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 4, 2013

Abstract

This research sought to analyze the patent data to track knowledge spillovers (one innovation leading to another) in the field of solar innovation. Through the use of the topic modeling toolkit in the MAchine Learning for Language Toolkit (MALLET), the patent text was placed into topics and mapped into real world categories. Once we found the hidden thematic structure of the patent data, we were able to correlate the topic models to other meta-data such as technology area, CPC Class, year and geographical location, and inferred relevant knowledge spillovers accordingly.

1 Introduction

The project proposed by Professor Rai involves analyzing a set of approximately 1,000 patents that pertain to solar energy technologies obtained from the US Patent and Trademark Office[1]. The information provided for each patent contains the following attributes:

1. inventor(s)
2. assignees

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

3. dates of filing, application, and issue

4. keywords based on standard patent classifications

5. geographical location of the assignee firm or inventor

The patents have been pre-classified based on the portion of the supply chain to which the patents apply. Additionally, a large amount of preprocessing and web crawling for corrected information was done in order to insure a reliable data source.

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 lead to other innovations. While this is intuitively easy to understand, it is not a trivial task to quantitatively track or measure. Traditionally, knowledge spillovers in patents is studied using citation linkages. However, this is not a reliable method; intellectual property attorneys often cite a wide variety of other patents to increase the chance of acceptance, which doesn't necessarily reflect the true relationships between innovations. We propose an alternative method to run a topic modeling algorithm over the abstract and claims texts of the patents. The patents are then sorted by the geographical location and the publishing date in the attempt to illustrate trends within certain topics.

1.1 Motivation

There is a significant amount of similar works performed in the area of copyright protection and IP portfolios management, including 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 coefficients [5].

By using a bag-of-words model such as topic modeling, it is interesting to analyze text-based hidden thematic structures 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. Although it is difficult to define and quantify such relations, we can reasonably verify our results with domain knowledges and geographical information [6].

1.2 Hypothesis

By using latent Dirichlet allocation (LDA), the simplest topic modeling algorithm, we make the following assumptions about our patent documents:

- words have no dependencies with respect to their surrounding neighbors
- document ordering is irrelevant
- the number of topic is fixed and has to be determined manually

Although these assumptions made by LDA aren't necessarily ideal and realistic to capturing the complete syntactic information, they are simple and efficient enough to help us model the hidden topics and thus infer our spillover assumptions.

2 Data

The data for this project was initially a database of approximately 1,000 entries containing 43 fields, including several forms of identification, classification, authors, owners, filing dates, and filing locations. The final analysis would use only 7 fields: patent number, filing year, city, state, country, tech area and CPC class. Additionally, the abstract and claims of the patent were added to the database after being scraped from the USPTO website.

2.1 Source

The database was assembled by a team working on the project for Professor Rai. The team manually read every patent that was returned when querying the patent office for a certain tech area. They then decided whether or not the patents were appropriate to the topic of this research. The team placed appropriate patents into tech areas based on their personal interpretation of the subject of the patent. Following this placement the patents were manually entered into the database along with all fields the team deemed important.

2.2 Preprocessing

Due to the hand-entered nature of the database, a fair portion of preprocessing was required. Firstly, many of the fields needed for the analysis contained inconsistently-formatted data. For example, many geographic locations would be surrounded by parentheses, have word orders switched, or

have stray blank space that imposed difficulties in our analysis. To overcome this problem, the database ran through several preprocessing scripts to remove these abnormalities. An additional script was used to add more granularity to the geographic location, sorted by cities for California, by states for the rest of the United States, and by country for the rest of the world.

The next issue with data normalization was the corrupted data. On one hand, human-generated data is prone to erroneous results. On the other hand, most of the links provided by Dr. Rai's team have expired and no longer reflects the true intended data source (an updated version was provided later on). In an effort to overcome these problems, we wrote a web crawler to scrape relevant data directly from the USPTO website and recreated a cleaner and more consistent database.

