Can big data fix a broken system for software patents?

By Derrick Harris

Legal scholars are always searching for ways to improve the U.S. patent system, sometimes via sweeping changes, but big data could help provide a technological fix to a big part of the problem.

The patent system is broken — on that almost everyone agrees. There's a backlog of applications that results in exorbitant wait times to get a patent issued, and the merit of patents that do get granted is often questionable. If you're forced to litigate a patent-infringement suit — an increasingly likely scenario – the costs can be crippling.

When it comes to software patents, the situation is particularly dire, which leads many critics arguing software patents should be abolished altogether. Patent trolls are a widely cited nuisance, but there's a more fundamental problem. Litigaton is expensive, but litigation is all too common because there are so many software patents out there and it can be very difficult — and very expensive — to find out whether a new invention possibly infringes on even one of them.

As we'll discuss in depth at our Structure: Data conference in New York later this month, techniques such as machine learning and natural-language processing are already having transformative effects in a number of fields. Why not the patent system, too?

Software patents don't scale …

Timothy B. Lee, a Cato Institute fellow (and frequent *Ars Technica* contributor), and Christina Mulligan of Yale's Information Society Project explore one big software-patent problem in a new research paper titled "Scaling the Patent System." The gist of Lee and Mulligan's argument is simple: software is such a wide-ranging and nebulous topic that it's nearly impossible to index software patents in a manner that would make it easier to search for them. The system just doesn't scale.

Current USPTO search engine

Property records are easily searchable because county recorders organize them in a logical manner based on geography. Even chemical patents, the authors point out, are relatively easy to search by chemical formula. With software patents, however, there's no such luck:

[I]n the absence of a precise, standardized scheme for classifying software inventions, patent applicants are free to use any terms they like — or even make up new ones — to describe their software inventions. The scope of a patent's claims will not always be obvious from a patent's title or abstract. And a single software patent can claim multiple applications that are only loosely connected to each other.

Lee and Mulligan's paper doesn't even touch on the problems that arise with *prior art*, generally defined as "all information that has been disclosed to the public in any form about an invention before a given date." It only compounds the issue of searching the USPTO database when attorneys or patent examiners are forced to search articles, presentations and anything else that might negate the novelty of a proposed invention.

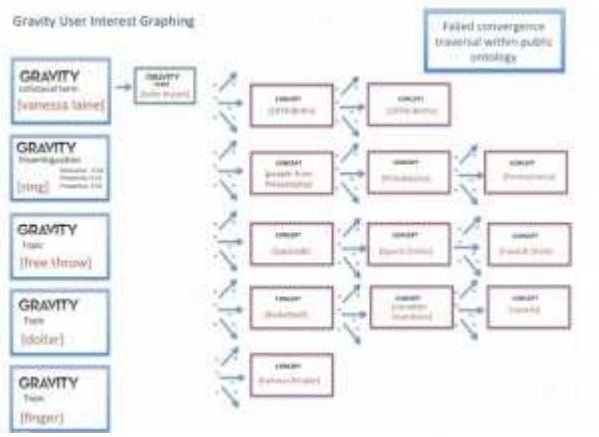
Unfortunately, the authors conclude, "Only dramatic reforms — such as excluding industries with high discovery costs from patent protection, establishing an independent invention defense, or eliminating injunctions — can return the patent system to its proper role of promoting innovation."

… but big data does

Looking outside the law, though, and into the world of big data analytics, one needn't look too hard to find some methods for making it easier to search for patents. The answer lies in semantics. If the problem is that keyword searches aren't effective, then build a search engine that addresses a wide variety of sources and that takes into account related terms based on how frequently they're linked, or based on the ontologies present in different industries.

- A startup called Apixio is already doing something similar in the field of medical records. It uses natural-language processing, machine learning and sematic association to make its Medical Information Navigation Engine (MINE) as easy to use as possible. Describing the service last April, I wrote that "when a doctor

types a patient's name and 'chest pain' into the search box, MINE is able to find ontological references to chest pain that bear little resemblance to the actual term."

- 

  Factually accurate, but irrelevant connections for Vanessa Laine

  Another method for doing this comes from Gravity, a startup that uses a hybrid man-machine process to personalize content for readers of sites such as the *Wall Street Journal* . Gravity's system is complex to say the least (here's a video tutorial that explains part of it) but the gist is that humans first serve as guides for machine-learning algorithms by determining connections between terms within large data sets, then the algorithms take over to complete the job faster than humans ever could. When they're done, the humans step in one more time to kill any bad connections between terms. The result is a system that can determine with high accuracy that a person tweeting about Vanessa Laine (Los Angeles Laker Kobe Bryant's ex-wife), for example, is probably more interested in basketball than about Laine's date of birth date or other accurate but irrelevant information.

- Even IBM's now-famous Watson question-answering machine could prove beneficial if the USPTO were to leverage its capabilities. The system has actually been suggested as an aid to help judges better interpret statutes against the Constitution, but loaded with patent data, it could help identify potential infringements and even answer with some certainty which ones might be the most relevant to any given application.

Indeed, a startup company called IP Street is already attempting to bring the benefits of semantic technology to bear on the patent field. By analyzing the entire library of patents issued by the USPTO, Founder and CEO Lewis Lee told me IP Street is able to extract meaning from patents using information from the patent claims. A succinct explanation on the company's web site explains that, "[The core] technology, known as LSI or latent semantic indexing, uses complicated mathematics and matrix decomposition (SVD) to identify similarities among documents. This allows you to enter an entire document (such as a product description, idea for a patent, etc.) and compare it to the universe of patents and patent applications—comparing across just the claims or the entire document."

Big data won’t solve all the complaints people have about patents, but it could make life a lot easier for the inventors, attorneys and examiners tasked with determining whether a patent infringes a previous patent, or is even patent-worthy in the first place. The question now is whether the USPTO wants to leave simplification of the process in the hands of private parties like IP Street, or if the agency wants to bring a few big data experts on board and improve what it’s able to offer those who rely on it.