TOWARDS PREDICTION OF EMERGING TECHNOLOGIES

by

Shalaka Thombare

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Computer Science

Charlotte

2018

Approved by:

____

Dr. Wlodek Zadrozny

____

Dr. Samira Shaikh

____

Dr. Erik Saule

Abstract

SHALAKA THOMBARE. Towards Prediction of Emerging Technologies. (Under the direction of DR. WLODEK ZADROZNY)

Predictive Analytics is an advanced branch of data analytics that deals with making future predictions that matter to the industry[6]. It has wide applications in the areas of health care, fraud detection, market analysis, cross-sell. Predicting emerging technologies thus can be of great advantage to the corporate and research community.

This thesis, has contributions mainly towards predicting emerging technologies using the US patents and creating an evaluation corpus appropriate for judging the quality of NLP and IR in technology domain. Applying the concept embedding to finding document similarity, this project strives to predict where emerging technologies will appear in the white space of a technology landscape and what those emerging technologies will look like.

I have applied numerical, and semantic techniques in proving the hypotheses hereby presented, and published the data obtained as a result of the data mining project performed to create the evaluation corpus using Patent Trial and Appeal Board's decision files.

ACKNOWLEDGEMENTS

I sincerely thank my advisor, Dr. Wlodek Zadrozny,for giving me an opportunity to work with him on a great idea like this and for all his mentoring, support, and patience during my research. He is a great mentor and continuously steered me in the right direction. I would like to thank the committee members, Dr. Erik Saule and Dr. Samira Shaikh, for agreeing to be part of my committee, and providing ideas and feedback.

A special thanks to my lab mates Dr. Walid Shalaby, Marcia Price for their consistent support, guidance, and contribution to this research. A big thanks to UNC Charlotte for giving me an opportunity to present this research. I would like to extend my thanks to my friends Abhishek Bhandwaldar, Sneha Garapati, Sean Gallagher, and Kshitij Palkar, Shubham Pawar for helping me process my thoughts and and make this idea a reality. Lastly, I sincerely express my profound gratitude to my parents and my sister for their support and encouragement, without whom this accomplishment would not have been possible.

DEDICATION

Dedicated to my dearest Parents, Swati and Chandrashekhar

Contents

List of Figures

LIST OF ABBREVIATIONS

BoW	Bag of Words
CRC	Concept Raw Context
GPS	Global Positioning System
HMM	Hidden Markov Models
IR	Information Retrieval
LED	Light Emitting Diode
MSA	Mined Semantic Analysis
NLP	Natural Language Processing
PTAB	Patent Trial Appeals Board
t-SNE	Stochastic Neighbor Embedding
USPTO	United States Patent and Trademark Office

CHAPTER 1: INTRODUCTION

For two hundred years the United States patent system has defined what is an invention and protected, enriched, and befuddled inventors. As a tool of corporate growth in a global economy, it is now more important than ever.
– American Heritage

1.1 What is Emergence

We define emergence to be a significant rise in the number of patents filed in a technology. Thus, not making emergence to be just about the invention of blend of two technologies.

The joining of patents in one technological area with those in other technological areas occurs as a function of time when a technology has moved into a new, industrially important area. A rapid increase in patent applications filed related to a specific classification can represent its dissemination into new and often unrelated technological areas. This rapid increase in patent application filing is a signal of technology emergence and industry acceptance[7].

However, this significant rise in the number of patents filed in a technology needs to be quantified with some of the numerical tools like standard deviation, etc. While this project only relies on the visual analysis of the technology rise graph, above mentioned limitation can be addressed in the future work on this project.

1.2 Why is prediction of emerging technologies important

Prediction of emerging technologies can be important to the research community as well as the corporate world for so many reasons. Predictive analytics is a branch of advanced analytics which is used to make predictions about the future predictions.

Applications of predictive analytics spread wide like in health care, risk management, direct marketing, fraud detection, cross sell etc [23]. Thus, the topic of this thesis can be seen of a huge application in detecting future competition for companies and their products, finding out about the potential industries for market investments, and for the research communities to find out about the future opportunities in the technological whitespace.

With this objective in mind, this project attempts to demonstrate the behavior of document-document similarity on the patent documents, using concept embeddings to vectorize the contextual meaning of a document. US patents are considered to a potential for many research tracks because of the huge amount of text, and images of almost every technology there is.

1.3 Problem Statement

With the idea of working towards prediction of emerging technologies, the first question in picture is

1. Can we predict emerging technologies?

The U.S. Patent and Trademark Office's formal categorization system has been used to classify literally millions of technical documents. The classification is not merely a single technical descriptor, but a categorization of all the areas of science and technology advanced in the document as recognized by the patent examiner. And thus when we think of exploring this database for prediction of emerging technologies,

2. Are current IR and NLP technologies capable of analyzing technology portfolios? If we do device one, 3. Do we have proper evaluation corpora and tools to answer this question?

This project attempts at addressing the above questions.

1.4 Contributions

The contributions of this project are as follows:

1. Demonstrating the predictive model for emerging technologies using four Case Studies

    - Hidden Markov Models into Genomics
    - Radio Frequency transmission into Smart Card
    - Lithium batteries into Mobile phones
    - Blockchain into Bitcoin
    - Building on prior work done at UNCC by Ankam et al 2013 on emergence of mobile phones

2. Confirming these findings using Semantic Techniques

3. Finding Cases of Emerging Technologies where the numerical and semantic approaches Fail

    - Precision Farming and Global Navigation Satellite System
    - LED and Displays
    - Deep Neural Network and Image Recognition

4. Creating an evaluation corpus appropriate for judging the quality of NLP and IR in technology domain

    - Extracting relationships between patents using PTAB report data
    - Publishing this corpus on data.world

CHAPTER 2: TOWARDS PREDICTION OF EMERGING TECHNOLOGIES

2.1 Objective

The objective is to be able to find the emergence of a mix of two technologies or two parts of the same technology. For instance, the way Hidden Markov Models, a technology for pattern recognition was seen to blend in the field of genomics, or, the use of authentication was seen to emerge in smart card readers, LED's were seen to migrate to Television. Let A, B and C be some existing technology or a concept or an application whose patents are being filed since some years. We predict the emergence of the mix of A and B together, by finding a common factor C which has been part of A and the occurrence of B in which is to cause A and B to merge.

2.2 Literature Survey

Not many attempts have been found made which directly intend to achieve the objective of this project, but this is project is based on the application of many machine learning and visualization techniques. We start by reviewing the papers forming the most important basis, word embeddings. Word embeddings form the basis of this project as the entire hypothesis relies on finding the relationship between the patent application documents.

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension.

Similarly, to convert the entire patent application document to a lower dimensional vector, such that it captures the essence of the document along with some latent concepts, we use the neural embedding language modelling techniques. Accurately representing a document (as big as 10 pages of text), requires that the neural network learns the distributional semantics for the concepts involved in the document. The idea of course starts from the successful introduction of neural probabilistic language models by Dr Bengio[15].

This learning as I understand would involve two important tasks, a) Learning the individual word embeddings, b) Aggregating those vectors into representing the entire document. Thus, we start by using the vectors generated by using the concept embeddings as elaborately described in the paper about Learning Concept Embeddings for Efficient Bag-of-Concepts Densification[2].

