

White Paper

Text Clustering on Patents

Manish Sinha
Chief Technology Officer
© 2009. Gridlogics Tech. Pvt Ltd

Overview

Amongst various analyses performed on patents, the area where specialized software helps immensely is text-mining and two of the most popular text mining techniques used over patent data are:

- Text segmentation / Tokenization
- Text Clustering / Topic identification

Text segmentation is a process of analyzing the patent text and identifying smaller meaningful segments from the text. These segments are also called as Tokens or keywords. Various analyses can be performed using these tokens; however for large patent sets the number of unique tokens can be very high thereby making it unsuitable for certain types of analyses.

Text clustering helps identify important topics or concepts (clusters) from a set of documents. Clustering of key patent data documents (such as Title, Abstract and Claims) has been used in various Patent Analysis tools and can help bring out the otherwise hidden insights within patents. Analyzing relationships between generated clusters or analyzing relationships between patent classifications and clusters are popular mechanisms used by researchers especially those in a competitive intelligence role.

Clustering Algorithms

Classical algorithms used for clustering are TF-IDF, K-Means or Bayesian Naïve. Their output is a set of topics (single level or hierarchical with multiple levels), each of which contain a group of documents cluster under the topic. The label (or name) of a topic is derived from the text of the patent and can be a combination of multiple words. Usually advanced algorithms understand parts-of-speech and interpret names accordingly. For instance, “tip of the probe...”, “using a probe tip...”, “with a probe whose tip is used for...”, and “probing the substrate with the tip...” will be represented as “probe tip”.

Algorithms can differ on their capability to cluster patent under single or multiple topics. For patents it is desirable to have algorithms that place a patent under more than one topic since an attempt make a best-match for the patent under a single topic may lead to errors in user interpretation.

One of the key challenges of clustering algorithms has been to make the overall process transparent for IP Professionals who usually find it hard to accept results of “black-box” clustering engines. Newer clustering algorithms provide the analyst control over most aspects of the clustering process thereby allowing them to fine tune the clustering output and train the algorithm to deliver more relevant results. The analyst can also now specify how deep s/he wants to go and choose to highlight broader or finer (more specific) topics.

Another challenge has been to make it easy for the analyst to influence the choice of the terms being picked up such to avoid picking up meaningless terms or to give importance to a certain *class* of terms. Apart from just being able to ignore some words (stop-words) it is

important to have the power to set advanced filters that decide the choice of the labels used to represent the topic. For instance these can be, whether to ignore or give more importance to - words in ALL Caps, words starting with or containing a number, words containing more than one hyphen (compound names) or just words that are too long or too short. Even a minor change using such filters can have a reasonable effect on the topics that are chosen.

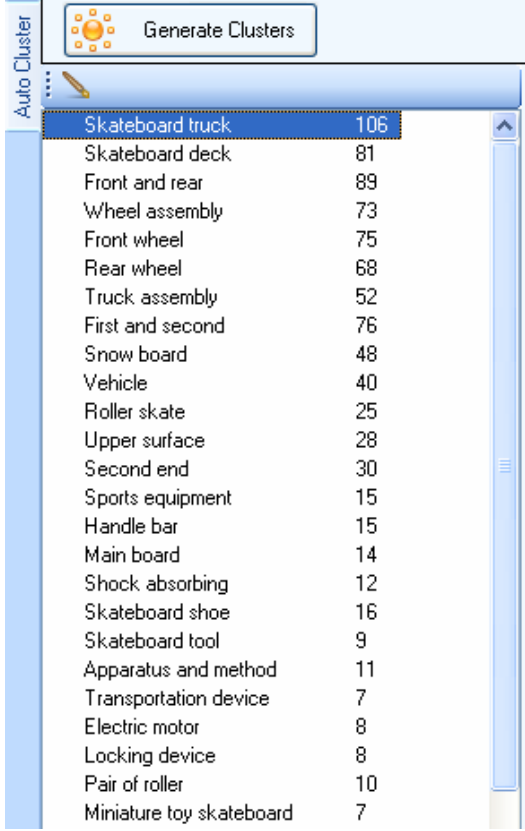
Benefits of Text Clustering on Patents (Use Case Scenario)

Lets understand the benefits of patent text clustering using a sample case use case scenario. on patents related to **skateboards**.

In this sample set, we did a simple search for the word “skateboard” in Title, Abstract and Claims of patents across key countries and then de-duplicated the results to only unique families. This resulted in 552 unique inventions.

Text clustering was then performed using Patent iNSIGHT Pro* over the Title, Abstract and Claims sections of these patents and the results obtained are illustrated below. We have used the sub-topics on Skateboards used in Wikipedia as a sample for cross-reference.

