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 3, 2013

Abstract

Knowledge spillovers

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].

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

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

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.

1.1 Motivation

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, there are previous works available, describing methods of textual data mining such as key phrase 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].

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 aren't 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].

1.2 Hypothesis

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 modeling. 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.

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.

2 Data

2.1 Source

2.2 Preprocessing

2.3 Chanllenges

3 Background

Given a collection of text-based patent documents, one intuitive idea to find out knowledge spillovers is to look at some underlying thematic structures hidden in the text. Based on the word usages and terminology distributions, we need to know what are the intrinsic topics implied in the relevant context, and one common way to do exactly that is called Probabilistic Topic Models formalized by Blei et al. [7]

3.1 Probabilistic Topic Models

Originally designed to facilitate text data visualization and browsing,

The main objective of topic modeling is to automatically discover the unobserved hidden structure–the topics, per-document topic distributions, and the per-document per-word topic assignments, with a collection of text documents as the only observable variables.

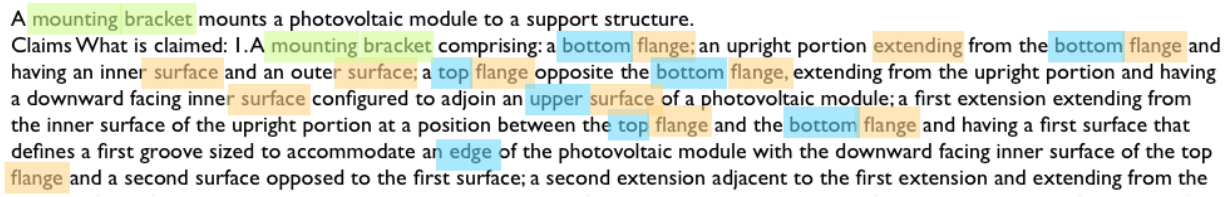

Figure 1: A sample patent document (partial)

For example, given a sample patent text, we assume that there is a set of topics associated with this patent. In Fig. 1, we have annotated a selection of words, with each topic distinguished by colors. For the orange topic, we get words such as flange, surface and extending, which could be interpreted as having something to do with attachment of hardware components such as pipes and cylinders. Similarly, the blue and green topics could be interpreted as installation and mounting respectively. By looking at the text, any human being with average comprehensive ability can easily tell what a document like Fig. 1 is about, and highlight the associated keywords that defines such topics.

Nonetheless, when we scale up the size and complexity of such patent documents, using human labor to do such tasks suddenly becomes erroneous and expensive. The goal of probabilistic topic modeling is to automate this process and to provide hidden insights and meaningful intelligence of big data. If we can successfully construct a reasonable thematic structure from our patent data, we can presumably infer influences or spillovers of patent authors within the same topics

3.2 Latent Dirichlet Allocation

Latent Dirichlet allocation, or LDA, is the simplest topic model [8] that assigns each word in the documents a distribution over a fixed number of topics. Instead of having a hard boundary between topic collections, LDA provides a distribution of topics per document, giving the likelihood of a mixed proportion of topic assignments. Namely, all patent documents share the same set of topic collection but with different proportions to each topic. For instance in Fig. 2, although there are $K = 100$ topics overall, only a few topics were actually activated.

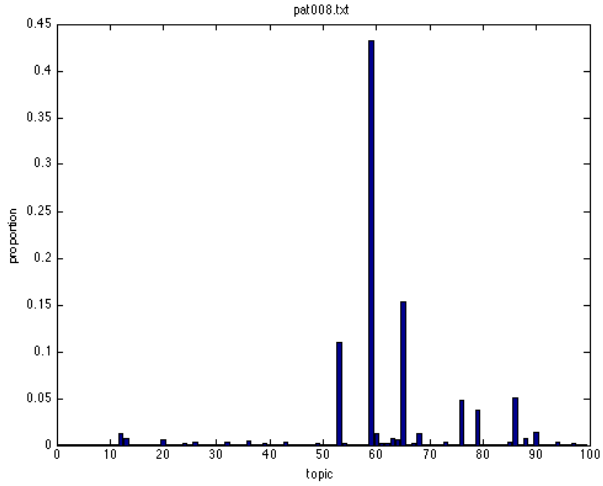

Figure 2: A sample topic proportion of a patent

To build the generative probabilistic model, we compute the joint distribution and use it to estimate the posterior probability. Before jumping into the actual calculation, let's formalize our notations:

