers and engineers exists. Some scholars started applying TRIZ and developing automatic tools to assist innovators in acquiring useful information from patents (Yanhong & Runhua, 2007).

Yanhong and Runhua have tried to develop a text mining technique which can be used by TRIZ users. Their objective is, basically, to classify patents according to contradictions and inventive principles. In their first publication, they claim that they have applied clustering algorithm and a multi-naïve Bayes algorithm, after applying text mining and indexing patents to keywords (Yanhong & Runhua, 2007). In their second report (Yanhong Liang et al., 2008), they claim that they have utilized semantic analysis by WordNet, followed by Latent Semantic Analysis (LSA) to reduce the dimensionality of the data set. At the end, the patents are classified into contradictions and inventive principles using text software.[7] Unfortunately, the authors have neither explained how the algorithms work, nor have they presented any empirical evidence or case study.

Verhaegen et al (Paul-Armand Verhaegen et al., 2011) have developed a method to extract product characteristics called Product Aspects (PA) to identify candidate products for design-by-analogy. They believe that there is a link between TRIZ and Design-by-Analogy. Design-by-Analogy enables a designer to find another product with similar functions. Similarly, TRIZ enables a designer to find solution (product characteristics or functionality) based on a specific problem. Therefore, there is similarity between their methodology, and this methodology, which is illustrated in Figure 19. Both methodologies address the product aspect as the pivot point to cluster products characteristics mentioned in patents. In the TRIZ context, PAs facilitate identifying specific products to supplement the abstracted solution in TRIZ tools.

[7] Waikato environment for knowledge analysis (WEKA) developed by the University of Waikato, New Zealand.

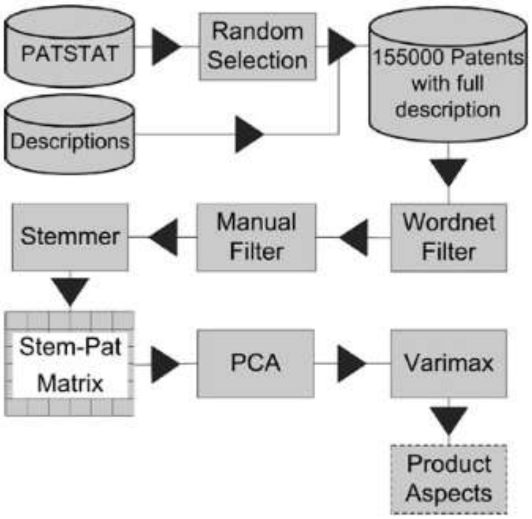


Figure 19- Methodology flowchart (P.-A. Verhaegen et al., 2010)

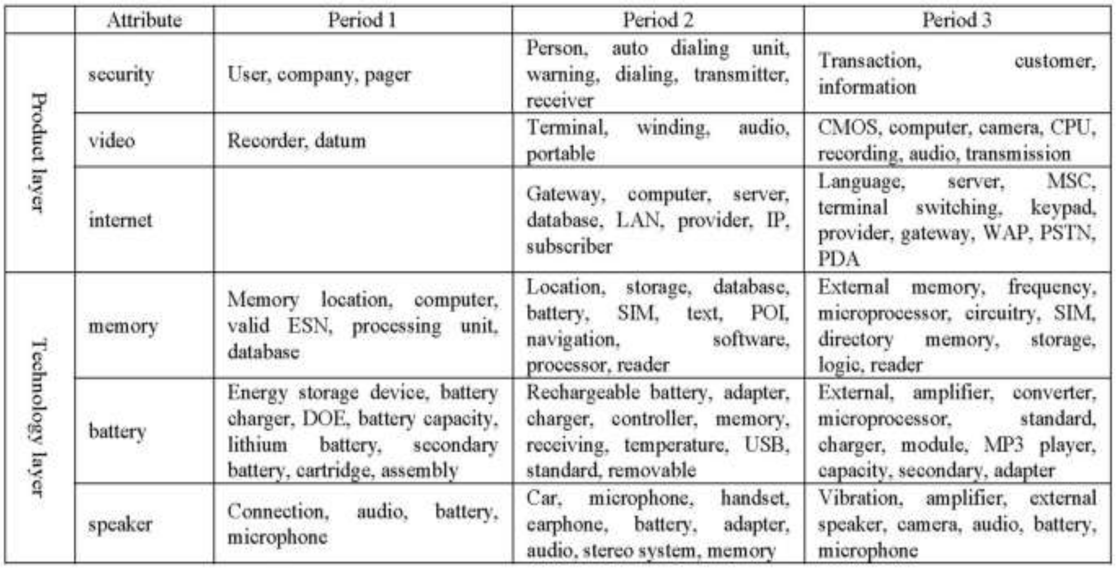
2.5.1.2. Application of Patent Mining in the Planning Stage

Only two of the papers mentioned in the previous section are applicable to planning in new product development. In the first paper (S. Lee et al., 2006), the authors have applied data mining and co-word analysis to develop a three-layer roadmap. In the second paper (Y.-R. Li et al., 2006), the authors propose a heuristic method for patent search to discover new opportunities.

In the first paper (S. Lee et al., 2006), the authors offer a keyword-based proves for technology roadmapping, which consists of three successive stages: 1) data elicitation, 2) data transformation, and 3) mapping. At the first stage, core keywords are identified, and then, in the second stage, a co-word matrix is constructed by measuring co-occurrence frequency of the keywords. The matrix is converted to a co-efficient matrix to show co-relation among the keywords. In the last stage, the keywords are mapped into two dimensions by applying the Multi-Dimensional Scaling (MDS) method on the co-efficient matrix. The result is a keyword network where keywords are connected based on their similarity. Also, it is worth explaining that 'roadmap' in this research is a three-layer map including product layer, technology layer, and R&D layer. The layers are respectively represented by a 'keyword portfolio map', a 'keyword relation map', and a 'keyword evolution map'.

In the second paper (Y.-R. Li et al., 2006), the authors claim that they have provided a novel tactic for discovering new opportunities. They have applied text mining to extract keywords from patents to figure out heterogeneity between technologies and firms. To filter out highly anomalous (*'significant rare*') keywords that only humans can recognize, the authors have developed two indexes: the entropy of the technology and the entropy of the firm. At the end, they have combined 'significant rare' keywords to a knowledge map, where the keywords present the potential opportunities. Also, they claim that this method helps researchers track knowledge spillover between different firms. This information can provide the chance to discover opportunities during the process of new product development.

The methods developed in these papers are potentially applicable in the planning stage. The keyword-based roadmap, Figure 20, reveals potential keywords in different period times. The keyword-based roadmap, the main output of the first paper (S. Lee et al., 2006), almost shows the evolution of keywords in product, technology, and R&D layers, and experts should interpret the evolution over period times for each of the attributes. The knowledge map, the main output of the second paper (Y.-R. Li et al., 2006), provides specific keywords in a patent system where researchers can track the sequence of ideas and decision makers can find more opportunities to get licensing in early stage of new product development. In summary, both methods claim that they can be applied in the NPD planning stage, but the methods do not have a cohesive relation with a specific NPD planning method.



	Attribute	Period 1	Period 2	Period 3
Product layer	security	User, company, pager	Person, auto dialing unit, warning, dialing, transmitter, receiver	Transaction, customer, information
	video	Recorder, datum	Terminal, winding, audio, portable	CMOS, computer, camera, CPU, recording, audio, transmission
	internet		Gateway, computer, server, database, LAN, provider, IP, subscriber	Language, server, MSC, terminal switching, keypad, provider, gateway, WAP, PSTN, PDA
Technology layer	memory	Memory location, computer, valid ESN, processing unit, database	Location, storage, database, battery, SIM, text, POI, navigation, software, processor, reader	External memory, frequency, microprocessor, circuitry, SIM, directory memory, storage, logic, reader
	battery	Energy storage device, battery charger, DOE, battery capacity, lithium battery, secondary battery, cartridge, assembly	Rechargeable battery, adapter, charger, controller, memory, receiving, temperature, USB, standard, removable	External, amplifier, converter, microprocessor, standard, charger, module, MP3 player, capacity, secondary, adapter
	speaker	Connection, audio, battery, microphone	Car, microphone, handset, earphone, battery, adapter, audio, stereo system, memory	Vibration, amplifier, external speaker, camera, audio, battery, microphone

Figure 20- Keyword-based technology roadmap (Product and technology layers) (S. Lee et al., 2006)

2.5.3. Summary of Section

The new product development process contains four general activities: planning, design, production, and post-production. Patent mining is inherently applicable only in planning and design activities. The literature shows that to date the main focus of the researchers has been to apply patent mining and clustering for specific design approaches, which are TRIZ and Design-by-Analogy. Also, no significant paper pertaining to patent mining applications for NPD planning has been published to date.

There is a significant difference between NPD design and NPD planning activities that leads us to apply different analysis methods. In NPD design approaches including TRIZ and DbA, the researchers intend to extract the features of the related patents and group the extracted features. In NPD planning, the decision makers need to identify specific technologies. These differences lead to apply clustering for design purposes and to apply classification methods for planning. It is the reason why I will apply classification methods to cover the gap in NPD planning.

	Design (TRIZ, DbA)	**NPD Planning**
Purpose	Features extraction	Feature identification
Analysis Method	Clustering	Classification

Table 6 - Different purposes and analysis methods in NPD design and planning activities

2.6. Chapter Summary

Section 2.2 serves as a general introduction to text mining. It mentions how text mining researchers can mine keywords from textual sources by applying natural language processing, extract information by deploying information retrieval, and learn to recognize patterns of the keywords by applying machine learning methods.

Section 2.3 introduces patent retrieval applications and methods. That section described how interactive patent retrieval improves the performance of keyword-based methods and methods based on semantics. The research proposed in this dissertation is an interactive patent retrieval method, which is built based upon both keywords and keyword semantics. The main difference between the proposed approach and currently deployed methods is that in the proposed approach queries are not the focal point of improving the performance of patent retrieval. Instead, classifiers act like queries. Classifiers are trained based on patents that human experts have judged either as relevant or irrelevant. Classifiers subsequently identify all relevant and irrelevant patents in a corpus. This distinction constitutes the primary contribution of the proposed research.

Section 2.4 discusses how patent analysis methods have evolved in the three stages shown in Figure 10. In the first stage, the researchers started analyzing metadata and citations. These analyses are called *bibliometric analysis* and *citation analysis*, respectively. In stage 2, the researchers started applying more sophisticated analyses, which are *cluster analysis* and *network analysis*, on metadata or citations. *Patent mining* appeared in the third stage due to the emergence of text mining methods, which allow researchers to extract technical information from the main body of patents. Researchers also utilized complementary methodologies including *ontology-based approaches* and *semantic analysis* to enhance the performance of their analysis. Section 2.4 also shows that statistical, lexical and corpus approaches are utilized to develop different information retrieval methods to extract information from patent contents. In addition, section 2.3 illustrates how network analysis and cluster analysis are applied to recognize the patterns of extracted information. Interestingly, no application of classification is recognized in the evolution of patent mining. Applying classification allows patent miners to have more accurately access to patent contents, which advances the state of the art of patent mining.

In section 2.5, I look at applications of patent mining in ***the NPD process***. In section 2.5.1, I review three prestigious sources in the NPD literature, which recognize that the general stages in the NPD process are planning, design, production, and post-production (see Table ***2-1***). Section 2.4.2 shows how patent mining can be applied to support each of these NPD stages (see Table ***2-2***). It turns out that patent mining is very applicable for *design* purposes. However, when it comes to applying patent mining to the *planning* stage of the NPD process to support *opportunity identification*, a huge gap in the literature has been identified--there is no significant patent mining research to match product features with enabling technologies.

My dissertation research intends to answer the management question “How can R&D engineers/managers that are engaged in product planning find patents that will provide new technological opportunities?” The literature review in this chapter has been performed to identify the gaps in the literature that pertain to this question. After stating this primary gap and decomposing it into sub-gaps, I pose the research questions that address these sub-gaps. My dissertation research will consist of answering these research questions, all of which have been stated in section 1.3.

CHAPTER 3: RESEARCH DESIGN AND RESEARCH METHODS

In this chapter, I introduce my research in detail. First, I describe a theoretical framework for the proposed research. I subsequently explain my research design; the variables and measures; the data collection procedure; issues related to validity and reliability; and my approach to data analysis.

3.1. Theoretical Framework

R&D engineers and managers are interested in patents as a rich technical source where they can extract the technical information to apply in many purposes like new product development planning. In this research, I have developed a method to extract required information, product feature and enabling technologies, from patents and identify available opportunities in USPTO patents based on how much the patents address the required information. To do so, I apply Natural Language Processing (NLP), Information Retrieval, and Machine Learning methods. In addition, to be able to address 'product feature' and 'enabling technologies', I apply ontological semantic analysis to enhance the effectiveness of this opportunity identification method. The theoretical framework for this endeavor is described below and illustrated in Figure 21.

Classification is deployed in my research to recognize specific *concepts* in patents. The concepts are product feature and enabling technologies. *Keywords* are the best representative of the concepts. In my research, there are three types of keywords: 1) regular keywords, 2) ontological keywords, and 3) ontological semantic keywords.

Regular keywords are generated through 'regular text extraction', i.e. without designing an ontology. *Ontological keywords* are keywords derived from designing an ontology that is based on interviews with experts. *Ontological semantic keywords* contain ontological keywords and their synonyms, which are extracted from the WordNet lexical database.

Figure 21 implies that ontological semantic keywords are a better representation of the concepts than ontological keywords are. If hypothesis 2 is confirmed, then the classifiers have a better performance with ontological semantic keywords than with regular keywords. If hypothesis 3 is confirmed, then the classifiers have a better performance with ontological semantic keywords than with ontological keywords. Similarly, ontological keywords are a better representation of the concepts than regular keywords are. That is, the classifiers have a better performance with ontological keywords than with regular keywords (hypothesis 1).

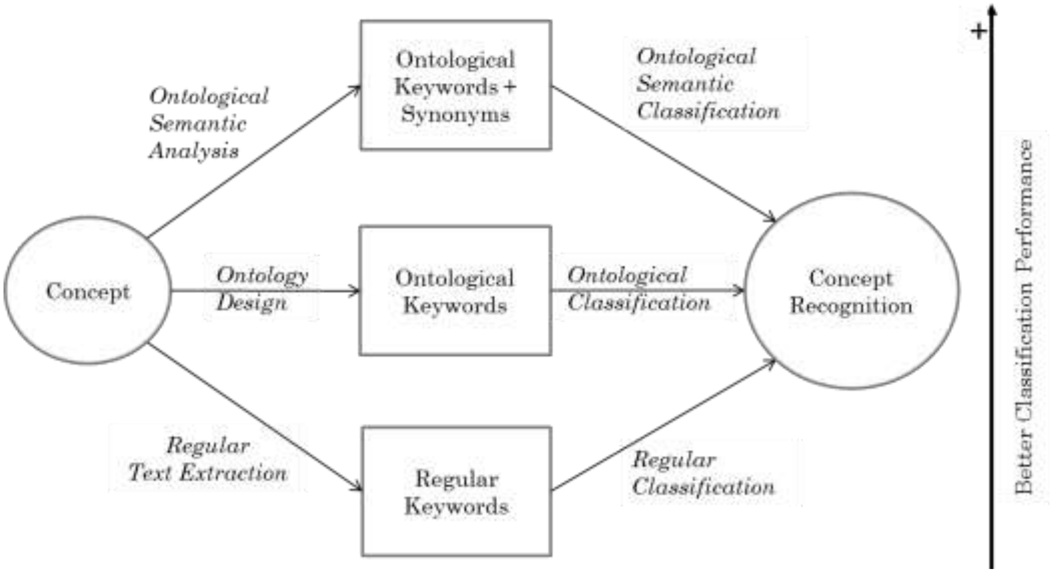

Ontological Keywords + Synonyms
Ontological Semantic Analysis
Ontological Semantic Classification
Concept
Ontology Design
Ontological Keywords
Ontological Classification
Concept Recognition
Regular Text Extraction
Regular Keywords
Regular Classification
+
Better Classification Performance

Figure 21- Theoretical framework

3.2. Research Setting and Unit of Analysis

To provide the setting of the research, an iron casing company producing traditional skillets is chosen. The name of the company is Finex; it is located on Portland (Oregon). The core technology of Finex is iron casting. Finex is chosen for the research because:

- Finex regularly surveys its market, so they know the needs of their customers in the market.
- Finex's experts have the required expertise in marketing, and iron casting (material science), so they are able to develop the ontologies of product features and enabling technologies, and participate in the patent search that is conducted as part of my dissertation research.
- There are thousands of iron-casting technologies in the USPTO database, so patents are a good source for identifying opportunities for product planning in Finex.

The US patents that potentially address 'product features' and 'enabling technologies' comprise the study's unit of analysis. To extract those patents from USPTO database that pertain to my unit of analysis, I use Cooperative Patent Classification (CPC), a patent classification system jointly developed by the European Patent Office and the USPTO that identifies the subject of patents. Finex experts select the CPC codes that pertain to their 'product features' and 'enabling technologies'. One patent often has several CPC codes.

3.3. Research Design

Finex acts as a case for testing the theoretical framework from section 3.1 under a variety of circumstances, which are depicted in Figure 22. Analysis of the method under development occurs along three dimensions: data set, classification scenarios and classifiers. Identifying technological opportunities depends upon two data sets that illustrate the basic concepts of new product development: one pertains to product features; the other pertains to attributes of enabling technologies that the product developers would consider integrating into their products. The two data sets are matched in three classification scenarios— I) regular classification, II) ontological classification and III) ontological semantic classification—which are described in section 1.3. These scenarios respectively require the following types of keywords as inputs for classification: 1) regular keywords that are recognized by NLP and information retrieval, 2) ontological keywords, and 3) ontological keywords and their synonyms. Three classifiers are benchmarked for performance in each scenario: k-NN, SVM and random forest. Going through both data sets, all three scenarios, and three classifiers yield a total of 18 instances under which the method to be developed in this dissertation is evaluated.

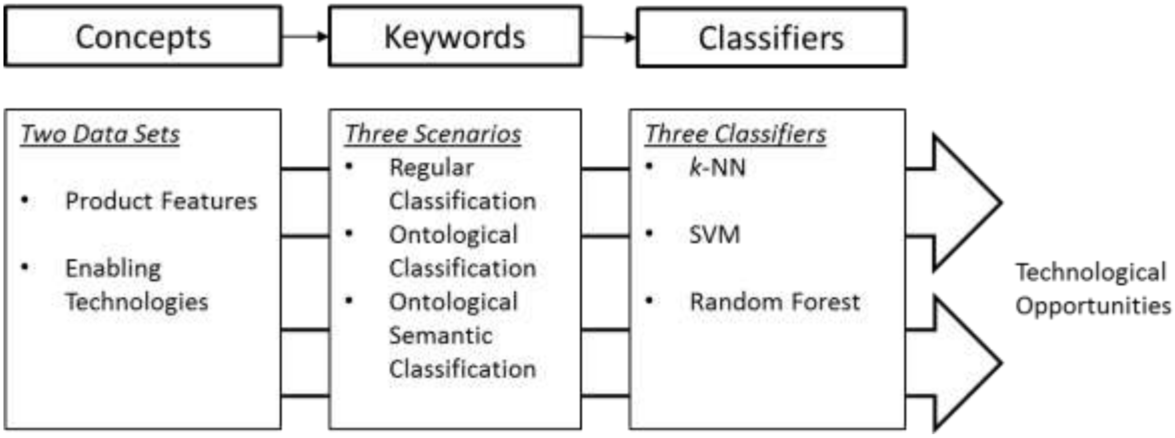


Figure 22- The combination of concepts, keywords, and classifiers.

All instances discussed above have gone through the same sequence of data collection and analysis activities. First, I supply human experts with a set of patent classes from the Cooperative Patent Classification system of patent classification. These experts identify patent classes that are of consequence to the product features and technical attributes for the product that they are trying to develop. I subsequently extract from the USPTO all patents related to the CPC classes that the experts have identified. Then, I apply Natural Language Processing to extract all words from the extracted patents and remove stop words (such as 'is', 'the', 'at' and 'which') and I call this set of words *raw keywords*. Next, I apply information retrieval to eliminate keywords deemed relatively unimportant according to the term frequency-inverse document frequency (TF-IDF) ranking. Information retrieval yields a reduced set of keywords and how often the keyword occurs in each patent (the term frequency number). Patents are subsequently classified according to whether they are related to specific product features or enabling technologies. The classifier assigns a '1' for relevant patents and a '0' for irrelevant patents. Different classifiers have different classification criteria. Opportunities are identified from resulting classifications of product features and enabling technologies. Finally, the identified opportunities are authenticated by one expert.

3.3.1 Two Data Sets: Product Features and Enabling technologies

Product features and enabling technologies are the two key factors for opportunity identification (Ulrich & Eppinger, 2011). The respective data sets that pertain to these factors both reside within the patent database. Information pertaining to product features and enabling technologies may even exist within the same patent. Experts that work in the company that is developing a product have to identify what the product features and enabling technologies are. Their collective understanding of product features of and enabling technologies of the product under development determines the scope and boundaries of the two data sets.

The current state of the art in patent mining does not delineate between product features and enabling technologies because it does not recognize the collective understanding of the developers. Developing ontologies that represent the collective understanding of the product developers could therefore enhance their ability to identify technological opportunities by making a clear distinction between the data sets for product features and enabling technologies.

3.3.2 Three Scenarios: RC, OC and OSC

Figure 23, Figure 24, and Figure 25 respectively detail the three scenarios under investigation: regular classification (RC), ontological classification (OC) and ontological semantic classification (OSC). All three consist of a data collection step, which involves elicitation of information from human experts and extraction of text from a patent database, and a data analysis step in which patents are classified as to whether they address product feature or enabling technologies. The difference between the scenarios consists of the depth of involvement of the human experts and the types of data for classification that result therefrom.

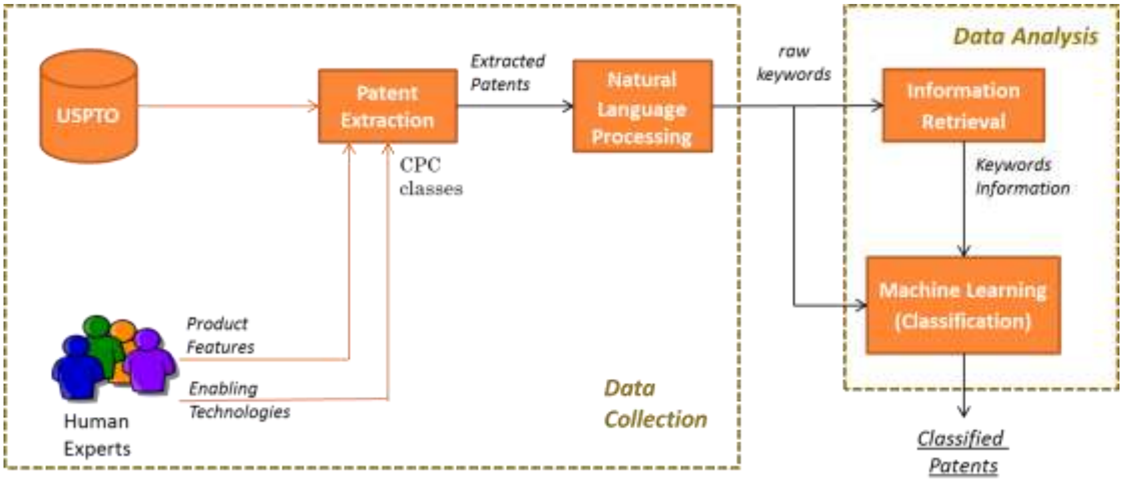


Figure 23- Research framework (scenario I or regular classification (RC))

In RC (Figure 23), the human experts identify section and the subordinate class of the USPTO to which their perceptions of the features for the product under development and attributes of the enabling technologies under consideration for incorporation apply. All patents within the USPTO that are contained the identified patent class are considered potentially relevant from the point of view of identifying technological opportunities for the product under development. They are consequently extracted from the USPTO. Natural language processing generates a set of raw keywords from the extracted patents. Information retrieval subsequently identifies the frequency numbers of the raw keywords, calculates the TF-IDF measure, and used the TF-IDF measure to remove the less important keywords. Finally, machine learning classifies the patents based on a pattern of frequency numbers of regular keywords.

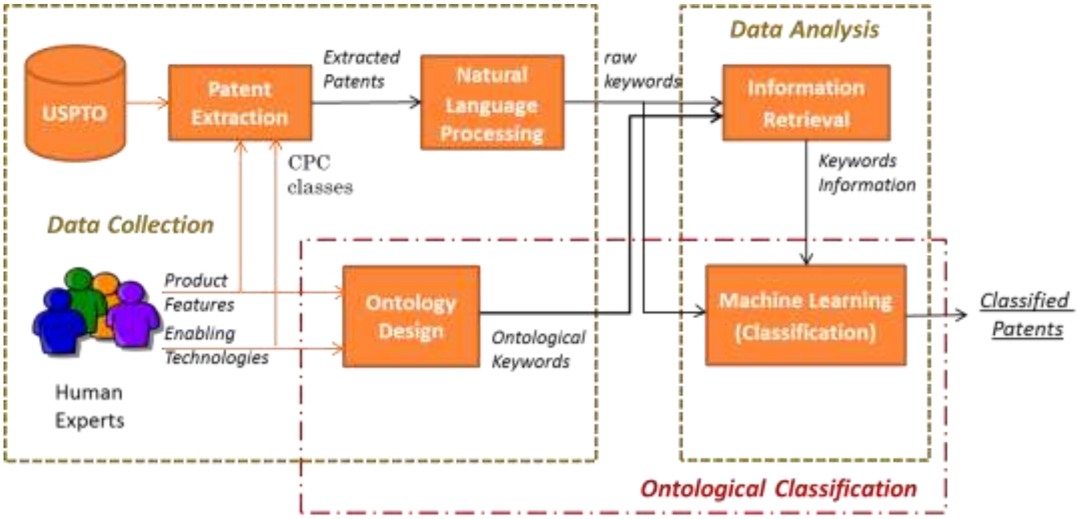


Figure 24- Research framework (scenario II—ontological classification)

OC (Figure 24) and OSC (Figure 25) proceed analogously. In OC and OSC, I engage with the human experts in an interview. I design an ontology that reflects their perceptions of product features and another that reflects their perception of enabling technologies. These ontologies yield ontological keywords for each data set. Information retrieval still identifies the frequency numbers of the raw keywords, calculates the TF-IDF measure. The TF's of raw keywords are used to remove the less important keywords, according to determining cut-off value based on the Zipf curve. However, I subsequently select the keywords that appear in the output of information retrieval AND on the list of ontological keywords for subsequent classification by machine learning.

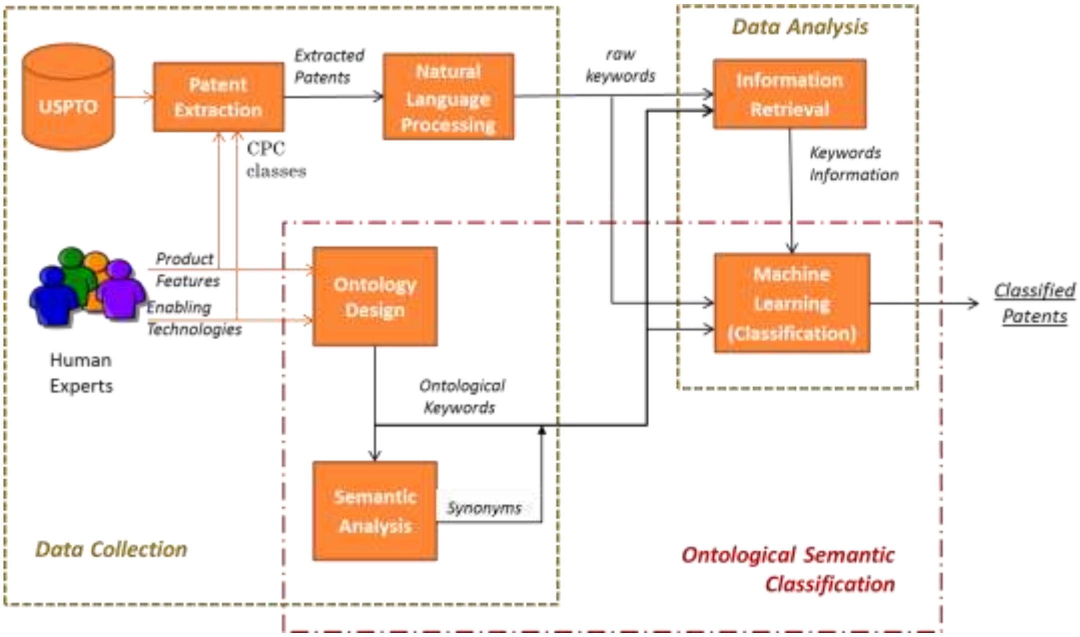


Figure 25- Research framework (scenario III—ontological semantic classification)

In OSC, I use the WordNet lexical database (Miller, 1995) to generate the synonyms of the ontological keywords. I select the keywords that appear in the output of information retrieval AND on the list of ontological keywords for subsequent classification by machine learning. I ALSO select the keywords that appear in the output of information retrieval AND on the list of synonyms of ontological keywords for subsequent classification by machine learning.

3.3.3 Classifiers

Three classifiers—*k*-NN, SVM and random forest—are selected to benchmark their performance in classification. *k*-NN is chosen because it is able to compare the similarity of documents. SVM is selected because SVM has yielded a substantial improvement of classification accuracy in many text classification studies (Joachims, 1998) (Sebastiani, 2002). SVM's main advantage is that SVM is not affected by the number of features encountered in the training data set (Kotsiantis, Zaharakis, &

Pintelas, 2007). Random forest is selected because it tends to generate accurate classifications when sample sizes are small (Tan et al., 2006, p. 278).

3.3.3.1. k-nearest neighborhood

The *k*-nearest Neighbor classifier is an instance-based learning algorithm method where the similarities between a document and the *k*-nearest neighbors of training data set are computed to determine the class of the document according to the similarities (Baharudin, Lee, & Khan, 2010). The training set is mapped into feature space while the feature space is partitioned into regions based on the classes of the training set, shown in Figure 26. A document is assigned to a class if it is the most frequent class among the k nearest training data. Nearest-neighbor methods have a tendency to work well when the class borders are somewhat complex; If class boundaries are nearly linear, other classification methods may perform better (Sutton, 2012).

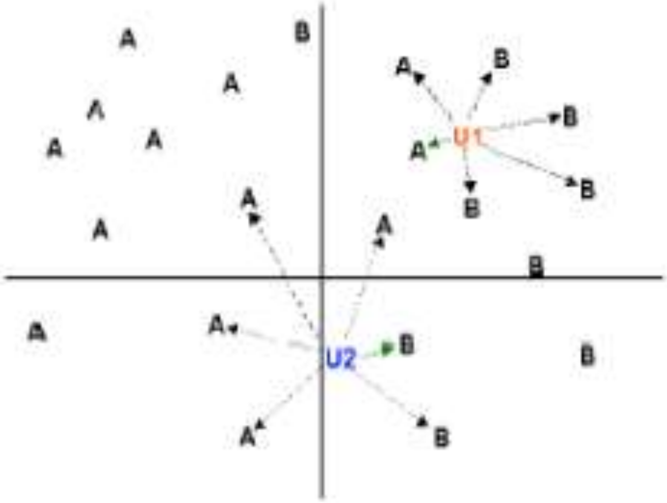


Figure 26- *k*-nearest neighbor (Baharudin et al., 2010)

In this research, TF-IDF weighting scheme and cosine similarity function are used instead of Euclidean distance to measure the similarity between two document vectors (Salton & Michael, 1986). Qian *et al* (Qian, Sural, Gu, & Pramanik, 2004) have theoretically and experimentally shown that Euclidean distance and cosine angel distance are similar when applied to high dimensional NN queries.

Given two documents D_1 and D_2, their corresponding weighted feature vectors are W_1 and W_2. The similarity between documents D_1 and D_2 is computed as shown below (Salton & Buckley, 1988):

$$S(D1, D2) = cos(D1\ , D2) = \frac{D1\ .\ D2}{\|D1\ \|\|D2\ \|}$$

$$= \frac{\sum_{i=1}^{n} d_{1i} .\ d_{2i}}{\sqrt{\sum_{i=1}^{n}(d_{1i})^2}\ \sqrt{\sum_{i=1}^{n}(d_{2i})^2}}$$

where d_{1i} and d_{2i} are the components of vectors D_1 and D_2 and $i=1\ldots,n$.

3.3.3.2. Support Vector Machine

The main principle of Support Vector Machine (SVM) is to determine the best separator (hyperplane) of different classes in the search space where the data set is also linearly separable (Aggarwal & Zhai, 2012). For example, as illustrated in Figure 27, there are infinite hyperplanes whose training error is zero, but there is no guarantee that they have equal generalization error on the test data set (Tan et al., 2006, p. 258). Each decision boundary, like B_1 is associated with a pair of hyperplanes like b_{11} and b_{12}, which are obtained by moving a parallel hyperplane away from the decision boundary until it touches the closest circle(s) or square(s). The distance between the two hyperplanes is called the *margin* of the classifier. Decision boundaries with large margins tend to have better generalization error than those with small margins (Tan et al., 2006, p. 258).

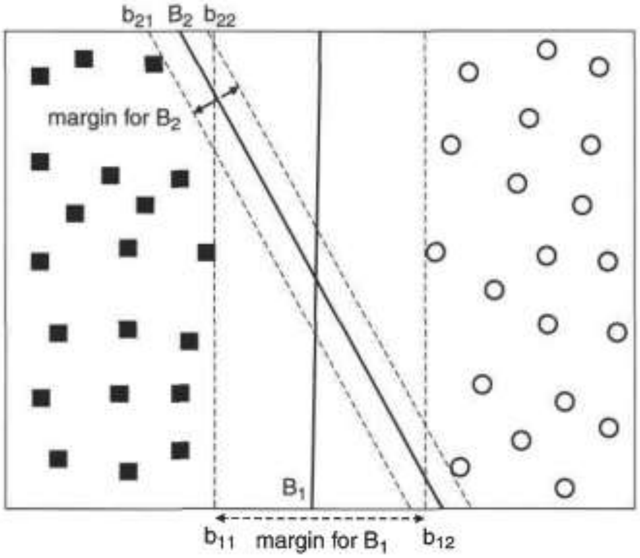


Figure 27- Possible decision boundaries for a linearly separable data set (Tan et al., 2006, p. 258)

The decision boundary of a linear SVM classifier can be written in this form:

$$\mathbf{w.x + b = 0}$$

The parameters **w** and **b** can be rescaled so that the two parallel hyperplanes, b_{i1} and b_{i2}, can be formed as below:

$$\boldsymbol{b_{i1}}\text{:}\ \ \mathbf{w.x + b \geq 1} \text{ if } y_i = 1\mathbf{;}$$

$$\boldsymbol{b_{i2}}\mathbf{:}\ \mathbf{w.x + b \leq -1} \text{ if } y_i = -1\mathbf{.}$$

The conditions impose the requirements, that all training instances from class y=1 must be located on or above the hyperplane w.x +b=1 while those instances from class y=-1 must be located on or below the hyperplane w.x+b=-1. Both equations can be summarized in this form:

$$\mathbf{y_i\ (w.x + b) \geq 1}$$

Let's consider **x** is a data point located on b_{i1} and $b_{i2.}$ The margin d can be computed by subtracting the second equation from the first equation.

$$\mathbf{W\ .\ (x_2 - x_1) = 2}$$

$$||\mathbf{w}|| \times d = 2$$

$$\therefore\ d = \frac{2}{||\mathbf{w}||}$$

Maximizing the margin is equivalent to minimize the following function:

$$\frac{||\mathbf{w}||^2}{\mathbf{2}}$$

In the training phase, the parameters **w** and **b** must be determined so that the abovementioned function is minimized and following two conditions are met:

$$\max_{w} \frac{||\mathbf{w}||^2}{\mathbf{2}}$$

Subject to: $\mathbf{y_i\ (w.x + b) \geq 1,} \quad i = 1,2,\ldots, N$

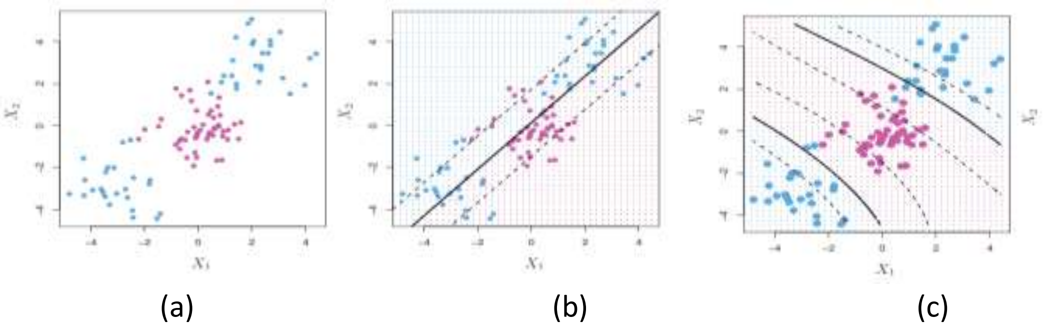
The objective function is quadratic and the constraints are linear, so the model is in form of **convex** multiplier method that can be solved by using the standard **Lagrange multiplier** method[8].

[8] For more details, refer to (Tan et al., 2006, pp. 262–264)

3.3.3.2.a. Kernel function selection

Sometimes data sets, such as the one shown in Figure 28-a, have nonlinear boundaries. If they are linearly transformed, the result cause to high variance as shown in Figure 28-b, but if a convenient nonlinear function is used, the data are transformed with low variance as shown in Figure 28-c.

Kernels are transformation functions that allow process input data to a right space. A kernel measures the similarity between two instances in the transformed space using the original attribute set (Tan et al., 2006, p. 273).



(a) (b) (c)

Figure 28- Transformed data from the original space to the transformed space by kernel function (James, G., Witten, D., Hastie, T., Tibshirani, 2013)

The radial basis function (RBF) kernel, also known as the Gaussian kernel, is a popular kernel function used in none-linear Support Vector Machine classification. Therefore, both the RBF kernel and the linear kernel will be considered in the *cross-validation* introduced in section 3.7.2., to discover which performs better.

3.3.3.3. Random Forest

In random forest classification, each decision tree uses a random vector that is generated from a fixed probability distribution. *N* input features are selected to split each node of the decision tree from these selected *N* features rather examining all the available features. This may help the bias present in the resulting tree. Random forest is an ensemble method that is based on decision tree classifiers. As shown in Figure 29, random forest combines the predictions of multiple decision trees, where each tree is generated based on the values of an independent set of random vectors (Tan et al., 2006, p. 290). Random vectors are created from a fixed probability distribution.

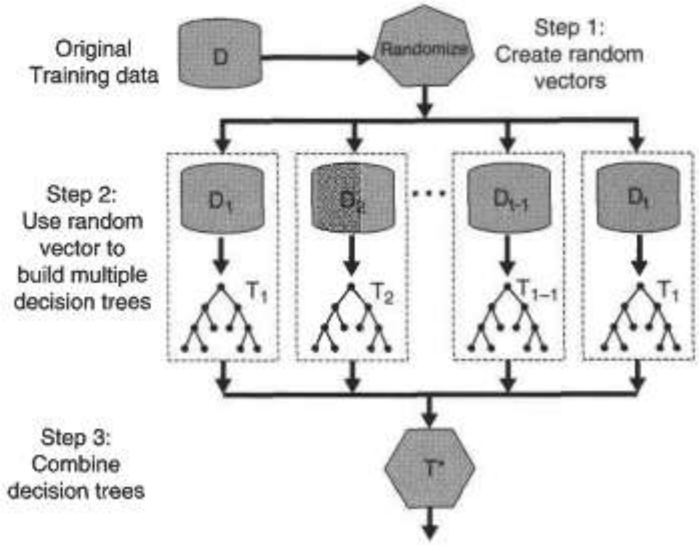

Original Training data
D
Randomize
Step 1: Create random vectors
Step 2: Use random vector to build multiple decision trees
D1
D2
Dt-1
Dt
T1
T2
T1-1
T1
Step 3: Combine decision trees
T*

Figure 29. Random forest (Tan et al., 2006, p. 292)

3.3.4. Research Activities

Figure 30 illustrates the research activities that are conducted as part of this dissertation. Finex's experts, a panel of external experts, a machine (a laptop and the main server of PSU), and the researcher (me) are the main players in the research process. Each performs a distinct set of activities.

3.3.4.1. Finex Activities

As shown in Figure 30, the research started with multiple meetings at Finex to determine the product features (PF) and the enabling technologies (ET). The meetings for the product features and for the enabling technologies were held separately. We discuss product features (PF) first. Once the product features are established, we proceed to work on the enabling technologies.

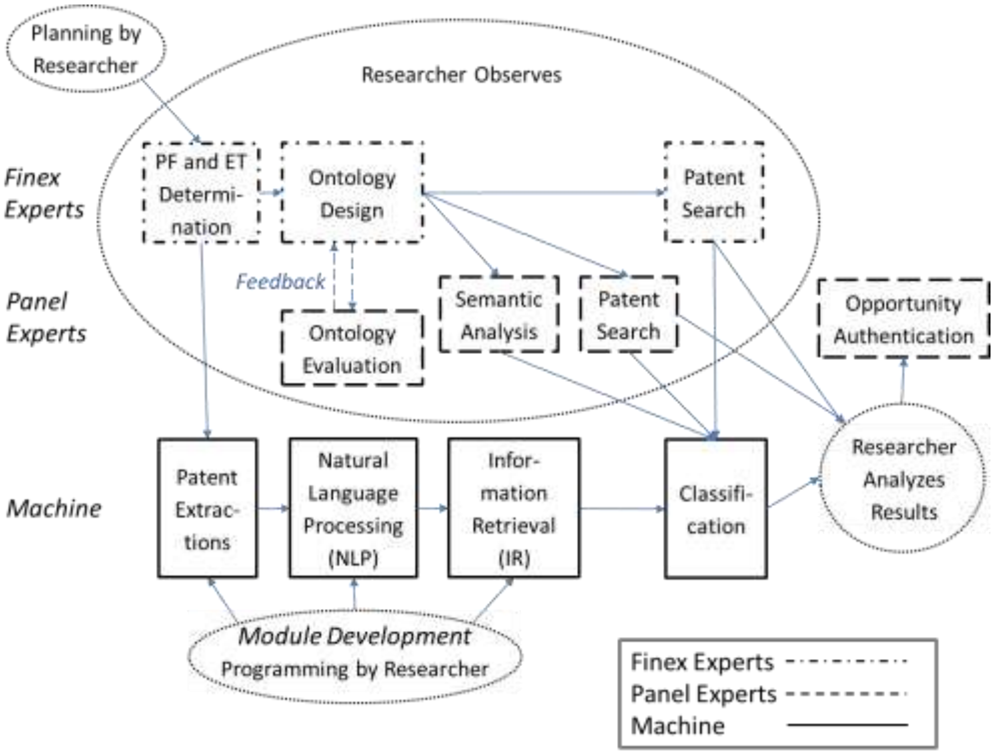


Figure 30- Research Activities

After determining the product features and the enabling technologies, Finex's experts design the ontologies for the PF and ET. The ontologies are revised multiple times until a consensus of satisfaction is achieved. Then, Finex's experts proceed with the patent search in the USPTO, in order to find US patents relevant to the ontologies. Again, the patent search sessions for PF and ET are conducted separately. In this endeavor, the Finex experts learn new keywords for PF and ET. The Finex experts expand the ontologies to improve the accuracy of their model.

3.3.4.2. Expert Panel Activities

In addition to Finex experts, expert E1, a panel of experts participated in the research for two purposes: 1) patent search (explained in section 3.4.3.), 2) ontology evaluation (explained in section 3.7.1.), and opportunity authentication.

Expert ID	Degree	Experience	Job
E1	Bsc	<10 years	Product Engineer
E2	PhD	>10 years	Professor
E3	PhD	>10 years	Professor
E4	Msc	>10 years	Product Engineer Manager
E5	PhD	<10 years	Product Engineer
E6	Msc	<10 years	PhD student
E7	Msc	<10 years	PhD student

Table 7- Profiles of the members of the expert panel

To select the members of the expert panel, two criteria are considered: 1) academic knowledge in materials science, and 2) professional experience in casting or foundry. More details about the expert panel and their performances are available in section 4.2. The general profiles of the participants in the panel members are shown in Table 7. Two are professors, two are PhD students, and two are professionals with a background in materials science.

