Norwegian School of Economics
Bergen, Autumn 2021

Enabling technology intelligence

An analytical, hybrid similarity framework
to generate practical insights from patent data

Janik Weigel

Supervisor: Steffen Juranek

Master Thesis

MSc. Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

Technology intelligence plays a crucial role for corporate strategy, especially during times of accelerating technological progress. Accompanying advancements in computing capacities and natural language processing tools give today’s decision-makers access to a broad range of market information. Particular deep insights are made possible through patent data, for example by measuring the similarity between individual documents, or whole IP portfolios.

Progress in the literature that focuses on accessing this wealth of information, termed IP analytics, is often mis-aligned with the requirements at a company-level: most research either use patent data as an exemplary case to showcase new algorithms for big (textual) data analytics, or they focus on insight-generation for policymakers and other researchers, leaving company decision-makers without practical, applicable solutions.

Therefore, the main contribution of this thesis is a practical and re-applicable framework to assess patent similarity via semantic and categorical means, combined into a hybrid model that can be used on a given portfolio of US patents. The outcome of this model is a weighted list of patent assignee organizations, which are strategically relevant to the initial portfolio.

Through a case study of the US patent portfolio of a medium-sized German firm, it is shown that the proposed hybrid similarity framework can automatically and accurately identify relevant market players, enabling company decision-makers with technology intelligence in a clear and concise way. Both measures of patent similarity were shown to be positively correlated with the strategic importance of the identified assignees to the target company.

All R code developed is available for replication, and application on new patent portfolios at: *https://github.com/janikweigel/IP_Similarity_Thesis*

Preface

While my idea of using patent data for market insights goes back over some considerable time already, sufficiently long enough that it could mature and be discussed with a range of intelligent people over time, the main work and writing was conducted during the last few weeks. It was indeed a long way from initial idea into reality (*or at least: into words and code*).

Especially during these final weeks, I need to thank my family and close ones around me, for enduring me and my deadline-approaching behaviour. Furthermore, the strategists and decision-makers from a range of industries that will not be named individually, from whom I was able to draw inspiration, and receive feedback for my initial concept, deserve a big *thank you*. Similarly, my good friends Aleksej and Egor from the Business Analytics major at NHH were my sounding board throughout three semesters of working on the thesis in Norway, Singapore, and Germany. Finally, I need to thank my academic supervisor Prof. Steffen Juranek specifically – not (only) for the content-related support and guidance in the field of law that I had no previous touchpoints with, but especially for the enormous flexibility and understanding that he showed over this extended period working with me, which started one year, seven months, and two weeks ago.

After all, a thought-through concept was transformed into something that is hopefully use- and helpful for people interested in the same idea I was – understanding and navigating an increasingly complex market environment.

Table of Contents

List of Abbreviations

AI	Artificial intelligence	Science and practise of developing computer programs able to perform a range of tasks that would traditionally require human intelligence
AO	Assignee organisation	Legal entity that owns the rights to an IPR class
API	Application programming interface	A software interface that enables other computer programs to communicate with it
CPC	Cooperative Patent Classification	Hierarchical classification system for patents, jointly developed by the EPO and the USPTO
CSV	Comma-separated values	A widely used data-interchange format
EPO	European Patent Office	Relevant patent office of the European Union
IP(R)	Intellectual property (rights)	Collection of legal, region-specific rights that protect individual intellectual achievements
IPA	Intellectual property analytics	The data science of analysing IPR information for different disciplines and use-cases
IPC	International Patent Classification	Hierarchical classification system of patents, developed by the World IP Organization
JSON	JavaScript object notation	A clearly structured data-interchange format that is easy for machines to generate and to parse
NLP	Natural language processing	Scientific discipline researching the interaction between computers and human language.
p2p	Patent-to-patent	Relationship between two individual patents
SME	Small- and medium-sized enterprise	Category of businesses with less than 1500 employees
TI	Technology intelligence	Strategic capability of a company to identify technological opportunities and threats
USPTO	United States Patent and Technology Office	Relevant patent office of the United States
USPN	United States patent number	Unique identifier number of USPTO patents

List of Figures and Tables

1. Introduction

1.1 Relevant Context

Technological progress is accelerating, with new products and whole new industries being created at a pace that has not been seen before (Crafts, 2021). Incumbent companies are under increasing pressure of becoming irrelevant, if they do not continue to innovate – both in regards to their products and services, but also their business model (Christensen et al., 2018). This change is particularly evident in two major areas, which mutually influence technological progress, and become an increasingly important aspect of corporate strategy: artificial intelligence (AI) and big data analytics on the one side, which enable completely new use-cases and drive innovation across industries, and intellectual property rights (IPRs) on the other side to protect innovation efforts, while simultaneously disclosing them.

One of the fundamental principles for innovation in the first place is the right to own IP. It is considered to be one of the key enablers for economic growth (van Looy & Magerman, 2019), by providing innovators (e.g., companies, research institutions, or individuals) with the incentive of a temporary monopoly in order to commercialize the result of a novel technology or process that includes an innovative step – or in other words: to exclusively sell an invention. In particular the IP category of patents plays an important role in safeguarding traditional innovation efforts, and is also considered an important measure of innovative activity, not only on a firm, but also on a sector, regional, or global level (Hain et al., 2021).

The polarizing appeal of IPRs became globally visible and evident during discussions at the beginning of 2021 regarding the lifting of patent protection of COVID-19 vaccines. A variety of groups argued for wider and faster availability of vaccines through these proposed waivers, including the US president and the pope of the Roman Catholic Church, prioritizing the global health needs during a pandemic over the right of a legal monopoly ((Zeferino De Menezes, 2021; Cokelaere, 2021)). What seems like an ethical choice, is refuted however by other groups, also those unaffiliated with the pharma companies producing the vaccines. They argue on grounds of global IP-law that removing IPRs not only removes further incentives to innovate, but in particular does not solve the manufacturing and distribution problem at hand, with a more practical solution being better use of existing licensing agreements and technology transfer (Haugen, 2021; Lindsey, 2021).

This showcases first of all the importance, but also the complexity around IPR issues. Increasingly, with the growth of internet usage and content sharing platforms, this complexity transfers to, and becomes more relevant to, everyday life. YouTube for example has been long criticised for its handling of copywrite claims when dealing with content creators, resulting in bans affecting millions of fans (Sands, 2018). Despite its shortcomings, IPRs have an overall positive, mutually influencing relationship with technological progress: one enables the other. The other big area with this dependent relationship on technological progress is AI.

A lot has been written about this megatrend that is already touching, or is about to, every aspect of today's life (Kim et al., 2021). It dominates both optimistic and pessimistic outlooks on our future, giving machine intelligence a higher, and human intelligence a less important role than today (Crafts, 2021). This includes a fundamental shift in the way companies work: Applied AI has been named "*the most applicable technology trend*" by a recent McKinsey & Company (2021) study, since it demonstrates the biggest potential to shape a wide range of industries. This can be broken down into transforming the very methods by which work is done across businesses, but also boost efficiency throughout existing processes, by applying these technologies. This implies that AI has both a macro- and micro-level impact on companies (Crafts, 2021). In reference to the earlier-mentioned importance of IPRs, this particular study also used die number of patents filed in a given area as a proxy for their analysis, showcasing the interlinkage of both themes with technological progress.

The wide availability of datasets and computing power that allowed for the growth of AI-based technologies also leads to some problems however. Firstly, since methods become increasingly complex, for example by shifting from supervised to unsupervised, or deep learning-based approaches, themselves and their results becomes less explainable ("*black box*"), which is a problem for transparent decision-making (Hsu et al., 2020; Krestel et al., 2021). Secondly, the sheer amount of data available leads to an information overload problem for organization, who struggle to keep up with it (X. Li et al., 2009). Finally, specialized AI knowledge clusters in specific regions and companies, leading to a large performance and digitalization gap between the top few percent of organization, often based out of the US or China, and the rest (Crafts, 2021). In line with research on disruptive innovation (Christensen et al., 2018), most of the product innovation in AI is not driven by market-leading incumbents, but rather from large technological companies like Google, Amazon, or Tencent, or venture-backed emerging start-ups. This means that a large part of technological progress is defined by companies that consider themselves "*software-first*", or even "*API-first*" (Marr, 2019).

Naturally, AI-based techniques have increasingly been applied to literature as well, which led to a fundamental shift in quantitative research itself: exponential growth in computing power available every year made more big data analytics and advanced machine learning-based algorithms available for researchers (Aristodemou et al., 2017). More recent advances in neural networks and deep learning techniques allowed for new benchmark performances on large visual and textual datasets (Arts et al., 2021; Yang et al., 2018). This enabled the discipline of IP analytics (IPA) to grow beyond measuring citations between patents, towards applying cutting-edge, AI-based models to work with the complexities of patent text, which is particularly hard to understand and extract meaning from (Fierro, 2013; Kim et al., 2021).

The effort is worth it however, since patents as a semantic carrier of technology itself, contain a large amount of the world's technology intelligence, most of it not published in any other form (van Looy & Magerman, 2019). This comes from the particular depth in information content and details provided while filing for such a patent, leading to the initially mentioned trade-off of getting a legal monopoly granted in exchange for making the underlying technological details of an invention public knowledge (Feldman, 2012). Thus, many strategic use-cases can be achieved by analysing patent data, such as technology forecasting (Kyebambe et al., 2017), generating competitive intelligence (Yan & Luo, 2017), or evaluating investment decisions (Sinan Erzurumlu & Pachamanova, 2020), among others. A cross-section of this use-cases is called technology intelligence (TI), which is the ability of a company to make use of the technical data available about the market and its competitors.

As a key method to enable TI, patent similarity measurement is considered one of fundamental building blocks of IPA (An et al., 2021). Hain et al. (2021) differentiate three main methods to measure similarity between technologies from IP data: classification-based, citation-based, and NLP-based approaches. Advantages and drawback of each will be expanded upon in Section 2. While significant efforts have been made to improve each of these fields over time with more advanced techniques, there is an increasing amount of evidence that combining approaches yields better results in measuring patent-to-patent (p2p) similarity (An et al., 2021; van Looy & Magerman, 2019; Zhang et al., 2016)**.** Therefore, a hybrid approach of semantic (text-based) and categorical (classification-based) patent features is proposed to enable decision-makers on a company-level with technology intelligence (TI).

1.2 Research Question

In order to advance the IPA literature towards interpretable similarity models, as well as to enable business practitioners with practical insights, the research question of this paper states:

How to generate market insights from public patent data that are
(i) *relevant on a company-level*, (ii) *easy to apply and understand*, and finally
(iii*) able to reduce bias from individual analysis methods,*
in order to equip decision-makers with technology intelligence?

Here, I hypothesize that a hybrid approach that uses both semantic, as well as categorical similarity information, based on the patent portfolio of a given company, is able to provide this strategic value (H1). This is based on recent findings from the IPA literature that will be discussed in section 2. Furthermore, I also hypothesize a non-linear relationship between patents and the strategic importance of those to the target firm, in particular I assume that high semantic, but medium categorical similarity (i.e., not too high; not too low) should have the highest importance to companies (H2). This is illustrated as a 2 x 3 matrix in figure 3, Appendix B. The reasoning is based additionally on the literature of disruptive innovation, which argues that radical new innovation, either technological, or via disruptive business models, is more likely to emerge from outside of incumbent's core industries, since it is claimed that those have no incentive for such novel innovation (Christensen et al., 2018).

In order to answer these questions, the remainder of this paper is structured as follows:

In Section 2, I review the literature on patent-based analysis methods, focusing on the mean to measure patent similarity. Here, the development of the IPA literature is set out, followed by a particular comparison of patent quality indicators and their usage in the respective papers. In Section 3, the methodological approach to assess patent similarity is described. Combining both a text-based approach to capture semantic meaning, as well as a categorial approach to use patent metadata, this hybrid approach builds on recently published databases, which are also described in detail. The described technique is applied on a case study of a German SME. In Section 4, first describing the context of the target firm and thereon identifying similar technologies to the firm's portfolio. Section 5 discusses the validated results from the case study, both quantitatively as well as qualitatively, and emphasises limitations of the thesis in three categories. Finally, Section 6 summarizes the approach, its findings, and concludes implications for the literature and decision-makers in companies.

2. Literature Review

2.1 History and Background of IPA

In varying forms, the science of using patents and other IP data for analysis (IPA) has been in existence since the 1930s, relying on the locally stored information of regional patent offices (C. Lee, 2021). Since the 1990s the world has witnessed a rapid growth in patenting activity, combined with a digitalisation push by patent offices, translating IP documents into computer-readable formats, stored in online accessible databases (Furman et al., 2018). IPA research has flourished since then, since the time and cost involved in retrieving large amounts of IP data was reduced significantly (Helmers et al., 2019).

From an analytical and methodological perspective, a wide range of tools were explored over time to deal with the specialities of patent documents, summarized by Kim et al. (2021) as:

a. A mix of structured (meta) and unstructured (textual) data within each document
b. A variety of classification systems employed for meta data
c. Expert-oriented, IP-specific language and general technicality used in texts

These aspects are representative of the history of IPA literature, which can be broadly classified into two categories: bibliographic information-based approaches (i.e., using meta data), and lexical based approaches (i.e., using textual data) (Zhang et al., 2016). Early on, analytics of IP documents focused on easily retrievable meta data, like citations and classifications, and advanced over time towards models that can analyse large amounts of textual data with increased computational power available (Aristodemou & Tietze, 2018).