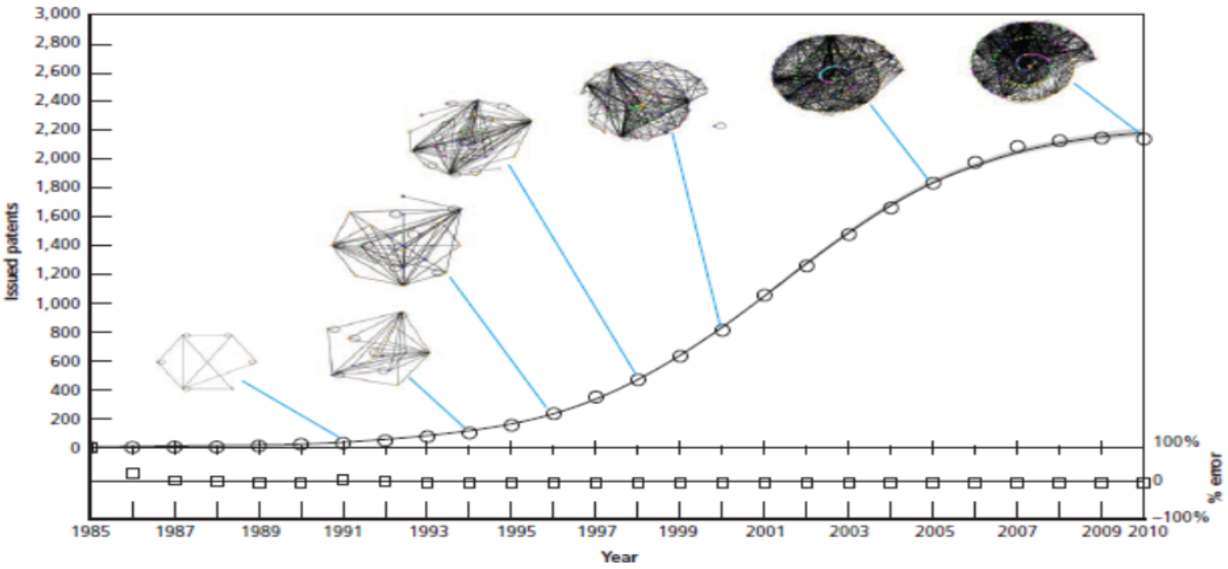
MSA utilizes a search index created using concept rich corpora (e.g., Wikipedia). The concept vector of a given text is constructed through two phases. First, an initial set of candidate concepts is retrieved from the index. Second, the candidates set is augmented with other related concepts from the discovered concept-concept association rules[6].

2.3 Related Work

As this project stands on a basic platform of patent analytics, it is important to have reviewed the techniques and approaches involved in the same. (Abbas*, et al 2014) [9] have extensively comprehended and presented a comparative analysis of these approaches. Semantic Analysis is identified as an important technique here.

An important and closely related work done on this topic is the one I find in [7]. Here, the authors have worked towards analyzing and identifying emerging technologies in terms of time and magnitude(number of patents filed), with the use of the disruption of classification codes of the US Patents. While detecting the emergence in terms of growth of the density of inter-classification network, i.e. when any one class/subclass starts appearing in other classes more and more, [7] clearly demonstrates the growth of GPS into vehicle navigation system over the years starting in 1980.

Figure below shows the increase in patents filed under the classification codes of GPS and vehicle navigation.



(a) S-curve as an indicator, identifies emergence of GPS into vehicle navigation [7]

I find a couple of papers quite closely related to the concept of prediction of emerging technologies using patents data. [4] and [5] refer to them.

Ankam, et al 2013 refers to prediction of emerging technologies using topic modelling of the patent application text. This demonstrates the use of topic modelling on a patent class under study, plots and visualizes the topics strength over time, whereby the emerging technologies are made evident. It is based on the fact that patent applications are published years ahead of being granted. Understanding that word embeddings will provide better results in discovering the latent concepts of patent documents, we propose to implement this project as an improvement over [4]. Moreover, it only limits to a specific class under study whereas it attempts to generalize to the emergence pattern by demonstrating over many case studies in various technology classes.

As we build this idea mainly based on the work in [4] which uses topic modelling to indicate the growth of smart phones technology, figure below, demonstrated by Ankam et al UNCC [4] can be seen to indicate topic modelling as an indicator of this emergence.

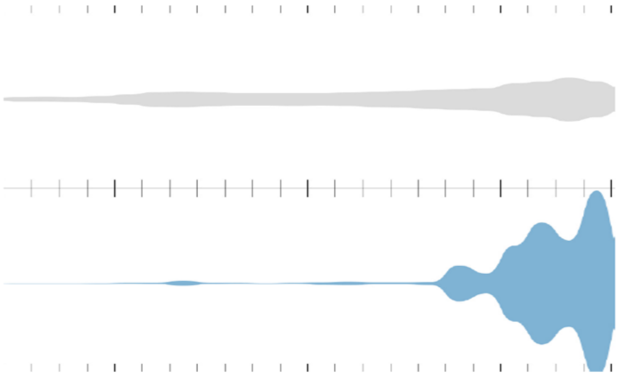


Fig.2. Showing the difference between a stable topic such as "transistor" vs. an emerging topic "storage, software..." in patent class 455 (telecommunication).

(b) Topic as an indicator of an emerging technology : Smart Phones [4]

Figure 2.1: Demonstrates how the technology of smart phones developed over the years and can be seen with the swell of the graph which grows as the number of patents filed in the class of telecommunication increases [4]

Erdi, et al 2013 refers to prediction of emerging technologies using the patent citation network in the patent applications. It builds what is called a citation vector for each patent document based on the citations mentioned in the application. Each coordinate of the citation vector is proportional to the relative frequency that the patent has been cited by the other patents in a particular technological category at a specific time. Changes in this citation vector over time reflect the changing role that a particular patented technology is playing as contributor to later development of the technology. Our methodology is similar to this paper in a way that we are using the patent application clusters over time to make the prediction of emerging technologies.

Moehrle, et al 2010, talks about the methods of evaluating patent similarities using text processing. Following on the approaches and addressing the limitation to extract the semantic meaning out of the patent text, in [12] and [14], we build this project to demonstrate the patent-patent similarity using concept embeddings [1].

Having stated an important limitation of our work in this project to be in selecting the candidate technologies to look up into in order to detect the emergence, I find [25] to be an important related work for this project's future work. The Emerging Clusters Model [25], is a tool for identifying the emerging technologies across multiple patent systems. This paper describes the first large scale test of the Emerging Clusters Model and reveals that patents in emerging clusters consistently have a significantly higher impact on subsequent technological developments than patents outside those clusters [25].

While we have worked on the basis of some case studies here in this project, work by H. Ernst, et al. 1997 is an important work where the author appropriately describes and demonstrates the use of CNC-technology in the machine tool industry[17].