3.3.4.3. Machine Activities

Machine activities consist of patent extraction, natural language processing (keyword extraction), information retrieval, and classification. The result of patent extraction is explained in section 4.3, and the results of information retrieval and classification are explained in sections 5.1 and 5.2.

3.3.4.4. Researcher Activities

The researcher plays three main roles. Firstly, the researcher plans the meetings with the experts. Then he observes the performance of the experts and gathers the required data. Secondly, the researcher provides all required modules for patent extraction, natural language processing, information retrieval, and classification. All modules are programmed in the Python environment. Thirdly, the researcher analyzes the results of the classifications and the results of the patent search. The researcher concludes which of the classifiers are the best and most appropriate for opportunity identification. Also, he compares the performance of the human experts to the performance of the machines in the patent classification.

3.4. Data Collection

The main activities of the data collection procedure are interviews that serve as the basis for designing ontologies; patent extraction from database; and keyword extraction by Natural Language Processing (NLP). The details of each of these activities are explained in the following sub-sections of this section.

3.4.1 Interviews

In this research, I conduct semi-structured interviews with experts who design the ontologies of product features and enabling technologies. I steer the interviews, but the experts generate the information based on their own expertise and knowledge. I meet twice with each expert. There are two sets of meetings to design the ontology of the product feature and of the enabling technologies. Each set contains at least two meetings. In the first meeting, the expert and I talk and discuss about what a product feature or enabling technologies is. Then, the expert creates a draft of the ontology for the product features or the enabling technologies. The expert performs a patent search about the ontology in the second interview. He/she revises the ontology in the next sessions based on his/her learning during the patent search. The details of each step are given below. They follow the procedure outlined by Jetter (Jetter, 2006).

1- *Identification of experts*: Experts are people who possess substantially more experience than average in a narrow field of expertise (Jetter 2006, p. 69). These experts are usually relatively easy to identify within their organizations because of their outstanding reputations (Jetter 2006, p. 69). I have a contact person in Finex under study. He identifies the experts for me. I subsequently ask the contact person to introduce me to the experts in marketing, R&D, product design and engineering, and other related expertise in the organization that is required for me to design the ontology that represents the collective expertise of these experts.

2- *Activation and capture of knowledge*: As mentioned above, the product features and the enabling technologies that pertain to the product under development are assumed to be understood by the experts within the company. To start knowledge elicitation session, I ask the expert to talk a little about the product features and the enabling technologies. To stimulate his knowledge, I obtain as much documented knowledge as possible from the company about the product to be developed. I use this documented knowledge as an ‘icebreaker’ to activate the interviewee, who subsequently volunteer information concerning product features and enabling technologies that pertain to the product. After activation, I focus the interview around specific questions about the product features and enabling technologies. The subjects of the questions about enabling technologies are type of technology, functionality, main components and parts, design features, materials and the manufacturing process. The subject of questions about product features are about application, shape, ergonomic attributes, material, functions, etc. Then, I ask the interviewee to write down his/her answers.

3- *Knowledge interpretation and Documentation*: In this step, the experts design the ontologies based on the procedure offered by Gavrilova *et al.* (Gavrilova et al., 2005). After the first round of the interview sessions, the expert has a textual description of product features and enabling technologies. The expert provides a glossary of the keywords, and then tentatively designs the high level of hierarchies of the product features or the enabling technologies. In the second round of sessions, the expert applies the ontology in a patent search to experience how the ontology can effectively address the ontology. He comes up with some points to modify the ontology. The expert finalizes the ontology by eliminating redundancies, extraneous synonyms, and contradictions. More details of ontology design are presented in section 3.4.2.

3.4.2 Ontology Design

As mentioned before, the purpose of this research is to find those opportunities in USPTO patents which address two specific concepts: 1) product features, and 2) enabling technologies. Ontology is applied to address these two concepts in my research.

There is a general agreement about the main components of an ontology. They are instances, classes, attributes, relationships, and hierarchical structure (Maglia, 2006). *Instances*, also called individuals, are the most basic component of ontology. Instances represent actual, concrete objects (e.g., animals, bones, cars, etc.) Ontology does not require the inclusion of instances, but a main purpose of ontology is to provide a means of classifying individuals, even if those instances are not explicitly part of the ontology. *Classes*, also called *concepts*, represent abstract groups, sets, or collection of objects. *Attributes*, also called *properties*, represent features, characteristics, or parameters that objects can have and share. Objects are described by assigning attributes to them. *Relationships* represent ways that objects interact with one another. The most common form of relation is '*is_a*'. *Hierarchical structure* is inherently the form of a classification system defined by *relationships* between *classes*. Most commonly used forms of relation in hierarchical structures are *is_a* and *part_of*. The ontology of the C programming language is illustrated in Figure 31 as an example of ontology.

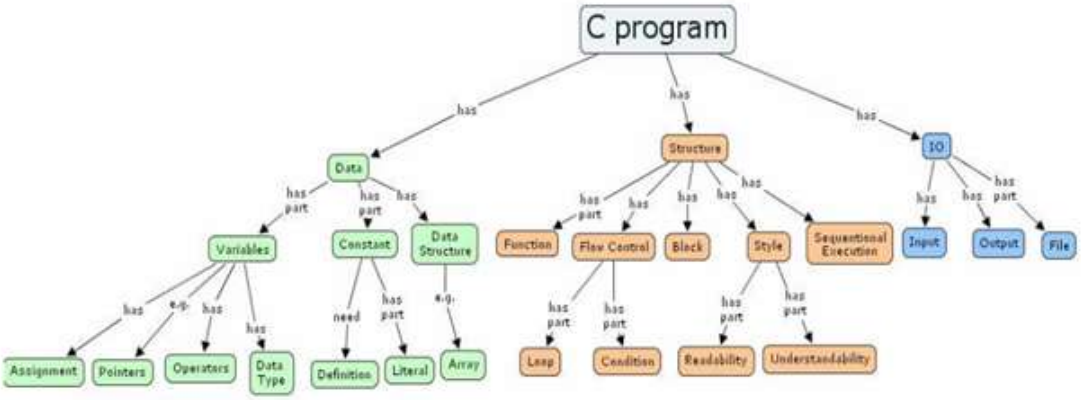


Figure 31- An example of ontology (Gavrilova et al., 2005)

All methodologies for designing ontologies contain the following general steps (Gavrilova et al., 2005):

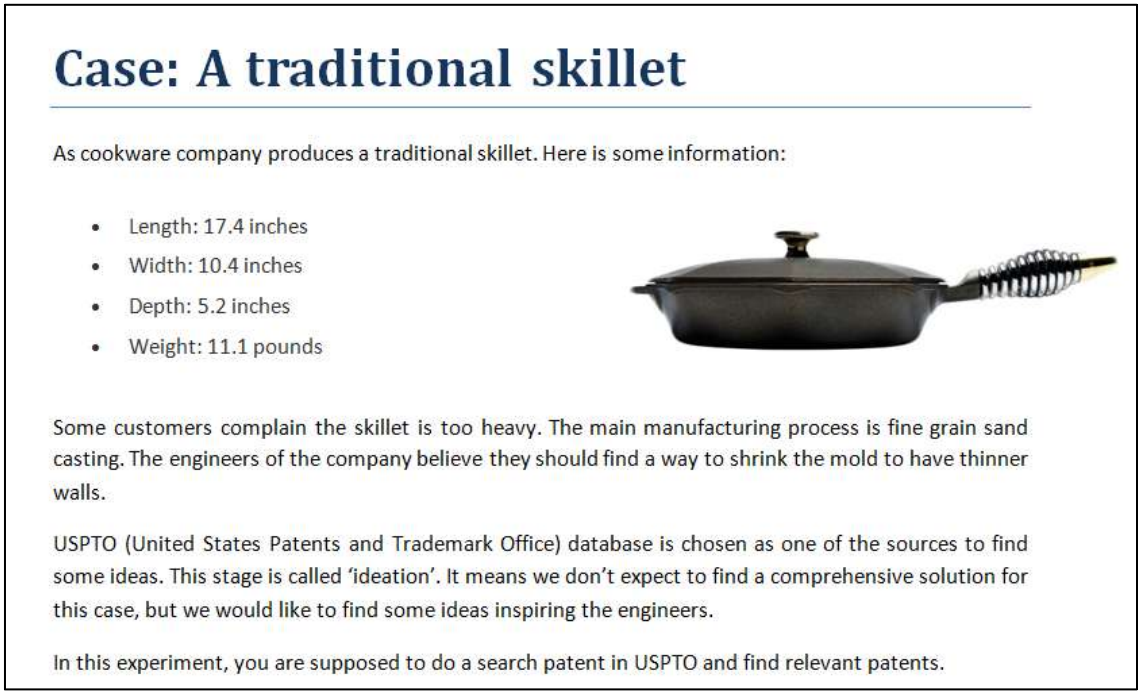
- *Glossary development*: gathering relevant information to select and describe all essential objects and concepts the domain.
- *Laddering*: defining and visually representing the high level of hierarchies among the concepts.
- *Disintegration*: hierarchically structuring the detailed concepts via a top-down strategy.
- *Refinement*: updating the visual structure by excluding the excessiveness (eliminating redundancies), synonymy, and contradictions.

3.4.3. Patent Search

The patent search is designed to observe and assess the performance of the experts versus of the classifiers in the RC, OC, and OSC scenarios. In addition, the search behavior of the members is studied to be able interpret their search performance. More details are available in sections 5.2.1 and 5.2.2. The main activities of the patent search experiment are:

1. introduction to the case
2. introduction to Google patent search engine
3. patent search
   - search behavior observation
   - search result observation

An ontology evaluation follows the patent search.



Case: A traditional skillet

As cookware company produces a traditional skillet. Here is some information:

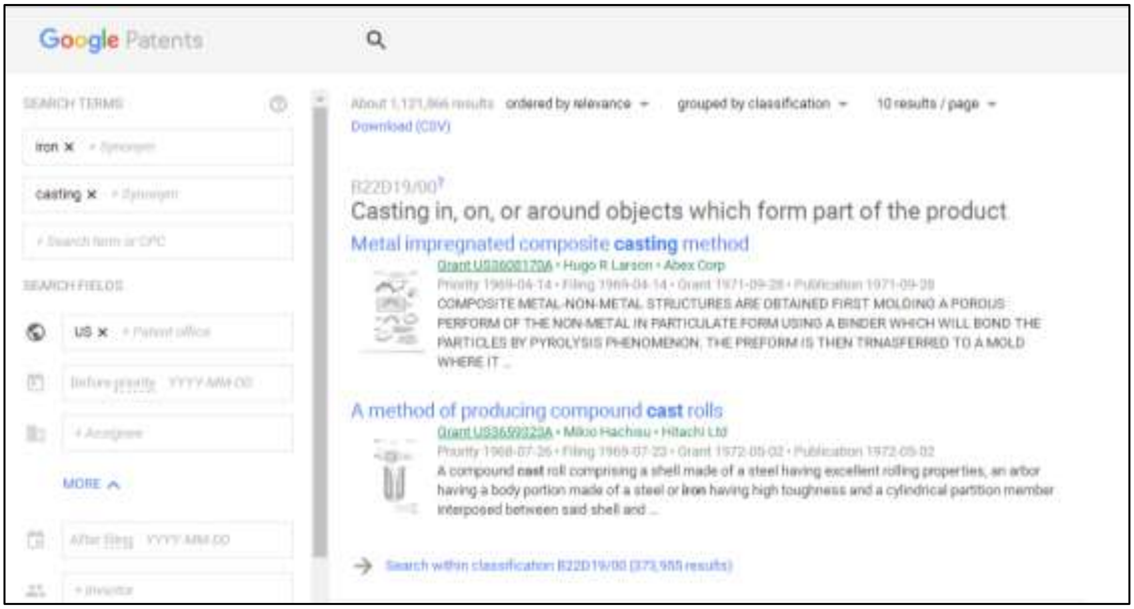
- Length: 17.4 inches
- Width: 10.4 inches
- Depth: 5.2 inches
- Weight: 11.1 pounds

Some customers complain the skillet is too heavy. The main manufacturing process is fine grain sand casting. The engineers of the company believe they should find a way to shrink the mold to have thinner walls.

USPTO (United States Patents and Trademark Office) database is chosen as one of the sources to find some ideas. This stage is called 'ideation'. It means we don't expect to find a comprehensive solution for this case, but we would like to find some ideas inspiring the engineers.

In this experiment, you are supposed to do a search patent in USPTO and find relevant patents.

Figure 32- Case introduction page



B22D19/00

Casting in, on, or around objects which form part of the product

Metal impregnated composite **casting** method

Grant US3608170A • Hugo R Larson • Abex Corp

Priority 1969-04-14 • Filing 1969-04-14 • Grant 1971-09-28 • Publication 1971-09-28

COMPOSITE METAL-NON-METAL STRUCTURES ARE OBTAINED FIRST MOLDING A POROUS PERFORM OF THE NON-METAL IN PARTICULATE FORM USING A BINDER WHICH WILL BOND THE PARTICLES BY PYROLYSIS PHENOMENON. THE PREFORM IS THEN TRNASFERRED TO A MOLD WHERE IT ...

A method of producing compound **cast** rolls

Grant US3659323A • Mikio Hachisu • Hitachi Ltd

Priority 1968-07-26 • Filing 1969-07-23 • Grant 1972-05-02 • Publication 1972-05-02

A compound **cast** roll comprising a shell made of a steel having excellent rolling properties, an arbor having a body portion made of a steel or **iron** having high toughness and a cylindrical partition member interposed between said shell and ...

Search within classification B22D19/00 (373,955 results)

Figure 33- A snap shot of Google patent search engine

The researcher engages in one-on-one interviews with each panel member. At the beginning of the patent search session, the Finex case is introduced to each individual expert on the panel via the page shown in Figure 32. The ontologies of the product features (light weight product) and the enabling technologies (thin wall iron casting) are introduced to the experts as well. Then, about five minutes are spent with the panel member, so that he/she could practice operating the Google patent search engine (www.patents.google.com). Some search tips are taught to the experts, so that they have enough skills to do a patent search efficiently on the Web site. Once the experts have acquired the requisite knowledge and skills, they start doing patent searches on their own. Two types of information are collected during the patent search:

- experts' search results including appropriate query numbers, patent numbers, relevant/irrelevant patents

- experts' search behavior including appropriate query numbers, keywords used in each query, starting and ending time of each query

3.4.4. Patent Extraction

To find, extract, and collect patents from USPTO patents database as the main source of this study, there are two approaches: 1) keyword-based approach, and 2) classification-based approach. In the first, keyword-based approach, the researcher either enters keywords into the USPTO Web site ("USPTO Database, Boolean Search," 2016) directly or introduces a more complicated query into the advanced-search Web page ("USPTO Database, Advanced Search," 2016). The query is built from keywords, Boolean operators, wildcard symbols, punctuation and special characters (*Search Help*, 2016). The search result is displayed as hyperlinks and the researcher can navigate them by clicking on the links. The researcher can save and download the Web pages of the patents one by one. If the researcher comes up with a large number of patents, (s)he faces three difficulties: 1) this process is long and time-consuming, 2) the downloaded data are unstructured and the researcher has to spend more time to structure the data in a database to be able to analyze them, and 3) (s)he may miss some patents due to inaccuracy or limitation of the keyword selection.

In the second, classification-based approach, the researcher chooses one of the available classification systems: either CPC or USPC[9]. In these classification systems, patents are hierarchically classified in terms of their technical features. Therefore,

[9] United States Patent Classification

patent researchers can search patents by using the codes (classification symbols) rather than keywords. In this approach, the researcher comes up with broader results, i.e. more patents in the sample. The researcher consequently loses some accuracy because he/she doesn't use keywords. Still, he/she suffers from the time-consuming downloading and structuring patents that characterize the keyword-based approach.

In this study, I collect my data in the classification-based approach by applying Cooperative Classification System (CPC). I look for those components of the CPC that address the two main concepts of my research: 'product features' and 'enabling technologies'. Once I follow these steps to extract the patents:

- Selecting CPC sections and their classes with the participation of experts as described in section 3.5.1.
- Extracting all patents under the selected CPC classes, i.e. downloading the patents and collecting them in a text file.

3.4.5. Natural Language Processing

Natural language processing (NLP) extracts all words from patents. Therefore, a patent is indexed according to a long list of keywords. The main steps of NLP are tokenization, stemming, Part-of Speech (POS), and parsing, which have already been explained in section 2.2.1. In this research, I only use the tokenization and stemming steps; I do not use the POS and parsing steps because I have constructed my research based on the relation between the keywords and ontologies, rather than the role of keywords in their sentences.

Before starting using the keyword lists, they still need to be purified. There are many keywords such as ‘the’, ‘as’, ‘is’, ‘the’, and ‘which’ that they are too frequented in the sentences and they lead to poor index terms (Fox, 1989). This type of keywords, called *stop words*, are not important for information retrieval purposes, so NLP methods apply a list of stop words to remove them from keyword lists (Wilbur & Sirotkin, 1992). A list of stop words is suggested for general texts in (Fox, 1989). In my research, I treat stop words as a control variable.

3.4.6. Semantic Analysis

Semantic analysis stands on two pillars: 1) the meaning of a word (lexical semantic analysis), and 2) the grammatical role of a word in a sentence such as verb, subject, object, etc. (formal semantic analysis) (Vikner & Jensen, 2002). Formal semantic analysis is used for applications like automatic translation, where the knowing the grammatical role of words is crucial. In my research, I apply lexical semantic analysis because classification in my research relies on the frequency numbers of keywords rather than their roles in a sentence. I use WordNet as the source of synonyms, as it is the most widely used lexical database for English semantic analysis.

Scenario III in my research, Ontological Semantic Classification, requires a list of synonyms for each keyword that results from the ontology (see also section 3.3.2). Therefore, I have two sets of lists of synonyms: one set will represent the ‘product features’ ontology and the other represent the ‘enabling technologies’ ontology. These synonyms are ‘stemmed’ (see section 2.2.1) by borrowing functions from natural language processing, because I need stemmed synonyms lists to calculate similarity between the ontology vectors and the keyword vectors.

3.5. Data Analysis

Data analysis occurs automatically. It consists of two steps: information retrieval and classification which is a type of machine learning method. Data analysis is the same for all scenarios (Regular Classification (RC), Ontological Classification (OC), and Ontological Sematic Classification (OSC)) except for the keywords that are classified in these scenarios. Raw keywords enter the information retrieval system in all three scenarios, and regular keywords comprise the output of information retrieval. Information retrieval also puts out information pertaining to the regular keywords, such as frequency numbers and TF-IDF. In RC, regular keywords are classified according to patterns that emerge from the three classifiers that are deployed (k-NN, SVM, Random forest, see section 3.3.3). OC and OSC do not use the regular keywords; keywords in these scenarios are provided by ontology, plus, in the case of OSC, the synonyms of these ontologies. Information retrieval only provides the frequency numbers and TF-IDF. The outputs of machine learning, and thus the output of data analysis, are the classifications of patents and the performance measures associated with these classifications.

3.5.1 Information Retrieval

To remove highly unimportant keywords, a cut-off value is applied. The cut-off value is empirically determined according to the frequency numbers of the keywords in the corpus. Applying cut-off value reduces the number of keywords, and the computations, consequently. In addition to applying the cut-off value, a term weighting scheme is deployed to lessen the impact of less important keywords. The term weighting in this research is based on the frequency of occurrence within a document (Luhn, 1958). Among different term weighting schemes developed, term frequency with inverse document frequency and length normalization have obtained the best recall and precision (Salton & Buckley, 1988).

3.5.1.1. Cut-off value

Keywords have different weight in describing an individual document. Some keywords are very determinant and important and some keywords are unimportant. To remove unimportant terms from the keywords list, it is necessary to apply a cut-off threshold value which is more introduced in section 3.6.3 as a moderator variable. Cut-off value is an empirically determined threshold value to discriminate between important and unimportant keywords in information retrieval. However, the cut-off value is identified empirically, Zipf's law (Zipf, 1949) (sometimes called the Power Law) can help to determine the cut-off value more systematically. Zipf law is an empirical law that shows many types of data including word frequency can be approximated by Zipfian distribution. Figure 34 illustrates some twelve reputed phenomena that follow Zipf's law (Newman, 2005). Based on Zipf's law, the frequency of a word is inversely proportional to its statistical rank (Newman, 2005). Similarly, term frequencies in patents follow Zipf's law. Murphy et al (J. Murphy, Fu, Otto, Jensen, et al., 2014) show

how the frequency of term in 61,000 patents follow the cumulative distribution of Zipf’s law.

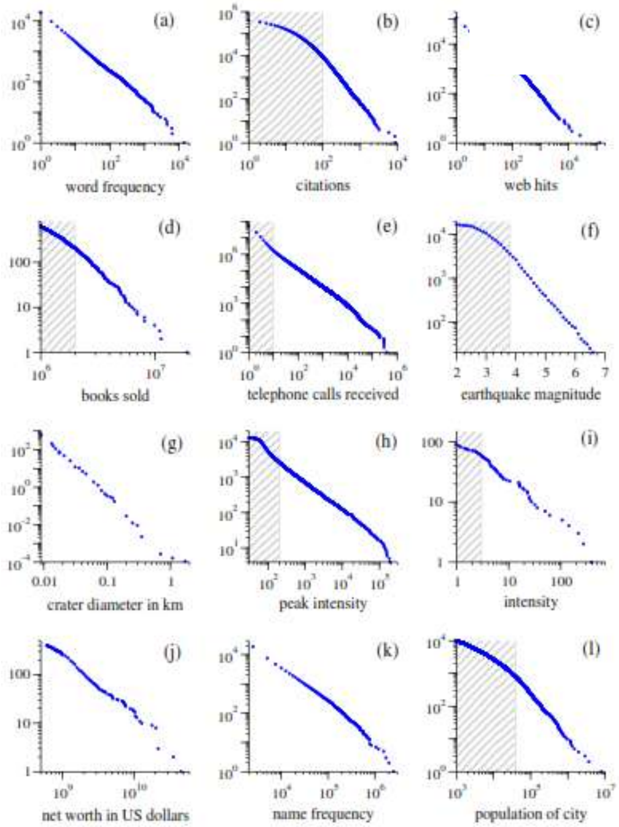


Figure 34- Cumulative distributions or “rank/frequency plots” of twelve quantities reputed to follow power laws (Newman, 2005)

3.5.1.2. TF-IDF

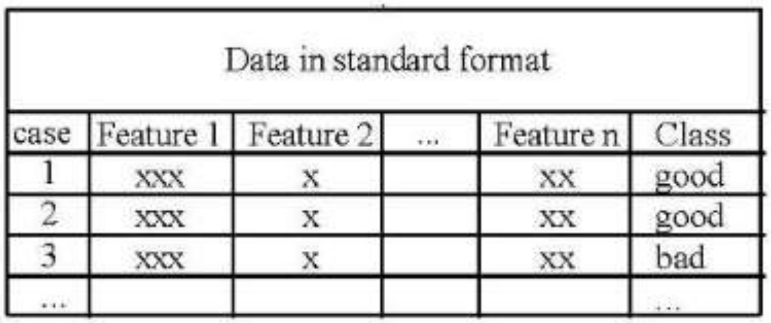
Term Frequency – Inverse Document Frequency (TF-IDF) is the most often used term weighting approach (Timonen, 2013). TF-IDF weights a given term based on how well the term describes an individual document within a corpus (Lott, 2012). TF-IDF positively weights a term for the *term frequency* which is the number of times that the term occurred within a specific document, and *inversely* weights the terms for the *document frequency* which is the number of documents which comprise the term (Lott, 2012). The document frequency inversely impacts the weigh in order to diminish the importance of highly frequented terms since this type of terms cannot describe an individual document within a corpus. The following formula is applied to calculate TF-IDF:

$$w_{ij} = \frac{tf_{ij} \times \log(\frac{N}{df_i})}{\sqrt{\sum_j tf_{ij}^2}}$$

where $w_{i,j}$ is the weight for term *i* in document *j*, *N* is the number of documents in the corpus, $tf_{i,j}$ is the term frequency of term *i* in document *j* and df_{i} is the document frequency of term *i* in the corpus. As it is shown in the formula, w_{ij}'s are normalized by the Euclidean distance of $\mathrm{tf}_{\mathrm{ij}}$'s.

3.5.2. Classification

In classification or supervised machine learning, briefly introduced in section 2.2.3, researchers investigate algorithms to produce general hypothesis which are reasoned from externally supplied instances, an example shown in Figure 35, in order to predict future instances (Kotsiantis, 2007). The goal of classification is "*to build a concise model of the distribution of class labels in terms of predictor features*" (Kotsiantis, 2007). The classifiers assign class labels to the testing data set where the values of the predictor features are known, but the value of the class label is unknown (Han, Kamber, & Pei, 2011).



Data in standard format					
case	Feature 1	Feature 2	...	Feature n	Class
1	xxx	x		xx	good
2	xxx	x		xx	good
3	xxx	x		xx	bad
...					...

Figure 35- Example of training data set (Kotsiantis, 2007)

Classification, like other machine learning methods, inductively learns a set of rules from instances which are examples in a training set. In another word, classification algorithms create classifiers (models) that can be used to generalize from new instances (Kotsiantis, 2007).

3.5.2.1. k-fold Cross Validation

k-fold Cross validation is a process used to study the generalizability of classification. In k-fold cross-validation, the data set is partitioned into k equal-sized segments (Tan et al., 2006, p. 187). Over each run, one of the segments is chosen for testing while the rest are used for training (Tan et al., 2006, p. 187). This procedure is repeated k times so that each segment is used for testing exactly once. The total error is found by summing up the errors for all k runs (Tan et al., 2006, p. 187). A sample of a 5-fold cross validation is shown in Figure 36. The cross validation will be iterated 100 times by shuffling the data set to avoid overfitting. To study the performance of the three classifiers, the average of the performance measures, introduced in section 3.5.5., are calculated after 100 iterations.

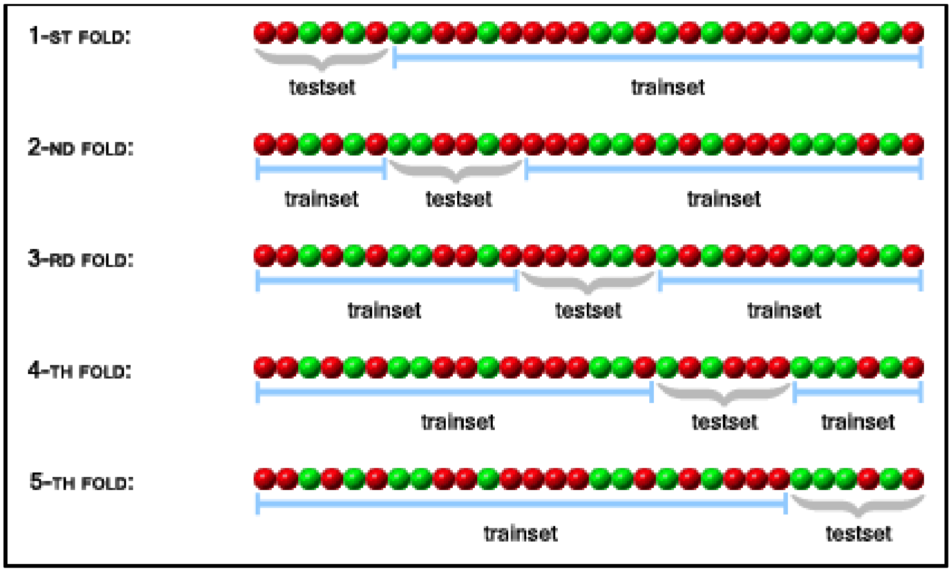


Figure 36- A sample of a 5-fold cross validation

3.6. Variables and Measures

The variables to be used in the proposed study are divided into five categories: independent variables, moderating variables, mediating variables, control variables and dependent variables. Researchers study to predict or explain the variations of ***dependent variables***, while they change the ***independent variables***. ***Moderating variables*** influence or moderate the relation between two or more other variables and thus produce an interaction effect. That means the relationship between two variables depends on the value of the moderator. The moderator either strengthens or weakens the *relationship* between the predictor and outcome (Peyrot, 1996). ***Mediating variables*** explain all or part of the relationship between two variables and provide a *causal link* between them (Peyrot, 1996). ***Control variables*** are extraneous variables that an investigator does not wish to examine in a study. Thus the investigator controls this variable. For example, controlling for gender means examining the original relationship (say between family functioning and diabetes control) separately among male and female (Peyrot, 1996).

Figure 37 illustrates the relationships between all variables in the research in this dissertation. Each variable is described in the following subsections.

3.6.1. Independent Variables

Independent variables come out of the interviews with experts. For all scenarios, independent variables include the patent classes of the CPC that pertain to the product under development, both from the point of views of product features and enabling technologies. In OC and OSC, ontological keywords also constitute input variables. Both input variables are expressed as lists.

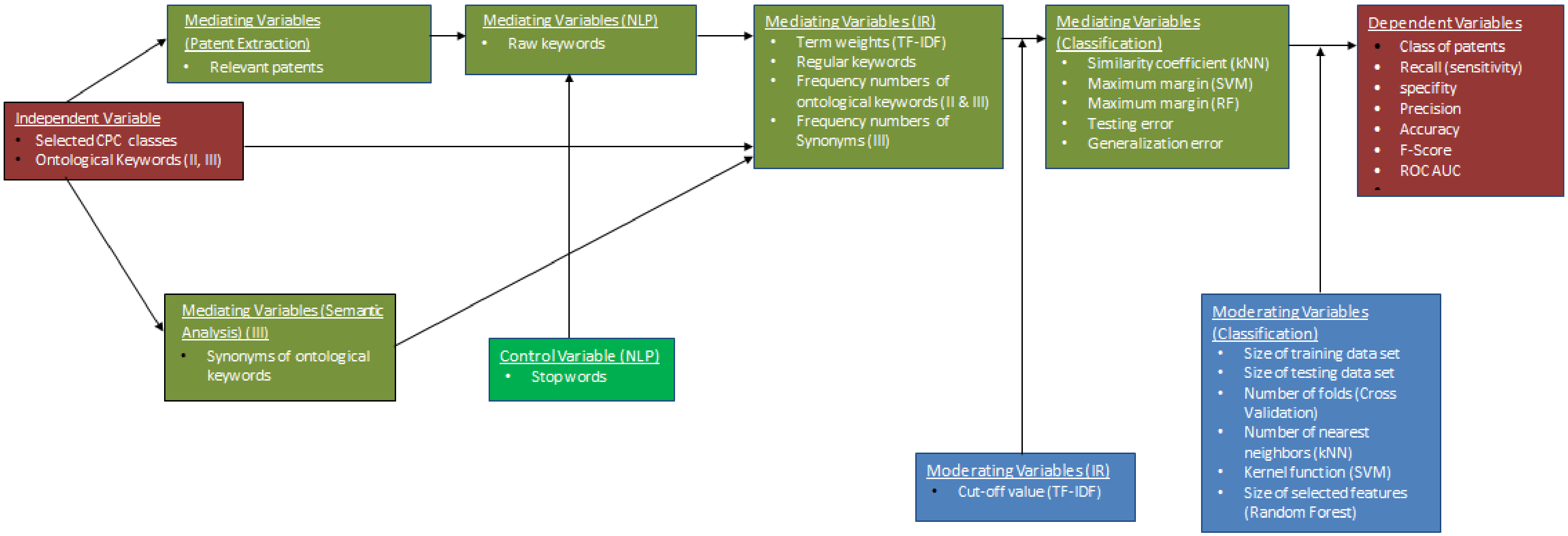

Mediating Variables (Patent Extraction)
• Relevant patents
Mediating Variables (NLP)
• Raw keywords
Mediating Variables (IR)
• Term weights (TF-IDF)
• Regular keywords
• Frequency numbers of ontological keywords (II & III)
• Frequency numbers of Synonyms (III)
Mediating Variables (Classification)
• Similarity coefficient (kNN)
• Maximum margin (SVM)
• Maximum margin (RF)
• Testing error
• Generalization error
Dependent Variables
• Class of patents
• Recall (sensitivity)
• specifity
• Precision
• Accuracy
• F-Score
• ROC AUC
Independent Variable
• Selected CPC classes
• Ontological Keywords (II, III)
Mediating Variables (Semantic Analysis) (III)
• Synonyms of ontological keywords
Control Variable (NLP)
• Stop words
Moderating Variables (IR)
• Cut-off value (TF-IDF)
Moderating Variables (Classification)
• Size of training data set
• Size of testing data set
• Number of folds (Cross Validation)
• Number of nearest neighbors (kNN)
• Kernel function (SVM)
• Size of selected features (Random Forest)

Figure 37- Research framework with variables

3.6.2. Mediating Variables.

Figure 37 shows that different components of the research process are associated with different mediating variables. In this section, mediating variables are introduced based on where they are created in the research process.

3.6.2.1 Mediating Variables Pertaining to Data Collection

- *Relevant patents:* the list of patents extracted from the CPC classes selected by experts.
- *Raw Keywords:* a list of words extracted from the relevant patents during natural language processing.
- *Synonyms of ontological keywords*: a set of lists of keywords that is extracted from WordNet for the purpose of Ontological Semantic Classification (OSC, Scenario III). Every list contains a list of synonyms for each of ontological keyword.

3.6.2.2 Mediating Variables Pertaining to Information Retrieval

These variables are listed below:

- *Term weight (TF-IDF):* a measure to determine the importance of a regular keyword in all relevant patents. It weights a given term based on how well the term describes an individual document within a corpus (Lott, 2012).
- *Regular Keywords*. Information retrieval takes a list of raw keywords and filters out unimportant words. It discriminates according the cut-off value of TF-IDF. The list of remaining keywords, which are deemed important, is known as the list of

regular keywords. It is particularly important in Regular Classification (RC, Scenario I).

- *Frequency numbers of ontological keywords*: a vector that represents the frequency of occurrence of ontological keywords in relevant patents. Every dimension of the vector denotes the frequency occurrence of ontological keywords in a particular patent. This variable is particularly important in OC (Scenario II) and OSC (Scenario III).
- *Synonyms of ontological keywords*: a set of vectors of synonyms of ontological keywords that are extracted from WordNet. Each ontological keyword may have zero or more synonyms. Each vector of synonyms is associated with one and only one ontological keyword.
- *Frequency numbers of synonyms*: a vector that represents the frequency of occurrence of synonyms of ontological keywords in relevant patents. Every dimension of the vector denotes the frequency of occurrence of synonyms of ontological keywords in a particular patent. This variable is particularly important in OSC (Scenario III).

3.6.2.3 Mediating Variables Pertaining to Classification

The following mediating variables pertain to classification. They are specific to particular classification schemes (*k*NN, SVM or random forest).

- *Similarity coefficient* (kNN): Given two relevant patents P_1 and P_2, their corresponding weighted feature vectors are W_1 and $W_{2.}$ The similarity between patents P_1 and P_2 is computed as below (Salton & Buckley, 1988):

$$S(P1,P2) = cos(P1\ ,P2) = \frac{P1\ .\ P2}{\|P1\ \|\|P2\ \|}$$

$$= \frac{\sum_{i=1}^{n} d_{1i}.\ d_{2i}}{\sqrt{\sum_{i=1}^{n}(d_{1i})^2}\ \sqrt{\sum_{i=1}^{n}(d_{2i})^2}}$$

where d_{1i} and d_{2i} are the frequency numbers of regular keywords in P_1 and P_2 and $i=1\ldots,n$ and n is the number of regular keywords.

- *Maximum margin (SVM):* The main principle of Support Vector Machine (SVM) is to determine the best separator (hyperplane) of different classes in the search space where the data set is also linearly separable (Aggarwal & Zhai, 2012). The distance between the two hyperplanes is called the *margin* of the classifier. Maximum margin refers to the maximum distance between the two hyperplanes.
- *Maximum margin (random forest)*: Random forest is an ensemble methods specifically developed based on decision trees classifiers. The strength of a set of classifiers is the average of the classifier's margin which is determined as below (Tan et al., 2006, p. 291):

$$margin,\ M(X,Y)\ = P(\bar{Y}_\theta = Y)\ \text{ - } \max_{Z \neq Y}(\bar{Y}_\theta = Z)$$

where $\bar{Y}_\theta$ is the predicted class of X according to a classifier built from some random vector θ.

- *Testing Error (kNN, SVM and random forest):* the rate of classification errors committed on a training data set.

- *Generalization error (kNN, SVM and random forest):* the expected error rate that the classifier yields when it is applied to the test data set (Tan et al., 2006, p. 172).

3.6.3. Moderating Variables

Moderating variables either pertain to information retrieval or classification. The variable 'cut-off value' is associated with information retrieval. All other variables are associated with classification. All moderating variables associated with classification also influence the relationship between the list of regular keywords and the performance measure of the classifier (recall, precision and F-score).

- *Cut-off value:* an empirically determined threshold value that discriminates between important and unimportant keywords in information retrieval based on TF-IDF. The cutoff-value affects the relationship between raw keywords and regular keywords and the number of regular keywords that are retrieved.
- *Size of training set:* the number of patents provided by experts for training the classifier. The size of training set moderates the relationship regular, ontological or ontological semantic keywords (depending on scenario) on the one hand, and the performance measures (recall, precision and f-score) on the other hand.
- *Size of test set:* the number of patents provided by experts for testing the performance of the classifier. The size of training set moderates the relationship regular, ontological or ontological semantic keywords (depending on scenario) on

the one hand, and the performance measures (recall, precision and f-score) on the other hand.

- *Number of folds*: the number of subsets of the patents provided by experts in cross validation. The size of training set moderates the relationship regular, ontological or ontological semantic keywords (depending on scenario) on the one hand and the performance measures (recall, precision and f-score) on the other hand.
- *Number of nearest neighbors (kNN):* the number of nearest patents to be compared to a particular patent for the purpose of classification. (The patents will have sets of regular keywords deemed the most similar by the kNN algorithm.) The number of nearest neighbors moderates the relationship between the similarity coefficient and the class of patents.
- *Kernel function (SVM)*: a similarity function that the domain expert provides to a machine learning algorithm. It determines the similarity between two sets of keywords in the training set. The first set of keywords is contained in patents that have been classified as related to a specific concept like product features or enabling technologies ('1'). The second set of keywords is contained in patents that have been classified as unrelated to a specific concept like product features or enabling technologies ('0'). The kernel function moderates the relationship between the maximum margin and the class of patents.

- *Size of selected feature (random forest):* the number of keywords (may be randomly chosen) used to generate decision trees. The size of selected feature moderates the relationship between maximum margin and the class of patents.

3.6.4 Control Variable: Stop Words

Keywords are the feature of the data set (patents) in my research. The only features (or keywords), which are controlled in my research, are *stop words*. Stop words are keywords such as 'the', 'as', 'is', 'the', and 'which'. They occur in the sentences very frequently, and they lead to poor index terms (Fox, 1989). S*top words* are not important for information retrieval purposes, so I will apply a list of stop words to remove them from keyword lists during natural language processing. A list of stop words is suggested for general texts in (Fox, 1989).

3.6.5. Dependent Variable

The output of the patent classification comprises the dependent variable in this research. It is a binary variable. Every patent is binned in to one of two possible classes of patents (CoPs): patents that are related to a specific concept (CoP=1) or patents that are not (CoP=0).

Recall, precision and the F-Score constitute a set of metrics that in conjunction determine the quality of a classifier. (They are analogous to the R^2 in a multiple regression.) They are based on the number of true positives (TPs), false positives (FPs), true negatives (TNs) and false negatives (FNs) that the classifier generates. Classifications can either yield true positives, false positives, true negatives or false negatives.

- *Recall (r) or sensitivity:*

$$r = \frac{TP}{TP+FN}$$

- *Specifity (s):*

$$s = \frac{TN}{TN+FP}$$

- *Precision (p):*

$$p = \frac{TP}{TP+FP}$$

- *Accuracy (A):*

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

- The *F-score* represents a harmonic mean between *recall* and *precision:*

$$F = \frac{2\,rp}{r+p} \quad \text{or} \quad F = \frac{2}{\frac{1}{r}+\frac{1}{p}}$$

- ROC AUC: is the area under a receiver operating characteristic curve, i.e. ROC curve, which is created by plotting the true positive rate (TPR) against the false positive rate (FPR).

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

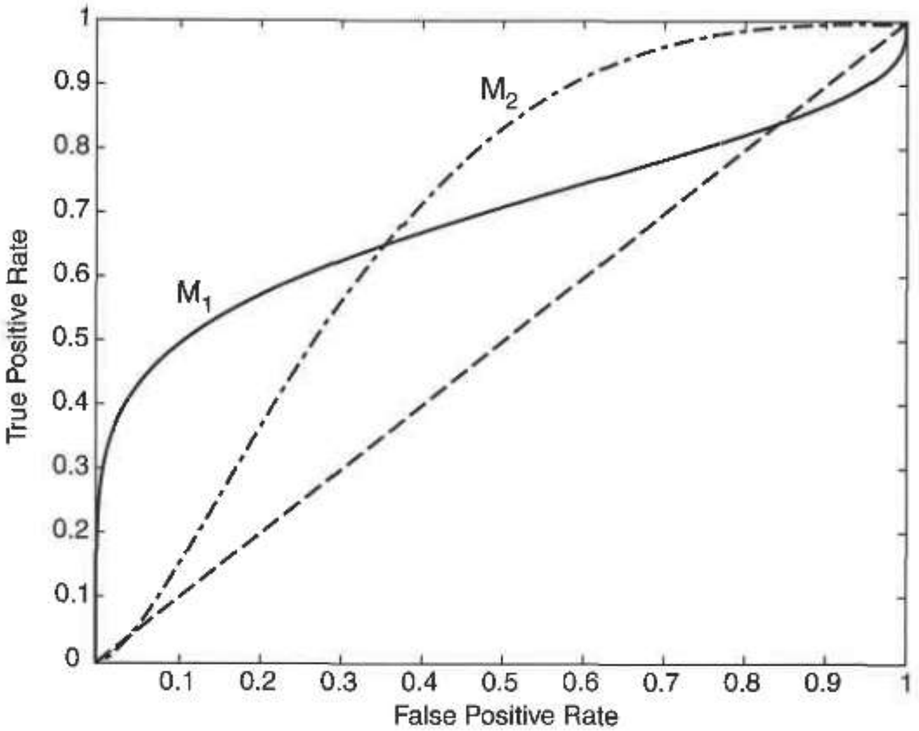


Figure 38- ROC curves for two different classifiers (Tan et al., 2006)

Figure 38 shows a sample ROC curve. A perfect classification model should be fitted to the upper left corner as much as possible. In theory, the best ROC AUC is 1. The main diagonal represents random guessing (chance) (Tan et al., 2006).

3.7. Validity and Reliability

In order to have viable research, it is important to show the *validity* and the *reliability* of the research design. “Validity refers to the extent to which a test measures what we actually wish to measure. Reliability has to do with the accuracy and precision of a measurement procedure.” (Thorndike & Hagen, 1986, p. 162). In other words, does a measurement (i.e. test, survey, observation, etc.) truly measure what it is intended to measure? The criteria that determine the validity of a research design are *content* *validity*, *criterion-related validity*, and *construct validity* (Kothari, 2004, p. 74) which are respectively discussed in sections 3.7.1, 3.7.2 and 3.7.3. *Reliability* has to do with the accuracy and precision of a measurement procedure (Thorndike & Hagen, 1986). It refers to the consistency of a measurement, i.e., a reliable measurement must yield the same results if it is repeated. Reliability is described in section 3.7.4.

3.7.1. Content Validity

Content validity measures the extent to which the questions provide adequate cover of the topic under study. Its determination is primarily judgmental and intuitive. It can be determined by using a panel of persons who shall judge how well the measuring instrument meets the standards, but there is no numerical way to express it (Kothari, 2004).

The content of my research consists of patents that address specific concepts, which are enabling technologies and product features determined by experts. Applying the ontologies of the concepts help more adequately cover the patents without considering all patents filed in the USPTO database. Therefore, having validated ontologies is a key to provide a valid content for the research. To validate the ontologies, experts *outside of Finex* validate the ontologies of enabling technologies and product features.

In order to validate the ontologies designed by Finex's experts, a semi-structured questionnaire (Appendix A) is designed based on seven criteria shown in Table 8. The questionnaire filled out by the external expert panel. They evaluated the ontologies after they did the patent search designed based on the Finex case. The patent search is explained in section 3.4.3.