Firstly, citations make up the “prior art” of patents an individual document refers to, and are performed in largely the same way as in scientific publications, which resulted in IPA using established methodologies of scholarly research and bibliometrics (Kim et al., 2021). Citations can be differentiated further, depending on the perspective of analysis: the target document referring to prior art (backward citations), or patents referring to the target document itself afterwards (forward citations). Especially the latter case received a lot of attention, since patents with a high number of forward citations were treated as particularly impactful for further innovation (Hain et al., 2021; Kyebambe et al., 2017). In general, citations have to be provided with a patent application, the relevant patent office can however add citations as well – the difference between both making up its own stream of research (Cotropia et al., 2013).

Secondly, patent classifications refer to one of several systems of hierarchical categories that classify patents according to their technology. Those systems are introduced and used by patent offices to allow a systematic arranging and retrieval of patent documents (Kim et al., 2021), for example the International Patent Classification (IPC) by the World IP Organisation, or the more recent Cooperative Patent Classification (CPC), jointly developed by the European Patent Office (EPO) and the United States Patent and Technology Office (USPTO). These structured systems include a lot of detailed, granular technical information and are subsequently used to map wider technological networks (Yan & Luo, 2017), but also allow for similarity checks between individual patent documents, since the different hierarchy levels of the IPC and CPC were designed based on the concept of proximity – meaning more similar technologies are placed closer to each other (Harris et al., 2010).

While some streams of the IPA literature continue to analyse citation-based and classification-based, a majority shifted their focus to also, or only, include textual content of patent documents (Aristodemou & Tietze, 2018), for instance by inferring semantical meaning from abstracts and claims, and comparing this from patent to patent (p2p). This is driven by simultaneous progress in computer science, data science, and related disciplines dealing with text mining techniques, which created the field of natural language processing (NLP) for a range of text-related tasks, drawing on much of the progress in overall AI research (Yang et al., 2018). The exact limitations of the NLP term are point of ongoing discussion and include a wide range of application fields beyond traditional text processing, such as transcription of human language from audio recordings (Arts et al., 2021). Lately, deep neural networks, a category of AI tools that is trained by deep learning techniques, enabled significant progress within NLP, since it does not require a manual selection of features, one of the major tasks in traditional machine learning, and especially important for textual data with potentially millions of features (Krestel et al., 2021).

The so far described differences within IPA methods show a diverse and heterogeneous field, which is also reflected in the way IPA literature is published, including a split of journals based in IP law, computer and data science, social science, or bibliometrics. Specifically, a large part of researchers come from, and publish outside of traditional legal-focused journals, and use patents as an application field, or “showcases” for their algorithmic advancements with new

ML-based textual data models (Aristodemou & Tietze, 2018). This is for example due to the earlier mentioned availability of large amounts of patent data, also including structured information, making it very accessible for big data analytics use-cases (Kim et al., 2021).

Similar to the methods applied, over time also a variety of use-cases for IPA have emerged, with specific research goals becoming increasing granular, testifying a growing importance of the field (Hain et al., 2021) – both on an individual patent-level as well as for overarching innovation topics. For example, IPA includes use-cases such as patent quality (Squicciarini et al., 2013), patent valuation (Hsu et al., 2020), and patent litigation analysis (Petherbridge, 2011), which aim to analyse individual IP assets, but also fields like emerging technology identification (C. Lee et al., 2018), and forecasting (Choi et al., 2021; Kyebambe et al., 2017), which try to infer larger trends from the technological information in IP assets.

This variety and fragmentation put a special focus on consolidated literature reviews, providing an overview of the different approaches and outcomes of IPA, and also providing joint taxonomy. The first comprehensive one has been performed by Abbas et al. (2014), which differentiate eight different use-cases for patent analysis: (i) novelty detection, (ii) trend analysis, (iii) forecasting technological developments, (iv) strategic technology planning, (v) technological road mapping, (vi) analysing patent infringement, (vii) competitor analysis, and finally (vii) identifying patent quality. More recently, Aristodemou & Tietze (2018) provided a more structured summary of use-cases:

a. Knowledge management (e.g., patent quality classification)
b. Technology management
    i. Technological patentability
    ii. Organizational R&D planning
    iii. Technological intelligence
c. Measuring economic value of IP
d. Extraction and management of information from IP

While generally applicable for the initial beneficiary of IPA, patent offices (Helmers et al., 2019), those fields have become especially important for companies who face ever-increasing market pressure globally, within and outside their core business (Christensen et al., 2018), and which are increasingly adopting methods and tools to use openly patent data strategically (Aristodemou & Tietze, 2018; personal communication, September 8, 2021).

Finally, Choi et at. (2021) summarized the shift in the IPA literature over time along three dimensions: from a retrospective to a prospective analysis perspective, from using ex-post to ex-ante evaluation, and shifting from unsupervised to supervised learning and analysis techniques. All of these changes aiming to provide more practical insights for a range of stakeholders: inventors, businesses, academia, politicians, and of course, patent offices.

2.2 Patent Similarity and Technological Intelligence

"*Patent similarity measurement, as one of the fundamental building blocks for patent analysis, is able to derive technical intelligence efficiently*" (An et al., 2021, p.1)

Technical or technology intelligence (TI) is not clearly defined in literature, but can broadly be categorized as the ability of an organization to identify and use technological opportunities and threats that may have a strategic impact on their current or future business (Aristodemou & Tietze, 2018; C. Lee, 2021). As shown earlier, TI is a major use-case of IPA, and the quote by An et al. (2021) above links it with the method required to enable it: patent similarity measures. Hain et al. (2021) provide a recent classification for p2p similarity measurement, differentiating three categories of approaches, described in Table 1, Appendix A, each with their identified advantages and drawbacks.

Specifically, they differentiate and define in line with the wider IPA literature:

a. Classification-based approaches as those measuring the co-occurrence of index classes
b. Citation-based approaches as those analysing co-citation (i.e., common forwards), bibliographic coupling (i.e., common backwards), or combined with indirect citations
c. Text-based approaches as those, which use either keyword-based methods, the analysis of the SAO-structure, ontology-based analysis, or based on machine/deep learning

As an example, for their own analysis, Hain et al. (2021) focus on measuring technological similarity semantically with embedding techniques and nearest-neighbour approximations, which are machine learning-based approaches used on text. They create vector representations of the main technology terms from abstracts, using only this source of data. With that they showcase on behalf of the electric vehicle industry three diverse research applications:

1. Firstly, visualizing the technology landscape of a given technology through internally homogeneous, externally heterogeneous patent clusters
2. Secondly, predicting patent quality measured by "*novelty*" and "*promisingness*"
3. Thirdly, mapping of knowledge flows and spill over effects between countries

Their choice of a textual-only approach using semantic features (i.e., category "C" in Table 1, Appendix A) is based on the premise that improvements in NLP techniques lead to superior performance of this category of techniques to assess novelty or similarity of patents, compared to classification- or citation-based approaches (Hain et al., 2021). Using only textual elements also allows for use-cases such as the second one, which is particularly important for academics, patent offices, and decision-makers on a country level.

For company-level decision making however, recent literature showed the importance of hybrid models, using more than one method to measure similarity between patents, for example IPC classifications with core terms from patent abstracts (Zhang et al., 2016), USPC classes with co-citation and backwards-citation (Kyebambe et al., 2017), or a combination of multiple categorial features (Hsu et al., 2020). While a general objective benchmark for patent similarity is still missing (Hain et al., 2021), it is argued that the approach to blend different methods reduces the bias that each single method introduces (Zhang et al., 2016). Within the IPA literature, besides measuring direct patent-to-patent (p2p) similarity, another successful example of applying a hybrid approach is predicting the quality, or value, of a patent: Combining bibliographic and textual features reduces of the overall mean average error (MAE), the key metric for prediction accuracy, compared to a range of the standalone predictors (Hsu et al., 2020).

Expanding the view towards commercialised research, those hybrid models are increasingly offered within a fast-growing industry of TI providers, as discussed with a tech start-up CTO who built such a model based on public patent data (personal communication, June 21, 2021), or confirmed by a business unit manager who buys such tools for his company (personal communication, September 8, 2021). Therefore, the following section will detail an approach to combine patent meta information (i.e., classification codes) and textual, semantic information (i.e., p2p cosine similarity scores), which can enable decision-makers on a company-level to retrieve relevant patent information for their business strategy.

3. Methodology

3.1 General Approach

3.1.1 Rational

Based on the research question of this thesis, a framework to construct a hybrid similarity model will be laid out, suitable to apply on any given USPTO patent portfolio. The core of the framework is based on recent deep learning-based approaches to measure p2p similarity by semantic means in the IPA literature (Arts et al., 2021; Hain et al., 2021; Whalen et al., 2020), but extending it for an important feature: using also structured (bibliographic) information as an additional metric, specifically the CPC classifications assigned to each patent.

As described earlier, recent work has shown improvement in model accuracy and further advantages like reducing bias by combining similarity approaches in this manner. In particular, in this way it is possible to combine advantages in NLP with the technological insights of an up-to-date classification system, reducing bias of each individual approach along the way, for example from relying too heavily on random semantic matches (Yang et al., 2018).

In order to construct a useful hybrid framework, shifting the focus to company decision-makers is important: instead of focusing on individual p2p similarity metrics, an aggregation of this data on assignee-level provides strategically relevant companies as a model output.

Due to the significant amount of computing resources and training time required for models from large-scale textual datasets, which can range from multiple days (Hain et al., 2021) to weeks (Whalen et al., 2020), to multiple months with trillions of entries (Younge & Kuhn, 2016), a practical framework has to reduce the scope of the analysis – concretely in three ways:

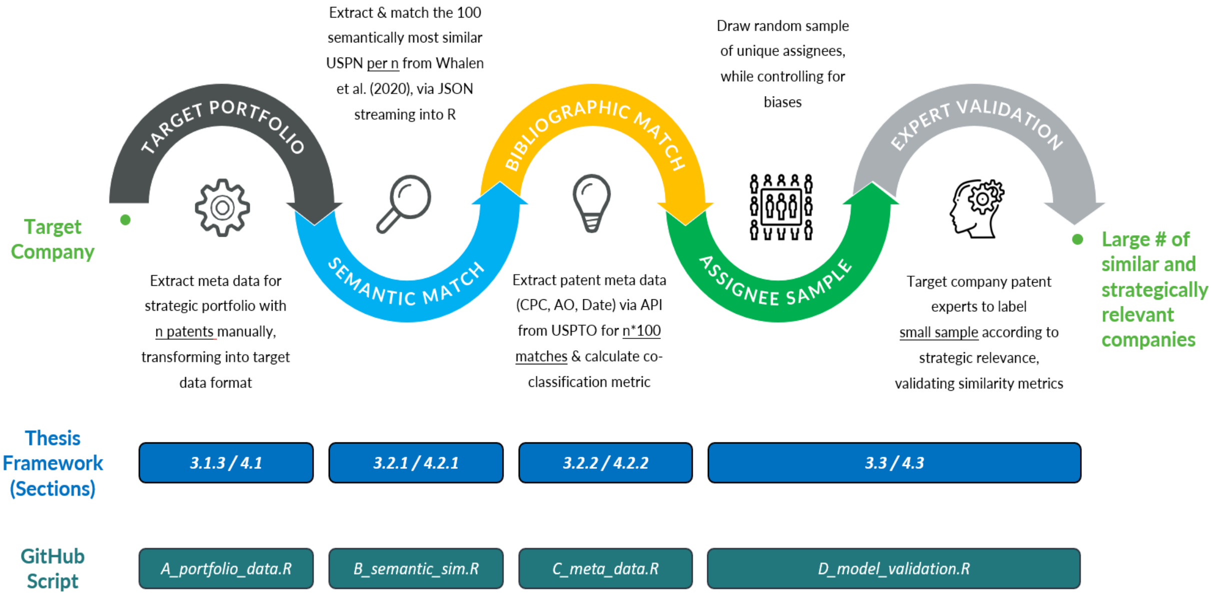
i. Instead of calculating new semantic similarity scores for patents, the approach leverages published databases of pre-calculated textual similarity metrics. Recently, two comprehensive, openly available sets have been made public, the first by Arts et al. (2018), including all US utility patents from 1970 to May 2018; the second by Whalen et al. (2020), covering 640 million similarity scores of US utility patents from 1976 to 2019. Both used p2p cosine distance as their similarity metric.

ii. Furthermore, "*n:n*" relationships, namely all possible p2p or portfolio-to-portfolio relations, which are important on a country-level to develop knowledge networks or to generate patent lanes (Niemann et al., 2017), are not the focus of this thesis. Instead, starting with a given patents in a company portfolio (“*1:n”* relationship), which not only reduces the computational load, but also aligns the outcome with the potential end-user of such a solution in a strategic company setting.

iii. For final validation of the outcome, expert assessment will be applied not on all of the identified similar patents, but on a random sample to avoid cognitive overload.

3.1.2 Procedure

The proposed framework follows loosely the de-facto standard procedure of IPA (Aristodemou & Tietze, 2018) as defined by Moerle et al. (2010), who applied business process modelling to make IP data useful in a business context. They differentiate the pre-processing, processing, and post-processing stages of patent data analytics. As shown in figure 1, the first step is to collect information on a target patent portfolio, for each of which the 100-most similar patents are extracted from a semantic similarity database in the second step. The third step involves the most data processing steps, enriching the semantically matched patents with meta data and constructing a second similarity measure based on CPC classifications. In the fourth step, a random sample of assignee organizations (AOs) is drawn, which are then validated with expert in the final step, based on their relevance to the initial target company. The outcome is a weighted list of AOs that are strategically relevant to the company. The analysis is performed in the programming language R and all code is available on GitHub.