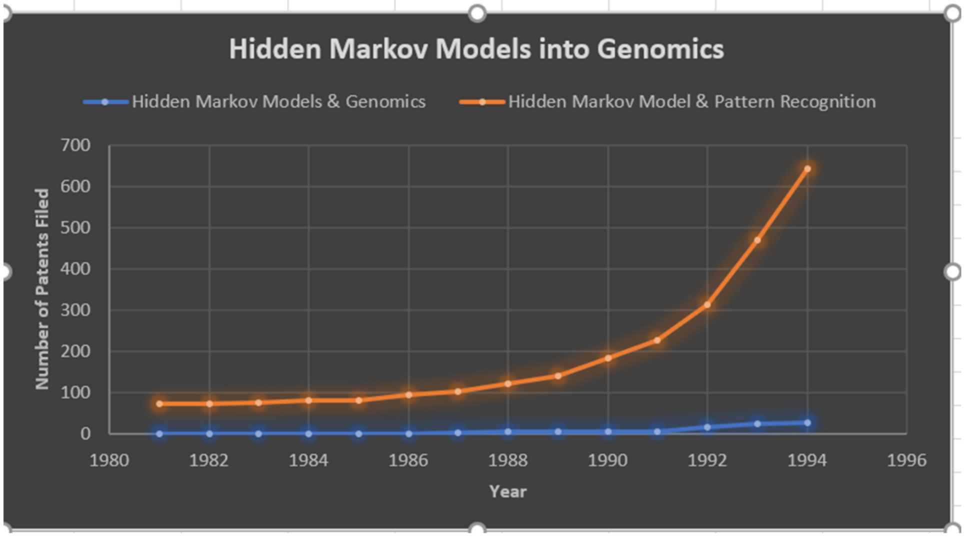
CHAPTER 3: Demonstrating the predictive model for emerging technologies using four Case Studies

3.1 Case Study 1: Emergence of Hidden Markov Model into Genomics

By looking into the patents data for a specific technology (Hidden Markov Models for this instance), our objective is to be able to predict emergence of a technology (for example HMMs in genomics). In the case study that we have for consideration, we are trying to see the emergence of use of HMMs in genomics by looking for the pattern in the growth of patents in HMMs+Genomics Vs patents in Genomics. By regressing this data, we are trying to predict the use of HMMs in Genomics, 5-10 years ahead of the actual emergence.

- We plot the number of patents filed in the application of HMMs in genomics. We start from the year 1981, since there are no patents filed in genomics+HMMs and HMMs+pattern recognition before the same. We see that in genomics+HMMS, the patents have begun to come up slowly, with peak rise during the year 2001.

- We plot the number of patents filed in the field of genomics. In the below chart,patents scaled - represents the patents filed in Genomics. Patents in Q3 and Q4 represent the patents filed in Genomics+HMMs



(a) Growth of Hidden Markov Models into Genomics: This figure clearly indicates how Hidden markov models emerged in pattern recognition five years before they emerged in Genomics

Figure 3.1

How I think HMM migrated to genomics:

1. During the year 1980-1995, the patents filed in the field of pattern recognition and Hidden Markov models substantiated, with peak rise in 1991-1992.

2. Patents in genomics always include the ones mentioning pattern recognition.

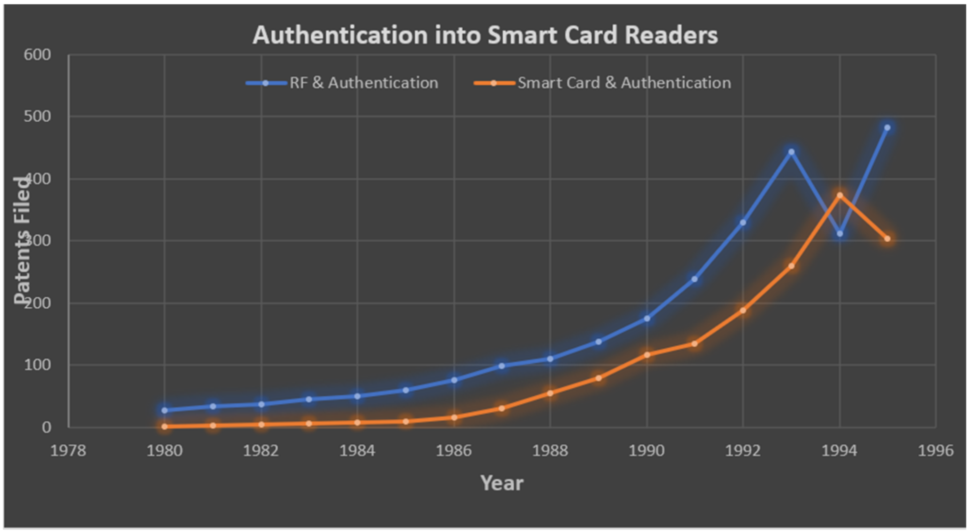
3. I find that only after the first patent of the use of HMMs in pattern recognition during 1982-85, was the use of HMMs in genomics was substantiated.

4. This is a statistical observation.

5. However, the main task of the project would be to be able to prove this hypothesis, by programmatically finding the fields in the patents which will hint at such emergence of technology.

3.2 Case Study 2: Emergence of Authentication into Smart Card Readers

Looking at the evolution of radio frequency transmission into authentication, smart card reader emergence can be tracked. I plotted the growth of authentication into smart card reader, and radio frequency transmission into authentication. This shows how the evolution of authentication technology into smart cards appeared. We plot the number of patents filed in technologies of smart card and authentication together over the years 1980-2000. Just like in example 1, we are trying to find the key to migration of authentication to smart card readers. For this, we track the growth of authentication and radio frequency together. In this case, the patents of authentication and smart cards started occurring together only after 3-4 years of occurrence of authentication and radio frequency transmission together.



(a) This figure visualizes how the emergence of RF and authentication occurred before the emergence of smart card reader and authentication

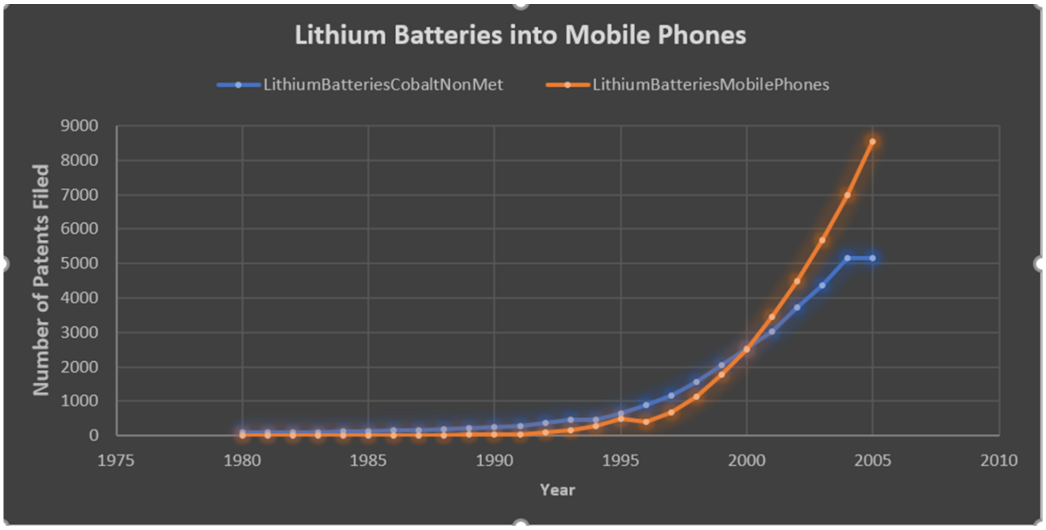
Figure 3.2

3.3 Case Study 3: Emergence of Lithium Ion Batteries into Mobile Phones

As we observe the growth of lithium batteries, they started getting used in mobile phones only after the occurrence of the invention of lithium ion batteries, that is, the ones which use non-metallic lithium in them. Thus, the growth of intersection of lithium batteries and mobile phones, having been constant till 1990, started rising after lithium-Cobalt dioxide or cobalt oxide. So the key term that I see is Cobalt, non-aqueous and non-metallic used in lithium batteries.

