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

May 1, 2013

Abstract

Content of the abstract.

1 Introduction

2 Data

3 Background

4 Methodology

The methodology we decided to use involved few carefully planned stages that put together will eventually produce reasonable results in terms of patent classification and knowledge spillover patterns in the field of solar energy. As mentioned previously we have focused on applying primarily methods of clustering to our solutions. The main tool that we used during the course of this project is MALLET (Machine Learning for Language Toolkit) discussed in previous sections. The main

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

features supplied by MALLET that we used are Topic Modelling and Clustering. MALLET toolkit is an open source project developed by the University of Massachusetts Amherst.

As we began working on the project, it became clear that the best way of tackling the problem is clustering, the nature of knowledge spillover is such that at the beginning there is no much data given on how the knowledge is being transferred from one field to another. This ruled out classification or regression methods at least in the initial stage. Furthermore, we decided to choose MALLET as the package that we will be working with. It provides very good tools for language and textual analysis and is easy in use and access, being written in Java and open source. Having many features, MALLET will allow us to try different approaches to tackling the problem and to limit the need to use other outside tools.

The data we used was supplied to us by a research group for the knowledge spillover in solar energy field working under Dr Rai. The database we received was handpicked by this group and one of our major concerns is how to approach handling these data. We ended up accessing United States Patent and Trademark Office (USPTO) website in order to get fuller and better representation of the data that we had obtained.

The above paragraphs show the biggest concerns and the main decisions we had to make in regards to the methodology that we used throughout this project. After deciding these key factors we could concentrate on actual use of these resources in order to come up with a method to find out patterns in the knowledge spillover.

First, issue that we encountered and that we had to spend considerable amount of time on is cleaning up the data and deciding which parts of patent texts are the most suitable to use for the task at our hands. We concluded that the best sections of patent documents to base spillover patent prediction are abstracts and claims. Abstracts show the shortest and most concise understanding of what the patent is about and therefore seem to be perfect for our needs. However, abstracts are usually not long, in terms of text size, enough to present on their own a considerable value in data mining applications and therefore they need to be supplemented. The best choice for the supplementary field turned out to be claims as they are the heart of the patent that in contrary to description or citations fields does not talk about previous or other patents and topics. Patents citations often contain redundant entries that are presented only to cover the patent to the fullest from the legal perspective. Similarly, description often involves the whole historical and technical background of the patent that is not something that we are interested in. Therefore, we decided to use the two patent document fields in our project, namely, abstract and claims.

To obtain these fields, a web crawling script was written in Java that crawled USPTOs website in order to get the required data. The script used patent number in case of patents and application

number in case of application to access the right entry. At this point abstract and claims fields were extracted from the entry to form the final entry that will be used from now on. The main difficulties with this step were different formats of application and patent numbers that did not work well with our script. Similarly, some patents had this number not provided, instead they had publication number, in these cases we had to get the right number first before crawling. Also, the navigation of USPTO website search engines is fairly complicated so getting accustomed to that took us awhile too.

After obtaining the data, the next step was to run topic modelling on the texts that we had obtained earlier. The details of what is topic modelling is have been provided in earlier paragraphs, so I will just briefly state how we applied topic modelling to our datasets. The data for each patent, i.e. abstract and claims, has been put into a separate file. Later all these files were supplied to MALLET together with standard stop words list in English language list to perform the topic modelling on the data. When running topic modelling we tried to come up with different values of n the number of topics created. After few runs and analysis of the outputs created, we decided that n equal to 10 is the most reasonable option. Furthermore, since topic modelling in MALLET is not deterministic, we ran the modelling with n=10 few times to make sure that this is the right choice and to arrive with the best range of topics.

Following the topic modelling we assigned each topic a very general class that it represented by looking at the topic models themselves. The classes we used for this purpose were SolarInventer, Solar Mounting/Rack, Solar Monitoring and Site Assesment. An example of a topic model can be seen on Fig. 1. It is the least of word in order of frequency as theyappear in each topic. This topic model was ran for n=15. To assign a class to a topic we looked at the list of words and arbitrarily decided which class given topic fits the most.

Next step was evaluating our results in attempt to see whether the topic models we arrived with are reasonable and reflect the actual patents distribution well. To approach this task we had to come up with a method that would compare our classes to some other method of classification. Since patent applications that arrive to USPTO are given classes by patent officers as they are filed we decide to use one of these as our method of evaluation. The patent classification scheme we decided to use is CPC Cooperative Patent Classification. Each patent is assigned one or more CPC classes as its application is filed to USPTO and we used these classes to see whether our topic models constitute a valid solution.

After ensuring that the topic models make sense we went on and grouped the data by geographies. Outside of the US it was done by continents. This is due to the fact that overwhelming number of patents unsurprisingly came from inside the US, the distribution of patents by coun-

tries can be seen on Fig. 2. Inside the US the clustering was done based on regions and states. We divided California into two regions, Bay Area and Southern California due to a great number of patents originating in that state. Following that we classified patents using our topic models obtained earlier into these topics. This gave us an opportunity to see which topics are popular in which areas and to form conclusions on whether certain topics are more popular in given geographies implying knowledge sharing and spillover.

Similarly, we divided patents by years and subsequently classified them into topics to see in which they were filed to see whether there are patterns of topics being more popular during certain time periods. This task was more difficult due to the fact that solar technology innovation field is very new and there is not many patents from the beginnings of the time period (2006-2012) we are working on. On Fig. 3 we can see all patents grouped by year and technology area, before running topic modelling on them.

Following the two procedures above we checked whether topics divisions inside the groups are relevant by seeing how much they overlap with CPC classes of the patents. This step was performed to ensure that our topic models, even when patents are divided into groups by geographies or years, still make sense and represent the actual topics that patents relate to.

Finally, the last step was to visualize the data to notice any patterns in knowledge spillover in the field of solar energy. It is much easier to see these patterns visually than in numerical and textual form. This allowed us to find any trends in the distribution of the topics and possibly to find out whether we can use data mining techniques to see whether and to what extent there is knowledge spillover in solar energy technology field. Next part of this paper will attempt analyse the results that we have arrived with.

5 Results

6 Conclusion

7 Future Works