Scientometric analysis of the emerging technology landscape

Ian I'Anson
Defence Science and Technology Laboratory, Fort Halstead, TN14 7BP

Abstract: For researchers and decision makers in any technical domain, understanding the state of their area of interest is of critical importance. This 'landscape' of emerging technologies is constantly evolving, and the sheer scale of research publication output in the modern era makes qualitative review increasingly difficult. Scientometric analysis is a valuable tool for the quantitative analysis of research output, and is employed by the Defence Science and Technology Laboratory (Dstl) Knowledge and Information Services in support of our research activities, for applications including identifying opportunities for academic collaboration, and technology watching/forecasting to identify emerging technologies and opportunities that may have implications for UK Defence.
This paper provides an overview of our approach to conducting scientometric analysis of research papers and patent submissions. The methods for extracting and disambiguating publications are described, and the qualitative inferences we seek to make, along with some of the associated limitations and potential pitfalls are also discussed.

Keywords: Scientometrics, technology watching, forecasting

1. Introduction

The volume of academic and patent literature produced annually has expanded greatly over the past few decades. Where once it was possible to aspire to understand the full breadth of human knowledge, it is now increasingly difficult even to maintain an overview of developments within one's own field. As early as 1892, Karl Pearson wrote "*Scarcely any specialist of today is really master of all the work which has been done in his own comparatively small field… It is as if individual workers in both Europe and America were bringing their stones to*

Received: 24.2.2016 Accepted: 21.3.2016 ISSN 2241-1925

one great building and piling them on cementing them together without regard to any general plan or to their individual neighbor's work." [Pearson (1897)]

Qualitative review of the literature- reading every new paper relevant to a technical domain has in most cases become impracticably time consuming, and without a good overview of one's domain strategic planning and the optimum use of finite resources becomes impossible.

Scientometrics- the statistical, quantitative analysis of scientific information offers a partial solution. Trends in research and commercial interest in specific topics can be evaluated based on publication volumes and their origins, versus time, allowing the overall 'landscape'- the distribution of research and development priorities in time and space, to be evaluated.

Scientometric methodology can be used to make inferences such as the imminent emergence of new technologies; the technological domains of interest to an individual, an institution or a country; or the best prospects for research collaboration, based on rankings of relevant output by a person or an institution. Scientometrics can also help to identify relationships between technologies and the key publications and authors relevant to a specific technology.

This paper provides an overview of Dstl's approach to conducting scientometric analysis of research papers and patent submissions, and some of the qualitative inferences that we seek to draw from the quantitative data. Like any other statistical technique, scientometric analysis is open to misinterpretation- the limitations of our methods are also discussed.

2. Methods and example applications

2.1 Obtaining and pre-processing data

For scientometric analysis to be meaningful it must be carried out using a representative corpus of data. The Web of Science (http://wok.mimas.ac.uk) provided by Thomson Reuters comprises several databases, containing bibliographic records of much of the world's academic and patent literature, including authors, titles and abstracts. Retrieving the corpus of records that is specific to a topic of interest is a matter of choosing the most appropriate databases for the topic and devising a search strategy that achieves an acceptable level of Precision and Recall for the topic. The Web of Science search engine permits free text word/phrase matching with Boolean logic, along with various refinement tools. Records can be exported from the databases as plain text.

To effectively analyse large sets of plain text data requires specialist software. Dstl employs the 'VantagePoint' data mining tool from Search Technology Inc. (http://www.searchtech.com/). This allows plain text data to be imported and specific fields extracted from it using bespoke filters. It also offers fuzzy logic disambiguation tools to programmatically identify variations on the names of

entities- for example recognising that 'Oxford University' and 'Uni. Of Oxford' are equivalent terms.

2.2 Analysing data

Once a corpus of records has been imported and pre-processed, a number of analyses can be carried out:

A time series showing number of new records per year can be generated in order to identify trends in output for a topic. Further sub-dividing these records by originating country (Figure 1), institution, or over-arching topic/application (Figure 2) may permit further comparison:

