Identify Company Permno Documentation

Junying Fang

Step 1 Create the First Patent Assignment Dataset

1. Since the raw patent record only contains the first patent assignment of each patent id, we create the first patent assignment dataset out of all parsed patent assignments to reduce the search range. The first patent assignment of each patent id is the patent assignment with the earliest recorded date. If the recorded date is same, we choose the one with the earliest last update date.
2. Note that the parsed patent assignments datasets from 1980-2015 are very large. Thus, we need to process two datasets consecutively to select the patent assignment with earlier record dates at each iteration.

Step 2 Standardize Company Names to Improve the Match

We use permnos of company names in the first patent assignment to find permnos for all company names of buyers and sellers in all patent assignments. However, these company names in patent assignments may not be same even if they represent the same company. To improve the match, we need to standardize the unique company names (from both buyers and sellers) in all patent assignments first.

1. Since each buyer and seller may have multiple patent assignments, we first extract unique company names from all buyers and sellers of patent assignments for efficient implementations. Then, we standardize these unique company names.
2. Remove meaningless symbols in company names such as ",", "."
3. Convert abbreviations to full names for standardization. For instance, we convert assn, ass,assoc, and associations to association. We convert cmte,comm, com to committee.
4. Remove any symbols at the beginning/end of words
5. Remove not meaningful words (stop words, company attributes, state name, country name)
6. Remove leading and ending trailing spaces
7. Drop missing values
8. Use cleanco functions twice to standardize company names

Note: Since each patent assignment dataset has over million entries, the multiprocessing is used to improve the running time by over 80%.

Step 3 Build the Standardized First Assignor Company Names and Permnos Crosswalk

1. We link Raw Patent Record and the First Patent Assignments with Standardized Company Names by Patent Numbers to build the crosswalk. In raw patent record, some patent numbers have multiple permnos. Then, we could not say that the assignor of the first patent assignment of such patent number has which permno.

2. To resolve the issue, we first only use raw patent records which relates each patent number to only one permno.

   After linking these raw patent records with first patent assignments by patent number, we find some multiple-on-multiple relationships between company names and permnos. We resolve such cases by the following:

   a. Multiple company names correspond to one permno: This occurs because one permno may correspond to multiple patent numbers. These patent numbers are supposed to have the same first assignor. However, the patent assignments datasets are not complete for 1980-2012. Thus, we may not find the correct first assignor for some patent ids. Such cases lead to multiple first assignors company names-one permno relationship. To resolve this, we could safely think that the company name with the most occurence related to the permno is the correct company name for the permno. Thus, we fuzzy match all the company names related to that permno. Then, we choose the company name which matches most company names related to the permno as the company name for that permno.

   b. Multiple permnos correspond to one company name: We use the new dataset crsp_directory to manually select the correct permno with respect to the company name.

3. Then, we link the raw patent records with one-on-multiple relationships between patent numbers and permnos to find additional links between the standardized company names and permnos.

   a. We skips the links if we already found the company name for the permno or found the permno for the company name after Step 2b.

   b. Multiple company names correspond to one permno: We fuzzy match all the company names related to that permno and then choose the company name which occurs most with respect to that permno.

   c. Multiple permnos correspond to one company name: This occurs because one patent id has multiple permnos in raw patent record. It is difficult to determine which permno correspond to the company name. Thus, we safely skip all such links.