The number of patents together filed in Mobile phones and lithium batteries were plotted from 1980 to 2005 as shown in figure 1, and the number of patents filed in lithium batteries and cobalt, non-aqueous and non-metallic were plotted

The below graph show the improvement of lithium batteries in mobile phones only after the growth of cobalt in lithium non-aqueous, non-metallic batteries.



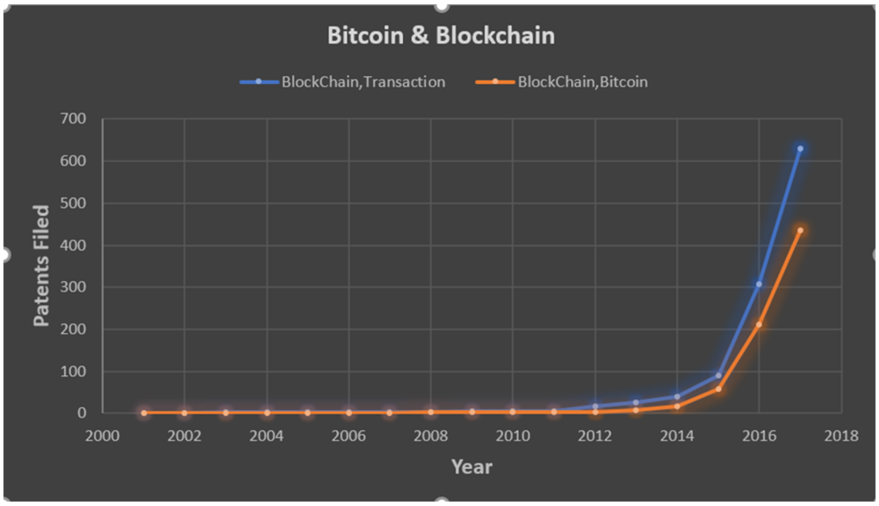
(a) Growth of Lithium Batteries into Mobile Phones: This visualizes the emergence of lithium and non-metallic, nickel and cobalt before the emergence of lithium batteries and mobile phones

Figure 3.3

3.4 Case Study 4: Emergence of Bitcoin and Blockchain

Looking into the patents of bitcoin, we see that the technology of blockchain for bitcoin transactions has emerged starting 2008- till date. So, I looked into the patents of bitcoin, individually, blockchain with the term transaction, and bitcoin and blockchain together. Following is the graph plot that is seen on plotting them in the timeline of 2001 to 2017.

Figure below is self-explanatory in explaining the relationship between growth of blockchain and bitcoin together, caused of the blockchain and transaction terms.



(a) Growth of Blockchain into Bitcoin

Figure 3.4

CHAPTER 4: Finding Cases of Emerging Technologies where the numerical approaches Fail

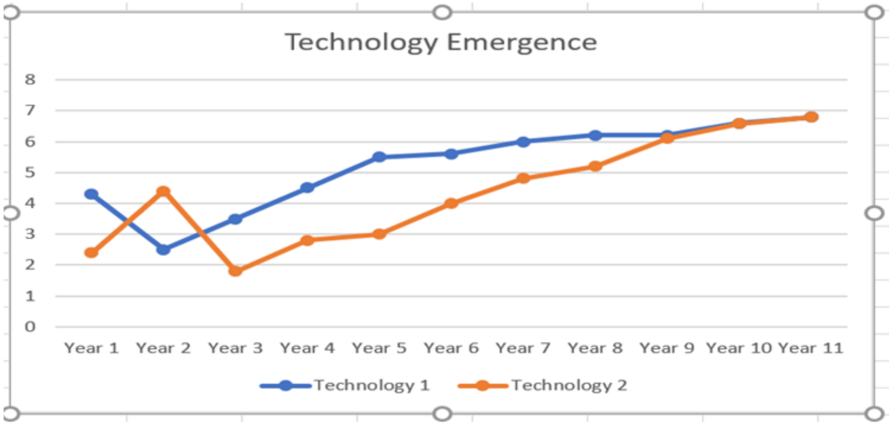
There have been more case studies under consideration like emergence of deep neural network for image recognition, bitcoin and blockchain, Precision farming, where the numerical hypothesis could not be confirmed. The failure of the hypothesis can be attributed to some of the limitations of this project. One important factor is being able to retrieve the results relevant to the technology under study for the search terms. As the number of words in the search phrase increase, the relevance gets awry.

Secondly, not all the technologies emerged are caused because of a limited number of and clearly defined terms. Like for the emergence of precision farming, the deductible factors are the farming and the use of global navigation satellite systems, but the involvement of other factors like database systems, data analysis, computer software, fail the hypothesis. Abundance of the available data and vectors being trained on the particular concepts under study also play an important role. In case of bitcoin and blockchain emergence, these factors fall weak, thus failing the hypothesis.

CHAPTER 5: METHODOLOGY FOR NUMERICAL ANALYSIS

5.1 Hypothesis

Having defined emergence as significant rise in the number of patents in a field, I hypothesize the pattern that I am looking for Looking at the number of patents filed in a technology field each year, I find the pattern where the advent of technology B to A, causes the emergence of the mix of A and C.



(a) Hypothesis of Technology Emergence: This chart illustrates how we define emergence to be two technology gra coming closer over the time in terms of the patents filed

Figure 5.1

5.2 Patent Search

5.2.1 Google Patents

Every data analytics project is known to dedicate most of the time to data gathering, data cleaning and other data preparation tasks. For this project, I relied on the Google patents search for the patents documents as Google patents provides a precise and user friendly, keyword search interface for searching the patents[24]

5.2.2 Why Google Patents

Patents can be searched for with Keywords and Key-phrases with years selection and also provides the immediate patent download options. Also worked with US Patent Advanced patent search tool and AcclaimIP. Google Patents returns more results for a particular search phrase within the three software since the searching is more accommodating, e.g : Same results are returned for Genomics and Genome, Hidden Markov Models and Hidden Markov Model