β_k, $k = 1 \cdots K$: the K topics, represented by a distribution over words

θ_d, $d = 1 \cdots D$: topic proportions for document d, where $\theta_{d,k}$ is the topic proportion of topic k for document d

z_d, $d = 1 \cdots D$: topic assignments for document d, where $z_{d,n}$ is the topic assignment for the n-th word in document d

w_d, $d = 1 \cdots D$: observed words for document d, where $w_{d,n}$ is the n-th word in document d

With the above notation, the LDA generative process can be formalized as the following joint probability of both hidden and observed random variables:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

which can alternatively be expressed as a graphical model:

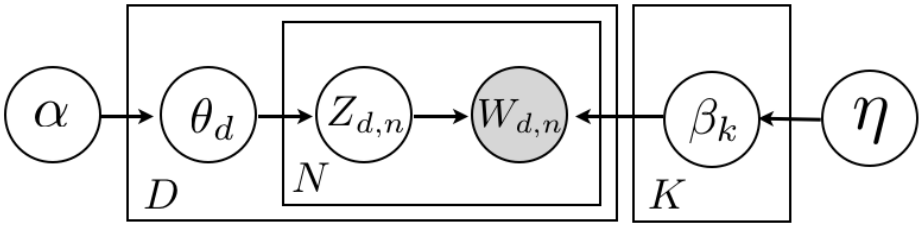

Figure 3: LDA graphical model. Nodes represent variables, while edges indicate the dependency relations. The shaded node is the only observed variable (document words), and all others are the hidden variables. The D plate denotes the replicated variables product over D documents, while the N plate denotes replication over N words in each document.

Note that there are several conditional dependencies implied in the graphical models, which reflects the main principles of how LDA "think" the documents are generated:

1. Randomly pick a distribution θ_d over topics.
2. For each word in the document
   (a) Randomly choose a topic from the previously-chosen distribution $\theta_{d,n}$.
   (b) Randomly choose a word from the corresponding distribution $Z_{d,n}$.

Assuming this generative process is how our documents are created, now LDA uses the graphical model in Fig. 3 to infer the posterior probability of the hidden structures given our observable:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

The computation of possible topic stuctures is often intractable and the posterior distribution can only be approximated in most cases. To form an approximation algorithm, topic modeling can generally be categorized as sampling-based algorithms and variational algorithms. The most popular sampling method for topic modeling is Gibbs sampling, which introduces a sequence of random variables to construct a Markov chain and collects samples from the limiting distribution to estimate the posterior. Instead of using samples to approximate the posterior, variational methods find the closest parameterized distribution candidate by solving optimization problems [8] [9].

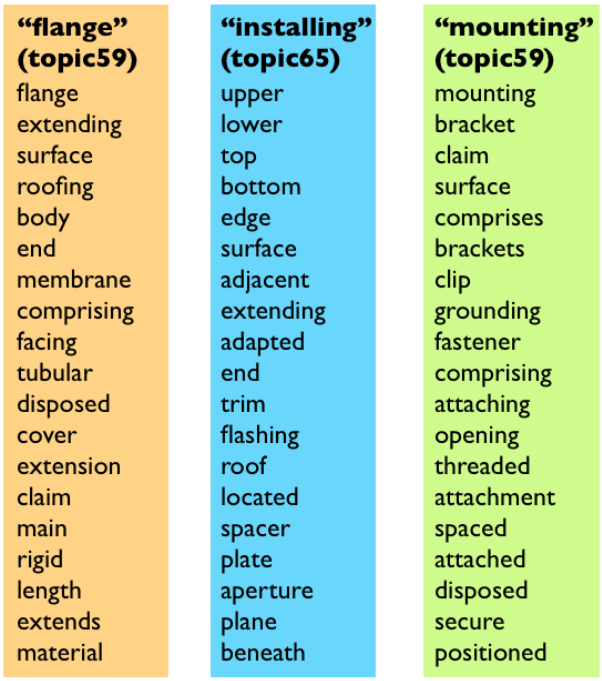

Figure 4: The top 3 topics of a sample patent

3.3 Limitations & Potential Improvements

Although LDA provides a powerful perspective to browsing and interpreting the implicit topic structures in our patent corpus, there are a few limitations it imposes against further discoveries. An extensive amount of research has been focused on relaxing some of the assumptions made by LDA to make it more flexible and suitable for various adaptations in different context.

