Dive deep into patent data

Gaining insight into the vast knowledge base of patent data is a Herculean task. Demand is thus growing for sophisticated patent analytics tools that can uncover business opportunities, identify gaps in R&D strategies and expose competitive threats

By **Matt Troyer** and **Art Nutter**, TAEUS International Corporation, Colorado Springs

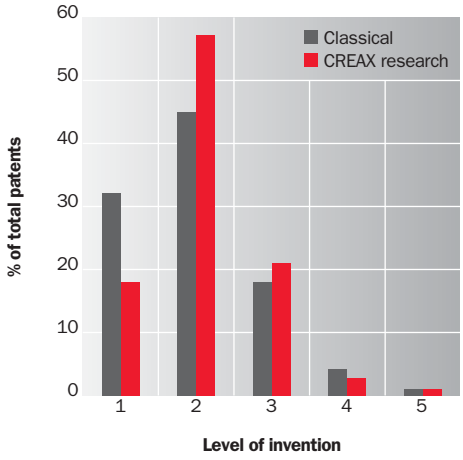
The financial impact of patents to corporations cannot be denied. Annual revenues stemming from patent licensing are conservatively estimated at well over US$120 billion in 2008. Patent licensing revenue often makes up over 10% to 20% of total income in R&D-centric companies. At the same time, patents are effectively pre-paid assets, providing a return on investment well over 95%, if you regard the R&D expenses as costs sunk well before the patent itself is monetised as a standalone asset – that is, detached from a particular product.

Moreover, thanks to international regulations on patent applications, patents are a valuable and unique source of up-to-date scientific and technological information. Patent applications are published 18 months after the date of filing, usually before the first action on the patent and well before the patent is granted. The consensus among experts we have spoken with suggests that the scientific knowledge contained in patents comprises about 80% of all scientific knowledge, making it by far the most useful corpus of data available to scientific or innovation researchers inside today's corporations. While 80% may sound reasonable (or not) at first glance, its accuracy can be further argued by analysing the real data in the body of patent citations. The importance of patent data to the corporate technology researcher is further amplified by the fact that nowhere else in the world is so much technological information available in one place.

Gaining insight into the growing knowledge base represented in patent literature is an immense task. The number of patent documents worldwide is at least 60 million, a figure which is growing at a rate of 1.5% year on year. Approximately 750,000 patents were issued worldwide in 2007 and almost twice as many patent applications published. (Roughly 25% of patent applications are abandoned before they are published, according to the most recent USPTO statistics from 2007 data.)

As the volume of patent documents grows, so too does the opacity of the patent data, making it more complicated for patent researchers to digest, absorb and glean insight (or actionable information). Put simply, unfiltered data is anarchy. It is therefore unsurprising that demand for advanced patent analysis software and document processing techniques that ensure easier access to patent data is on the increase.

Today, there are a growing number of techniques for analysing patent literature, each with their own benefits and limitations. The field is rapidly changing, due to larger market forces. More than any other factor, patent data has only recently become available to researchers in machine-readable format at a reasonable cost. Moreover, raw computing power is greater today than ever before and access to supercomputer-like processing power is available at a practical cost through massively parallel clustered servers such as the Amazon EC2 cloud. Today Thomson Reuters still controls almost 80% of the patent information and



processing market, but the genie has been unleashed and innovative new approaches are being undertaken by researchers and entrepreneurs interested in patent processing.

With the exception of search or document retrieval, automated patent processing techniques have not proven dependable and therefore there is a tendency among potential users to dismiss new entrants into the marketplace as more of the same. However, we believe that this attitude is risky in light of the amount of work that has gone into patent data analysis in the last few years.

There will never be a magic bullet that will slay all the dragons, but tools available today and in the near future will bring new efficiencies and a competitive edge to those who effectively embrace them.