Term	Definition
Accuracy	The criterion for determining is the asserted knowledge in the ontology agrees with the expert's knowledge about the domain. A higher accuracy will typically results from correct definitions and descriptions of classes, properties, and individuals.
Adaptability	Measures the ease of use of an ontology in different contexts possibly by allowing it to be extend and specialized monotonically, i.e. without the need to remove axioms
Clarity	Clarity Measures how effectively the ontology communicates the intended meaning of the defined terms.
Cohesion	From an ontology point of view, cohesion refers to the relatedness of elements in ontologies. It is intended to measure modularity. An ontology would have high cohesion if its classes are strongly related therefore, high cohesion is a desirable property.
Completeness	Measures if the domain of interest is appropriately covered. All questions the ontology should be able to answer can be answered.
Conciseness	Intended to reflect if the ontology defines irrelevant elements with regards to the domain to be covered or redundant representations of the semantics.
Consistency	Describes that the ontology does not include or allow for any contradictions.

Table 8- Criteria used for ontology evaluation (Hlomani & Stacey, 2014)

3.7.2. Criterion-Related Validity

A criterion is a measure used to determine the accuracy of a decision. The validity of a criterion is a measure of how well a variable or a set of variables predicts what it intended to measure and to approximate the truthfulness of the results (K. R. Murphy & Davidshofer, 2005). Criterion-related validity actually refers to (i) *concurrent validity*, and (ii) *predictive validity* (Kothari, 2004, p. 74). Concurrent validity refers to the usefulness of a test in closely relating to other measures of known validity. Predictive validity refers to the usefulness of a test in predicting some future performance. In my research, I validate the patent extraction (information retrieval) and the classifiers (classification methods) separately.

In patent extraction, I applied CPC classes, selected and validated by experts, to extract a set of patents that address enabling technologies or product features. Also, I randomly select 100 patents—50 patents for the training set and 50 for the test set. The panel of external experts also determines whether each of the 100 patents is related to product features or enabling technologies. Therefore, I validate training and test data sets for classification.

In classification methods, performance measures including recall, precision and F-score, are applied to assess the performance of the classifiers in terms of how they accurately identify true positive and negative instances in the training data set (concurrent validation), and how they accurately predict test data set (predictive validation).

Criterion validity in the dissertation research extends beyond concurrent and predictive validation. I also determined whether any of the approaches examined in this research performed as well as, or perhaps better than, current approaches, which rely exclusively on human judgement. To do so, I asked the panel experts to try to identify technological opportunities by considering the enabling technologies and product features that they have provided in the interview. I compared the technological opportunities identified by my three scenarios to those identified by the panel experts.

This determines whether any of the approaches studied in my research will perform as well as or better than current approaches to identify the technological opportunities.

3.7.3. Construct Validity

A *construct* is an initial concept, notion, question or hypothesis that determine *which data* is to be gathered and how it is to be gathered (Golafshani, 2003). A construct is an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories (Dean Brown, 2000). Construct validity has traditionally been defined as the experimental demonstration that a test is measuring the construct it claims to be measuring (Kothari, 2004, p. 74). Such an experiment could take the form of a differential-groups study, wherein the performances on the test are compared for two groups: one that has the construct and one that does not have the construct. If the group with the construct performs better than the group without the construct, that result is said to provide evidence of the construct validity of the test (Dean Brown, 2000).

According to my hypotheses mentioned in section 1.3, I will study the performance of the application of two constructs which are *ontology* and *semantics analysis*. In order to validate the performance of these two constructs, I will examine the performance of the classifiers (kNN, SVM, and random forest), with the presence of the constructs in scenarios II and III, and, with the absence of them in scenario I, to see how their presence will improve the performance of the classifiers based on measures such as recall, precision, and F-score measures.

3.7.4. Reliability

Reliability is defined as the extent to which results are consistent over time and the extent to which the constitute an accurate representation of the total population under study (Golafshani, 2003). Reliability is the consistency and repeatability of the measurement. The main measurement of my research is the performance of the classifiers, for which recall, precision and the F-Score act as reliability measures, which have been mentioned in section 3.4.5. The ROC-Curve (Tan, et al., 2006) has also been used as reliability measure for classification methods. They require me to keep track of the true positives, the false positives, the true negatives and the false negatives for all classifications.

3.8. Opportunity Identification

The performances of the three classifiers under study (*k*NN, SVM, random forest) is tested for two data sets (product features and enabling technologies), and three classification scenarios (RC, OC and OSC). The performances are measured by sensitivity (recall), specificity, recall, accuracy and F-score measures, and the reliability of the classifiers is determined by the area under the curve of Receiver Operating Character (AUC ROC) curve to decide which classifier can classify the patents more effectively.

Classifier	*Scenario I (RC)*			*Scenario II (OC)*			*Scenario III (OSC)*		
	Sensitivity	…	F-score	Sensitivity	…	F-score	Sensitivity	…	F-score
k-NN									
SVM									
random forest									

Table 9- Sample table to compare the performance of the classifiers under the three states

A sample table for recording these results for a specific data set is shown in Table 9. Four of these tables are created: one for each combination of data set (product features or technological attribute). I use the ROC AUC and F-Score as a metric for opportunity identification for each combination of data set, because they constitute a composite score of the other measures. I identify the best classifier in each of the scenarios (RC, OC, and OSC) that yields the highest ROC-AUC and F-score. After choosing the best classifier for classifying the patents for the product features and for the enabling technologies, machine classifies the patents separately for the product features and for the enabling technologies and I have two lists of patents classified. The intersection of the two lists represented the technological opportunities for the Finex case.

4.1. Results of Ontology Design

As mentioned previously, experts from Finex designed two ontologies: one for product features and another for enabling technologies. These are described in this section.

4.1.1. Ontology of the Product Features

As explained previously, Finex's customers complained about the heaviness of the skillets. Accordingly, the product features are represented by 'light weight product'. As shown in Figure 39, the 'light-weight product' refers to a product that:

- is made of cast iron which is thermally conductive.
- is a cookware device (e.g. pot, pan, plate, skillet, and griddle),
- or is a part or a components used in car (e.g. brake rotor, brake drum, brake disk, and engine block)
- and has a surface, or wall, or side which is thin or lightweight.

One might question why car parts and components are considered in the ontology. The reason is that car manufacturers encounter similar problems when they try to reduce the weight of iron cast parts and components that are used in a brake or an engine. In the process, they have generated many solutions to these problems. Therefore, considering this group of patents can increase the chances of finding opportunities relevant to weight reduction in many kinds of iron cast objects in other industries, including the manufacture of skillets.

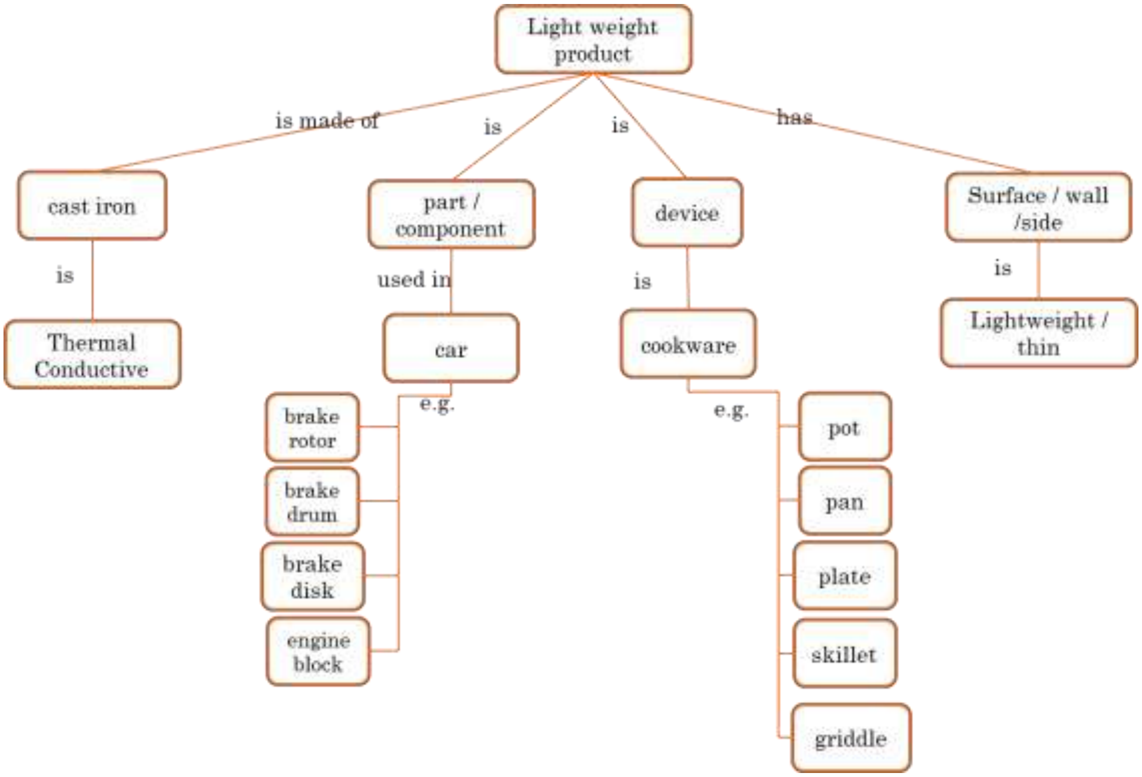


Figure 39- The ontology of the product features developed by Finex experts

4.1.2. Ontology of the Enabling Technologies

Finex experts believe mold shrinkage is the key factor for weight reduction, and five casting methods are considered as candidate technologies for weight reduction through mold shrinkage. The five methods are gravity casting, pressure casting, die casting, vacuum casting, and powder metallurgy. The concept of mold shrinkage, as well as the five methods of achieving it, is reflected in the ontology of thin wall iron casting shown in Figure 40. This ontology has been revised multiple times during ontology design sessions, and, as discussed in Section 4.1.2 and in Chapter 6, during ontology evaluations made by experts outside of Finex.

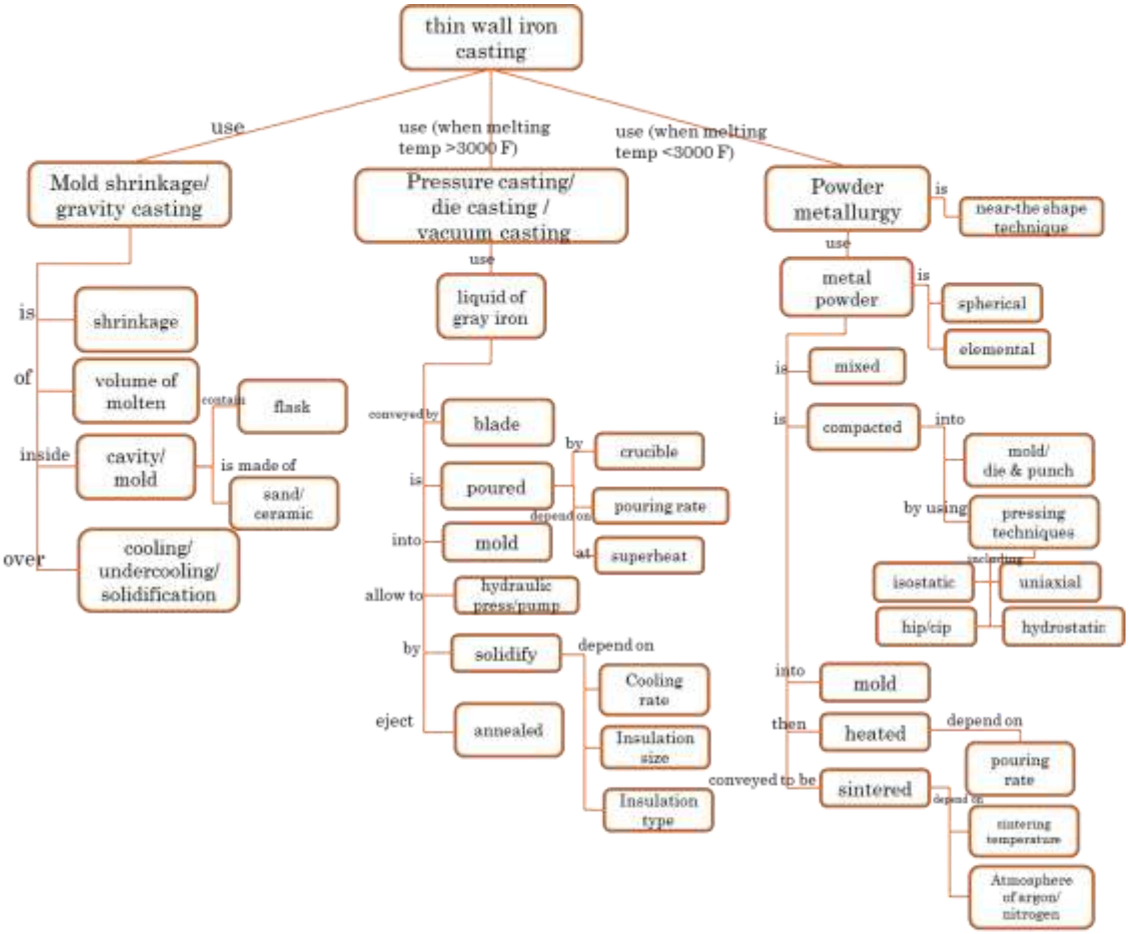
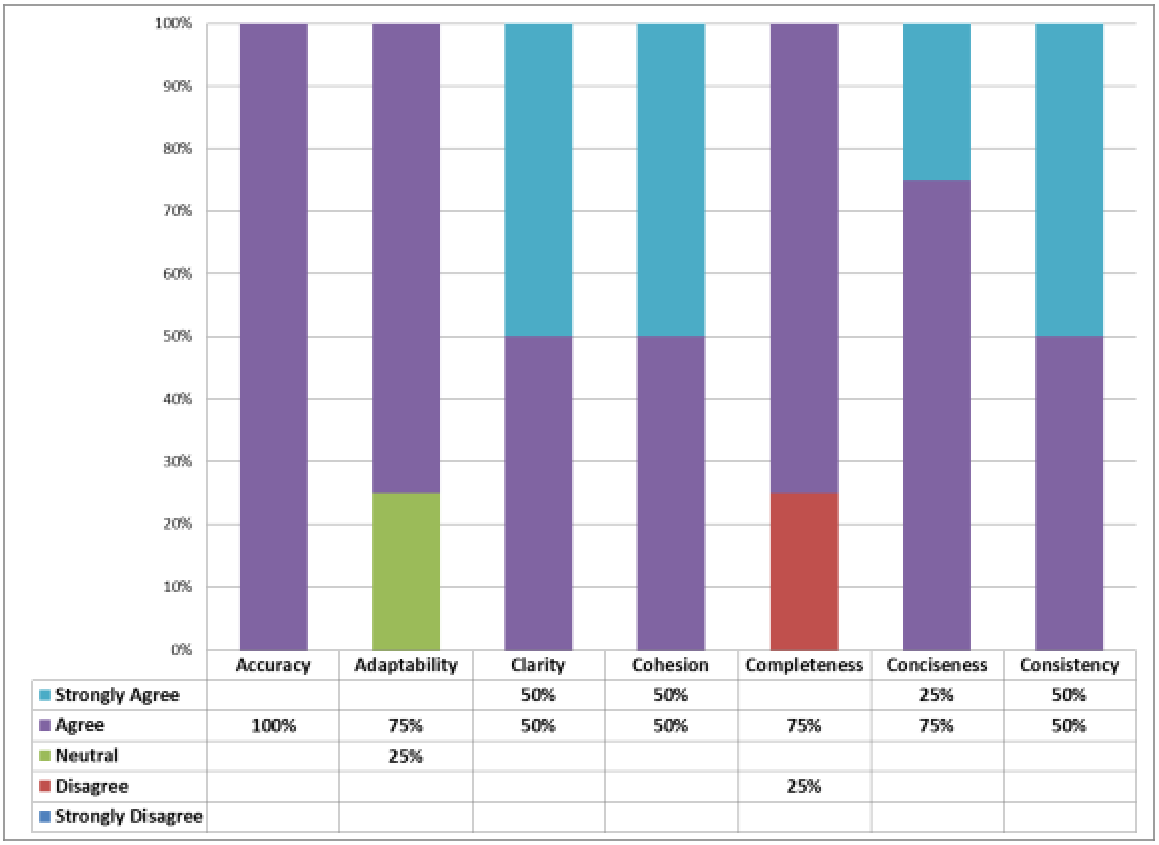


Figure 40- The ontology of enabling technologies (thin wall iron casting) developed by Finex experts

4.1.3. Ontology Evaluation

As explained in section 3.8.1., I designed a semi-structured questionnaire (Appendix A) to through which the members of the expert panel validated the ontologies generated by Finex. The questionnaire is built based on seven criteria including accuracy, adaptability, clarity, cohesion, completeness, conciseness, and consistency, which are defined in Table 8 (Hlomani & Stacey, 2014).

As shown in Figure 41, the panel experts agree or strongly agree that the ontology of thin wall iron casting is accurate based upon the seven criteria, except for one item pertaining to completeness. The external experts made minor modifications to the ontology of thin wall iron casting due to this evaluation.



	Accuracy	Adaptability	Clarity	Cohesion	Completeness	Conciseness	Consistency
🟦 Strongly Agree			50%	50%		25%	50%
🟪 Agree	100%	75%	50%	50%	75%	75%	50%
🟩 Neutral		25%					
🟥 Disagree					25%		
🟦 Strongly Disagree							

Figure 41- results of thin wall casting ontology evaluation

For the product features ontology (a light weight product), there is a consensus that this ontology is strongly accurate. All of the external experts respond with '*strongly agree*' to all seven criteria.

4.2. Patent Search

In this section, patent search behavior and patent search performance are presented. For this purpose, a group of measures are described, and are applied to the data gathered from the patent searches. The relationship between the performance measures and search behavior measures will be discussed in section 6.1.

4.2.1. Experts' Performance in Patent Search

Table 10 shows the basic data of the patent search experiment performed by Finex and six other experts. The participants spent between 40 and 60 minutes in one session; expert E1 spent 150 minutes in two sessions. The participants found a different number of patents within a range of 13 to 60. Some participants may have judged one patent at least two times. Such a patent is called a *duplicate*. Also, they may have judged one patent two times, but with opposite results (relevant and irrelevant). Such a patent is called a *contradiction*. The net number of patents is calculated by deducting the number of duplicates and contradictions from the number of patents judged.

Data	E1	E2	E3	E4	E5	E6	E7
Spent time (min)	150	50	40	60	43	53	60
Number of patents judged	25	38	14	60	13	15	19
Net number of patents	24	33	14	46	13	15	18
Number of duplicates	1	5	0	14	0	0	1
Number of contradictions	0	2	0	0	0	0	0
Number of relevant patents	8	13	12	5	8	8	13
Number of irrelevant patents	16	23	2	41	5	7	5

Table 10- Basic data of the performance of the experts in the patent search

4.2.1.1. Reliability

The reliability of a research instrument concerns the extent to which the instrument yields the same results on repeated trials (Golafshani, 2003). According to the definition of reliability, this concept is generalized to the patent search experiment. We expect that, when two experts use *similar keywords* for a similar case, they should come up with *similar results*. This means the two experts should find similar relevant or irrelevant patents. Therefore, in order to evaluate the reliability of a patent search, I extend the concept of reliability in the patent search as below:

$$reliability = 1 - failure\ rate$$

$$reliability = 1 - rate(finding\ unsimilar\ patents | using\ similar\ keywords)$$

$$= rate(finding\ similar\ patents | using\ similar\ keywords)$$

$$reliability = \frac{Average\ of\ similarity\ of\ patents\ judged\ in\ two\ patent\ searches}{Average\ of\ similarity\ of\ keywords\ applied\ in\ two\ patent\ searches}$$

To calculate the similarity measures mentioned in the reliability definition, the Jaccard index (Niwattanakul, Singthongchai, Naenudorn, & Wanapu, 2013) is used. The Jaccard index compares similarity between two sets A and B as shown below.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

In the numerator of the reliability equation, A and B denote the respective sets of patents judged by two different experts (a and b), regardless of the results of the judgments (relevancy or irrelevancy). In the denominator of the reliability equation, A and B denote the respective sets of keywords applied by two different experts (a and b). It should be stressed that the reliability measure is a *ratio*, not a *probability* function.

The similarity indexes between the patent searches are calculated based on the keywords used and based on the relevant patents found. The similarity indexes are shown in Table 11 and Table 12.

Experts	E1	E2	E3	E4	E5	E6	E7
E1	--	10%	5%	11%	14%	15%	12%
E2	10%	--	19%	29%	16%	35%	31%
E3	5%	19%	--	21%	10%	17%	24%
E4	11%	29%	21%	--	13%	23%	24%
E5	14%	16%	10%	13%	--	17%	28%
E6	15%	35%	17%	23%	17%	--	30%
E7	12%	31%	24%	24%	28%	30%	--
Ave. of similarity	**11%**	**23%**	**16%**	**20%**	**16%**	**23%**	**25%**

Table 11- Similarity indexes between the patent searches based on keywords used

Expert	E1	E2	E3	E4	E5	E6	E7
E1	--	0.0%	0.0%	1.4%	0.0%	2.6%	0.0%
E2	0.0%	--	9.5%	8.3%	0.0%	2.2%	0.0%
E3	0.0%	9.5%	--	1.7%	0.0%	0.0%	0.0%
E4	1.4%	8.3%	1.7%	--	1.7%	3.4%	0.0%
E5	0.0%	0.0%	0.0%	1.7%	--	0.0%	3.3%
E6	2.6%	2.2%	0.0%	3.4%	0.0%	--	0.0%
E7	0.0%	0.0%	0.0%	0.0%	3.3%	0.0%	--
Ave. of similarity	**0.68%**	**3.34%**	**1.87%**	**2.77%**	**0.84%**	**1.37%**	**0.56%**

Table 12- Similarity indexes between the patent searches based on the relevant patents found

In order to calculate the reliability measure, the average of the similarity indexes shown in Table 11 and Table 12 are utilized. The result of calculating the reliability measure is shown in Figure 42.

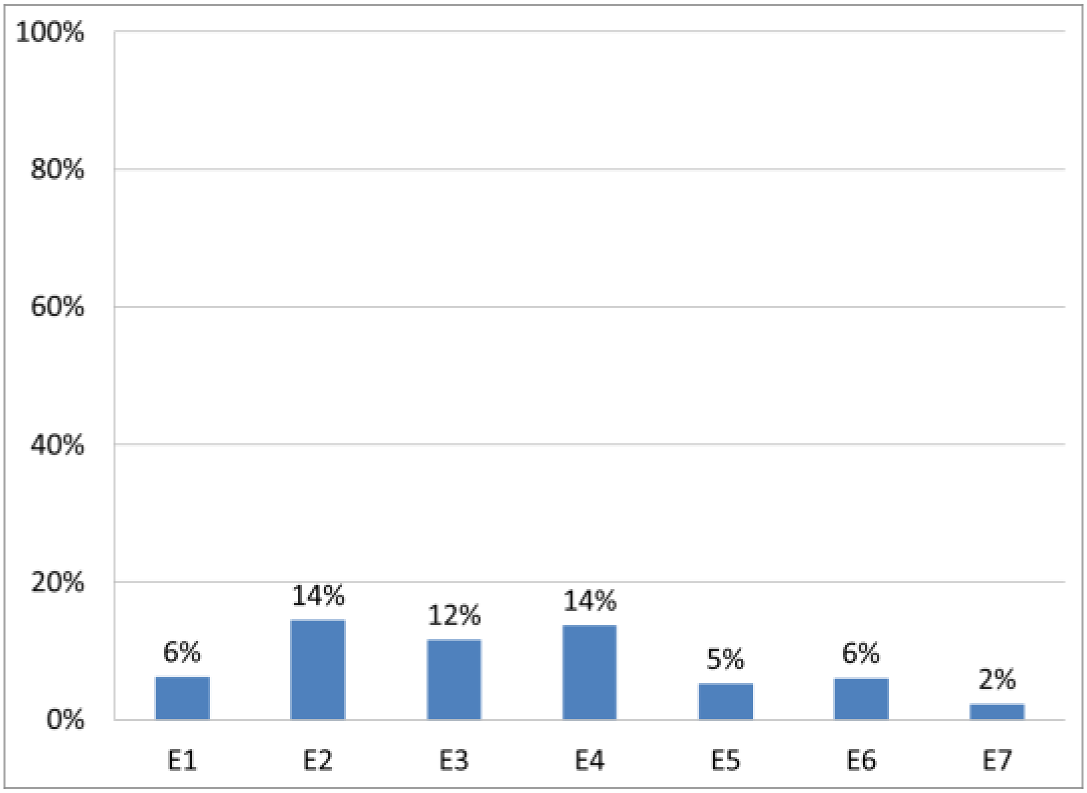


Figure 42- Reliabilities measured of the patent searches

4.2.1.2. Efficiency

Efficiency means how well resources are expended in a process (Machado & Davim, 2017). In a patent search, *time* is the main resource used to find relevant patents. Therefore, efficiency is defined in this research as below:

$$Efficiency = \frac{Number\ of\ relevant\ patents\ found\ in\ a\ patent\ search}{Total\ time\ spent\ (minute)}$$

As shown in Figure 43, the expert E2 and the expert E3 (the professors in the expert panel) have higher efficiency than the others, and the efficiency of expert E1 is the lowest.

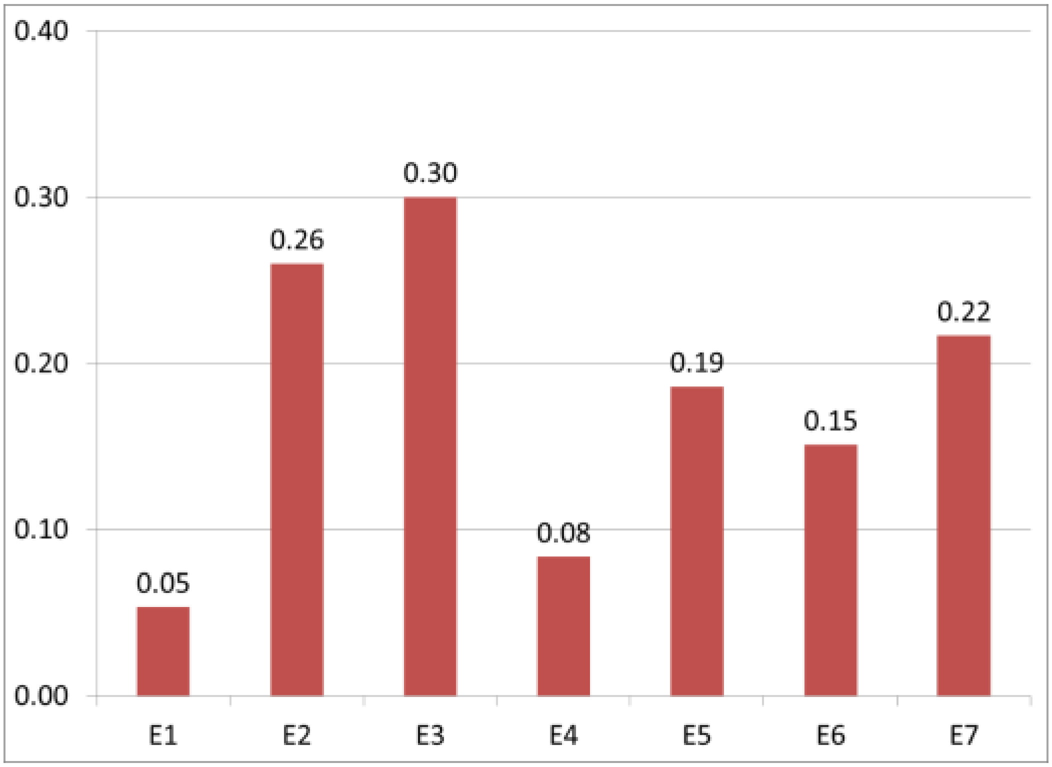


Figure 43- Efficiencies of the patent searches

4.2.1.3. Effectiveness

The concept of effectiveness refers to the degree of achieving possible objectives. Since there is no standard for the effectiveness of a patent search, the best performance in finding relevant patents among the experts is considered the base objective. The effectiveness of patent searches can be determined relative to that base objective. Therefore, the definition of effectiveness is given as follows:

$$Effectiveness = \frac{number\ of\ relevant\ patents\ found\ in\ a\ patent\ search}{number\ of\ patents\ judged\ in\ a\ patent\ search}$$

The effectiveness of the patent searches conducted by the experts is shown in Figure 44. The expert E3 has the best performance in finding relevant patents. E5, E6, and E7 have a moderate effectiveness, and the rest have weak effectiveness.

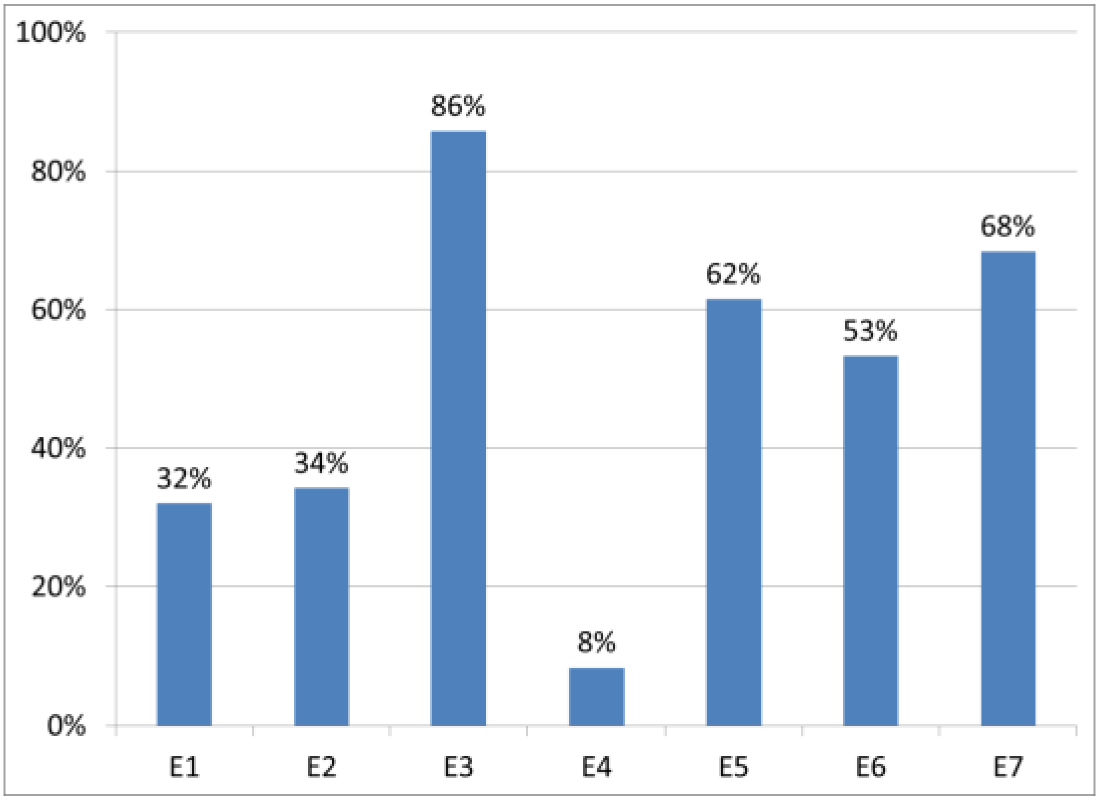


Figure 44- Relative effectiveness of the patent searches

4.2.2. Experts' Search Behavior

The raw data for the search behavior of all experts are shown in Table 13. The experts exhibit different behavior in terms of spent time, number of queries, total number of keywords, and total number of distinct keywords. Distinct keywords refer to all keywords used in the queries without considering their repetition.

Data	E1	E2	E3	E4	E5	E6	E7
Spent time (min)	150	50	40	60	43	53	60
Number of Queries	12	16	18	13	12	12	14

number of keywords	40	76	88	45	145	83	106
number of distinct keywords	19	15	22	12	22	20	20

Table 13- Basic data of the experts' search behaviors

An analysis of the search behavior of experts should help explain the experts' search results. This analysis will be discussed in detail in section 6.1. It is based upon, three different measures from above: the time spent, the keywords used, and the queries designed in the patent searches. The measures are keyword diversity, query complexity, and search speed.

4.2.2.1. Keyword Diversity

The keyword diversity measure is defined as follows:

$$keyword\ diversity = \frac{number\ of\ distinct\ keywords\ used\ in\ a\ patent\ search}{number\ of\ keywords\ used\ in\ a\ patent\ search}$$

This measure illustrates the variety of ways through which experts utilize their knowledge to use keywords in query design. In other words, this measure compares the innovativeness of the different experts in query design. As shown in Figure 45, all experts, except E1, use distinct keywords from 15% to 25%, whereas E1 uses 47.5% distinct keywords in their patent search. This result illustrates how E1 develops his queries by applying more diversified knowledge (keywords), meaning he is more innovative than the other experts.

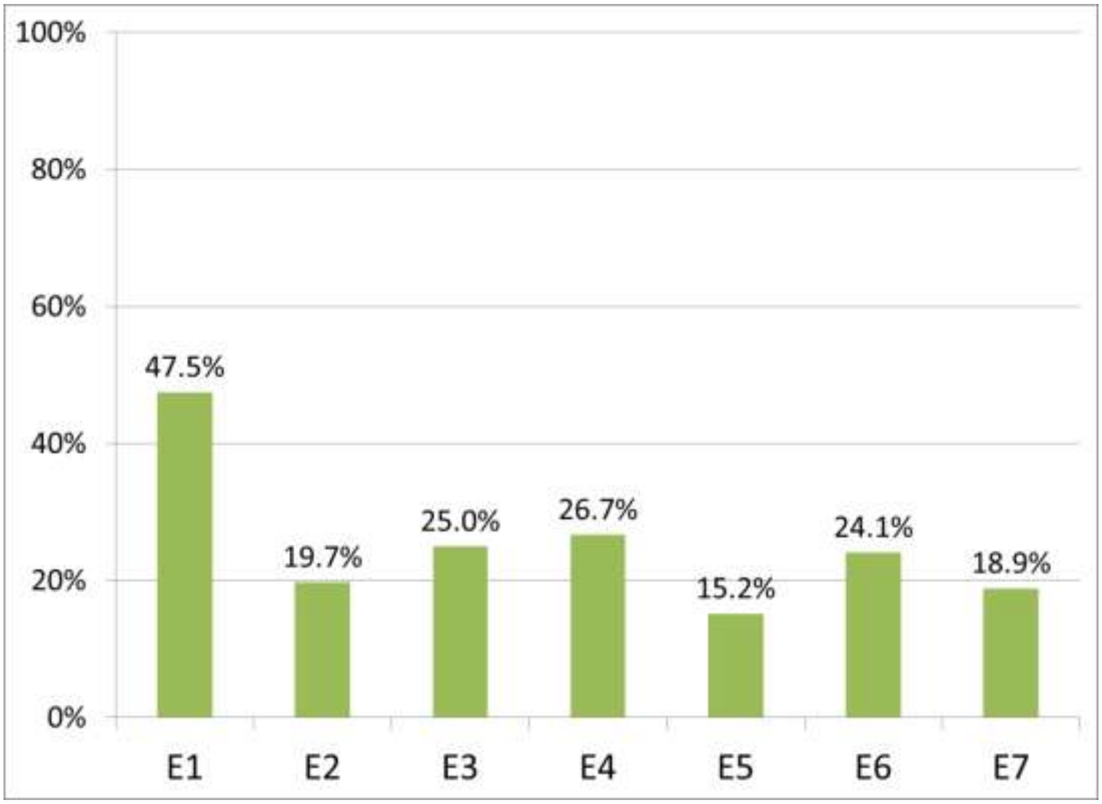


Figure 45- Keyword diversity in the patent searches

4.2.2.2. Query Complexity

The experts have different behaviors in the patent searches in terms of using the number of keywords in each query. To address this behavior, query complexity is defined as below:

$$query\ complexity = \frac{number\ of\ keywords\ used\ in\ a\ patent\ search}{number\ of\ qeries\ applied\ in\ a\ patent\ search}$$

This measure shows how complex queries are designed. The more keywords used in a query, the more complex the query is.

As shown in Figure 46, experts E1, E2, E3, and E4 have used 3 to 5 keywords per query. On the other hand, experts E5, E6, and E7 have used 7 to 12 keywords per query.

The first group includes experts E1E2, E3, and E4, and the second group includes experts E5, E6, and E7. This difference between the two groups shows that the more experienced experts used less complex queries.

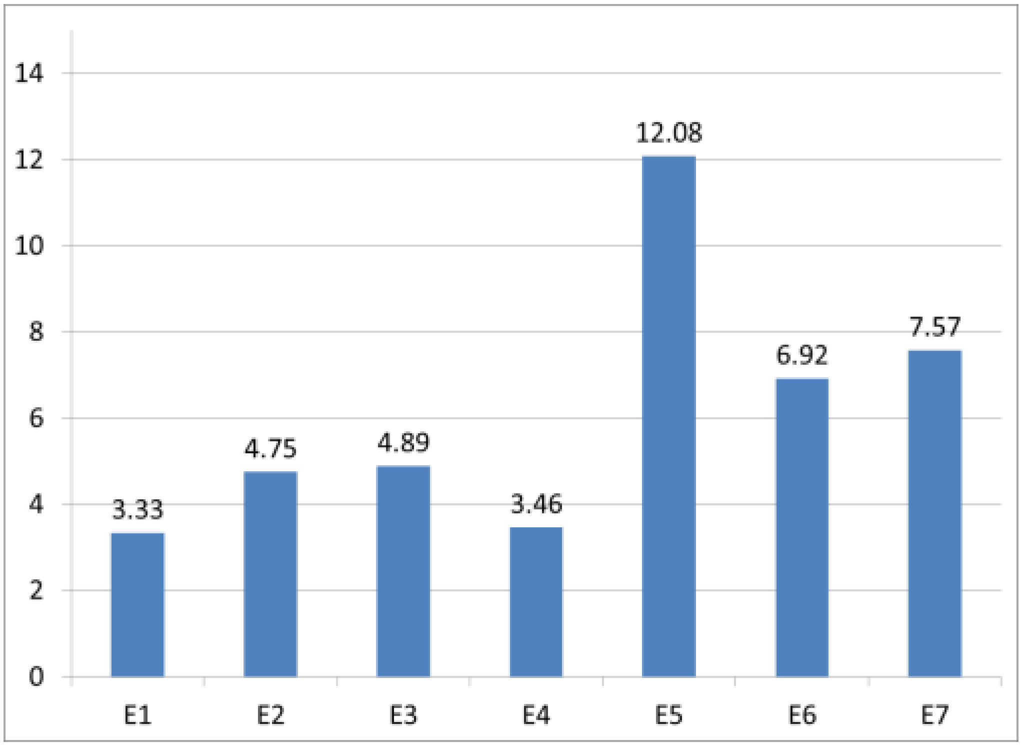


Figure 46- Query Complexity

4.2.2.3. Search Speed

Human experts constitute an expensive resource, so their time spent in a patent search should be considered in the study of the behavior of the experts in the patent searches. To do so, speed search is defined as below:

$$search\ speed\ = \frac{number\ of\ patents\ judged\ in\ a\ patent\ search}{time\ (minutes) spent\ to\ do\ a\ patent\ search}$$

As shown in Figure 47, the experts had different speed in patent search. Expert E1 has the lowest speed. If expert E1 would want to judge 100 patents, he would have to spend almost 10 hours.

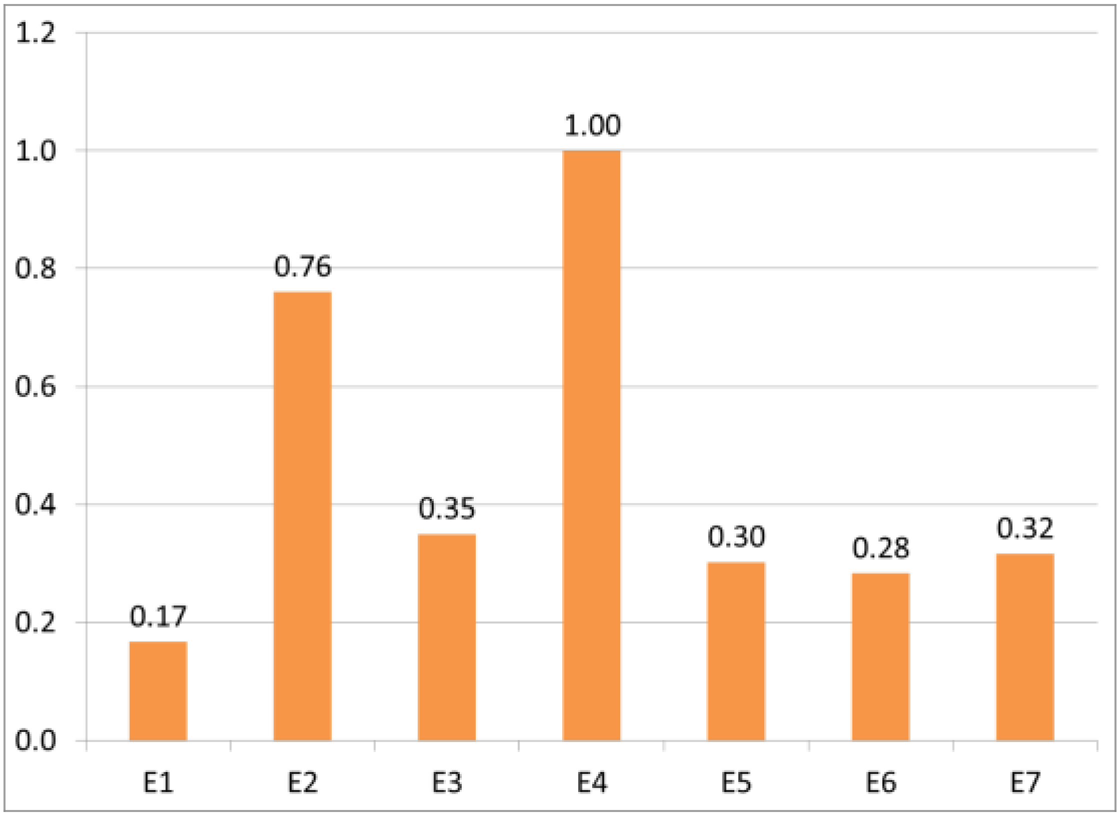


Figure 47- Search Speed in the patent searches

4.2.2.4. Error

Humans may make errors in activities like patent search quite frequently. The experts have demonstrated two different errors in the patent search. The may have reviewed a patent at least two time, called ‘duplicate’, and they may have judged a patent oppositely relevant and irrelevant in two independent judgements, called contradictions.

The error of the experts is calculated as below:

$$error = \frac{(number\ of\ duplicates + number\ of\ contradictions)\ in\ a\ patent\ search}{number\ of\ judged\ patents\ in\ a\ patent\ search}$$

According to Table 10 and the definition of error, the measured errors of the experts in patent searches are shown in Figure 48. The error of four of the experts is between 4% and 23%. Three panel members made no errors.

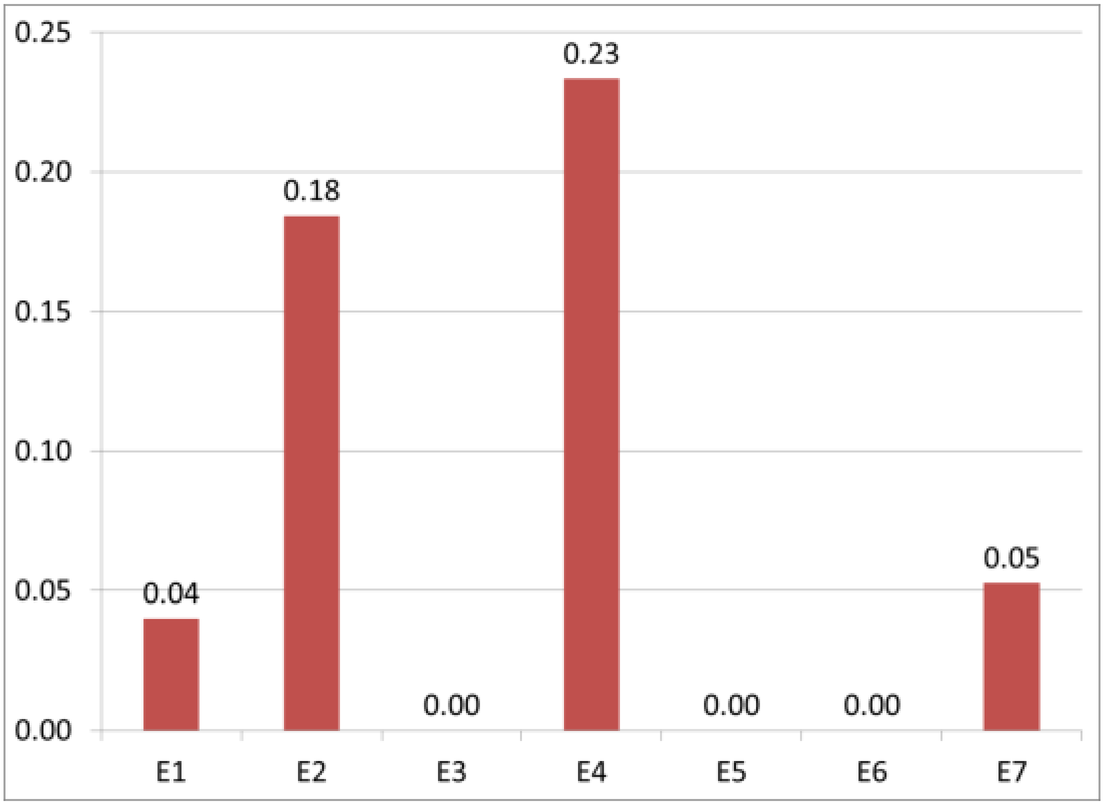


Figure 48- Error in the patent searches

4.3. Patent Extraction

As described in section 3.4.3., CPC classification is used to extract patents from the USPTO database, which are potentially relevant to ontologies. According to the two ontologies, two CPC sub-classes are chosen: sub-classes 'ALLOYS' (C22C) and 'CASTING

OF METALS’ (B22D). All C22C and B22D patents issued between 1/1/1991 and 1/1/2017 are downloaded from USPTO Website. The total number of downloaded patents is 19,525.

