Structuring the unstructured

Creating knowledge through visual analytics and the use of Tibco Spotfire with Attivio for text analytics of scientific patents

August 11th, 2014
Philip Skinner, PhD
Executive Director, Regional Marketing
Patent Searching through the USPTO

The following is a Seven Step Strategy for conducting preliminary searches of U.S. patents and published applications using the online resources of the USPTO and the EPO (European Patent Office). Search Engine Optimization (SEO) is available nationwide to provide training on the Seven Step Strategy.

PTO provides access to the PPTs (Patents Full-Text) and Adept (Applications Full-Text) databases on the USPTO website and the EPO Patent Full-Text Database on the EPO website. In addition, PTO provides access to additional USPTO patent databases through PTO/Alert (Public version of the examiner assisted search tool) and TIPS (Public version of the EPO/USPTO examiner tool). Always call your nearest PTO or TIPS to make an appointment before visiting the USPTO.

The use of patent classification in searching U.S. patents and published applications usually results in a more comprehensive search than one solely done by keyword (text) searching. Since keyword searching is challenging with the often technical and legal language of patents, these steps are designed to supplement a patent classification search. Patent classification searching is the focus of the Seven Step Strategy.

The USPTO is currently transitioning from the over 100-year-old U.S. Patent Classification system to a new classification system jointly developed with the European Patent Office called Cooperative Patent Classification (CPC). Beginning in 2013, the USPTO will only use CPC for classifying utility patent documents, but will continue to use U.S. Patent Classification for design and plant patents. The current version of the Seven Step Strategy reflects the transition between the two classification systems.

The Seven Steps in a Preliminary Search of U.S. Patents and Published Patent Applications

1. Brainstorm terms to describe the inventors.
2. Use these terms to find initial US Patent Classification (USPC) class/subclass in the Index to the US Patent Classification (USPC).
3. Verify the relevance of USPC class/subclasses using the US Patent Classification Schedules.
5. Retrieve and review complete U.S. patent documents and published applications by USPC classification in the PPTs (Patents Full-Text) and Adept (Applications Full-Text) databases.
6. Using the USPC classification(s) you previously identified, find relevant Cooperative Patent Classifications (CPCs) using Statistical Mapping from USPC to CPC, retrieve and review U.S. patent documents and published applications with these CPC classifications in the PPTs and Adept databases.
7. Conduct a Classification Search of CPC Class Schemes on the PTO’s transparent website to expand your CPC searching to additional relevant CPC classifications. You can then use these additional CPC classifications to find additional relevant patent documents and published applications on the PPTs and Adept databases on the USPTO website.
Once your search is complete

Searches result in PDFs which contain lots of data but in an unstructured format
The Challenge

There is a wealth of untapped knowledge in published patents and patent applications globally related to biochemistry and other fields

- Chemical Reactions
- Yield Rates from various catalysts
- Chemical/Compound Drawings and Diagrams
- Variety of valuable business information (Inventor, Assignee, Geographic Location, etc.)
- Similar, referenced, and related patents

How do you structure that unstructured data to mine it?

The Challenge: Develop a POC to illustrate a greater understanding of Key Opinion Leaders, Centers of Excellence, any trends in reaction yields, patterns in concepts etc. through structuring data from a selection of patents
The Plan

Identify target patents

Download target patents

Load into Attivio Active Intelligence Engine & extract structured data

Enable visualizations and advanced analytics

Develop & execute additional extraction scripts
Identify Target Patents

Identified a series of 100 US patents within key topics (Suzuki, Sonogashira, Heck, Ullman couplings) from Scifinder
Identify Target Patents

Patents were loaded into the Attivio Active Intelligence Engine and core metadata extraction scripts executed:

Patent Number, Inventors, Assignees, Locations
Additional Scripts

Additional extraction scripts developed
e.g. Identify Yield where *palladium* is proximate
Search Capability enabled through Attivio interface
Search Capability enabled through Attivio interface
Tibco Spotfire enables advanced analytics

Extracting Key Opinion Leader and Key Concepts from US Patents

The following visualization contains an extract of around 200 US patents, identified from key concepts of certain palladium-catalysed reactions. Patents were identified using SciFinder, extracted using Attivo and additional extraction procedures developed by Bardess and examined and visualized in Tibco Spotfire from PerkinElmer.
Catalyst system provides most efficient, general, and mild catalytic combination reported to date for the amination of aryl iodides, aryl bromides and aryl chlorides, and in forming carbon-carbon bonds between aryl boronic acids and aryl halides by displacement of the halide and boronic acid moieties

US 7385058 B1

ABSTRACT

Methods and compounds are provided for the formation of carbon-nitrogen or carbon-carbon bonds comprising reacting an amine or an aryl boronic acid with an aryl halide in the presence of a palladium catalyst, a base, and a compound of formula I:

\[
\begin{align*}
R_2 & \quad R_1 \\
\end{align*}
\]

DESCRIPTION

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims the benefit under 35 U.S.C. 119(e) of U.S. Provisional Application No. 60/013,303, filed on Sep. 27, 2004, which is hereby incorporated

CLAIMS (18)

1. A compound of formula I:
Explore which patents came from where

The relevant patents are shown
Explore which patents came from where

Choose the assignees of interest

The relevant locations are shown
Is there a difference in average yields?

- Yields of the form x% were extracted when proximate to *Palladium*.
- Average yields by reaction class were then displayed on a boxplot with 95% confidence circles.
- No difference was observed for this sample set.
Explore Key Concepts from patents

- Key Concepts were extracted using the Attivio Active Intelligence Engine and built into a second table within Spotfire.
- An Open Source R data function, using the wordcloud & tm packages, converted the key concepts table to a cleaned, single word corpus.
- The corpus was then returned as a wordcloud image and an additional table of single concept words.
Explore Key Concepts from patents

- The corpus was then returned as a wordcloud image and an additional table of single concept words.
Next Steps

- Deepen set of patents to be able to see more time based metrics
- Expand to WO and other non US patents and beyond palladium catalysts
- Match catalysts names to yields to see if catalyst types had different yield characteristics
- Extract other reaction properties (Solvent, time, temperature)
Conclusions
Acknowledgements

PerkinElmer
- Jeremy Rualo
- Gerard Conway
- Joshua Bishop
- Megean Schoenberg
- Michael Swartz

Attivio
- Rik Tamm-Daniels
- Matt Connon

Bardess
- Michael Prorock