Bag of Words LDA is essentially a bag-of-words probabilistic model. Namely, it constructs a word frequency vector for each document but disregards the word ordering or the surrounding context. Although this assumption looses the syntactic information and sometimes seems unrealistic when processing natural language, it is usually good enough when capturing the document semantics and simplifying hidden structural inferences. Nonetheless, for more sophisticated tasks such as language generation or writing style modeling, the bag-of-words assumption is apparently insufficient and needs to be relaxed. In such cases, there are variants of topic models that generate topic words conditional on the previous word [10], or switches between LDA and hidden Markov models (HMM) [11].

Document Ordering The LDA graphical model in Fig. 3 is invariant to the ordering of our patent documents, which could be inappropriate if the hidden thematic structure is actually de-

pendent on sequential information such as years published, which is typical in document collections spanning years, decades or centuries. To discover how the topics change over time, the dynamic topic model [12] treats topics as a sequence of distributions over words and tracks how they change over time.

Fixed Number of Topics In either LDA or more sophisticated dynamic topic models [12], the number of topics $\beta_{1:K}$ is determined manually and assumed to be fixed. One elegant approach provided by the Bayesian nonparametric topic model [13] is to find a hierarchical tree of topics, in which new documents can now imply previously undiscovered topics.

Meta-data To include additional attribute information associated with the documents such as authorships, titles, geolocation, citations and many others, an active branch of research has been performed to incorporate meta-data in topic models. The author-topic model [14] associates author similarity based on their topic proportions, the relational topic model [15] assumes document links are dependent on their topic proportion distances, and more general purpose methods such as Dirichlet-multinomial regression models [16] and supervised topic models [17].

Others Many other extensions of LDA are available, including the correlated topic model [18], pachinko allocation machine, [19], spherical topic model [20], sparse topic models [21] and bursty topic models [22].

3.4 MALLET

The Java-based package we used for topic model training is called MAchine Learning for LanguagE Toolkit (MALLET), developed by the team led by McCallum at the University of Massachusetts Amherst [2]. It covers various algorithms for statistical natural language processing, document classification, clustering, topic modeling, information extraction and many other text-based machine learning applications, including Hidden Markov Models (HMM), Conditional Random Fields (CRF), decision trees and others. More importantly, the MALLET topic modeling toolkit provides sample-based implementations of LDA, pachinko allocation and Hierarchical LDA.

4 Methodology

The methodology we decided to use involved few carefully planned stages that put together eventually produced 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 features supplied by MALLET that we took advantage of are Topic modeling 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 modeling on the texts that we had obtained earlier. The details of what is topic modeling is have been provided in earlier paragraphs, so I will just briefly state how we applied topic modeling 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 modeling on the data. When running topic modeling 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 15 is the most reasonable option. Furthermore, since topic modeling in MALLET is not deterministic, we ran the modeling with n=15 few times to make sure that this is the right choice and to arrive with the best range of topics.

Following the topic modeling 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 Solar Inverter, Solar Mounting/Rack, Solar Monitoring and Site Assessment. An example of a topic model can be seen on Fig. 51 It is the least of word in order of frequency as they appear in each topic. This topic model was ran for n=15. To assign a class to a topic we looked at the bag of words and arbitrarily decided which technology area each topic fits to 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