5.3 Process

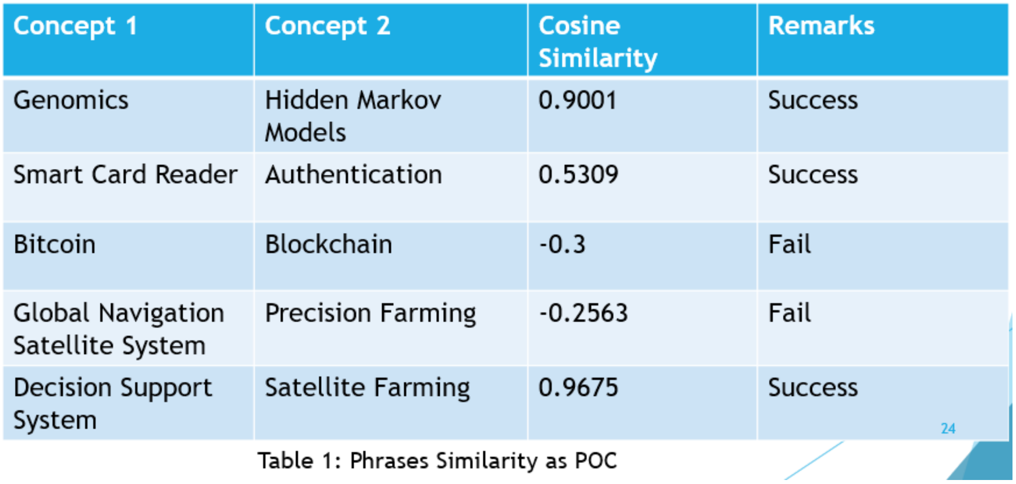
The python script takes as input the terms we want to search for, and the year range we want, it then, using Selenium, accesses Google patents and sends it queries year by year. Sample data formed can be seen below.

CHAPTER 6: CONFIRMING THE NUMERICAL FINDINGS USING SEMANTIC TECHNIQUES

6.1 Proof of Concept

The idea of finding semantic similarity between the documents based on their concepts, relies on the fact that the concepts under study are similar to the model. For example, if the two documents contain just the terms “Genomics” and “Hidden Markov Models”, their similarity score should be at least a positive score that we can rely upon.

Thus, I worked on finding the similarity between the concepts involved in our case studies so that their scores indicate on the possible success of that case study.



Concept 1	Concept 2	Cosine Similarity	Remarks
Genomics	Hidden Markov Models	0.9001	Success
Smart Card Reader	Authentication	0.5309	Success
Bitcoin	Blockchain	-0.3	Fail
Global Navigation Satellite System	Precision Farming	-0.2563	Fail
Decision Support System	Satellite Farming	0.9675	Success

Table 1: Phrases Similarity as POC

(a) Phrase Concept Similarities [6]

Figure 6.1

Fig.6.1 shows the similarity between the phrases/concepts of the concepts under the study. Since Genomics and Hidden Markov Models show a cosine similarity of 0.9001, we can deduce that the model has discovered the concepts hidden in those terms correctly. Similarly, the concepts in the terms of bitcoin and blockchain are not correct for the experiment since they show negative similarity score. This failure is also seen in the numerical hypothesis.

6.2 Approach and Process

For the scope of this project, we are only taking US patents under consideration. This project is about trying to predict emerging technologies by tracking the progress of patents of two specific seemingly independent technologies 10-15 years before the actual emergence. We have observed that the most occurring terms in the description and introduction field of a patent released by the USPTO hint us at the future emerging technologies. So, the important steps to this project as I identify them are,

1. Pick a target patent of the desired technology.
2. Find the semantically significant terms/phrases (one word or multiple words) out of the patent text of that patent.
3. Searching the US patents which contain the semantically similar set of key terms over a given time period of 5 to 10 years.
4. Thus, matching the candidate documents and the target documents based on the key terms, and predicting the emerging technology as a blend of the candidate and target technologies is the final step.
5. One important challenge is to statistically estimate the years over which we choose the candidate patents and the years over which we predict the emergence to occur. Question is, how do we find the exact candidate patents in which we look for the so called “key terms”

6.2.1 Approach to proving hypothesis

Primary case under study is of the emergence of use of hidden Markov models into genomics. Conceptualizing the patent documents using Concept-Concept Semantic

Space models thus obtaining vector representation for each of them [20],[21]. Visualizing the vectors using the stochastic neighbor embedding technique for visualization of high dimensional data. We hypothesize that this plot will help see the emergence of patents and highlight the relationship between target and candidate patents.

6.3 About the Vectors

The main idea is to embed textual structures into a semantic space of concepts which captures the main topics of these structures. The paper proposes two neural embedding models to learn continuous concept vectors based on the skip-gram model. Moreover, the concept raw context (CRC) model, maps both words and concepts into the same semantic space. Therefore, making it easy to measure word-word, word-concept, and concept-concept semantic similarities.

This model uses the Wikipedia articles 'corpora. The vectors are trained using context size of 9 and embedding size 500. These vectors are special cause they don't just use words as a model but also the concepts fetched from the articles. This collection of concepts and words is then modelled using the skip-gram method and then trained on the neural network with embeddings size of 500 and context size 9. Following Mikolov et al. [8], it utilizes negative sampling to approximate the softmax function.

It works in two phases, Search and Expansion. The search phase outputs the explicit concepts and the expansion phase takes these concepts and looks for the implicit concepts obtained by connecting an article with it's See Also section in the Wikipedia page.

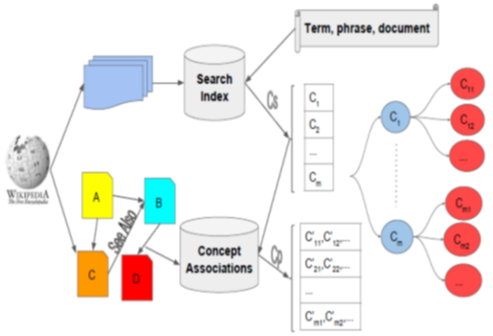


Figure 1. *MSA* generates the concept vector of a given textual structure through: 1) explicit concept retrieval from the index (top), and 2) concept expansion from the concept-concept associations repository (bottom).

(a) Mined Semantic Analysis [6]

Figure 6.2

MSA is used here for the document to vector conversion looking at MSA's performance on finding the implicit concepts of a text. Fig. 6.2 clearly demonstrates the implicit concepts and explicit concepts of a sample concept of Computational Linguistics.

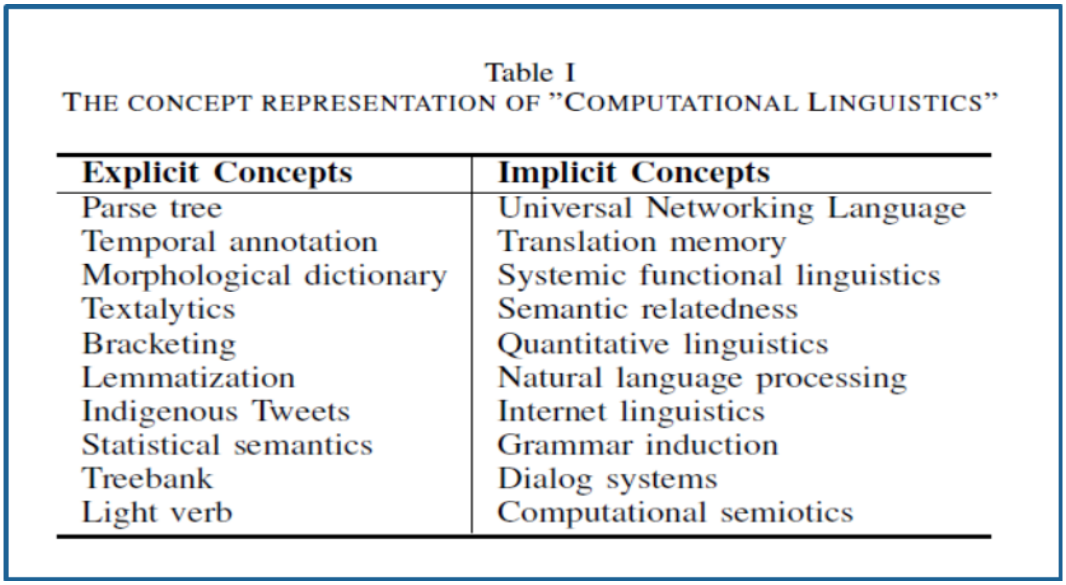


Table I
THE CONCEPT REPRESENTATION OF ”COMPUTATIONAL LINGUISTICS”