4. Final Check the special cases (multiple permnos correspond to one company name or multiple company names correspond to one permno) are reasonable by using USPTO Patent Assignment Search: Confirm that the first assignor of the patent id is correct even under special cases

Step 4 Build the All Unique Standardized Company Names and Permnos Crosswalk

1. The crosswalk after Step 3 allows us to find permnos of the first assignor company names. However, these company names may vary in patent assignments even if they represent the same company. Thus, we find the permnos of all standardized company names of both buyers and sellers in all patent assignments by exactly matching and fuzzy matching these standardized company names from patent assignments with standardized company names from the crosswalk.
2. We first implement the exact match and the fuzzy match on company names.
    a. The TF-IDF with N-Grams Algorithm is used for the Fuzzy Match.
    b. This algorithm considers substrings matching by using N-Grams.
    c. It assigns more weights to less frequent words, which serve as the unique identifiers in the company names.
    d. It uses cosine similarity to measure the similarity between two strings. Doing so allows us to use the matrix operations to calculate the similarity between one string and the alternative set of strings simultaneously, which speeds up the algorithm.
    e. The range of similarity is between 0 and 1. The 0.5 cosine similarity is used as the threshold to find the fuzzy match. If the similarity is below 0.5, no fuzzy match is founded.
3. We only keep founded permnos for company names from exact match. We apply fuzzy match results to process company names again. Then, we apply exact match and fuzzy match on further processed company names to find permnos. Doing so improves both the number and accuracy of exact matches and fuzzy matches.

Step 5 Improve Exact Matches and Fuzzy Matches on Company Names

1. Use Fuzzy Match Results to further Process Company Names
    a. We find the end words with ≥ 100 occurrences in the fuzzy matched company names from all patent assignments.
    b. These words are common in all fuzzy matched company names. They are not the unique identifiers of company names. Thus, we remove these end words.

c. This allows us to match company names such as nokia networks with nokia, bookham technologies with bookham. This also leads to one-on-multiple relationship between raw company names and permnos.

2. Improve the Number of the Exact Match

   a. We implement the exact match again based on the processed company names after Step 5-1 to find permnos for additional company names.

3. Improve the Accuracy of the Fuzzy Match

   a. Implement the second fuzzy match by using the processed company names after Step 5-1

   b. Keep the fuzzy matched results if the similarity ≥ 0.8. We choose the relatively high threshold to ensure the accuracy.

   c. Drop the fuzzy matched results if the company names only have the one word

4. Develop Additional Fuzzy Match Algorithms by Textual Analysis

   a. Keep the matched results regardless of the similarity if company names are overlapped to certain extent. Identify overlap:

      i. The company names overlap if one is the complete subset of the other.

      ii. The company names overlap if more than half of words in two company names match exactly

5. Keep the Exact Match and Fuzzy Match Results in 2, 3, 4 above to Find Additional Permnos of Company Names to build the complete All Standardized Company Names and Permnos Crosswalk.

Step 6 Find Permnos of All Buyers and Sllers in Patent Assignment

1. Apply the complete Unique Standardized Company Names and Permnos Crosswalk after Step 5 to find permons of all buyers and sellers in all patent assignments

Codes Descriptions

1. 01_Create_First_Patent_Assignment.ipynb: Step 1

2. 02_Standardize_Company_Names_Both_Dataset.ipynb: Step 2

3. 03_Build_FirstAssignorName_Permno_Crosswalk.ipynb: Step 3

4. 04_Build_AllCompanyNames_Permno_Crosswalk.ipynb: Step 4, 5

5. 05_Find_Permnos.ipynb: Step 6

Output Descriptions

1. first_patent_assignment_1980_2015.csv: first patent assignments after Step 1
2. first_patent_assignment_1980_2015_standardized.csv: first patent assignments with standardized company names after Step 2
3. patenttransfer_companynames.csv: unique company names from both buyers and sellers from all patent assignments after Step 2-1
4. patenttransfer_companynames_standardized.csv: unique standardized company names from both buyers and sellers from all patent assignments after Step 2-2 to Step 2-8
5. permno_processedfirstassignorname_crosswalk.csv: the Standardized First Assignor Company Names and Permnos Crosswalk after Step 3
6. permno_processedallcompanyames_crosswalk.csv: the All Unique Standardized Company Names of Patent Assignments and Permnos Crosswalk after Step 4 and 5
7. patent assignments datasets end with _final: final patent assignment datasets with permnos included after Step 6