0	0.07401	load disposed plug wall lines diodes potential air efficiency converting switched automatically car
1	0.16949	inverter grid time bus mode component torque element based frequency disclosed structure maximum ph
2	0.12674	production array air structural shingle local lead perimeter tilt pamcc strut period fault modular
3	0.10942	pv dc plurality electric member housing connected metal side thermal end arm battens cycle tab cabl
4	0.07682	cell clips bracket bottom slider elongate circuit surface beams member structures conductive facade
5	2.06081	system power includes plurality module panel configured support coupled base array output grid pane
6	0.9473	photovoltaic mounting structure rail support side provided members roof device adapted parallel mou
7	0.07656	switching circuit generator housing bridge connection terminals side part fastening terminal interm
8	0.82828	connected module voltage cell unit device direct elements connecting invention element box point me
9	1.60119	solar modules mounting frame surface assembly pv energy provide module methods panel embodiment com
10	2.12657	power dc inverter solar photovoltaic voltage ac current converter output control input electrical e
11	0.03738	splices rack phase spacer sun link truss male body adapted current stress leg bases lag lip concent
12	0.18096	panel channel clamp rails pipe main rotate flange motor window sun variable legs sequence chain mem
13	0.07403	power conversion array device switch achieve circuitry configurable apparatus elements cell convert
14	0.16116	bracket frame number assemblies portions material cap pier sheet roofing aperture rotating ballast

Figure 5: An example of a topic model

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.

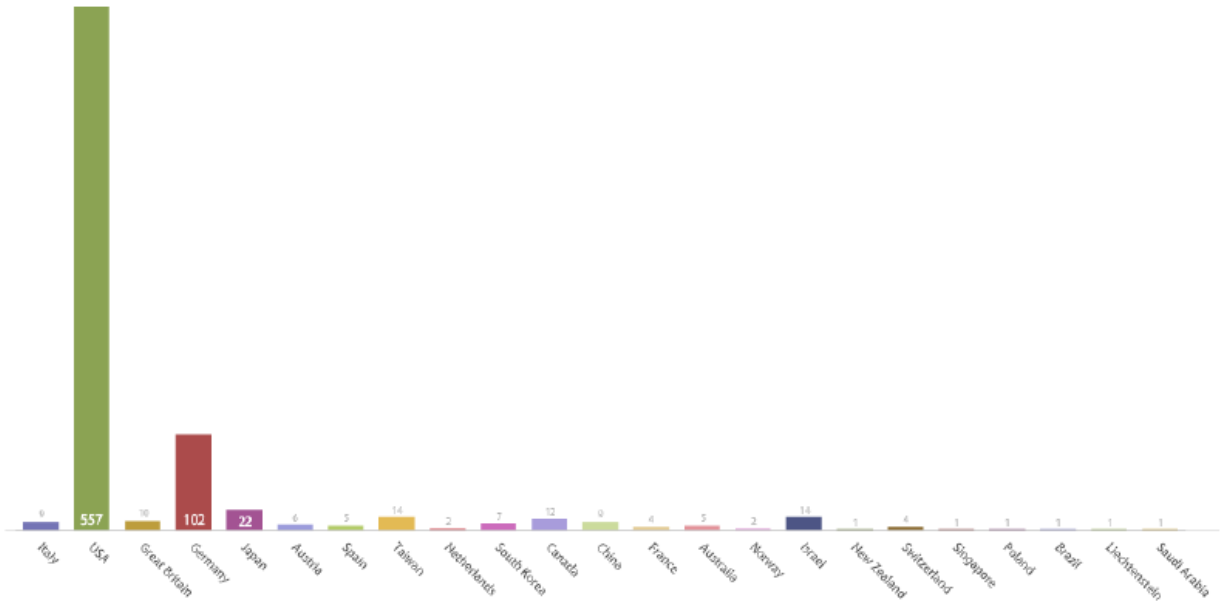

Figure 6: A distribution of patents by countries

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 countries can be seen on Fig. 6 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. 7 we can see all patents grouped by year and technology area, before running topic modeling on them.

