Mining for spillovers in patents

Jacob Calder,[*] Iwo Dubaniowski ,[†] C. Vic Hu ,[‡] Subhashini Venugopalan[§]

EE380L Data Mining
University of Texas at Austin
https://github.com/cvhu/ee380l-ghosh-project

March 5, 2013

1 Introduction

The project proposed by Professor Rai involves analyzing a set of approximately 1000 patents that pertain to solar energy technologies obtained from the US Patent and Trademark Office[1]. The information provided for each patent contains the inventor(s); assignees; dates of filing, application, and issue; keywords based on standard patent classifications; geographical location of the assignee firm or inventor; etc. The patents have been pre-classified based on the portion of the supply chain the patents apply to. Additionally, the data has undergone simple topic modeling/LDA analysis based using the MAchine Learning for LanguagE Toolkit (MALLET) Java-package[2].

2 Motivation

The overall goal of the project is to develop a method of clustering the patents that will allow for the discovery of temporal and spatial connections between groups of patents linked to spillovers in innovation. In other words, it is not uncommon for an innovation to leads to other innovations. One breakthrough that will quickly be implemented into other problems and lead to other breakthroughs. While this is intuitively easy to understand it is not easy to track or measure. Traditionally, this is tracked using citations, though this is not a reliable method because patents often cite a wide variety of other patents, many of which are not pertinent. This project will attempt to analyze the abstracts and bodies of the patents to find common phrases and attempt to use these links as a way of analyzing the spillovers from one patent to another.

3 Related Work

There is significant amount of similar work performed in the areas of copyright protection and IP portfolios management, which both also depend on models that most adequately represent patents. Consequently,

[*] jcalder@utexas.edu
[†] mduban@gmail.com
[‡] vic@cvhu.org
[§] vsubhashini@utexas.edu

there are previous works available, describing methods of textual data mining such as keyphrase extraction algorithms and co-word analysis in the context of patents [3]. Furthermore, there are works available on evaluating these methods to best fit the problem at hand [4]. Moreover, the idea of textual patents similarity is a subject of an ongoing research applicable to the cognitive science and legal (patents infringements) fields resulting in different concepts of similarity measurements and similarity coefficients available [5].

4 Hypothesis and Approach

Other than bag-of-word topic modeling, it will be interesting to analyze the writing styles and term usages between correlated patent publications. Our hypothesis assumes some underlying linguistic patterns introduced by knowledge spillover that arent necessarily apparent in the context of citations or word frequency modeling. Although it is difficult to define and quantify such relations, we can verify our results with domain knowledges and geographical information [6].

As a first step, we would like to identify and define some metrics to measure the performance of clustering with respect discovering spillovers. The most challenging aspect of the project is to come up with a language model to represent the patents. We will begin by attempting to apply known language models such as Bigram and Probabilistic Context Free Grammar (PCFG) for topic modelling. We will then cluster the patents and attempt to find links between the clusters and patents within individual clusters and verify the performance based on the defined metrics.

5 Outcome

The expected outcome of the project is to identify a language model (or a combination of models) and clustering techniques that can best capture knowledge transfer and spillover. The end result is intended to be a general method that can find connections between patents regardless of type of technology as well as an analysis of the provided solar power patents.

References

[1] The United States Patent and Trademark Office, http://www.uspto.gov/patents/process/search/. 2012

[2] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[3] Yuen-Hsien Tseng, Chi-Jen Lin, Yu-I Lin. Text mining techniques for patent analysis. Information Processing and Management, Volume 43, Issue 5, September 2007. http://www.sciencedirect.com/science/article/pii/S0306457306002020

[4] Tao Liu, Shengping Liu, Zheng Chen, and Wei Y. Ma. An evaluation on feature selection for text clustering. Proc. 20th International Conference on Machine Learning (ICML03), August 2003.

[5] Martin Moehrle. Measures for textual patent similarities: a guided way to select appropriate approaches. Scientometrics, May 2010. http://link.springer.com/article/10.1007%2Fs11192-010-0243-3?LI=true#

[6] Peter Thompson. Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-added Citations The Review of Economics and Statistics. 2006