Extract & match the 100 semantically most similar USPN per n from Whalen et al. (2020), via JSON streaming into R
Draw random sample of unique assignees, while controlling for biases
TARGET PORTFOLIO
BIBLIOGRAPHIC MATCH
EXPERT VALIDATION
Target Company
Large # of similar and strategically relevant companies
SEMANTIC MATCH
ASSIGNEE SAMPLE
Extract meta data for strategic portfolio with n patents manually, transforming into target data format
Extract patent meta data (CPC, AO, Date) via API from USPTO for n*100 matches & calculate co-classification metric
Target company patent experts to label small sample according to strategic relevance, validating similarity metrics
Thesis Framework (Sections)
3.1.3 / 4.1
3.2.1 / 4.2.1
3.2.2 / 4.2.2
3.3 / 4.3
GitHub Script
A_portfolio_data.R
B_semantic_sim.R
C_meta_data.R
D_model_validation.R

Figure 1 - Steps of constructing the hybrid similarity framework

3.1.3 Criteria for Target Patent Portfolio

In order to provide a useable and insightful framework or a given target portfolio, the respective company owning this portfolio should fulfill following criteria:

a. *Strategic IP assets:* Firstly, the need for TI must be clear to the company. A core part of this is based on the fact that they are using IPRs strategically, and not just patenting randomly over time. Someone in charge of market or patent strategy in the company would be a good indicator for this. Alternatively, the company is already using another tool to enable TI, from either patents or other market data.

b. *Company size*: While big corporates often have whole departments for TI (personal communication, November 11, 2020), small firms are often too resource-constrained to have a special focus on overarching strategic insight initiatives. In general, both categories of companies can benefit, but in particular the proposed framework is tailored to SMEs, which might have a small to medium sized patent portfolio, but not necessarily the right tools to make use of strategic insights based on those. Furthermore, there is a growing importance of IP assets for high-growth SMEs (EPO & EUIPO, 2021).

c. *US-focus:* Since the semantic similarity metrics used for the model are based on USPTO patents, the company must be active on this market, and possess a US-portfolio of patents. Furthermore, the US should play a strategic role within the overall company strategy, insofar that the TI gained is actually effective.

d. *Industry:* Especially companies in competitive industries that are undergoing change can benefit from patent-based insights (Aristodemou et al., 2017). However, the industry must not be too novel or too fast changing, otherwise the information found in patents will be outdated too soon (van Looy & Magerman, 2019).

3.2 Constructing the Hybrid Similarity Model

3.2.1 Semantic Similarity – Leveraging Open Data

Instead of taking the computational-heavy route of calculating new semantic similarity scores for any given patent pair, a key part of making this model easy to use, is to leverage recent research that published datasets of those scores, as a first step. Due to the more recent availability of patent data until end of 2019, the approach by Whalen et al. (2020) is preferred.

While not performed as part of the analysis per se, it is therefore critical to understand the steps undertaken by the researchers to arrive at the p2p similarity score presented, since those choices matter significantly for later model outcomes (van Looy & Magerman, 2019).

Whalen et al. (2020) aimed to provide an accessible resource for further research that provides insights on the similarity of inventions. They structure their approach analysing the patent text structure in three main steps that will be further described:

i. Downloading raw patent data and converting into a SQLite database (input corpus)
ii. Using the text as input for a deep learning-based distance model (“Doc2Vec”)
iii. Calculating a range of cosine similarity scores, in particular also the 100-most similar patents to each given one

Firstly, the required data was bulk-downloaded directly from the USPTO with a Python script, including all granted US patents from 1976 to 2019. Taking a couple of days of run-time, and consisting of multiple tables, another script combined all patent descriptions and claims into a joint database, using the open-source standard SQLite.

Secondly, the critical part of working with the large text corpus is based on an approach for vector space modelling, a popular technique within NPL (Yang et al., 2018). Simplified, it is possible to represent a corpus of *n* documents as a *n* x *m* matrix, with *m* representing each unique terms in those documents, resulting in a mostly sparse term frequency matrix. The more common terms two documents have, the more similar they are considered in terms of their vocabulary, measured by the cosine of each row *n*. This is the basis for simple analytic techniques like “term frequency-inverse document frequency” or slightly more advanced approaches like “Latent Dirichlet Allocation”, each assigning individual terms a specific weight or probability (C. Lee, 2021). Whalen et al. (2020) decided to use a neural network called “Doc2Vec” because of recently shown performance advantages for assessing technological similarity on large patent corpora (Helmers et al., 2019). This ML-based approach represents each patent as a 300-dimension vector as input.

Thirdly, based on this input, for any given patent the 100-most similar ones were identified, based on the cosine similarity as roughly described above. This results in a 21-gigabyte large

JSON file, openly available to download,[1] but requiring further pre-processing steps to be really useful. Most notably, no patent meta information is included in the set besides the USPNs. Also, due to the large file size, specific file processing steps have to be performed to work with it in any non-cloud, personal computer setting. Finally, in terms of validity it can be noted that outside of the sub-set used for this framework, the authors performed several tests, for example by comparing the average p2p similarity between non-citing patents (0.09) with backward- or forward-citing patents (0.26), validating their calculated cosine distance as a significant similarity metric.

3.2.2 Categorical Similarity – Bibliographic Meta Data

To map similar technological fields, and to facilitate prior art search, all patents are classified according to one or more hierarchical schemes: patent classification systems (EPO & USPTO, 2017). Based on the IPC, the CPC was agreed upon and jointly created by the EPO and the USPTO in 2010, officially launched in 2013, and its applicability expanded ever since (EPO, 2021). Its main difference to, and advantage over the IPC, especially in terms of TI, is the fact that CPC classifies according to all technical information available in a patent, not just via the published claims, which often leads to a large number of sub-classes per patent (Fierro, 2013).

To increase understandability of the many sub-classification layers of the scheme, I will refer to a class like “B30” as (3-digit) class, and more granular levels with their respective number of letters – the higher the number, the more granular the classification.

As to how exactly to measure the CPC class matching, I keep it simple in order to enhance the clarity of the model making it more understandable overall, and propose to measure the percentage of overlapping (matching) CPC classification for any two given individual patents – a co-classification measurement (Hain et al., 2021). This, however, has to be performed on a uniform level as per the schema, since on the lowest level they are too far spread out, and while insightful for themselves, barely comparable (X. Li et al., 2009). For example, the CPC lists over 250,000 (9-digit) subgroups on the most granular level (EPO & USPTO, 2017).

Depending on the heterogeneity of the target portfolio, this decision has to be taken individually, on a level that makes sense and yields balanced results. A metric for this can be

[1] From here: https://zenodo.org/record/3552078

a mean and/or median score of 0.5 for all CPC scores. More concretely, one also needs to differentiate in which ways classes are matched, depending on the total number of unique classifications per patent. This plays a role if a portfolio patent has four classes, of which are three matching (high score of 0.75), but the matched patent has a total of ten classes (indicating a low score of 0.3). In such cases is might makes sense to take the average of both, and in order to reduce bias for any given set, I suggest to calculate all three scores independently:

a. Percentage of portfolio CPC classes matched by the semantic similar patent
b. Percentage of sematic similar patent classes matched by the portfolio patent
c. The average of a. and b. to generate a balance

Besides the CPC classification, there are two more important categories of meta data needed: one is the priority or grant date of a patent. This has a variety of reasons, although mainly to be able to quickly sort and assess the most recent technological developments that might be more relevant from an ex-ante perspective of evaluation, today.

Concretely: companies are more interested in more recent patents, since they should include more recent technological developments that are of interest to them. The similarity matching algorithm described in section 3.2.1 will find the most similar patents across the whole range between 1970 and 2019. A useful cut-off date for relevance has to be set here, for example based on to the fundamental fact that patent validity is capped at 20 years after filing date (Squicciarini et al., 2013).

The other piece of bibliographic information needed is the assignee – a specific one in particular. Because from a strategic perspective, companies first and foremost care about other market participants, not about individual patents. While not all patents are assigned to a legal entity that owns the rights to commercialize it, those that are, are of special interest to a given target company, since those patent assignees are the either known, or previously unknown and potential future, competitors, applying similar technologies in their market offering. In this sense, the AO is the single most important data asset to retrieve, since it actually unlocks TI hidden in patents (An et al., 2021), and should therefore be the core of any result. This importance will be further highlighted in the following section 3.3.

One of the core problems when working with this piece of bibliographic meta data can be traced back to the method or way of the patent application process: any given legal or personal entity can apply for a patent, without a centrally coordinated unique identifier being assigned by the USPTO, resulting in disambiguation of assignees and authorship in general (G.-C. Li et al., 2014). This poses an ongoing challenge for IPA, with a variety of approaches being applied to solve it (e.g. Morrison et al., 2017), but for the scope of this thesis, will mainly rely on manual filtering.

Finally, as to how to retrieve the relevant meta information, the original patent offices are the most unbiased and reliable sources. Both the EPO and the USPTO are accessible via means of SQL queries (de Rassenfosse et al., 2014) APIs (Baker, 2021), or simply bulk downloads.

3.3 Model Validation

When it comes to validating a model that measures patent similarity in general, researchers today are still facing a problem: while a range of established performance measures for semantic similarity exist, Jaccard measure is one, the cosine similarity of two documents performed better on larger sets (Aristodemou & Tietze, 2018), there is no general, labelled benchmark dataset, as would be normally be usual for NLP tasks (Hain et al., 2021). The very domain-specific language used for patents is currently still assessed by patent experts as best practise, which is a very labour-intense process, especially when it comes to labelling large datasets of patens (Arts et al., 2018). This lack of ground-truth in labelled data for patent similarity is seen as a major hurdle in IPA (Aristodemou et al., 2017).

The proposed hybrid framework will follow the best practise in this regard, meaning an evaluation of the most relevant patents via a, or multiple, patent experts. This group of company-specific people would be involved into the approach in any way, since the target company in question must possess an at least sizeable patent portfolio. Furthermore, this approach eliminates the problem of having two separate evaluations performed for each of the two similarity dimensions. Instead, validation of the model should be performed by analysis of the importance of individual variables in the model. In the case of a simple regression, one can look at correlation metrics, for more advanced non-linear models, individual feature importance must be analysed (Hain et al., 2021). Throughout, the focus of the validation must be on reducing bias that might be introduced, quantitatively, qualitatively, or by design.

The most important area to consider here is the grouping by AO. As described earlier, an expert-based evaluation implies that individual patents will have to be grouped by their assignee at some stage of the analysis, which brings with it some advantages, but also certain disadvantages. Firstly, it is beneficial from a general interpretability perspective, since company names of potential competitors are simply more accessible for people, especially for strategically minded decision-makers.

Therefore, validation and labelling of data into "*relevant*" or "*irrelevant*" can be performed more accurately. Conversely, considering the fact that the similarity scores and matches are based on individual patents, the variance in matched patents per assignee becomes a critical factor. Assuming the matched number of individual AOs is too large to be evaluated by the patent experts in total, validation will be performed by taking a sample of the overall matched patents. Depending on the mapping from assignees to patents (and its variance), there are three options to evaluate an assignee based on its individual patent similarity scores:

a. Take all respective patents of this assignee into account
b. Take only a single respective patent per assignee
c. Take the average of the relevant scores across all patents

This will be decided based on the particular target portfolio, and its specific advantages and disadvantages will be discussed in section 5. Finally, from a statistical validity perspective, any random sample should fulfil the requirement of an equal representation of the relevant metrics, for example by having a similar mean and variance. Alternatively, a cluster sample could be taken (Krippendorff, 2018).

Once the to be-to-evaluated sample has been taken, company experts with knowledge of both, company strategy and IP strategy, will be asked to rank the perceived importance of each AO in an experimental setting that minimizes biases introduced by the interview and the data. For this, the experts should only be shown the name of the randomly selected patent assignee organisations. All other information has to be excluded, especially the dependent similarity metrics. To allow for potential look ups of those companies in internal systems, one USPN per AO is also shown, however also randomly selected. This should reduce the perception of relevance for AOs with high numbers of matched patents. For the ranking itself, a Likert-scale from 1-4 is proposed, representing numbered categories on a continuum where 1 = not relevant at all, and 4 = very relevant to the company. Due to the proposed small sample size and a single evaluator, using this scale that does not providing a neutral option, and therefore enforces a choice, allows for later binary coding of the results, into "*not relevant*" (1-2) or "*relevant*" (3-4) (Bailey et al., 1994), which will be evaluated alongside the raw ratings.

4. Case Study

4.1 Target Company

4.1.1 Company Background and Industry

The firm chosen for my case study is a Germany-based SME producing extrusion-based plastic materials for the automotive industry, Plastic AG. The company has been chosen for the following reasons, based on the requirements laid-out in section 3.1.3: *IP assets*, *size*, *US focus*, and *industry*.

a. The company owns a small 3-digit number of patent families and is still actively patenting.
b. It has circa 1300 employees globally, being significantly large enough to consider strategic management of those IP assets a factor for decision-making, but not too large to have a big department and sophisticated strategies for it.
c. Besides its home country, the US is a key market for the firm, having also 25 patents registered with the USPTO currently (October 2021).
d. Finally, the automotive industry is a complex, diverse, and very interesting market for innovation, which I will expand and elaborate upon a little.

While the industry is distributed globally, it is also separated into different tiers, with tier-1 suppliers directly selling to original equipment manufacturers (OEM) like Mercedes or Volkswagen, often as component suppliers (J. Lee & Berente, 2011). Tier-2 suppliers deliver to tier-1, etc. Within this system, the target firm is considered a tier-2 supplier, but supplies some products also directly (tier-1) to OEMs.

For those global clients, they offer extrusion-based plastic products for car and truck doors, as well as water tanks for a variety of vehicle types (personal communication, September 8, 2021)[2]. These factors become imported for research, innovation, and patenting activities, since this complex environment produces a double-edged sword effect for suppliers: on the one hand, product innovation projects are often done in cooperation across the tier level, since

[2] Over the last few weeks, I was in contact with a senior manager, and the patenting team of the SME, based in Germany.

there is a strong dependence along the value chain. On the other hand, there is also fierce competition between the different levels, with tier-1 suppliers competing directly with OEMs for core technologies in the components that they deliver (J. Lee & Berente, 2011).