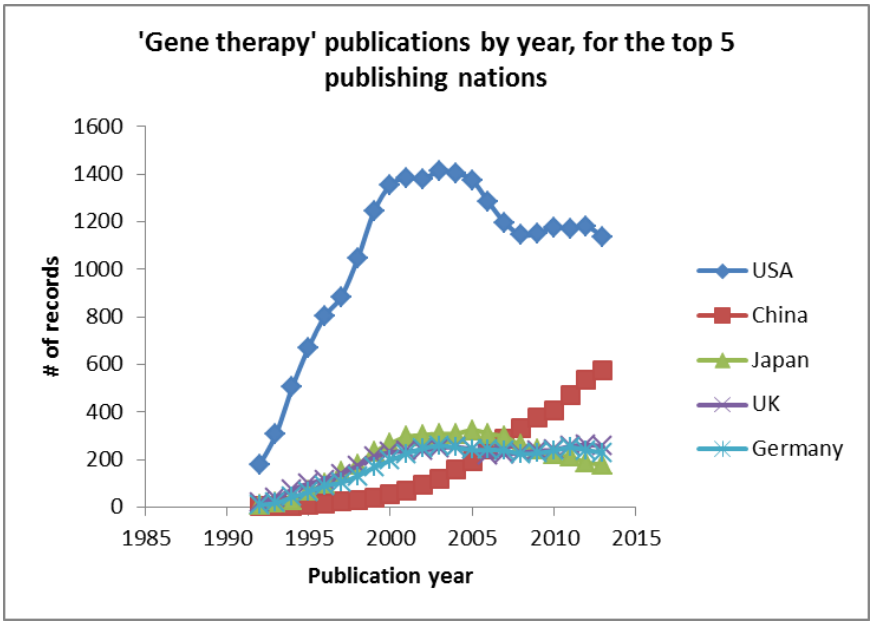

Figure 1 – Time series of publications relating to 'Gene therapy' for the top 5 publishing nations (from Web of Science Core Collection). (3 point moving-average)

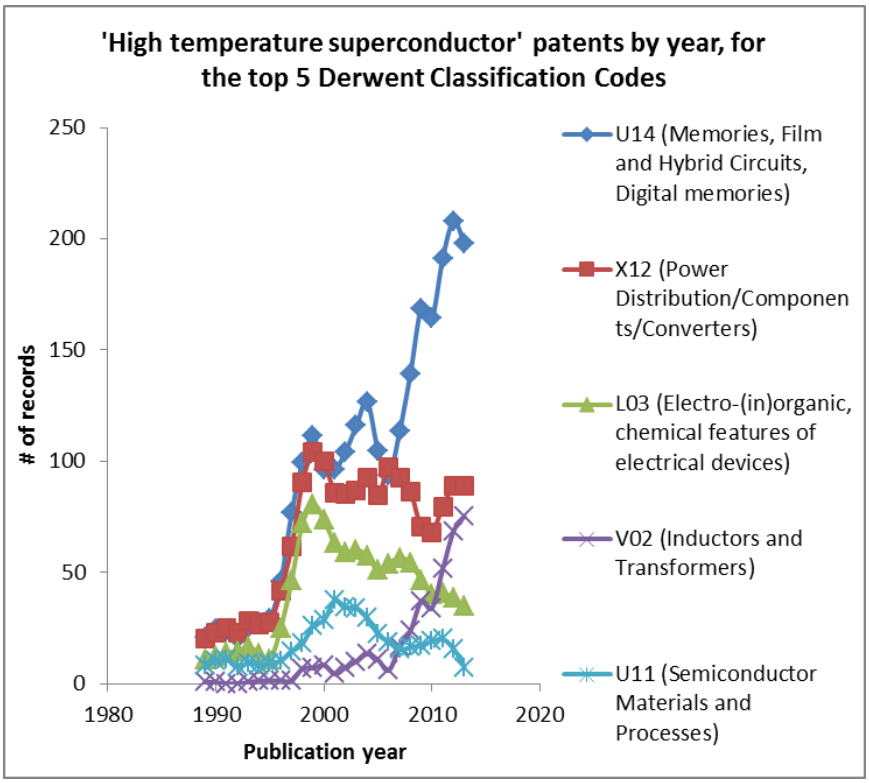

Figure 2 – Time series of priority patents relating to 'High temperature superconductors' within the top 5 Derwent Classification Codes (from Derwent Innovation Index).(3 point moving-average).

These time series trends can be used to make inferences, with varying levels of confidence. A sudden proliferation of research or patent output pertaining to a topic is suggestive of a significant development- either in the topic of interest or in a related topic on which the topic of interest depends, but further investigation is necessary in order to identify the nature of the aforementioned development.