4.4. Natural Language Processing (NLP)

The 19,525 downloaded patents are processed by natural language processing and converted to a huge set of keywords. After filtering out the 728 stop words listed in Appendix B, 65,922 keywords are obtained. The keywords are stemmed by a snowball algorithm.[10] Finally, the term frequency numbers of the stemmed keywords are calculated in each patent and the frequency numbers are stored in a matrix in an Excel file.

4.5. Semantic Analysis

Three types of keywords are created in this research. First, raw keywords are extracted directly from patents downloaded from C22C and B22D sub-classes. The number of raw keywords is 65,922. Therefore, the matrix of term frequency of raw keywords has 19,525 rows (patents) and 65,922 columns (raw keywords). The second type of keywords is ontological keywords, which are extracted directly from the ontologies. The ontology of ‘light weight product’ has 27 ontological keywords, and the ontology of ‘thin wall iron casting’ has 98 ontological keywords. The ontological keywords are shown in Figure 39 and Figure 40. The third type of keywords is semantic ontological keywords, which consist of ontological keywords and their synonyms, which

[10] Snowball is a small string processing language designed for creating stemming algorithms for use in Information Retrieval.

are extracted from the WordNet database. The ontological keywords pertaining to the product features ontology ‘light weight product’ yielded 63 keywords that are synonyms. The ontological keywords pertaining to the ontology for the enabling technologies called ‘thin wall iron casting’ yielded 76 keywords that are synonyms. These semantic ontological keywords of the product features and of the enabling technologies are shown in Table 14 and Table 15, respectively. The synonyms that do not exist in the patents are not reflected in Table 14 and Table 15.

ontological keywords	Synonyms	ontological keywords	Synonyms	ontological keywords	Synonyms
light	lighter	device	gadget	conductive	conduct
	lightest		equipment		transmit
	lighten		tool		transfer
	Reduce		implement		convey
	reduction		instrument		dissipate
	lessen		means	Iron	ferrous
	lesser		article	pan	griddle
	decrease		apparatus	pot	--
	lightweight	plate	dish	part	--
weight	heaviness		platter	component	element
	weightiness	skillet	frypan		piece
	heft	surface	face	car	automobile
	mass		top		machine
cookware	utensil	wall	side	brake	--
	kitchen		rim	rotor	--
	cook		bowl	drum	--
	kitchenware	side	edge	disk	--
cast	mold		rim	engine	--
	mould	thin	slim	block	--
	pour		slender		
cook	make		narrow		
	food		thinwall		
	fix	thermal	warmth		
	ready		thermic		
	prepare		caloric		
frying	sautéing		heat		

Table 14- Ontological keywords of ‘light weight product’ and their synonyms

Ontological Keywords	Synonyms	Ontological Keywords	Synonyms	Ontological Keywords	Synonyms	Ontological Keywords	Synonyms
shrink	reduce	size	large	punch	slug	spherical	--
	contract		big	gray	grey	steel	--
	shrivel		measure	die	stamp	superheat	--
molten	melt		extent	quench	chill	temperature	--
	liquefied		scale	temper	tough	thick	--
cavity	hole		circumference		elasticity	thin	--
	pit		perimeter		snap	tool	--
cool	chill	melt	smelt	anneal	normal	type	--
	cold		liquidify		harden	undercool	--
	frigid		thaw	argon	--	uniaxial	--
	refrigerate		fuse	blade	--	volume	--
	freeze		dissolve	castiron	--	wall	--
	frost	powder	particle	ceramic	--		
solidify	solid		fine	cip	--		
	consolidate		dust	diecast	--		
	crystallize		talc/talcum	eject	--		
pressure	press	metal	alloy	element	--		
vacuum	vacant	compact	compress	flask	--		
	empty		dense	furnace	--		
	blank		condense	gravity	--		
	hollow		squeeze	hip	--		
liquid	liquidify	sinter	fuse	hydraulic	--		
	fluid			hydrostat	--		
	viscous	heat	hot	iron	ferrous		
pour	spill		warmth	isostatic	--		
	shed		warm	mold/mould	--		
	transfuse	atmosphere	ambiance	near	--		
	effuse		ambience	net	--		
	decant	cast	foundry	nitrogen	--		
rate	pace	crucible	pot	pump	--		
	speed	mix	blend	sand	--		
	velocity		intermix	shape	--		
insulate	heatproof		intermingle	slag	--		
	insulant		combine				

Table 15- Ontological keywords of ‘thin wall iron casting and their synonyms

4.6. Summary

This chapter has described how the data of this research are collected. The data include three components, which are 1) two ontologies, 2) two sets of patents judged by the Finex and panel experts, and 3) patents extracted from the USPTO database according to CPC classes selected by the Finex experts.

Two *ontologies* are designed by Finex experts to address the concepts of the product features and enabling technologies. These ontologies are designed in an ad-hoc process. They are reviewed and redesigned to characterize the concepts of product features and enabling technologies. These ontologies are validated an expert panel. (The panel also modified the ontology for enabling technologies. This activity is discussed in Chapter 6.)

A total of four sets of *judged patents* are formed in two different patent search experiments. The first experiment is conducted by Finex; the second by a panel of outside experts. Each group of experts classified the patents either as relevant or as irrelevant with respect to product features or the enabling technologies. The four sets of patents are used for training the machine learning classifiers.

The *extracted patents* form the corpus of this research. Thus, they constitute the main material for this research. The opportunities are supposed to be identified from the extracted patents.

In this chapter, the main results of data analysis are presented. The results are organized in the order in which the main activities of data analysis, which are information retrieval and classification, transpired.

5.1. Information Retrieval

As mentioned in section 3.7.1.2., *Term Frequency – Inverse Document Frequency* (TF-IDF) is chosen as a measure for information retrieval because it is the most often used term-weighting approach (Timonen, 2013). TF-IDF weights a given term based on how well the term describes an individual document within a corpus (Lott, 2012).

There are five TF-IDF matrices in this research. The first TF-IDF matrix contains the TF-IDF of regular keywords with 19,525 rows and 65,382 columns. Each of the ontologies has two TF-IDF matrices: one for ontological keywords, and one for ontological semantic keywords. The dimensions of the TF-IDF matrices are shown in Table 16.

TF-IDF matrix	Keyword type	Number of rows	Number of columns	Ontology	
				Light weight product	Thin wall iron casting
#1	regular	19,525	65,382	✓	✓
#2	ontological	19,525	30	✓	
#3	ontological semantic	19,525	72	✓	
#4	ontological	19,525	70		✓
#5	ontological semantic	19,525	141		✓

Table 16- The information of the five TF-IDF matrices

5.5.1. Cut-off value

Zipf law is an empirical law that shows many types of data including word frequency can be approximated by Zipfian distribution (Zipf, 1949). The cut-off value is an empirically determined threshold value that discriminates between important and unimportant keywords. Figure 49 shows the Zipf curve drawn based on the term frequency of 65,922 keywords in the corpus, which is the collection of patents downloaded from C22C and B22D sub-classes. There are a small number of keywords whose frequency is greater than 100,000. After reviewing the keywords and their frequencies, 20,000 is considered as the appropriate cut-off value for this case because this number is located in the middle of the area of generic terms as shown in Figure 49. Besides, a threshold of 20,000 keywords means that keywords whose average frequency number in the corpus is less than one will be removed. Remember, there are almost 20,000 patents (19,525 patents) in the corpus. Given the cut-off value, 540 keywords, less than 0.8% of the keywords, are removed. Since the curve of term frequencies of the keywords is too close to both axes in Figure 49, the logarithmic curve of the term frequencies is drawn as well. The logarithmic scale is shown in the second vertical axis in the right side of Figure 49.

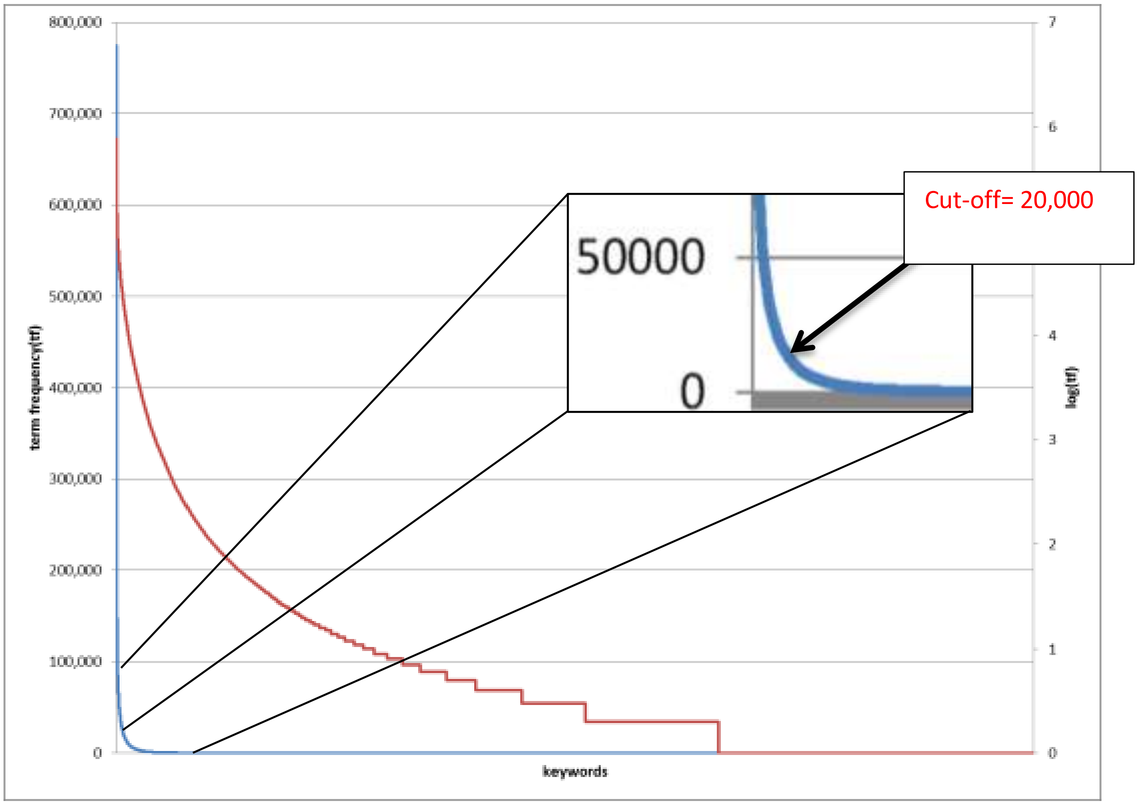


Figure 49- Zipf curve drawn based on the term frequency of all keywords in the corpus

5.2. Classification

As promised in chapter 3, K-NN, SVM, and random forest classifiers are applied to classify the three keyword types—regular keywords, ontological keywords, and ontological semantic keywords—for each of the two ontologies. These three types of classification are called regular classification (RC), ontological classification (OC), and ontological semantic classification (OSC). Therefore, the results of six classifications are presented in this section.

5.2.1 Training set and test set preparation

To train the classifiers and then examine their performance, it is required to provide two data sets: training set and test set. These sets are the results of the patent search by Finex experts. The patent search is performed in multiple sessions, and the experts have judged the patents as either relevant or irrelevant to the related ontology. Relevant patents are assigned ‘1’, and irrelevant patents are assigned ‘0’. The information of classified patents by Finex experts is shown in Table 17.

Ontology	**Number of relevant patents**	**Number of irrelevant patents**	**Total number of data set**
Thin wall iron casting (Enabling technologies)	17	36	53
Light weight product (product features)	29	63	92

Table 17- The data sets used for training and testing (classified patents by experts)

The training and test sets are created by setting the number of folds to 5 in the cross-validation process. For example, 53 classified patents of the technological attribute ontology are divided into five folds. Then, four folds (80% of the data) are considered for training the classifiers and one fold (20% of the data) is kept for testing the performance of the classifiers.

5.2.2 Cross Validation settings

The performance of the classifiers depends on:

- the combination of training folds (training sets);
- the combination of data in the training folds (training sets).

To avoid any random error, the cross validation process is repeated 100 times by:

- randomly shuffling the data sets before creating the folds;
- randomly shuffling the folds for each iteration of the cross validation.

The performance measures including specificity, sensitivity (recall), ROC AUC, accuracy, and precision are calculated based on the average number of measures in each iteration. The measures are defined in Table 18.

Measure	Definition	Formula
Sensitivity (recall*)*	the proportion of positives that are correctly identified	(TP)/(TP+FN)
Specifity	the proportion of negatives that are correctly identified	(TN)/(TN+FP)
ROC AUC	the area under the receiving operating characteristics (ROC) curve which created by plotting sensitivity against specificity	N/A
Accuracy	the proportion of true results (both true positives and true negatives) among the total number of cases examined	(TP+TN)/(TP+TN+FP+FN)
Precision	the proportion of true positives among the total number of positives determined by classifier	(TP)/(TP+FP)
F-score	harmonic mean of recall and precision. A high value of F-measure ensures that both precision and recall are reasonably high.	2x(precision x recall)/ (precision + recall)

Table 18- The definitions of the performance measures

1- The number of irrelevant patents is almost two times greater than the number of relevant patents. This means the data sets in the research are inherently imbalanced. In this situation, if someone randomly select patents and classifies them as ‘0’ or ‘1’, this random classifier (let’s call it ‘chance’) may perform well and yield effective performance measures. Therefore, a chance classifier is considered as the baseline in each iteration, and all performance measures are calculated for the baseline (chance), as well as for k-NN, SVN and random forest. The baseline allows for comparing the performance of the classifiers with respect to chance.

2- The following parameters are tuned in the cross validation process:

  - Number of neighbors (between 3 and 9) for *k*-NN classifier
  - Number of trees (between 4 and 19) for random forest classifier
  - C (include, 1E-9, 1e-7,1e-5, 1e-3, .01, .1, , 10, 20, 50, and 100) and gamma (include 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1) parameters as well as the type of kernel (include RBF and linear) for SVM classifier.

5.2.3 Product Features Classification

In this section, *k*-NN, random forest, and SVM classifiers are applied to classify the data set provided for the product features. The data set contain 92 patents for the product features, which are judged by the experts as 29 relevant patents and 63 irrelevant patents.

5.2.3.1. Regular Classification

As shown in Table 19, the k-NN classifier with n=9 has the best performance, according to ROC AUC measure which is 0.78. However, k-NN with n=8 has a performance that is very close; its precision and accuracy measures are even a little better. Nonetheless, k-NN with n=9 is considered the best because of its overall performance, according to ROC AUC and F-score.

Random forest couldn't show any impressive performance on regular keywords. ROC AUC varies between 0.51 and 0.53, and the F-score is too low in all classifications (see Table 20).

Number of Neighbors	**Ave. of Sensitivity**	**Ave. of Specificity**	**Ave. of ROC AUC**	**Ave. of Accuracy**	**Ave. of Precision**	**Ave. of F-score**
n=3	0.59	0.83	0.71	0.77	0.69	0.52
n=4	0.55	0.91	0.73	0.83	0.74	0.56
n=5	0.6	0.85	0.73	0.8	0.68	0.55
n=6	0.55	0.93	0.74	0.85	0.78	0.59
n=7	0.6	0.9	0.75	0.83	0.71	0.59
n=8	0.58	0.96	0.77	0.87	0.82	0.64
n=9	**0.63**	**0.93**	**0.78**	**0.86**	**0.78**	**0.65**

Table 19- k-NN performance: regular classification for the product features

As expected, SVM shows a good performance in regular classification. As shown in Table 21, among all classifications made base on combinations of C and gamma parameters and RBF and linear kernels, SVM with C= 10, Gamma=0.1, and kernel= 'rbf' has the best performance. ROC AUC is reasonably high (0.78), and F-score is not very low (0.64). More details about the performances of the SVM classifiers in regular classification of the product features are available in Table and Table in Appendix C.

Number of Trees	Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
4	0.07	0.97	0.52	0.76	0.51	0.09
5	0.13	0.93	0.53	0.75	0.52	0.16
6	0.05	0.98	0.51	0.77	0.45	0.07
7	0.10	0.95	0.53	0.76	0.49	0.12
8	0.04	0.98	0.51	0.77	0.44	0.06
9	0.08	0.97	0.52	0.77	0.47	0.10
10	0.03	0.99	0.51	0.77	0.43	0.04
11	0.06	0.98	0.52	0.77	0.46	0.08
12	0.03	0.99	0.51	0.77	0.43	0.04
13	0.05	0.98	0.52	0.77	0.45	0.06
14	0.03	0.99	0.51	0.77	0.45	0.04
15	**0.04**	**0.99**	**0.52**	**0.77**	**0.44**	**0.06**
16	0.03	0.99	0.51	0.78	0.45	0.04
17	0.04	0.99	0.51	0.78	0.44	0.05
18	0.02	1.00	0.51	0.77	0.43	0.02
19	0.03	0.99	0.51	0.77	0.45	0.04

Table 20- Random forest performance: regular classification for the product features

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
10	0.1	RBF	0.66	0.91	0.78	0.85	0.77	0.64

Table 21- SVM classifier performance: regular classification for the product features

By comparing the best performance of the classifiers in Table 22, it is observed that SVM is the best classifier for regular classification for the product features. However, k-NN exhibits very close performance in terms of ROC AUC and F-score, by comparing sensitivity (recall). Recall is more important than precision to Finex, because Finex wants to identify as many opportunities as possible. Thus, SVM is picked as the best classifier. The sensitivity of SVM is 0.66, while the sensitivity of k-NN is 0.58. Also, it is worth mentioning that regular classification for the product features is not very sensitive to ‘chance’. The performance of all classifiers is better than chance.

Classifier	Sensitivity	Specifity	ROC AUC	Accuracy	Precision	F-score
Chance	0.01	0.69	0.35	0.54	0.38	0.01
KNN(n=9)	**0.58**	**0.96**	**0.77**	**0.87**	**0.82**	**0.64**
RF (n=15)	0.04	0.99	0.52	0.77	0.44	0.06
SVM (('C=', 10, 'gamma=0.1', 'kernel=', 'rbf'))	**0.66**	**0.91**	**0.78**	**0.85**	**0.77**	**0.64**

Table 22- The best performance of the classifiers in regular classification for the product features

5.2.3.2. Ontological Classification

In this section, the classifiers are applied to classify the 92 patents judged by the experts for the product features. Unlike for regular classification, only ontological keywords are considered for this type of classification, which is called ontological classification. The ontological keywords are shown in Table 14.

As described in section 4.4, the TF-IDF matrix is calculated based on the term frequency of 30 ontological keywords in 19,525 patents. The training and test sets actually contain the TF-IDF of the ontological keywords extracted from the TF-IDF matrix.

All k-NN ontological classifications resulted in reasonable performances, as shown in Table 23. All ROC AUC’s are between 0.74 and 0.78. Also, F-score for n=8 is 0.65. All in all, k-NN with n=8 has the best performance; in particular, its specificity and accuracy show a very reasonable performance.

Number of Neighbors	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
n=3	0.68	0.88	0.78	0.83	0.78	0.63
n=4	0.61	0.93	0.77	0.86	0.83	0.65
n=5	0.70	0.80	0.75	0.77	0.70	0.57
n=6	0.58	0.90	0.74	0.83	0.78	0.58
n=7	0.62	0.88	0.75	0.82	0.74	0.59
n=8	**0.62**	**0.94**	**0.78**	**0.87**	**0.82**	**0.65**
n=9	0.62	0.93	0.78	0.86	0.78	0.64

Table 23- k-NN performance: ontological classification for the product features

The performance of random forest in the ontological classification is much better than the performance of random forest in the regular classification. All measures have improved tremendously; however, sensitivity is around 0.5, which is not high enough. As shown in Table 24, the random forest classifier with n=15 has the best performance. Its ROC AUC is 0.72, and its other measures have slightly better performance than those of classifications with similar ROC AUC.

SVM has better performance in the ontological classification. As shown in Table 25 and, SVM with C= 100, gamma=0.9, and RBF kernel has the best performance among all of the other SVM ontological classifications. ROC AUC with 0.87 shows that this classifier has performed nearly perfectly. Sensitivity (recall) is 0.81, which is a reasonable performance for this research. More details about the performance of the SVM classifiers are available in Table and Table in Appendix C.

Number of Trees	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
4	0.42	0.92	0.67	0.81	0.68	0.53
5	0.51	0.91	0.71	0.82	0.74	0.55
6	0.46	0.92	0.69	0.81	0.75	0.53
7	0.51	0.91	0.71	0.82	0.72	0.54
8	0.48	0.92	0.70	0.82	0.75	0.53
9	0.52	0.91	0.71	0.82	0.77	0.55
10	0.50	0.92	0.71	0.83	0.77	0.52
11	0.52	0.92	0.72	0.83	0.75	0.53
12	0.49	0.93	0.71	0.83	0.75	0.52
13	0.51	0.92	0.72	0.83	0.74	0.53
14	0.50	0.92	0.71	0.82	0.74	0.45
15	**0.52**	**0.93**	**0.72**	**0.83**	**0.77**	**0.52**
16	0.48	0.93	0.70	0.83	0.75	0.49
17	0.50	0.92	0.71	0.82	0.73	0.53
18	0.49	0.93	0.71	0.83	0.73	0.50
19	0.51	0.92	0.72	0.83	0.69	0.54

Table 24- Random forest performance: ontological classification for the product features

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
10	0.1	RBF	0.66	0.91	0.78	0.85	0.77	0.64

Table 25- Best SVM classifier in ontological classification for the product features

The performances of the classifiers in the ontological classification for the product features are compared in Table 26. The SVM shows remarkably higher performance than the other classifiers. Although the sensitivity of the k-NN is better, the specificity of the SVM has performed better. In this condition, the ROC AUC and F-score of the SVM show that the general performance of the SVM is much better than of the others. The specificity of the SVM, which is 0.92, shows this classifier has been able to classify perfectly irrelevant patents, and that enhances the overall performance of the SVM because the data is inherently imbalanced and the majority of the data contain irrelevant patents.

Classifier	Sensitivity	Specificity	ROC_AUC	Accuracy	Precision	F-score
Chance	0.69	0.01	0.35	0.53	0.38	0.01
KNN(n=8)	0.94	0.62	0.78	0.87	0.82	0.65
RF (n=15)	0.93	0.52	0.72	0.83	0.77	0.55
SVM ('C=', 100, 'gamma=0.9', 'kernel=', 'RBF')	0.81	0.92	0.87	0.84	0.78	0.73

Table 26- The best performance of the classifiers: ontological classification for the product features

5.2.3.3. Ontological Semantic Classification

In this section, the classifiers are applied to classify the 92 patents judged by the experts for the product features. In ontological semantic classification, 30 ontological keywords, as well as 42 of their synonyms, are considered. This amounts to 72 ontological semantic keywords in total, which are displayed in Table 14.

As described in section 4.4, the TF-IDF matrix is calculated based on the term frequency of 72 ontological keywords in 19,525 patents. The training and test sets actually contain the TF-IDF of the ontological semantic keywords extracted from the TF-IDF matrix.

The k-NN classifiers have reasonable performance in ontological semantic classifications, as shown in Table 27. The ROC AUC values are greater than 0.75, and their sensitivity is around 0.90. The k-NN with n=3 has the best performance. Its ROC

AUC (0.82) and its F-score (0.70) show that this classifier performs significantly better than the other k-NN classifiers shown in Table 27. Also, its sensitivity is 0.90, and its specificity is 0.74. Thus, k-NN with n=3 performs with high accuracy.

Number of Neighbors	Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
n=3	0.90	0.74	0.82	0.86	0.82	0.70
n=4	0.92	0.61	0.77	0.85	0.75	0.62
n=5	0.88	0.64	0.76	0.82	0.74	0.61
n=6	0.93	0.59	0.76	0.85	0.81	0.62
n=7	0.92	0.62	0.77	0.85	0.76	0.63
n=8	0.94	0.62	0.78	0.86	0.81	0.65
n=9	0.93	0.62	0.77	0.86	0.82	0.64

Table 27- k-NN performance: ontological semantic classification for the product features

The random forest classifiers exhibit similar performances in the ontological semantic classification (OSC) and in the ontological classification (OC); the ROC AUC of OSC has only improved slightly in comparison to that of the OC. Nonetheless, as shown in Table 28, random forest classifier with n=17 has the best performance in the OSC. Its ROC AUC is 0.76 and its other measures have slightly better performance than of random forest with n=19.

SVM with C= 100, gamma=0.1, and RBF kernel has the best performance among all the SVM ontological semantic classifications. The ROC AUC of the SVM is 0.87; however, it is similar to of the best SVM of the ontological classification; the other measures and specifically the ROC AUC show how perfectly this classifier has performed in the ontological semantic classification (sensitivity=0.91, specificity=0.82, accuracy=0.89, precision=0.87, and F-score=0.77). More details about the other SVM classifiers with different parameters are available in Table and Table in Appendix C.

Number of Trees	Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
4	0.41	0.96	0.69	0.84	0.75	0.47
5	0.53	0.93	0.73	0.84	0.72	0.55
6	0.45	0.96	0.71	0.84	0.74	0.50
7	0.54	0.94	0.74	0.85	0.75	0.57
8	0.45	0.96	0.71	0.84	0.74	0.51
9	0.54	0.95	0.74	0.85	0.79	0.58
10	0.48	0.95	0.72	0.85	0.75	0.53
11	0.56	0.94	0.75	0.86	0.77	0.59
12	0.51	0.96	0.73	0.85	0.77	0.56
13	0.55	0.95	0.75	0.86	0.78	0.59
14	0.51	0.96	0.73	0.86	0.80	0.56
15	0.56	0.95	0.75	0.86	0.77	0.60
16	0.54	0.96	0.75	0.86	0.79	0.60
17	**0.58**	**0.95**	**0.76**	**0.86**	**0.79**	**0.61**
18	0.52	0.96	0.74	0.86	0.80	0.58
19	0.56	0.95	0.76	0.86	0.78	0.61

Table 28- Random forest performance: ontological semantic classification for the product features

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
100	0.1	RBF	0.91	0.82	0.87	0.89	0.87	0.77

Table 29- The best SVM classifier in ontological semantic classification for the product features

The performances of the classifiers in the ontological semantic classification of the product features are compared in Table 30. The ROC AUC (0.87) and F-score (0.77) of the SVM show how the SVM performs remarkably better than the other classifiers in the ontological semantic classification. Although the sensitivity of the random forest is better, the specificity of the SVM has performed much better. The specificity of the SVM, which is 0.82, shows that this classifier has been able to classify perfectly irrelevant patents, and it enhances the overall performance of the SVM, because the data are inherently imbalanced and the majority of the data contain irrelevant patents.

Classifier	Sensitivity	Specificity	ROC AUC	Accuracy	Precision	F-score
Chance	0.69	0.01	0.35	0.54	0.38	0.01
KNN(n=3)	0.90	0.74	0.82	0.86	0.82	0.70
RF (n=17)	0.95	0.58	0.76	0.86	0.79	0.61
SVM('C=', 100, 'gamma =0.1', 'kernel=', 'rbf')	**0.91**	**0.82**	**0.87**	**0.89**	**0.87**	**0.77**

Table 30- The best performance of the classifiers: ontological semantic classification for the product features

5.2.3.4. Best classifier for the Product Features

Table 31 shows the best classifiers in regular classification (RC), ontological classification (OC), and ontological semantic classification (OSC) for the product features. In all three scenarios (RC, OC, and OSC) SVM is the best classifier. Also, the SVM of the OSC is the best scenario to classify the product features. Therefore, the SVM-OSC scenario will be applied to classify the product features to identify the opportunities.

Scenario	Best Classifier	Sensitivity	Specifity	ROC AUC	Accuracy	Precision	F-score
RC	SVM (('C=', 10, 'gamma=0.1', 'kernel=', 'rbf'))	0.66	0.91	0.78	0.85	0.77	0.64
OC	SVM ('C=', 100, 'gamma=0.9', 'kernel=', 'rbf')	0.81	0.92	0.87	0.84	0.78	0.73
OSC	SVM('C=', 100, 'gamma=0.1', 'kernel=', 'rbf')	0.91	0.82	0.87	0.89	0.87	0.77

Table 31- The best classifiers in the RC, the OC, and the OSC of the product features

5.2.4 Enabling technologies Classification

In this section, *k*-NN, random forest, and SVM classifiers are applied to classify the data set provided for the enabling technologies (thin wall iron casting). The data set contains 17 patents that the Finex experts judged as relevant to the enabling technologies and 36 patents that the Finex experts judged as irrelevant to the enabling technologies.

5.2.4.1. Regular Classification

As shown in Table 32, none of the k-NN classifiers perform well. All measures are low, and ROC AUC and F-score show the overall performance of the k-NN classifiers are weak. Nonetheless, k-NN with n=3 is the best one among these weak classifiers.

The random forest classifiers couldn’t perform well in regular classification. As shown in Table 33, however, the accuracy measures are slightly high (more than 0.63), the ROC AUC and F-score measures show the overall performance of the random forest classifiers is low. Nonetheless, random forest with n=5 has the best performance in the regular classification.

The random forest classifiers couldn’t perform well in regular classification. As shown in Table 33, however, the accuracy measures are slightly high (more than 0.63), the ROC AUC and F-score measures show the overall performance of the random forest classifiers is low. Nonetheless, random forest with n=5 has the best performance in the regular classification.

Number of Neighbors	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
n=3	**0.27**	**0.75**	**0.51**	**0.59**	**0.54**	**0.28**
n=4	0.07	0.90	0.49	0.63	0.39	0.09
n=5	0.19	0.81	0.50	0.61	0.46	0.21
n=6	0.10	0.92	0.51	0.65	0.44	0.13
n=7	0.18	0.84	0.51	0.63	0.48	0.22
n=8	0.09	0.93	0.51	0.66	0.45	0.12
n=9	0.13	0.88	0.51	0.63	0.47	0.17

Table 32- *k*-NN performance: regular classification for the enabling technologies

The SVM classifier performs weakly in regular classification (as did the k-NN classifiers and the random forest classifiers). As shown in Table 34, the SVM classifier with C= 20, Gamma=0.3, and linear kernel perform better than the other SVM classifiers. Their accuracy measures are 0.65, but ROC AUC (0.52) and F-score (0.52) are still low. More details about the other SVM classifiers with different parameters are available in Table and Table in Appendix D.

Number of Trees	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
4	0.10	0.90	0.50	0.64	0.44	0.12
5	**0.19**	**0.84**	**0.52**	**0.63**	**0.46**	**0.21**
6	0.10	0.93	0.51	0.66	0.43	0.12
7	0.17	0.88	0.52	0.65	0.50	0.19
8	0.09	0.93	0.51	0.66	0.44	0.12
9	0.13	0.90	0.51	0.65	0.49	0.16
10	0.07	0.95	0.51	0.66	0.43	0.10
11	0.11	0.91	0.51	0.65	0.45	0.14
12	0.06	0.96	0.51	0.66	0.45	0.09
13	0.09	0.93	0.51	0.66	0.47	0.12
14	0.06	0.96	0.51	0.67	0.42	0.09
15	0.08	0.94	0.51	0.66	0.42	0.11
16	0.05	0.97	0.51	0.67	0.40	0.07
17	0.06	0.95	0.51	0.66	0.44	0.08
18	0.03	0.97	0.50	0.67	0.40	0.05
19	0.06	0.95	0.51	0.66	0.44	0.08

Table 33- Random forest performance: regular classification for the enabling technologies

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
20	0.3	linear	0.15	0.88	0.52	0.64	0.52	0.19

Table 34- SVM classifier performance: regular classification for the enabling technologies (kernel: linear)

5.2.4.2. Ontological Classification

In this section, the classifiers are applied to classify the 53 patents that the Finex experts judged for the enabling technologies. Unlike for regular classification, only ontological keywords are considered for the ontological classification. The ontological keywords are shown in Table 15.

As described in section 4.4, the TF-IDF matrix is calculated based on the term frequency of 70 ontological keywords in 19,525 patents. The training and test sets actually contain the TF-IDF of the ontological keywords extracted from the TF-IDF matrix.

When compared to regular classification, the k-NN classifiers performed slightly better in ontological classification. In particular, the k-NN classifier with n=5 remarkably performed well. Its ROC AUC (0.57) and accuracy (0.62) have improved in comparison to regular classification. Nonetheless, its sensitivity (0.43) is too low, and it impacted the F-score (0.41) negatively.

Number of Neighbors	Ave. of Sensitivity	Ave. of Specifity	Aver. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
n=3	0.29	0.74	0.52	0.59	0.49	0.28
n=4	0.13	0.84	0.48	0.60	0.40	0.13
n=5	**0.43**	**0.71**	**0.57**	**0.62**	**0.61**	**0.41**
n=6	0.15	0.84	0.50	0.62	0.47	0.18
n=7	0.32	0.76	0.54	0.61	0.54	0.32
n=8	0.17	0.83	0.50	0.61	0.48	0.19
n=9	0.30	0.74	0.52	0.60	0.51	0.30

Table 35- k-NN performance: ontological classification for the enabling technologies

The random forest classifiers did not perform well in the ontological classification (just like they did not perform well in the regular classification). The ROC AUC's are less than 0.50 meaning the overall performance of the random classifiers are even worse than 'chance'.

Number of Trees	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
4	0.21	0.71	0.46	0.55	0.42	0.21
5	0.25	0.67	0.46	0.53	0.44	0.23
6	0.23	0.70	0.46	0.55	0.44	0.22
7	0.23	0.69	0.46	0.54	0.42	0.22
8	0.22	0.70	0.46	0.54	0.44	0.21
9	0.24	0.69	0.47	0.55	0.45	0.23
10	0.22	0.71	0.46	0.55	0.44	0.21
11	0.24	0.68	0.46	0.54	0.44	0.23
12	0.22	0.70	0.46	0.55	0.43	0.22
13	0.24	0.70	0.47	0.55	0.45	0.23
14	0.22	0.70	0.46	0.55	0.43	0.22
15	0.23	0.69	0.46	0.54	0.42	0.22
16	0.22	0.71	0.46	0.55	0.46	0.22
17	0.23	0.70	0.47	0.55	0.44	0.23
18	0.22	0.71	0.46	0.55	0.46	0.22
19	0.22	0.71	0.47	0.55	0.45	0.22

Table 36- Random forest performance: ontological classification for the enabling technologies

The overall performance of the SVM classifiers in the ontological classification (OC) has improved slightly in comparison to their performance in the regular classification (RC). In comparison to RC, the sensitivity of the SVM's has improved remarkably, but their specificity has degraded. In general, the SVM F-score measures of OC are 6% to 8% higher than those of RC. All in all, the SVM with C= 0.001, Gamma=1, and RBF kernel has the best performance in the ontological classification by the SVM classifiers. More details about the other SVM classifiers with different parameters are available in Table and Table in Appendix D.

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
0.001	1	RBF	0.71	0.33	0.52	0.47	0.53	0.44

Table 37- Best SVM classifier performance: ontological classification for the enabling technologies

Comparing the best performance of the classifiers in Table 38 suggests that the k-NN is the best classifier for the ontological classification for the enabling technologies; however, its performance is not much better than chance. The other classifiers even performed worse than chance.

Classifier	Ave. of Sensitivity	Ave. of Specifity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
Chance	0.42	0.72	0.57	0.62	0.58	0.42
KNN (n=5)	0.43	0.71	0.57	0.62	0.61	0.41
RF (n=9)	0.24	0.69	0.47	0.55	0.45	0.23
SVM (C=100, gamma= 0.1, kernel=RBF)	0.49	0.55	0.52	0.53	0.51	0.39

Table 38- The best performance of the classifiers: ontological classification for the enabling technologies

5.2.4.3. Ontological Semantic Classification

In this section, the classifiers are applied to classify the 53 patents that the experts judged for the enabling technologies. In ontological semantic classification (OSC), 70 ontological keywords, as well as 71 of their synonyms, are considered. These ontological semantic keywords (141 in total) are shown in Table 15.

As described in section 4.4, the TF-IDF matrix of OSC is calculated based on the term frequency of 141 ontological keywords in 19,525 patents. The training and test sets actually are extracted from the TF-IDF matrix of OSC.

Number of Neighbors	Ave. of Sensitivity	Ave. of Specifity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
n=3	**0.45**	**0.82**	**0.63**	**0.70**	**0.64**	**0.48**
n=4	0.19	0.92	0.56	0.69	0.59	0.25
n=5	0.31	0.87	0.59	0.69	0.59	0.38
n=6	0.15	0.95	0.55	0.69	0.55	0.21
n=7	0.23	0.93	0.58	0.70	0.64	0.30
n=8	0.11	0.96	0.54	0.68	0.53	0.16
n=9	0.16	0.93	0.54	0.67	0.55	0.21

Table 39- k-NN performance: ontological semantic classification for the enabling technologies

As shown in Table 39, the performance of the k-NN classifiers has notably improved in OSC in comparison to their performances in RC and OC. The ROC AUC of three k-NN classifiers are close to 0.6, and that of the k-NN with n=3 is 0.63. Also, the accuracy measures of the k-NN classifiers have improved by around 10%. The accuracy of the best k-NN classifier with n=3 is 0.70. Despite the improvement of the sensitivity measures in some k-NN classifiers, these measures are still low; they affect the F-score negatively. For example, the sensitivity of the best k-NN (n=3) is 0.45 and its F-score is 0.48, while its specificity is 0.82. This means this classifier can successfully recognize irrelevant patents, but its performance in recognizing relevant patents is still low.

As shown in Table 40, the ROC AUC scores of the random forest classifiers in the ontological semantic classification (OSC) have improved slightly in comparison to their

ROC AUC scores in the ontological classification (OC). The ROC AUC's are around 0.5 (chance), and F-score measures are too low. The main reason for this is that the sensitivity measures are low. Nonetheless, the random forest with n=5 has the best performance in the OSC.

Number of Trees	Average of Sensitivity	Average of Specifity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
4	0.14	0.87	0.50	0.63	0.44	0.17
5	**0.26**	**0.80**	**0.53**	**0.62**	**0.52**	**0.27**
6	0.15	0.87	0.51	0.64	0.45	0.18
7	0.24	0.81	0.53	0.63	0.50	0.26
8	0.16	0.87	0.52	0.64	0.48	0.19
9	0.23	0.83	0.53	0.63	0.53	0.25
10	0.16	0.89	0.52	0.65	0.46	0.19
11	0.22	0.83	0.53	0.64	0.52	0.25
12	0.17	0.87	0.52	0.64	0.49	0.20
13	0.19	0.84	0.52	0.63	0.49	0.22
14	0.16	0.88	0.52	0.64	0.51	0.19
15	0.19	0.85	0.52	0.64	0.51	0.23
16	0.15	0.89	0.52	0.65	0.50	0.19
17	0.19	0.86	0.52	0.64	0.47	0.22
18	0.15	0.88	0.52	0.64	0.46	0.18
19	0.18	0.84	0.51	0.63	0.48	0.20

Table 40- Random forest performance: ontological semantic classification for the enabling technologies

As shown in Table 41, the overall performance of the SVM classifiers in the ontological semantic classification (OSC) has improved slightly in comparison to their performance in the ontological classification (OC) and the regular classification (RC). On average, the ROC AUC measures of the SVM's are 3% to 4%, and their accuracy measures have improved by 9%. More details about the performance of the SVM classifiers are available in Table and Table in Appendix D.

SVM Parameters			Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma	Kernel						
100	0.8	RBF	0.42	0.74	0.58	0.63	0.61	0.4

Table 41- Best SVM classifier performance: ontological semantic classification for the enabling technologies

5.2.4.4. Best Classifier for the Enabling technologies Classification

Table 42 shows the best classifiers in the regular classification (RC), the ontological classification (OC), and the ontological semantic classification (OSC) for the enabling technologies. None of the classifiers could perform well in the RC; they performed worse than chance. The K-NN with n=5 is the best classifier in the OC and k-NN with n=3 is the best classifier in the OSC.

Scenario	Classifier	Ave. of Sensitivity	Ave. of Specifity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
RC	**Chance**	0.40	0.71	0.55	0.61	0.58	0.40
OC	**KNN (n=5)**	0.43	0.71	0.57	0.62	0.61	0.41
OSC	**KNN (n=3)**	**0.45**	**0.82**	**0.63**	**0.70**	**0.64**	**0.48**

Table 42- The best classifiers in the RC, the OC, and the OSC of the enabling technologies

5.2.5 Opportunity Identification

The patents relevant to both the product features (light weight product) and enabling technologies (thin wall iron casting) are potential opportunities that can be considered for new product planning. The best scenarios (introduced in Table 31 and

Table 42, respectively) are applied to identify the opportunities for new product planning in the Finex case. They are presented in Table 43.

Ontology	Classification	Classifier	Parameters
Product features	OSC	SVM	C= 100, Gamma=0.1, kernel= RBF
Enabling technologies	OSC	k-NN	N=3

Table 43- The scenarios used for the opportunity identification

In order to identify the opportunities, the following steps have been taken:

- The SVM classifier (C= 100, Gamma=0.1, kernel= RBF) is trained by using the TF-IDF of the OSC of the product features, and then the trained SVM is used to classify the 19,525 patents. Consequently, 750 patents are classified as relevant to the product features.
- Similarly, the k-NN classifier (n=3) is trained by using the TF-IDF of the OSC of the enabling technologies, and then the trained k-NN is used to classify the 19,525 patents. Consequently, 5415 patents are classified are classified as relevant to the enabling technologies.
- The joint list of two sets of the classified patents contains 188 patents, which are known as the opportunities for new product planning in the Finex case. The patent numbers of the opportunities are shown in Table in Appendix E.

5.3. Opportunity Authentication

OSC classifies patents either as relevant or as irrelevant, according to the subject of an ontology. Still, experts should step in and review the classified patents as opportunities to identify the ones that are truly valuable.

In order to authenticate some of the identified opportunities, 20 patents are randomly selected out of 188 opportunities identified and evaluated more deeply by expert E3 who has the best performance among the experts. The patents are evaluated from three angles:

- *Relatedness*: how much the technology is related to the case.
- *Innovativeness*: how much the technology can be considered innovative.
- *Usability*: how much the technology is applicable for the Finex case.

The result of the evaluation is shown in Table 44. The used scale for evaluation is as below:

1=very weak, 3=weak, 5= moderate, 7=strong, and 9=very strong.

As shown in Table 44, the relatedness of 15 authenticated patents is scored greater than seven, meaning they are highly relevant to the subject of the case. Also, the innovativeness of seven authenticated patents is scored higher than seven, meaning they contain technologies which can be considered highly innovative. At the end, the usability of 15 of the authenticated patents is scored more than 5, meaning they are usable for solving the problem stated in the Finex case.