Furthermore, apart from automotive, Plastic AG also focuses on the lighting industry, producing plastic profiles for industrial lamps. While this industry is irrelevant for the US market, it shows that their product portfolio is not fully homogenous. In order to classify their industry focus further, the North American Industry Classification System (NAICS) could be used in general, but the target firm is classified as 326199 "*All Other Plastics Product Manufacturing*", which is too broad an index to generate meaningful insights from.

Therefore, I will refer to the CPC classification of the patents to quantify the connectiveness of industries, is necessary. Finally, as per request of the company, its real name will be anonymized and no assigned USPN will be published, neither in the thesis, nor on GitHub[3].

4.1.2 Descriptive Analysis of Patent Portfolio

As the first step, I will take a close look at the US patent portfolio of the target firm, in particular the technology classification of its individual patents. In the case of Plastic AG, it contains a total of n = 14 utility patents granted by the USPTO, that were active in December 2019, the last month of similarity scoring data available for the analysis. All are either unique patents or representing a patent family solely, but never two from the same. They are either assigned to the target firm directly (n = 11), or to their relevant subsidiaries (n = 3). Since they are all equally relevant to the US market of Plastic AG, those 14 patents make up the target portfolio from here on. The relevant bibliographic meta data (*USPN, CPC class, CPC subclass, publication date, AO*) for those was extracted manually via a "*Lens.org*" query, and imported into R via comma-separated values (CSV) format.

Figure 2 below shows the split of the target portfolio according to CPC class. This was achieved by splitting the 14 patents into 31 individual CPC subclasses (4-digit) that are assigned to them, and then plot the results grouped by their CPC class (3-digit) with the R

[3] For the analysis only the real USPNs have been used, and the matches shown are the real ones, just without the target firm.

package “*treemap*” (Tennekes, 2021). An indexed portfolio overview with CPC split can also be found in Figure 4, Appendix C, also including the number of unique classes.

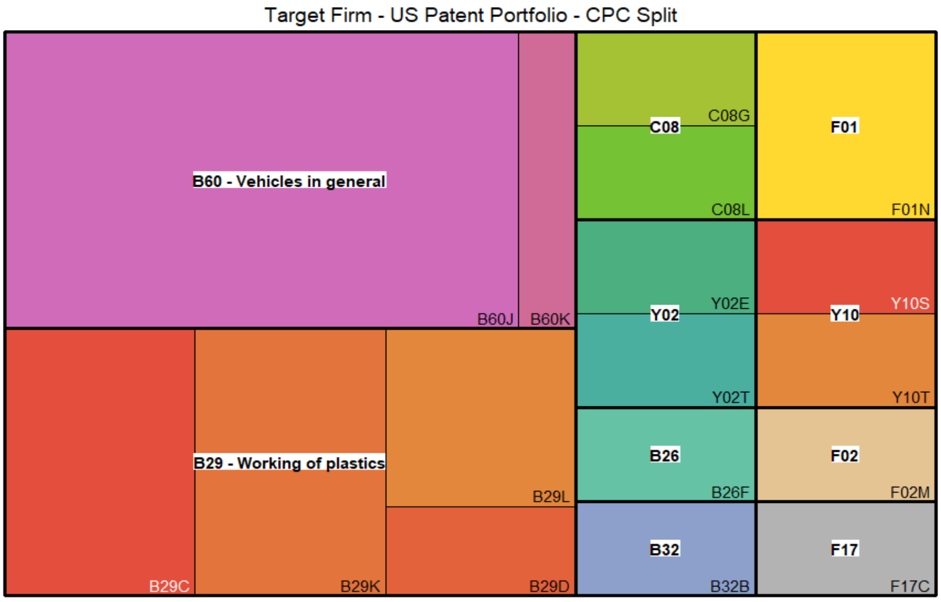


Figure 2 - Treemap with CPC split of target portfolio

This results in an average of 2.21 subclasses (1.74 classes) per portfolio patent with a minimum of 1 (1), a maximum of 7 (4), and median of 2 (1.5) classifications per patent. Uniquely, a total of 17 subclasses und 10 classes are represented in the portfolio, shown in Figure 4, Appendix C. Furthermore, Table 2 in Appendix C shows the definitions of CPC classes (3-digit) and subclasses (4-digit), according to the official naming scheme. Overall, it can be seen that the portfolio is neither very heterogenous, nor overly homogenous. It does, however, focus strongly on the automotive industry, and appliances of extrusion-based plastic materials within it, rather than representing a wide range of different technologies.

Regarding the options described in section 3.3 to use either classes or sub-classes, it can be seen in Table 2, Appendix C that especially for class B29 “*Working of Plastics*” on a subclass level, only B29C provides further relevant information, since B29K and B29L are both so-called orthogonal indexing codes (EPO & USPTO, 2017). In case those are matched (or not matched) by the model with another 4-digit class, it is not clear if it actually makes sense from a technology perspective, without further digging into the (8-digit) group level. Combined with the fact that the vast majority of class B60 consists of a single sub-class, it therefore seems logical to perform further analysis based on the class-level (3-digit) classification.

The respective code for these steps can be found in the script “*A_portfolio_data.R*” on GitHub.

4.2 Applying the Hybrid Similarity Model

4.2.1 Extracting Semantic Similarity Metrics

Starting point of constructing the proper analytical framework is the 21-gigabyte large semantical similarity JSON file from Whalen et al. (2020). In order to facilitate the insight generation process, I aim to extract the 100 most semantically similar USPN and their respective cosine similarity score, for each of the n = 14 target patents. The relevant database in JSON format has to be streamed into R, since its too large to keep in the working memory (R's standard way of computing) even for a high-end computer. The "*stream_in*" function of the "*jsonlite*" package (Ooms, 2014) can be used for this, enabling line-by-line execution of pre-defined functions for such large files via streaming of the data. Therefore, I design an empty *m* x *n* matrix in the right output format and fill it iteration by iteration via a loop that checks if the current USPN is of the 14 target ones, saved in a separate vector. With only a single computing core active, this step is by far the longest-lasting of the whole analyses[4], and parallelization (i.e., running calculations on multiple cores at the same time) is not supported by the package as of now. Once done, transforming the data from a nested list into a "*data.table*" format, using the equally named package, performs particularly fast for larger matrix operations (Dowle and Srinivasan, 2021).

The final *m* x *n* matrix results in four columns (*portfolio_id*, *patent_no, similar_no,* and *cosine)* with 1400 rows or observations (1400 x 4 matrix). Since each row represents one of the 100 most semantically similar patents to each of the firm's portfolio patents, there is a certain amount of overlap, i.e., patents that are similar to more than one of the portfolio patents. In order to enrich the matched patent numbers with the meta information as our target portfolio, I temporarily filtered for unique USPNs, resulting in 864 individual patents that have been semantically matched to the portfolio. This relatively large number (61.7% uniqueness) can be an indicator of heterogeneity of the target portfolio. Interestingly, within it, seven patents are similar to at least one of the other ones, and were matches themselves, while seven others are not among the top 100 most similar ones. Finally, the respective code for these steps can be found in the script "*B_semantic_sim.R*" on GitHub.

[4] The extraction function takes ca. 26 hours to run on a single core. The datafile has an estimated 16 million list items.

4.2.2 Applying Bibliographic Meta Data

After the lengthy step of extracting similarity measures from the provided database is done, for bibliographic information, i.e., date, CPC classes, and AOs, a different approach is required. To automatically enrich each of the 864 unique patents with the relevant meta information, the USPTO offers an application programming interface (API) to access data. This is computationally effective compared to downloading bulk patent data and extracting only some required information, and can easily be achieved in R through yet another package, "patentsview" (Baker, 2021).

First, three queries were made for each USPN, to retrieve the three required pieces of bibliographic information as a list, and afterwards a loop function iterated over each and combined them in a joint meta data matrix. During this process it also counted the distinct number of (3-digit) CPC classes of each patent, resulting in a 4 x 864 matrix, a snapshot of which is shown in Figure 7 in Appendix E. Some things to note here: there are both missing values (“*NAs*”), as well as multiple values for both, AOs and CPC classes.

While the former case does not require specific adjustments beyond excluding them from further analysis, the latter one requires some conditional processing of the data, i.e., summarizing these values in a single cell. Concretely, this means for AO that some patents have either multiple assignees, for example USPN "10309537", which is assigned to both “Kia Motors Corporation” and “Hyundai Motor Company”, or no AO at all, which includes a range of patents that are fully owned by their inventor, for example. For CPC, “*NAs*” were rare and can be explained for example by granted patents that were withdrawn, like USPN “10480924”. After filtering 39 “*NAs”*, the next step is to merge all three tables via a joint “key”:

a. Initial meta data of target portfolio: 14 x 7 matrix
b. 100 semantic matches per portfolio patent: 1400 x 4 matrix
c. Bibliographic data per unique patent (see above): 826 x 4 matrix

Since a “*merge*” function is R to similar to a SQL “*INNER-JOIN*” statement, only data that is matched against a “key” in both tables is kept, for the first merge of (a) and (b), this is the USPN of the target portfolio patent, and for the final merge with (c), it is the USPN of the semantic similarity match is. For this main table, some important decisions have to be taken here in order to allow for an outcome-focused (i.e., provide firms with TI) analysis:

- **Assignee organisation** (AO) – From here on, all data will be grouped by AO.

This is done in order to generate more practical, and a wider variety of market insights. Firstly, it is easier for humans to recognize and work with company names, which helps for data wrangling, but mainly for the expert validation. Furthermore, as a possible extension of this approach (that is not covered in this thesis), grouping by legal entity makes it is easier to enrich the model with further meta-information, which might not be linked to the patent number, for example press releases, 10-K reports, financial data, etc.

For our data, we find a total of 449 unique AOs with 825 patents matched to the target portfolio after grouping (1.84 patents per AO). "Uniqueness" in this sense is defined as being a unique string that R can recognize; this filter approach is not intelligent enough to differentiate multiple subsidiaries of the same corporate, or different legal entities of the same company.

- **Date** – Filter all patents that were granted before January 1st, 2000.

From a time-perspective, it is critical to keep up with recent technology changes, which includes that old patents should be excluded from the model. Since a patent duration is usually 20 years, and the similarity data is available until the end of 2019, setting the cut-off date for technologies at 01.01.2000 seems reasonable. This excludes 155 unique patents that were semantically matched, but granted before this date from further analysis (17.9%), leaving us with 671 patents from 340 assignees.

- **CPC similarity metrics** – Calculate three different measures (*portfolio, similar, average*), based on the total number of joint class matches.

Before we are able to calculate the similarity score of CPC similarity, we need to ungroup the both per patent, and per assignee grouped classes. Fortunately, I thought ahead and anticipated this potential problem earlier on in the code, and included created a solution to switch between a "wide" (i.e., grouped per AO), and a "long" (i.e., each CPC class has its own row) data format: since the separator used for grouping of CPC classes (";;") is the same for both level of grouping, the two functions "*separate_rows*" and "*unwrap_cols*" enable a fast switch between "long" and "wide". Now its possible to match each individual CPC class of target portfolio and its semantic matches against each other, and aggregating this sum per patent match. Following the instructions from section 3.2.2, I then calculate the three matching scores per assignee: percentage of portfolio CPC classes matched by the semantic similar patent (*portfolio*), percentage of sematic similar patent classes matched by the portfolio patent (*similar*), and the average of both (*average*).

- **Additional semantic similarity metrics** – Calculate two more metrics (*port_no_matched, sim_no_matched*), based on the total number of semantically matched patents, again from both sides.

To achieve these final metrics, I simply count the unique USPNs per AO, both for the initial portfolio, as well as for the semantic patent matches. This leaves us finally with a 341 x 16 matrix, which is the main outcome of the model. This outcome has to be validated first, before being able to make assumptions about its usefulness for TI. Also for this step, the respective code can be found in script "*C_meta_data.R*" on GitHub.

4.3 Validation of Results

As part of the expert assessment, a number of measures were taken to allow for an unbiased assessment of perceived strategic relevance to the company. First, quantitatively, random sampling was performed. From 340 unique assignees in our similarity dataset, a basic sampling algorithm in R selected n = 36 assignees (10.6%). The selected AOs hold a total of 95 unique, semantically similar patents, which result in 181 of the 1209 relevant patent matches (15.0%) – all patents from the year 2000 onwards. Since these ratios imply a significant higher average of patent matches per assignee (5.03 in sample versus 3.56), it is critical to eliminate outlier bias here.

Following the three laid-out options in section 3.3 to validate the patent-specific scores from the performed expert assessment on an assignee-level, I first look into the possibility to treat each evaluated assignee, as if their individual patents have been evaluated (i.e., "*option a*"). However, what was already found in section 4.1.2 for the whole dataset also holds true for the sample: its strongly skewed upwards, with 88 of the 181 (48.6%) patent matches belonging to the largest three assignees, 103 (56.9%) to the largest four, shown in figure 6, Appendix D. Because of this uneven split in matched patents, "*option b*" i.e., using only a single patent per assignee would be equally biased. Therefore, going forward with "*option c*" enables us to give assignees a more equal representation, however the main disadvantage is that the signal from the individual patents for highly matched assignees gets weaker, which could lead to: a large number of semantically similar patents of an AO match the CPC classes of a portfolio patent exactly, many others not at all, which will then even out and the signal is lost. With a large variance in matched patents per assignee present in the sample, this should still be the least harmful method. The mean for the "*average*" CPC metric, as seen in section 4.2.2, is 0.48 and the standard deviation 0.31, in line with an expected distribution for the whole dataset.