2.3 Challenges

The primary challenges encountered were the previously mentioned hand-entered nature of the data and the poorly-formatted HTMLs on the USPTO website. A significant amount of time was dedicated to preprocessing and authenticating this data to ensure accurate results. However, there were several mislabeled patents which could not be corrected. Although all the identified conflicts were removed, it is likely that some of them remained undetected. Additionally, crawling the patents was difficult because the USPTOs website did not follow a standardized template and the patents had inconsistent fields. To overcome this, a large number of fringe cases needed to be handled correctly in our crawler.

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

The main objective of topic modeling is to automatically discover the unobserved hidden structures–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.

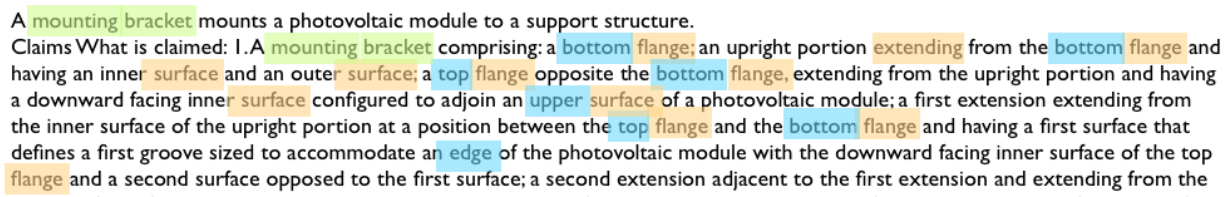

Figure 1: A sample patent document (partial)

For example, given a sample patent text, we assume that there exists a set of topics associated with this patent. In Fig. 1, we have annotated a selection of words, with topics distinguished by colors. For the orange topic, we get words like 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 common comprehensive ability can easily tell what a document similar to Fig. 1 is about, and highlight the associated keywords that defines such topics.

Nonetheless, when we scale up the size and complexity of these 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.

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

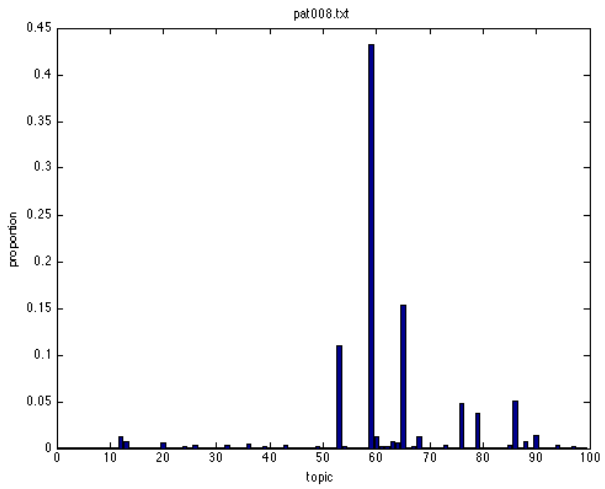

Figure 2: A sample topic proportion of a patent

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:

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 $Z_{d,n}$.

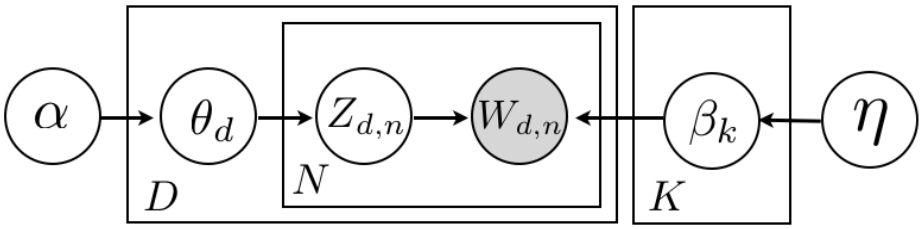

Figure 3: LDA graphical model. Nodes represent variables, while edges indicate the dependency relations. The shaded node is the only observed variable (document words), and all others are the hidden variables. The D plate denotes the replicated variables product over D documents, while the N plate denotes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

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 structures 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].