For instance, research output pertaining to High Temperature Superconductors (HTSCs) exhibited sudden, rapid growth beginning in 1987, see figure 3:

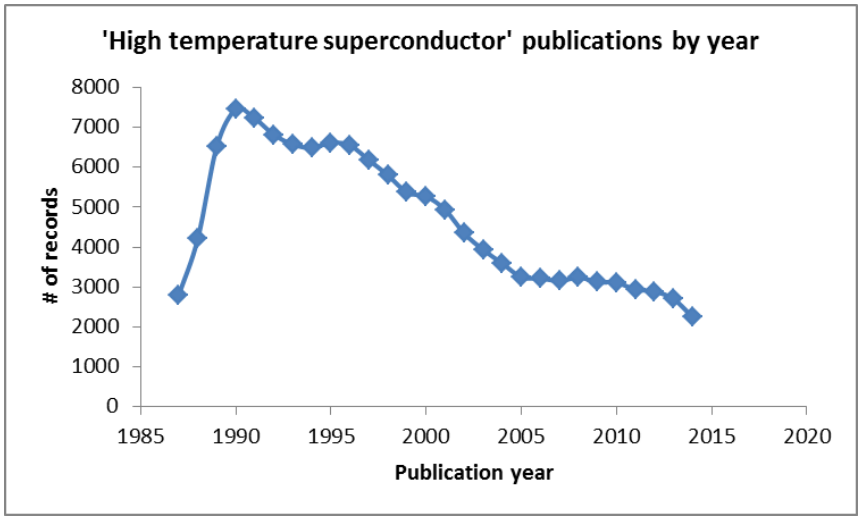

Figure 3 – Time series of publications relating to 'High temperature superconductors' (from Web of Science Core Collection and Inspec). (3 point moving-average).(3 point moving-average).

The sudden growth in research output is easily explained- High Temperature Superconductors were not known to exist prior to a discovery by Müller and Bednorz, in 1986, (for which they were awarded the 1987 Nobel Prize for Physics). [Bednorz and Müller (1986)]
Similarly, research output pertaining to Internet Telephony ('Voice Over IP') also exhibited fairly rapid growth from ~1998 until 2008:

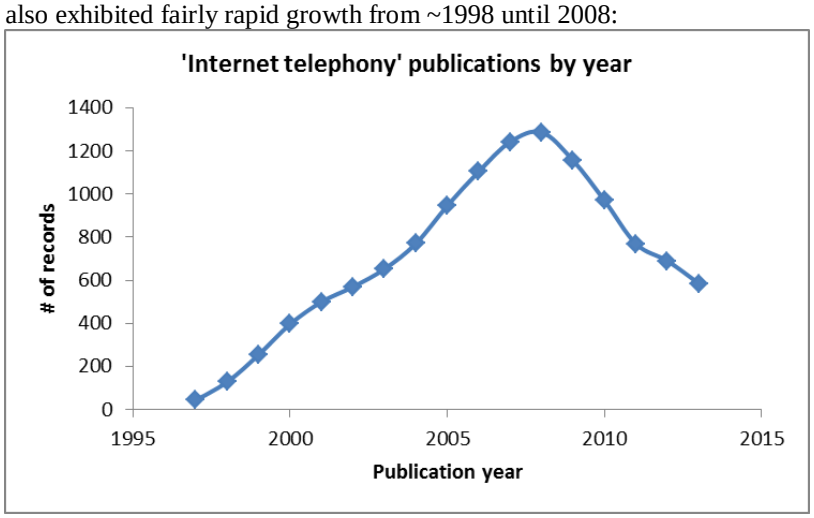

Figure 4 – Time series of publications relating to 'Internet Telephony' (from Web of Science Core Collection and Inspec). (3 point moving-average).(3 point moving-average).

In this case however, the trend does not imply a fundamental discovery- rather, Internet Telephony lacked mass market appeal until generally available internet bandwidth was suitably developed- the demand, and therefore the rationale for Research and Development (R&D) effort, developed along with the capacity of broadband networks. [Van Lente (2013)]

The observed trends indicate that 'something' happened to bring them about, and may also indicate at approximately what point in time, but they do not permit any inference as to the nature of the 'something'.

It can also be instructive to directly compare trends in patent versus academic literature output for a given topic. For instance, the trend in HTSC research shown in Figure 3 appears indicative of a field in slow decline, but the growth in patents over a similar period (Figure 2) suggests that the technology is increasingly finding practical applications. The nature of these applications can be inferred from the Derwent Classification codes.

2.3 Technology forecasting

An obvious use one might attempt to make of an apparent trend in a time series is to forecast future developments by extrapolation. This should be treated with extreme caution however, since short-to-medium term trends are subject to external influences that cannot necessarily be inferred from what came before.

