Semantic Indexing of the Green Technology Patent Literature

George M. Garrity, Charles T. Parker and Catherine Lyons
NamesforLife, LLC

Under a Phase I/II STTR, NamesforLife, LLC created a suite of software tools and techniques to manage dynamic terminologies and an underlying term set (an up-to-date list of over 14,000 validly published names of bacteria and archaea, including all of the synonyms and homonyms, links to appropriate taxonomic literature, key genetic and genomic data). The company's N4L tools can automatically detect and tag bacterial names in HTML and XML documents with a high degree of precision. An interactive browser-based application (N4LGuide) provides end users direct access to correct nomenclatural information along with links to key data (16S sequence and genome sequence data) while reading the literature. It uses ISO standard Digital Object Identifier (DOI) technology to create links at each occurrence of a validly published name in HTML documents. The company has also developed batch tools (the N4L Semantic Tagger) that can embed N4L-DOIs into XML versions of scientific articles that are created as part of the contemporary publishing workflow and used to create human readable content in various forms (e.g., HTML, PDF, ink-on-paper). The company has also developed a unique way of tracking the occurrence of biological names in the literature, based on the usage of our tools (the N4L Contextual Index).

While initially intended as a tool for readers, authors, and publishers of scientific literature, N4L tools can also be applied to other documents where bacterial names appear. As proof of principle, the company processed approximately 250,000 US patents and patent applications with the Semantic Tagger and then indexed the tagged documents using Apache Lucene to provide end users with additional search and retrieval capabilities. Simple graphical tools were added to support limited on-demand analyses of search results. These tools are designed to support data mining by non-commercial organizations, highlighting trends in commercialization of biodiversity research. This work also led to the discovery of "terminological fingerprints" that could be used to classify patents and other documents using externally managed term sets.

To validate the concept of "terminological fingerprinting", the company processed the EPO Green Technology collection of patents, which consists of approximately 362,000 documents. In addition to detecting bacterial names, the N4L Semantic Tagger was modified to recover patent classifications (IPC and ECLA), applicants, assignees, inventors, patent references, patent metadata, and patent titles, as is common in patent landscaping.

A total of 3,845 patents were found that made reference to 3,385 distinct bacterial and archaeal names held in the NamesforLife database. Of these, 626 names were unique to non-US patents. The number of names per patent (name vectors) ranged from 1 – 1,290, with an average of 13 names and a median of 5 names. In addition to name occurrence, frequency data for each name occurrence per patent was tabulated. The resulting name vectors were then used to further examine the associations among the

patents based on the IPC and ECLA classifications. Simple associations could be derived directly from the captured data. However, more complex patterns involving multiple many-to-many relationships could only be ascertained from the cross-products of underlying contingency and frequency data.

The results were then examined using routine approaches for exploratory data analysis and visualization (e.g., principal components analysis, robust clustering, 2D scatter plots, 3D spin plots and heatmaps). Each of these methods revealed strong evidence of terminological fingerprints in the patents. However, those methods did not scale well or suffered from other limitations. Hexagonal binning was, however, found to be suitable for visualizing the complex relationships inherent in the patent data. The company is currently developing interactive hexagonal bin plots as a means of selecting subsets of patents that involve related technologies and microorganisms.

As DOE research on biofuels, bioremediation and carbon sequestration moves from the laboratory into production or commercial environments, a number of important policy and business decisions must be made that demand correct information. These include establishing the patentability of a given technology, freedom to operate, and potential infringement of patents held by competitors, both in the US and abroad. Failure to pay careful attention to these issues can have serious consequences beyond the payment of stiff penalties for infringement. These include lost opportunities arising for technology licensing, failure to detect and understand regional disparities, rapid growth in patent coverage of technologies by competitors and migration of technology across international borders. The scientific and technical literature provides an incomplete view of any field having commercial potential because the underlying technologies are typically not revealed in public until absolutely necessary, and then only after patent applications have been filed. While patents with corresponding papers are not uncommon as a means of announcing important new developments, they are not obligatory. Therefore, an awareness of developments in the field requires a thorough review of both bodies of literature. NamesforLife is building tools to simplify such searches, using its proven approach to indexing through the creation of persistent links to externally managed terminologies that common to both bodies of literature. This approach integrates well with existing commercial, academic and USPTO data mining capabilities.