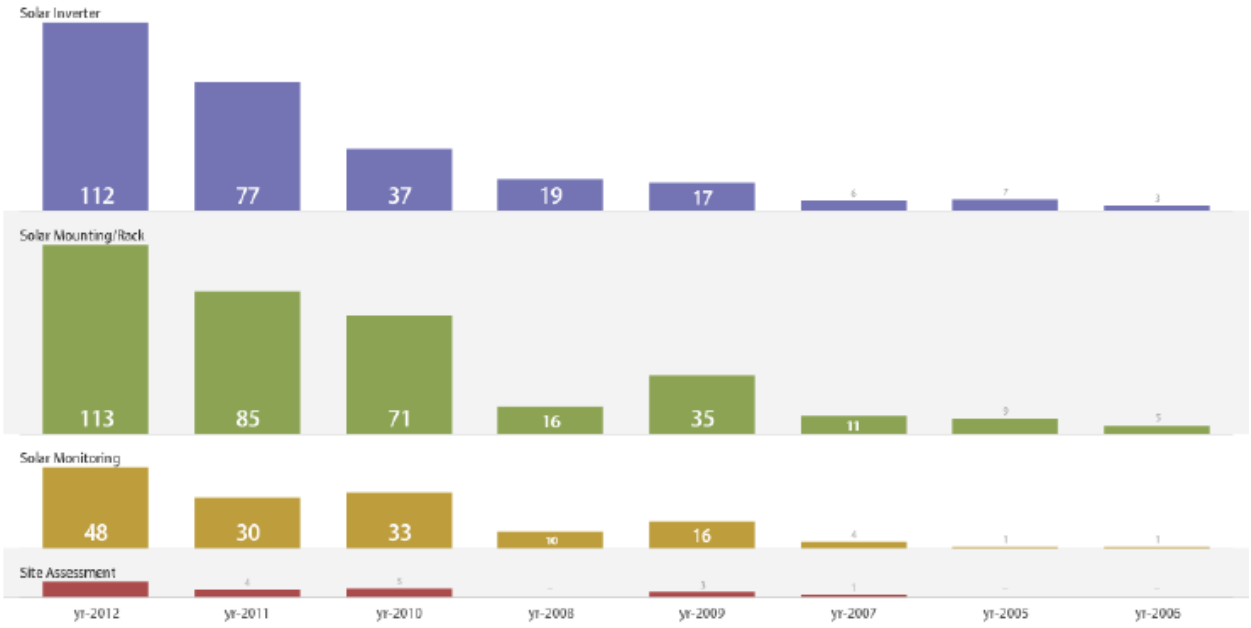

Figure 7: A distribution of patents by years and technology areas

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

Coarse-grained (technology area)	Fine-grained (CPC Class)
79.49% Exact match	81.01% Match 1-out-of-3 predicted to top 3

Table 1: Results

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 to analyze the results that we have arrived with and evaluate them.

5 Results

We have run the MALLET topic modeling on the patent data (claims and abstracts) carefully handpicked and crawled from the USPTO website. The objective was to determine whether data mining techniques could be used to predict and see knowledge spillover patterns in the field of solar power technology. The method we used consisted of running topic modeling on the patents then evaluating topic models by seeing whether they overlap with CPC classes that each patent is assigned by the USPTO at the point of filing the patent application. Grouping the patents by geographies and years allowed us to see whether there appear to be any patterns in the development of patents that would imply knowledge spillover. This section of the paper is presenting the results that we have obtained.

The first question that is crucial to any further consideration of the results is whether we can use topic modeling to classify patents in the first place. This could be evaluated by computing whether patent topics obtained through topic modeling overlap with either technology area or/and CPC classes. There are 4 broad technology areas that patents cover: Solar Monitoring, Solar Mounting/Rack, Solar Inverter and Site Assessments. There are 23 CPC classes that patents composing our data represent. The results of overlapping between the true class and modeled class obtained for topic modeling are presented in Table 1. Fig. 8 is a graphic representation of the confusion matrix of patents classification using topic modeling and technology areas.

The above data suggests that the classification of patents based on topic modeling is good and we can say with a significant degree of confidence that topic modeling can be used for the purpose of patent classification.