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 more sophisticated 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 and the neighboring 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

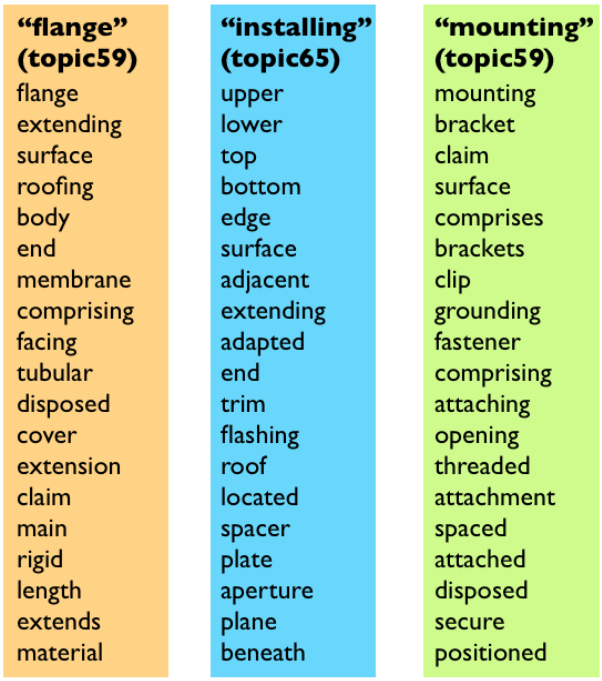

Figure 4: The top 3 topics of a sample patent

---

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 these cases, there are variants of topic models that generate topic words conditioned 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 dependent 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 Prof. 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 consists of a few carefully planned stages that put together reasonable results for patent classification and analysis of knowledge spillover patterns in the field of solar energy. As mentioned previously, we have focused primarily on applying methods of clustering in our solutions. The main tool that we used throughout 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 Classification.

Very soon into the project, it became clear that the best way of tackling the problem is through clustering. The nature of knowledge spillover is such that at the beginning there is not much data given on how the knowledge is being transferred. This ruled out classification or regression methods at least in the initial stage. We decided to choose MALLET as the toolkit package that we would be working with. Written in Java and open source, it provides very good tools for textual

analysis and is relatively easy in use and access. Having a wide range of features, MALLET would allow us to try different approaches and to limit the need to use other outside tools.

As mentioned before, our data source was supplied by a research group supervised by Dr Rai. The database we received was handpicked by this group, and one of our major concerns was how to handle inconsistent and incomplete entries. We ended up accessing United States Patent and Trademark Office (USPTO) website in order to obtain fuller and better representation of the data that we had initially received. After resolving our initial concerns and cleaning up our database, we could concentrate on the actual use of available resources in order to come up with a method to find out the hidden patterns and to infer knowledge spillovers.

4.1 Feature Selection & Construction

First issue that we encountered and that we had to spend considerable amount of time on was data pre-processing. Cleaning up the data and deciding which parts of patent text are the most suitable to use for the task at our hands turned out to be quite tricky. Though discussion with Dr. Rai's group, we concluded that the most representative sections of patent documents are the abstract and claims. Abstracts show the shortest and the most concise understanding of what the patent is about, and therefore seem to be suitable for our needs. Nonetheless, abstracts are usually not long enough to present a considerable value in data mining applications alone. The best choice for the supplementary field turned out to be the "claims" section in a patent document, as they are the essence of what an innovation has accomplished. The citations often contain redundant entries that are presented only to protect the patent from lawsuits. Similarly, description often involves the whole historical and technical background of the patent and the inventor, which are not as relevant as the abstract and claims to our objectives.