For example, 'Gene therapy' patent submissions showed steady or accelerating growth from ~1990 to 2002. Following the death of Jesse Gelsinger in 1999 [Van Lente (2013)], and the resulting loss of investor confidence, an abrupt collapse in interest in the field was observed. Crucially, the observed trend from 1990 to 2002 could not be used to predict the subsequent trend from 2002 onwards.

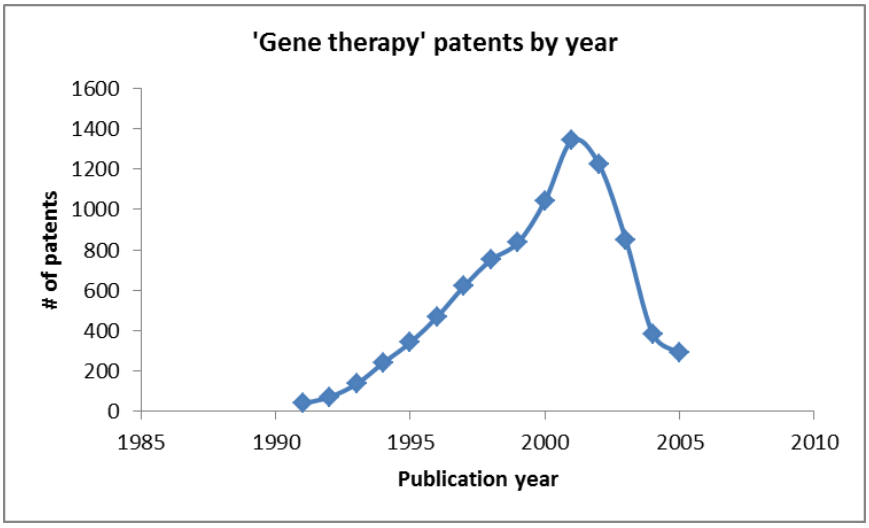

Figure 5– Time series of priority patents relating to 'Gene therapy' (from Derwent Innovations Index). (3 point moving-average).

Technology forecasting is a complicated undertaking, success in which requires a detailed understanding of the nature of the technology of interest, the drivers behind its development and the dependencies on which continued development rests. Trends up to the present time are not reliable indicators of future developments, but can contribute to developing the background understanding of a technological domain that is necessary in order to make informed predictions.

3. Limitations and caveats

Scientometric analysis features numerous pitfalls, both in mathematically determining the significance of trends and in determining their real implications.

3.1 Interpreting trends: Sources of quantitative error

There are a number of factors that limit the accuracy of scientometric analyses of the type discussed above.

The initial acquisition of a corpus of records representative of one's chosen topic is often challenging- when relying on text matching via a search engine there is always a trade-off between Precision (the proportion of all records retrieved *that are relevant* to the topic) and Recall (the proportion of all relevant records *that exist* that have been retrieved by the search). Since perfect Precision *and* Recall is not usually achievable in practice, any corpus of records analysed must strictly be considered to be a sub-set of the totality of relevant material that exists, *plus* a certain proportion of material that is not relevant to

the topic. The extent to which the Precision-Recall trade-off can be optimised for any given topic is largely contingent on the researcher's skill.

Unpublished data is unavoidably excluded from consideration, and one could argue that this is especially problematic in the military domain, where security considerations may keep relevant research from being made public. To put this problem in context however, according to the Battelle Memorial Institute, overall aerospace/defence/security R&D spending for 2014 accounted for some US$26.4 billion, out of a total worldwide spend of some US$1.6 trillion, or ~1.65% (over half of which is contributed by the United States). [Batelle Memorial Institute, 2013] The statistical significance of undisclosed military R&D is therefore unlikely to be problematic for most topics.

Pre-processing the data creates further scope for error- fuzzy logic disambiguation is not perfect and import filters extract fields from plain text using Regular Expressions, which are also fallible in this context.

The inherent sources of error discussed thus far can be controlled sufficiently so as not to undermine the credibility of subsequent statistical analysis. However, statistical trends observed in scientometric data must also be considered in the context of external factors, and require an appreciation of the limits of statistical significance.