Finally, in terms of the interview set-up, the bias-reducing setup as described in section 3.2.2 was applied, and two company experts were selected on their ability to judge both, company strategy and patent strategy, having access to internal TI tools as well. They were shown the list of 36 matched AOs with reduced information, shown in Appendix E, and asked to evaluate each company on scale from one to four in terms of their relevance to Plastic AG. It was explained that relevance in this context means that companies are perceived as an existing competitor, or if they were flagged from previous market or patent insight tools. Part of the limited information shown was also a randomly selected USPN per assignee, which is useful to reduce outlier bias for AOs with up to 50 patent matches (Appendix D), independent of how the actual dependent metrics were calculated ("*option c*" above). Lastly, AOs were shown in the same format they were extracted, for example as double-assigned.

Having stated all these, the results look promising in a way that we see medium to strong correlation between our dependent variable ("*relevance to company"*) and the different independent variables of the model, shown in Appendix G. Surprisingly, at least initially, is that the highest correlation is not matched CPC classes, but for the total number of initially matched patents via the semantic similarity metric alone – both in terms of portfolio patents matched per assignee, as well as total number of assignee patents that were matched. Regression results, also shown in Appendix G, confirm this, showing a higher R^2 for the semantic matched model. Still, both key metrics are significantly correlated with the relevance.

To put these results into context: they are based on 36 randomly selected assignees, each of which consisting of a number of individual patents, that were evaluated by experts at Plastic AG in terms of their perceived strategic relevance to the company. It was possible to show that both similarity scores of an AO, semantic and categorical, are positively correlated with an increased importance of this particular assignee. The semantic similarity alone is more predictive of a higher importance, however. This will be further discussed following. Finally, as in previous sections, the respective code for these steps can be found in the script "*D_model_validation.R*" on GitHub. For replicability on a given portfolio, the code includes a "*set.seed*" function, which allows multiple random draws selecting the same sample.

5. Discussion

5.1 Evaluation of Findings

To start off the discussion, let me refer back to the beginning and restate the research question:

How to generate market insights from public patent data that are
(i) *relevant on a company-level*, (ii) *easy to apply and understand*, and finally
(iii*) able to reduce bias from individual analysis methods,*
in order to equip decision-makers with technology intelligence?

The previous validation showed the successful construction of an analytical framework to assess patent similarity, in a way that (i) useful insights are generated at a company-level, that (ii) it can be applied to various patent portfolios with the provided code on GitHub and laid-out methodology, using methods that are clearly explained, and finally in a way that (iii) incorporates two distinct measures of patent similarity, which combined do not eliminate all bias, but at least combat the reliance on a single metric In this short synopsis, all three research goals that have been set out were achieved, showing positive correlation of both similarity metrics with the perceived usefulness to a target company, indicating that H1 holds true.

On a second note, however, it was not possible to validate H2, which hypothesized based on the literature of disruptive innovation that in particular patents (and assignees) with a medium level of CPC similarity are of particular interest to the target firm, given a high semantic similarity. This is due to a range of factors that will be further discussed in more depth, but mainly due to the small sample size of n = 36 randomly selected assignees that have been evaluated per the laid-out methodology prohibited the fitting of relevant non-linear models that would have been needed to prove a non-linear relationship between the CPC similarity and the relevance for a given firm.

The decision to work on an assignee-level had practical reasons, specifically to achieve research questions (i) and (ii). In an optimized setting where patents would have been evaluated on an individual basis, and with a significant deeper involvement of patent experts at a target company, it would have been possible to set up a more stringent model validation method: having more labelled data available would allow a split in a training and testing set for non-linear predictive models like support vector machines (X. Li et al., 2009), latent dirichlet allocation (Aristodemou & Tietze, 2018), or even a deep neural network setup (Hsu et al., 2020; Krestel et al., 2021; Yang et al., 2018). Especially the latter IPA category made good progress with semi-supervised approaches, requiring only partially labelled data. This was outside the scope of the thesis however, and therefore the hybrid similarity framework can serve as a practical blueprint for academics, as well as company decision-makers.

5.2 Limitations

5.2.1 Validation and Methodological Setup

Overall, the proposed method does not consist of overly complicated models, rather that it largely leverages recent research. In one aspect this makes it comprehensible for people without profound knowledge in either legal or statistical matters, at least to a certain degree. In another aspect this simplified approach did not deliver an accurate quantification of the model's accuracy. First and foremost, a bigger sample size of evaluated assignees would have been helpful to get deeper statistical insights, in particular the potential to fit a non-linear model to validate the second hypothesis.

Because grouping on assignee-level was necessary to account for the variance in total number of matched patents, the approach lost signal in the data, especially for those companies with a large number of matches. As shortly discussed in 4.3 already, I tried to assess to the best of my knowledge and capabilities that the advantage of having a less-skewed patent sample where the top-four assignees (11% of 36) make up 57% of the matched patents prevails the alternative set-up in which each patent would count individually. Since in fact three out of these four assignees were ranked 3 or 4 on the relevance scale during validation, the analysis would have looked completely different, and would have simply confirmed that the CPC match is significantly predictive of relevance for the target company.

Finally, in terms of the overall methodological set-up, a broader categorization of sample patents could have been performed. This means specifically in regards to the 2 x 3 matrix in Appendix C to also include non-semantically similar patents into the expert evaluation, in order to have a more robust experimental set-up. But even in this way it would be impossible to cover the potential blind spots of high categorical similarity, but no semantic one. Therefore, the model results cannot be seen as exhaustive, also considering the following reasons below.

5.2.2 Patent Similarity Measures

On a technical level, several limitations exist for the categorical and semantic similarity measures. When judging the CPC classification index used, four points have to be noted. Firstly, a decision that was taken early on, while analysing the target portfolio, was to focus on the rather high-level (3-digit) class, which led to a higher matching percentage with the semantically similar patents. It would be interesting to see if a more granular classification has a higher correlation, or better predictive performance for target firm relevance.

The relevant data up to (8-digit) subgroup level can be pulled via the same API used in script "C". The matching process would take slightly longer, and data processing steps would be more cumbersome, but especially for very homogenous portfolios this might yield more insightful results.

Secondly, besides CPC also other classification schemes could be used, especially when focusing only on a specific region like the US. Thirdly, when comparing classifications in general, it could be an option to use more a more sophisticated measure of similarity than the average of joint classes, for example based on the actual technological closeness between these classes. A different direction across these three areas can be relevant for further research.

Finally, a more general, and often-cited downside of categorical measures is based on the fact that they get assigned by a central institute (the patent office), which relies on a centrally-updated index. This index however cannot be updated, meaning having to introduce new classes and shift existing ones within the system, at the same speed to keep up with the speed of technological progress – a problem that has been shown especially for newly evolving industries like bioinformatics (van Looy & Magerman, 2019).

In fact, there is research that uses the creation of new classification sub-classes as a proxy for emerging technologies, showing that emerging patent clusters can be identified ex post in this way (Kyebambe et al., 2017). Patent offices have to anticipate the rapid change obviously, and they try to keep the classification indexes constantly up-to-date (EPO, 2021), but in particular for newly granted patents, using novel technologies, it is not completely accurate.

Coming to the other axis of the graph, textual similarity. My proposed hybrid approach circumvented two often-cited downsides of semantic measures quite nicely, the first one being the size of the patent portfolio: semantic measures alone don't work well for small portfolios, since the model needs large amounts of input data to be trained and finetuned (Zhang et al., 2016). Since the measures were calculated p2p on the whole range of USPTO data from 1976 to 2019, the sample was more than sufficiently large. The other prevented downside is even more obvious: instead of performing the computing-intense exercise ourselves, the method leverages recently published data to extract the relevant scores, which is significantly faster. Being able to parallelize the JSON extraction step, for example via specific Python packages, or on a Linux-based operating system that make use of a different way of reading and indexing files, this process of streaming the 21-gigabyte file (taking ca. 24-30 hours) should be able to be accelerated by a factor of ten.

At the same time, this is also the largest weakness of the method: the limited availability of similarity data until the end of 2019. In case of Plastic AG for example, the company got three new patents granted after this cut-off date, which were included in the analysis, since the relevant similarity score are missing. While it would be possible to re-calculate all similarity scores regularly, significant computing time and cost would have to be taken into account, which would stand against the main value of this approach: easy to apply.

Overall, as for the CPC index, there are a lot of different methods available to assess semantic similarity. The neat feature of the proposed hybrid approach is that it is modular in general, meaning further research can simply use another basis of similarity calculations at basis for the matching, especially one that calculated scores with more recent, up-to-date patents. Equally worthwhile thinking about would be an extension of the model to include also citations, the third category of similarity measure techniques. Semantically matched patents could then be compared with the citations provided within a target portfolio in an extra step.

Finally, the validation showed that the total number of semantically matched patents has the highest correlation with perceived relevance for the company, which is first of all an acknowledgement of the method used by Whalen et al. (2020) for their similarity database. To assess a causal relationship however, one should compare relative numbers here, since it is straightforward to assume that a very large company that is at least a little similar to the target company, has some of its patents matched semantically.

To generate a more meaningful similarity metric on an assignee-level, there is one data point missing: the total patent portfolio size of this given assignee. If, for example, a company's total US patent portfolio is 1000 patents strong, of which 50 are matched, is this company more, or less relevant as one with 10 patents of which 5 are matched? With this information available, it would be easy to calculate a relative importance score, based on total number of matched patents. This can be a very worthwhile extension of the approach for further research.

5.2.3 Strategic Technology Intelligence

Since a large part of the initial research question was focused on applicability for, and usability by business practitioners, it is important to evaluate the strategic insights, or TI that can actually be provided. First of all, from a time perspective, a major shortcoming has already been mentioned: semantic similarity scores that end in December 2019. This is contrary to the need of having up-to-date information on the one hand, but not a KO-criteria, especially considering the early stage of the product development where patents play a main role.

I would argue that companies who apply the proposed method in 2021 or the near future, are still able to flag important, strategically relevant AOs for their organisation, simply because of the fact that patents are normally valid for 20 years and are granted during early stages, even before market entry for example (Woo et al., 2018).

Likewise, above are the benefits of a modular approach described in which a different dataset of similarity measures can be used within the existing framework. The other perspective of time in the particular dataset is also relevant: having patents from 1970 might add performance to the model by Whalen et al. (2020), but take unnecessary time within the proposed approach to generate insights, since we filter out patents before 2000. Within the overall picture it is still a good trade-off, but further research or even company practitioners might use a more fitting semantic database as a matching basis.

Another strategic consideration is the focus on the US market. Firstly, as a single market, this focus leaves out out many global innovation leaders in relevant fields who have not patented in the market at all (Morrison et al., 2017). This critique could however be generalized to all companies that have not filed patents, or those that patent only certain, but not every innovation out of strategic considerations. Within the proposed approach this makes rather little sense, since it specifically states to be “one tool in the toolbox of TI”, and its advantages lie in fast applicability, not in complete comprehensiveness.

A more detailed aspect of the general USPTO patenting system is however interesting to discuss in this regard: the role of software patents and handling of computer programs as IP –

especially its difference to the EPO regularly, or to Europe culturally (van Looy & Magerman, 2019).

Since the case study firm is based in Germany, as is its acting patenting department, one could argue that the firm might have a historically different approach towards patenting software-related patents compared to a US firm, especially for patent families where the patents are split globally, but the invention and core claims are the same. Therefore, this basic differentiation between physical and software product might not be respected here. A more throughout analysis of CPC sub-classifications could give an answer here.

The final point to discuss is regarding IP as a market insight tool in general. Firstly, the already mentioned ambiguity of naming and attribution limits the practicality of any framework that relies on unique organization names, or in other words: assigning the technological potential to real competitors. While it is not finally solved, some promising and creative approaches have been made (G.-C. Li et al., 2014; Morrison et al., 2017), which should be further applied to the outcome of this framework as well.

Furthermore, while the framework used a sub-class of IP, utility patents, a range of other IP assets are also freely available to research, i.e., design patents, copyright, and trademark data, especially since the literature on these is growing fast (Kim et al., 2021; C. Lee, 2021). For long-term successful TI an integrated approach is needed, not only focusing on data that is easily available, but rather covering a broad range of innovation fields.

This idea can be expanded beyond IPRs as well: if data is grouped by a (nearly) unique assignee, why not add further meta data outside of the patent ecosystem? Financial data, press releases, and market reports are just some of the example data categories that can be linked to an individual assignee to further enrich the model with relevant insights, and bring any company a significant step closer to achieving an integrated TI strategy.

6. Concluding thoughts

This thesis set out to use information from patent data for insights into the market and its competitors. By setting up a hybrid framework of semantic and categorical similarity, I showed that, albeit less significant than my initial hypotheses were proposing, a clean and comprehensible way to assess patent similarity. I took into account various biases and pitfalls that are present when working with large amounts of data on a granular level, for example by adjusting for patents per assignee or in the way the validation was set-up: randomly and bias-conscious. Still, it would be a far stretch to claim a perfect, or overly robust approach – various limitations are still present at the current approach and it was not possible to assess potential non-linear relations in the data.

However overall, the presented approach was designed stringently to the best of my knowledge in statistics, data analytics and IPRs. From setting out to validation, I kept the initial research question as top priority and created a hybrid similarity framework that generates practical insights from freely available patent data. Especially compared with expensive market insight and TI tool on the market, this approach may be very useful for a range of company decision makers, especially in industries where working with IPRs is relevant. This should be an important goal to keep in mind and advice for all researchers on business-focused use cases: keep your (potential) end-users in mind.

In terms of further literature implications, I discussed a range of potential extensions of the model, as well as its limitations in section 5.2, both of which can be complemented by future research. Of particular interest would be two topics:

a. Firstly, a larger-scale assessment of the available data, in particular by labelling (assessing) a higher number of assignees or individual patents, in order to properly investigate a potential non-linear relationship between the CPC similarity and the relevance, at high semantic similarity.
b. Secondly, the replacement of the semantic similarity scores used with a more up-to-date version, proving a true modular approach of the framework.