Explicit Concepts	**Implicit Concepts**
Parse tree	Universal Networking Language
Temporal annotation	Translation memory
Morphological dictionary	Systemic functional linguistics
Textalytics	Semantic relatedness
Bracketing	Quantitative linguistics
Lemmatization	Natural language processing
Indigenous Tweets	Internet linguistics
Statistical semantics	Grammar induction
Treebank	Dialog systems
Light verb	Computational semiotics

(a) Concept Representation of “Computational Linguistics” [6]

Figure 6.3

The motive of this project is to find the conceptual similarity between patent documents[20], thus to visualize the pattern of emergence of technologies. Thus, the vectors which look for the implicit concepts in the documents are of importance [9]. Thereby, when picking the word embeddings model, I picked Mined Semantic Analysis amongst Word2Vec, BoW etc.

The performance of mined semantic analysis over other neural embedding techniques is elaborately described in the paper [9][2]. Thus, we use the model generated from this experiment to convert the patent documents to representative vectors.

Once all the patents under review are converted to vectors, we pick one case at a time and visualize it on a three-dimensional graph with horizontal axis as time. Effective representation of high dimensional data utilizes the t-SNE technique [4].

CHAPTER 7: EXTRACTION OF THE SEMANTIC KEY TERMS

7.1 Topic Modelling

Latent Dirichlet allocation(LDA) is a generative topic model to find latent topics in a text corpus. It can be trained via collapsed Gibbs sampling[13]. Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Mallet is a JAVA based tool that, uses Gibbs sampling to find the topics from the given documents. In topic modelling, the user should make the judgment on the number of topics that he wants out of the documents, just like in K-means clustering where the user should choose the number of clusters.

To track the trending terms of Key terms out of the patents in Genomics, I used the patents between the years 1975-80 and chose 50 topics out of the 100 documents. Out of the fifty topics that mallet found out, three seem to contain pattern and recognition as their significant keywords which indicates the semantic importance of "Pattern" in the field of Genomics.

Below are the topic details out of the topic phrase report generated by mallet.

- <topic id="1" alpha="0.1" totalTokens="3889" titles="antenna, signal, frequency, modulation, modulating, pattern, directional, response, array, side">

- <topic id="8" alpha="0.1" totalTokens="4477" titles="detectable levels, active, reaction, template, poly, system, specific, complete, recognition, viral">

- <topic id="15" alpha="0.1" totalTokens="2193" titles="sequence, id, acid, nu-

cleic, ul, hcv, base, reaction, pairs, recognition">

CHAPTER 8: RESULTS AND ANALYSIS OF RESULTS

8.1 Detailed Experiment

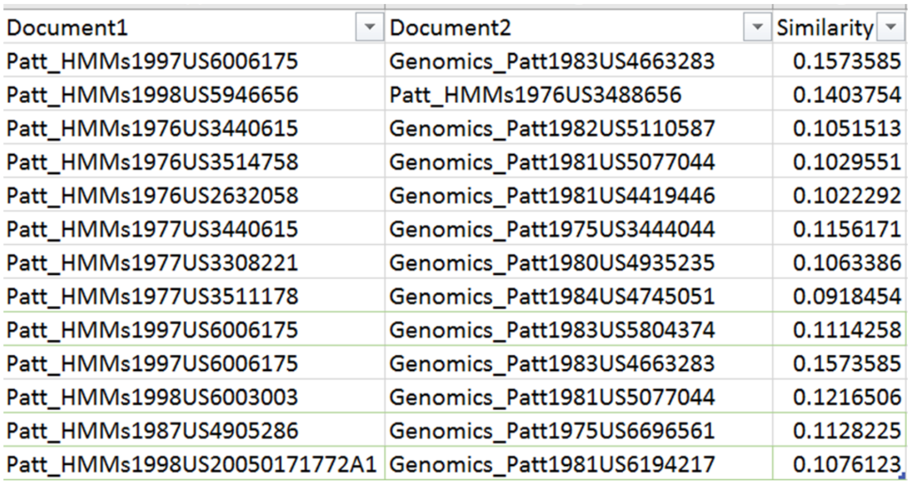
I executed this process in two versions, Version 1: All the patents of a year in a field are merged and kept as one document Version 2: Every Patent of every year is a different field is one document

After the data, i.e. the patent documents have been preprocessed, cleaned and shortened, I associate the patent number of a document as it's id. Example : Version 1: GenomicsPatt1984, PattHMMs1981 Version 2: PattHMMs1977US3308221, GenomicsPatt1983US4663283

Version 1 can be seen as a cluster of the documents of Version 2

8.2 Results

This experiment basically is about finding document-document similarity between patents. Out of 5K similarities between documents, if we pick the highest few scores, the following results are seen,



Document1	Document2	Similarity
Patt_HMMs1997US6006175	Genomics_Patt1983US4663283	0.1573585
Patt_HMMs1998US5946656	Patt_HMMs1976US3488656	0.1403754
Patt_HMMs1976US3440615	Genomics_Patt1982US5110587	0.1051513
Patt_HMMs1976US3514758	Genomics_Patt1981US5077044	0.1029551
Patt_HMMs1976US2632058	Genomics_Patt1981US4419446	0.1022292
Patt_HMMs1977US3440615	Genomics_Patt1975US3444044	0.1156171
Patt_HMMs1977US3308221	Genomics_Patt1980US4935235	0.1063386
Patt_HMMs1977US3511178	Genomics_Patt1984US4745051	0.0918454
Patt_HMMs1997US6006175	Genomics_Patt1983US5804374	0.1114258
Patt_HMMs1997US6006175	Genomics_Patt1983US4663283	0.1573585
Patt_HMMs1998US6003003	Genomics_Patt1981US5077044	0.1216506
Patt_HMMs1987US4905286	Genomics_Patt1975US6696561	0.1128225
Patt_HMMs1998US20050171772A1	Genomics_Patt1981US6194217	0.1076123

(a) Cosine Similarity between patents in HMMs and Genomics starts rising after 1981-82. This shows how, years ahead of emergence of HMMs and Genomics, the patents in the two fields begin to gain similarity

Figure 8.1

8.3 Analysis of the results

As the emergence of a technology approaches, the similarity between the patents from the two technologies keep increasing and the patents in both the technologies with higher similarity score increases. That's when we can predict the emergence to be 4-5 years away. In case of Hidden Markov Models and Genomics, the peak of emergence was seen during the 1990-1992, and thus the cross-similarity can be seen to occur beginning 1983-84.