#	Patent Number	Relatedness	innovativeness	usability
1	US5127467	9	3	7
2	US5165464	9	3	7
3	US5501833	9	3	7
4	US5664619	9	3	9
5	US5800902	5	7	3
6	US6129134	9	7	7
7	US6309743	9	9	9
8	US6328820	3	3	1
9	US6427755	3	9	3
10	US6537395	9	3	7
11	US6582533	7	9	5
12	US6719104	1	3	1
13	US6908590	7	3	7
14	US6913062	9	3	9
15	US7045022	9	3	7
16	US7056598	9	7	7
17	US7086151	9	7	7
18	US7793703	9	3	5
19	US8840738	9	3	5
20	US8905203	1	3	1

Table 44- Evaluation of 20 opportunities by the expert panel

CHAPTER 6: DISCUSSION

As reported in Chapter 5, applying ontologies can improve the performance of classifications. The SVM classifier has a *perfect* performance in the OSC of the product features, while the k-NN classifier has a *moderate* performance in the OSC of the enabling technologies. In section 6.1 of this chapter, I discuss how human experts' search behavior affects their performance in the patent search. In section 6.2, the performance of the classifiers in RC, OC, and OSC are examined for a different data set. This examination reconfirms applying ontologies is a reliable method to improve the performance of classifiers. In section 6.3., I discuss why different ontologies come up with different classification performances.

6.1. Human Expert's Patent Search Analysis

Alongside of developing the method of this research, there has been an important question: Do experts basically need an intelligent patent search method? The patent search introduced in section 3.4.3 is designed and conducted to answer this question. It is studied in the patent search how well experts perform in patent searches without any intelligent tool. The results of the patent search experiment are illustrated in section 4.2.

Basically, the results of the patent search experiment showed that the experts do not have a good performance in the patent searches. The performance of the experts can be summarized as below:

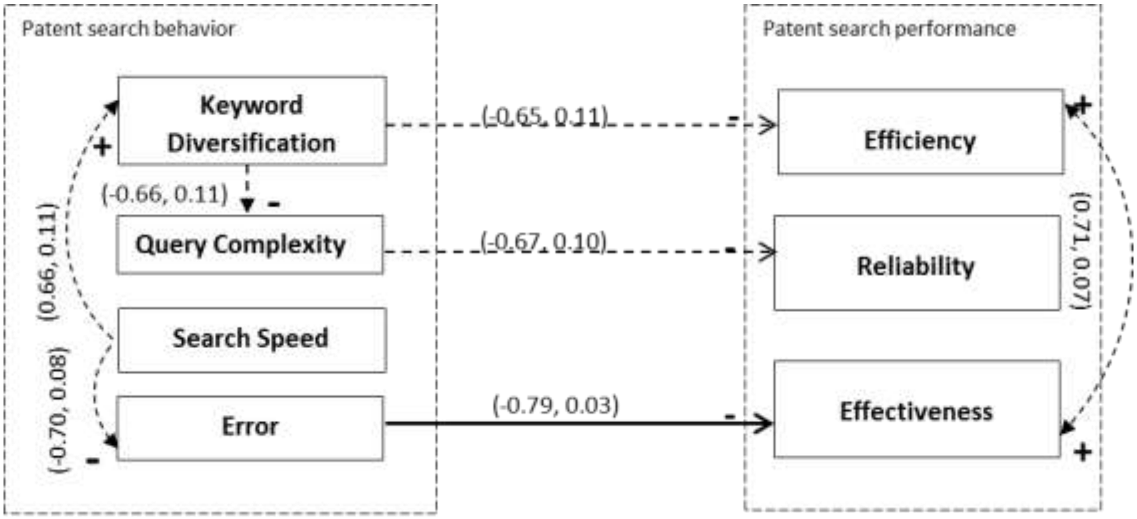
- The *reliability* of the patent searches is between 2% and 14.45%. These numbers show that the reliability is *very low*. The average of similarities between keywords (shown in Table 11) and the average of similarities between relevant patents found in the patent searches (shown in Table 12) illustrate that the experts used very low similar keywords and consequently found very low similar relevant patents, despite having the ontological keywords at their disposal.
- The *efficiency* of the patent searches varies between 0.05 and 0.30, and it is 0.18 on average. An expert who wants to find 100 *relevant* patents would have to spend 9.26 hours, if he/she were to perform at a rate that reflects the average of these efficiencies. Considering cases in which there are thousands of relevant patents, it would take a long time to find them. Thus, the efficiency of the experts is *too low*.
- The *effectiveness* of the patent searches relatively *low*. Half of the experts have an effectiveness of more than 50%, and half of the experts have an effectiveness of less than 50%.

To figure out how the patent search behavior affects the patent search performance, a correlation analysis is applied to the measures introduced in section 4.2. A table that represents the results of the correlation analysis is available in Appendix F. The result of the correlation analysis is shown in Figure 50. The strong correlations between patent search behavior and patent search performance can be taken as *causation* relation, because their search behavior (as expressed by keyword diversity, query complexity, search speed and error rate) influences their ability to retrieve patents and judge them as relevant or irrelevant. The explanations of the correlations between patent search behavior and patent search performance follow:

- ***Keyword Diversification***: There is a moderate negative correlation between keyword diversification and efficiency (-0.65). The correlation is moderately significant because its p-value is 0.11. When an expert applies more diversified keywords, it means he/she performs a search on a greater variety of subjects, so he/she has to spend more time to judge patents that he/she has found. Therefore, more diversified keywords negatively impact the efficiency of his/her patent search.
- ***Query Complexity:*** There is moderate correlation (-0.67) between query complexity and reliability. The correlation is moderately significant with 0.10 for p-value. To explain this correlation, let's imagine two imaginary experts; expert A and expert B who both perform a patent search for a similar concept. When expert A applies more complex queries than expert B does, the similarity between keywords used by expert A and expert B is decreased. Therefore, these two experts come up with less similar patents in their search results. Therefore, more complex queries lead to less reliable patent searches.
- ***Search Speed***: There is no strong correlation between search speed and the search performance measures. Nonetheless, there are moderate correlations between search speed and two other search behavior measures, which are keyword diversification and error.

    - There is moderate positive correlation (0.66) between search speed and keyword diversification with moderate statistical significance (p-value is 0.11). When an expert does a patent search faster, he/she applies more keywords and most likely more distinct keywords. Therefore, there is a moderate positive correlation between search speed and keyword diversification.
    - There is a moderate negative correlation between search speed and error (-0.70) with moderate statistical significance (p-value is 0.08). When an expert does a patent search faster, he/she retrieves and judges more patents. The error rate becomes smaller because the number of duplicates and contradictions (numerator of error rate) grows slower than the number of patents judged (denominator of error rate). This explains the negative correlation between search speed and error.
- ***Error:*** Some of the experts came up with some duplicates in their patent retrieval and contradictions in their judgments. These errors, especially the contradictions, negatively affect effectiveness because the experts have misjudged some patents. Thus, they consider some irrelevant patents relevant and some relevant patents irrelevant. As shown in Figure 50, there is a strong negative correlation (-0.79) between error and effectiveness. The correlation is strongly significant with 0.03 for p-value.

In addition to the abovementioned explanations of correlations between search behavior measures and patent search performance measures, there is a moderate positive correlation between efficiency and effectiveness. The number of relevant patents constitutes the numerator of both factors. Therefore, there is a positive correlation between them.



* Numbers in the parentheses are correlation, and p-value, respectively.

Figure 50- Relations between patent search behavior and patent search performance

According to the facts presented in section 4.2 and the discussions of this section, there are some conclusions about the patent searches observed in this study:

1- The patent searches are ***highly unreliable***.

2- The patents searches are ***highly inefficient*** in finding the relevant patents.

3- The patent searches are ***relatively ineffective***.

4- The method of this research significantly improves the performance of patent searches because:

a. The method doesn’t rely directly on queries and keywords selected by a human expert. The method uses a classification models to process all retrieved patents in a corpus.

b. The method doesn’t have human error. The method doesn’t retrieve duplicate patents and doesn’t judge contradictorily one patent to relevant and irrelevant.

c. The method performs patent classification extremely faster than an expert.

6.2. Ontological Semantic Classification

As a remedy for the low performance of human experts in patent search, ontological semantic classification (OSC) can mitigate the effects of this problem. Therefore, three hypotheses are introduced in section 1.3 and three scenarios are designed in section 3.3 to examine what kind of keywords (regular, ontological, and ontological semantic) and which of the classifiers (k-NN, SVM, and random forest) have the best performance in patent classification. In these hypotheses, it is assumed that OSC would outperform RC and OC. In the Finex case, ontologies for the product features and the enabling technologies are developed and applied to two data sets: the Finex experts’ data set and expert panel’s data set. Therefore, the hypotheses are literally examined two times, and consequently, the performance of OSC is examined four times, as shown in Table 45.

Data set	**Finex Experts**						**Expert Panel**					
Ontology	**PF**			**TA**			**PF**			**TA**		
Classification	RC	OC	OSC	RC	OC	OSC	RC	OC	OSC	RC	OC	OSC
Hypotheses examination	#1						#2					

RC: Regular Classification, OC: Ontological Classification, OSC: Ontological Semantic Classification
PF: Product Features, ET: Enabling technologies

Table 45- the combinations of data sets, ontologies, classifications, and hypotheses examinations

The hypotheses are confirmed in examinations by two sets of experts. Therefore, OSC has the best performance in patent search for the product features ontology and for the enabling technologies ontology. Also, SVM classifier is the best classifier in the three of the OSCs, and k-NN is only in one OSC. Therefore, SVM can be considered as the number one priority in choosing classifiers for OSC; however, k-NN should be considered as the back-up classifier in case of the weak performance of SVM.

6.2.1. Reliability of OC and OSC

As mentioned in Chapter 3, a group of external experts participated in the patent search; their performance and behavior are reflected in section 4.2. Since the panel experts applied the ontology of the enabling technologies of Finex, the patents classified by the panel experts are deployed to reexamine the hypotheses mentioned in chapter 1. The final results are shown in Table 46.

Scenario	**Classifier**	**Ave. of Sensitivity**	**Ave. of Specificity**	**Ave. of ROC AUC**	**Ave. of Accuracy**	**Ave. of Precision**	**Ave. of F-score**
RC	**Chance**	**0.38**	**0.68**	**0.53**	**0.59**	**0.56**	**0.36**
	k-NN (n=9)	0.21	0.81	0.51	0.63	0.50	0.23
	random forest (n =11)	0.07	0.93	0.50	0.66	0.44	0.10
	SVM (C= 10, Gamma= 0.1, kernel= RBF)	0.35	0.68	0.52	0.58	0.50	0.32
OC	**Chance**	0.37	0.68	0.52	0.58	0.54	0.35
	k-NN (n=5)	0.11	0.90	0.50	0.65	0.48	0.14
	Random forest (n =5)	0.27	0.79	0.53	0.63	0.52	0.28
	SVM (C= 10, Gamma= 0.1, kernel= RBF)	**0.57**	**0.59**	**0.58**	**0.58**	**0.58**	**0.45**
OSC	**Chance**	0.37	0.68	0.52	0.58	0.54	0.35
	k-NN (n=5)	0.18	0.86	0.52	0.65	0.51	0.22
	random forest (n=9)	0.21	0.85	0.53	0.65	0.52	0.24
	SVM(C= 20, Gamma= 0.1, kernel= RBF)	**0.58**	**0.64**	**0.61**	**0.62**	**0.62**	**0.48**

Table 46- The performance of the classifiers based on the patents classified in the patent search experiment

As shown in Table 46, the SVM classifier with ROC AUC 0.52 has the best performance in the regular classification (RC); however, all of the classifiers performed worse than ‘chance’ in RC. The SVM classifier with C= 10, Gamma= 0.1, RFB kernel has the best performance in the ontological classification (OC). The ROC AUC of the SVM is 0.58, and its sensitivity and F-score is much higher than the other classifiers in the OC. Similarly, the SVM classifier with C= 20, Gamma= 0.1, and RFB kernel has the best performance in the ontological semantic classification (OSC). The ROC AUC and the sensitivity of the SVM are 0.61 and 0.58, respectively. Also, the other measures of the SVM are generally better than those of the other classifiers in the OSC.

By comparing the performance measures of the best classifiers in the RC, OC, and OSC scenarios, the hypotheses are reconfirmed. The OSC performs better than the RC and the OC. Also, the OC performs better than the RC.

6.3. Human Expert's roles in OSC

Even though OSC can be considered an intelligent method, it still relies on human experts in three activities: 1) ontology design; 2) data set (training set and test set) preparation; and 3) opportunity authentication. The performance of human experts in activities 1 and 2 impacts the performance of OSC directly, as experts are the only ones who can authenticate the opportunities identified by OSC.

6.3.1. Ontology Design and Evaluation

The ontological keywords of the product features and the enabling technologies are the main inputs of OC and OSC, as shown in Figure 37. If an ontological keyword has a high frequency number in a corpus, this keyword is not discriminative enough for classification purposes.

The cut-off value is a good criterion to see if a keyword is discriminative enough. As shown in Figure 51, 50% of the ontological keywords of the product features have a total term frequency more than the cut-off value (of 20,000). On the other hand, as shown in Figure 52, 66% of the ontological semantic keywords have a total term frequency more than the cut-off value. This means the ontological semantic keywords are more discriminative than the ontological keywords. This fact explains why ontological semantic classification (OSC) performs better than ontological classification (OC).

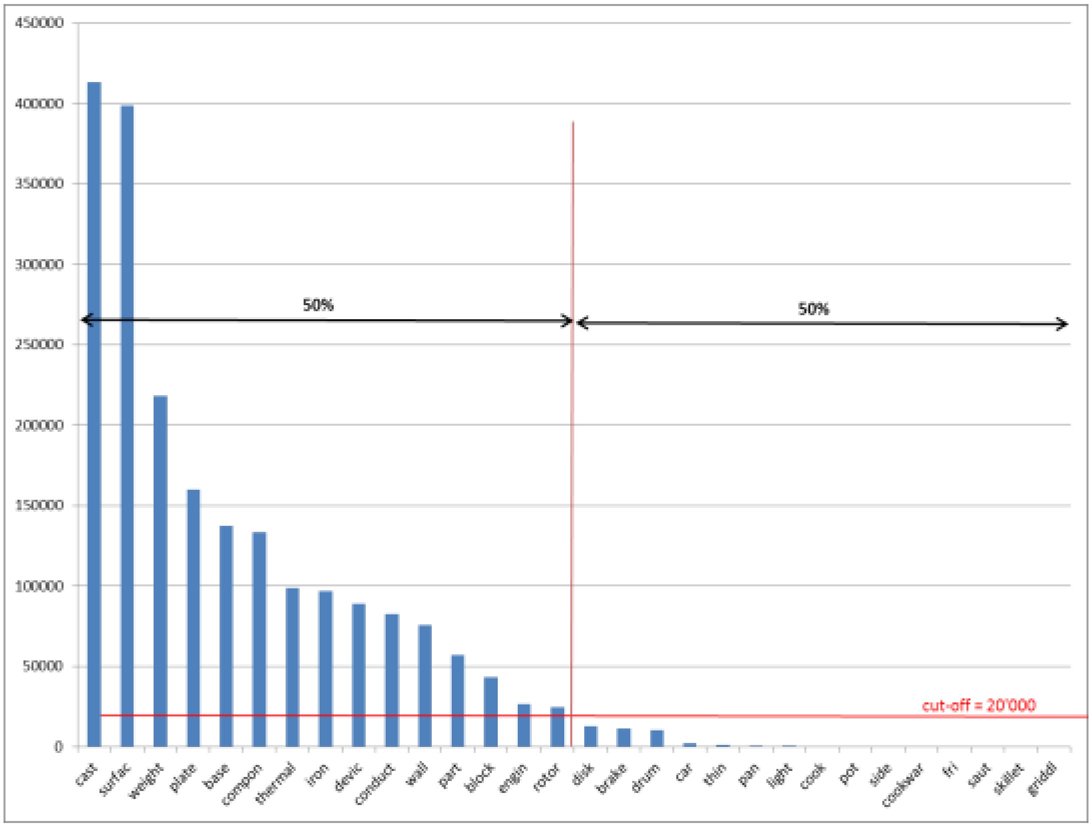


Figure 51- The ontological keywords of the product features (ranked based on their term frequencies in the corpus)

If an ontology does not contain enough discriminative ontological keywords, the ontological keywords cannot perform well in classification. The first ontology of the enabling technologies, Figure 53, has this deficiency. This problem prevents reasonable performance of OC and OSC. Only 13% of the ontological keywords are under the cut-off value, as shown in Figure 54. By developing the second edition of the ontology of the enabling technologies (shown in Figure 55), the percentage of the ontological keywords whose term frequency is greater than the cut-off value increased to 37%. This improvement could not lead to an acceptable performance in the OC; see Table 42 in chapter 4.

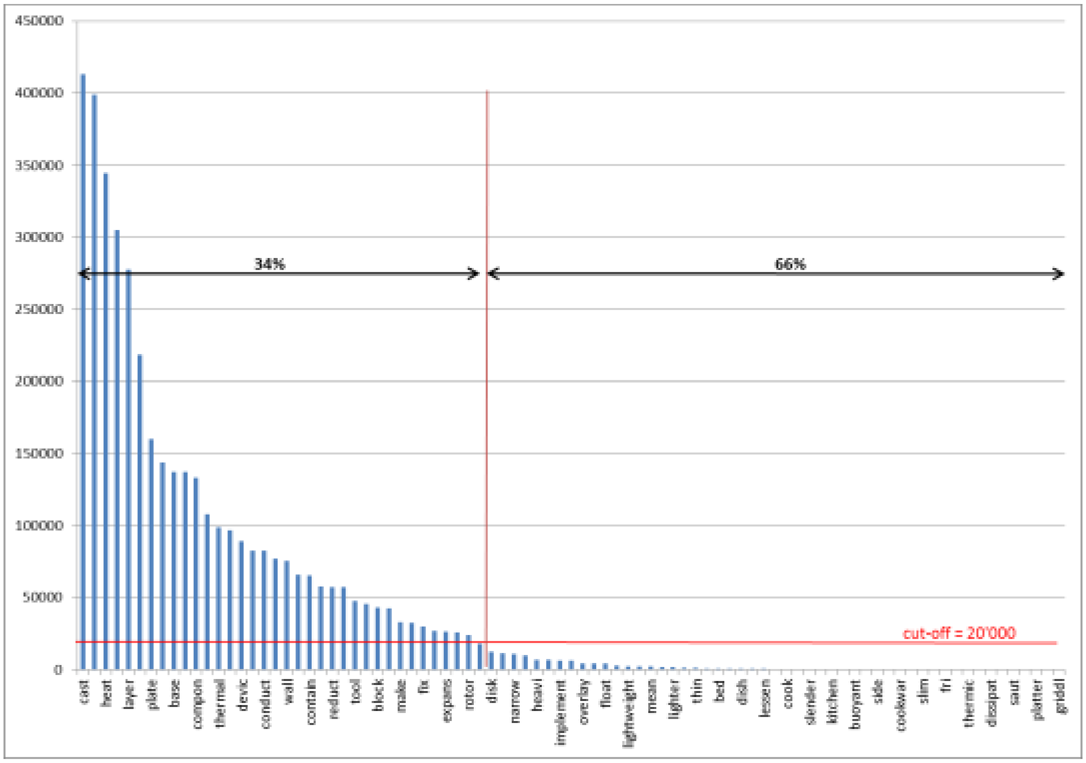


Figure 52- The ontological semantic keywords of the product features (ranked based on their term frequencies in the corpus)

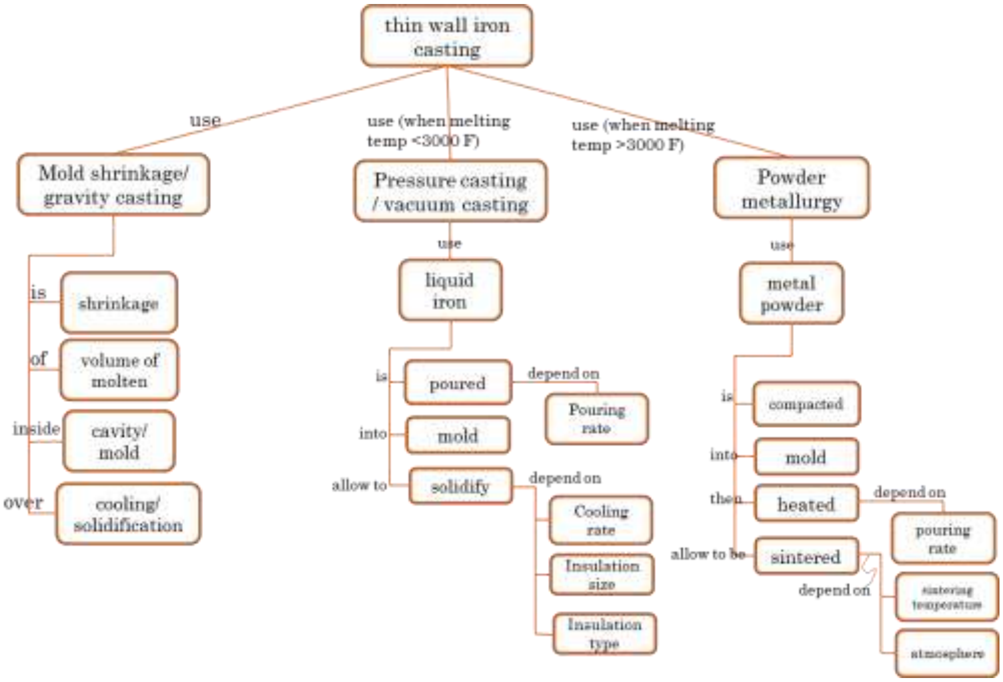


Figure 53- The first edition of the ontology of the enabling technologies (thin wall iron casting)

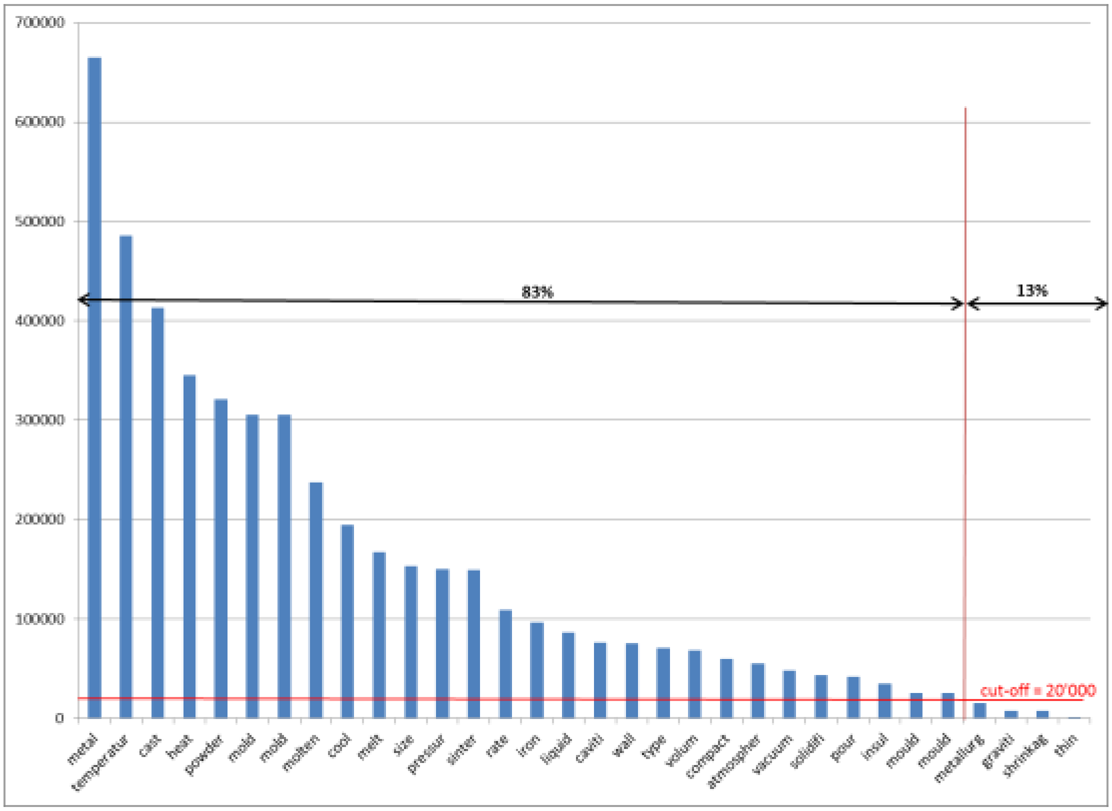


Figure 54- The ontological keywords of the first enabling technologies (ranked based on their term frequencies in the corpus)

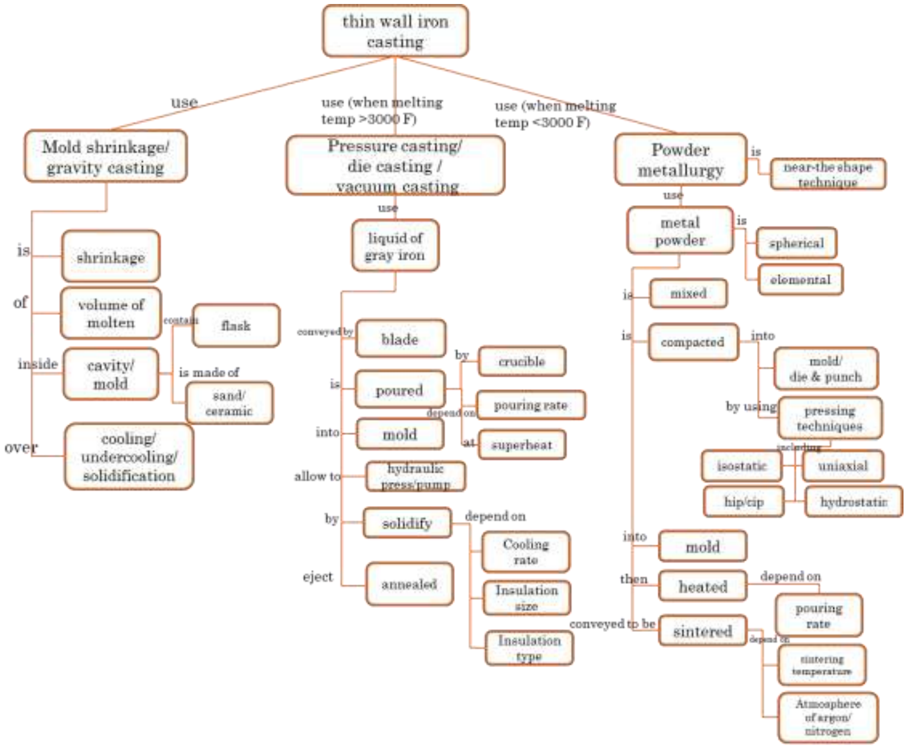


Figure 55- The second edition of the ontology of the enabling technologies (thin wall iron casing)

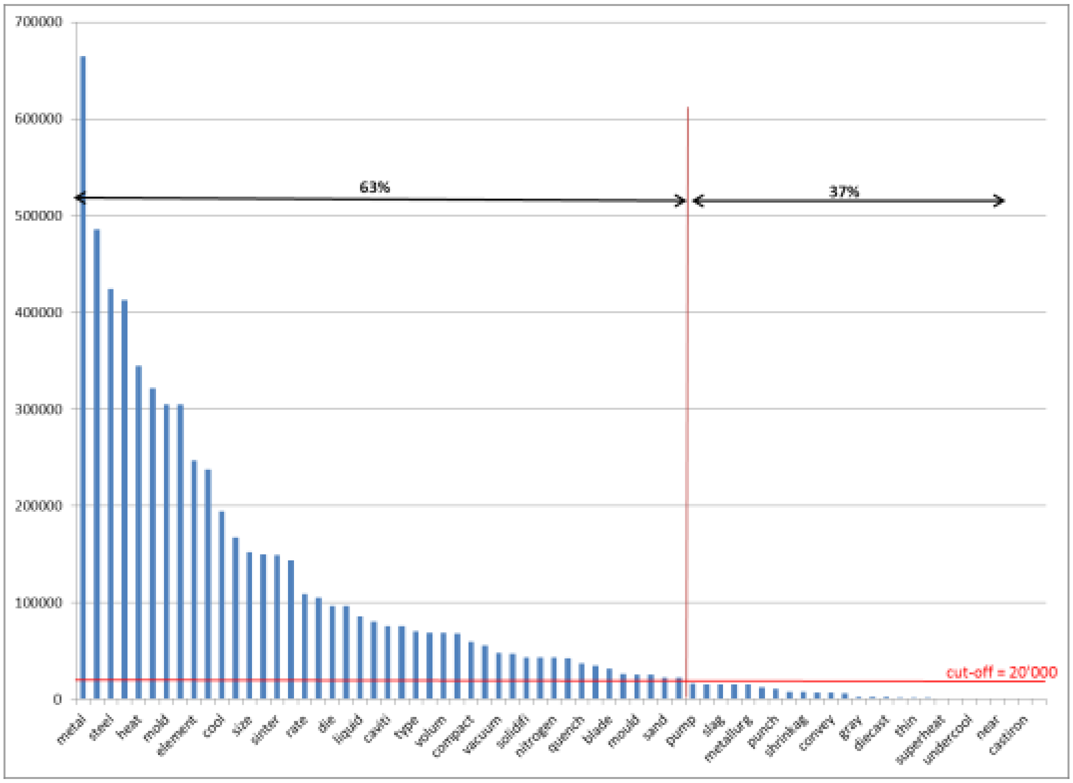


Figure 56- The ontological keywords of the 2nd edition of the enabling technologies (ranked based on their term frequencies in the corpus)

The ontological semantic keywords of the 2nd edition contain more discriminating keywords. The frequency of 59% of these keywords is greater than the cut-off value. Therefore, the OSC of the enabling technologies performs reasonably (see Table 42 in chapter 5).

In order to increase the performance of the proposed method, experts need to know the cut-off value and the total term frequency of ontological keywords during ontology design sessions to pick stronger keywords. This ability improves the efficiency of the method to achieve better performance in OC and OCS.

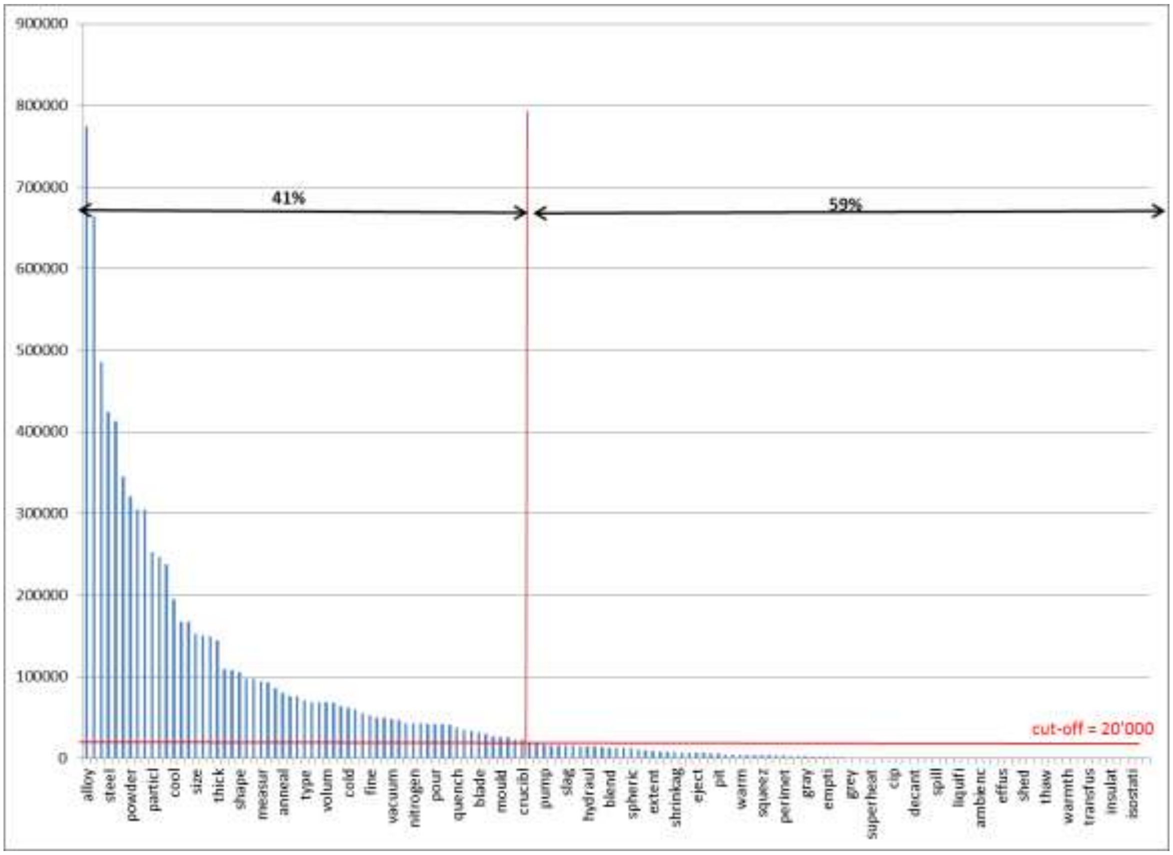


Figure 57- The ontological semantic keywords of the 2nd edition of the enabling technologies (ranked based on their term frequencies in the corpus)

6.3.2. Data set preparation

Experts should spend enough time to classify both relevant and irrelevant patents to the subject of an ontology (e.g., based on the findings of this dissertation, 30 minutes for 10 patents). Often, the number of relevant patents in a patent search is much less than the number of irrelevant patents. Finding irrelevant patents is not a big challenge because they can often be recognized just by looking at the title of the patent. On the other hand, finding relevant patents could be a challenge, since experts need to read parts of the content (e.g. the abstract, a few paragraphs of the description or a few paragraphs of the claims) of the patent to recognize its relevancy. Therefore, experts may initially provide just a small number of relevant patents. However, they should be ready to find more relevant patents during patent search, if they used weak performance measures during cross validation.

6.4. Wrap-up

1. Human experts are unreliable, inefficient, and not very effective when it comes to performance in patent search. Ontological semantic (OSC) classification is suggested to improve the performance of the experts in the patent search.
2. OSC has two characteristics that differentiate it from methods deployed by human experts in patent search. The characteristics are: 1) using computer capabilities including natural language processing, information retrieval, and classification as a machine learning method; and 2) relying on ontologies. These capabilities help improve the reliability, efficiency, and effectiveness of human experts' efforts in patent search.

3. The classifiers exhibit different performances for product features and enabling technologies. The classifiers perform very well for the product features, and they perform moderately well for the enabling technologies. The ontological keywords play a major role in the performance of the ontological classifications (OCs) and consequently in the performance of the ontological semantic classifications (OSCs). The total term frequency number of the ontological keywords determines how discriminative each ontological keyword is. Comparing the total term frequency number with the cut-off value determines the status of every ontological keyword. To boost the performance of OCs and OSCs, experts should consider the total term frequency number of ontological keywords during ontology design.
4. The more discriminative the ontological keywords, the better the classification performance. In order to have discriminative ontological keywords during ontology design, human experts should preferably take keywords whose total term frequency is lower than the cut-off value. Therefore, term frequency numbers should be calculated and cut-off value should be determined before ontology design.
5. Ontology evaluation (validation) is very important when experts do a patent search for a complex concept like the enabling technologies in the Finex case. The evaluation by other experts improves the quality of an ontology, and consequently improves the performance of a classification.

CHAPTER 7: CONCLUSIONS

In this chapter, I draw conclusions pertaining to the hypotheses described in section 1.3, which are based upon the results of the classifications for the product features and the enabling technologies that are discussed in chapters 4, 5 and 6. I also identify the limitations of applying natural language processing, information retrieval, and machine learning based on the results of the study. In addition, the theoretical contributions and practical implications of the study are reviewed. At the end, some extensions of the study are suggested for future research.

7.1. Main Findings

According to the Hypothesis 1, applying ontological keywords (ontological classification) improves the performance of classification over the base line (regular classification). The ROC AUC and F-score measures of the classifications for the product features (Table 31) and the ROC AUC and F-score measures of the classifications for the enabling technologies (Table 42) indicate that the performance of the ontological classification is better than that of regular classification. Therefore, the Hypothesis 1 is confirmed.

According to the Hypothesis 2, applying ontological keywords and their synonyms (ontological semantic classification) improves the performance of classification over the base line (regular classification). The ROC AUC and F-score measures of the classifications for the product features (Table 31) and the ROC AUC and F-score measures of the classifications for the enabling technologies (Table 42) indicate that the performance of the ontological semantic classification is better than that of regular classification. Therefore, Hypothesis 2 is confirmed.

According to Hypothesis 3, applying ontological keywords and their synonyms (ontological semantic classification) improves the performance of classification over that of ontological classification. The ROC AUC and F-score measures of the classifications for the product features (Table 31) and the ROC AUC and F-score measures of the classifications for the enabling technologies (Table 42) indicate that the performance of the ontological semantic classification is better than that of ontological classification. Therefore, Hypothesis 3 is confirmed.

According to the patent search results described in section 4.2 and discussed in section 6.1, experts performed unreliably, inefficiently and with relatively low effectiveness. Applying OSC in to patent search improves all factors pertaining to search behavior. OSC eliminates '*error*' (duplicates and contradictions), improves '*search speed*', and expands 'keywords diversity' by considering ontological keywords and their synonyms. Therefore, OSC increases patent search performance in terms of reliability, efficiency, and effectiveness.

7.2. Contributions

Fulk and Steinfeld's discussion on the uses of theory reveals the following potential contributions to theory that an academic study can make (Fulk & Steinfield, 1990).

1) to provide a framework for identifying empirical patterns;
2) to resolve inconsistencies across studies;
3) to generate hypotheses by which generalizable conclusions may be tested;
4) to provide perspective on larger issues;
5) to recommend directions for future research; and
6) to help integrate knowledge from related fields.

In alignment with Steinfeld and Fulk (1990), this dissertation has made the contributions to theory listed below. Each of the numbered items below corresponds to the item with the same number in Fulk and Steinfeld (1990).

1- The method developed in this dissertation has yielded the search framework, which can identify a concept in a textual corpus, as has demonstrated for specific CPC classes within the USPTO. Figure 58 compares this framework to the state of the art in patent search.

   As shown in Figure 58, patent searches that reflect the state of the art contain four cyclic steps. The expert that conducts the patent search starts with query design. He/she subsequently retrieves some patents and judges them. He/she

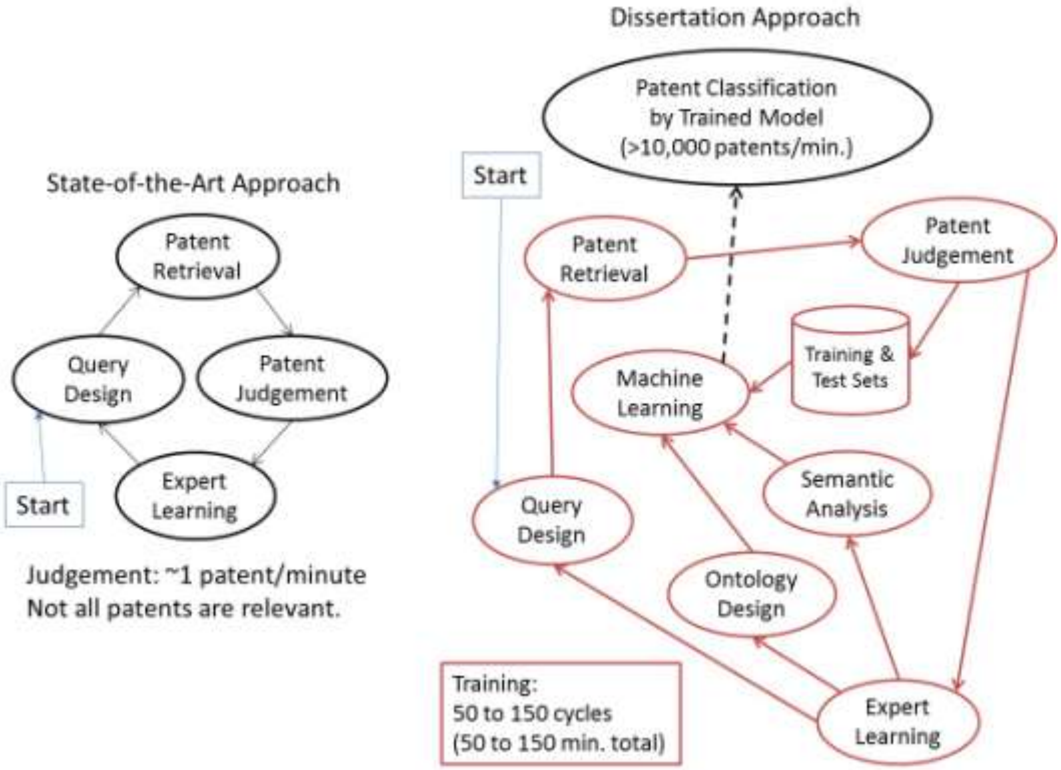


Figure 58- State of the Art vs. OSC Framework

revises his/her queries depending on what he/she learns. This process continues until the expert believes that he/she cannot find any more relevant results. The results from the patent search experiment indicate that a patent search that reflects the current state of the art has a judgement rate of about one patent per minute. Thus, an expert cannot retrieve and judge all patents in a corpus because of time and cost limitations. Furthermore, the majority of patents that are judged may end up being irrelevant.

The method developed in this dissertation enhances the performance of patent searches by applying machine learning. In this approach, the keywords used in the ontology (ontological keywords) and their synonyms are considered the main features that train classifiers such as k-NN and SVM. Classifiers determine whether the documents in the corpus are relevant or irrelevant to the concepts in the ontology. The expert still has to invest about 50 to 150 cycles in training the classifier, which takes from 50 to 150 minutes. After that, the machine classifier classifies patents at a rate in excess of 10,000 per minute. And, unlike human experts, the proposed method retrieves and figures out the relevance of all patents in a corpus. Thus, the method developed in this dissertation significantly improves upon the performance of patent searches that represent the current state of the art.

The approach developed in this dissertation is not restricted to the USPTO. Further research, which is described in section 7.5, would allow this method to be deployed in other patent databases (such as the European Patent Offices database). It could also be applied to texts that are not patents.

2- In many patent analysis studies, including those based on patent mining and citation analysis, data sets are provided based on two patent extraction methods: 1) retrieving patents based on queries; and 2) retrieving patents based on related CPC classes. In reality, the data sets provided based on the abovementioned methods contain many patents that are irrelevant to the subject of the study. In other words, those methods come up with data sets with lots of *noise*, which reduces the accuracy of the studies. The method described in this dissertation can be applied as a complementary method to reduce irrelevant patents (noise) from the data sets and improve the performance of patent analysis.

3- The study in this dissertation has shown that ontological semantic classification is the best scenario for patent classification. This approach is generalizable to a variety of topics that involve the analysis of many kinds of texts, not just patents. It is thus possible to generate hypotheses that test the approach in a variety of contexts.

4- To date, network approaches to patent analysis are restricted to citation analysis and keyword analysis. Ontological semantic classification allows us to look at patents from the network perspective. This dissertation may thus serve as the impetus for applying the network perspective to the content (abstract, description and claims) of patents.

5- The research conducted for this dissertation may motivate future researchers to investigate the relations between patents in terms of the concepts introduced in their corpuses. The researchers can look at the patent networks in terms of the extent to which the patents under investigation have covered the whole concept or just some components of the concept (as modeled by ontologies).

6- Finally, the classification scheme proposed in this dissertation integrates three fields of computer science—Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning—with ontological semantic analysis.

7.3. Implications for Practitioners

The method developed in this dissertation can be applied widely in areas where practitioners need to recognize one or multiple concepts in patent databases that contain a large number of documents. Therefore, the proposed method provides the capability to recognize technologies in USPTO database for a variety of purposes such as:

- *Opportunity identification*: identifying technologies which match with specific product features and specific enabling technologies.
- *Technology landscaping*: identifying technologies which match specific product features.
- *Technology Acquisition*: identifying companies that developed a specific technology.

In addition, the proposed method can be applied in areas outside of technology management such as infringement analysis, which is a method to identify patents that ignore prior intellectual property.

7.4. Limitations

This proposed research is subject to the following limitations:

1- The data set (patents of C22C and B22D sub-classes) is provided by the USPTO database, while experts in this study use the Google patent search engine. The content of some patents identified by the Google engine are not available in the USPTO database. This problem is not limited to old patents. Even, some recently issued patents are not available in USPTO database. To solve this problem, the unavailable patents are manually downloaded from the Google database and added manually to the data set.

2- WordNet database is used as the main source of the synonyms of the ontological keywords. Some keywords may have been common synonyms in the field of the study, but they are not considered as a synonym in WordNet. For example, ‘lightweight’ is a common term used in the patents, but it doesn’t exist in WordNet, so it is not available as a synonym of ‘light’.

3- The more time is spent by the experts, the more patents are classified by experts, and the better the performance of the classifiers become. The time constraint of experts is always an important limitation in this study.