What patent analysis can do for your company

If you accept that approximately 80% of technological innovation is represented in patent data, surely you also recognise that there is tremendous value in mining patent literature to further your company's goals. Patent analysis, used correctly, can uncover information that saves or makes today's corporations millions of dollars. There follows a partial list of some of the benefits it offers:

- Patents constitute an important source of information about competitors. They represent the only publicly accessible indicator (albeit a lagging indicator) of competitors' future strategies. As a result, analysis of patent data is one of the primary tools for technological and business intelligence.
- Patents can be visualised as part of a larger landscape of similar and overlapping technologies. A careful patent landscape study can point towards companies that are good candidates for M&A; identify new competitors in the market; uncover potential future threats from existing competitors; and flag up outlying patents from smaller firms or individual inventors for in-licensing or acquisition.
- Patents constitute about 80% of all cited prior art and cited documents in state-of-the-art studies. They are the most accessible source of information when trying to invalidate a competitor's patents for both anticipation (single document) and, more commonly, obviousness (combining several documents).
- Patent analytics can show you what you have in your portfolio and identify technology groups to which they belong. Analytics is a valuable first step when performing an internal IP audit.
- By substituting a patent characteristic for value (there is no transparent market for patents and therefore no way to find comparables), such as renewal rates, litigated patents or patents that have international counterparts, you can develop profiles for all your patents, which can help to prioritise your investigation strategy.
- Patent analytics can uncover potential buying opportunities to fill gaps in your existing portfolio.
- Corporate licensing executives can identify potential licensees by uncovering who else is patenting in the general area of the innovation. It's like a GAP analysis in reverse, where your technology may fill holes in a competitor's technology.
- Advanced linguistic processing techniques can be used to make reading patents far easier for the lay businessperson, unaccustomed to reading the style and legal jargon presented in patent claims.
- Targeting specific technologies that read on your patents can be accomplished when patent analytics is combined with an analysis of non-patent literature.

Challenges in patent processing and document retrieval

Many patent search and retrieval systems exist on the market today. Most systems, however, use general-purpose search engines that do not take into account the important characteristics of a patent document – especially in the area of semantic analysis (using software to uncover meaning).

One the one hand, the macro-structure of a patent document is characterised consistently from document to document. Each patent has a number, application date, issue date, citations, inventor(s), assignees (owners), classification(s) and other structured data that can be exploited by analytics software. Most of the patent analytics software commercially available today has focused on this area, helping users with both patent landscaping and patent profiling techniques.

However, the bulk of every patent document, and the inventions disclosed therein, constitute free-form text written by

the patent agent or attorney who prosecuted the patent. Each agent has a unique style and vocabulary, making the language used inconsistent from patent to patent. Text analysis is further complicated by the fact that patents are written in vague, general legalese, especially in the claims, to enlarge the scope of the invention – making it difficult for linguistic processing systems to handle.

To illustrate, by way of example, a wheel in the preamble of a claim may be described as "A circular translocation facilitation device comprising…". It is possible to create a completely different statement having ostensibly the same meaning: "A round transportation assisting machine comprising…" You many never find the word "wheel" in a patent that applies to wheels, or RFID in an RFID patent.

In theory, but not yet reduced to practice, phrases such as these can be normalised using hand-crafted lexicons and concordances in various fields on invention, and further using text processing techniques such as parts of speech tagging, stemming (finding word roots) and lemmatisation (converting mice to mouse).

Other technical challenges in processing text include variables around the location and sub-location of the found phrases. Consider a patent's claims. Do you look more closely at independent claims than dependent claims? Do you weigh terms in the preamble of a claim differently from those in the body of the claim? Each section of the text of a patent document has significances that can be exploited for different tasks.

A further challenge is to capture the output of the analytics and create a useful weighting scheme that exploits statistical information from multiple variables:

- Term frequency – how often the term (or phrase) occurs in the document.
- Inverse document frequency – a measure of the importance of the term by comparing the frequency of the term to how often it appears in a larger collection of similar documents. Terms that occur in few documents are thought to be more important than terms that occur in many.
- Document length – terms that appear the same number of times in a short document can be weighted heavier than terms appearing the same number of times in a longer one.

Language processing will never be perfect, even using the most advanced statistical methods, sophisticated rules engines and custom lexicons tuned for various parts of the patent literature. Natural language is simply too variable and too complicated.