To obtain these fields, a web crawling script was created in Java that crawled USPTOs website in order to get the required data fields. The script used patent or application numbers to access the right patent or application webpage. At this point, the abstract and claims fields were extracted from the webpage to form the final entry. The main difficulties with this step were the inconsistent markup formats and invalid patent numbers.

4.2 Topics Inferences

After getting the data, the next step was to run topic modeling on the obtained patent text. The details about how topic modeling works are provided in the previous section, so in this section we will focus on how topic modeling was applied to our dataset.

The data for each patent (i.e. the abstract and claims) is separated into individual files under a common directory. All of these files were supplied to MALLET together with a standard "stop words" list for English to perform the topic modeling on the data. When running topic modeling, we tried to come up with different number of topics K. After a few runs and result analysis, we decided to pick $K = 15$ topics. Since topic modeling converges to slightly different results for each trials, we ran the algorithm multiple times to confirm that choosing $K = 15$ gives us a consistent result of topics. An example of a topic model can be seen on Fig. 5

```
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

4.3 Evaluation

To see whether the topic models we arrived with are reasonable and reflect the actual patents distribution well, we came up with a method that would compare our results from topic models with the assigned CPC class numbers. CPC (Cooperative Patent Classification) classes are assigned by patent officers at USPTO before they are published; each patent is assigned one or more CPC classes as its application is filed to USPTO, and we use these class information to measure if our topic models constitute a valid thematic structure.

4.4 Geographic & Temporal Analysis

The distribution of patents by countries can be seen on Fig. 6 We divided geographical locations with various granularities according to their patent population. For instance, California was divided into two regions–the Bay Area and Southern California, while countries with very few samples are grouped into regions or even continents. Using the topic models we obtained in the previous steps, we can observe whether certain topics are more popular in a selection of locations, and thus inferring relevant patterns of knowledge spillovers.

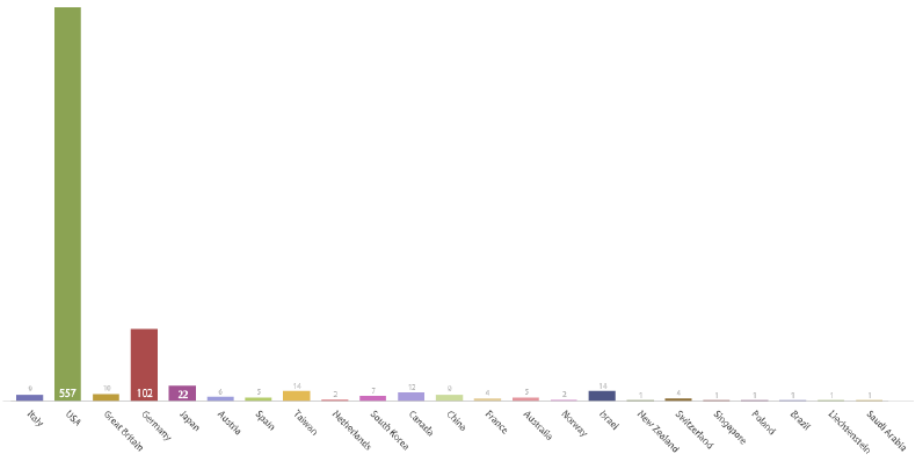

Figure 6: A distribution of patents by countries

Similarly, we divided patents by years and subsequently classified them into topics to observe trending patterns over time. This task was more challenging in that the solar technology innovation field is relatively new, and thus the distribution over the years was rather not uniform (in Fig. 7, we can see all patents grouped by year and the technology area.)

