Universidade Estadual de Campinas
Instituto de Computação

Pablo Alejandro Fonseca Arroyo

Metric Learning for Patent Similarity

Aprendizado Métrico para Similaridade entre Patentes

CAMPINAS
2015

Pablo Alejandro Fonseca Arroyo

Metric Learning for Patent Similarity

Aprendizado Métrico para Similaridade entre Patentes

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Jacques Wainer

Este exemplar corresponde à versão final da Dissertação defendida por Pablo Alejandro Fonseca Arroyo e orientada pelo Prof. Dr. Jacques Wainer.

CAMPINAS
2015

Agência(s) de fomento e nº(s) de processo(s): CNPq, 133548/2013-9

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

F733m

Fonseca Arroyo, Pablo Alejandro, 1987-
Metric learning for patent similarity / Pablo Alejandro Fonseca Arroyo. – Campinas, SP : [s.n.], 2015.

Orientador: Jacques Wainer.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Mineração de dados (Computação). 2. Patentes. 3. Inteligência artificial. 4. Aprendizado de máquina. I. Wainer, Jacques,1958-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Aprendizado métrico para similaridade entre patentes
Palavras-chave em inglês:
Data mining
Patents
Artificial intelligence
Machine learning
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Jacques Wainer [Orientador]
Eduardo Alves do Valle Junior
Anderson de Rezende Rocha
Data de defesa: 11-12-2015
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Pablo Alejandro Fonseca Arroyo

Metric Learning for Patent Similarity

Aprendizado Métrico para Similaridade entre Patentes

Banca Examinadora:

- Prof. Dr. Jacques Wainer
  Instituto de Computação - UNICAMP
- Prof. Dr. Anderson de Rezende Rocha
  Instituto de Computação - UNICAMP
- Prof. Dr. Eduardo Alves do Valle Junior
  Facultade de Engenharia Elétrica e de Computação - UNICAMP

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 11 de dezembro de 2015

To Gustavo, Vicky, Rocio and Daniel and to the loving memory of my grandmother Francisca.

Lonely star at night, are you alone? or the friends you had, all over the infinite sky, are pretending to hide between the clouds and wind. Lonely star at night, I don't know much of this place, nor the sky above us. Can you tell me more? you must know it well. I can tell about the sky back at home, where once in a while, the cloudy night hides every lonely star that wants to shine. Lonely star at night, you are the one to tell me: what is light? why it hides? and appears suddenly, in the lonely night?

Agradecimentos

I'm convinced that coming to do the Masters at UNICAMP was one of the better things that ever happened to me. Many dreams came true here, studying Machine Learning and Image Processing was a goal since while time ago. So many adventures came together as well as academic knowledge: meeting great friends, traveling and some free time to explore music and painting.

First, many thanks to my advisor, Dr. Jacques Wainer for all the help, understanding and encouragement. Also many thanks to Prof. Isaac Yrigoyen for his involvement in the patent distances research. Also thanks to Dr. Benjamin Castaneda and people at Medical Innovation and Technology and Oncosalud in Peru, specially Jose Ferrer, for the collaboration on the Breast Density project.

Also, infinite thanks to my family: chato, ranita, renoise and rechicken (or better known as Gustavo, Vicky, Daniel and Rocio). You are my life. Your love and support makes every adventure in the world worthwhile. On good days, we will be happy together; on hard ones, you are the light that illuminates the day.

I also want to thank Dr. Juan Arroyo, my uncle, for being a great source of inspiration and my academic reference in the family, and also for teaching me how to program when I was still a kid.

Also many thanks to my aunt Rocio Fonseca, who always encouraged me to pursue new challenges. I think is not late to thank you for giving me my first computer and for supporting me in many dimensions. Distance was never an issue for us.

Also to my relatives living in Brazil, for the support and encouragement and the time we spent whenever we got the chance. Thanks a lot uncle Pedro and aunt Alicia.

Last but not least, thanks to my friends here in Brazil, Peru and everywhere else, without you none of this would have been that much fun. It was never easy to have the heart divided in more than one country, but you made that worth it.

Resumo

Hoje em dia, obter uma melhor visão de um campo de tecnologia é crucial para a estratégia nos negócios, na universidade e no governo. As patentes são uma fonte muito importante de informação ao respeito. A similaridade textual entre patentes é um dos tipos de similaridade em que os analistas de patentes estão interessados, a fim de melhor compreendê-las. As técnicas comuns para medir a similaridade entre documentos de texto incluem representações bag-of-words ou distribuições de tópicos não supervisionadas, em combinação com varias opções possíveis para distâncias. No entanto, estes métodos não incorporam a informação do domínio de conhecimento, que pode ser crucial para um corpus difícil como as patentes são. Nesta dissertação de mestrado, uma abordagem para a aprendizagem de similaridade entre patentes é apresentada. O método utiliza aprendizado métrico e aproveita parte do processo legal que as patentes passam antes de serem concedidas. Os resultados do método proposto foram comparados com distâncias padrão, não supervisionadas como KL-divergence, a distância do coseno e a distância euclidiana com a obtenção de resultados superiores e mais confiáveis.

Abstract

Nowadays, gaining insight into a technology field is crucial for business, academy and government strategy. Patents are a great source of information in this regard. Textual patent similarity is one of the kinds of similarities in which patent analysts are interested in order to better understand them. Common techniques to measure similarity across text documents include bag-of-words representations or unsupervised topic distributions in combination with several possible options for distances. However, these methods do not incorporate information of the domain of knowledge, which might be crucial for approaching the challenging corpus patents are. In this master thesis, an approach for learning pairwise similarity between patents is presented. The method uses metric learning and takes advantage of some of the artifacts of the legal process patents undergo before being granted. The results of the proposed method were compared to standard, but unsupervised, distances (KL-Divergence, Cosine distance and Euclidean distance) obtaining superior and yet more trustful results.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this master thesis, an approach for learning pairwise similarity between patents is presented. Patents are very complex subjects of study because they are legal, and as well, technical documents. Often there are also strategical intentions behind them. Of the many dimensions from which similarity can be obtained, this work is only concerned with textual features; however, patents might also have images, chemical formulas, assembly plans among many other kinds of content.