Next step is evaluating whether patents are associated with specific geographies and can topic

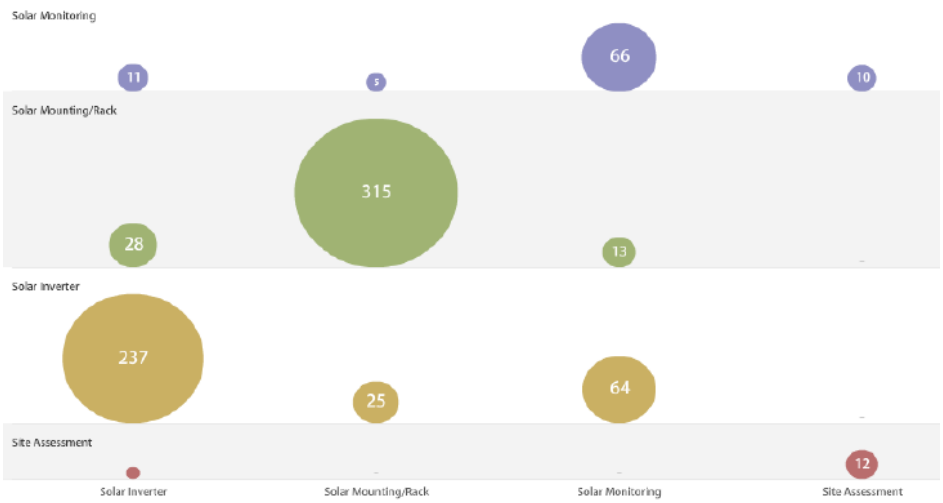

Figure 8: The confusion matrix of our predicted technology area results.

modeling predict these associations. To answer these questions patents are divided into geographies and then classification based on topic models is performed on them. In Fig. 9 and Fig. 10 we can see the most popular actual technology area and the most popular predicted technology area using topic modeling. Both seem to overlap almost completely and therefore, we can confidently give a positive answer to this problem.

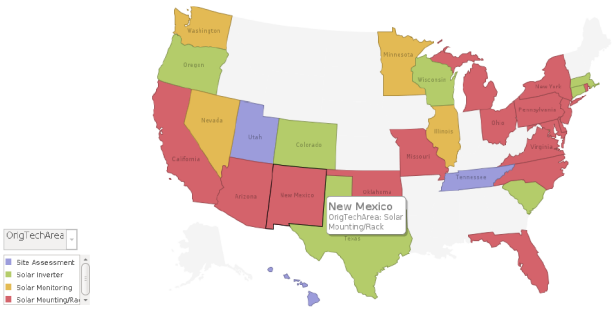

Figure 9: The original technology area distributed over geolocation

As we observed above there are patents associated with specific geographies and topic modeling can identify that quite accurately. This results assures us that topic modeling is a valid method of classifying patents.

However, the real question that we are interested in for the purpose of this paper is whether we can detect knowledge spillovers in the field of solar energy using data mining techniques. For this part we have focused mostly on the data pertaining to the United States, since there is simply not enough data in other places and if there is then there is not enough granularity in the data. To

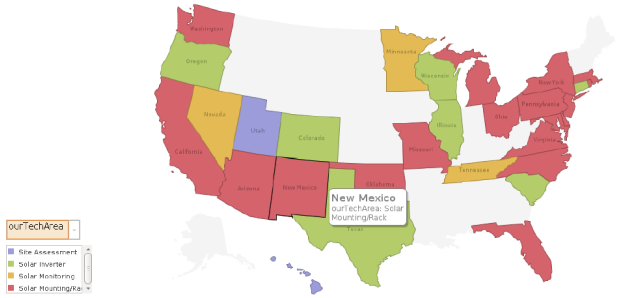

Figure 10: Our predicted technology area distributed over geolocation

answer the question of detecting the knowledge spillovers it will be very helpful to visualize the results in the form of the US map that is color-coded by different topics that are the most popular in each state. This graph can be seen in the Fig. 11

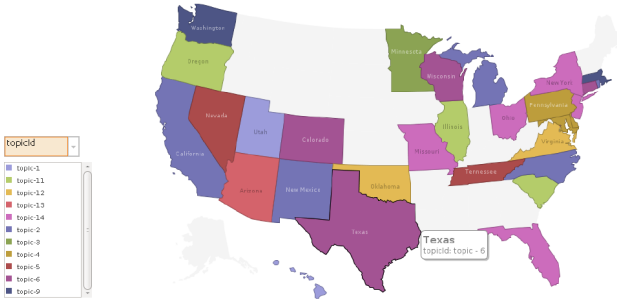

Figure 11: The most popular topics in each state of the U.S.