To illustrate, in developing Keyhole™, our company's natural language patent analytics software, we developed a statistical model to extract meaning from patent literature. On top, we also developed a rules engine that processes exceptions generated by the statistical model. In US patents claims, we counted 533 predicates (verbs) that make up our statistical core – the core is defined as the verbs that account for 98% of those used in patent claims. The other 2% of predicates include several thousand infrequently used terms. For our purposes, we developed a set of rules around the core (98%), and for practical reasons were forced to reserve the other 2% for future updates to our software.

Put in another context, one of the fathers of Latent Semantic Analysis, Tomas K Landauer, currently at the University of Colorado, ran a statistical proof published in the *Handbook of Latent Semantic Analysis* (Landauer *et al* 2007), where he suggests that statistical techniques can mimic the vocabulary and "understanding" on English texts as well as applicants to US colleges from non-English speaking countries. Not perfect, but remarkably good.

As stated above, the magnitude of patent data is staggering and there is simply no way that any individual or team can rifle through it without the aid of high-powered search and document processing techniques. The benefit of patent analysis is not that it is spot on, but rather that it can cut the research time by trained experts dramatically and allow them to perform tasks that would be impossible without it.

What are we really dealing with?

It is important to recognise that patents vary dramatically in quality. According to the classical TRIZ research and a follow-on study by Darrell Mann and Simon DeWulf (Updating TRIZ: 1985-2002 Patent Research Findings), 75% of patents cover routine or minor improvements to existing inventions, 21% are fundamental improvements to existing inventions and less than 4% represent a new generation of system or a pioneering innovation. Our own experience, having reviewed over 50,000 patents for their technical quality, correlates to these numbers. A single level 4 or level 5 invention

can affect the corporate bottom line more than all your level 1 and level 2 inventions combined.

It is important to recognise that patent analytics do not determine patent quality. Today there is no system in place that can effectively do that. Patent analytics can uncover likely suspects, but human judgement is required to measure a patent's significance in the market.

Classes of technique

As the marketplace has developed for patent processing software, it should be recognised that several packages make use of several, all or one of the following general techniques to mine patent data. Each implementation is proprietary or custom and the details are not disclosed to the public, but it is a good idea to understand some of the basic techniques applied in software, and their benefits, limitations and best uses by your organisation.

Boolean and "smart-Boolean" techniques

Most board-level executives are familiar with Boolean searching: when you use Google or Yahoo! or another search engine, your initial query is a Boolean query. If you type in "video-on-demand" + "interface" you will get a list of all documents containing those two keyword phrases. They are not dumb Boolean queries, however. Based on the famous algorithm that determines a document's relevance based on the number and quality of back-links to that page (the precise algorithm is far more complex and is a trade secret to discourage both competitors and gamers), a list of the most relevant web pages is returned which increases the usefulness of the result set.

Most patent-specific search engines are tuned using various techniques to return the most significant results based on a proprietary algorithm as well.

Thomson's Delphion is the most famous and one of the most widely used examples of a smart-Boolean search engine. Its algorithm uses several smart techniques to improve relevancy of the documents returned, including weighting based on frequency of the terms in the document; automatic stemming so that a search for the word "drive" returns documents with the terms "driver", "drives", "driving" and so on; automatic phrase searching; term weighting; and special operators such as wildcards and "accrue".

Statistical correlations to patent quality

Analytics software is used to measure patent quality. It is a fascinating part of patent analytics and involves determining what measures of a specific patent compare favourably to patent quality or value.

We are aware of at least 30 factors that are generally considered when profiling patents, but there are probably more than 100 less significant indicators of patent quality.

For example, a large number of forward citations (future patents that cite the target patent) significantly corresponds to high renewal rates. A large number of forward citations is also an indication of technological importance, a measure of fundamentality of the patent and value.

Other factors that correlate to value include number of claims, length of claims, number of inventors, backward citations, length of specification and more.

These types of analysis are indicators or profilers. They suggest that a patent may be valuable (or not) without regard to the claimed invention in a patent. The problem, when applied to an individual patent, is that the results are indiscriminate. Patents are unique by statute. The variance in quality is considerable. Taking valid statistics across large populations of patents, then applying them to a single patent, is inaccurate if not used for their expected purpose.