In consequence, the present work is situated just in a small part of the big intellectual property system. However, given that patent retrieval often starts by text queries, this small part is also of central interest. From the point of view of the machine learning techniques studied, this work is concerned with metric learning, a subfield of supervised learning that aims to learn better distances by incorporating supervision. It also can be situated within the field of text mining; specially, because this work extends standard procedures with metric learning.

The objective of this work is to compare how such an approach improves results over standard similarity or distances and to provide confidence margins on such measures.

1.1 An overview of the patent world

To better understand this work, its important to know what are patents, what makes them important as a subject of study and which characteristics they have that could be exploited for supervised learning. A patent gives its holder the right to exclude others from making, using or selling the invention claimed in the patent deed for approximately 17 to 18 years, provided that certain fees are paid [18]. Patents are a unique kind of documents as they hold both legal and technical value. On one side, their legal importance resides at the protection they confer, as no one else will be allowed to produce the protected invention if the patent is granted. At the same time, the legal value is highly related to the technical one: patents need to describe in a precise manner the inner details of the inventions in order to warrant complete protection and thus their content is also highly technical.

The process that ends up with a granted patent is very complex, but more or less similar across many countries due to international treaties, a fact exploited for this re-

search. The process starts with a patent application. From that point, it undergo a strict legal process that involves patent analysts to evaluate if the presented invention meets the patentability requirements. One of the most important requirements is novelty: no one else should have presented it before.

A patent application has four important sections, that vary a lot in how they are written. This list is presented in table 1.1.

Section	Content
Abstract	A reduced description of the invention.
Description	Description of the invention.
Claims	What is protected by the patent.
Citations	Reference to other patents or scientific literature.

Table 1.1: Patent sections

The Abstract is regarded to be the most valuable and informative section and its written for general information. The Description is technical and claims are legal. Citations on the other hand are bibliographical information and can be generated by the applicant as well as the patent analysts reviewing the application at IP offices.

1.2 Challenges in patent analysis

Patents are a key resource while analyzing the development of technology both in academia and industry. Moehrle [15] points out that patent analysis is important for technological management, however it presents the three big challenges: (I) the number of patents in the world grows steadily, (II) trying to understand a patent is a time consuming task, that can be handled only with considerable manpower and (III) patent analysis at intellectual property offices is not as good as it could be [4] [17]. For those three reasons, Moehrle [15] states that the usage of automatic tools for patent analysis seems useful. Moreover, patents are tricky because of the usage of non standard terms when a technology is at an early stage of development: there is not standardization yet producing high variability within the names used to describe elements in the technology [12].

Common patent analysis tasks are **Prior art analysis**, or finding similar documents to a new patent document which was not presented before. **Infringement analysis** is concerned with finding other overlapping patents, starting with an infringed patent. **Patent mapping** aims to use a matrix of similarities for getting insight into a landscape of patents.

Textual similarity is defined as a form of association, relationship or resemblance which is based on textual elements within patent documents. Textual similarity, therefore implies some shared or common textual elements across patents. However, it does not guarantee that the purpose of two described inventions are similar.

1.3 Search Report

The search report is an artifact produced by part of the legal process patent applications undergo at the European Patent Office before being granted. The search report produces a list of citations the patent analyst might found of interest in order to question the novelty of a given patent application.

1.3.1 Coded Citations

The references that appear in patents may be added to the document for different reasons, at different times and by different people [13]. For the purposes of this thesis, patent citations can be divided into two main groups. Those made by the applicant and those made by patent analysts at the IP office to which the application was presented. Citations made by analysts have the aim of pointing out possible conflicting patents presented before. However, other kind of citations might appear as well, for instance those related to the technological background of the invention. Coded citations present valuable information for patent analysis as they offer a categorization of the citations that might represent a notion of ranked relative similarity among them. In the table 1.2 the codes along their meaning are presented.

Code	Meaning
&	Corresponding document (from the same family)
A	Technological background
D	Document cited in the application
E	Earlier patent document, but published on or after the filing date
L	Document cited for other reasons (miscellaneous category) Non-written
O	Non-written disclosure
P	Intermediate document
T	Theory or principle underlying the invention
X	Particularly relevant
Y	Particularly relevant, when combined with another document

Table 1.2: Coded Citations [13]

Relevant patents for questioning novelty are those marked with **X** or **Y** in the citations. Patents marked with **&** are expected to be very similar as well, as they belong to the same family of patents. A family of patents protect the same invention, but the same company or person, but might have been presented to different patent authorities.

1.4 The PATSTAT Database

The PATSTAT database[1] is published twice a year by the European Patent Office (EPO). It has about 80 million patents, from several patenting authorities worldwide including the EPO, USPTO and JPO. It worth noticing that it only includes the abstracts of the

[1] http://www.epo.org/searching/subscription/raw/product-14-24.html

patents. Information regarding claims only include the number of them. No text is available for the description as well. However, the citation network is available along the coded citations, making feasible to obtain relative similarity information from them.

Chapter 2

Approaches to patent similarity

Patent similarity research have targeted different kinds of similarity. The aim was always to take advantage of the information in patents as well as in the artifacts produced by the legal process they undergo at patent offices. Elements in patents are rich: bibliographical, textual, graphical content as well as citation networks are sources for the analysis. Textual similarity, the main focus of this work, is just one the kinds of similarities in which patent analysts are interested. How textual similarity is used in the world of patent analysis is discussed extensively by Moehrle [15]. That work specially, has served to this research as a theoretical framework.

Moehrle addresses the measurement of textual patent similarities, stating that they are crucial for the most important tasks in patent management discussed in the previous chapter: prior art search, infringement analysis and patent mapping. The main motivation to pursue research in that field is related to the aspects that reduce the technology manager ability to deal with patents efficiently.