Each color in Fig. 11 represents the topic that is the most popular in each state. We can see that on the graph above topics themselves are not directly meaningful when it comes to geographical distribution of the patents. The colors in general are distributed all over the place. However, we can see that purple-pink color is more popular in the South and East while blue is more popular in the West and green in the North. These assumptions are, however very weak and based on subjective heuristic perceptions. In overall, although we can see some patterns forming to some degree, we cannot see a clear pattern of knowledge spillover that would clearly support any thesis on knowledge spillover in solar energy field. Therefore, we concluded that although it looks like it would be possible to use topic modeling to mine for knowledge spillovers, we are not yet there and some further iterations and more complex models would need to be applied in order to obtain an acceptably meaningful and refined solution.

Another question that we are trying to find an answer to is whether we can mine for knowledge spillover across different years. This is whether there are any topics that become especially popular

after given years or in given time periods. We can see the distribution of topics among years on graph on Fig. 12

Figure 12: The topics population over years

On the graph it can be clearly seen that the distribution of the topics does not change much from year to year. It basically stays the same throughout the whole period under consideration in this paper. It should be noted that the overall number of patents in solar energy field is growing from year to year and this is most likely caused by more incentives to move towards renewable energy sources coupled with rising prices of conventional energy sources. However, there seems to be no possibility of measuring knowledge spillover across the years using topic modeling methods, apart from the fact that there is similar distribution of topics from year to year and establishing what this distribution is.

6 Conclusion

The projects objective was to analyze around 800 patents from solar energy field and come up with data mining techniques that could help with knowledge spillover analysis and prediction in this field. We have crawled the data from USPTO website, using for the purpose of this project fields of abstract and claims in the text of patents. After this step, we clustered patents based on topics that they correspond to. We evaluated the topics by checking whether based on information about topics the original classification of patents can be obtained. Finally, we visualized and interpreted the results with respect to patent CPC classes (classes assigned by patent officers at USPTO), geographical distribution and time i.e. when patent has been invented. The analysis of the results

allowed us to come up with answers to the vital questions pertaining the issue at hand.

The analysis of the results gives an idea of whether the knowledge spillover can be measured using topic modeling methods. We have used MALLET (Machine Learning for Language Toolkit) to perform topic modeling and classification based on topic models tasks. The data that we used this is claims and abstracts of patents and applications for patents from solar energy fields from years 2006-2012 show that Topic Modeling is a viable method of patents classification. Using Topic Modeling gives rate of success on the data we had crawled from USPTO website of up to 81%. This is a good rate of success suggesting that most of the patents are classified correctly.

As for geographical distribution, we can see that there are interesting patterns in the geographical distribution of patents that can be noted. These patterns can be detected using topic modeling at least to some extent. We showed that it is possible to show which technology is the most popular in each area based on topic modeling and that this data overlaps with the actual technology areas popular in these regions. Furthermore, we can see that some very weak patterns are being formed even at the level of single topic models (not assembled into technology area groups). These features of topic models can be of significant use when analyzing temporal changes in the field of solar energy. Similarly, they can be of use when trying to figure out local interests in research in the field. And so we know that Texas for example is more interested in research into Solar Inverters and California is more focused on Solar Mounting and Racks.

As for the temporal analysis of patents there has not been much data that would support the thesis that topic modeling could be used to predict and notice patterns in knowledge spillover. The distribution of topics across different years has not been changing much. The only reasonable conclusion that we can draw from the data under consideration is the fact that number of patents in the solar energy fields is growing but this is not a surprising fact as alternative energy sources are becoming more and more popular in recent years. Also, there is no great range on the time scale in the field of solar energy as this is a very new field.

Therefore, the temporal analysis could be performed only on the period of 6 years and this could affect the outcome of our results. Analyzing more established field could lead to more meaningful conclusion in terms of how knowledge spillovers can be measured using topic modeling of patents. Furthermore, the data was mostly handpicked, which means that there is a possibility that some entries were discarded or added to the data set based on subjective opinions or simply because of human error. This could also impact our results.

The hardest tasks that we had to deal with throughout the course of this project were: data pre-processing, inferences and visualizing the data to confirm the hypotheses. Data pre-processing including crawling the data from the USPTO website and sorting out the right Application and

Patent numbers took us significant time and effort. Also visualizing the data to show the results in the nice fashion using clear and understandable graphics was not a very easy task. The color-coded maps where a real challenge to master and develop.