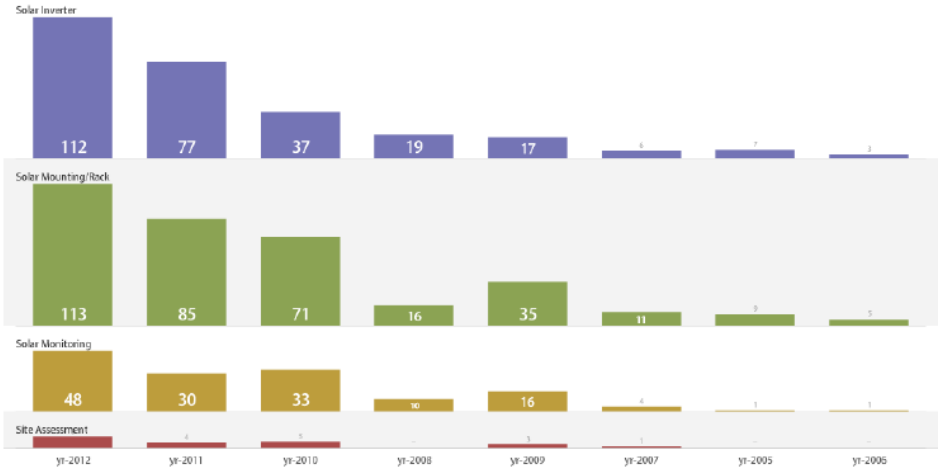

Figure 7: A distribution of patents by years and technology areas

Following the procedures above, we measured our topic modeling results by comparing how they mapped to the CPC classes within spatial and temporal groups. 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 the patents relate to.

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

4.5 Visualization

Finally, we visualize the data to discover patterns of knowledge spillovers in the field of solar energy. It is much easier to see these patterns visually than in numerical and textual forms. This allowed us to find any trends in the distribution of topics and to see whether or to what extent there is knowledge spillover in solar energy technology field. The next part of this paper will analyze and evaluate the results we have arrived with.

5 Results

We have run MALLET topic modeling on the solar energy patents data (restricted to 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 and analyzing whether they overlap with CPC classes to which each patent is assigned by the USPTO. Grouping the patents by geographical locations and years allowed us to see whether there appear to be patterns in the development of patents that would imply knowledge spillover.

To see whether we can use topic modeling to classify patents in the first place, we evaluate our results from topic model by computing how the topics agree with either technology area or/and the CPC classes. There are 4 broad technology areas our patents covered: Solar Monitoring, Solar Mounting/Rack, Solar Inverter and Site Assessments. There are 23 CPC classes that our patent data represented. The overlap between the true class and modeled class obtained from topic modeling is presented in Table 1. Fig. 8 is a graphic representation of the confusion matrix of patents' classification using topic modeling and technology areas, which confirms our assumption that topic modeling is capable of a simple classification task that aligned with our meta-data.

Next, we evaluate whether patents are associated with specific geographical locations, and that topic modeling can successfully predict these associations. To answer these questions, patents were sliced by the geographical locations before running the topic modeling algorithms over these location slices. In Fig. 9 and Fig. 10, we can see the most popular original technology area and

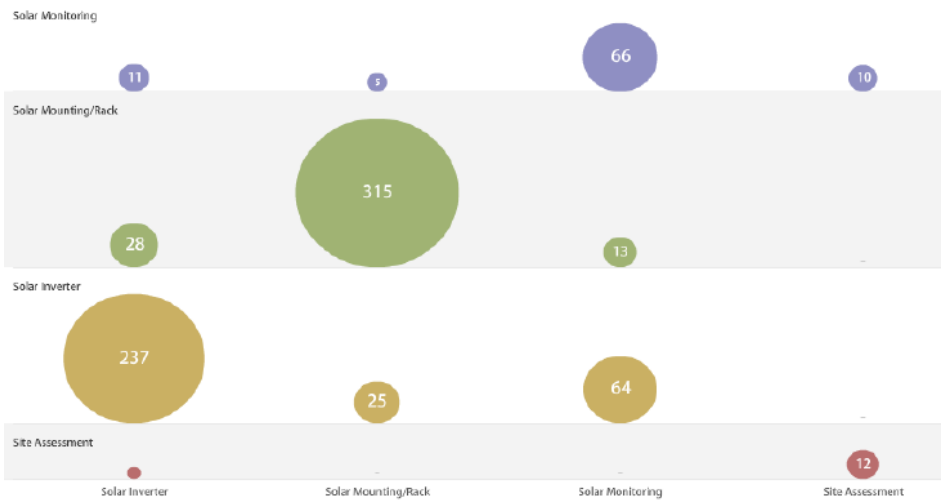

Figure 8: The confusion matrix of our predicted technology area results. The data suggests that the classification of patents based on topic modeling is good and we can say with confidence that topic modeling can be used for the purpose of patent classification.