Similarity can be defined as an increasing function of commonality and a decreasing function of differences among the compared objects [9]. Patent similarity has two levels for Moehrle.

- **Formal oriented level similarity** - regarding "formal" elements, such as the text of a patent, or the included images. It is related to how it was presented.
- **Content oriented level similarity** - regarding the described elements, they true nature: purpose, which problem it solves. It is related to the idea behind the presented description.

Textual similarity is, of course, just an instance of the **formal oriented level similarity** on this two-level similarity model. Both levels of similarity are connected; however, a high similarity of the purpose of two inventions (that is: their technological advantage or even the problem they intend to solve) does not necessarily lead to a high textual similarity and viceversa. In this thesis, the objective is try to overcome that limiting fact.

2.1 A high level view of previous approaches to the problem

To the best of the knowledge gained in the earlier stages of the research, the techniques used for approaching the problem can be divided into three kind of approaches:

- **Text mining** - These techniques are well known and include: bag-of-words representation, were each document is represented by its histogram of words of n-grams (words that appear together), stemming (for obtaining the root of a term), stop-word removal (for eliminating common words that don't offer new information) as well as TF-IDF weighting scheme and cosine distances.
- **Semantic tagging** - are mostly related to the extraction of SAO structures (Subject-action-object) which for the authors working within this approach encode well the technological information in patents; for instance "new device (Subject) performs transformation (Action) on matter (Object)".
- **Topic modeling** - include both Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) based techniques.

|

All the papers reviewed in the section 2.2 follow one of these approaches.

2.2 Literature Revision

Below a list of relevant research on patent similarity is presented, nevertheless, not all of the reviewed papers are related to textual patent similarity which is the actual target of this research.

- **Sternitzke et al. (2008)** - Similarity measures for document mapping: A comparative study on the level of an individual scientist [19].
- **Li et al. (2011)** - Extracting the significant-rare keywords for patent analysis [12].
- **Moldovan et al. (2005)** - Latent Semantic Indexing For Patent Documents [16].
- **Magerman et al. (2010)** - Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications [14].
- **Bergmann et al. (2008)** - Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips [2].
- **Choi et al. (2012)** - An SAO-based text mining approach to building a technology tree for technology planning [6].
- **Tang et al. (2012)** - PatentMiner: Topic-driven Patent Analysis and Mining [20].

In the following subsection each one of the papers is briefly reviewed.

2.2.1 Sternitzke et al. (2008) - Similarity measures for document mapping: A comparative study on the level of an individual scientist

This paper reports to address the mapping of documents (not precisely patents) following the traditional approaches in the bibliometric community. The approach is largely similar to the one presented in Bergmann et al. [2]. It can be summarized in three steps.

- **Step 1:** Bibliographic elements are selected for serving as a basis for comparing documents: backward citations, forward citations, words as item sets to describe similarity.
- **Step 2:** Similarities are computed based on the above mentioned items. They mention in this step: Pearson correlation coefficient, Salton's cosine formula, Jaccard's index and the Inclusion index.
- **Step 3:** The computed distances are then visualized with the help of cluster analysis and multidimensional scaling (MDS).

This paper is concerned with mapping, which basically is projecting similarity information onto a 2D map.

2.2.2 Li et al. (2011) - Extracting the significant-rare keywords for patent analysis

In this paper, authors propose a version of the traditional TF-IDF weighting using the number of assignees (companies that hold the rights of a patent) to weight how "popular" is a keyword within a set of companies. It worth noticing, that this paper contributes a weighting scheme to discover significant but rare (not frequent) terms. Some techniques based on keywords or bag-of-words could benefit from this scheme.

2.2.3 Magerman et al. (2010) - Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications

This paper explore the usage of Latent Semantic Analysis (LSA) for similarity measurement between patents and papers. The idea is to discover which patents are related to which scientific paper for a particular academic inventor. The setup of the experiment included building a dataset of patents and scientific articles from six academic inventors in the same institution as the authors. Field experts were asked to evaluate the relatedness of patent documents and scientific papers based on three possible categories: "highly related", "unrelated", and "somewhat related". Then, the scores of each patent-paper pair were compared using the kappa metric.

The result of this experiment yielded interesting conclusions:

- SVD performed worst under all circumstances, especially with a limited number of dimensions.
- For a small dataset, parameter options that respect the richness of the underlying data and also the application of weighting schemes produce better results.
- For a set of small datasets, a global applied indexing and dimensionality reduction does not yield worst results than a per-case-based approach. The explanation for the bad performance of SVD within the experiment is the small number of documents in the sample.

2.2.4 Moldovan et al. (2005) - Latent Semantic Indexing For Patent Documents

This paper uses Latent Semantic Indexing (LSI) with Single Value Decomposition (SVD) for bag-of-words representation of patent documents. In this paper, the aim is to retrieve documents of the same patent class using only textual information. For the evaluation of the method the precision-recall metric was used and documents that share at least one patent class were considered relevant. They report to obtain better results with 80 singular values contrary to what they found in literature (recommended between 100 and 300). Another interesting conclusion is that LSI produced an improvement of 5% in average for seven of the ten classes in the dataset.

2.2.5 Bergmann et al. (2008) - Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips

This paper presents an analysis of patent infringement detection by using a patent map, built up with distances computed by the use of Semantic Tagging and SAO structures extraction as described below.

- Stage 1: SAO structures are extracted from patent documents using a semantic processor.
- Stage 2: The usage of domain-specific speech filters is intended for standardization and minimization of the highly differentiated language of the domain; SAO structures are then modified to use synonyms (synonymizing filter) and concept hierarchies (generalizing filter) into account.
- State 3: The similarity is measured by the number of identical SAO structures.
- Stage 4: Determination of the significance of the similarity coefficients.

For the visualization of the data, they use the multidimensional scaling (MDS).