7 Future Works

Due to the time constraints we were unable to address all possible solutions and to approach the problem from every possible angle. There is still plenty of room left for improvements and further extensions to the procedures we performed. Similarly, there are few other approaches that we could have tried in order to receive even more satisfying solutions. One obvious approach that we could use and adapt to our solution in order to refine our topic modeling is the use of Natural Language Processing (NLP) and phrases analysis. This would mean mining for the topics based on phrases and pairs of words rather than single words alone. Such a development would hopefully lead to more accurate topic models and in turn better classification method.

Furthermore, a good idea to research would be attempting and testing the method with patents from different areas. This would allow us to see whether the algorithm is applicable to different technology areas and check its universality of the algorithm. Extending the project for possibly patents from any area would be a great research opportunity that would take a significant amount of time but could produce very interesting results.

Defining a more adequate stop words definition is an aspect of our method that we could further work on. This would hopefully improve our topic modeling classification method since words that are specific to all patents would be omitted while constructing the topic models. We were initially going to attempt this s improvement, however, it turned out that regular stop words list for English language delivered satisfying results so we left this as an option towards the end of our time schedule and we simply did not have time at the end. Nevertheless, it would be an interesting opportunity to work on a development of patent specific stop words list.

Another factor that could contribute to spillovers and possibly mining in it could give us some insight on knowledge transfer is size and type of the owner of a patent assignee i.e. the entity that owns the patent. Therefore, researching the distribution of patents based on the type and size of their assignees was suggested to us as another factor that we could consider in our project. Definitely work on it could lead to some meaningful insights on knowledge spillover.

Lastly, we could use a data set that represented patents more uniformly distributed in time and over wider domain (more than 6 years). This could be achieved at least in theory by the use of stratified sampling. Having wider domain and more uniformly distributed patents could lead

to a better conclusions regarding the knowledge spillover in time. However, we have to keep in mind that this may not be possible at all in the first place as the solar energy field is very new field and there is not many patents available that were granted or filed before 2005. Attempting a different, more established, field and a wider time domain could be a good substitute for this further development.

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

[7] Blei, D. Introduction to Probabilistic Topic Models. Princeton University. 2011.

[8] Blei, D., Ng, A. and Jordan, M. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, January 2003.

[9] Hoffman, M., Blei, D. and Bach, F. On-line learning for latent Dirichlet allocation. In Neural Information Processing Systems, 2010.

[10] Wallach, H. Topic modeling: Beyond bag of words. In Proceedings of the 23rd International Conference on Machine Learning, 2006.

[11] Griffiths, T., Steyvers, M., Blei, D. and Tenenbaum, J. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17, pages 537-544, Cambridge, MA, 2005. MIT Press.

[12] Blei, D. and Lafferty, J. Dynamic topic models. In International Conference on Machine Learning, pages 113-120, New York, NY, USA, 2006. ACM

[13] Teh, Y., Jordan, M., Beal, M. and Blei, D. Hierarchical Dirichlet process. Journal of the American Statistical Association, 101(476):1566-1581, 2006.

[14] Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smith, P. The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pages 487-494. AUAI Press, 2004.

[15] Chang, J. and Blei, D. Hierarchical relational models for document networks. Annals of APplied Statistics, 4(1), 2010.

[16] Mimno, D. and McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In Uncertainty in Artificial Intelligence, 2008.

[17] Blei, D. and McAuliffe, J. Supervised topic models. In Neural Information Processing Systems, 2007.

[18] Blei, D. and Lafferty, J. A correlated topic model of Science. Annals of Applied Statistics, 1(1):17-35, 2007.

[19] Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In International Conference on Machine Learning, pages 577-584, 2006.

[20] Reisinger, J., Waters, A. ,Silverthorn, B. and Mooney, R. Spherical topic models. In International Conference on Machine Learning, 2010.

[21] Wang, C. and Blei, D. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22, pages 1982-1989. 2009.

[22] Doyle, G. and Elkan, C. Accounting for burstiness in topic models. In International Conference on Machine Learning, pages 281-288. ACM, 2009.