8.4 CHALLENGES

8.4.1 Limitations of this work

- Inadequate Evaluation techniques for the semantic similarity results
- Curse of the generality : The patents classification database is just so wide, that generalizing the equation to prediction would need a lot of experiments covered.
- How many keywords would you choose ?

- This project has not quantified the rate of change of the number of patents filed in the emerged technology field. Although visually analyzed, and relative to the number of patents filed earlier, the change that we address as a rise of emergence needs to be fixated.

CHAPTER 9: PTAB Data Mining Tool

Post grant review and Inter Partes Review (IPR) is conducted at the USPTO Patent Trial and Appeal Board (PTAB) and is aimed at reviewing the patentability of one or more claims in a patent. The board releases a decision document with each of its decision which specifies the decision, the related patents/applications which have already been rejected in the past and form a reasoning for the decision, the rulings involved etc. The aim of this project is mining various information fields from this Patent Trial and Appeal Board (PTAB) data. We have coded an information extraction program which successfully extracts information fields from the available data.

Next steps: Evaluation of the new code:

1. Choose 25 random ptab zip files
2. Run the new data extraction programs on these 25 zip files
3. Choose 100 random pdf files from the 25 zip files– 4 random pdfs per zip file
4. Evaluate results of step 2 on the pdfs from step 3
5. Put results in the table

Within this data, each case has a relatively small collection of highly relevant documents used as evidence. The outcomes are clear and the reasoning can be modeled. There's enough data for statistical inference (although perhaps not enough to train a neural net from scratch). Also, the PTAB data set might represent better the practitioner needs, as contrasted with using citations as such representation.

Patent Trial and Appeal Board (PTAB) publicly available dataset, as of Jan 2017 has about 100 zip files containing 10 GB of data (compressed)[4]. Each decision pertains to the validity of claims of one patent.

The data is in form of PDF text and dated. Approximate size of data is around 20 GB. Each PDF contains trial of separate patent application. The fields of interest are the Application Number, the decision, prior art and rulings.

Process of Extraction: As stated above, the original PTAB data in the bulk folder ranging from year 1997 to 2016 (about 20GB) consists of PDF format.

- PTAB data in the bulk folder ranging from year 2017 to 2017 till date(about 1GB).
- In order to perform text analysis and bulk extraction, each of these files need to be converted to text format so that we could search and in-bulk be able run test cases.
- Using Adobe Acrobat 9 pro, batch processing over PDF files can be done using Optical Character Recognition(OCR) which converts the images in PDF, the content of each patent in PTAB, to readable and searchable text format.
- A Replica of each image in the PDF is converted to text and saved with the same file name which can be accessed and read from.
- The converted files are then passed through above java code to extract the fields. The output is saved in CSV file. The data separated by comma separators is then placed into different columns for easier readability.

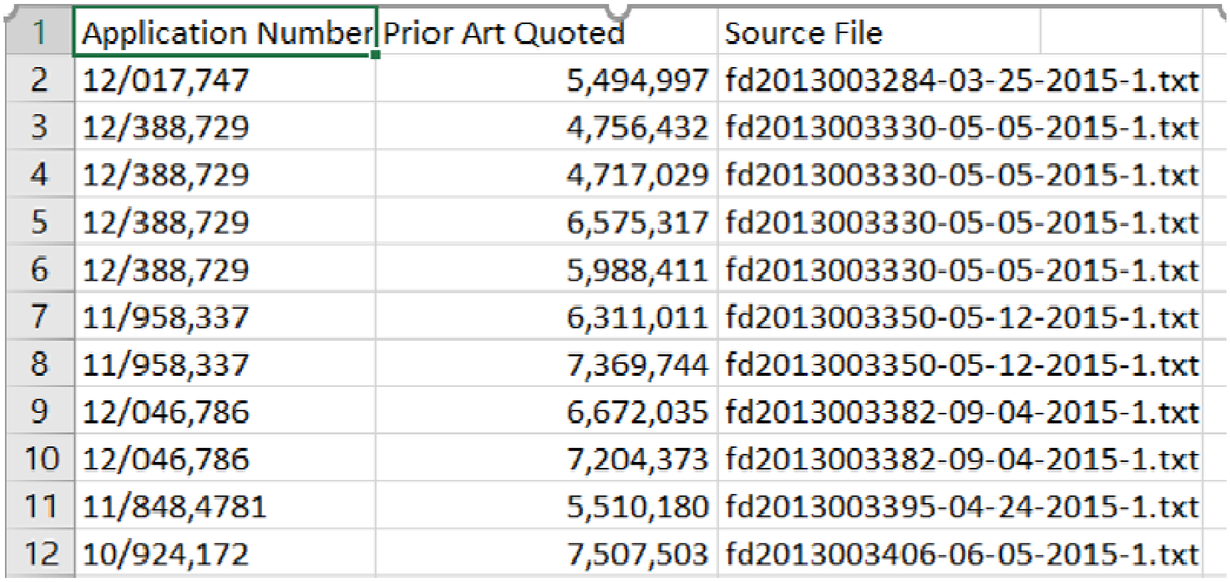
PTAB data in the bulk folder ranging from year 2017 to 2017 till date(about 1GB).

In order to perform text analysis and bulk extraction, each of these files need to be converted to text format so that we could search and in-bulk be able run test cases.

Using Adobe Acrobat 9 pro, batch processing over PDF files can be done using Optical Character Recognition(OCR) which converts the images in PDF, the content of each patent in PTAB, to readable and searchable text format.

A Replica of each image in the PDF is converted to text and saved with the same file name which can be accessed and read from.

The converted files are then passed through above java code to extract the fields. The output is saved in CSV file. The data separated by comma separators is then placed into different columns for easier readability.



1	Application Number	Prior Art Quoted	Source File
2	12/017,747	5,494,997	fd2013003284-03-25-2015-1.txt
3	12/388,729	4,756,432	fd2013003330-05-05-2015-1.txt
4	12/388,729	4,717,029	fd2013003330-05-05-2015-1.txt
5	12/388,729	6,575,317	fd2013003330-05-05-2015-1.txt
6	12/388,729	5,988,411	fd2013003330-05-05-2015-1.txt
7	11/958,337	6,311,011	fd2013003350-05-12-2015-1.txt
8	11/958,337	7,369,744	fd2013003350-05-12-2015-1.txt
9	12/046,786	6,672,035	fd2013003382-09-04-2015-1.txt
10	12/046,786	7,204,373	fd2013003382-09-04-2015-1.txt
11	11/848,4781	5,510,180	fd2013003395-04-24-2015-1.txt
12	10/924,172	7,507,503	fd2013003406-06-05-2015-1.txt

(a) Prior Art Extracted as a result of Data Mining tool

Figure 9.1

9.1 Future Work

9.1.1 Semantic Analysis of Patent Applications and Patents

The patent grant process is a very tedious and expensive process. It often involves lawyers and manual labor for the examination. Patent examiners review patent applications to determine whether the invention(s) claimed in each of them should be granted a patent or whether the application should instead be refused. One of the most important tasks of a patent examiner is to review the disclosure in the application and to compare it to the prior art. This involves reading and understanding a patent application, searching the prior art (including prior patent applications and patents, scientific literature databases, etc.) to determine what contribution the invention makes over the prior art, and issuing office actions to explain to the applicants and their representatives (i.e., patent attorneys or agents) any objections that may exist against the grant of a patent. In other words, an examiner reviews a patent application substantively to determine whether it complies with the legal requirements for granting of a patent. A claimed invention must meet patentability requirements of novelty, inventive step or non-obviousness, industrial application (or utility) and sufficiency of disclosure. Examiners are expected to be efficient in their work and to determine patentability within a limited amount of time.