2.2.6 Choi et al. (2012) - An SAO-based text mining approach to building a technology tree for technology planning

This paper focus on building "Technology Trees" (TechTrees), a branching diagram that represents relationships among technologies in its different kinds: product taxonomy trees, technology taxonomy trees and function taxonomy trees. Even if this problem is not strictly close to patent similarity measures, the approach taken by the usage of SAO structures is of major interest for this research because it proposes similarity measures between SAO structures after being obtained from a patent.

To perform similarity measure of SAO structures, a Wordnet-based sentence similarity measure is used: First sentences are tokenized, then words are stemmed, then part of speech tagging is performed, determining the most likely meaning of each sentence.

2.2.7 Tang et al. (2012) - PatentMiner: Topic-driven Patent Analysis and Mining

This paper presents a topic modeling technique very similar to the Latent Dirichlet Allocation (LDA) method. They add observed values to the topic inference: inventor and company. The model tries to describe the generative process of patent writing. With that in mind, it assumes that a patent document has a vector of words that it was developed by a group of inventors (the way words are chosen reflects the expertise of these inventors) and its owned by a company (topics suggested are relevant to the company that owns the patent).

Chapter 3

Metric Learning

Metric learning is an actively researched topic in Machine Learning. It aims to learn task specific distance functions in a supervised way [10]. The area started back in 2003 with the paper by Xing et al. [21], however it can be traced back to earlier works [1]. To the date there are three main surveys that sums up the development in the field. In 2006, a review by Yang [22] and more recently in 2012 the survey by Kulis [10] and in 2013 the survey by Bellet et al. [1]. This section is mainly based on these three surveys.

3.1 Overview

Metric distance learning aims at improving prediction capabilities of machine learning algorithms that are dependent on distances or similarity measures. The learning process tries to capture the idiosyncrasies of the data in order to parametrize a standard metric that should behave better than general purpose ones. In that sense, metric distance learning is a supervised learning task. More often, this supervision is weak. For many problems, explicit pairwise distances are not available neither from experts in the domain field and thus relative similarity training samples are used by many algorithms to learn the metrics.

3.1.1 Pairwise similarity

Bellet et al. [1] describe the importance of having pairwise similarity functions as highly important in machine learning. For instance, traditional algorithms such as k-nearest neighbors and k-means depend on the measurement of distances between data points, and therefore, the performance of the setup is highly dependent on the used metrics. General purpose distances exist: euclidean distance, cosine distance, earth movers distance. However, they are not application-specific and do not incorporate supervision.

3.1.2 Training samples

In metric distance learning, the supervision is obtained from three kinds of training samples:

- **Positive/negative pairs** (eg. belongs or not to a certain class). In positive/negative pairs the algorithm will have two sets available: S and D, where $S = \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be similar }\}$ and $D = \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be dissimilar }\}$.
- **Relative similarity triplets** (a more similar object, a less similar object given a starting object). Relative similarity set R, where $R = \{(x, x^+, x^-) : x \text{ is more similar to } x^+ \text{ than to } x^-\}$.
- **Quadruplets made by two pairs**, with a known relative similarity between the inner groups. The idea of quadruplets was introduced by Law et al. [11]. Each training sample is a quadruplet or two pairs were the similarity of a pair should be greater that the similarity between the elements of the other pair.

3.1.3 Optimization

Most machine learning algorithms can be seen as optimization problems. In the case of metric learning, the aim is to produce distances that conform to certain properties. Bellet [1] argues that most state-of-the-art methods in metric learning fit in the following optimization problem form.

$$\underset{M}{\text{minimize}} \quad \ell(M, S, D, R) + \lambda R(M)$$

where $\ell(M, S, D, R)$ is a loss function that has penalties when training constraints are not met, $R(M)$ a regularizer, and M the learned metric. Most of the times, M is an standard distance, such as the euclidean or the cosine distance, however parametrized.

One specific parametrization of great interest for this research is the Mahalanobis-like distances. Details will vary from one approach to other: specially how loss function is constructed and which regularization is used; however, the basic construction is shown below.

$$d_{mahalanobis}(M, x, x') = \sqrt{(x - x')^T M (x - x')}$$

Where M is the parametrization in the form of an square matrix with d^2 elements where d is the dimension of the features vectors used. If M is a positive semidefinite, the properties of a pseudo-distance are respected. Moreover, M induces a linear projection of the feature space where constrains are better respected using the euclidean distance that in the original one. To prove this, consider the following way of writing the euclidean distances between x and x'.

$$d_{euclidean}(x, x') = \sqrt{(x - x')^T (x - x')}$$

Replacing $M = L^T L$ in the expression for $d_{mahalanobis}$ we obtain the following rewriting of the distance.

$$d_{Mahalanobis}(M, x, x') = \sqrt{(x - x')^T L^T L (x - x')}$$

which lead us to

$$d_{Mahalanobis}(M, x, x') = \sqrt{(Lx - Lx')^T(Lx - Lx')}$$

which can be seen as the euclidean distance, but computed on x and x' on a linear projection induced by matrix L.

Another distance of great interest is the bilinear distance. It is also parametrized by the W matrix, and of course, the trick described above is also applicable. It is the distance used in [5] which is the base for the algorithm presented in this chapter.

$$d_{bilinear}(W, x, x') = x^T W x'$$

which can also be written as

$$d_{bilinear}(W, x, x') = (Lx)^T(Lx')$$
$$W = L^T L$$

which is helpful for obtaining an alternative feature space were inner product makes sense to the problem as a measure of distance. It worth noticing that the cosine distance can be written as.

$$d_{cosine}(x, x') = \frac{x^T x'}{||x||||x'||}$$

which is also parameterizable.

3.2 The OASIS algorithm

OASIS stands for Online Algorithm for Scalable Image Similarity. This method was proposed by Chechik et al. [5]. It was tailored to work with images and aiming an scalable approach, however it is applicable to a wide range of feature vectors. It learns bilinear distances in an online way, which is one of the methods to deal with scalability according to [1]. OASIS can handle web scale datasets in order to learn semantic representations over feature vectors of images. Authors, who are concerned with retrieval, pointed out that the distance can be computed in $\mathbb{O}(k_1 k_2)$ when the number of non-zero entries in sparse vectors x_1 and x_2 are k_1 and k_2 respectively. This, regardless of the size dxd of the matrix.