To illustrate, statisticians study populations of humans (a variable population, like patents) and, based on gender, race, age, income, education and other factors, can determine whether the person is statistically more likely to be a terrorist, a consumer of sugared cereal or a Nobel Prize winner. While these metrics work well for marketers and help them to target advertising to a large demographic that is more likely to buy their product, it says nothing about the individual – he or she is free to buy the product or not. The rule of large numbers must be applied to leverage statistics of this nature for commercial value.

It is unreliable to apply statistics gleaned from large populations to individual people or patents. In the case of criminal profiling, stereotyping, gender biasing and cultural biasing, the practice is controversial and can lead to misuse by perpetrators of the act.

In the case of patents (arguably a more variable population than humans), statistical regressions are best used when applied to large populations of patents – for example, comparing portfolios from multiple electronics firms side by side. Statistical regressions can give you a data point to consider, but – dumb luck aside – it will

never find that nuclear patent. Level 4 and 5 innovations are as rare as Nobel Prize winners, and even if regression analysis can pick them up, they are almost equally as likely to be missed.

Natural language processing and indexing

The most complicated, most interesting and potentially most useful technique for evaluating patent text and mining patents is also the most underutilised, especially when applied to patents. An entire field of research is called natural language processing. One well-known statistical technique, mentioned earlier, is called Latent Semantic Analysis (Indexing) (LSA(I)). It can be weighted with factors such as citation analysis or PTO classification system as part of a broader analytics tool, but its primary advantage is finding hidden meaning in bodies of text. The first patent on LSA was issued in 1988 and it has been an active area of linguistics research ever since.

LSA is the most widely known variant of the statistical approaches to uncover meaning in bodies of text. Implemented well, searches leveraging natural language processing of patent text can almost always uncover patents that are missed by trained experts using smart Boolean searches – and you never know whether these are some of those level 4 or 5 patents.

Our experience is that statistical analysis of patent text gets us 80% of the way to our goal. We found that to increase accuracy and develop better relationships between patents, we have had to create a rules engine to process exceptions. Rules engines are intricate and are developed around a single lexicon – in our case, patent claim language.

Now show me

The final challenge in patent analytics is finding a visualisation that presents data in useful ways, providing quick and meaningful access to large sets of data.

Several vendors in our industry have developed some truly innovative visualisation for complex data sets. Thomson Aureka uses a ThemeScape visualisation to show relationships among patents in a landscape that looks like a tropical island, with the mountains representing clustered patents in technology areas. Innography uses interesting block diagrams to group assignees in groups. Spark-IP provides data in an innovative clustered display. Our software, Keyhole™, clusters related patents in a movable 3D display, allowing users to view relationships among patents from multiple angles.

Conclusion

Patent analytics is an essential tool for IP managers to uncover new business opportunities, find gaps in the company's R&D strategy and assess the company's exposure to competitive threats. Because of the sheer volume of patent data, good patent analytic tools are required to do the job. Patent analytics groups, clusters and finds relationships among patents, but it cannot tell you whether your patent is good or not. For that, you still have to use old-fashioned human judgement.

Matt Troyer is director of online strategy of TAEUS. He manages the development of TAEUS's online suite of patent management applications. He is a lifelong entrepreneur, having started three successful e-commerce companies in the 1990s (Firesale.com, Techforless.com and PhysiqueTransformation.com). Mr Troyer has a BA in biochemistry from the University of Colorado, Boulder.

Matt Troyer
Martketing Director
Email: mtroyer@taeus.com
Tel: +1 719 325 5000

TAEUS International Corporation
United States
www.taeus.com

Entrepreneur, engineer, instrument pilot and businessman, Art Nutter founded TAEUS International Corporation in 1992 as the world's first engineering company dedicated to helping patent owners make money from their patents.

Art Nutter
President
Email: art@taeus.com
Tel: +1 719 325 5000

TAEUS International Corporation
United States
www.taeus.com