averbis
text analytics

Offer of Presentation for the 2013 II-SDV Meeting Nice, 15 -16 April 2013

Large-Scale Patent Landscaping@Roche Diagnostics

Experiences and Lessons Learned

Manuel Dietrich, Roche Diagnostics GmbH
Markus Bundschus, Roche Diagnostics GmbH
Katrin Tomanek, Averbis GmbH
Philipp Daumke, Averbis GmbH

Monitoring of competitors, contractors, or certain products is a high demand in various industries. Analyzing patent portfolios is one strategy to complement this complex puzzle. Such an analysis comprises the detailed assessment of large patent corpora to specific technology fields. Assessments such as technology fields, however, go often beyond the typical patent meta-data such as International Patent Classifications (IPCs) or other classification schemes. Furthermore, the technology fields to be analyzed may be taken from a company internal thesaurus with possibly large number of concepts as well as hierarchical relations between them. Still, the classification of a considerable fraction of patent portfolios according to such application-specific thesauri is a non-trivial task that requires a significant amount of human expert man power. Therefore, there is a pressing need for decision support systems that assist human experts in this complex annotation task and at the same time learn from previous expert annotations to speed up the landscaping process.

The here presented contribution describes a large-scale patent landscaping initiative within Roche Diagnostics. Besides standard patent landscaping requirements and tasks the project stakeholders (i.e. strategic decisions makers and IP professionals) prioritized the following requirements:

- Analyze patents from a huge fraction of different companies ranging from small companies to conglomerates with a highly diverse patent portfolio.
- Classify the patent portfolios with respect to a company internal terminology covering more than sixty hierarchical classes.
- Substantially reduce and for some categories even eliminate the manual expert annotation effort.
- Support of different patent data forms and formats.
- Ability to use interfaces to in-house visualization tools.

Since the company internal terminology in the here described initiative is very detailed and the diversity of companies is very high, manually crafted automatic rules or simple keyword search for automatic assignment to technology fields could not reach the classification accuracy required by the strategic decision makers. Thus, machine learning techniques were identified as being a high demand. Computers learn from the experience of patent professionals and produce useful predictions on new, unseen examples after being trained on a learning data set. Automatically categorizing large data sets of patents with a high number of (hierarchical) categories while still opting for excellent prediction quality requires a sufficient number of learning data. The concept of active learning minimizes the effort of manual creation of such data by intelligent data sampling and iterative supervised learning. Another demand of patent professionals is the possibility to bring in custom requirements regarding design, ergonomics and functionality into a patent classification system. Roche finally decided to cooperate with Averbis, a market leader on text mining, search technologies and machine learning approaches with a special focus on Life Science.

The Averbis Patent Classification Suite is a new system co-developed with Roche's IP professionals to assist them in patent landscape analyses. From a set of training examples (i.e., patents manually assigned to the relevant categories), which reflects the patent professional's or the company-internal policy of category assignment, the system learns to automatically assign patents in the same manner as the user would do. A key-feature of the software is the support of efficient training data creation. On the one hand, faceted and semantic search support retrieval of possibly relevant documents. Automatic pre-annotation can also be used to point patent professionals to interesting documents. The user can then accept or review the system's category suggestion. And finally, the system also provides intelligent data sampling techniques such as Active-

Learning for rapid model improvement guiding the user to documents which, when manually assigned to categories, will result in large gains of classification performance.

The interface has been designed to Roche needs to provide efficient access to large collections of patents through faceted browsing, semantic full-text search, and navigation based on application-specific terminologies. Another focus of usability is on rapid, manual category assignment to patents. For selected patents, a preview consisting of their abstract, customer-defined key fields (such as patent assignees, IPCs, etc.), and also additional data (such as images, original patent PDF, etc.) is shown.

Under the hood the system hosts a fully-fledged machine learning library optimized for patent classification. Users can define both flat and hierarchical category systems which the system learns to assign automatically to new patents. The system also supports multi-label classification – a common need in patent retrieval scenarios where documents may fall into multiple categories and assignment to one distinct category may be impossible. The classification module itself is highly configurable with respect to the learning algorithm, the features (which sections of the patent to be used in the classification process as well as automatic feature selection techniques), and the method how the classifier's predictions are translated into the user-defined hierarchy of categories.

In a pilot project, the performance of the machine learning library was prototypically tested on different patent data with different category systems. For the "medical indications" branch (having about dozen subcategories), a classification performance of about 90% accurracy was reached, for another category branch ("instruments", also about a dozen subcategories), the performance was only slightly lower with 89% accurracy. Experiments proved that the amount of training data available is crucial for accurate automatic classification. In a first setting with only about 1/5th of the training data, the overall system performance was about 20 percentage points lower (around 70% accuracy), especially because of significant class imbalance and sparse data problems. This agains emphasizes the need for a system such as the Averbis Patent Classification Suite which supports annotation of training data, especially in scenarios like patent classification where class imbalance and data sparsity is often a result of complex category systems.

Experiments run during the pilot project also revealed, that the features (i.e., parts of the patent to be considered) best-suited for automatic classification are not necessarily the same ones as humans would identify as most relevant for manual classification. While, for example, property agents indicated the abstracts, claims and Derwent titles to be extremely relevant, these features proved suboptimal for the machine learning system. In contrast, best performance was achieved on the original title and IPC classes (pruned to the first three levels)!

Detailed benchmark analyses are currently run to measure the performance of patent classification with machine learning in contrast to manual annotation tasks. This includes an evaluation about how much manual time can be saved for the patent expert with this classification task.

Short Biography of the proposed speakers

Manuel Dietrich (Roche Diagnostics GmbH)

Manuel works at Scientific & Business Information Services within Roche where he is responsible for the development of innovative applications for the integration and management of scientific information to derive potential new insights for biomedical researchers and decision makers. Hereby he also focuses on machine learning, visualization and text mining algorithms. Manuel received his Master's Degree in Bioinformatics from the Center for Bioinformatics at the University of Tübingen. From 2006 to 2008 he worked with Insilico Biotechnology AG in the field of systems biology where he focused on the development of algorithms for flux and pathways analysis in genome-scale metabolic networks. End of 2008 Manuel joined Roche Applied Science, where he developed an Internet portal with innovative pathway visualization components for customized pre-plated real-time qPCR assays.

Katrin Tomanek (Averbis)

From 2006 to 2010 Katrin Tomanek worked as research assistant at the computational linguistics lab of the University of Jena. In 2010, she received her PhD in computer science / artificial intelligence from the University of Dortmund. Her thesis entitled "Resource-Aware Annotation through Active Learning" investigates approaches to intelligent data sampling with the goal to decrease time needed to produce accurate classification models. Since 2011, Katrin works with Averbis GmbH where she is responsible for the development of systems for automatic classification and keywording. She is also working in the PETRUS project with the German National Library where the Averbis Extraction Platform is applied for machine-based cataloguing processes.