Authors has made the code available online [1]. The way used to produce PSD matrices in an online setup is to project the matrix, after a number of iterations, to the PSD cone, using the nearest point in the cone according to the Frobenius norm. This operation implies that an eigenvector decomposition operation is performed. More on this procedure is discussed in [8].

The OASIS algorithm is based on a family of algorithms called passive-aggressive [7]. It uses the hinge loss in the following way, given the relative similarity triplet.

$$l_W(p, p^+, p^-) = max(0, 1 - dist_W(p, p^+) + dist_W(p, p^-))$$

[1] http://ai.stanford.edu/gal/Research/OASIS/

Algorithm 1 OASIS

$n \leftarrow NumberOfSteps$
$W \leftarrow I$
$C \leftarrow 0.01$
$EnforceSimetry \leftarrow True$
$i \leftarrow 0$
while $i \leq n$ **do**
  $x \leftarrow sample(element)$
  $x^+ \leftarrow sample(similar)$
  $x^- \leftarrow sample(dissimilar)$
  $l_W \leftarrow max(0, 1 - dist_W(x, x^+) + dist_W(x, x^-))$
  $V^i \leftarrow [x^1(x^+ - x^-), ..., x^d(x^+ - x^-)]^T$
  $T_i \leftarrow min(C, l_W / \|V^i\|^2)$
  $W \leftarrow W + T_i V^i$
  if $EnforceSimetry$ **then**
    $W \leftarrow (W + W^T)/2$
  end if
  $i++$
end while

where $dist_W(x, x') = x^T W x'$ is the parametrized bilinear distance. This is the loss for a single triplet. The goal is to minimize the global loss L_W.

$$L_W = \sum_{(p,p^+,p^-) \in P} l_W(p, p^+, p^-)$$

If $l_W(p, p^+, p^-) = 0$ it means that the constraint has been already satisfied, therefore in an online algorithm the matrix W^i and W^{i-1} will be the same. However, if $l_W > 0$ a lagrangian is defined as follows.

$$\mathcal{L}(W, \tau, \xi, \lambda) = \frac{1}{2}\|W - W_{i-1}\|^2 + C\xi + \tau(1 - \xi - p^T W(p^+ - p^-)) - \lambda\xi$$

The optimal solution is found when the gradient vanishes ($\frac{\partial \mathcal{L}(W,\tau,\xi,\lambda)}{\partial W} = 0$).

$$\frac{\partial \mathcal{L}(W, \tau, \xi, \lambda)}{\partial W} = W - W^{i-1} - \tau V_i = 0$$

where the gradient matrix $V_i = [p^1(p^+ - p^-), ..., p^d(p^+ - p^-)]^T$

Chapter 4

A metric learning approach to patent similarity

Often machine learning algorithms require representation of objects as feature vectors in order to predict some behavior or conduct some simulation on those objects. A classic way of representing text documents for text mining algorithms is to use a histogram of the terms appearing in the document. This is also known as bag-of-words representation or vector space representation. However, this way of describing a text document often lead to very high dimensional feature vectors, which might present two problems: metric learning often learns a matrix with as many elements as the squared dimensionality of the feature space and that over-fitting is more likely to appear in such high dimensional spaces.

In metric learning a dimensionality reduction technique is often applied, algorithms such as PCA o K-PCA [1] usually improve the behavior of the learned metric. However, for text documents there are other techniques that might be of interest as well. Those techniques, such as topic modeling, have not yet been explored in the literature. Latent Dirichlet Allocation (LDA) can learn more succinct representations of text documents: probability distribution of topics, where the topics are inferred in an unsupervised manner and often can be seen as a dimensionality reduction technique as well.

4.1 Training data from the search report

Explicit distances for patents are not available in datasets. Moreover, even specialists might not be capable of producing an specific distance value for textual similarity between two patents. For strategical reasons the textual information might not be precise. in order to gain relative similarity information that could be used foor building a metric learning triplet sample we must see the search report that produces a list of coded-citations. This report is made by documents an analyst found to be relevant for assessing the novelty of a patent. If a patent is very similar to a pre-existing one it is going to be marked with a **X** code, if it is relevant but in combination with other documents, it will get an **Y** code. There are other codes as shown in table 1.2, for instance, the **A** code means that the cited document belongs to the technological background of the invention. We propose to build

a general ranked structure from those coded citations. This ranking is shown in table 4.1.

Rank	Codes
1	**& - Same family of patents**
2	**X - Very relevant document**
3	**Y - Relevant document in combination with other documents**
4	**Other Cited documents**
5	**Other Documents in the class/subclass**

Table 4.1: Relative similarity from coded citations

As mentioned in section 3.1.2, training samples for metric learning algorithms can be pairs, triplets or quadruplets and should present relative similarity constraints. With the ranked similarity information that can be derived from the coded citations of the Search Report, it is possible to build training samples for metric learning algorithms. In the triplet scenario, each training sample becomes $(patent, patent^+, patent^-)$ where a *patent* and two related patents are supplied: $patent^+$ that should be more similar to *patent* than $patent^-$. As shown in equation 4.1, these triplets can be built using the relative similarity information available in the coded citations.

$$d(patent, patent^+) < d(patent, patent^-) \tag{4.1}$$

That way, a patent cited with code **&** is expected to be more similar because is from the same patent family. **X** is expected to be highly relevant, and sure more relevant than **Y** citations. Then, other cited documents in the patent should be more relevant than other documents in the IPC class or subclass on which the training is performed. Although this citations are sometimes coded on a claim basis, the PATSTAT database shows them at the level of citations.