---

the most popular predicted technology area using topic modeling. Both seem to overlap almost completely. Therefore, we can confidently give a positive answer to this problem.

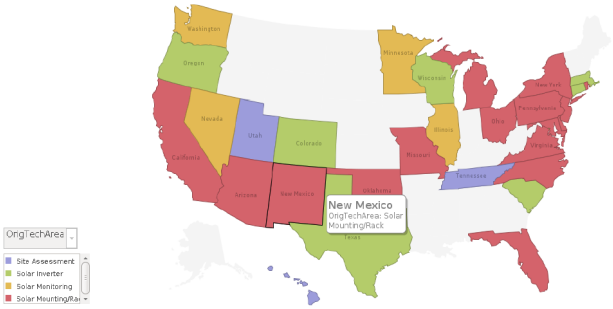

Figure 9: The original technology area distributed over geolocation

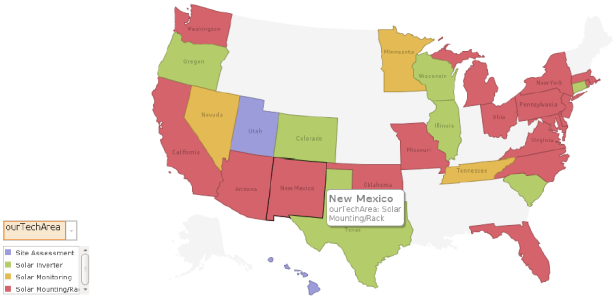

Figure 10: Our predicted technology area distributed over geolocation

As we observed above, there are patents associated with specific geographical locations, and topic modeling can identify that quite accurately. This result assures us that topic modeling is a valid method to classify our selected patents.

However, the real question that we are interested in for the purpose of this paper is to detect knowledge spillovers in the field of solar energy using data mining techniques. We have focused mostly on the data pertaining to the United States since patents from other countries simply don't have the ideal granularities we need to generalize our analysis. To detect knowledge spillovers, we

visualized our results by plotting the most popular topics of each states on the US map in Fig. 11.

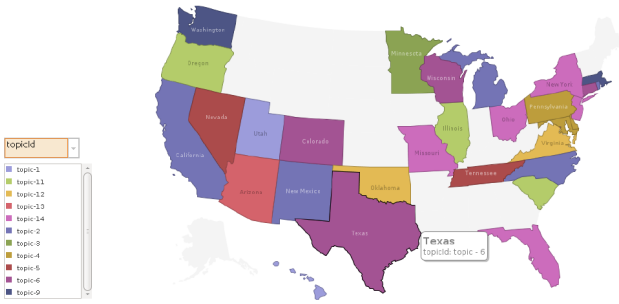

Figure 11: The most popular topics in each state of the U.S.

From Fig. 11, the topics themselves are not necessarily 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. However, these qualitative discoveries are subjective to heuristics and individual perspectives. Although there are some patterns, we cannot find a quantitative evidence to support a concrete relation to knowledge spillovers in the field of solar energy. Therefore, we concluded that although it might be possible to use topic modeling to mine for knowledge spillovers, further iterations and more complex models are required to obtain a meaningful and solid conclusion.

Figure 12: The topics population over years

On the other hand, it can be clearly seen from Fig. 12 that the distribution of 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.

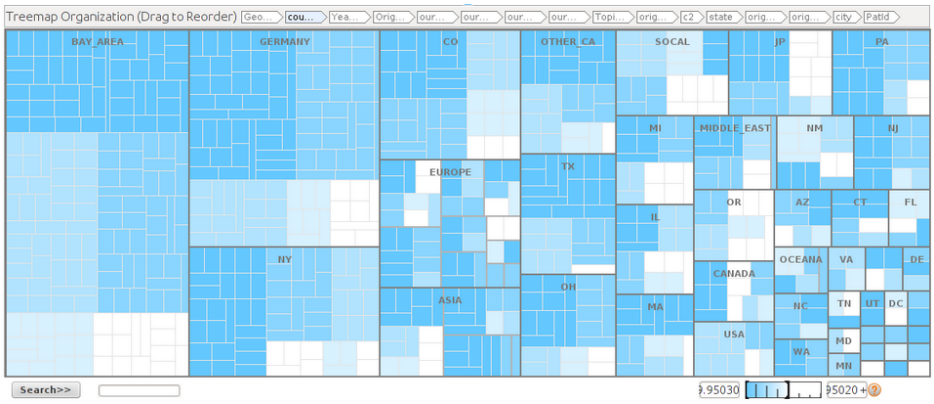

Figure 13: A treemap to illustrate patent population over geographical location and year. We can see that almost one third of the patent population is dominated by the Bay Area, Germany and New York, while years are mostly concentrated in 2010–2012 (the darker means more recent.)

6 Conclusion

The projects objective was to analyze around 800 patents from solar energy field and to 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 are the claims and abstracts sections of patents and applications for solar energy patents from years