Taking all the results and discussion into account, I can conclude by saying that in order to succeed with ever-increasing technological progress, companies require capabilities for technology intelligence. My contribution to this growing challenge is a framework to measure and leverage patent similarity data that is easy to apply and understand. While it cannot, and was never intended to be, a one-size-fits-all solution, I see it as another useful technique in a toolbox of working with patent data and market insights in general, both for the growing IPA literature community as well as for business practitioners.

Finally, within the overall theme of technological progress, I can note that IPRs continue to play an important role, not only to safeguard innovation, but also to generate insights about it. An ever-increasing trend from physical towards digital tools and business model, the legacy global patent system must continue to constantly challenge and re-invent itself to keep up with the pace of change – applying machine or deep learning powered approaches themselves might be a great start. Still, the question remains to be answered what role patents will play for platform-based or API-first software companies that rely heavily on the integration of external solutions. If patent offices want to fulfil on their role to foster innovation, they must remove roadblocks for those companies, while allowing space and time for physical-first companies to adjust their business model to the "*new digital normal*". At the same time, patent systems must stay robust in times of crisis to allow the fast pace of innovation that we saw with developing and producing the COVID-19 vaccines.

For every affected company, an intelligent and scalable strategy for navigating increasingly complex markets, with increasing amounts of data available, is now needed more than ever.

Appendix

A. Literature Overview for Patent Similarity Analysis

Patent Information Category	**Analytic approaches**	**Short description**	**Advantages**	**Drawbacks**	**Literature Examples**
A. **Classifications**	Co-classification, Classification overlap	Distance measure of technological fields by analysing the co-occurrence of patent classes according to IPC, CPC or USPC	+ Easily applicable technological insights + Easy access and structured data base	- Comparison only between predefined classes - Classification systems can be either too narrow or too coarse - Technological change requires frequent updates	Harris et al. (2010); Yan & Luo (2017)
B. **Citations**	Co-citation, bibliographic coupling, direct/indirect citation	Analysing common citation pathways (backward, forward), also indirectly	+ Established in the literature + Easy access and structured data base	- Forward citations suffer from time delay, not most recent patents usable - Subjectivity problem, e.g., home bias and strategic citations - No full comprehensibility given	X. Li et al. (2009); Rodriguez et al. (2015)
C. **Text** (e.g., title, abstract, claims)	Keyword-based, SAO-based, Ontology-based, ML-based (NLP)	Comparison of textual features like frequently used or common words, sentence structure, or semantic meaning	+ Content and meaning of patents + Access to semantic information	- Large textual data bases; computational expensive - Context-sensitivity and specifics of the legal language - Advanced approaches not straight-forward to apply	Arts et al. (2021); Hain et al. (2021)

Table 1 – Classification of patent similarity literature by category

B. Reasoning for a Hybrid Similarity Approach

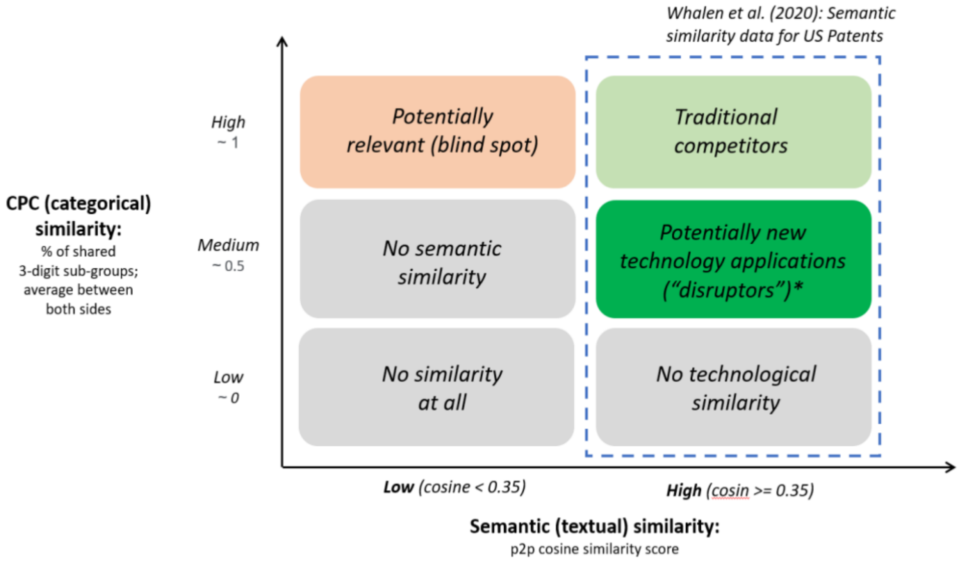
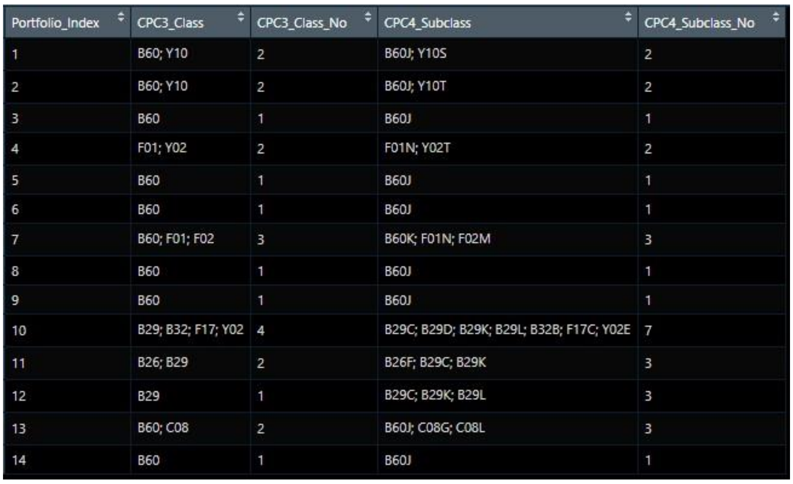

Whalen et al. (2020): Semantic similarity data for US Patents
CPC (categorical) similarity:
% of shared 3-digit sub-groups; average between both sides
High ~ 1
Medium ~ 0.5
Low ~ 0
Potentially relevant (blind spot)
Traditional competitors
No semantic similarity
Potentially new technology applications ("disruptors")*
No similarity at all
No technological similarity
Low (cosine < 0.35)
High (cosin >= 0.35)
Semantic (textual) similarity:
p2p cosine similarity score

Figure 3 - Expected outcome for patent-to-patent similarity in the hybrid model

**Note:* Literature does not imply that disruption is coming from specifically this category; the radical nature of disruption implies that it can create whole new industries. H2 theorizes that the impact of something like this happening would affect the target most in this "somewhat" similar market setting.

C. CPC Classifications of Target Portfolio



Portfolio_Index	CPC3_Class	CPC3_Class_No	CPC4_Subclass	CPC4_Subclass_No
1	B60; Y10	2	B60J; Y10S	2
2	B60; Y10	2	B60J; Y10T	2
3	B60	1	B60J	1
4	F01; Y02	2	F01N; Y02T	2
5	B60	1	B60J	1
6	B60	1	B60J	1
7	B60; F01; F02	3	B60K; F01N; F02M	3
8	B60	1	B60J	1
9	B60	1	B60J	1
10	B29; B32; F17; Y02	4	B29C; B29D; B29K; B29L; B32B; F17C; Y02E	7
11	B26; B29	2	B26F; B29C; B29K	3
12	B29	1	B29C; B29K; B29L	3
13	B60; C08	2	B60J; C08G; C08L	3
14	B60	1	B60J	1

Figure 4 - Split of CPC classifications in target portfolio (anonymized)

Note: The 3. and 5. column show the sum of unique CPC classes and subclasses.

Source: Screenshot from R analysis, script: "A_portfolio_data.R"

CPC Class	**Class Name**	**CPC Sub-class** (*Selected*)	**Subclass Name**
B29	WORKING OF PLASTICS; WORKING OF SUBSTANCES IN A PLASTIC STATE IN GENERAL	B29C	SHAPING OR JOINING OF PLASTICS; SHAPING OF MATERIAL IN A PLASTIC STATE, NOT OTHERWISE PROVIDED FOR; AFTER-TREATMENT OF THE SHAPED PRODUCTS, e.g. REPAIRING
		B29K	INDEXING SCHEME ASSOCIATED WITH SUBCLASSES B29B, B29C OR B29D, RELATING TO MOULDING MATERIALS OR TO MATERIALS FOR {MOULDS, } REINFORCEMENTS, FILLERS OR PREFORMED PARTS
		B29L	INDEXING SCHEME ASSOCIATED WITH SUBCLASS B29C, RELATING TO PARTICULAR ARTICLES
B60	VEHICLES IN GENERAL	B60J	WINDOWS, WINDSCREENS, NON-FIXED ROOFS, DOORS, OR SIMILAR DEVICES FOR VEHICLES; REMOVABLE EXTERNAL PROTECTIVE COVERINGS SPECIALLY ADAPTED FOR VEHICLES
C08	ORGANIC MACROMOLECULAR COMPOUNDS; THEIR PREPARATION OR CHEMICAL WORKING-UP; COMPOSITIONS BASED THEREON	G08G	MACROMOLECULAR COMPOUNDS OBTAINED OTHERWISE THAN BY REACTIONS ONLY INVOLVING UNSATURATED CARBON-TO-CARBON BONDS
F01	MACHINES OR ENGINES IN GENERAL; ENGINE PLANTS IN GENERAL; STEAM ENGINES	F01N	GAS-FLOW SILENCERS OR EXHAUST APPARATUS FOR MACHINES OR ENGINES IN GENERAL; GAS-FLOW SILENCERS OR EXHAUST APPARATUS FOR INTERNAL COMBUSTION ENGINES
Y02	TECHNOLOGIES OR APPLICATIONS FOR MITIGATION OR ADAPTATION AGAINST CLIMATE CHANGE	Y02T	CLIMATE CHANGE MITIGATION TECHNOLOGIES RELATED TO TRANSPORTATION
Y10	TECHNICAL SUBJECTS COVERED BY FORMER USPC	Y10T	TECHNICAL SUBJECTS COVERED BY FORMER US CLASSIFICATION

Table 2 - CPC names of relevant portolio classes and subclasses

Source: *www.uspto.gov*

D. Outlier Detection: Patents per Assignee Matched

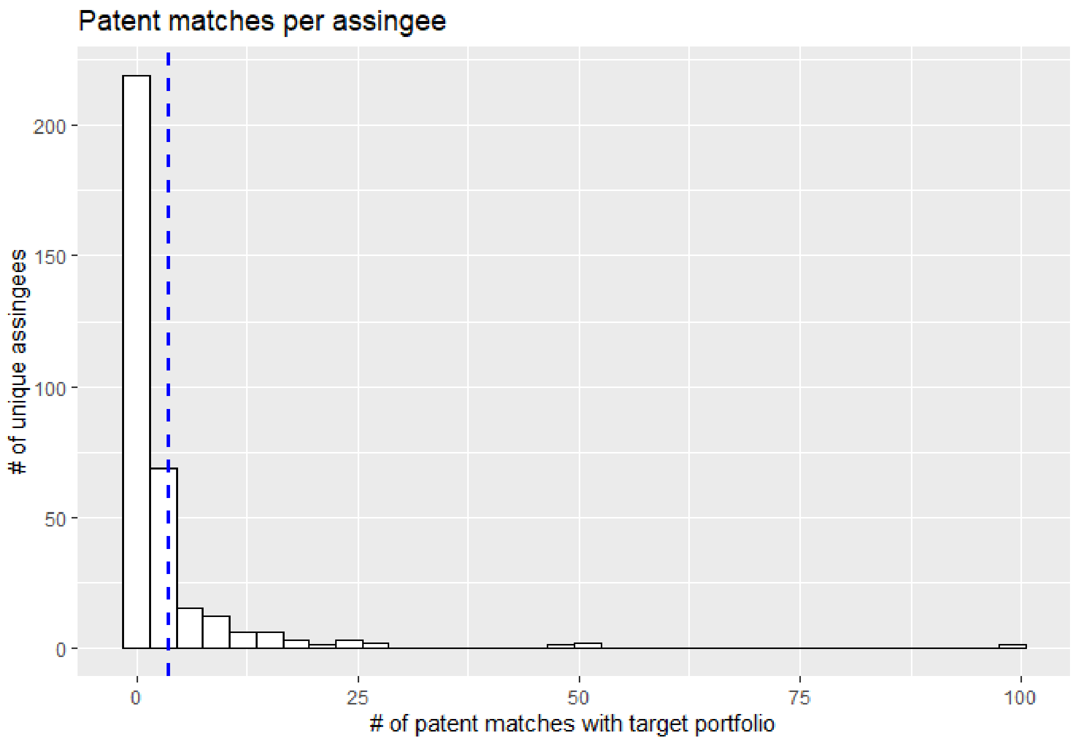
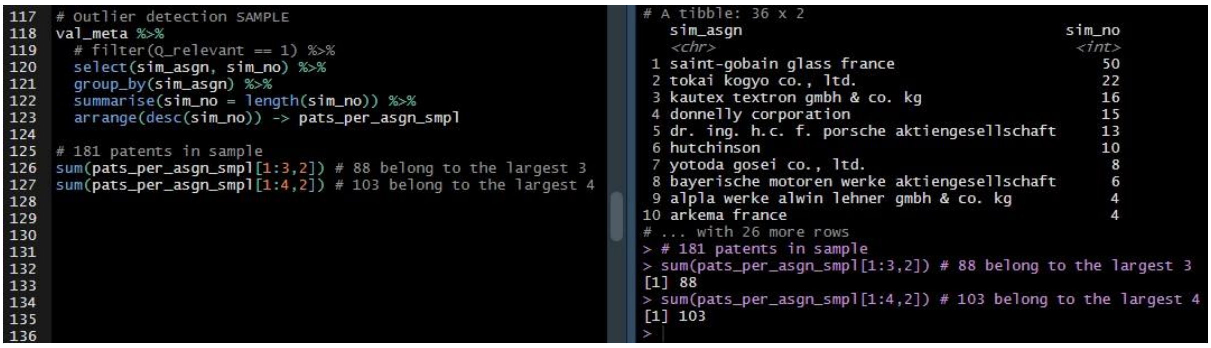


Figure 5 - Histogram of patent matches per assignee

Note: 340 unique assignees organisations.