4.2 Topic modeling: lower dimensional text features

Topic models such as Latent Dirichlet Allocation (LDA) presented by Blei et al. [3] are inference models that obtain in an unsupervised way a latent topic distribution based on observed words per document. LDA is parametric method, were the number of topics to infer is one of them. Also the α prior to the Dirichlet distribution and also its know to control the sparsity of the produced topic distributions.

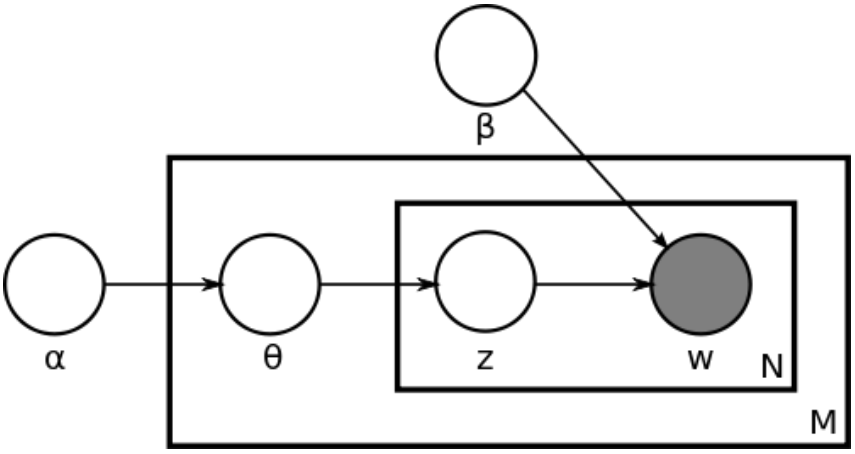

Figure 4.1: Latent Dirichlet Allocation Graphical Model. α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ_i is the topic distribution for document i, φ_k is the word distribution for topic k, z_{ij} is the topic for the j-th word in document i, and w_{ij} is the observed word.

LDA can also be seen as a robust text feature as there is a lot of noise that can be expected in Bag-of-words models due to non-standard use of words, the usage of synonyms, and the like. It also can be seen as dimensionality reduction technique working on an unsupervised way. It also has running time advantages over other methods such as Latent Semantic Analysis (LSA).

Chapter 5

Experiments, results and conclusions

In this chapter, the experiments are described. A tricky patent subclass even for specialists was used to validate the results. An overview of the experimental protocol is shown below.

- Text features will be LDA@100 topics using just the abstracts of roughly 5k patents from the C12N subclass obtained with the SQL Query presented in Apprendix A in the October 2013 edition of the PATSTAT Database.
- 200 patents and their citations will be used for validation only and will not be part of the training.
- OASIS will be trained with a sampling strategy designed for patents for forming triplets on the fly.
- Two sets of patents are defined for each query patent in the validation set. A set of relevant patents made by the cited patents and a non-relevant set make with 50 randomly sampled patents from the C12N class.
- The biserial correlation is computed using these two sets for the bilinear distance (trained with OASIS), KL-Divergence, Cosine Distance and Euclidean Distance, also for a control group of random generated distances.
- Confidence intervals are computed by means of the bootstrapping technique.

5.1 Experimental protocol

In order to validate the performance of the developed method, the C12N subclass was chosen. It is known to be a difficult class because it is related to biotechnology. Therefore, an interesting field for evaluating the performance of the method. In the figure 5.1 the overview of the method is presented. At the last stage, the biserial correlation is measured for 4 distances: (1) The OASIS Algorithm parametrized distance, (2) The Euclidean distance, (3) the Cosine similarity and (4) The KL-divergence, which is used standardly for distances between inferred topic distributions. Any cited patent is considered relevant, while not cited ones are considered irrelevant for measuring the correlation coefficients.

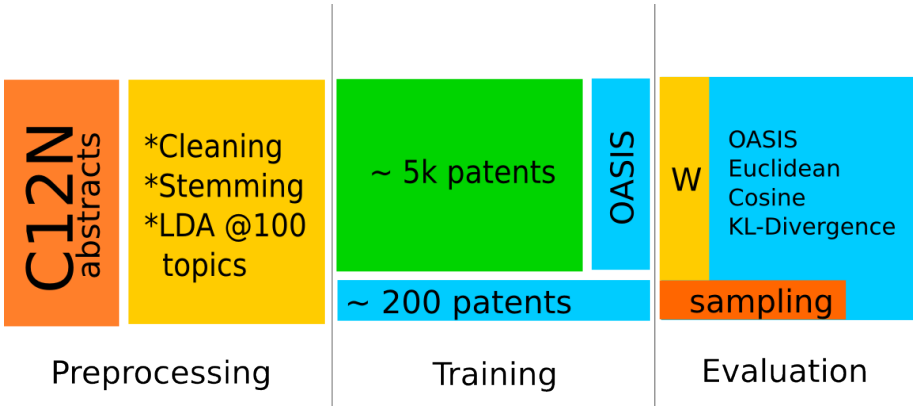

Figure 5.1: Overview of the method and the validation protocol. Here the C12N abstracts extracted from the PATSTAT database is preprocessed for obtaining a 100 topics feature vector. Then the dataset is splited and the block extracted for testing is not visible to the training stage. Finally, while evaluating sampling is performed on the reduced subset multiple times and OASIS, Euclidean, Cosine and KL-divergence is computed and the correlation with relevance is measured.

The C12N patent subclass

The C12N subclass include patents related to micro-organisms, DNA technology and the like. The list below is a more comprehensive one as reported by USPTO [1].

