Patent Analysis and Visualization: Today's Challenges and Opportunities

Search Technology

Nils Newman
newman@searchtech.com
The last two decades have been very good for patent analysis and visualization

- Fantastic growth in computing power
- Significant progress in supporting software
- Major improvements in analysis algorithms
- Substantial advances in visualization technology

We can now process more information faster, and get more useful stuff out of it than we ever could before.
However.....

We still don’t have the magic bullet.

There is no real easy button for patent analysis.

The process still takes skill, time, and a little patience.

Part of the complexity is inherent in the process....
And part is in the tools

The bits that are assembled together to make analytical tools are still under development.

When you are working with analysis tools, you are working with technologies that are still growing.

There is no “right” tool and there is no real standardization within the tool building community.

Thus as a user, you are working with a moving target.
So don’t trust the salesman

Since the technological ground truth is changing constantly within the tool building community, you as users cannot take what we as tool builders say at face value.

To understand the capability and issues with any analytical tool you need to look under the hood at the main components that make an analytical process work.
The major elements in Analytics

There are many different ways to chop up the analytic process.

One way is to follow the knowledge flow from data source to final reporting.
The Import Process

Homogeneity  Structure
Data Homogeneity

• Data can come from
  • One source (easier)
  • Many sources (harder)
• Homogeneous single source databases have the potential to allow for direct connection of the analytics tool to the data
• Can connect at the API level or at the Database level
• XML is making these kinds of connections easier but is inflating record sizes

• Data from many sources requires additional effort, but XML is making things easier as well

Trend: Most patent data providers are starting to provide some level of analytics directly on their platform. Scientific information providers are following too, but slower.
Data Structure

- Structured content (e.g., a patent front page) is relatively easy to work with
  - Finding content within structured text can be done with technologies like Regular Expression Pattern Matching

- Unstructured content (e.g., the full text of a patent) is a bigger challenge
  - Creating structure in unstructured content requires technologies like Content Tagging
  - Tagging systems work well with general classes of content (e.g., geographic place names, people’s names, company names) because the training system can be generic
  - Tagging topic-specific content usually requires specialized training sets (and someone to train and maintain it)

Trend: Generic tagging systems are impacted by economies of scale. The more data they have to work with, the better they work. Thus they tend to be incorporated within the data production side of the business.
The Data Manipulation Process

Extract  Clean  Mashup
Extraction

- Technologies like Named Entity Recognition (NER) can be used to further extract key user-specific topics from within chunks of text (e.g. Claims). At this level most NER is topic-specific, requiring supervised training sets. (Very different from generic NER systems.)

- Tools like Natural Language Processing (NLP) require less maintenance but produce more generic results - pulling out things like noun phrases.

- Most extraction tools require some kind of training and can be language-dependent.

Trend: Topic-specific NER systems require expert knowledge to build and effort to maintain. As such, they are rarely given away for free.
Cleaning

- Data from even the best curated source usually require some cleaning
- Technologies for cleaning include:
  - Fuzzy String Matching
  - Rule Sets
  - Thesauri and Ontologies
  - Disambiguation
  - Manual Editing
- The ability to clean is as dependent on the interface and the ability to store and recall changes as it is on the underlying cleaning algorithm

Trend: Cleaning data is not just a correction process. Its results often reflect standards and perceptions that are unique to a specific organization. The cleaning process allows a user to imprint their point of view on the data.
Mashup

- The process of creating hybrid content enables the user to augment standard sources with additional data.
- The technology behind merging data together is fairly simple but the analytical ramifications of working with hybrid content can be quite complex.
- As a result you will see many systems that either let you add content before the analysis or will not let you fully analyze the added content.
- Ideally a user should be able to add and analyze content at any point during the process.
- The added content can be additional text or numerical data.