This project provides us with a patent application number and the prior art(s) based on which some of the claims disclosed in the application have been rejected using a decision file. We can use this data to get the contents of the respective application and the related prior arts. Using this data, we could look for semantically or contextually similar words, phrases or sentences. This obtained information could be used to form word and phrase embeddings. For example, in Application No. 10/016,935, the phrases message "filtering policy" and "filtering rules" are semantically similar to the phrases "filtering criteria", "rule-based scheme" (Patent No. US 08/902,400) and "filter and forward options", "a plurality of rules to apply to said messages" (Patent No. US 09/157,818) that appear in the referenced Prior Arts. Similarly, "automatically determining if there is at least one alias for the search keyword by searching a first database using the search keyword" (Application No. 11/841,405) and "multistage queries, making inferences from a first set of result items as to likely additional sources of information which are then queried using the original query terms and/or terms derived from the first set of result items" have the same connotation of the process of finding the aliases of the search keyword and then, using the original keyword and the associated aliases, executing the search. In some cases, the words/phrases are not conspicuously semantically similar. For example, "access control device" (Application No. US 11/002,077), "remote controller" (Patent No. 6,675,300), "proximity badge" (Patent No. US 6,189,105), "security device" (Application No. 10/608,459) and "portable device" (Patent No. 6,819,219) refer to the same kind of security device contextually. The word or phrase embeddings should help in determining these kind of similarities between phrases in related patent applications and patents. A Machine Learning algorithm can be developed which trains on this data and can be used to automatically look for related prior art(s) in a vast collection of prior patent applications and patents in the review process of a patent application. This algorithm can also be used to simplify a process of reasoning the rejection of certain disclosed claims in a patent application based on identified prior art(s) by determining the underlying patent laws or statutes. Such an algorithm significantly automates the work performed by a patent examiner which in turn saves valuable time and money.

CHAPTER 10: CONCLUSION

10.1 Contributions

With the main purpose of demonstrating the case studies that can be used in predicting emerging technologies using IR and NLP techniques, and evaluating the sufficiency of available IR and NLP techniques for the same, this project has successfully achieved the following.

- Successfully automated the patent scraping, data cleaning and data creation processes.
- Successfully automated the process of creating the numerical analysis graphs to view the technology growth. All of the scripts are available in a public repository, on Github, which makes reproducing these results smooth.
- Worked on analyzing six case studies of technology emergence.
- Performed Topic Modelling to demonstrate validity of the key terms.

10.2 FUTURE WORK AND EXTENSION

After the successful results of the first hypothesis, I further my hypothesis in being able to mark a pattern of the growth of similarity scores between cross-patents. This pattern, should then be generalized to predict the emergence of two technologies as a function of time.

Bibliography

[1] W. A. Shalaby, W. W. Zadrozny, and K. Rajshekhar, "Natural language relatedness tool using mined semantic analysis," Jan. 30 2018. US Patent 9,880,999.

[2] W. Shalaby and W. Zadrozny, "Learning concept embeddings for efficient bag-of-concepts densification," *arXiv preprint arXiv:1702.03342*, 2017.

[3] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[4] S. Ankam, W. Dou, D. Strumsky, D. X. Wang, T. Rabinowitz, and W. Zadrozny, "Exploring emerging technologies using patent data and patent classification," in *Proc. IEEE VIS Workshop on Interactive Visual Text Analytics*, 2012.

[5] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, and L. Zalányi, "Prediction of emerging technologies based on analysis of the us patent citation network," *Scientometrics*, vol. 95, no. 1, pp. 225–242, 2013.

[6] W. Shalaby and W. Zadrozny, "Mined semantic analysis: A new concept space model for semantic representation of textual data," in *Big Data (Big Data), 2017 IEEE International Conference on*, pp. 2122–2131, IEEE, 2017.

[7] C. A. Eusebi and R. S. Silberglitt, *Identification and analysis of technology emergence using patent classification.* JSTOR, 2014.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] A. Abbas, L. Zhang, and S. U. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, pp. 3–13, 2014.

[10] T. C. Stratopoulos, "Emerging technology adoption and expected duration of competitive advantage," 2016.

[11] A. J. Trippe, "Patinformatics: Tasks to tools," *World Patent Information*, vol. 25, no. 3, pp. 211–221, 2003.

[12] M. G. Moehrle, "Measures for textual patent similarities: a guided way to select appropriate approaches," *Scientometrics*, vol. 85, no. 1, pp. 95–109, 2010.

[13] K.-Y. M. Chen and Y. Wang, "Latent dirichlet allocation," 2007.

[14] K. Younge and J. Kuhn, "Patent-to-patent similarity: a vector space model," 2016.

[15] Y. Bengio, P. Ducharme, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[16] H. Youn, D. Strumsky, L. M. Bettencourt, and J. Lobo, “Invention as a combinatorial process: evidence from us patents,” *Journal of The Royal Society Interface*, vol. 12, no. 106, p. 20150272, 2015.

[17] H. Ernst, “The use of patent data for technological forecasting: the diffusion of cnc-technology in the machine tool industry,” *Small business economics*, vol. 9, no. 4, pp. 361–381, 1997.

[18] B. P. Abraham and S. D. Moitra, “Innovation assessment through patent analysis,” *Technovation*, vol. 21, no. 4, pp. 245–252, 2001.

[19] I. Bergmann, D. Butzke, L. Walter, J. P. Fuerste, M. G. Moehrle, and V. A. Erdmann, “Evaluating the risk of patent infringement by means of semantic patent analysis: the case of dna chips,” *R&d Management*, vol. 38, no. 5, pp. 550–562, 2008.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[21] W. Shalaby and W. Zadrozny, “Innovation analytics using mined semantic analysis.,” in *FLAIRS Conference*, pp. 597–601, 2016.

[22] E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using wikipedia-based explicit semantic analysis.,” in *IJCAI*, vol. 7, pp. 1606–1611, 2007.

[23] Predictive Analytics Today, “Predictive analytics.”

[24] Wikipedia contributors, “Google patents.”

[25] A. Breitzman and P. Thomas, “The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems,” *Research policy*, vol. 44, no. 1, pp. 195–205, 2015.

[26] W. Shalaby, W. Zadrozny, and H. Jin, “Beyond word embeddings: Learning entity and concept representations from large scale knowledge bases,” *arXiv preprint arXiv:1801.00388*, 2018.