A multidimensional approach to visualising and analysing patent portfolios

Edwin Horlings

e.horlings@rathenau.nl
Rathenau Instituut (Netherlands)

Introduction

Innovation is a social process. Patents are an important source of information on invention – a vital part of innovation – which is one of the reasons why patent databases are harvested. Patterns extracted from the metadata in patent databases reveal emergent outcomes of the complex process of innovation. Mapping, visualising and statistically analysing patent portfolios is a powerful tool for technology forecasting (e.g. Kim, Suh & Park, 2008; Daim et al. 2006; Ernst, 1997), university-industry knowledge transfer (e.g. Gurney et al., 2014), and other applications in science and innovation studies. However, most social scientists lack the technical expertise to use patent data.

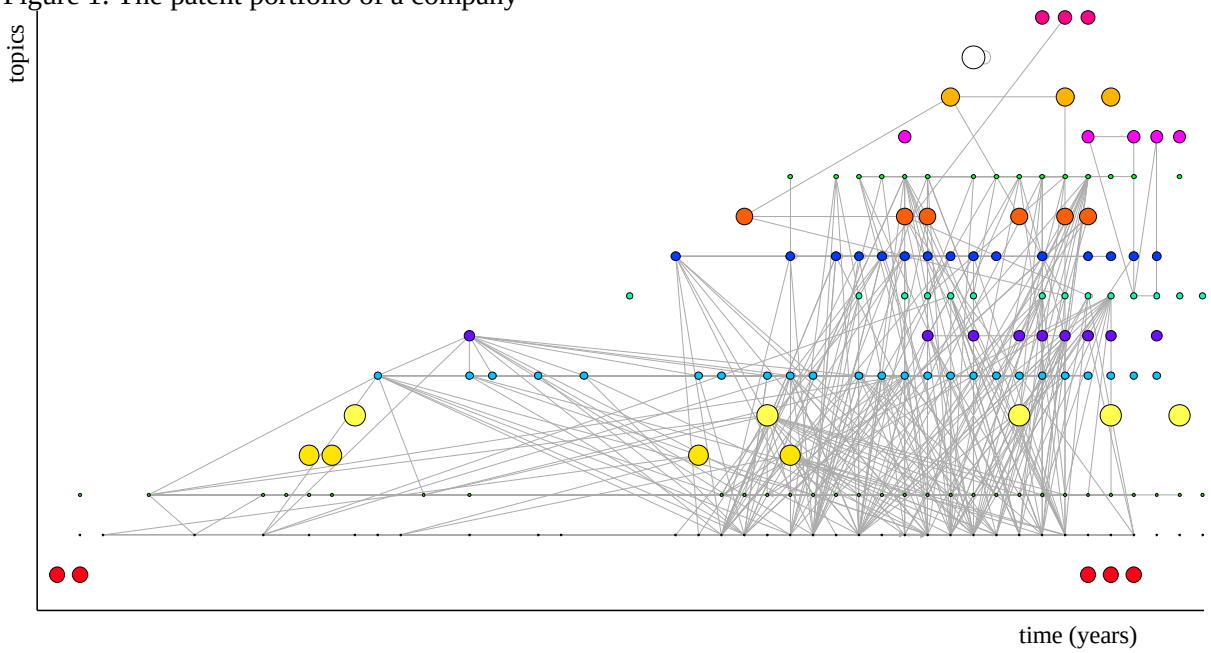
There is a vast reservoir of tools and techniques for patent analysis and visualisation. Applications include citation networks and co-inventor/co-assignee networks (Garfield 1979; Sternitzke et al. 2008; Gress 2010); topic maps (Huang et al., 2003; Boyack & Klavans, 2008; Leydesdorff, 2008; Kay et al., 2013); innovation geography (Leydesdorff & Bornmann, 2012); and links between inventors in patents and authors of scientific publications (Cassiman 2007; Gurney et al., 2014). Existing methods for visualising and analysing patent portfolios do, however, have a number of shortcomings. They tend to focus on one dimension at a time (e.g. citation networks or topic maps). They rarely provide a longitudinal perspective, other than on the volume of patenting activity. Most importantly, there is a lack of information on the quality or value of patents (Lissoni, 2012). Last but not least, there is a need for transparent and shared procedures and source codes, available to all researchers of science and innovation.

In this paper, we develop a new method for visualising and analysing patent portfolios that combines multiple dimensions, specifically those of time, citation, topical similarity, diversity, and quality. The aim is to develop a general-purpose data infrastructure that is open to the science and innovation studies community. All sql queries as well as a manual for generating and using the data infrastructure will be made available upon publication on the expectation that users will reciprocate by also making available their improvements and adjustments to the community.

Methods and data

The data infrastructure is based on PATSTAT (April 2013 release). The first step was to generate an version aggregated at the level of INPADOC patent families. This includes quality indicators proposed by the OECD (Squicciarini et al. 2013). The second step was to develop queries for mining the aggregated PATSTAT database that can be applied to any entry dataset to produce statistical tables containing data for each individual patent family, aggregated statistics per topic and year, and files that can be used to produce visualisations in Gephi (Bastian, Heymann, & Jacomy, 2009). By first aggregating the entire PATSTAT database (step 1), we gain substantial time savings with every partial analysis (step 2).
The visualisation is based on a method developed by Horlings and Gurney (2013) for mapping the portfolios of scientists. It has been applied to inventor-author relations in Gurney et al. (2014). The figure presents an example of a visualisation of the patent portfolio of a company. Similar visualisations can be made for countries, IPC classes, technologies, and so on.

Figure 1: The patent portfolio of a company



The maps that can be produced with our data infrastructure are highly versatile. They are available with one node for each individual patent families or with nodes as clusters of patent families per topic per year. Edges are citation relations within the portfolio or similarities among patent families in the set. In the above figure the size of nodes is scaled to the number of patent families per node, while colours indicate the topic. Topics are identified by calculating similarity among patent families in the dataset in terms of IPC code co-occurrence and then applying the Infomap clustering algorithm (Rosvall & Bergstrom, 2008) to the resulting similarity matrix, using the SAINT Toolkit (Somers et al. 2009). Size and colours can be adjusted to other indicators, such as radicalness (a quality indicator) or the presence of a patent application in a particular country or region (e.g. WIPO or China). Thus, the visualisation can simultaneously capture the dimensions of time, citation, topical similarity, and quality. In addition, the data infrastructure includes queries to expand the entry dataset to all patents that cite the patent families in the entry set (forward citations) and all patents that are cited by those patents (backward citations).

Applications

The data infrastructure can serve many purposes, such as the analysis of technological search dynamics (the scope and diversity of activity and the nature and value of subsequent patents); technological forecasting, tracing the growth rate and quality of specific topics; comparing cohorts such as large and small firms or universities and PROs, for example to assess patent quality indicators; and so on and so forth. In the full paper, we will present three examples of applications, namely the portfolio of Apple (following up on Jun & Park, 2013), the portfolio of a university and the patents that cite its patents; and forecasting the development of an emerging technology (bone grafting). The purpose of the three examples is to examine the potential uses of the statistical information and visualisations that can be produced. By making the data infrastructure available to the community, we hope to encourage broader use of patent analysis for understanding innovation.