7.5. Future Studies

There are a few directions suggested for the future studies:

1- It is expected that this research will be applicable on European patents because the structure of these patents is similar to those of the USPTO database, and these patents are written in English. Further research is required to assess whether the approach developed in this dissertation can be to other foreign patent databases. Chinese and Japanese patents are of special concern because of the large linguistic differences between English and these Far Eastern languages, as well as the fundamental differences between western and kanji-based in writing.

2- A literature review is a part of almost every research study. Researchers search the literature to identify multiple concepts that are relevant to their research. They can use the proposed method to classify the papers based on every concept they identify. This method improves the performance of researchers to review a large number of papers in a short time. They can use complementary methods such as keyword network analysis and clustering analysis to come up with more tangible results. This has not been done to date, and consequently should be considered a subject of future research.

3- The proposed method provides a background to do network analysis based on the content of papers. After recognizing relevant patents by using the proposed method, a researcher can generate a network of patents based on the similarity between patents and the ontology under study. The nodes are the patents and the links are the similarities measured between patents. The similarities are determined based on how much two patents are similar according to the ontology under study. Core-Periphery structure analysis (Madani, Daim, & Weng, 2015) can be applied to identify novel opportunities.

Developing technology intelligence tools, such as the abovementioned suggestions and the method of this dissertation, will make technical and academic information more available and accessible to practitioners and decision makers. As a consequence, practitioners and decision makers reduce their reliance on experts to extract data from data sources, to convert the data to information, and to generate knowledge from information, in order to make wise decisions.

References

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, *37*, 3–13. http://doi.org/10.1016/j.wpi.2013.12.006

Acosta, M., & Coronado, D. (2003). Science-technology flows in Spanish regions - An analysis of scientific citations in patents. *Research Policy*, *32*(10), 1783–1803.

Acosta, M., Coronado, D., & Angeles Martinez, M. (2012). Spatial differences in the quality of university patenting: Do regions matter? *Reseach Policy*, *41*(4), 692–703.

Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Classification Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 163–222). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4614-3223-4

Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–519. http://doi.org/10.1016/0166-4972(96)00031-4

Atkinson, K. H., & H., K. (2008). Toward a more rational patent search paradigm. In *Proceeding of the 1st ACM workshop on Patent information retrieval - PaIR '08* (p. 37). New York, New York, USA: ACM Press. http://doi.org/10.1145/1458572.1458582

Baeza Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Adison-Wesley. Retrieved from http://dialnet.unirioja.es/servlet/libro?codigo=369152

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, *1*(1), 4–20. http://doi.org/10.4304/jait.1.1.4-20

Belkin, N. J., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., … Cool, C. (2003). Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03* (p. 205). New York, New York, USA: ACM Press. http://doi.org/10.1145/860435.860474

Bermudez-Edo, M., Noguera, M., Hurtado-Torres, N., Hurtado, M. V., & Garrido, J. L. (2013). Analyzing a firm's international portfolio of technological knowledge: A declarative ontology-based OWL approach for patent documents. *Advanced Engineering Informatics*, *27*(3), 358–365. http://doi.org/10.1016/j.aei.2013.02.003

Bloehdorn, S., Cimiano, P., & Hotho, A. (2006). Learning Ontologies to Improve Text Clustering and Classification. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, & W. Gaul (Eds.), *From Data and Information Analysis to Knowledge Engineering* (pp. 334–341). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-31314-1

Bonaccorsi Andrea, F. G. (2007). Expanding the Functional Ontology in Conceptual Design. In *ICED07: 16th International Conference of Engineering Design*. Retrieved from https://www.designsociety.org/publication/25716/expanding_the_functional_ontology_in_conceptual_design

Bonino, D., Ciaramella, A., & Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information*, *32*(1), 30–38. http://doi.org/10.1016/j.wpi.2009.05.008

Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, *109*(2), 155–172. http://doi.org/10.1108/02635570910930073

Braschler, M., Harman, D. K., Pianta, E., & CLEF. (2010). Prior art retrieval using the different sections in patent documents. In *Proceedings of the Conference on*

Multilingual and Multimodal Information Access Evaluation. Clef-ip 2010: . Retrieved from https://www.narcis.nl/publication/RecordID/oai:repository.ubn.ru.nl:2066%2F85991

Breitzman, A., & Thomas, P. (2002). Using Patent Citation Analysis to Target/Value M&A Candidates. *Research Technology Management*, *45*(5). Retrieved from http://www.ingentaconnect.com/content/iri/rtm/2002/00000045/00000005/art00006

Brenner, M. S. (1996). Technology Intelligence and Technology Scouting. *Comepetitiv Intelligence Review*, *7*(3), 20–27.

Brockhoff, K. (1991). Competitor technology intelligence in German companies. *Industrial Marketing Management*, *20*(2), 91–98. http://doi.org/10.1016/0019-8501(91)90027-D

Cascini, G., Fantechi, A., & Spinicci, E. (2004). *Natural language processing of patents and technical documentation*. *Document analysis systems VI*. Springer Berlin Heidelberg.

Cascini, G., Russo, D., & Zini, M. (2007). Computer-Aided Patent Analysis: finding invention peculiarities. In *INTERNATIONAL FEDERATION FOR INFORMATION PROCESSING-PUBLICATIONS-IFIP* (Vol. 250).

Cascini, G., & Zini, M. (2011). Computer-aided comparison of thesauri extracted from complementary patent classes as a means to identify relevant field parameters. In *Global Product Development* (pp. 555–566). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-15973-2_56

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems*, *14*(1), 20–26. Retrieved from http://www.computer.org/csdl/mags/ex/1999/01/x1020.pdf

Chang, S.-B., Lai, K.-K., & Chang, S.-M. (2009). Exploring technology diffusion and classification of business methods: Using the patent citation network. *Technological Forecasting and Social Change*, *76*(1), 107–117.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.

Chen, C. (2004). *Information Visualization*. Springer Berlin Heidelberg.

Chen, C. (2006). CiteSpace II : Detecting and Visualizing Emerging Trends. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. http://doi.org/10.1002/asi

Chen, C. (2014). CiteSpac. Retrieved from http://cluster.cis.drexel.edu/~cchen/citespace/

Chen, C., Zhang, J., & Vogeley, M. S. (2009). Visual Analysis of Scientific Discoveries and Knowledge Diffusion. In *the 12th International Conference on Scientometrics and Informetrics* (pp. 14–17).

Chen, Y.-L., & Chiu, Y.-T. (2013). Cross-language patent matching via an international patent classification-based concept bridge. *Journal of Information Science*, *39*(6), 737–753.

Chiavetta, D., & Porter, A. (2013). Tech mining for innovation management. *Technology Analysis & Strategic Management*, *25*(6), 617–618. http://doi.org/10.1080/09537325.2013.802933

Choi, J., & Hwang, Y. (2014a). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting & Social Change*, *83*, 170–182. http://doi.org/10.1016/j.techfore.2013.07.004

Choi, J., & Hwang, Y.-S. (2014b). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting and Social Change*, *83*, 170–182. http://doi.org/10.1016/j.techfore.2013.07.004

Choi, S., Kim, H., Yoon, J., Kim, K., & Lee, J. (2013). An SAO-based text-mining approach for technology roadmapping using patent information. *R&D Management*, *43*(1), 52–74.

Choi, S., Yoon, J., Kim, K., & Kim, C.-H. (2011). SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics*, *88*(3).

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. http://doi.org/10.1002/aris.1440370103

Cooper, R. (1987). New products: What separates winners from losers? *Journal of Product Innovation Management*, *4*(3), 169–184. http://doi.org/10.1016/0737-6782(87)90002-6

Cooper, R. G. (1979). The Dimensions of Industrial New Product Success and Failure. *Journal of Marketing*, *43*(3), 93–103.

Coronado, D., & Acosta, M. (2005). The effects of scientific regional opportunities in science-technology flows: Evidence from scientific literature in firms patent data. *The Annals of Regional Scienc*, *39*(3), 495–522.

Dean Brown, J. (2000). What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter*, *4*(2), 8–12.

Droge, C., Jayaram, J., & Vickery, S. K. (2004). The effects of internal versus external integration practices on time-based performance and overall firm performance. *Journal of Operations Management*, *22*(6), 557–573. http://doi.org/10.1016/j.jom.2004.08.001

Dunlop, M. (2000). Reflections on Mira: Interactive evaluation in information retrieval.

Journal of the American Society for Information Science, *51*(14), 1269–1274. http://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1042>3.0.CO;2-7

Economics and statistics administration, and unitated states patent and trademark office. (2012). Intellectual property and the U.S. economy: industries in focus. Retrieved November 26, 2014, from http://www.uspto.gov/news/publications/IP_Report_March_2012.pdf

Enkel, E., Gassmann, O., & Chesbrough, H. (2009). Open R&D and open innovation: exploring the phenomenon. *R&D Management*, *39*(4), 311–316. http://doi.org/10.1111/j.1467-9310.2009.00570.x

Fafalios, P., & Tzitzikas, Y. (2017, January 2). Patent Retrieval: A Literature Review. arXiv preprint arXiv. Retrieved from http://arxiv.org/abs/1701.00324

Fattori, M., Pedrazzi, G., & Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, *25*(4), 335–342. http://doi.org/10.1016/S0172-2190(03)00113-3

Fox, C. (1989). A stop list for general text. *ACM SIGIR Forum*, *24*(1–2), 19–21. http://doi.org/10.1145/378881.378888

Fu, K., Cagan, J., Kotovsky, K., & Wood, K. (2013). Discovering Structure in Design Databases Through Functional and Surface Based Mapping. *Journal of Mechanical Design*, *135*(3), 31006. http://doi.org/10.1115/1.4023484

Fu, K., Chan, J., Schunn, C., Cagan, J., & Kotovsky, K. (2013a). Expert representation of design repository space: A comparison to and validation of algorithmic output. *Design Studies*, *34*(6), 729–762. http://doi.org/10.1016/j.destud.2013.06.002

Fu, K., Chan, J., Schunn, C., Cagan, J., & Kotovsky, K. (2013b). Testing the Basis for an Automated Design-by-Analogy Tool Through Comparison to Expert Thinking. In *Volume 5: 25th International Conference on Design Theory and Methodology; ASME*

2013 Power Transmission and Gearing Conference (p. V005T06A026). ASME. http://doi.org/10.1115/DETC2013-12128

Fu, K., Murphy, J., Yang, M., Otto, K., Jensen, D., & Wood, K. (2013). Investigating the Effect of Functionality Level of Analogical Stimulation on Design Outcome. In *Korea-Japan Design Engineering Workshops (DEWS)*. Kitakyushu, Fukuoka, Japan.

Fulk, J., & Steinfield, C. (1990). The Theory Imperative. In *Organizations and communication technology*. Sage Newbury Park. Retrieved from http://sk.sagepub.com/books/organizations-and-communication-technology/n1.xml

Gavrilova, T., Farzan, R., & Brusilovsky, P. (2005). One practical algorithm of creating teaching ontologies. *12th International Network- …*. Retrieved from http://lauda.ulapland.fi/bitstream/handle/10024/59474/nbe_book2005.pdf?sequence=1#page=36

General Patent Statistics Reports. (2016). Retrieved from http://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm

Gerken, J. M. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645.

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*. Retrieved from http://nsuworks.nova.edu/tqr/vol8/iss4/6

Gordon, W. J. J. (1961). *Synectics: The development of creative capacity.*

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. http://doi.org/10.4304/jetwi.1.1.60-76

Han, J., Kamber, M., & Pei, J. (2011). Classification: Basic Concepts. In *Data Mining: Concepts and Techniques* (pp. 327–392). Elsevier. Retrieved from https://books.google.com/books?hl=en&lr=lang_en&id=pQws07tdpjoC&pgis=1

Hanbury, A., Bhatti, N., Lupu, M., & Mörzinger, R. (2011). Patent image retrieval. In *Proceedings of the 4th workshop on Patent information retrieval - PaIR '11* (p. 3). New York, New York, USA: ACM Press. http://doi.org/10.1145/2064975.2064979

Harman, D. (2011). Information Retrieval Evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *3*(2), 1–119. http://doi.org/10.2200/S00368ED1V01Y201105ICR019

Hiemstra, D. (2009). Information Retrieval Models. In *Information Retrieval: Searching in the 21st Century* (p. 320). John Wiley & Sons. Retrieved from https://books.google.com/books?hl=en&lr=lang_en&id=nH9G7LCR1-UC&pgis=1

Hlomani, H., & Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, *1*, 1–5. Retrieved from http://www.semantic-web-journal.net/system/files/swj657.pdf

Hong, L. Y., Hua, T. R., & Hong, J. (2013). Study on patent text classification for product innovative design. *Computer Integrated Manufacturing Systems*, *19*(2), 382–390. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTOTAL-JSJJ201302020.htm

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*, 19–62. http://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Hull, F. M. (2004). A Composite Model of Product Development Effectiveness: Application to Services. *IEEE Transactions on Engineering Management*, *51*(2), 162–172. http://doi.org/10.1109/TEM.2004.826015

Intellectual Property Owners Accosiation. (2013). Top 300 organizations granted U.S. patents in 2012. Retrieved November 26, 2014, from http://www.ipo.org/wp-content/uploads/2013/06/Top-300_6.23.13.pdf

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Jarczyk, A. P. J., Loffler, P., & Shipmann, F. M. (1992). Design rationale for software engineering: a survey. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences* (Vol. ii, pp. 577–586 vol.2). IEEE. http://doi.org/10.1109/HICSS.1992.183309

Jeon, J., Lee, C., & Park, Y. (2011). How to Use Patent Information to Search Potential Technology Partners in Open Innovation. *Journal of Intellectual Property Rights*, *16*(5), 385–393.

Jetter, A. (2006). Elicitation – Extracting Knowledge from Experts. In *Knowledge Integration* (pp. 65–76). Physica-Verlag HD.

Jing, L., Zhou, L., Ng, M. K., & Huang, J. Z. (2006). Ontology-based distance measure for text clustering. In *Proc. of SIAM SDM workshop on text mining*. Maryland, USA.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features (pp. 137–142). Springer Berlin Heidelberg. http://doi.org/10.1007/BFb0026683

Joho, H., Azzopardi, L. A., & Vanderbauwhede, W. (2010). A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceeding of the third symposium on Information interaction in context - IIiX '10* (p. 13). New York, New York, USA: ACM Press. http://doi.org/10.1145/1840784.1840789

Jun, S., Park, S. S., & Jang, D. S. (2012). Technology forecasting using matrix map and patent clustering. *Industrial Management & Data Systems*, *112*(5–6), 786–807.

Kajikawa, Y., & Takeda, Y. (2008). Structure of research on biomass and bio-fuels: A citation-based approach. *Technological Forecasting and Social Change*, *75*(9), 1349–1359.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, *75*(6), 771–782. http://doi.org/10.1016/j.techfore.2007.05.005

Karvonen, M., & Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting and Social Change*, *80*(6), 1094–1107. http://doi.org/10.1016/j.techfore.2012.05.006

Kerr, C. I. V., Mortara, L., Phaal, R., & Probert, D. R. (2006). A conceptual model for technology intelligence. *International Journal of Technology Intelligence and Planning*, *2*(1). Retrieved from http://inderscience.metapress.com/content/acr6wk60tcn9bybm/

Kogut, B., & Zander, U. (1992). Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. *Organization Science*, *3*(3), 383–397. http://doi.org/10.1287/orsc.3.3.383

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques*. New Age International. Retrieved from https://books.google.com/books?hl=en&lr=lang_en&id=hZ9wSHysQDYC&pgis=1

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies* (pp. 249–268). IOS Press. Retrieved from https://books.google.com/books?hl=en&lr=lang_en&id=vLiTXDHr_sYC&pgis=1

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–

190. http://doi.org/10.1007/s10462-007-9052-3

Lancaster, F. W. (1968). *Evaluation of the MEDLARS Demand Search Service.* Washington D.C.: National Library of Medicine. Retrieved from https://eric.ed.gov/?id=ED022494

Lawrence, P. R., & Lorsch, J. W. (1967). Differentiation and Integration in Complex Organizations. *Administrative Science Quarterly*, *12*(1), 1. http://doi.org/10.2307/2391211

Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, *79*(1), 16–29.

Lee, C., Jeon, J., & Park, Y. (2011a). Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach. *Technological Forecasting and Social Change*, *78*(4), 690–702. http://doi.org/10.1016/j.techfore.2010.11.010

Lee, C., Jeon, J., & Park, Y. (2011b). Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach. *Technological Forecasting and Social Change*, *78*, 690–702. http://doi.org/10.1016/j.techfore.2010.11.010

Lee, C., Park, H., Kim, C., & Park, Y. (2011). Monitoring Evolutionary Trends in Electronic Business Models: A Dynamic Patent Analysis Approach. *MANAGEMENT AND SERVICE SCIENCE*, *8*, 28–32. Retrieved from http://apps.webofknowledge.com.proxy.lib.pdx.edu/full_record.do?product=WOS&search_mode=AdvancedSearch&qid=7&SID=4CqwGX1FL8hLeYN2u4n&page=1&doc=7

Lee, J., & Lai, K.-Y. (1991). What's in Design Rationale? *Human-Computer Interaction*, *6*, 251–280.

Lee, P.-C., Su, H.-N., & Chan, T.-Y. (2010). Assessment of ontology-based knowledge network formation by Vector-Space Model. *Scientometrics*, *85*(3), 689–703.

Lee, P.-C., Su, H.-N., & Wu, F.-S. (2010a). Quantitative mapping of patented technology - The case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, *77*(3), 466–478.

Lee, P.-C., Su, H.-N., & Wu, F.-S. (2010b). Quantitative mapping of patented technology — The case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, *77*(3), 466–478. http://doi.org/10.1016/j.techfore.2009.08.006

Lee, S., Kang, S., Oh, M., Kim, K., Park, E., Lee, S., & Park, Y. (2006). Using Patent Information for New Product Development: Keyword-Based Technology Roadmapping Approach. In *2006 Technology Management for the Global Future - PICMET 2006 Conference* (Vol. 3, pp. 1496–1502). IEEE. http://doi.org/10.1109/PICMET.2006.296714

Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change*, *76*(6), 769–786. http://doi.org/10.1016/j.techfore.2009.01.003

Lelescu, A., Langston, B., Louie, E., Cheng, I., Labrie, J., Colino, J., … Chen, Y. (2014). The Strategic IP Insight Platform (SIIP): A Foundation for Discovery. In *2014 Annual SRII Global Conference* (pp. 27–34). IEEE. http://doi.org/10.1109/SRII.2014.14

Li, Y.-R., Tong, G.-E., Hong, C.-F., & Wang, L. (2006). Exploring Cognitive Difference in Education using Average Path Length of Concept Map. In *2006 IEEE International Conference on Systems, Man and Cybernetics* (Vol. 3, pp. 2133–2136). IEEE.

http://doi.org/10.1109/ICSMC.2006.385176

Li, Z., Tate, D., Lane, C., & Adams, C. (2012). A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics. *Computer-Aided Design*, *44*(10), 987–1010. http://doi.org/10.1016/j.cad.2011.12.006

Liang, Y., & Liu, Y. (2013). Rationale-Based Patent Analysis for Corporate Product Design. In *Volume 2B: 33rd Computers and Information in Engineering Conference* (p. V02BT02A013). ASME. http://doi.org/10.1115/DETC2013-12792

Liang, Y., Liu, Y., Kwong, C. K., & Lee, W. B. (2012). Learning the "Whys": Discovering design rationale using text mining — An algorithm perspective. *Computer-Aided Design*, *44*(10), 916–930. http://doi.org/10.1016/j.cad.2011.08.002

Liang, Y., & Tan, R. (2007). A text-mining-based patent analysis in product innovative process. In *Trends in computer aided innovation* (Vol. 250, pp. 89–96). Retrieved from http://www.springerlink.com/index/3153t4701wh48k73.pdf

Liang, Y., Tan, R., & Ma, J. (2008). Patent analysis with text mining for TRIZ. In *2008 4th IEEE International Conference on Management of Innovation and Technology* (pp. 1147–1151). IEEE. http://doi.org/10.1109/ICMIT.2008.4654531

Linsey, J. S., Laux, J. P., Clauss, E., Wood, K. L., & Markman, A. B. (2007). Increasing Innovation: A Trilogy of Experiments Towards a Design-by-Analogy Method. In *Volume 3: 19th International Conference on Design Theory and Methodology; 1st International Conference on Micro- and Nanosystems; and 9th International Conference on Advanced Vehicle Tire Technologies, Parts A and B* (pp. 145–159). ASME. http://doi.org/10.1115/DETC2007-34948

Liu, L., Kang, J., Yu, J., & Wang, Z. (2005). A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering. In *International Conference on Natural*

Language Processing and Knowledge Engineering (pp. 597–601). IEEE. http://doi.org/10.1109/NLPKE.2005.1598807

Liu, Y., Liang, Y., Kwong, C. K., & Lee, W. B. (2010). A New Design Rationale Representation Model for Rationale Mining. *Journal of Computing and Information Science in Engineering*, *10*(3), 31009. http://doi.org/10.1115/1.3470018

Lott, B. (2012). *Survey of Keyword Extraction Techniques*. *UNM Education*. Retrieved from http://www.cs.unm.edu/~pdevineni/papers/Lott.pdf

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165. http://doi.org/10.1147/rd.22.0159

Machado, C., & Davim, P. (2017). *Productivity and Organizational Management - Google Books*. De Gruyter. Retrieved from https://books.google.com/books?id=rWmtDgAAQBAJ&pg=PA132&dq=productivity+management+definition+efficiency+effectiveness&hl=en&sa=X&ved=0ahUKEwjG0re6y-vWAhVpxFQKHT8gCyEQ6AEIJzAA#v=onepage&q=productivity management definition efficiency effectiveness

Madani, F. (2014). *The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis*.

Madani, F., & Weber, C. (2016). The evolution of patent mining: applying bibliometrics analysis and keyword network analysis. *World Patent Information*, *46*, 32–48.

Madani, F., & Zwick, M. (2017). Smart Building Technology Impact Analysis via Information Theory. *Technological Forecasting & Social Change*, *Under revi*.

Magdy, W., Leveling, J., & Jones, G. J. F. (2010). Exploring Structured Documents and Query Formulation Techniques for Patent Retrieval (pp. 410–417). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-15754-7_48

Maglia, A. (2006). What is an ontology? In *AmphibiaTree 2006 Workshop*.

Mahdabi, P., Keikha, M., Gerani, S., Landoni, M., & Crestani, F. (2011). Building Queries for Prior-Art Search (pp. 3–15). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-21353-3_2

Mann, D. (2001). An Introduction to TRIZ: The Theory of Inventive Problem Solving. *Creativity and Innovation Management*, *10*(2), 123–125. http://doi.org/10.1111/1467-8691.00212

McDonald-Maier, L. (2009). esp@cenet®: Survey reveals new information about users. *World Patent Information*, *31*(2), 142–143. http://doi.org/10.1016/j.wpi.2008.10.006

Melin, G., & Danell, R. (2000). A bibliometric mapping of the scientific landscape on Taiwan. *Issues & Studies*, *36*(5), 61–82.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://doi.org/10.1145/219717.219748

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Montobbio, F., & Sterzi, V. (2011). Inventing together: exploring the nature of international knowledge spillovers in Latin America. *Journal of Evolutionary Economics*, *21*(1), 53–89.

Moreno, D. P., Hernández, A. A., Yang, M. C., Otto, K. N., Hölttä-Otto, K., Linsey, J. S., … Linden, A. (2014). Fundamental studies in Design-by-Analogy: A focus on domain-knowledge experts and applications to transactional design problems. *Design Studies*, *35*(3), 232–272. http://doi.org/10.1016/j.destud.2013.11.002

Murphy, J., Fu, K., Otto, K., Yang, M., Jensen, D., & Wood, K. (2014). Facilitating Design-by-Analogy: Development of a Complete Functional Vocabulary and Functional Vector Approach to Analogical Search. In *Volume 2A: 40th Design Automation*

Conference (p. V02AT03A010). ASME. http://doi.org/10.1115/DETC2014-34491

Murphy, J., Fu, K., Otto, Ke., Jensen, D., Wood, K., & Yang, M. (2014). Facilitating Design-By-Analogy: Development Of A Complete Functional Vocabulary And Functional Vector Approach To Analogical Search. In *the ASME 2014 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference* (pp. 1–11). Buffalo, New York, USA.

Murphy, K. R., & Davidshofer, C. O. (2005). Psychological Testing: Principles and Applications. Retrieved March 20, 2016, from http://www.amazon.com/Psychological-Testing-Principles-Applications-Edition/dp/0131891723

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–51. http://doi.org/10.1136/amiajnl-2011-000464

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the international conference on Knowledge capture - K-CAP ’03* (p. 70). New York, New York, USA: ACM Press. http://doi.org/10.1145/945645.945658

Newman, M. (2005). Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, *46*(5), 323–351. http://doi.org/10.1080/00107510500052444

Newton, D. (2000). A survey of users of the new British Library Patent Information Centre. *World Patent Information*, *22*(4), 317–323. http://doi.org/10.1016/S0172-2190(00)00068-5

Nirenburg, S., & Raskin, V. (2004). Ontological Semantics. Retrieved April 12, 2016, from https://mitpress.mit.edu/books/ontological-semantics

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

Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*.

Pahl, G., Beitz, W., Feldhusen, J., & Grote, K.-H. (2007). *Engineering Design: A Systematic Approach*. Springer Science & Business Media. Retrieved from http://books.google.com/books?hl=en&lr=lang_en&id=57aWTCE3gE0C&pgis=1

Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, *40*(7), 2373–2390.

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics*, *90*(2), 515–529.

Park, Y., & Lee, S. (2012). Patent analysis for promoting technology transfer in multi-technology industries: the Korean aerospace industry case. *Journal of Technology Transfer*, *37*(3), 355–374.

Petrelli, D. (2008). On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing & Management*, *44*(1), 22–38. http://doi.org/10.1016/J.IPM.2007.01.024

Peyrot, M. (1996). Causal Analysis: Theory and Application. *Journal of Pediatric Psychology*, *21*(1), 3–24. http://doi.org/10.1093/jpepsy/21.1.3

Porter, A. L. (2005). QTIP: Quick technology intelligence processes. *Technological Forecasting and Social Change*, *72*(9), 1070–1081. http://doi.org/10.1016/j.techfore.2004.10.007

Porter, A. L., & Cunningham, S. W. (2005). *Tech mining : exploiting new technologies for competitive advantage*. N.J: Wiley.

Prahalad, C. K., & Hamel, G. (2006). The Core Competence of the Corporation. In D. Hahn & B. Taylor (Eds.), *Strategische Unternehmungsplanung — Strategische Unternehmungsführung* (pp. 275–292). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/3-540-30763-X

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on Applied computing - SAC '04* (p. 1232). New York, New York, USA: ACM Press. http://doi.org/10.1145/967900.968151

Quinn, G. (2017). Enhance Product Development: Ideation, Design and Licensing for Inventors - IPWatchdog.com | Patents &amp; Patent Law. Retrieved November 28, 2017, from http://www.ipwatchdog.com/2017/09/09/enhance-product-development-ideation-design-licensing-inventors/id=87819/

Regli, W. C., Hu, X., Atwood, M., & Sun, W. (2014). A Survey of Design Rationale Systems: Approaches, Representation, Capture and Retrieval. *Engineering with Computers*, *16*(3–4), 209–235. http://doi.org/10.1007/PL00013715

Ruffaldi, E., Sani, E., & Bergamasco, M. (2010). Visualizing perspectives and trends in robotics based on patent mining. In *2010 IEEE International Conference on Robotics and Automation* (pp. 4340–4347). IEEE. http://doi.org/10.1109/ROBOT.2010.5509648

Russo, D. (2011). Knowledge Extraction from Patent: Achievements and Open Problems. A Multidisciplinary Approach to Find Functions. In A. Bernard (Ed.), *Global Product Development* (pp. 567–576). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-15973-2

Russo, D. (2014). Function-based patent search: achievements and open problems. *International Journal of Product Development*, *19*(1–3). Retrieved from http://inderscience.metapress.com/content/46744p7736j54812/

Russo, D., & Montecchi, T. (2011). A Function-Behaviour Oriented Search for Patent Digging. In *Volume 2: 31st Computers and Information in Engineering Conference, Parts A and B* (pp. 1111–1120). ASME. http://doi.org/10.1115/DETC2011-47733

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. http://doi.org/10.1016/0306-4573(88)90021-0

Salton, G., & Michael, J. (1986). *Introduction to Modern Information Retrieval*. McGill.

Salton, G., Wong, A., & Yang., C. S. (1975). A vector space model for information retrieval. *Communications of the ACM*, *18*(11), 613–620.

Schilling, M. A., & Hill, C. W. . (1998). Managing the new product development process: Strategic imperatives. *The Academy of Management Executive (1993-2005)*, *12*(3), 67–81.

Scopel, F., GREGOLIN, L. A. R., & FARIA, L. I. L. (2013). Tendências tecnológicas do uso do sisal em compósitos a partir da prospecção em documentos de patentes. *Polímeros Ciência E Tecnologia*, *23*(4), 514–520. http://doi.org/10.4322/polimeros.2013.044

Search Help. (2016). Retrieved from http://www.uspto.gov/trademarks/resources/Search_Help.pdf

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. http://doi.org/10.1145/505282.505283

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., & Spiteri, R. (2007). Document Representation and Dimension Reduction for Text Clustering. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* (pp. 770–779).

IEEE. http://doi.org/10.1109/ICDEW.2007.4401066

Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology - Case study of a solar cell. *Technological Forecasting and Social Change*, *77*(7), 1147–1155.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011). Detecting potential technological fronts by comparing scientific papers and patents. *Foresight*, *13*(5), 51–60. http://doi.org/10.1108/14636681111170211

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*(11), 758–775.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2009). Early detection of innovations from citation networks. *2009 IEEE International Conference on Industrial Engineering and Engineering Management*, 54–58. http://doi.org/10.1109/IEEM.2009.5373444

Shih, M.-J., Liu, D.-R., & Hsu, M.-L. (2010). Discovering competitive intelligence by mining changes in patent trends. *Expert Systems with Applications*, *37*(4), 2882–2890. http://doi.org/10.1016/j.eswa.2009.09.001

Shin, J., & Park, Y. (2007). Building the national ICT frontier: The case of Korea. *Information Economics and Policy*, *19*(2), 249–277.

Singhal, A. (2001). Modern information retrieval: a brief overview. *IEEE Data Engineering Bulletin*, *24*(4), 35–43. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.7676

Souder, W. E. (1988). Managing relations between R&D and marketing in new product development projects. *Journal of Product Innovation Management*, *5*(1), 6–19. http://doi.org/10.1016/0737-6782(88)90029-X

Sutton, C. (2012). Nearest-neighbor methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(3), 307–309. http://doi.org/10.1002/wics.1195

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison Wesley.

Tang, B., Shepherd, M., Milios, E., & Heywood, M. (2005). Comparing and combining dimension reduction techniques for efficient text clustering. In *In Proceeding of SIAM International Workshop on Feature Selection for Data Mining* (pp. 17–26).

Thorleuchter, D., & Van den Poel, D. (2014). Semantic compared cross impact analysis. *Expert Systems with Applications*, *41*(7), 3477–3483. http://doi.org/10.1016/j.eswa.2013.10.051

Thorndike, R. L., & Hagen, E. P. (1986). *Measurement and Evaluation in Psychology and Education*. Macmillan. Retrieved from https://books.google.com/books/about/Measurement_and_Evaluation_in_Psychology.html?id=P-DfiaSbhyoC&pgis=1

Tijssen, R. J. . (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows. *Research Policy*, *30*(1), 35–54. http://doi.org/10.1016/S0048-7333(99)00080-3

Timonen, M. (2013). *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/37924/timonen_dissertation.pdf?sequence=1

Trappey, A. J. C., Trappey, C. V., Wu, C.-Y., Liaw, Y.-C., & Zhang, F.-X. (2013). Development of innovative product design process using patent multi-scale analysis and TRIZ methodology. In *Proceedings of International Conference on Computers and Industrial Engineering, CIE*.

Trappey, A. J. C., & Trappey, C. V. (2008). An R&D knowledge management method for patent document summarization. *Industrial Management and Data Systems*, *108*(1–2), 245–257.

Trappey, A. J. C., Trappey, C. V., Chiang, T.-A., & Huang, Y.-H. (2013). Ontology-based neural network for patent knowledge management in design collaboration. *International Journal of Production Research*, *51*(7), 1992–2005. http://doi.org/10.1080/00207543.2012.701775

Trappey, A. J. C., Trappey, C. V., & Wu, C.-Y. (2009). Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, *18*(1), 71–94.

Trappey, A. J. C., Trappey, C. V., & Wu, C.-Y. (2009). Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, *18*(1), 71–94. http://doi.org/10.1007/s11518-009-5100-7

Trappey, A., Trappey, C., & S. Kao, B. (2006). Automated Patent Document Summarization for R&D Intellectual Property Management. In *2006 10th International Conference on Computer Supported Cooperative Work in Design* (pp. 1–6). IEEE. http://doi.org/10.1109/CSCWD.2006.253004

Trappey, C. V., Trappey, A. J. C., & Wu, C.-Y. (2010). Clustering patents using non-exhaustive overlaps. *Journal of Systems Science and Systems Engineering*, *19*(2), 162–181.

Trappey, C. V., Wu, H.-Y., Taghaboni-Dutta, F., & Trappey, A. J. C. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics*, *25*(1), 53–64.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis.

Information Processing & Management, *43*(5), 1216–1247. http://doi.org/10.1016/j.ipm.2006.11.011

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, *9*(3), 187–222. http://doi.org/10.1145/125187.125188

Ulrich, K., & Eppinger, S. (2011). *Product Design and Development*. McGraw-Hill. Retrieved from http://www.amazon.com/Product-Design-Development-5th-Edition/dp/0073404772

Urban, G. L., Hauser, J. R., & Dholakia, N. (1987). *Essentials of new product management*. Englewood Cliffs : Prentice-Hall. Retrieved from http://library.wur.nl/WebQuery/clc/507612

USPTO. (2015). U.S. Patent Statistics Chart Calendar Years 1963 - 2014. Retrieved from http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm

USPTO Database, Advanced Search. (2016). Retrieved from http://patft.uspto.gov/netahtml/PTO/search-adv.htm

USPTO Database, Bollean Search. (2016). Retrieved from http://patft.uspto.gov/netahtml/PTO/search-bool.html

Verberne, S., & D’hondt, E. (2010). Prior Art Retrieval Using the Claims Section as a Bag of Words (pp. 497–501). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-15754-7_60

Verhaegen, P.-A., D’hondt, J., Vandevenne, D., & Dewulf, J. R. D. S. (2010). Automatically Characterizing Products through Product Aspects. In *Global Product Development* (pp. 595–605).

Verhaegen, P.-A., D’hondt, J., Vandevenne, D., Dewulf, S., & Duflou, J. R. (2011). Automatically characterizing products through product aspects. In *Global Product*

Development (pp. 595–605). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-15973-2_60

Verhaegen, P.-A., D'hondt, J., Vandevenne, D., Dewulf, S., & Duflou, J. R. (2011). Identifying candidates for design-by-analogy. *Computers in Industry*, *62*(4), 446–459. http://doi.org/10.1016/j.compind.2010.12.007

Veugelers, M., Bury, J., & Viaene, S. (2010). Linking technology intelligence to open innovation. *Technological Forecasting and Social Change*, *77*(2), 335–343. http://doi.org/10.1016/j.techfore.2009.09.003

Vikner, C., & Jensen, P. A. (2002). A Semantic Analysis of the English Genitive. Interaction of Lexical and Formal Semantics. *Studia Linguistica*, *56*(2), 191–226. Retrieved from http://web.a.ebscohost.com.proxy.lib.pdx.edu/ehost/detail/detail?sid=fa94454c-8709-4f39-8f20-a6f299c6259f%40sessionmgr4007&vid=0&hid=4109&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3D%3D#db=ufh&AN=6950681

Wang, J., Lu, W. F., & Loh, H. T. (2011). P-SMOTE: One Oversampling Technique for Class Imbalanced Text Classification. In *Volume 2: 31st Computers and Information in Engineering Conference, Parts A and B* (pp. 1089–1098). ASME. http://doi.org/10.1115/DETC2011-47313

Wang, W. M., & Cheung, C. F. (2011a). A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis. *Engineering Applications of Artificial Intelligence*, *24*, 1510–1520. http://doi.org/10.1016/j.engappai.2011.05.009

Wang, W. M., & Cheung, C. F. (2011b). A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis. *Engineering Applications of Artificial Intelligence*, *24*(8), 1510–1520.

Weng, S., & Chang, H. (2008). Using ontology network analysis for research document recommendation. *Expert Systems with Applications*, *34*(3), 1857–1869.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55. http://doi.org/10.1177/016555159201800106

Xu, Y. (2009). Apply text mining in analysis of patent document. In *2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design* (pp. 2350–2352). IEEE. http://doi.org/10.1109/CAIDCD.2009.5375302

Yang, C. H., Park, H. W., & Heo, J. (2010). A network analysis of interdisciplinary research relationships: the Korean government's R&D grant program. *Scientometrics*, *83*(1).

Yanhong, L., & Runhua, T. (2007). A Text-Mining-based Patent Analysis in Product Innovative Process. In *Trends in Computer Aided Innovation* (pp. 89–96).

YOON, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications*, *35*(1–2), 124–135. http://doi.org/10.1016/j.eswa.2007.06.022

Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, *15*(1), 37–50. http://doi.org/10.1016/j.hitech.2003.09.003

Yoon, J., Choi, S., & Kim, K. (2010). Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. *Scientometrics*, *86*(3), 687–703. http://doi.org/10.1007/s11192-010-0303-8

Yoon, J., Choi, S., & Kim, K. (2011). Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. *Scientometrics*, *86*(3), 687–703.

Yoon, J., & Kim, K. (2011). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, *90*(2), 445–461. http://doi.org/10.1007/s11192-011-0543-2

Yoon, J., & Kim, K. (2012). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications*, *39*(3), 2927–2938. http://doi.org/10.1016/j.eswa.2011.08.154

Zhang, D., & Tsai, J. J. P. (2003). Machine Learning and Software Engineering. *Software Quality Journal*, *11*(2), 87–119. http://doi.org/10.1023/A:1023760326768

Zhao, X., Huo, B., Selen, W., & Yeung, J. H. Y. (2011). The impact of internal integration and relationship commitment on external integration. *Journal of Operations Management*, *29*(1–2), 17–32. http://doi.org/10.1016/j.jom.2010.04.004

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey*.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Appendix A: The Questionnaire of Ontology Evaluation

1- Accuracy

The asserted knowledge in the ontology of thin wall casting agrees with the expert's knowledge about the domain.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more accurate:

……………………………………………………………………………………………

……………………………………………………………………………………………

2- Adaptability

The ontology of thin wall casting can be <u>used easily</u> in <u>different contexts</u> possibly by allowing it to be extended and specialized monotonically, i.e. without the need to remove axioms

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more adaptable:

……………………………………………………………………………………………

……………………………………………………………………………………………

3- Clarity

The ontology of thin wall casting communicates effectively the intended meaning of thin wall casting.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more clear:

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

4- Cohesion

The elements of the ontology of 'thin wall casting' are strongly related.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more cohesive:
……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………

5- ***Completeness***

The elements of the ontology of thin wall casting completely cover all aspects of thin wall casting.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more completed:
……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………

6- ***Conciseness***

The elements of the ontology of thin wall casting don't reflect any irrelevant or redundant aspects of thin wall casting.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write you're your suggestions to make the ontology more concise:
……………………………………………………………………………………………
……………………………………………………………………………………………
……………………………………………………………………………………………

7- ***Consistency***

The ontology of thin wall casting doesn't include or allow any contradictions.