2006-2012. It showed that Topic Modeling is a viable method of patents classification. Topic Modeling gives up to 81% accuracy on the data we had crawled from the USPTO website. This was an ideal preliminary results 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 popular technology areas in these regions. Furthermore, we can see very weak patterns formed even at the level of single topic models (not assembled into technology area groups). These features of topic models can be relevant when analyzing temporal changes in the field of solar energy. For example, we know that Texas is more interested in research into Solar Inverters and California is more focused on Solar Mounting and Racks.

For the temporal analysis of patents, there has not been much data that would support the hypothesis that topic modeling could be used to predict or recognize patterns in knowledge spillovers. The distribution of topics across different years didn't change much. The only reasonable conclusion that we can draw from the data under consideration is the fact that the number of patents in the solar energy fields is growing. However, this is not a surprising result, as alternative energy sources are becoming more and more popular in recent years. Also, there isn't a uniformly-distributed patents over a wide time scale since the field of solar energy is relatively new.

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 a more meaningful conclusion in terms of how knowledge spillovers can be measured using topic modeling of patents. Furthermore, the data was mostly handpicked, so it is possible that some entries were discarded or added to the dataset based on subjective opinions or simply due to human error, which could also impact our results.

The most challenging parts of this project were data pre-processing, inferences and visualization. Data pre-processing such as crawling the data from the USPTO website and sorting out the right Application/Patent numbers took us significant time and effort. Furthermore, visualizing the data on the corresponding geographical location and choosing the right way to illustrate our results in a meaningful and crystal-clear fashion weren't trivial.

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.

Another possible modification to our topic modeling would be to utilize a supervised topic modeling algorithm. These algorithms, such as the supervised latent Dirichlet allocation algorithm (sLDA) [17] allow users to manually influence their topic modeling. This serves the purpose of modeling text more accurately for a specific purpose. For example, a standard topic model of the text of an entire patent may return several topics specific to patents but not the text area (such as the terms figure or table). Through the use of a supervised method, the user can guide the topics towards terms such more specific to their goals. In our case these would be terms relevant to solar technologies.

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 the 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.

8 Acknowledgements

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.