Source: Plotted from R analysis, script: "*C_meta_data*.R"



```
# Outlier detection SAMPLE
val_meta %>%
  # filter(Q_relevant == 1) %>%
  select(sim_asgn, sim_no) %>%
  group_by(sim_asgn) %>%
  summarise(sim_no = length(sim_no)) %>%
  arrange(desc(sim_no)) -> pats_per_asgn_smpl

# 181 patents in sample
sum(pats_per_asgn_smpl[1:3,2]) # 88 belong to the largest 3
sum(pats_per_asgn_smpl[1:4,2]) # 103 belong to the largest 4
```

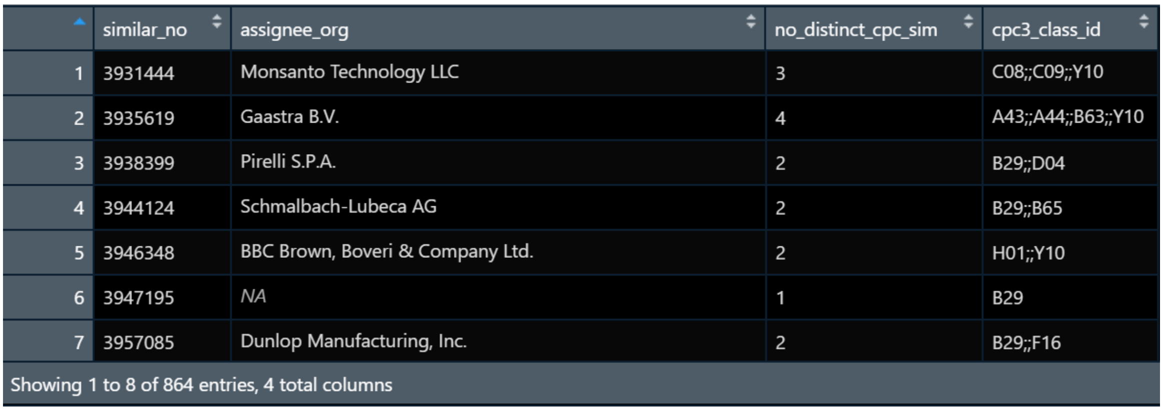
```
# A tibble: 36 x 2
   sim_asgn                                       sim_no
   <chr>                                           <int>
 1 saint-gobain glass france                          50
 2 tokai kogyo co., ltd.                              22
 3 kautex textron gmbh & co. kg                       16
 4 donnelly corporation                               15
 5 dr. ing. h.c. f. porsche aktiengesellschaft        13
 6 hutchinson                                         10
 7 yotoda gosei co., ltd.                              8
 8 bayerische motoren werke aktiengesellschaft         6
 9 alpla werke alwin lehner gmbh & co. kg              4
10 arkema france                                       4
# ... with 26 more rows
> # 181 patents in sample
> sum(pats_per_asgn_smpl[1:3,2]) # 88 belong to the largest 3
[1] 88
> sum(pats_per_asgn_smpl[1:4,2]) # 103 belong to the largest 4
[1] 103
> 
```

Figure 6 - Analysis of patent matches per assignee in sample

Note: Table shows a total of 36 unique assignee organisations.

Source: Screenshot from R console and analysis sheet, script: "*D_model_validation*.R"

E. Data Transformation Steps



	similar_no	assignee_org	no_distinct_cpc_sim	cpc3_class_id
1	3931444	Monsanto Technology LLC	3	C08;;C09;;Y10
2	3935619	Gaastra B.V.	4	A43;;A44;;B63;;Y10
3	3938399	Pirelli S.P.A.	2	B29;;D04
4	3944124	Schmalbach-Lubeca AG	2	B29;;B65
5	3946348	BBC Brown, Boveri & Company Ltd.	2	H01;;Y10
6	3947195	*NA*	1	B29
7	3957085	Dunlop Manufacturing, Inc.	2	B29;;F16

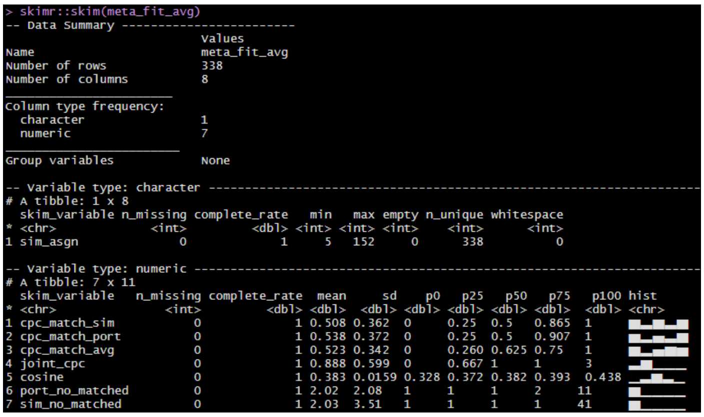
Showing 1 to 8 of 864 entries, 4 total columns

Figure 7 - Unique patent matches enriched with meta information

Note: Table shows in total 864 individual USPNs before filtering for “*NAs*”, like in row six

Source: Screenshot from R analysis, script: “*C_meta_data*.R”

F. Validation Sample of Assingees



```
> skimr::skim(meta_fit_avg)
-- Data Summary ------------------------
                           Values
Name                       meta_fit_avg
Number of rows             338
Number of columns          8
_______________________
Column type frequency:
  character                1
  numeric                  7
________________________
Group variables            None

-- Variable type: character --------------------------------------------------------------------
# A tibble: 1 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>             <int>         <dbl> <int> <int> <int>    <int>      <int>
1 sim_asgn              0             1     5   152     0      338          0

-- Variable type: numeric ----------------------------------------------------------------------
# A tibble: 7 x 11
  skim_variable   n_missing complete_rate  mean     sd    p0   p25   p50   p75   p100 hist
* <chr>               <int>         <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <chr>
1 cpc_match_sim           0             1 0.508 0.362  0     0.25  0.5   0.865 1      ▇▂▇▂▇
2 cpc_match_port          0             1 0.538 0.372  0     0.25  0.5   0.907 1      ▇▂▅▂▇
3 cpc_match_avg           0             1 0.523 0.342  0     0.260 0.625 0.75  1      ▇▂▅▇▇
4 joint_cpc               0             1 0.888 0.599  0     0.667 1     1     3      ▃▇▁▁▁
5 cosine                  0             1 0.383 0.0159 0.328 0.372 0.382 0.393  0.438 ▁▂▇▂▁
6 port_no_matched         0             1 2.02  2.08   1     1     1     2     11     ▇▁▁▁▁
7 sim_no_matched          0             1 2.03  3.51   1     1     1     1     41     ▇▁▁▁▁
```

Figure 8 - Analysis of patent matches per assignee in sample

Note: 36 unique assignee organisations.

Source: Screenshot from R analysis, script: “*D_model_validation.R*”

Assignee Organisation	**Example USPN**	**Rele-vance Rank**	**Relation**	**CPC Match “Avg.”**	**Semantic Matches Portfolio**
ALPLA WERKE ALWIN LEHNER GMBH & CO. KG	10093472	1		0.51	2
Arkema France	10040889	3	Existing Competitior	0.52	2
Bayerische Motoren Werke Aktiengesellschaft	10260678	3	Potential Competitor	0.52	3
BREVETTI ANGELA S.R.L.	10315788	1		0.56	2
BUNDESDRUCKEREI GMBH	10255515	1		0.00	1
Coloplast A/S	10105254	1		0.75	1
COMMISSARIAT À L'ÉNERGIE ATOMIQUE ET AUX ÉNERGIES	10044068	4	IP Lawsuit	0.38	1
COMPAGNIE GENERALE DES ETABLISSEMENTS MICHELIN	10323118	1		0.75	1
CONTINENTAL DENTAL CERAMICS, INC.	10182895	1		0.00	1
Creative Balloons GmbH	10456953	1		0.75	1
Discma AG	10350815	1		0.75	1
DONNELLY CORPORATION	6691464	2	Potential Competitor	0.80	8
Dr. Ing. h.c. F. Porsche Aktiengesellschaft	10232889	3	Potential Competitor	0.62	9
ENGEL AUSTRIA GMBH	10293549	3	IP Monitoring	0.75	1
Evonik Roehm GmbH	10207435	2	IP Monitoring	0.29	2
FESTO SE & CO. KG	10316983	1		0.00	1
FUNDACIÓN TECNALIA RESEARCH & INNOVATION	10179429	1		0.63	1
Heraeus Noblelight America LLC	10324232	1		0.42	1
Hutchinson	9096114	4	Existing Competitior	0.89	8

KAUTEX TEXTRON GmbH & Co. KG	10000003	3	Existing Competitior	0.56	3
KENNAMETAL INC.	10300537	1		0.00	1
KRONES AG	10279939	2	IP Monitoring	0.42	1
Marbleous World B.V.	6592706	1		0.00	1
MERCK PATENT GMBH	10279520	1		0.50	1
Muehlemann IP GmbH	10315830	1		0.00	1
OSRAM GMBH	9447943	3	IP Monitoring	0.63	1
OSRAM Opto Semiconductors GmbH	10290784	1		0.00	1
PHP FIBERS GMBH	10265885	1		0.75	1
SAFRAN AERO BOOSTERS SA	10245766	1		0.48	2
SAINT-GOBAIN GLASS FRANCE	9694659	4	Existing Competitior	0.78	9
SELLE ROYAL S.P.A.	6136426	1		0.00	1
thyssenkrupp AG;;ThyssenKrupp Federo und Stabilisatoren GmbH	10479031	1		0.63	1
TOKAI KOGYO CO., LTD.	9327585	4	Existing Competitior	0.92	8
TOYOTA BOSHOKU KABUSHIKI KAISHA	9132717	3	Potential Competitor	1.00	1
VOLKSWAGEN AKTIENGESELLSCHAFT	10256703	3	Potential Competitor	0.25	2
Yotoda Gosei Co., Ltd.	6679003	1		0.94	8

Table 3 - Sample assignees and relevance ranking

G. Correlation of Model Variables

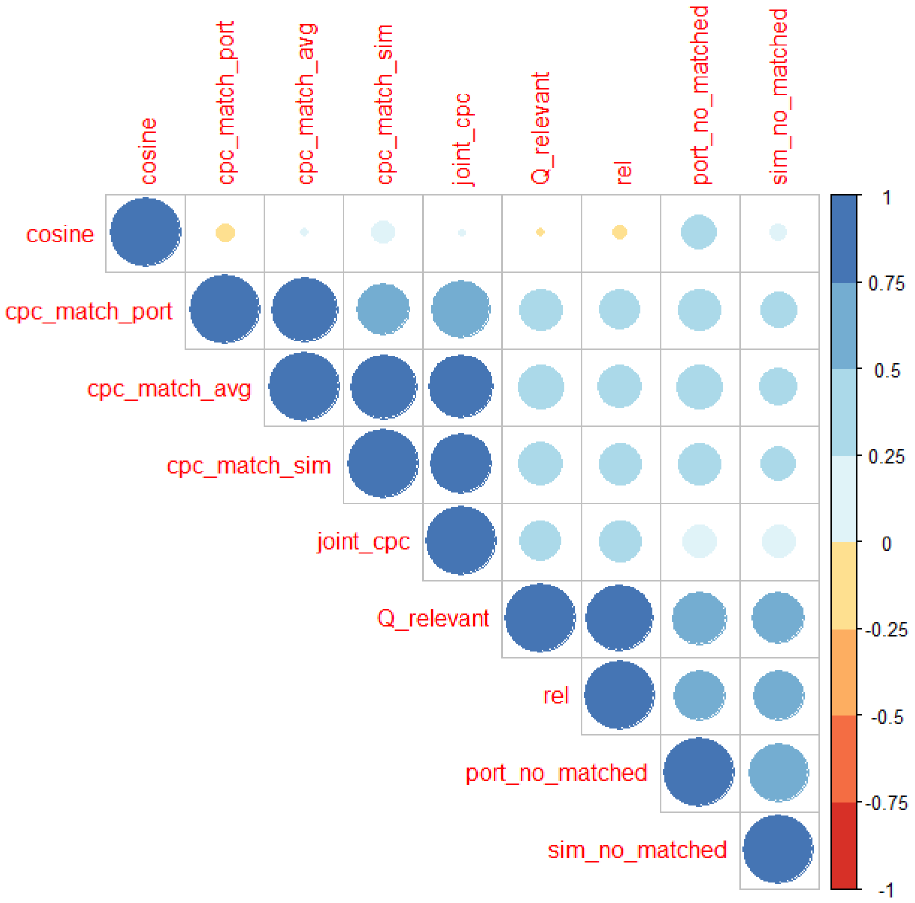


Figure 9 - Correlation analysis between all model variables

Note: "*Q_relevant*" and "*rel*" are the expert evaluations, coded numerically (1-4) and categorically ("useful/not useful"), therefore the independent variables in a regression setting.