Strongly Disagree:□ Disagree:□ Neutral:□ Agree: □ Strongly Agree:□

Please write your suggestions to make the ontology more consistent:
…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

Appendix B: Stop Words

A	but	from	keeps	often	seen	thru	you're
able	by	further	kept	oh	self	thus	yours
about	c	furthermore	kg	ok	selves	to	yourself
above	ca	g	km	okay	sensible	together	yourselves
abst	came	gave	know	old	sent	too	you've
accordance	can	get	known	omitted	serious	took	z
according	cannot	gets	knows	on	seriously	toward	zero
accordingly	cant	getting	l	once	seven	towards	january
across	can't	give	largely	one	several	tried	february
act	cause	given	last	ones	shall	tries	march
actually	causes	gives	lately	only	she	truly	april
added	certain	giving	later	onto	shed	try	may
adj	certainly	go	latter	or	she'll	trying	june
affected	changes	goes	latterly	ord	shes	t's	july
affecting	clearly	going	least	other	should	twice	august
affects	c'mon	gone	less	others	shouldn't	two	september
after	co	got	lest	otherwise	show	u	october
afterwards	com	gotten	let	ought	showed	un	november
again	come	greetings	lets	our	shown	under	december
against	comes	h	let's	ours	showns	unfortunately	xii
ah	concerning	had	like	ourselves	shows	unless	xiii
ain't	consequently	hadn't	liked	out	significant	unlikely	xiv
all	consider	happens	likely	outside	significantly	until	xix
allow	considering	hardly	line	over	similar	unto	xxiv
allows	contain	has	little	overall	similarly	up	vii
almost	containing	hasn't	'll	owing	since	upon	viii
alone	contains	have	look	own	six	us	HR
along	corresponding	haven't	looking	p	slightly	use	BR
already	could	having	looks	page	so	used	
also	couldnt	he	ltd	pages	some	useful	
although	couldn't	hed	m	part	somebody	uses	
always	course	hello	made	particular	somehow	using	
am	c's	help	mainly	particularly	someone	usually	

among	currently	hence	make	past	somethan	v	
amongst	d	her	makes	per	something	value	
an	date	here	many	perhaps	sometime	various	
and	definitely	hereafter	may	placed	sometimes	very	
announce	described	hereby	maybe	please	somewhat	via	
another	despite	herein	me	plus	somewhere	viz	
any	did	heres	mean	poorly	soon	vs	
anybody	didn't	here's	means	possible	sorry	w	
anyhow	different	hereupon	meantime	possibly	specifically	want	
anymore	do	hers	meanwhile	potentially	specified	wants	
anyone	does	herself	merely	pp	specify	was	
anything	doesn't	hes	mg	predominantly	specifying	wasn't	
anyway	doing	he's	might	present	still	way	
anyways	done	hi	million	presumably	stop	we	
anywhere	don't	hid	miss	previously	strongly	we'd	
apart	down	him	ml	primarily	sub	welcome	
apparently	downwards	himself	more	probably	substantially	well	
appear	due	his	moreover	promptly	successfully	we'll	
appreciate	during	hither	most	proud	such	went	
appropriate	e	home	mostly	provides	sufficiently	were	
approximately	each	hopefully	mr	put	suggest	we're	
are	ed	how	mrs	q	sup	weren't	
aren	edu	howbeit	much	que	sure	we've	
arent	effect	however	mug	quickly	t	what	
aren't	eg	hundred	must	quite	take	whatever	
arise	eight	i	my	qv	taken	what's	
around	eighty	id	myself	r	tell	when	
as	either	i'd	n	ran	tends	whence	
a's	else	ie	na	rather	th	whenever	
aside	elsewhere	if	name	rd	than	where	
ask	end	ignored	namely	re	thank	whereafter	
asking	ending	i'll	nay	readily	thanks	whereas	
associated	enough	im	nd	really	thanx	whereby	
at	entirely	i'm	near	reasonably	that	wherein	

auth	especially	immediate	nearly	recent	thats	where's	
available	et	immediately	necessarily	recently	that's	whereupon	
away	et-al	importance	necessary	ref	the	wherever	
awfully	etc	important	need	refs	their	whether	
B	even	in	needs	regarding	theirs	which	
back	evenly	inasmuch	neither	regardless	them	while	
be	ever	inc	never	regards	themselves	whither	
became	every	indeed	nevertheless	related	then	who	
because	everybody	index	new	relatively	thence	whoever	
become	everyone	indicate	next	research	there	whole	
becomes	everything	indicated	nine	respectively	thereafter	whom	
becoming	everywhere	indicates	ninety	resulted	thereby	who's	
been	ex	information	no	resulting	therefore	whose	
before	exactly	inner	nobody	results	therein	why	
beforehand	example	insofar	non	right	theres	will	
begin	except	instead	none	run	there's	willing	
beginning	f	into	nonetheless	s	thereupon	wish	
beginnings	far	invention	noone	said	these	with	
begins	few	inward	nor	same	they	within	
behind	ff	is	normally	saw	they'd	without	
being	fifth	isn't	nos	say	they'll	wonder	
believe	first	it	not	saying	they're	won't	
below	five	itd	noted	says	they've	would	
beside	fix	it'd	nothing	sec	think	wouldn't	
besides	followed	it'll	novel	second	third	www	
best	following	its	now	secondly	this	x	
better	follows	it's	nowhere	section	thorough	y	
between	for	itself	o	see	thoroughly	yes	
beyond	former	i've	obtain	seeing	those	yet	
biol	formerly	j	obtained	seem	though	you	
both	forth	just	obviously	seemed	three	you'd	
brief	found	k	of	seeming	through	you'll	
briefly	four	keep	off	seems	throughout	your	

Table 47- Stop Words

Appendix C: SVM Performance in RC, OC, and OSC for the Market Need

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.1	0.30	0.96	0.63	0.81	0.40	0.32
0.001	0.2	0.30	0.96	0.63	0.82	0.40	0.32
0.001	0.3	0.30	0.95	0.62	0.81	0.40	0.32
0.001	0.4	0.31	0.96	0.63	0.82	0.40	0.33
0.001	0.5	0.31	0.96	0.63	0.82	0.40	0.33
0.001	0.6	0.29	0.96	0.63	0.82	0.40	0.32
0.001	0.7	0.30	0.96	0.63	0.81	0.40	0.32
0.001	0.8	0.30	0.96	0.63	0.82	0.40	0.32
0.001	0.9	0.30	0.96	0.63	0.82	0.40	0.32
0.001	1	0.30	0.96	0.63	0.82	0.40	0.33
0.01	0.1	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.2	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.3	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.4	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.5	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.6	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.7	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.8	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.9	0.00	1.00	0.50	0.77	0.40	0.00
0.01	1	0.00	1.00	0.50	0.77	0.40	0.00

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.1	0.1	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.2	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.3	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.4	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.5	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.6	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.7	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.8	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.9	0.20	0.80	0.50	0.65	0.40	0.06
0.1	1	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-05	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.2	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.3	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.8	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.9	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-09	0.1	0.21	0.80	0.50	0.65	0.40	0.07

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1.00E-09	0.2	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.3	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.4	0.21	0.80	0.50	0.65	0.40	0.07
1.00E-09	0.5	0.21	0.80	0.50	0.65	0.40	0.07
1.00E-09	0.6	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.7	0.21	0.80	0.50	0.65	0.40	0.07
1.00E-09	0.8	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.9	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	1	0.20	0.80	0.50	0.65	0.40	0.06
10	0.1	0.47	0.95	0.71	0.84	0.77	0.53
10	0.2	0.49	0.95	0.72	0.84	0.78	0.54
10	0.3	0.49	0.95	0.72	0.85	0.78	0.54
10	0.4	0.48	0.95	0.71	0.84	0.80	0.54
10	0.5	0.48	0.95	0.72	0.84	0.77	0.53
10	0.6	0.48	0.95	0.71	0.84	0.76	0.54
10	0.7	0.48	0.95	0.71	0.84	0.75	0.53
10	0.8	0.47	0.95	0.71	0.84	0.77	0.53
10	0.9	0.48	0.95	0.71	0.84	0.76	0.53
10	1	0.47	0.95	0.71	0.84	0.77	0.53
100	0.1	0.49	0.95	0.72	0.84	0.77	0.53
100	0.2	0.49	0.95	0.72	0.84	0.76	0.53

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
100	0.3	0.47	0.95	0.71	0.84	0.78	0.53
100	0.4	0.48	0.95	0.71	0.84	0.77	0.53
100	0.5	0.48	0.95	0.72	0.84	0.78	0.54
100	0.6	0.47	0.95	0.71	0.84	0.72	0.52
100	0.7	0.47	0.95	0.71	0.84	0.77	0.53
100	0.8	0.48	0.95	0.71	0.84	0.75	0.53
100	0.9	0.49	0.95	0.72	0.84	0.74	0.54
100	1	0.48	0.95	0.72	0.85	0.78	0.55
20	0.1	0.48	0.95	0.72	0.84	0.78	0.54
20	0.2	0.47	0.95	0.71	0.84	0.78	0.53
20	0.3	0.48	0.95	0.71	0.84	0.74	0.53
20	0.4	0.49	0.95	0.72	0.85	0.81	0.55
20	0.5	0.49	0.95	0.72	0.85	0.75	0.55
20	0.6	0.49	0.95	0.72	0.85	0.77	0.55
20	0.7	0.47	0.95	0.71	0.84	0.78	0.53
20	0.8	0.48	0.95	0.72	0.84	0.77	0.54
20	0.9	0.47	0.95	0.71	0.84	0.74	0.53
20	1	0.48	0.95	0.71	0.84	0.77	0.53
50	0.1	0.48	0.95	0.71	0.84	0.76	0.53
50	0.2	0.47	0.95	0.71	0.84	0.77	0.53
50	0.3	0.47	0.95	0.71	0.84	0.72	0.52

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
50	0.4	0.48	0.95	0.71	0.84	0.72	0.53
50	0.5	0.49	0.95	0.72	0.85	0.79	0.54
50	0.6	0.47	0.95	0.71	0.84	0.76	0.53
50	0.7	0.49	0.95	0.72	0.85	0.77	0.54
50	0.8	0.48	0.94	0.71	0.84	0.72	0.52
50	0.9	0.48	0.95	0.71	0.84	0.74	0.53
50	1	0.49	0.95	0.72	0.84	0.79	0.53
1.00E-07	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.2	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.3	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.8	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.9	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	1	0.00	1.00	0.50	0.77	0.40	0.00

Table 48- SVM classifier performance: regular classification for the product features (kernel: linear)

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
0.001	0.1	0.28	0.96	0.62	0.81	0.40	0.31
0.001	0.2	0.29	0.96	0.62	0.81	0.40	0.32
0.001	0.3	0.28	0.96	0.62	0.82	0.40	0.32
0.001	0.4	0.28	0.96	0.62	0.81	0.40	0.31
0.001	0.5	0.27	0.96	0.62	0.81	0.40	0.30
0.001	0.6	0.25	0.96	0.61	0.81	0.40	0.29
0.001	0.7	0.24	0.97	0.60	0.81	0.40	0.28
0.001	0.8	0.23	0.97	0.60	0.81	0.40	0.26
0.001	0.9	0.22	0.97	0.60	0.81	0.40	0.25
0.001	1	0.20	0.97	0.59	0.81	0.40	0.24
0.01	0.1	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.2	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.3	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.4	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.5	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.6	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.7	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.8	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.9	0.00	1.00	0.50	0.77	0.40	0.00
0.01	1	0.00	1.00	0.50	0.77	0.40	0.00
0.1	0.1	0.20	0.80	0.50	0.65	0.40	0.06

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
0.1	0.2	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.3	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.4	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.5	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.6	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.7	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.8	0.20	0.80	0.50	0.65	0.40	0.06
0.1	0.9	0.20	0.80	0.50	0.65	0.40	0.06
0.1	1	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-05	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.2	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.3	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-	0.8	0.00	1.00	0.50	0.77	0.40	0.00

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
05							
1.00E-05	0.9	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-05	1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-09	0.1	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.2	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.3	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.4	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.5	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.6	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.7	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.8	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	0.9	0.20	0.80	0.50	0.65	0.40	0.06
1.00E-09	1	0.20	0.80	0.50	0.65	0.40	0.06
10	0.1	0.66	0.91	0.78	0.85	0.77	0.64

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
10	0.2	0.45	0.93	0.69	0.82	0.72	0.48
10	0.3	0.42	0.94	0.68	0.82	0.71	0.46
10	0.4	0.40	0.95	0.67	0.83	0.68	0.46
10	0.5	0.39	0.95	0.67	0.83	0.67	0.45
10	0.6	0.39	0.95	0.67	0.83	0.69	0.45
10	0.7	0.38	0.96	0.67	0.83	0.71	0.46
10	0.8	0.37	0.96	0.67	0.83	0.68	0.45
10	0.9	0.38	0.97	0.67	0.83	0.73	0.46
10	1	0.38	0.97	0.68	0.84	0.74	0.46
100	0.1	0.43	0.95	0.69	0.83	0.76	0.50
100	0.2	0.43	0.94	0.68	0.83	0.71	0.48
100	0.3	0.40	0.95	0.67	0.83	0.66	0.46
100	0.4	0.41	0.95	0.68	0.83	0.74	0.47
100	0.5	0.40	0.95	0.67	0.83	0.70	0.45
100	0.6	0.40	0.95	0.67	0.83	0.69	0.46
100	0.7	0.37	0.96	0.67	0.83	0.67	0.45
100	0.8	0.38	0.96	0.67	0.83	0.73	0.45
100	0.9	0.38	0.97	0.67	0.83	0.68	0.46
100	1	0.38	0.97	0.67	0.83	0.71	0.46
20	0.1	0.46	0.93	0.69	0.82	0.72	0.48
20	0.2	0.42	0.95	0.69	0.83	0.74	0.49

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
20	0.3	0.41	0.95	0.68	0.83	0.73	0.47
20	0.4	0.41	0.95	0.68	0.83	0.73	0.46
20	0.5	0.40	0.96	0.68	0.83	0.66	0.46
20	0.6	0.39	0.95	0.67	0.83	0.70	0.46
20	0.7	0.38	0.96	0.67	0.83	0.72	0.45
20	0.8	0.39	0.96	0.67	0.83	0.73	0.45
20	0.9	0.38	0.97	0.67	0.83	0.69	0.46
20	1	0.38	0.97	0.68	0.84	0.69	0.46
50	0.1	0.42	0.95	0.69	0.83	0.68	0.49
50	0.2	0.42	0.95	0.68	0.83	0.76	0.48
50	0.3	0.41	0.95	0.68	0.83	0.71	0.46
50	0.4	0.40	0.95	0.67	0.82	0.68	0.46
50	0.5	0.41	0.95	0.68	0.82	0.73	0.47
50	0.6	0.40	0.96	0.68	0.83	0.72	0.46
50	0.7	0.39	0.96	0.67	0.83	0.73	0.46
50	0.8	0.39	0.97	0.68	0.83	0.77	0.47
50	0.9	0.38	0.97	0.67	0.84	0.70	0.46
50	1	0.38	0.97	0.67	0.83	0.70	0.45
1.00E-07	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.2	0.00	1.00	0.50	0.77	0.40	0.00

C	Gamma	Average of Sensitivity	Average of Specificity	Average of ROC_AUC	Average of Accuracy	Average of Precision	Average of F-score
1.00E-07	0.3	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.8	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	0.9	0.00	1.00	0.50	0.77	0.40	0.00
1.00E-07	1	0.00	1.00	0.50	0.77	0.40	0.00

Table 49- SVM classifier performance: regular classification for the product features (kernel: rbf)

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.1	0.38	0.9	0.64	0.80	0.40	0.35
0.001	0.2	0.40	0.89	0.65	0.79	0.40	0.36
0.001	0.3	0.38	0.89	0.63	0.79	0.40	0.34
0.001	0.4	0.39	0.89	0.64	0.79	0.40	0.35
0.001	0.5	0.39	0.9	0.64	0.79	0.40	0.35
0.001	0.6	0.37	0.89	0.63	0.79	0.40	0.33
0.001	0.7	0.40	0.89	0.65	0.79	0.40	0.36
0.001	0.8	0.39	0.9	0.64	0.79	0.40	0.35
0.001	0.9	0.38	0.9	0.64	0.79	0.40	0.35
0.001	1	0.39	0.89	0.64	0.79	0.40	0.35
0.01	0.1	0.00	1	0.50	0.77	0.40	0.00
0.01	0.2	0.00	1	0.50	0.77	0.40	0.00
0.01	0.3	0.00	1	0.50	0.77	0.40	0.00
0.01	0.4	0.00	1	0.50	0.77	0.40	0.00
0.01	0.5	0.00	1	0.50	0.77	0.40	0.00
0.01	0.6	0.00	1	0.50	0.77	0.40	0.00
0.01	0.7	0.00	1	0.50	0.77	0.40	0.00
0.01	0.8	0.00	1	0.50	0.77	0.40	0.00
0.01	0.9	0.00	1	0.50	0.77	0.40	0.00
0.01	1	0.00	1	0.50	0.77	0.40	0.00
0.1	0.1	0.64	0.85	0.75	0.80	0.81	0.59
0.1	0.2	0.65	0.86	0.76	0.81	0.80	0.60
0.1	0.3	0.66	0.86	0.76	0.81	0.80	0.60
0.1	0.4	0.65	0.86	0.76	0.81	0.77	0.60
0.1	0.5	0.65	0.85	0.75	0.80	0.82	0.59
0.1	0.6	0.64	0.86	0.75	0.81	0.83	0.60
0.1	0.7	0.65	0.86	0.76	0.81	0.76	0.60
0.1	0.8	0.65	0.85	0.75	0.80	0.80	0.60
0.1	0.9	0.65	0.87	0.76	0.81	0.81	0.60
0.1	1	0.65	0.85	0.75	0.80	0.82	0.59
0.00001	0.1	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.2	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.3	0.00	1	0.50	0.77	0.40	0.00

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.00001	0.4	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.5	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.6	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.7	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.8	0.00	1	0.50	0.77	0.40	0.00
0.00001	0.9	0.00	1	0.50	0.77	0.40	0.00
0.00001	1	0.00	1	0.50	0.77	0.40	0.00
1.E-09	0.1	0.49	0.72	0.60	0.66	0.40	0.32
1.E-09	0.2	0.50	0.72	0.61	0.66	0.40	0.32
1.E-09	0.3	0.49	0.72	0.60	0.66	0.40	0.32
1.E-09	0.4	0.50	0.71	0.61	0.66	0.40	0.33
1.E-09	0.5	0.47	0.73	0.60	0.66	0.40	0.31
1.E-09	0.6	0.49	0.72	0.60	0.66	0.40	0.31
1.E-09	0.7	0.48	0.72	0.60	0.66	0.40	0.31
1.E-09	0.8	0.48	0.72	0.60	0.66	0.40	0.31
1.E-09	0.9	0.49	0.71	0.60	0.66	0.40	0.32
1.E-09	1	0.48	0.72	0.60	0.66	0.40	0.31
10	0.1	0.86	0.78	0.82	0.80	0.75	0.65
10	0.2	0.86	0.78	0.82	0.80	0.72	0.65
10	0.3	0.86	0.78	0.82	0.80	0.76	0.65
10	0.4	0.85	0.78	0.82	0.80	0.74	0.64
10	0.5	0.86	0.78	0.82	0.80	0.74	0.65
10	0.6	0.86	0.78	0.82	0.80	0.73	0.65
10	0.7	0.86	0.78	0.82	0.80	0.73	0.65
10	0.8	0.86	0.78	0.82	0.80	0.74	0.65
10	0.9	0.85	0.78	0.82	0.80	0.74	0.65
10	1	0.86	0.78	0.82	0.80	0.75	0.65
100	0.1	0.91	0.78	0.84	0.81	0.76	0.69
100	0.2	0.91	0.78	0.84	0.81	0.74	0.68
100	0.3	0.90	0.78	0.84	0.81	0.76	0.68
100	0.4	0.91	0.78	0.85	0.81	0.76	0.69
100	0.5	0.92	0.78	0.85	0.81	0.75	0.69
100	0.6	0.91	0.78	0.84	0.81	0.76	0.68

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
100	0.7	0.90	0.78	0.84	0.81	0.76	0.68
100	0.8	0.91	0.78	0.85	0.81	0.74	0.69
100	0.9	0.90	0.78	0.84	0.80	0.74	0.68
100	1	0.90	0.78	0.84	0.81	0.76	0.68
20	0.1	0.89	0.78	0.83	0.80	0.76	0.67
20	0.2	0.87	0.78	0.82	0.80	0.75	0.66
20	0.3	0.88	0.78	0.83	0.80	0.74	0.67
20	0.4	0.87	0.78	0.82	0.80	0.74	0.66
20	0.5	0.89	0.78	0.84	0.80	0.76	0.68
20	0.6	0.87	0.78	0.82	0.80	0.73	0.66
20	0.7	0.88	0.78	0.83	0.80	0.76	0.66
20	0.8	0.89	0.78	0.83	0.80	0.75	0.67
20	0.9	0.88	0.78	0.83	0.81	0.76	0.67
20	1	0.87	0.78	0.83	0.80	0.73	0.66
50	0.1	0.91	0.77	0.84	0.80	0.75	0.68
50	0.2	0.89	0.78	0.83	0.80	0.76	0.67
50	0.3	0.91	0.78	0.84	0.80	0.75	0.68
50	0.4	0.90	0.78	0.84	0.80	0.75	0.68
50	0.5	0.89	0.78	0.84	0.80	0.74	0.67
50	0.6	0.91	0.77	0.84	0.80	0.74	0.68
50	0.7	0.89	0.77	0.83	0.80	0.74	0.67
50	0.8	0.90	0.78	0.84	0.80	0.75	0.68
50	0.9	0.89	0.78	0.84	0.80	0.75	0.68
50	1	0.89	0.78	0.84	0.80	0.74	0.67
1.E-07	0.1	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.2	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.3	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.4	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.5	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.6	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.7	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.8	0.00	1	0.50	0.77	0.40	0.00
1.E-07	0.9	0.00	1	0.50	0.77	0.40	0.00

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1.E-07	1	0.00	1	0.50	0.77	0.40	0.00

Table 50 - SVM classifier performance in ontological classification for the product features (kernel: linear)

C	Gamma	Average of Specificity	Average of Sensitivity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
0.001	0.1	0.90	0.37	0.64	0.79	0.40	0.34
0.001	0.2	0.90	0.39	0.64	0.80	0.40	0.35
0.001	0.3	0.90	0.38	0.64	0.79	0.40	0.34
0.001	0.4	0.90	0.39	0.64	0.80	0.40	0.36
0.001	0.5	0.90	0.40	0.65	0.80	0.40	0.36
0.001	0.6	0.90	0.39	0.65	0.80	0.40	0.36
0.001	0.7	0.90	0.40	0.65	0.80	0.40	0.36
0.001	0.8	0.89	0.41	0.65	0.80	0.40	0.36
0.001	0.9	0.88	0.42	0.65	0.79	0.40	0.35
0.001	1	0.87	0.44	0.66	0.79	0.40	0.36
0.01	0.1	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.2	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.3	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.4	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.5	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.6	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.7	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.8	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.9	1.00	0.00	0.50	0.77	0.40	0.00
0.01	1	1.00	0.00	0.50	0.77	0.40	0.00
0.1	0.1	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.2	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.3	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.4	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.5	0.81	0.22	0.52	0.66	0.43	0.10
0.1	0.6	0.87	0.34	0.60	0.73	0.67	0.28
0.1	0.7	0.90	0.48	0.69	0.80	0.77	0.47
0.1	0.8	0.91	0.57	0.74	0.83	0.86	0.57
0.1	0.9	0.91	0.63	0.77	0.84	0.86	0.62
0.1	1	0.90	0.66	0.78	0.84	0.87	0.64
0.00001	0.1	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.2	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.3	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.4	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.5	1.00	0.00	0.50	0.77	0.40	0.00

C	Gamma	Average of Specificity	Average of Sensitivity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
0.00001	0.6	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.7	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.8	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	0.9	1.00	0.00	0.50	0.77	0.40	0.00
0.00001	1	1.00	0.00	0.50	0.77	0.40	0.00
1.E-09	0.1	0.71	0.52	0.61	0.66	0.40	0.34
1.E-09	0.2	0.72	0.51	0.61	0.66	0.40	0.34
1.E-09	0.3	0.72	0.51	0.61	0.66	0.40	0.33
1.E-09	0.4	0.72	0.51	0.61	0.66	0.40	0.33
1.E-09	0.5	0.73	0.49	0.61	0.66	0.40	0.32
1.E-09	0.6	0.73	0.49	0.61	0.66	0.40	0.32
1.E-09	0.7	0.72	0.52	0.62	0.66	0.40	0.34
1.E-09	0.8	0.72	0.50	0.61	0.66	0.40	0.31
1.E-09	0.9	0.70	0.55	0.63	0.66	0.40	0.36
1.E-09	1	0.70	0.55	0.62	0.66	0.40	0.35
10	0.1	0.79	0.84	0.81	0.80	0.73	0.64
10	0.2	0.79	0.86	0.83	0.81	0.77	0.66
10	0.3	0.79	0.86	0.83	0.81	0.75	0.66
10	0.4	0.79	0.89	0.84	0.82	0.77	0.68
10	0.5	0.79	0.89	0.84	0.81	0.74	0.68
10	0.6	0.81	0.88	0.85	0.83	0.77	0.69
10	0.7	0.81	0.89	0.85	0.83	0.79	0.70
10	0.8	0.81	0.89	0.85	0.83	0.79	0.71
10	0.9	0.81	0.90	0.86	0.83	0.79	0.71
10	1	0.81	0.91	0.86	0.83	0.79	0.72
100	0.1	0.79	0.89	0.84	0.81	0.74	0.69
100	0.2	0.81	0.90	0.85	0.83	0.78	0.71
100	0.3	0.81	0.91	0.86	0.83	0.76	0.72
100	0.4	0.81	0.91	0.86	0.83	0.78	0.71
100	0.5	0.81	0.91	0.86	0.83	0.79	0.72
100	0.6	0.81	0.92	0.86	0.83	0.81	0.72
100	0.7	0.81	0.92	0.86	0.83	0.77	0.72
100	0.8	0.81	0.91	0.86	0.84	0.78	0.72
100	0.9	0.81	0.92	0.87	0.84	0.78	0.73
100	1	0.81	0.91	0.86	0.83	0.78	0.72

C	Gamma	Average of Specificity	Average of Sensitivity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
20	0.1	0.79	0.86	0.83	0.81	0.76	0.67
20	0.2	0.79	0.85	0.82	0.81	0.72	0.66
20	0.3	0.79	0.89	0.84	0.81	0.78	0.68
20	0.4	0.81	0.89	0.85	0.82	0.77	0.69
20	0.5	0.81	0.89	0.85	0.83	0.77	0.70
20	0.6	0.81	0.91	0.86	0.83	0.79	0.72
20	0.7	0.81	0.91	0.86	0.83	0.78	0.72
20	0.8	0.81	0.90	0.86	0.83	0.78	0.71
20	0.9	0.81	0.91	0.86	0.84	0.79	0.72
20	1	0.81	0.91	0.86	0.83	0.80	0.72
50	0.1	0.80	0.87	0.84	0.81	0.75	0.68
50	0.2	0.80	0.89	0.84	0.82	0.76	0.68
50	0.3	0.81	0.89	0.85	0.83	0.78	0.70
50	0.4	0.81	0.90	0.85	0.83	0.78	0.71
50	0.5	0.81	0.91	0.86	0.83	0.78	0.72
50	0.6	0.81	0.91	0.86	0.83	0.78	0.72
50	0.7	0.81	0.92	0.86	0.83	0.78	0.72
50	0.8	0.81	0.92	0.86	0.83	0.80	0.73
50	0.9	0.81	0.92	0.86	0.84	0.78	0.73
50	1	0.81	0.91	0.86	0.84	0.80	0.72
1.E-07	0.1	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.2	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.3	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.4	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.5	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.6	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.7	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.8	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	0.9	1.00	0.00	0.50	0.77	0.40	0.00
1.E-07	1	1.00	0.00	0.50	0.77	0.40	0.00

Table 51- SVM classifier performance: ontological classification for the product features (kernel: RFB)

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
0.001	0.1	0.38	0.93	0.66	0.82	0.40	0.37
0.001	0.2	0.37	0.93	0.65	0.82	0.40	0.37
0.001	0.3	0.38	0.94	0.66	0.82	0.40	0.37
0.001	0.4	0.36	0.93	0.65	0.81	0.40	0.36
0.001	0.5	0.37	0.93	0.65	0.81	0.40	0.37
0.001	0.6	0.37	0.94	0.65	0.82	0.40	0.37
0.001	0.7	0.38	0.93	0.65	0.82	0.40	0.37
0.001	0.8	0.38	0.93	0.66	0.82	0.40	0.37
0.001	0.9	0.37	0.94	0.65	0.82	0.40	0.37
0.001	1	0.37	0.93	0.65	0.82	0.40	0.37
0.01	0.1	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.2	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.3	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.4	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.5	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.6	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.7	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.8	0.00	1.00	0.50	0.77	0.40	0.00
0.01	0.9	0.00	1.00	0.50	0.77	0.40	0.00
0.01	1	0.00	1.00	0.50	0.77	0.40	0.00
0.1	0.1	0.64	0.91	0.78	0.85	0.80	0.63
0.1	0.2	0.64	0.91	0.78	0.85	0.82	0.64
0.1	0.3	0.64	0.90	0.77	0.84	0.81	0.62
0.1	0.4	0.63	0.91	0.77	0.84	0.81	0.62
0.1	0.5	0.64	0.91	0.77	0.85	0.80	0.63
0.1	0.6	0.64	0.91	0.77	0.85	0.79	0.63
0.1	0.7	0.64	0.91	0.77	0.84	0.80	0.63
0.1	0.8	0.64	0.91	0.78	0.85	0.81	0.63
0.1	0.9	0.65	0.91	0.78	0.85	0.83	0.64
0.1	1	0.63	0.91	0.77	0.84	0.83	0.62
1E-05	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.2	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.3	0.00	1.00	0.50	0.77	0.40	0.00

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
1E-05	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.8	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	0.9	0.00	1.00	0.50	0.77	0.40	0.00
1E-05	1	0.00	1.00	0.50	0.77	0.40	0.00
1E-09	0.1	0.37	0.77	0.57	0.67	0.40	0.23
1E-09	0.2	0.35	0.77	0.56	0.66	0.40	0.20
1E-09	0.3	0.37	0.77	0.57	0.67	0.40	0.23
1E-09	0.4	0.37	0.77	0.57	0.67	0.40	0.23
1E-09	0.5	0.38	0.77	0.57	0.67	0.40	0.23
1E-09	0.6	0.35	0.77	0.56	0.66	0.40	0.21
1E-09	0.7	0.36	0.77	0.57	0.67	0.40	0.22
1E-09	0.8	0.35	0.77	0.56	0.66	0.40	0.21
1E-09	0.9	0.38	0.77	0.58	0.67	0.40	0.24
1E-09	1	0.35	0.77	0.56	0.66	0.40	0.21
10	0.1	0.84	0.89	0.86	0.88	0.83	0.75
10	0.2	0.83	0.89	0.86	0.87	0.83	0.75
10	0.3	0.82	0.89	0.86	0.87	0.84	0.74
10	0.4	0.83	0.89	0.86	0.88	0.84	0.75
10	0.5	0.82	0.89	0.85	0.87	0.83	0.74
10	0.6	0.83	0.89	0.86	0.87	0.84	0.74
10	0.7	0.83	0.89	0.86	0.87	0.82	0.75
10	0.8	0.83	0.89	0.86	0.87	0.86	0.75
10	0.9	0.83	0.89	0.86	0.87	0.85	0.75
10	1	0.83	0.88	0.86	0.87	0.81	0.74
100	0.1	0.81	0.89	0.85	0.88	0.86	0.75
100	0.2	0.80	0.89	0.85	0.87	0.82	0.73
100	0.3	0.80	0.89	0.85	0.87	0.83	0.73
100	0.4	0.80	0.89	0.85	0.87	0.83	0.73
100	0.5	0.80	0.89	0.85	0.87	0.84	0.74
100	0.6	0.81	0.89	0.85	0.87	0.81	0.74

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
100	0.7	0.82	0.89	0.86	0.88	0.85	0.75
100	0.8	0.80	0.89	0.85	0.87	0.82	0.73
100	0.9	0.82	0.89	0.86	0.88	0.86	0.75
100	1	0.81	0.90	0.85	0.88	0.82	0.74
20	0.1	0.81	0.89	0.85	0.87	0.82	0.74
20	0.2	0.81	0.89	0.85	0.88	0.82	0.74
20	0.3	0.80	0.89	0.85	0.87	0.82	0.73
20	0.4	0.81	0.89	0.85	0.87	0.82	0.73
20	0.5	0.81	0.89	0.85	0.87	0.87	0.74
20	0.6	0.82	0.89	0.85	0.87	0.84	0.74
20	0.7	0.80	0.89	0.84	0.87	0.83	0.73
20	0.8	0.81	0.89	0.85	0.87	0.84	0.74
20	0.9	0.81	0.89	0.85	0.87	0.81	0.73
20	1	0.81	0.89	0.85	0.87	0.82	0.73
50	0.1	0.79	0.89	0.84	0.87	0.80	0.73
50	0.2	0.79	0.89	0.84	0.87	0.82	0.72
50	0.3	0.78	0.89	0.84	0.87	0.80	0.72
50	0.4	0.79	0.89	0.84	0.87	0.83	0.72
50	0.5	0.80	0.89	0.85	0.87	0.83	0.73
50	0.6	0.80	0.89	0.85	0.87	0.81	0.73
50	0.7	0.79	0.89	0.84	0.87	0.82	0.73
50	0.8	0.80	0.90	0.85	0.87	0.83	0.73
50	0.9	0.80	0.89	0.85	0.87	0.83	0.74
50	1	0.78	0.89	0.84	0.87	0.83	0.73
1E-07	0.1	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.2	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.3	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.4	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.5	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.6	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.7	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.8	0.00	1.00	0.50	0.77	0.40	0.00
1E-07	0.9	0.00	1.00	0.50	0.77	0.40	0.00

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
1E-07	1	0.00	1.00	0.50	0.77	0.40	0.00

Table 52- SVM classifier performance in ontological semantic classification for the product features (kernel: linear)

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
0.001	0.1	0.93	0.38	0.65	0.82	0.40	0.37
0.001	0.2	0.94	0.37	0.66	0.82	0.40	0.38
0.001	0.3	0.95	0.37	0.66	0.83	0.40	0.38
0.001	0.4	0.95	0.37	0.66	0.83	0.40	0.38
0.001	0.5	0.95	0.36	0.66	0.83	0.40	0.38
0.001	0.6	0.96	0.37	0.66	0.83	0.40	0.39
0.001	0.7	0.96	0.36	0.66	0.83	0.40	0.38
0.001	0.8	0.96	0.37	0.67	0.84	0.40	0.39
0.001	0.9	0.96	0.36	0.66	0.83	0.40	0.39
0.001	1	0.97	0.37	0.67	0.84	0.40	0.40
0.01	0.1	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.2	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.3	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.4	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.5	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.6	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.7	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.8	1.00	0.00	0.50	0.77	0.40	0.00
0.01	0.9	1.00	0.00	0.50	0.77	0.40	0.00
0.01	1	1.00	0.00	0.50	0.77	0.40	0.00
0.1	0.1	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.2	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.3	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.4	0.80	0.20	0.50	0.65	0.40	0.06
0.1	0.5	0.80	0.21	0.51	0.65	0.41	0.08
0.1	0.6	0.82	0.27	0.54	0.68	0.55	0.16
0.1	0.7	0.83	0.35	0.59	0.70	0.61	0.28
0.1	0.8	0.86	0.42	0.64	0.75	0.75	0.38
0.1	0.9	0.87	0.47	0.67	0.77	0.78	0.46
0.1	1	0.89	0.48	0.69	0.79	0.84	0.48
1E-05	0.1	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.2	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.3	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.4	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.5	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.6	1.00	0.00	0.50	0.77	0.40	0.00

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
1E-05	0.7	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.8	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	0.9	1.00	0.00	0.50	0.77	0.40	0.00
1E-05	1	1.00	0.00	0.50	0.77	0.40	0.00
1E-09	0.1	0.78	0.34	0.56	0.66	0.40	0.20
1E-09	0.2	0.78	0.31	0.55	0.66	0.40	0.17
1E-09	0.3	0.78	0.34	0.56	0.67	0.40	0.20
1E-09	0.4	0.78	0.32	0.55	0.66	0.40	0.19
1E-09	0.5	0.79	0.31	0.55	0.66	0.40	0.18
1E-09	0.6	0.79	0.32	0.55	0.67	0.40	0.19
1E-09	0.7	0.79	0.29	0.54	0.66	0.40	0.16
1E-09	0.8	0.79	0.30	0.54	0.66	0.40	0.16
1E-09	0.9	0.79	0.28	0.53	0.66	0.40	0.14
1E-09	1	0.79	0.28	0.54	0.66	0.40	0.14
10	0.1	0.87	0.86	0.87	0.87	0.83	0.75
10	0.2	0.89	0.86	0.87	0.88	0.85	0.77
10	0.3	0.89	0.84	0.87	0.88	0.83	0.76
10	0.4	0.91	0.80	0.85	0.88	0.85	0.74
10	0.5	0.92	0.78	0.85	0.89	0.83	0.74
10	0.6	0.93	0.77	0.85	0.90	0.87	0.76
10	0.7	0.94	0.77	0.86	0.90	0.88	0.77
10	0.8	0.95	0.76	0.86	0.91	0.88	0.77
10	0.9	0.96	0.76	0.86	0.91	0.90	0.77
10	1	0.95	0.75	0.85	0.91	0.89	0.77
100	0.1	0.91	0.82	0.87	0.89	0.87	0.77
100	0.2	0.92	0.79	0.85	0.89	0.87	0.76
100	0.3	0.92	0.79	0.86	0.89	0.85	0.75
100	0.4	0.92	0.79	0.86	0.89	0.85	0.76
100	0.5	0.93	0.79	0.86	0.90	0.87	0.77
100	0.6	0.93	0.78	0.85	0.89	0.88	0.76
100	0.7	0.94	0.78	0.86	0.91	0.91	0.78
100	0.8	0.95	0.76	0.85	0.90	0.91	0.76
100	0.9	0.96	0.76	0.86	0.91	0.87	0.77
100	1	0.95	0.75	0.85	0.91	0.87	0.77
20	0.1	0.89	0.85	0.87	0.88	0.84	0.76
20	0.2	0.90	0.84	0.87	0.89	0.87	0.77

SVM Parameters		Ave. of Sensitivity	Ave. of Specificity	Ave. of ROC AUC	Ave. of Accuracy	Ave. of Precision	Ave. of F-score
C	Gamma						
20	0.3	0.91	0.81	0.86	0.89	0.87	0.76
20	0.4	0.92	0.78	0.85	0.89	0.85	0.75
20	0.5	0.93	0.79	0.86	0.90	0.85	0.76
20	0.6	0.93	0.78	0.86	0.90	0.87	0.76
20	0.7	0.95	0.77	0.86	0.91	0.90	0.78
20	0.8	0.94	0.76	0.85	0.90	0.87	0.76
20	0.9	0.95	0.75	0.85	0.91	0.89	0.76
20	1	0.96	0.76	0.86	0.91	0.91	0.78
50	0.1	0.90	0.84	0.87	0.89	0.84	0.77
50	0.2	0.92	0.78	0.85	0.89	0.85	0.74
50	0.3	0.92	0.78	0.85	0.89	0.85	0.75
50	0.4	0.93	0.78	0.85	0.89	0.86	0.75
50	0.5	0.93	0.79	0.86	0.90	0.86	0.77
50	0.6	0.93	0.78	0.86	0.90	0.86	0.76
50	0.7	0.95	0.77	0.86	0.91	0.89	0.78
50	0.8	0.95	0.77	0.86	0.91	0.88	0.77
50	0.9	0.95	0.76	0.86	0.91	0.89	0.77
50	1	0.96	0.76	0.86	0.91	0.86	0.77
1E-07	0.1	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.2	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.3	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.4	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.5	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.6	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.7	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.8	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	0.9	1.00	0.00	0.50	0.77	0.40	0.00
1E-07	1	1.00	0.00	0.50	0.77	0.40	0.00

Table 53- SVM classifier performance in ontological semantic classification for the product features (kernel: RBF)

Appendix D: SVM Performance in RC, OC, and OSC for the Enabling technologies

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.1	0.52	0.49	0.51	0.53	0.51	0.34
0.001	0.2	0.53	0.50	0.51	0.53	0.50	0.35
0.001	0.3	0.52	0.50	0.51	0.53	0.47	0.34
0.001	0.4	0.52	0.50	0.51	0.53	0.49	0.34
0.001	0.5	0.51	0.49	0.50	0.52	0.47	0.33
0.001	0.6	0.52	0.50	0.51	0.53	0.50	0.34
0.001	0.7	0.52	0.50	0.51	0.53	0.48	0.34
0.001	0.8	0.53	0.49	0.51	0.53	0.48	0.35
0.001	0.9	0.52	0.49	0.51	0.53	0.48	0.34
0.001	1	0.53	0.50	0.51	0.53	0.49	0.34
0.01	0.1	0.52	0.50	0.51	0.53	0.49	0.34
0.01	0.2	0.53	0.49	0.51	0.53	0.49	0.35
0.01	0.3	0.52	0.49	0.51	0.53	0.47	0.34
0.01	0.4	0.53	0.49	0.51	0.53	0.49	0.35
0.01	0.5	0.53	0.49	0.51	0.53	0.49	0.35
0.01	0.6	0.52	0.49	0.50	0.52	0.47	0.34
0.01	0.7	0.52	0.50	0.51	0.53	0.49	0.34
0.01	0.8	0.52	0.49	0.51	0.53	0.49	0.34
0.01	0.9	0.52	0.50	0.51	0.53	0.49	0.34
0.01	1	0.52	0.49	0.51	0.53	0.51	0.34
0.1	0.1	0.12	0.90	0.51	0.64	0.49	0.13
0.1	0.2	0.12	0.90	0.51	0.64	0.47	0.13
0.1	0.3	0.11	0.90	0.51	0.64	0.49	0.11
0.1	0.4	0.12	0.90	0.51	0.64	0.50	0.12
0.1	0.5	0.12	0.89	0.51	0.64	0.47	0.13
0.1	0.6	0.13	0.89	0.51	0.63	0.49	0.13
0.1	0.7	0.12	0.90	0.51	0.64	0.49	0.12
0.1	0.8	0.13	0.89	0.51	0.63	0.47	0.13
0.1	0.9	0.13	0.90	0.51	0.64	0.55	0.13
0.1	1	0.12	0.90	0.51	0.64	0.46	0.12
1E-05	0.1	0.12	0.90	0.51	0.64	0.46	0.13
1E-05	0.2	0.13	0.89	0.51	0.64	0.49	0.14
1E-05	0.3	0.12	0.90	0.51	0.64	0.48	0.12

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1E-05	0.4	0.13	0.89	0.51	0.64	0.50	0.14
1E-05	0.5	0.14	0.89	0.51	0.64	0.48	0.14
1E-05	0.6	0.13	0.90	0.51	0.64	0.48	0.14
1E-05	0.7	0.12	0.89	0.51	0.64	0.49	0.13
1E-05	0.8	0.12	0.90	0.51	0.64	0.49	0.13
1E-05	0.9	0.13	0.89	0.51	0.64	0.46	0.13
1E-05	1	0.12	0.90	0.51	0.64	0.45	0.13
1E-09	0.1	0.52	0.50	0.51	0.53	0.46	0.34
1E-09	0.2	0.53	0.49	0.51	0.53	0.47	0.34
1E-09	0.3	0.51	0.50	0.50	0.53	0.47	0.33
1E-09	0.4	0.52	0.50	0.51	0.53	0.51	0.34
1E-09	0.5	0.52	0.49	0.51	0.53	0.49	0.34
1E-09	0.6	0.53	0.50	0.51	0.53	0.49	0.35
1E-09	0.7	0.52	0.49	0.51	0.53	0.47	0.34
1E-09	0.8	0.53	0.50	0.52	0.53	0.49	0.35
1E-09	0.9	0.52	0.49	0.51	0.53	0.48	0.34
1E-09	1	0.53	0.50	0.52	0.53	0.48	0.35
10	0.1	0.15	0.88	0.52	0.64	0.52	0.19
10	0.2	0.16	0.88	0.52	0.64	0.50	0.20
10	0.3	0.15	0.88	0.52	0.64	0.51	0.19
10	0.4	0.15	0.87	0.51	0.64	0.46	0.19
10	0.5	0.15	0.88	0.52	0.64	0.48	0.19
10	0.6	0.15	0.88	0.52	0.65	0.50	0.19
10	0.7	0.16	0.89	0.52	0.65	0.47	0.20
10	0.8	0.14	0.88	0.51	0.64	0.51	0.17
10	0.9	0.16	0.88	0.52	0.65	0.46	0.20
10	1	0.15	0.88	0.51	0.64	0.53	0.18
100	0.1	0.16	0.88	0.52	0.64	0.50	0.20
100	0.2	0.15	0.88	0.52	0.64	0.46	0.19
100	0.3	0.14	0.88	0.51	0.64	0.50	0.18
100	0.4	0.16	0.87	0.51	0.64	0.51	0.19
100	0.5	0.15	0.88	0.52	0.64	0.47	0.19
100	0.6	0.15	0.87	0.51	0.64	0.50	0.19
100	0.7	0.15	0.88	0.51	0.64	0.44	0.19
100	0.8	0.15	0.88	0.51	0.64	0.51	0.18

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
100	0.9	0.16	0.87	0.52	0.64	0.49	0.20
100	1	0.15	0.87	0.51	0.64	0.46	0.19
20	0.1	0.16	0.88	0.52	0.64	0.51	0.19
20	0.2	0.15	0.88	0.51	0.64	0.44	0.18
20	0.3	0.15	0.88	0.52	0.64	0.52	0.19
20	0.4	0.15	0.88	0.51	0.64	0.47	0.19
20	0.5	0.15	0.87	0.51	0.64	0.50	0.18
20	0.6	0.14	0.88	0.51	0.64	0.44	0.18
20	0.7	0.15	0.88	0.52	0.64	0.49	0.19
20	0.8	0.15	0.88	0.52	0.64	0.48	0.18
20	0.9	0.16	0.88	0.52	0.65	0.45	0.20
20	1	0.15	0.88	0.51	0.64	0.51	0.19
50	0.1	0.15	0.87	0.51	0.64	0.47	0.19
50	0.2	0.15	0.88	0.51	0.64	0.49	0.18
50	0.3	0.15	0.88	0.51	0.64	0.52	0.19
50	0.4	0.14	0.88	0.51	0.64	0.46	0.18
50	0.5	0.15	0.88	0.51	0.64	0.52	0.19
50	0.6	0.14	0.87	0.51	0.63	0.49	0.18
50	0.7	0.15	0.88	0.51	0.64	0.47	0.18
50	0.8	0.15	0.87	0.51	0.64	0.46	0.19
50	0.9	0.16	0.88	0.52	0.64	0.46	0.20
50	1	0.14	0.87	0.51	0.63	0.45	0.18
1E-07	0.1	0.52	0.50	0.51	0.53	0.51	0.34
1E-07	0.2	0.52	0.49	0.51	0.53	0.44	0.34
1E-07	0.3	0.52	0.49	0.51	0.53	0.45	0.34
1E-07	0.4	0.52	0.49	0.51	0.53	0.48	0.34
1E-07	0.5	0.53	0.50	0.52	0.54	0.48	0.35
1E-07	0.6	0.52	0.49	0.51	0.53	0.48	0.34
1E-07	0.7	0.53	0.49	0.51	0.53	0.49	0.34
1E-07	0.8	0.53	0.49	0.51	0.53	0.46	0.35
1E-07	0.9	0.52	0.50	0.51	0.53	0.45	0.34
1E-07	1	0.52	0.50	0.51	0.53	0.46	0.34

Table 54- SVM classifier performance: regular classification for the enabling technologies (kernel: linear)