- Micro-organisms (e.g. protozoa, bacteria, fused plant cells, hybridomas, viruses, animal cells or tissue, stem cells, tumour cells) and enzymes or proenzymes and compositions containing micro-organisms and enzymes or proenzymes.
- Processes for preparing, activating, inhibiting, separating, or purifying enzymes.
- Treatment of micro-organisms or enzymes with electrical or wave energy.
- Processes of reproducing, maintaining, or preserving microorganisms or compositions thereof.
- Processes of preparing or isolating a composition containing micro-organisms.
- Preparing mutants and screening processes therefor.
- Processes of fusing two or more cells to each other.
- Recombinant DNA-technology including:
- Processes for manipulating genetic material;
- Processes of preparing, isolating and purifying nucleic acids;

[1] http://www.uspto.gov/web/patents/classification/cpc/html/defC12N.html

- Methods for the introduction of genetic material into microorganisms using vectors or other expression systems, using micro-encapsulation, using micro-injection, and other ways;
- Methods of regulating gene expression;
- Non-coding nucleic acid sequences, e.g. Promoters, operators, enhancers, suppressors, silencers, locus control regions, antisense nucleic acids, and aptamers, used in regulating gene expression or in other recombinant DNA technology related methods.
- Genes, per se; and vectors and expression systems, per se.
- Media for supporting or sustaining the growth of micro-organisms.

Biotechnology patents has been regarded as more difficult to evaluate, even for patent analysts, and thus it can help while evaluating the presented approach. In order to obtain the citations along the code and the patent abstracts in English the following SQL code was run on the October 2013 version of PATSTAT Database from the European Patent Office.

Text Features

The presented approach uses the LDA topic distribution per document as feature vectors for the English abstracts of the retrieved patents. It worth noticing that all the documents must be in the same language for topic extraction. Because of that, the query (shown on Appendix A) used to extract the patents from the PATSTAT database filters out abstracts in other languages. LDA is a parametric algorithm, for this experiment the number of topics chosen is 100.

Sampling strategy

In order to provide triplets to the OASIS algorithm, and without labeling, the relative similarity is derived from the citations. Coded citations can be ranked on a similarity basis. On training time, a list of patents of the same subclass constitute a pool of patents. The original input of the OASIS algorithm are label images; then, sampling two elements from the same class and another from other class would provide an easy way to obtain triplets. However, in this case, the coded citations are used. The method for sampling is shown below.

Algorithm 2 Sampling Strategy

$p \leftarrow SampleRandomPatent(pool)$
$p^+ \leftarrow SampleRandomPatent(citations(p))$ // X, Y, other codes (in this order)
$p^- \leftarrow getLowerSimilarityCitation(citation\_code(p^+), citations(p))$
if $isEmpty(p^-)$ **then**
  $p^- \leftarrow SampleRandomPatent(pool)$ //Sample random patent from the pool
end if

Patents without enough coded citations are removed from the pool in order to guarantee that the algorithm above will work.

Biserial Correlation

The biserial correlation is a measure of relationship between two scores: a continuous variable and a dichotomous variable, with the requirement that the last one could be regarded as fundamentally continuous. The continuous variable is not required to be normally distributed and the dichotomous one is a categorical variable with two possible classes. For the purposes of evaluating the performance of the presented method (which means to evaluate how well the learned distance correlates with the actual relevance of another patent for prior-art-search) the distance is chosen as the continuous variable and the relevance is dichotomized as following: cited patents are considered relevant while non-cited patents are considered irrelevant.

5.2 Results

The OASIS algorithm was run on the training set using sampling with reposition approximately 10^6 triplets from the pool. This process generated the W matrix shown on figure 5.2. The matrix is presented as a heat-plot in order to make sense of the high dimensional 100x100 matrix. Some qualitative effects can be seen on the graphic, for instance, the stronger the diagonal, the more close to the cosine distance. A higher magnitude in a given cell will imply that the interaction of both dimensions (or topics in our case) is more important for determining the distance between the feature vectors.

It is in this part that the supervision improves the correlation with the relevance. The biserial correlation coefficients obtained for the bilinear distance trained with the OASIS algorithm is shown in the table below.

Distance Function	**C.I. for Mean Biserial Correlation (95%)**
Bilinear (OASIS)	[-0.93, -0.87]
KL-Divergence	[-0.42, -0.31]
Euclidean	[-0.39, -0.28]
Cosine	[+0.14, +0.23]
Random Distances *(for control)*	[-0.03, +0.09]

Table 5.1: Distance comparison: For a set of 200 unseen documents at training stage, the correlation between two variables was computed. (I) Distance - a continuous variable produced by the proposed distance and a few other standard distances such as: KL-divergence, Euclidean and Cosine Similarity and (II) the dichotomous variable of relevance - a patent was considered relevant if it was cited. Patents belong to the C12N subclass. It can be seen that the Bilinear Distance trained with OASIS produce higher and less disperse correlation with the dichotomous variable of relevance. The confidence interval was obtained by bootstrapping.

The distance obtained by the presented method is higher and less disperse than the

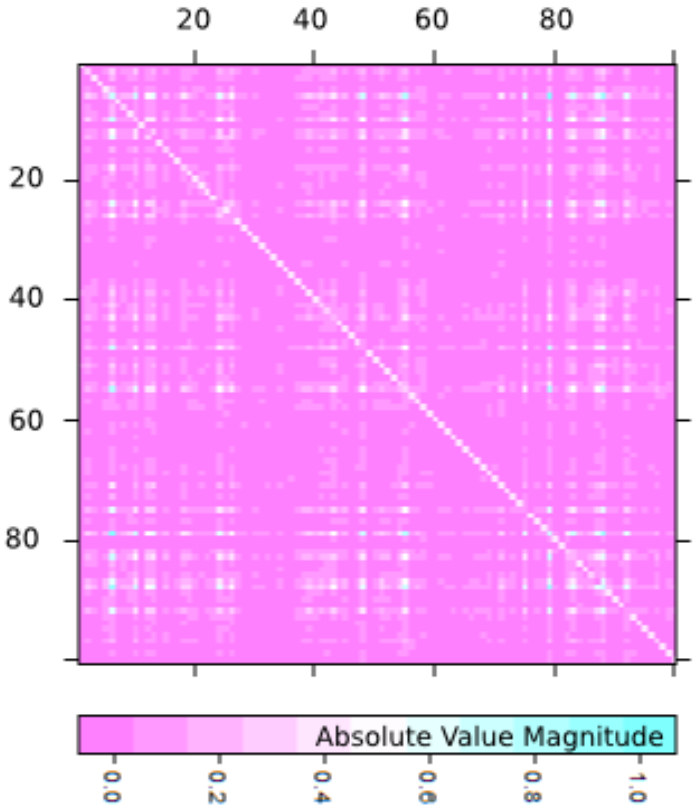

Figure 5.2: Learned matrix W for the C12N class. The stronger the diagonal, the more close to the cosine distance. A higher magnitude in a cell implies that the interaction of both dimensions its more important for determining the distance between the feature vectors.

ones produced by KL-Divergence, Euclidean, Cosine. Random distances were also used as a control measure.