Source: Plotted from R analysis with "*corrplot*" and "*RColorBrewer*" packages, script: "*D_model_validation*.R"

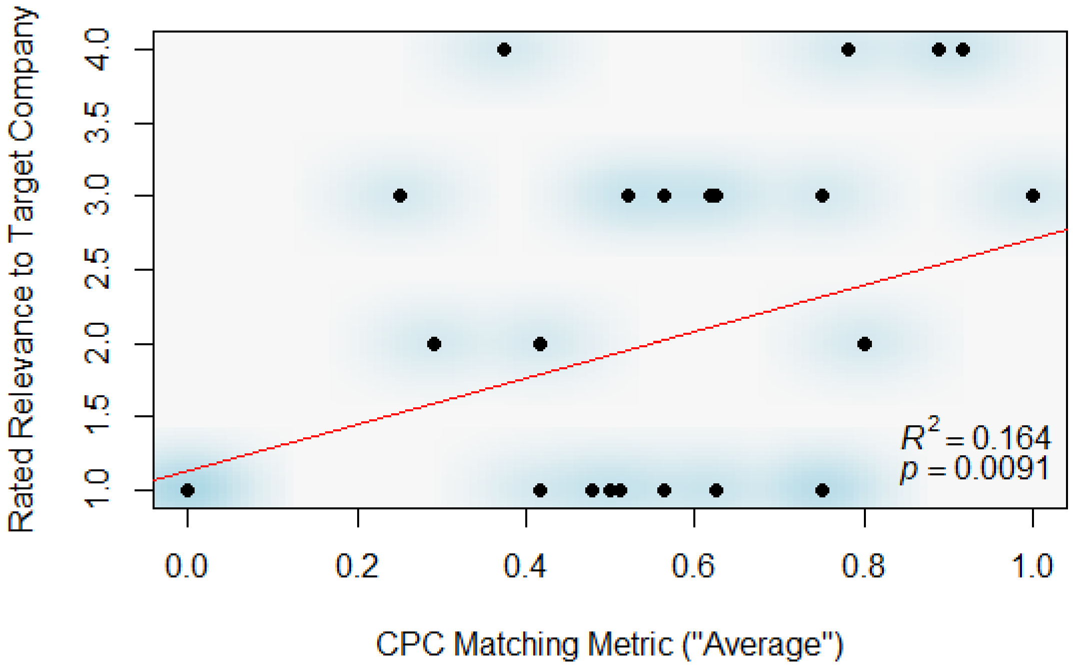


Figure 10 - Regression result main CPC metric

Source: Plotted from R analysis, script: "*D_model_validation*.R"

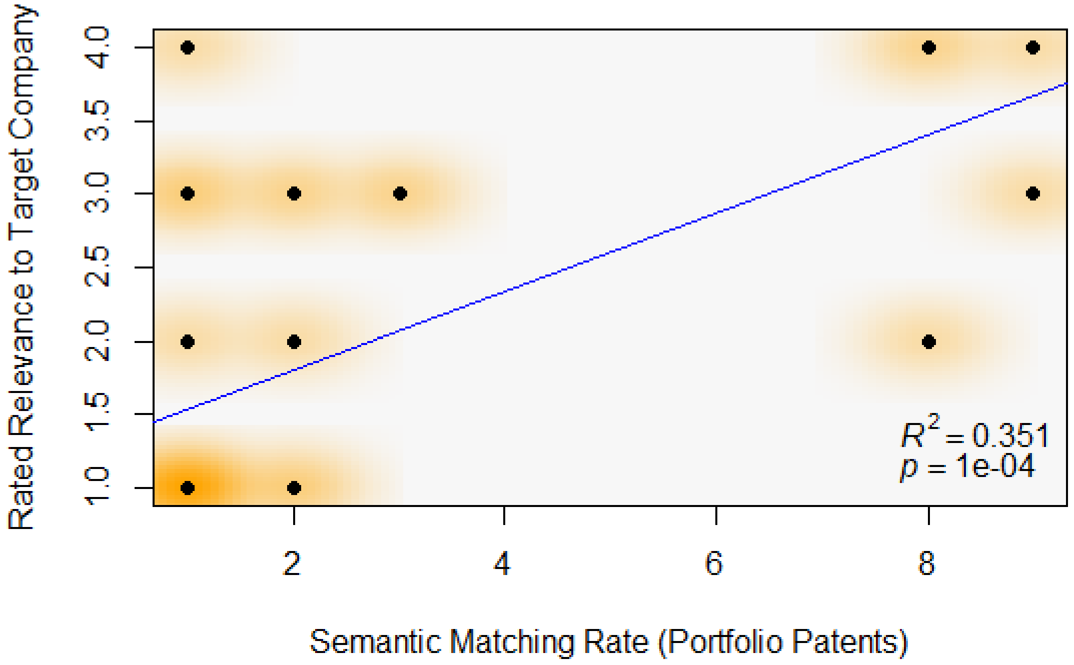


Figure 11 - Regression result main semantic metric

Source: Plotted from R analysis, script: "*D_model_validation*.R"

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. https://doi.org/10.1016/j.wpi.2013.12.006

An, X., Li, J., Xu, S., Chen, L., & Sun, W. (2021). An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics*, *15*, 101135. https://doi.org/10.1016/j.joi.2021.101135

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, *55*, 37–51. https://doi.org/10.1016/j.wpi.2018.07.002

Aristodemou, L., Tietze, F., Athanassopoulou, N., & Minshall, T. (2017). *Exploring the Future of Patent Analytics: A Technology Roadmapping approach* (No. 5; Centre for Technology Management Working Paper Series).

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, *39*(1), 62–84. https://doi.org/10.1002/SMJ.2699

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*(2), 104144. https://doi.org/10.1016/J.RESPOL.2020.104144

Bailey, K. M., Hatch, E., & Lazaraton, A. (1994). The Research Manual: Design and Statistics for Applied Linguistics. *TESOL Quarterly*, *28*(1), 209. https://doi.org/10.2307/3587214

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. *Scientometrics*, *126*, 5431–5476. https://doi.org/10.1007/s11192-021-04001-1

Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive Innovation: An Intellectual History and Directions for Future Research. *Journal of Management Studies*, *55*(7), 1043–1078. https://doi.org/10.1111/JOMS.12349

Cokelaere, Hanne. (2021, May 8). Pope backs coronavirus vaccine patent waivers. Politico. Retrieved from: https://www.politico.eu/article/pope-francis-backs-coronavirus-vaccine-patent-waivers/

Cotropia, C. A., Lemley, M. A., & Sampat, B. (2013). Do applicant patent citations matter? *Research Policy*, *42*(4), 844–854. https://doi.org/10.1016/J.RESPOL.2013.01.003

Crafts, N. (2021). Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy*, *37*(3), 521–536. https://doi.org/10.1093/OXREP/GRAB012

de Rassenfosse, G., Dernis, H., & Boedt, G. (2014). An Introduction to the Patstat Database with Example Queries. *The Australian Economic Review*, *47*(3), 395–408. https://doi.org/https://doi.org/10.1111/1467-8462.12073

EPO, & EUIPO. (2021). *Intellectual property rights and firm performance in the European Union*. https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/IPContributionStudy/IPR_firm_performance_in_EU/

EPO, & USPTO. (2017). *Guide to the CPC (Cooperative Patent Classification)*. https://www.cooperativepatentclassification.org/publications/miscellaneous/miscellaneous_index

Feldman, R. (2012). *Rethinking Patent Law*. Harvard University Press. https://doi.org/doi:10.4159/harvard.9780674064966

Fierro, G. (2013). *Extracting and Formatting Patent Data from USPTO XML*. http://www.funginstitute.berkeley.edu/sites/default/les/Extracting_and_Formatting.pdf

Furman, J. L., Nagler, M., & Watzinger, M. (2018). *Disclosure and Subsequent Innovation: Evidence from the Patent Depository Library Program* (No. 24660; Working Paper Series). https://doi.org/10.3386/W24660

Hain, D., Jurowetzki, R., Buchmann, T., & Wolf, P. (2021). *Text-based Technological Signatures and Similarities: How to create them and what to do with them*. https://arxiv.org/abs/2003.12303v3

Harris, C., Arens, R., & Srinivasan, P. (2010). Comparison of IPC and USPC classification systems in patent prior art searches. *International Conference on Information and Knowledge Management, Proceedings*, 27–31. https://doi.org/10.1145/1871888.1871894

Haugen, H. M. (2021). Does TRIPS (Agreement on Trade-Related Aspects of Intellectual Property Rights) prevent COVID-19 vaccines as a global public good? *Journal of World Intellectual Property*, *24*(3–4), 195–220. https://doi.org/10.1111/JWIP.12187

Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Müller, K.-R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PLOS ONE*, *14*(3), e0212103. https://doi.org/10.1371/JOURNAL.PONE.0212103

Hsu, P.-H., Lee, D., Tambe, P., & Hsu, D. H. (2020). Deep Learning, Text, and Patent Valuation. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3758388

Kim, J., Jeong, B., & Kim, D. (2021). Patent Analytics. Transforming IP Strategy into Intelligence. In *Springer Books* (1st ed.). Springer.

Krestel, R., Chikkamath, R., Hewel, C., & Risch, J. (2021). A survey on deep learning for patent analysis. *World Patent Information*, *65*. https://doi.org/10.1016/j.wpi.2021.102035

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change*, *125*, 236–244. https://doi.org/10.1016/j.techfore.2017.08.002

Lee, C. (2021). A review of data analytics in technological forecasting. *Technological Forecasting and Social Change*, *166*, 120646. https://doi.org/10.1016/j.techfore.2021.120646

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators.

Technological Forecasting and Social Change, *127*, 291–303. https://doi.org/10.1016/j.techfore.2017.10.002

Lee, J., & Berente, N. (2011). Digital Innovation and the Division of Innovative Labor: Digital Controls in the Automotive Industry. *Https://Doi.Org/10.1287/Orsc.1110.0707*, *23*(5), 1428–1447. https://doi.org/10.1287/ORSC.1110.0707

Li, G.-C., Lai, R., D'amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z., & Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy*, *43*, 941–955. https://doi.org/10.1016/j.respol.2014.01.012

Li, X., Chen, H., Zhang, Z., Li, J., & Nunamaker, J. F. (2009). Managing Knowledge in Light of Its Evolution Process: An Empirical Study on Citation Network-Based Patent Classification. *Journal of Management Information Systems*, *26*(1). https://doi.org/10.2753/MIS0742-1222260106

Lindsey, Brink. (2021, June 3). Why intellectual property and pandemics don't mix. Brookings. Retrieved from: https://www.brookings.edu/blog/up-front/2021/06/03/why-intellectual-property-and-pandemics-dont-mix/

Marr, B. (2019). *Artificial intelligence in practice: how 50 successful companies used AI and machine learning to solve problems*. John Wiley & Sons.

McKinsey & Company. (2021). The top trends in tech. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-top-trends-in-tech

Moehrle, M. G., Walter, L., Bergmann, I., Bobe, S., & Skrzipale, S. (2010). Patinformatics as a business process: A guideline through patent research tasks and tools. *World Patent Information*, *32*(4), 291–299. https://doi.org/10.1016/J.WPI.2009.11.003

Morrison, G., Riccaboni, M., & Pammolli, F. (2017). *Data Descriptor: Disambiguation of patent inventors and assignees using high-resolution geolocation data*. https://doi.org/10.1038/sdata.2017.64

Niemann, H., Moehrle, M. G., & Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technological Forecasting and Social Change*, *115*, 210–220. https://doi.org/10.1016/j.techfore.2016.10.004

Petherbridge, L. (2011). On predicting patent litigation. *Tex. L. Rev. See Also*, *90*, 75.

Rodriguez, A., Kim, B., Turkoz, M., Lee, J., Scientometrics, B. C.-, & 2015, undefined. (2015). New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network. *Springer*, *103*, 565–581. https://doi.org/10.1007/s11192-015-1531-8

Sands, Mason. (2018, December 30). Why Copyright Will Be The Biggest Issue For Youtube In 2019. Forbes. Retrieved from: https://www.forbes.com/sites/masonsands/2018/12/30/why-copyright-will-be-the-biggest-issue-for-youtube-in-2019/?sh=7bde42a11c12

Sinan Erzurumlu, S., & Pachamanova, D. (2020). Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. *Technological Forecasting and Social Change*, *156*. https://doi.org/10.1016/j.techfore.2020.120041

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). *Measuring Patent Quality: Indicators of Technological and Economic Value* (No. 3; Technology and Industry Working Papers).

van Looy, B., & Magerman, T. (2019). Using Text Mining Algorithms for Patent Documents and Publications. *Springer Handbooks*, 929–956. https://doi.org/10.1007/978-3-030-02511-3_38

Whalen, R., Lungeanu, A., DeChurch, L., & Contractor, N. (2020). Patent Similarity Data and Innovation Metrics. *Journal of Empirical Legal Studies*, *17*(3), 615–639. https://doi.org/10.1111/JELS.12261

Woo, H.-G., Yeom, J., & Lee, C. (2018). Screening early stage ideas in technology development processes: a text mining and k-nearest neighbours approach using patent information. *Https://Doi.Org/10.1080/09537325.2018.1523386*, *31*(5), 532–545. https://doi.org/10.1080/09537325.2018.1523386

Yan, B., & Luo, J. (2017). Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology*, *68*(2), 423–437. https://doi.org/10.1002/ASI.23664

Yang, C., Jacome, W., Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75. http://veredshwartz.blogspot.sg.

Zeferino De Menezes, H. (2021). *The TRIPS waiver proposal: an urgent measure to expand access to the COVID-19 vaccines* (No. 129).

Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics, 10*(4). https://doi.org/10.1016/j.joi.2016.09.006