Table: Automatic Clustering vs Manual Categorization

Single level Categorization by Patent iNSIGHT Pro Keyword: “Skateboard” in T-A-C	Categorization in Wikipedia Reference: http://en.wikipedia.org/wiki/Skateboard
	<p>Contents</p> <ul style="list-style-type: none"> 1 History 2 Parts <ul style="list-style-type: none"> 2.1 Deck 2.2 Trucks 2.3 Wheels 2.4 Bearings 2.5 Hardware 3 Optional components <ul style="list-style-type: none"> 3.1 Rails 3.2 Slip Tape 3.3 Lapper 3.4 Nose guard 3.5 Tail guard 3.6 Angled risers 4 References

* Note- A new Advanced Clustering engine has been recently added to Patent iNSIGHT Pro

As can be seen, clustering algorithm outputs do overlap to a large extent with standard human developed classifications.

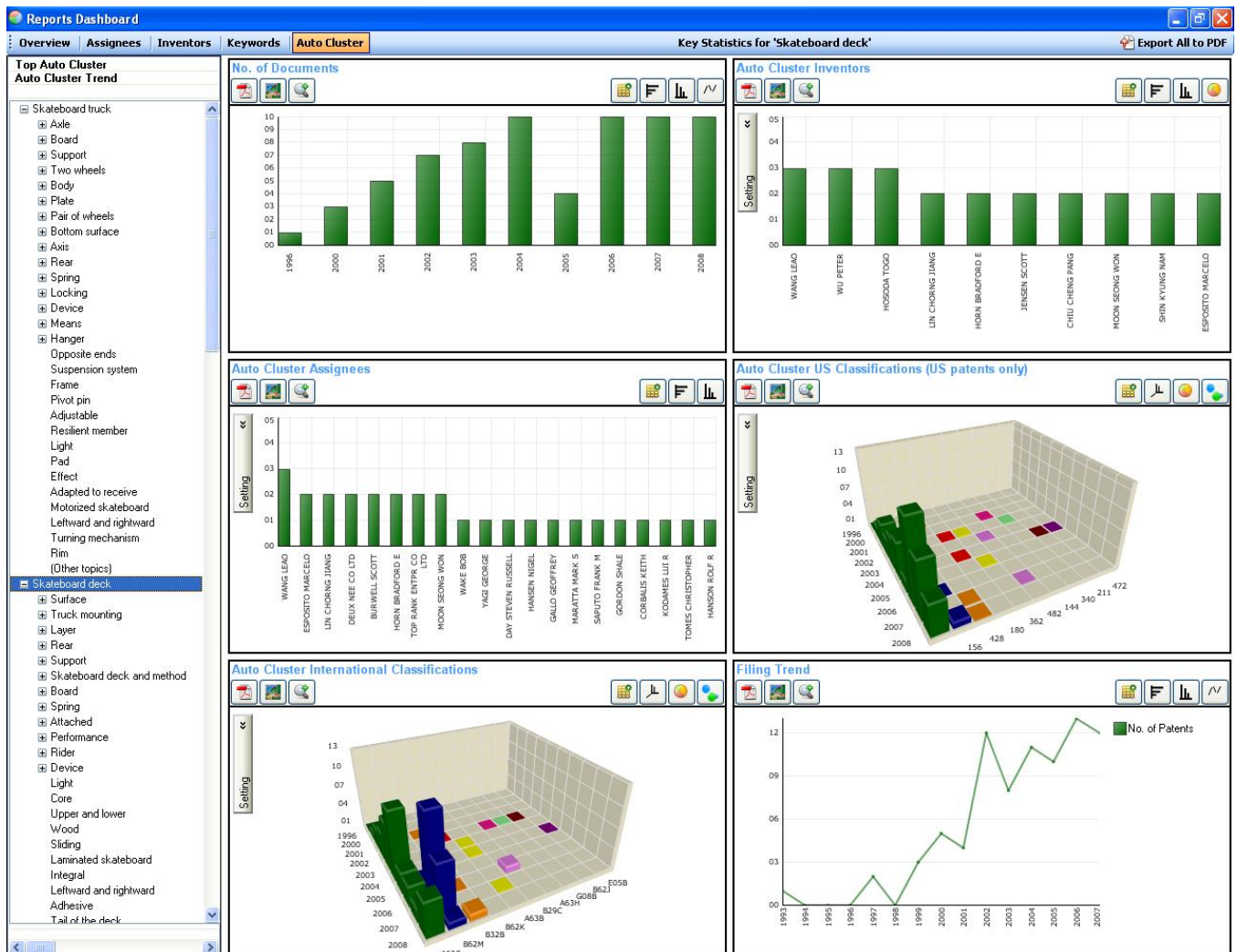
Multi-level (hierarchical) clustering algorithms allow analyst to further sub-divide these clusters and see how each topic is further broken down. In the current example, we set the clustering engine to generate 2-level and 3-level clusters:

Table: Hierarchical Clustering

2-level categorization by Patent iNSIGHT Pro		3-level categorization by Patent iNSIGHT Pro	
Skateboard truck	106	Skateboard truck	106
Axle	32	Axle	32
Board	30	Plate	9
Support	22	Support	8
Two wheels	18	Axis	8
Body	17	Axle bracket	6
Plate	15	Pair of wheels	7
Pair of wheels	15	Pivot pin	5
Bottom surface	13	Hanger	5
Axis	14	Nut	5
Rear	13	Deck	5
Spring	13	Suspension	5
Locking	11	Riding	4
Device	11	Bottom surface	3
Means	12	Adjustable	3
Hanger	10	Mechanism	3
Opposite ends	8	Bearing assembly	2
Suspension system	6	Ice skateboard	2
Frame	7	Skateboard or roller	2
Pivot pin	6	Obstacle	2
Adjustable	7	Elastomer	2
Resilient member	5	Frame	2
Light	5	Pad	1
Pad	5	Board	30
Effect	5	Wheel support	8
Adapted to receive	4	Bottom surface	7
Motorized skateboard	4	Two wheels	7
Leftward and rightward	3	Platform	7
Power source	3	Body	7
Turning mechanism	3	Skate	6
Runner	2	Device	6
Laminated skateboard	1	Sports board	5
Carrying strap	1	Attached	6
Rim	1	Rear	5
(Other topics)	6	Ice skateboard	4
Skateboard deck	81	Deck	4
Front and rear	89	Means	4
Wheel assembly	73	Suspension system	3
Front wheel	75	Roller board	3
Rear wheel	68	Skateboard locking	3
Truck assembly	52	Plastic	3
First and second	76	Spring	2
Snow board	48	(Other topics)	1
Vehicle	40	Support	22
Roller skate	25	Two wheels	18
Upper surface	28	Body	17
		Plate	15
		Pair of wheels	15
		Bottom surface	13
		Axis	14
		Rear	13
		Spring	13
		Locking	11
		Device	11

Hierarchies (like above) are of immense help when you are developing your own internal classifications (also referred to as taxonomy). They help you cross check and refine the categories you created. Having IP portfolios mapped with well-defined industry specific internal classification is popular within large corporations.

An obvious use is to quickly view the trends (as shown below) across clusters and get the big picture on how your companies portfolio is spread out, or understand which technologies are becoming mature and phasing out and which are the new “focus areas” that are coming up.



A lateral benefit is that these clusters can also be used as means to speed up shifting through large sets of patent data or narrowing down to the specific technology or area you are interested in.

When categorizing patents, it is sometimes important to map a portfolio to actual market products to see how various technologies and IPs are spread or to see a bigger picture by correlating parameters such as the market size of each product or your financials (costs/revenue) behind each product. For creating such product-patent mapping, the analyst can merge or map algorithm generated clusters rather than individually mapping each patent to the products and in this way, drastically reduce the time taken to categorize large patent sets.

Summary of Benefits

To sum up the benefits, patent text clustering can be used:

- To reduce time taken to categorize large sets of patents
- To enhance scanning efficiency and to narrow down on a specific area
- To identify sub-topics, discover hidden topics and understand the trends and competitive focus in each topic
- To develop internal taxonomies or as a reference to cross-check and improve your manually prepared categories
- To reduce time taken to develop advanced categorizations such as product-patent mapping or market-patent mapping
- As a new dimension to analyze along with standard bibliographic fields (for trend analysis and competitive landscaping)
- And, as a reference to refine your search strategy.

About Patent iNSIGHT Pro

Patent iNSIGHT Pro™ is a comprehensive patent analysis platform that allows you to accelerate your time-to-decision from patent analysis activities. Designed from inputs by experienced patent researchers, Patent iNSIGHT Pro easily blends into your existing Research workflow. Patent iNSIGHT Pro is used by leading legal services, Pharmaceutical & biotech, electronics companies and research organization across US, Europe, South America and India with more than 150 end users. Patent iNSIGHT Pro is developed and marketed by Gridlogics, a research driven IT Company specializing in providing intellectual property analysis and visualization solutions to aid R&D and corporate strategy.

Gridlogics is headquartered in Pune, India and has a sales presence in Delhi, Mumbai and USA.

For more information:

Visit us at: www.patentinsightpro.com

Or call us at: 1-408-786-5524

Or mail us at: contact@patentinsightpro.com