5.3 Conclusions

Patents are documents that contain: text, bibliographic information and images. Its sections are written very differently and targeting distinct sides of its dual nature as technical and legal documents. They are very powerful documents with direct impact in the society, that why its writing might be of strategical interest to organizations. All of the above mentioned plus careful use of language and images would even make it difficult even for specialists, and therefore it is a greater challenge for computer algorithms. In this chapter, the problem was approached from a text similarity side. Classic text features such as topic distributions were combined with metric learning deriving relative similarity information from part of the legal process patents undergo at the European and World Patent Offices: The search report. This information was of vital importance for the metric learning based method, as the supervision comes from a sense of ranking between coded citations. Moreover, it allowed to avoid the need for building a custom dataset and requiring specialists inputs. It is known that patent analysts take more or less a day to produce a search report. On the presented C12N subset it will imply of more or less one year and a half of specialists time in order to build the dataset.

Results have shown that obtaining semantic information from coded citations is feasible and valid, and moreover, very promising. The distinction made by Moerhle [15] about the textual similarity and the purpose of the invention might be a little bit shortened by incorporating the metric learning algorithm to parametrize common distances.

Bibliography

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. page 59, June 2013.

[2] Isumo Bergmann, Daniel Butzke, Lothar Walter, Jens P Fuerste, Martin G Moehrle, and Volker A Erdmann. Evaluating the risk of patent infringement by means of semantic patent analysis: The case of dna chips. *R&d Management*, 38(5):550–562, 2008.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] Paul F Burke and Markus Reitzig. Measuring patent assessment quality—analyzing the degree and kind of (in) consistency in patent offices' decision making. *Research Policy*, 36(9):1404–1430, 2007.

[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. An online algorithm for large scale image similarity learning. In *Advances in Neural Information Processing Systems*, 2009.

[6] Sungchul Choi, Hyunseok Park, Dongwoo Kang, Jae Yeol Lee, and Kwangsoo Kim. An sao-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, 39(13):11443–11455, 2012.

[7] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.

[8] Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103 – 118, 1988.

[9] Buhwan Jeong, Daewon Lee, Hyunbo Cho, and Jaewook Lee. A novel method for measuring semantic similarity for xml schema matching. *Expert Systems with Applications*, 34(3):1651–1658, 2008.

[10] Brian Kulis. Metric learning: A survey. *Foundations & Trends in Machine Learning*, 5(4):287–364, 2012.

[11] Marc T Law, Nicolas Thome, and Matthieu Cord. Quadruplet-wise image similarity learning, 2013.

[12] Yan-Ru Li, Leuo-Hong Wang, and Chao-Fu Hong. Extracting the significant-rare keywords for patent analysis. *Expert Syst. Appl.*, 36(3):5200–5204, April 2009.

[13] Jane List. An A to X of patent citations for searching. *World Patent Information*, 32(4):306–312, December 2010.

[14] Tom Magerman, Bart Van Looy, and Xiaoyan Song. Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2):289–306, 2010.

[15] Martin G Moehrle. Measures for textual patent similarities: a guided way to select appropriate approaches. *Scientometrics*, 85(1):95–109, 2010.

[16] Andreea Moldovan, Radu Ioan Bot, and Gert Wanka. Latent semantic indexing for patent documents. *International Journal of Applied Mathematics and Computer Science*, 15(4):551, 2005.

[17] Minoo Philipp. Patent filing and searching: Is deflation in quality the inevitable consequence of hyperinflation in quantity? *World Patent Information*, 28(2):117–121, 2006.

[18] David Pressman. *Patent It Yourself, 11th Edition*. Nolo, 11 edition, 2005.

[19] Christian Sternitzke and Isumo Bergmann. Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, 78(1):113–130, 2008.

[20] Jie Tang, Bo Wang, Yang Yang, Po Hu, Yanting Zhao, Xinyu Yan, Bo Gao, Minlie Huang, Peng Xu, Weichang Li, et al. Patentminer: topic-driven patent analysis and mining. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1374. ACM, 2012.

[21] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance Metric Learning with Application to Clustering with Side-Information. In S Becker, S Thrun, and K Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 521–528. MIT Press, 2003.

[22] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.

Appendix A

SQL Query to obtain the patent abstracts and citations from the C12N subclass

```
SELECT distinct
CC.citn_categ, P.appln_auth, P.appln_nr, A.appln_abstract,
P2.appln_auth, P2.appln_nr, A2.appln_abstract
FROM tls212_citation C
inner join tls201_appln P ON P.appln_id =  C.pat_publn_id
left outer join tls215_citn_categ CC on CC.pat_publn_id=P.appln_id
and C.citn_id = CC.citn_id
inner join tls203_appln_abstr A ON  A.appln_id = P.appln_id
inner join tls201_appln P2 ON C.cited_pat_publn_id = P2.appln_id
inner join tls203_appln_abstr A2 ON P2.appln_id = A2.appln_id
inner join tls209_appln_ipc IPC on P.appln_id = IPC.appln_id
WHERE
IPC.ipc_subclass_symbol = 'C12N' and
(P.appln_auth = 'EP' OR P.appln_auth='WO') and
(P.appln_abstract_lg = 'EN' and P2.appln_abstract_lg = 'EN')
ORDER BY
P.appln_auth, P.appln_nr
```