SVM Parameters	Average	Average	Average	Average	Average	Average

C	Gamma	of Sensitivity	of Specificity	of ROC AUC	of Accuracy	of Precision	of F-score
0.001	0.1	0.52	0.50	0.51	0.53	0.50	0.34
0.001	0.2	0.52	0.50	0.51	0.53	0.50	0.34
0.001	0.3	0.51	0.51	0.51	0.53	0.47	0.34
0.001	0.4	0.51	0.51	0.51	0.53	0.51	0.33
0.001	0.5	0.52	0.50	0.51	0.53	0.48	0.34
0.001	0.6	0.50	0.51	0.51	0.53	0.46	0.33
0.001	0.7	0.52	0.50	0.51	0.53	0.49	0.34
0.001	0.8	0.52	0.52	0.52	0.54	0.49	0.35
0.001	0.9	0.50	0.52	0.51	0.54	0.46	0.33
0.001	1	0.52	0.51	0.51	0.54	0.52	0.35
0.01	0.1	0.53	0.49	0.51	0.53	0.51	0.35
0.01	0.2	0.53	0.50	0.52	0.54	0.49	0.35
0.01	0.3	0.52	0.50	0.51	0.53	0.48	0.34
0.01	0.4	0.52	0.51	0.51	0.54	0.51	0.34
0.01	0.5	0.51	0.51	0.51	0.53	0.48	0.33
0.01	0.6	0.52	0.51	0.51	0.53	0.47	0.34
0.01	0.7	0.51	0.51	0.51	0.54	0.49	0.34
0.01	0.8	0.49	0.52	0.51	0.54	0.46	0.32
0.01	0.9	0.50	0.51	0.51	0.54	0.46	0.33
0.01	1	0.50	0.51	0.50	0.53	0.49	0.33
0.1	0.1	0.13	0.90	0.51	0.64	0.48	0.14
0.1	0.2	0.12	0.90	0.51	0.64	0.47	0.13
0.1	0.3	0.11	0.90	0.51	0.64	0.49	0.12
0.1	0.4	0.11	0.90	0.50	0.64	0.49	0.11
0.1	0.5	0.11	0.91	0.51	0.64	0.47	0.12
0.1	0.6	0.11	0.91	0.51	0.65	0.49	0.12
0.1	0.7	0.12	0.92	0.52	0.65	0.52	0.13
0.1	0.8	0.11	0.92	0.51	0.65	0.50	0.12
0.1	0.9	0.12	0.92	0.52	0.65	0.48	0.14
0.1	1	0.11	0.92	0.51	0.65	0.49	0.12
1E-05	0.1	0.12	0.90	0.51	0.64	0.51	0.13
1E-05	0.2	0.11	0.90	0.51	0.64	0.48	0.12
1E-05	0.3	0.12	0.91	0.51	0.64	0.49	0.12
1E-05	0.4	0.11	0.90	0.51	0.64	0.50	0.12
1E-05	0.5	0.12	0.90	0.51	0.64	0.48	0.13
1E-05	0.6	0.11	0.91	0.51	0.64	0.50	0.12

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1E-05	0.7	0.12	0.91	0.51	0.65	0.50	0.14
1E-05	0.8	0.11	0.91	0.51	0.64	0.48	0.12
1E-05	0.9	0.11	0.91	0.51	0.65	0.46	0.12
1E-05	1	0.11	0.92	0.52	0.65	0.52	0.13
1E-09	0.1	0.52	0.50	0.51	0.53	0.49	0.34
1E-09	0.2	0.53	0.50	0.52	0.54	0.53	0.35
1E-09	0.3	0.51	0.50	0.51	0.53	0.46	0.33
1E-09	0.4	0.51	0.51	0.51	0.53	0.48	0.33
1E-09	0.5	0.52	0.51	0.52	0.54	0.50	0.35
1E-09	0.6	0.51	0.50	0.51	0.53	0.48	0.33
1E-09	0.7	0.50	0.51	0.51	0.54	0.49	0.32
1E-09	0.8	0.51	0.51	0.51	0.54	0.49	0.33
1E-09	0.9	0.51	0.52	0.51	0.54	0.51	0.34
1E-09	1	0.50	0.52	0.51	0.54	0.53	0.33
10	0.1	0.15	0.88	0.51	0.64	0.49	0.18
10	0.2	0.13	0.90	0.52	0.65	0.48	0.17
10	0.3	0.10	0.90	0.50	0.64	0.45	0.13
10	0.4	0.10	0.91	0.50	0.64	0.44	0.13
10	0.5	0.09	0.92	0.51	0.65	0.46	0.13
10	0.6	0.07	0.93	0.50	0.65	0.47	0.10
10	0.7	0.06	0.92	0.49	0.64	0.44	0.09
10	0.8	0.04	0.93	0.49	0.64	0.41	0.06
10	0.9	0.03	0.94	0.48	0.64	0.37	0.04
10	1	0.02	0.93	0.48	0.64	0.38	0.03
100	0.1	0.14	0.89	0.52	0.65	0.46	0.18
100	0.2	0.12	0.89	0.51	0.64	0.48	0.16
100	0.3	0.11	0.90	0.51	0.65	0.47	0.15
100	0.4	0.10	0.91	0.50	0.64	0.45	0.13
100	0.5	0.09	0.92	0.51	0.65	0.48	0.13
100	0.6	0.08	0.92	0.50	0.65	0.46	0.11
100	0.7	0.06	0.92	0.49	0.64	0.46	0.09
100	0.8	0.04	0.93	0.49	0.64	0.42	0.06
100	0.9	0.03	0.93	0.48	0.64	0.37	0.04
100	1	0.02	0.94	0.48	0.64	0.38	0.03
20	0.1	0.14	0.88	0.51	0.64	0.46	0.17

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
20	0.2	0.13	0.89	0.51	0.64	0.47	0.16
20	0.3	0.11	0.90	0.51	0.65	0.49	0.15
20	0.4	0.10	0.91	0.51	0.65	0.43	0.14
20	0.5	0.09	0.92	0.50	0.65	0.41	0.12
20	0.6	0.07	0.92	0.50	0.64	0.42	0.10
20	0.7	0.07	0.93	0.50	0.65	0.45	0.09
20	0.8	0.05	0.93	0.49	0.64	0.41	0.06
20	0.9	0.02	0.93	0.48	0.64	0.37	0.03
20	1	0.02	0.93	0.48	0.64	0.37	0.03
50	0.1	0.14	0.89	0.52	0.65	0.48	0.18
50	0.2	0.12	0.90	0.51	0.64	0.48	0.16
50	0.3	0.10	0.90	0.50	0.64	0.47	0.14
50	0.4	0.10	0.91	0.51	0.65	0.44	0.13
50	0.5	0.10	0.92	0.51	0.65	0.48	0.13
50	0.6	0.08	0.92	0.50	0.64	0.43	0.10
50	0.7	0.07	0.92	0.50	0.64	0.44	0.09
50	0.8	0.05	0.93	0.49	0.64	0.38	0.06
50	0.9	0.02	0.93	0.48	0.63	0.39	0.03
50	1	0.02	0.94	0.48	0.64	0.37	0.03
1E-07	0.1	0.52	0.49	0.50	0.52	0.48	0.33
1E-07	0.2	0.54	0.50	0.52	0.53	0.50	0.36
1E-07	0.3	0.52	0.50	0.51	0.53	0.48	0.34
1E-07	0.4	0.53	0.51	0.52	0.54	0.50	0.36
1E-07	0.5	0.51	0.51	0.51	0.53	0.49	0.34
1E-07	0.6	0.52	0.51	0.51	0.54	0.51	0.34
1E-07	0.7	0.51	0.52	0.51	0.54	0.51	0.34
1E-07	0.8	0.50	0.51	0.51	0.54	0.49	0.33
1E-07	0.9	0.51	0.51	0.51	0.54	0.48	0.33

Table 55- SVM classifier performance: regular classification for the enabling technologies (kernel: RBF)

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.1	0.65	0.34	0.49	0.46	0.51	0.40

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.2	0.63	0.34	0.49	0.45	0.44	0.38
0.001	0.3	0.63	0.34	0.49	0.45	0.46	0.39
0.001	0.4	0.64	0.34	0.49	0.46	0.49	0.40
0.001	0.5	0.64	0.34	0.49	0.46	0.46	0.39
0.001	0.6	0.64	0.34	0.49	0.46	0.47	0.39
0.001	0.7	0.64	0.33	0.49	0.45	0.49	0.39
0.001	0.8	0.65	0.33	0.49	0.45	0.47	0.40
0.001	0.9	0.63	0.34	0.49	0.45	0.46	0.38
0.001	1	0.64	0.34	0.49	0.46	0.48	0.39
0.01	0.1	0.64	0.34	0.49	0.45	0.45	0.39
0.01	0.2	0.64	0.34	0.49	0.45	0.45	0.39
0.01	0.3	0.65	0.34	0.49	0.46	0.46	0.40
0.01	0.4	0.65	0.34	0.49	0.46	0.48	0.40
0.01	0.5	0.65	0.34	0.50	0.46	0.49	0.40
0.01	0.6	0.65	0.34	0.49	0.45	0.48	0.40
0.01	0.7	0.63	0.34	0.49	0.45	0.47	0.39
0.01	0.8	0.65	0.34	0.50	0.46	0.48	0.39
0.01	0.9	0.65	0.33	0.49	0.46	0.48	0.40
0.01	1	0.65	0.33	0.49	0.45	0.45	0.39
0.1	0.1	0.26	0.74	0.50	0.57	0.49	0.19
0.1	0.2	0.25	0.73	0.49	0.56	0.45	0.18
0.1	0.3	0.25	0.74	0.49	0.57	0.48	0.18
0.1	0.4	0.24	0.73	0.49	0.56	0.47	0.17
0.1	0.5	0.25	0.74	0.49	0.57	0.48	0.18
0.1	0.6	0.25	0.73	0.49	0.56	0.47	0.18
0.1	0.7	0.25	0.74	0.49	0.56	0.48	0.18
0.1	0.8	0.23	0.74	0.49	0.56	0.45	0.17
0.1	0.9	0.26	0.73	0.50	0.57	0.48	0.19
0.1	1	0.25	0.73	0.49	0.56	0.48	0.18
1E-05	0.1	0.24	0.73	0.49	0.56	0.48	0.18
1E-05	0.2	0.25	0.73	0.49	0.56	0.48	0.18
1E-05	0.3	0.24	0.74	0.49	0.57	0.48	0.18
1E-05	0.4	0.24	0.74	0.49	0.56	0.46	0.18
1E-05	0.5	0.24	0.74	0.49	0.57	0.48	0.18
1E-05	0.6	0.24	0.74	0.49	0.57	0.47	0.18

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1E-05	0.7	0.23	0.76	0.49	0.57	0.47	0.17
1E-05	0.8	0.25	0.74	0.50	0.57	0.48	0.19
1E-05	0.9	0.24	0.74	0.49	0.56	0.46	0.18
1E-05	1	0.25	0.74	0.49	0.56	0.47	0.18
1E-09	0.1	0.63	0.34	0.48	0.45	0.45	0.38
1E-09	0.2	0.64	0.34	0.49	0.46	0.47	0.40
1E-09	0.3	0.64	0.35	0.49	0.46	0.43	0.39
1E-09	0.4	0.66	0.34	0.50	0.46	0.49	0.40
1E-09	0.5	0.65	0.34	0.50	0.46	0.46	0.39
1E-09	0.6	0.64	0.34	0.49	0.45	0.46	0.39
1E-09	0.7	0.64	0.35	0.49	0.46	0.48	0.40
1E-09	0.8	0.65	0.32	0.49	0.45	0.48	0.40
1E-09	0.9	0.65	0.34	0.50	0.46	0.49	0.40
1E-09	1	0.64	0.34	0.49	0.45	0.46	0.40
10	0.1	0.50	0.53	0.51	0.52	0.51	0.38
10	0.2	0.51	0.52	0.51	0.51	0.51	0.39
10	0.3	0.49	0.52	0.51	0.51	0.50	0.38
10	0.4	0.50	0.51	0.51	0.51	0.50	0.38
10	0.5	0.51	0.52	0.52	0.51	0.55	0.39
10	0.6	0.51	0.51	0.51	0.51	0.51	0.39
10	0.7	0.49	0.52	0.51	0.51	0.50	0.38
10	0.8	0.50	0.53	0.51	0.52	0.52	0.39
10	0.9	0.51	0.51	0.51	0.51	0.52	0.39
10	1	0.51	0.52	0.51	0.51	0.53	0.39
100	0.1	0.49	0.51	0.50	0.50	0.49	0.37
100	0.2	0.48	0.51	0.49	0.50	0.49	0.37
100	0.3	0.48	0.51	0.49	0.50	0.49	0.37
100	0.4	0.50	0.51	0.50	0.51	0.52	0.38
100	0.5	0.48	0.51	0.50	0.50	0.50	0.37
100	0.6	0.49	0.51	0.50	0.51	0.51	0.38
100	0.7	0.49	0.51	0.50	0.50	0.50	0.38
100	0.8	0.48	0.51	0.50	0.50	0.50	0.37
100	0.9	0.46	0.50	0.48	0.49	0.50	0.35
100	1	0.49	0.51	0.50	0.50	0.50	0.37
20	0.1	0.49	0.52	0.50	0.51	0.49	0.38

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
20	0.2	0.51	0.52	0.51	0.51	0.51	0.39
20	0.3	0.50	0.52	0.51	0.52	0.49	0.39
20	0.4	0.51	0.54	0.52	0.53	0.53	0.40
20	0.5	0.51	0.52	0.52	0.52	0.52	0.39
20	0.6	0.50	0.52	0.51	0.51	0.52	0.38
20	0.7	0.49	0.52	0.51	0.51	0.53	0.38
20	0.8	0.51	0.52	0.51	0.51	0.52	0.39
20	0.9	0.49	0.52	0.50	0.51	0.50	0.38
20	1	0.50	0.52	0.51	0.52	0.53	0.39
50	0.1	0.49	0.52	0.50	0.51	0.50	0.38
50	0.2	0.51	0.52	0.51	0.51	0.52	0.39
50	0.3	0.49	0.53	0.51	0.52	0.51	0.38
50	0.4	0.49	0.52	0.51	0.51	0.53	0.38
50	0.5	0.49	0.52	0.51	0.51	0.50	0.38
50	0.6	0.50	0.51	0.51	0.51	0.53	0.39
50	0.7	0.51	0.51	0.51	0.51	0.51	0.39
50	0.8	0.50	0.51	0.51	0.51	0.50	0.38
50	0.9	0.49	0.52	0.51	0.51	0.51	0.38
50	1	0.49	0.52	0.50	0.51	0.51	0.37
1E-07	0.1	0.66	0.33	0.50	0.45	0.51	0.40
1E-07	0.2	0.63	0.33	0.48	0.45	0.47	0.38
1E-07	0.3	0.64	0.34	0.49	0.46	0.49	0.39
1E-07	0.4	0.65	0.33	0.49	0.46	0.47	0.39
1E-07	0.5	0.66	0.32	0.49	0.45	0.49	0.40
1E-07	0.6	0.64	0.34	0.49	0.45	0.46	0.39
1E-07	0.7	0.64	0.34	0.49	0.45	0.44	0.39
1E-07	0.8	0.66	0.32	0.49	0.45	0.48	0.40
1E-07	0.9	0.65	0.33	0.49	0.45	0.47	0.40
1E-07	1	0.64	0.34	0.49	0.45	0.48	0.39

Table 56- SVM classifier performance: ontological classification for the enabling technologies (kernel: linear)

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
0.001	0.1	0.64	0.34	0.49	0.45	0.45	0.39
0.001	0.2	0.67	0.34	0.50	0.46	0.49	0.40
0.001	0.3	0.67	0.33	0.50	0.46	0.48	0.41
0.001	0.4	0.68	0.32	0.50	0.46	0.48	0.42
0.001	0.5	0.69	0.32	0.50	0.46	0.51	0.42
0.001	0.6	0.69	0.33	0.51	0.46	0.51	0.43
0.001	0.7	0.69	0.33	0.51	0.46	0.50	0.42
0.001	0.8	0.69	0.32	0.50	0.46	0.50	0.42
0.001	0.9	0.69	0.33	0.51	0.46	0.51	0.43
0.001	1	0.71	0.33	0.52	0.47	0.53	0.44
0.01	0.1	0.64	0.34	0.49	0.45	0.46	0.38
0.01	0.2	0.64	0.34	0.49	0.46	0.46	0.39
0.01	0.3	0.67	0.33	0.50	0.46	0.51	0.41
0.01	0.4	0.66	0.34	0.50	0.46	0.48	0.40
0.01	0.5	0.66	0.34	0.50	0.46	0.49	0.41
0.01	0.6	0.68	0.33	0.51	0.46	0.50	0.42
0.01	0.7	0.68	0.33	0.51	0.46	0.52	0.42
0.01	0.8	0.67	0.33	0.50	0.46	0.49	0.41
0.01	0.9	0.70	0.33	0.52	0.47	0.52	0.44
0.01	1	0.69	0.33	0.51	0.46	0.55	0.42
0.1	0.1	0.24	0.75	0.50	0.57	0.49	0.18
0.1	0.2	0.26	0.73	0.50	0.57	0.50	0.19
0.1	0.3	0.25	0.73	0.49	0.56	0.47	0.18
0.1	0.4	0.27	0.74	0.51	0.57	0.50	0.20
0.1	0.5	0.27	0.73	0.50	0.57	0.48	0.19
0.1	0.6	0.28	0.73	0.50	0.57	0.51	0.21
0.1	0.7	0.30	0.72	0.51	0.57	0.51	0.22
0.1	0.8	0.28	0.73	0.50	0.57	0.50	0.20
0.1	0.9	0.29	0.73	0.51	0.57	0.54	0.21
0.1	1	0.30	0.72	0.51	0.57	0.53	0.22
1E-05	0.1	0.26	0.74	0.50	0.57	0.49	0.19
1E-05	0.2	0.27	0.74	0.50	0.57	0.52	0.20
1E-05	0.3	0.26	0.72	0.49	0.56	0.47	0.19
1E-05	0.4	0.26	0.73	0.50	0.57	0.49	0.19
1E-05	0.5	0.28	0.73	0.50	0.57	0.52	0.21

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
1E-05	0.6	0.28	0.73	0.50	0.57	0.47	0.20
1E-05	0.7	0.30	0.71	0.51	0.56	0.50	0.21
1E-05	0.8	0.28	0.73	0.50	0.57	0.51	0.20
1E-05	0.9	0.30	0.73	0.51	0.57	0.49	0.22
1E-05	1	0.30	0.72	0.51	0.57	0.53	0.22
1E-09	0.1	0.66	0.34	0.50	0.46	0.50	0.41
1E-09	0.2	0.67	0.33	0.50	0.46	0.48	0.41
1E-09	0.3	0.66	0.33	0.50	0.46	0.50	0.41
1E-09	0.4	0.66	0.33	0.50	0.46	0.49	0.41
1E-09	0.5	0.68	0.32	0.50	0.45	0.50	0.41
1E-09	0.6	0.69	0.33	0.51	0.46	0.50	0.42
1E-09	0.7	0.68	0.34	0.51	0.47	0.51	0.42
1E-09	0.8	0.69	0.33	0.51	0.46	0.52	0.43
1E-09	0.9	0.69	0.33	0.51	0.46	0.54	0.43
1E-09	1	0.70	0.33	0.51	0.47	0.52	0.44
10	0.1	0.47	0.51	0.49	0.50	0.50	0.36
10	0.2	0.47	0.52	0.50	0.51	0.50	0.37
10	0.3	0.45	0.55	0.50	0.52	0.50	0.36
10	0.4	0.45	0.55	0.50	0.51	0.50	0.36
10	0.5	0.47	0.53	0.50	0.51	0.50	0.37
10	0.6	0.46	0.53	0.49	0.51	0.45	0.36
10	0.7	0.43	0.53	0.48	0.50	0.47	0.34
10	0.8	0.44	0.54	0.49	0.51	0.49	0.35
10	0.9	0.44	0.55	0.50	0.52	0.47	0.35
10	1	0.43	0.56	0.50	0.52	0.50	0.35
100	0.1	0.49	0.55	0.52	0.53	0.51	0.39
100	0.2	0.44	0.54	0.49	0.51	0.51	0.35
100	0.3	0.39	0.55	0.47	0.49	0.50	0.31
100	0.4	0.41	0.53	0.47	0.49	0.48	0.33
100	0.5	0.40	0.53	0.47	0.49	0.47	0.32
100	0.6	0.42	0.53	0.47	0.49	0.46	0.33
100	0.7	0.42	0.53	0.48	0.49	0.47	0.33
100	0.8	0.43	0.54	0.48	0.50	0.49	0.35
100	0.9	0.42	0.54	0.48	0.50	0.47	0.33
100	1	0.43	0.55	0.49	0.51	0.48	0.34

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-score
C	Gamma						
20	0.1	0.48	0.54	0.51	0.52	0.51	0.38
20	0.2	0.47	0.54	0.51	0.52	0.49	0.37
20	0.3	0.46	0.54	0.50	0.51	0.51	0.36
20	0.4	0.46	0.53	0.50	0.51	0.50	0.37
20	0.5	0.43	0.54	0.48	0.50	0.49	0.34
20	0.6	0.43	0.54	0.49	0.51	0.50	0.35
20	0.7	0.42	0.55	0.49	0.51	0.47	0.34
20	0.8	0.40	0.55	0.48	0.50	0.49	0.33
20	0.9	0.41	0.54	0.47	0.50	0.45	0.33
20	1	0.41	0.56	0.49	0.51	0.48	0.34
50	0.1	0.49	0.54	0.52	0.53	0.51	0.38
50	0.2	0.47	0.55	0.51	0.52	0.47	0.38
50	0.3	0.43	0.54	0.48	0.50	0.49	0.34
50	0.4	0.42	0.56	0.49	0.51	0.48	0.33
50	0.5	0.40	0.55	0.47	0.50	0.50	0.33
50	0.6	0.43	0.53	0.48	0.50	0.47	0.34
50	0.7	0.42	0.53	0.48	0.49	0.46	0.33
50	0.8	0.41	0.55	0.48	0.51	0.50	0.34
50	0.9	0.43	0.53	0.48	0.50	0.45	0.34
50	1	0.42	0.55	0.48	0.51	0.45	0.34
1E-07	0.1	0.65	0.33	0.49	0.45	0.46	0.39
1E-07	0.2	0.66	0.33	0.49	0.45	0.48	0.40
1E-07	0.3	0.67	0.33	0.50	0.46	0.47	0.41
1E-07	0.4	0.67	0.32	0.50	0.45	0.46	0.41
1E-07	0.5	0.67	0.34	0.51	0.46	0.50	0.41
1E-07	0.6	0.68	0.33	0.50	0.46	0.50	0.42
1E-07	0.7	0.70	0.32	0.51	0.46	0.52	0.43
1E-07	0.8	0.70	0.34	0.52	0.47	0.52	0.43
1E-07	0.9	0.70	0.34	0.52	0.47	0.52	0.43
1E-07	1	0.70	0.32	0.51	0.46	0.52	0.43

Table 57- SVM classifier performance: ontological classification for the enabling technologies (kernel: RBF)

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
0.001	0.1	0.68	0.42	0.55	0.53	0.59	0.46
0.001	0.2	0.66	0.42	0.54	0.52	0.55	0.44
0.001	0.3	0.67	0.43	0.55	0.53	0.58	0.44
0.001	0.4	0.67	0.42	0.55	0.52	0.61	0.45
0.001	0.5	0.67	0.43	0.55	0.53	0.56	0.45
0.001	0.6	0.68	0.42	0.55	0.52	0.57	0.45
0.001	0.7	0.67	0.42	0.55	0.52	0.59	0.45
0.001	0.8	0.67	0.42	0.54	0.52	0.57	0.44
0.001	0.9	0.66	0.43	0.54	0.52	0.56	0.44
0.001	1	0.68	0.42	0.55	0.52	0.58	0.45
0.01	0.1	0.66	0.41	0.54	0.51	0.56	0.43
0.01	0.2	0.67	0.42	0.54	0.52	0.56	0.44
0.01	0.3	0.67	0.42	0.54	0.52	0.57	0.45
0.01	0.4	0.67	0.41	0.54	0.51	0.58	0.44
0.01	0.5	0.66	0.43	0.54	0.52	0.55	0.44
0.01	0.6	0.67	0.42	0.55	0.52	0.59	0.45
0.01	0.7	0.67	0.41	0.54	0.51	0.55	0.44
0.01	0.8	0.65	0.42	0.54	0.52	0.55	0.43
0.01	0.9	0.66	0.43	0.54	0.52	0.57	0.44
0.01	1	0.66	0.43	0.54	0.52	0.55	0.44
0.1	0.1	0.25	0.82	0.54	0.63	0.55	0.22
0.1	0.2	0.27	0.81	0.54	0.62	0.55	0.23
0.1	0.3	0.28	0.83	0.55	0.64	0.59	0.25
0.1	0.4	0.25	0.83	0.54	0.63	0.55	0.22
0.1	0.5	0.27	0.82	0.55	0.63	0.59	0.23
0.1	0.6	0.27	0.82	0.54	0.63	0.57	0.23
0.1	0.7	0.25	0.82	0.53	0.62	0.55	0.22
0.1	0.8	0.27	0.82	0.54	0.63	0.57	0.23
0.1	0.9	0.27	0.82	0.54	0.63	0.56	0.23
0.1	1	0.27	0.82	0.55	0.63	0.60	0.23
1E-05	0.1	0.26	0.83	0.54	0.63	0.55	0.23
1E-05	0.2	0.26	0.83	0.54	0.63	0.58	0.23
1E-05	0.3	0.27	0.84	0.55	0.64	0.56	0.24

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
1E-05	0.4	0.27	0.82	0.54	0.63	0.58	0.23
1E-05	0.5	0.27	0.82	0.55	0.63	0.59	0.24
1E-05	0.6	0.26	0.82	0.54	0.63	0.57	0.22
1E-05	0.7	0.27	0.83	0.55	0.64	0.61	0.24
1E-05	0.8	0.28	0.82	0.55	0.63	0.57	0.24
1E-05	0.9	0.27	0.82	0.55	0.63	0.58	0.24
1E-05	1	0.28	0.82	0.55	0.63	0.55	0.24
1E-09	0.1	0.67	0.42	0.55	0.52	0.58	0.44
1E-09	0.2	0.67	0.42	0.54	0.52	0.57	0.44
1E-09	0.3	0.66	0.42	0.54	0.52	0.59	0.43
1E-09	0.4	0.67	0.42	0.54	0.52	0.57	0.44
1E-09	0.5	0.66	0.42	0.54	0.52	0.56	0.44
1E-09	0.6	0.67	0.43	0.55	0.53	0.58	0.45
1E-09	0.7	0.69	0.42	0.55	0.52	0.57	0.45
1E-09	0.8	0.65	0.43	0.54	0.52	0.58	0.44
1E-09	0.9	0.67	0.42	0.55	0.52	0.60	0.45
1E-09	1	0.66	0.43	0.55	0.53	0.57	0.45
10	0.1	0.39	0.67	0.53	0.58	0.55	0.36
10	0.2	0.39	0.65	0.52	0.57	0.52	0.35
10	0.3	0.39	0.67	0.53	0.58	0.54	0.35
10	0.4	0.38	0.67	0.52	0.57	0.56	0.34
10	0.5	0.39	0.66	0.53	0.58	0.51	0.35
10	0.6	0.38	0.66	0.52	0.57	0.49	0.34
10	0.7	0.39	0.66	0.53	0.57	0.55	0.36
10	0.8	0.40	0.66	0.53	0.58	0.50	0.36
10	0.9	0.39	0.67	0.53	0.57	0.53	0.35
10	1	0.39	0.66	0.53	0.57	0.53	0.35
100	0.1	0.39	0.69	0.54	0.59	0.53	0.36
100	0.2	0.38	0.70	0.54	0.59	0.54	0.35
100	0.3	0.38	0.70	0.54	0.59	0.54	0.35
100	0.4	0.38	0.70	0.54	0.60	0.56	0.35
100	0.5	0.38	0.69	0.53	0.59	0.54	0.35
100	0.6	0.38	0.70	0.54	0.59	0.58	0.35

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
100	0.7	0.39	0.70	0.55	0.60	0.55	0.36
100	0.8	0.39	0.69	0.54	0.59	0.50	0.36
100	0.9	0.37	0.70	0.54	0.59	0.54	0.34
100	1	0.39	0.69	0.54	0.59	0.52	0.37
20	0.1	0.38	0.67	0.53	0.58	0.51	0.35
20	0.2	0.38	0.67	0.52	0.57	0.51	0.35
20	0.3	0.39	0.66	0.52	0.57	0.51	0.35
20	0.4	0.38	0.66	0.52	0.57	0.53	0.34
20	0.5	0.39	0.66	0.53	0.57	0.54	0.35
20	0.6	0.40	0.66	0.53	0.57	0.49	0.35
20	0.7	0.38	0.67	0.52	0.57	0.53	0.34
20	0.8	0.39	0.67	0.53	0.58	0.53	0.35
20	0.9	0.38	0.67	0.52	0.57	0.54	0.35
20	1	0.38	0.66	0.52	0.57	0.52	0.34
50	0.1	0.40	0.69	0.55	0.60	0.54	0.37
50	0.2	0.42	0.69	0.55	0.60	0.55	0.38
50	0.3	0.40	0.70	0.55	0.60	0.54	0.37
50	0.4	0.39	0.69	0.54	0.59	0.52	0.36
50	0.5	0.37	0.68	0.53	0.58	0.51	0.34
50	0.6	0.40	0.69	0.54	0.59	0.52	0.36
50	0.7	0.38	0.69	0.53	0.59	0.51	0.35
50	0.8	0.37	0.69	0.53	0.58	0.51	0.34
50	0.9	0.39	0.69	0.54	0.59	0.52	0.36
50	1	0.40	0.69	0.55	0.60	0.53	0.36
1E-07	0.1	0.66	0.42	0.54	0.52	0.56	0.44
1E-07	0.2	0.66	0.42	0.54	0.52	0.55	0.44
1E-07	0.3	0.67	0.42	0.54	0.52	0.58	0.44
1E-07	0.4	0.66	0.43	0.55	0.52	0.57	0.44
1E-07	0.5	0.68	0.42	0.55	0.52	0.56	0.45
1E-07	0.6	0.65	0.43	0.54	0.52	0.57	0.44
1E-07	0.7	0.67	0.42	0.55	0.52	0.59	0.45
1E-07	0.8	0.66	0.42	0.54	0.52	0.57	0.44
1E-07	0.9	0.66	0.41	0.54	0.51	0.56	0.44

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
1E-07	1	0.67	0.43	0.55	0.52	0.60	0.45

Table 58- SVM classifier performance: ontological semantic classification for the enabling technologies (kernel: linear)

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
0.001	0.1	0.64	0.43	0.54	0.52	0.55	0.43
0.001	0.2	0.66	0.42	0.54	0.52	0.57	0.43
0.001	0.3	0.65	0.44	0.54	0.53	0.60	0.44
0.001	0.4	0.66	0.44	0.55	0.53	0.61	0.45
0.001	0.5	0.65	0.44	0.55	0.53	0.56	0.44
0.001	0.6	0.64	0.45	0.55	0.53	0.58	0.43
0.001	0.7	0.64	0.45	0.55	0.53	0.59	0.43
0.001	0.8	0.64	0.44	0.54	0.53	0.59	0.43
0.001	0.9	0.64	0.46	0.55	0.54	0.58	0.44
0.001	1	0.63	0.46	0.54	0.53	0.58	0.42
0.01	0.1	0.67	0.42	0.55	0.52	0.57	0.44
0.01	0.2	0.65	0.43	0.54	0.52	0.55	0.43
0.01	0.3	0.65	0.44	0.55	0.53	0.54	0.44
0.01	0.4	0.66	0.44	0.55	0.53	0.59	0.45
0.01	0.5	0.65	0.43	0.54	0.52	0.55	0.43
0.01	0.6	0.63	0.45	0.54	0.53	0.57	0.43
0.01	0.7	0.65	0.45	0.55	0.53	0.56	0.44
0.01	0.8	0.64	0.45	0.55	0.53	0.57	0.43
0.01	0.9	0.64	0.46	0.55	0.54	0.60	0.43
0.01	1	0.62	0.47	0.54	0.54	0.59	0.42
0.1	0.1	0.25	0.83	0.54	0.63	0.54	0.22
0.1	0.2	0.28	0.82	0.55	0.63	0.58	0.24
0.1	0.3	0.27	0.82	0.55	0.63	0.60	0.23
0.1	0.4	0.25	0.83	0.54	0.63	0.56	0.22
0.1	0.5	0.27	0.84	0.56	0.65	0.61	0.24
0.1	0.6	0.24	0.85	0.55	0.64	0.57	0.22
0.1	0.7	0.23	0.86	0.55	0.65	0.57	0.22
0.1	0.8	0.23	0.86	0.55	0.65	0.58	0.22
0.1	0.9	0.22	0.87	0.54	0.65	0.57	0.21
0.1	1	0.22	0.87	0.54	0.65	0.58	0.21
1E-05	0.1	0.26	0.82	0.54	0.63	0.56	0.23
1E-05	0.2	0.26	0.84	0.55	0.64	0.58	0.24

SVM Parameters		Average of Sensitivity	Average of Specificit y	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
1E-05	0.3	0.26	0.83	0.55	0.64	0.58	0.23
1E-05	0.4	0.24	0.86	0.55	0.65	0.59	0.23
1E-05	0.5	0.26	0.84	0.55	0.64	0.55	0.23
1E-05	0.6	0.25	0.86	0.55	0.65	0.61	0.23
1E-05	0.7	0.24	0.85	0.54	0.64	0.59	0.22
1E-05	0.8	0.23	0.86	0.54	0.64	0.56	0.21
1E-05	0.9	0.24	0.85	0.54	0.64	0.58	0.22
1E-05	1	0.21	0.86	0.54	0.64	0.54	0.20
1E-09	0.1	0.66	0.43	0.54	0.52	0.54	0.44
1E-09	0.2	0.66	0.43	0.55	0.53	0.56	0.44
1E-09	0.3	0.65	0.43	0.54	0.52	0.55	0.43
1E-09	0.4	0.64	0.44	0.54	0.53	0.59	0.43
1E-09	0.5	0.64	0.45	0.55	0.53	0.57	0.43
1E-09	0.6	0.64	0.45	0.54	0.53	0.58	0.43
1E-09	0.7	0.65	0.45	0.55	0.53	0.59	0.43
1E-09	0.8	0.64	0.47	0.55	0.54	0.61	0.44
1E-09	0.9	0.63	0.47	0.55	0.54	0.56	0.43
1E-09	1	0.63	0.47	0.55	0.54	0.56	0.43
10	0.1	0.42	0.66	0.54	0.58	0.55	0.38
10	0.2	0.39	0.68	0.54	0.59	0.54	0.36
10	0.3	0.38	0.69	0.53	0.59	0.53	0.35
10	0.4	0.36	0.71	0.54	0.60	0.53	0.35
10	0.5	0.35	0.73	0.54	0.60	0.55	0.34
10	0.6	0.33	0.74	0.53	0.60	0.50	0.33
10	0.7	0.33	0.73	0.53	0.60	0.52	0.32
10	0.8	0.32	0.74	0.53	0.61	0.56	0.32
10	0.9	0.31	0.75	0.53	0.61	0.53	0.31
10	1	0.32	0.76	0.54	0.62	0.54	0.33
100	0.1	0.39	0.70	0.55	0.60	0.55	0.37
100	0.2	0.36	0.73	0.55	0.61	0.54	0.35
100	0.3	0.38	0.73	0.55	0.61	0.52	0.37
100	0.4	0.38	0.72	0.55	0.60	0.54	0.36

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
100	0.5	0.40	0.71	0.56	0.61	0.55	0.37
100	0.6	0.40	0.74	0.57	0.63	0.56	0.38
100	0.7	0.40	0.74	0.57	0.63	0.57	0.39
100	0.8	0.42	0.74	0.58	0.63	0.61	0.40
100	0.9	0.39	0.75	0.57	0.63	0.56	0.38
100	1	0.39	0.75	0.57	0.63	0.53	0.38
20	0.1	0.41	0.67	0.54	0.59	0.55	0.37
20	0.2	0.38	0.69	0.53	0.59	0.53	0.35
20	0.3	0.36	0.72	0.54	0.60	0.54	0.34
20	0.4	0.35	0.73	0.54	0.61	0.56	0.34
20	0.5	0.36	0.73	0.54	0.61	0.54	0.35
20	0.6	0.33	0.73	0.53	0.60	0.53	0.32
20	0.7	0.32	0.73	0.53	0.60	0.53	0.32
20	0.8	0.31	0.73	0.52	0.59	0.52	0.30
20	0.9	0.33	0.74	0.53	0.60	0.51	0.32
20	1	0.33	0.74	0.54	0.61	0.53	0.33
50	0.1	0.40	0.68	0.54	0.59	0.55	0.36
50	0.2	0.38	0.73	0.56	0.62	0.56	0.38
50	0.3	0.37	0.72	0.55	0.61	0.55	0.36
50	0.4	0.35	0.71	0.53	0.59	0.52	0.33
50	0.5	0.35	0.72	0.53	0.60	0.52	0.34
50	0.6	0.35	0.72	0.54	0.60	0.52	0.34
50	0.7	0.35	0.72	0.54	0.60	0.55	0.34
50	0.8	0.36	0.73	0.54	0.61	0.54	0.35
50	0.9	0.36	0.73	0.55	0.61	0.54	0.35
50	1	0.34	0.74	0.54	0.61	0.53	0.34
1E-07	0.1	0.64	0.42	0.53	0.51	0.56	0.43
1E-07	0.2	0.66	0.43	0.54	0.52	0.60	0.44
1E-07	0.3	0.65	0.44	0.55	0.53	0.54	0.44
1E-07	0.4	0.64	0.44	0.54	0.53	0.54	0.43
1E-07	0.5	0.65	0.44	0.55	0.53	0.58	0.44
1E-07	0.6	0.65	0.45	0.55	0.53	0.56	0.44

SVM Parameters		Average of Sensitivity	Average of Specificity	Average of ROC AUC	Average of Accuracy	Average of Precision	Average of F-Score
C	Gamma						
1E-07	0.7	0.64	0.45	0.55	0.53	0.57	0.44
1E-07	0.8	0.63	0.46	0.54	0.53	0.60	0.43
1E-07	0.9	0.62	0.47	0.54	0.54	0.54	0.43
1E-07	1	0.62	0.47	0.55	0.54	0.56	0.42

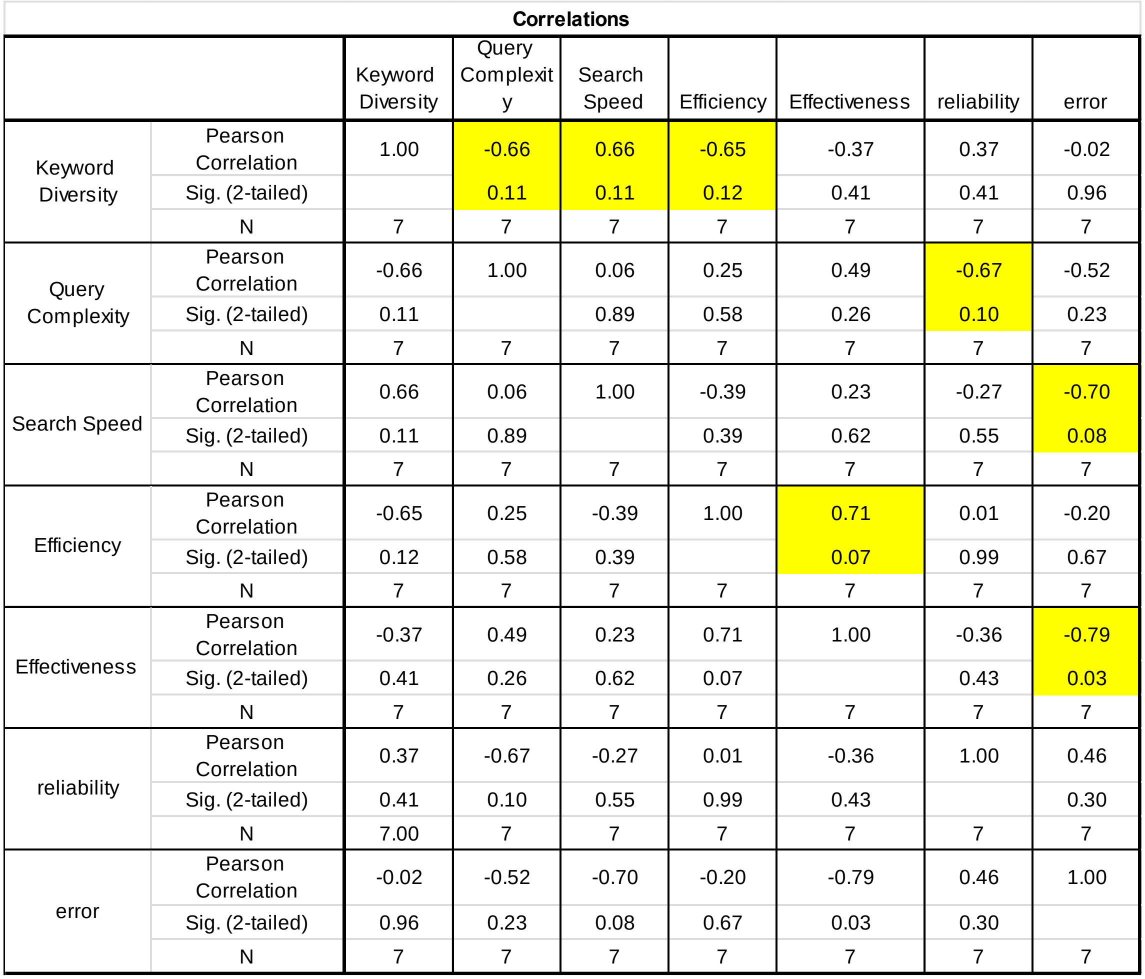
Table 59- SVM classifier performance: ontological semantic classification for the enabling technologies (kernel: RFB)

Appendix E: Patents Identified as Opportunities

38555	5387272	5800637	6129134	6427755	6800244	7165598	7767040	8420011	9016443
5111873	5391243	5800902	6180258	6460602	6805827	7168529	7776257	8435670	9028959
5127467	5423925	5837387	6187119	6467528	6817859	7175719	7776486	8453711	9057121
5131144	5451352	5841042	6196363	6468673	6841011	7226641	7793703	8535764	9089893
5135041	5460639	5855239	6241056	6491772	6844085	7258154	7824607	8550145	9109275
5157136	5462107	5858127	6271162	6505716	6880681	7258209	7845918	8607941	9193411
5165464	5476554	5887684	6280541	6524405	6908590	7261951	7861832	8656982	9194033
5211500	5501833	5902511	6290784	6537395	6913062	7267882	7879129	8657972	9228244
5224535	5509728	5948353	6299834	6558815	6918427	7331373	7879460	8672077	9285169
5234080	5515905	5979538	6309743	6564856	6962189	7429301	7892369	8701948	9293232
5244517	5524696	5979614	6321826	6572712	7000677	7438770	8016018	8709124	9352388
5246056	5535857	5980651	6328093	6582533	7045022	7494552	8025747	8714232	9376738
5253398	5613184	5988260	6328820	6610247	7045207	7559353	8091609	8802243	9403574
5261511	5658400	6021842	6329075	6627340	7056598	7588179	8132612	8840738	9453272
5282374	5664619	6088906	6337455	6705848	7074282	7608156	8168011	8905203	9487848
5318094	5705125	6109334	6368427	6712124	7081151	7628196	8192561	8939266	9488238
5323883	5728638	6110268	6386271	6719104	7086151	7644750	8276647	8945466	3421886
5326384	5746268	6110299	6395107	6745819	7087318	7648594	8349096	8962163	--
5344606	5782324	6125916	6409966	6773664	7163594	7766073	8360134	8997945	--

Table 60- The patent numbers of the 188 opportunities identified in the Finex case

Appendix F: Correlation Analysis Results



Correlations

		Keyword Diversity	Query Complexity	Search Speed	Efficiency	Effectiveness	reliability	error
Keyword Diversity	Pearson Correlation	1.00	-0.66	0.66	-0.65	-0.37	0.37	-0.02
	Sig. (2-tailed)		0.11	0.11	0.12	0.41	0.41	0.96
	N	7	7	7	7	7	7	7
Query Complexity	Pearson Correlation	-0.66	1.00	0.06	0.25	0.49	-0.67	-0.52
	Sig. (2-tailed)	0.11		0.89	0.58	0.26	0.10	0.23
	N	7	7	7	7	7	7	7
Search Speed	Pearson Correlation	0.66	0.06	1.00	-0.39	0.23	-0.27	-0.70
	Sig. (2-tailed)	0.11	0.89		0.39	0.62	0.55	0.08
	N	7	7	7	7	7	7	7
Efficiency	Pearson Correlation	-0.65	0.25	-0.39	1.00	0.71	0.01	-0.20
	Sig. (2-tailed)	0.12	0.58	0.39		0.07	0.99	0.67
	N	7	7	7	7	7	7	7
Effectiveness	Pearson Correlation	-0.37	0.49	0.23	0.71	1.00	-0.36	-0.79
	Sig. (2-tailed)	0.41	0.26	0.62	0.07		0.43	0.03
	N	7	7	7	7	7	7	7
reliability	Pearson Correlation	0.37	-0.67	-0.27	0.01	-0.36	1.00	0.46
	Sig. (2-tailed)	0.41	0.10	0.55	0.99	0.43		0.30
	N	7.00	7	7	7	7	7	7
error	Pearson Correlation	-0.02	-0.52	-0.70	-0.20	-0.79	0.46	1.00
	Sig. (2-tailed)	0.96	0.23	0.08	0.67	0.03	0.30	
	N	7	7	7	7	7	7	7

Table 61- Correlation analysis results