

Patentics.com

20370 Town Center Ln, Suite 139
Cupertino, CA 95014

Patentics Multiple Library Technology

*Partitioning Patent Documents Into Multiple
Libraries to Solve Homonyms in Different Languages*

Homonyms in the English language can pose a difficult challenge for information retrieval algorithms when some words with the same spelling have different definitions depending on the context. These multiple meanings can affect an algorithm's precision by returning more irrelevant results since it has to take into account each of the different definitions. This added noise will clutter and obfuscate the relevant results and the user must spend time filtering through it.

Patentics has found a unique solution to this problem by partitioning the patent data into several, smaller "libraries". These libraries are constructed using patent clustering so that the resulting groups represent 16 domain-level classifications, such as biological, chemical, computer hardware, etc. When a user issues a search query, Patentics will determine which library to search based on the semantic of the search query. Using the full-text of the patent to search will provide the best results since the algorithm will be able to extract the domain-level classification to determine the library in which to search for results, and limiting the domain in which to search and rank results will yield in more accurate ranking.

However, suppose a patent may have unique applications beyond its traditional field. In this case, only using the libraries to determine results would be potentially limiting and miss important results that are still pertinent to the original patent but are in different domains. To solve this type of problem of cross-functional patents, Patentics has further developed the algorithm to also take into account the ranking results using the full document set. By merging the rankings from the library and the rankings from the comprehensive domain, Patentics can present rankings that are tailored for the specific knowledge contained within a patent but also the general applicable knowledge. This merging process is itself a sophisticated algorithm that takes into account the ranks of a particular result using a domain-specific library and using the full document set so that the final ranking of the results are weighted accordingly.

Patentics goes one step further and also constructs these libraries for the different languages that are supported. As an example, there are domain-specific libraries for CN patents in native Chinese as well as CN patents in machine-translated English. By using these additional libraries, Patentics can mine the slight language nuances that are introduced during translation. Using the native Chinese libraries and the machine-translated libraries, we conducted a test to compare SIPO examiner search reports with our relevancy ranking resulting from using the full-text of the patent in question. The results from the 2 libraries can be found in the table below:

	Successful Ranking of Patent in SIPO Search Report
Native Chinese Library	59%
Machine-Translated to English Library	62%

As the table shows, there is a slight difference between using the native Chinese library and the translated English library, with indications that using English to search results in slightly better results, even when the text has been machine translated from the original language.

During the process of merging the domain library results with the full document set results, Patentics will also consider the patent rank from using the different language libraries, with the final ranking taking into account language translations and domain knowledge thus using every single possible combination to provide the best ranking for our user. This is only a fraction of our efforts and we are constantly working on our technologies and will continue to mine as much data as possible and present the best for our user.