Trend: As analytic systems become more data platform-specific, mashup technologies might become harder to find.
The Analysis Process

Navigate  Interact  Cluster  Visualize
Data Navigation

- There are four main navigation paradigms
  - The Record
    - Based on traditional search results
  - The Meta Tag Header
    - Designed to mimic common web search engine results
  - The Table
    - Drawn from the Excel experience
  - The List
    - Has roots in Data Mining and Statistical Analysis

Trend: There is no trend. Dominance of a navigation paradigm will arise when there is some standardization in text analytic processes.
Data Interaction

- There are two main approaches to data interaction: one can impose order on the data or one can let the data self-organize
  - Self organization typically means a **Statistical** approach
    - Has its roots in Data Mining and Bibliometrics
    - Uses tools like co-occurrence, correlation, and TFIDF
  - Imposing order typically means an **Artificial Intelligence (AI)** approach
    - Has its roots in Machine Learning and Computational Linguistics
    - Uses tools like relevance ranking, Natural Language Processing, and Sentiment Analysis
    - Typically requires a supervised or semi-supervised approach
- Both techniques have the goal of facilitating your understanding of the content, increasing your ability to digest large amounts of data

Trend: The statistical side of data interaction is the most stable with the data mining field acting as a proving ground for new ideas. AI-based systems are more volatile with more sources of change.
Clustering: Organizing Data

- Clustering can come in two forms which analytically mean very different things:
  - Document Clustering -> Categorical approach
  - Term Clustering -> Descriptive approach
- Clustering covers a broad range of technological approaches which all have a common goal of organizing the data.
- The technological approach will vary depending on the data interaction approach.
  - Statistical approaches include LSA, LSI, PCA, pLSA, AS/PI
  - AI approaches include SVM and TM

Trend: Technology in this area is rapidly changing. Some of the techniques have been around for less than a decade. Some have been around for years but are constantly being tweaked by the combination of additional algorithms. Implementation is key in this area.
Analytical Visualizations

- Visualization software falls into two very different types:
  - Visualizations for Exploration
  - Visualizations for Communication
- Visualization for text analytics has a checkered past
  - Tool makers have been trying to make visualizations that you will use to make decisions with little success
  - For now, visualizations that help you explore and navigate within the data are more effective
- The tools are pretty generic – the bar graph, the donut, hyperbolic trees, surface maps, heat maps, network diagrams, etc… but implementation can be tricky

Trend: Pretty much all the visualization technologies we use in text analytics were designed for someone else doing something completely different. Text analytics as an area is too small to have its own dedicated visualization designers.
The Reporting Process

Visualize  Workflow
Visualization for Reporting

- Visualization for Reporting boils down to “what would you show your boss”
- The constraints are organizational not technological
- As tool developers we see a wide variety of behavior within text analytics
- So we are left to try a variety of different visualizations hoping that they stick
- However, we ultimately know there is no universally accepted way to visualize information
- As the community matures it will hopefully move toward norms

Trend: On a technical level there has always been a tension between visualizations for the Web and visualizations for the Microsoft Office world. However, the tablet movement and Windows 10 could be very disruptive.
Workflow

- Text Analytics fits in a workflow
- Thus the analytics process and output should not put up barriers to integrating the content within the workflow
- The inclusion of analytics in the workflow requires access to infrastructure and technologies that are usually outside the scope of core analytics developers
- For now, we rely on the data producers and IP workflow software developers to facilitate this union
- In the future, standards have the potential of creating open systems, where different analytical systems, data sources, and management systems could all talk to one another

Trend: Fully integrated workflow solutions will appear first on closed unified systems (single source data-analytics-ip management). Open systems will take a bit longer.
How to decide on tools

- Begin by looking at the quality of your data

Note: Data homogeneity plays a role in this decision step. If you are working with many types of data, you might need a solution for each type of data or might get by with mixing all together first and then using one approach.
How to decide on tools

• You also need to decide if you want to let the data self-organize or if you want to impose your own (or someone else’s) order.

Note: Remember that imposing order typically comes at a maintenance cost. The more order is imposed, the more someone has to spend to curate the externally created order. Self organization is typically cheaper upfront but requires more expertise by the analyst to interpret results.
How to decide on tools

- Reality will usually be an engineered hybrid approach

Note: Tool developers often build in a variety of approaches into their systems but tend to hide them to make tools look simpler than they actually are.
What does this mean to you?

• There is no “one size fits all” when it comes to text analytics
• Analytical techniques still need to be matched to your data and your problems
• The state of the art is rapidly evolving
• You need to have a good sense of how the tools that you use actually work
• Otherwise, you have to take it on faith that what we are doing fits your analytical situation
In Summary

The underlying technologies that support Patent Analytics are continually improving.

However - generally, patent analytics doesn’t create its own technologies. It borrows them from other areas.

Almost all the engines of change are all outside of patent analytics.

This means technical change can be very quick.

It also means you as a user must have some understanding of what is under the hood within our tools.
Questions?

Nils Newman
newman@searctech.com

Search Technology

www.theVantagePoint.com