Disambiguating Patent Inventors, Assignees, and their Locations in PatentsView

Nicholas Monath\(^2\), Sarvo Madhavan\(^1\), Christopher DiPietro\(^1\), Andrew McCallum\(^2\), Christina Jones\(^1\)

1. American Institutes for Research
2. University of Massachusetts Amherst

\{nmonath, mccallum\}@cs.umass.edu
\{smadhavan, cdipietro, cjones\}@air.org

Abstract

PatentsView, an initiative supported by the Office of Chief Economist in the US Patent & Trademark Office (USPTO), is a tool for patent search, analysis, and visualization. It provides a curated and easy-to-use database of pre-granted and granted USPTO patents from 1976 to present. Not only does PatentsView carefully process and clean raw patent data collections from USPTO’s bulk XML files (https://bulkdata.uspto.gov/), but it also performs entity resolution of the ambiguous inventor names, assignee names, and the location names of the inventors and assignees. This process disambiguates which inventor, assignee, and location names refer to the same entity in PatentsView. In this work, we describe the entity resolution models and algorithms used in PatentsView, highlight their technical and empirical strengths, and provide examples of studies that have benefited from PatentsView’s disambiguation.

Keywords: Entity resolution, de-duplication, disambiguation, patent data

1 Introduction

Granted patents as well patent applications in the USPTO do not record unique identifiers for the patent’s inventors nor assignees. Instead, the names of each inventor and assignee appear as they would in natural language and therefore can be ambiguous. For example, there are many inventors who share common names such as John Smith and the assignee Eastman Kodak Company is listed on some patents as Eastman Kodak Co., on others Eastman Kodak and Eastman Kodak Comapny\[sic\]. The task of resolving these ambiguities, determining which mentions of a particular inventor / assignee refer to the same real world entity, is the task of entity resolution (also known as disambiguation).

Entity resolution of inventors and assignees in patent records is crucial for a variety of applications. Disambiguated entities provides for more efficient search of patent records, allowing users to find records by only those inventors/assignees that they are interested in. It also allows users to more easily follow recent applications by particular inventors/assignees. Entity resolution also facilitates research in scence-of-science, economics, and innovation as it can provide more accurate statistics of which inventors are prolific and the mobility of inventors.

PatentsView\(^1\) provides both a patent search and visualization tool as well as a collection of curated and cleaned databases of patent records. Entity resolution is one of the services provided by PatentsView. Databases containing the disambiguation of inventor names and assignee names as well as the string valued geographic locations associated with each. These databases are continuously updated with each PatentsView data update. We describe the methodology used by PatentsView, how it compares to previous iterations of the PatentsView entity resolution methodology, and provide examples of use cases of PatentsView.

2 Methods

Entity resolution is a clustering task. We are given as input a set of ambiguous mentions of entities (inventor names, assignee names, location strings) and produce as output groups of these mention for each each group/cluster contains mentions that all refer to the same real world entity.

Each of our clustering approaches are based on hier-

---

\(^1\)https://www.patentsview.org/
architectural agglomerative clustering with learned (or rule-based) similarity functions. The similarity functions take as input a pair of mentions and produce a scalar indicating how likely the pair refer to the same real world entity. We develop three separate such similarity models for inventors, assignees, and locations. The inventor model is a linear model using features such as matching names initials of the names of the inventors, the number of overlapping co-inventors, the similarity of patent titles using a SentVec model (Gupta et al., 2019), and the locations of the inventors. This model is trained on publicly available labeled data from the 2015 PatentsView Workshop. The assignee model is based on a tf-idf character n-gram string similarity model that uses data from PermiD². The location model uses both string similarity features as well as disambiguated inventor/assignee entity with which the location is associated via a rule-based model. A flat clustering is selected from the hierarchical structure using a threshold on the similarity function.

Incremental Updates Given an existing hierarchical clustering of entity mentions, we can incrementally update this tree structure with mentions newly filed patent application newly granted patents using incremental clustering methods such as Grinch (Monath et al., 2019).

Canopies To reduce the scale of the clustering problems required, we partition the data into (possibly overlapping) canopies (McCallum et al., 2000). The assumption is that two mentions must be in the same canopy to refer to the same entity. For inventors, we use the first initial and last name as the canopy. For assignees, we use character 4-grams. For locations, we use the disambiguated inventor/assignee and string spelling features.

Data Processing. The PatentsView data pipeline collects patent data provided by USPTO and other sources and transforms them into a relational database format, ultimately making the data available through the PatentsView website. The pipeline consists of four major phases: Data Collection phase to collect and clean messy XML data from various sources; Disambiguation Phase (the subject of this abstract); Post Processing Phase to prepare data for user consumption and Data Delivery Phase to update the website with disambiguated and processed data.

3 Experiments

We apply our method to the entire collection of patent applications and granted data. In total, this corresponds to 26.1 million inventor mentions, 9.5 million assignee mentions, and 35.6 million location mentions.

In earlier versions of PatentsView, a sampling-based clustering approach was used for the resolution of inventor mentions (Monath and McCallum, 2015) and a simpler Jaro-Winkler-based model was used for assignees. Here we present an empirical comparison of the proposed approaches in this abstract and the previous ones.

We evaluate the inventor disambiguation using pairwise precision/recall/F1 of labeled mention data. For assignees, we evaluate using labeled assignee string names instead of mention level data. This means each unique assignee string name is considered once in the evaluation regardless of how many patents the name appears on.

4 Impact

PatentsView disambiguation is available to the public to enable longitudinal analysis of innovation and invention. Numerous researchers have used PatentsView data for their analysis such as these works from 2020 (Breschi et al., 2020; Chandra et al., 2020; Forman and Goldfarb, 2020).

5 Conclusion

We present the entity resolution approaches for inventor, assignee, and location disambiguation used in PatentsView. In future work, we hope to support user feedback on the entity resolution decisions.

References


<table>
<thead>
<tr>
<th></th>
<th>ALS-C</th>
<th>ALS</th>
<th>ENS</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prev. PV</td>
<td>0.965</td>
<td>0.984</td>
<td>0.972</td>
<td>0.978</td>
</tr>
<tr>
<td>This work</td>
<td><strong>0.985</strong></td>
<td><strong>0.986</strong></td>
<td><strong>0.981</strong></td>
<td><strong>0.978</strong></td>
</tr>
</tbody>
</table>

Table 1: Inventor Results. Pairwise F1 results on the datasets used in the 2015 Inventor Disambiguation workshop.

<table>
<thead>
<tr>
<th></th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prev. PV</td>
<td>0.998</td>
<td>0.331</td>
<td>0.497</td>
</tr>
<tr>
<td>This work</td>
<td><strong>0.992</strong></td>
<td><strong>0.372</strong></td>
<td><strong>0.541</strong></td>
</tr>
</tbody>
</table>

Table 2: Assignee Results. Pairwise metric results on labeled assignee disambiguation data.

²https://permid.org/