3.2 Interpreting trends: Statistical significance of apparent trends

When evaluating an apparent trend in publication output, it is necessary to consider the circumstance in which it occurs. For example, annual academic research output across all topics worldwide has grown substantially and fairly consistently for decades, see figure 6:

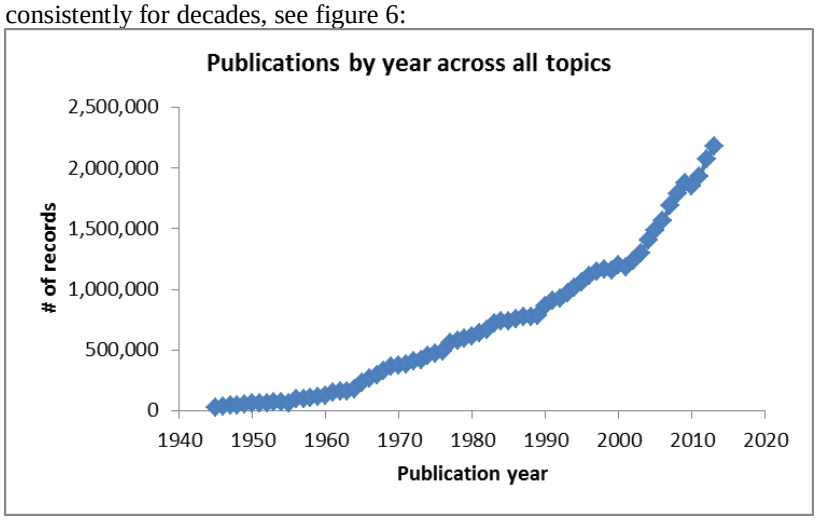

Figure 6 – Time series of all academic publications from Web of Science Core Collection

Therefore the mere fact that research output within a specific topic, or originating from a specific country appears to have increased year-on-year is not necessarily indicative of an exceptional level of interest in that topic.
For the same reason an apparent 'flatline' in research output *may* be better interpreted as a real-terms decline. It is safer and arguably more useful to infer meaning from *relative* trends between different countries or regions.

Furthermore, when investigating a relatively obscure topic or a sub-set of data that otherwise yields only a small number of records, the statistical significance of any apparent trend must be considered carefully, as 'noise' begins to dominate when the publications for individual years become sparse.

3.3 Interpreting trends: Wider context, external factors

Even when a trend is sufficiently pronounced that its significance can be inferred confidently, the implications of the trend still need to be understood if it is to have value. For this reason, scientometric analysis reporting is best accompanied by an assessment by appropriate Subject Matter Experts (SMEs).

The previously discussed example topics: Room Temperature Superconductors, Gene Therapy and Internet Telephony illustrate this- there are reasonable explanations for the observed trends in research and patent output in every case, but identifying them and their implications requires a reasonable background knowledge of the topic.

A further confounding element is the tendency for research and development efforts to be driven by 'Hype': An initial period of intense interest in a topic, followed by a pronounced decline in interest, as expectations give way to realistic assessments of development potential. [Gartner (2016)]

Intense and growing interest in a topic can be detected by scientometric analysis, but determining whether this is motivated by 'Hype' or by something more tangible is not possible by Scientometric analysis alone- again the SME's assessment is vital.

4. Conclusions

Scientometric analysis of the type discussed here is a useful tool to aid qualitative assessment of a technical domain. It can be used to indicate the overall vitality of a field, the maturity of technologies, key events, publications and players in order to aid Subject Matter Experts in targeting their finite resources, in order to best understand and influence the state-of-the-art. It can also be used as an aid to planning R&D activities. Identifying potential research collaborations is an obvious example, but an overview of the research landscape also permits the maturity of specific developments in a field to be identified, along with gaps that may benefit from targeted R&D investment.

The value of scientometric analysis in isolation is limited, since apparent trends offer only a basis for making general inferences- they are at best indicative, and never represent sufficient information to draw definitive conclusions. Considerable care is needed even when making general inferences concerning trends that seem superficially apparent from time series visualisations, (this caveat is applicable to statistical analysis in general).

As a complement to the qualitative assessment of research output- gained by actually reading and understanding the relevant publications- and other statistical measures such as citation metrics- scientometric analysis can make a significant contribution to overcoming the increasingly common problem of 'information overload', in order to achieve a level of current-awareness that might not otherwise be possible for a reasonable commitment of time and effort.

References

Pearson, K., (1892). *The Grammar of Science*, Dover Publications, Mineola, New York.

Bednorz, J.G. and Müller, K.A. (1986). Possible high Tc superconductivity in the Ba−La−Cu−O system, *Zeitschrift für Physik B Condensed Matter*, Vol 64, Issue 2, 189-193.

Van Lente, H., Spitters, C. and Peine, A. (2013). Comparing technological hype cycles: Towards a theory, *Technological Forecasting and Social Change*, Vol 80, Issue: 8, 1615-1628.

Battelle Memorial Institute: 2014 Global R&D Funding Forecast, (2013). URL: https://www.battelle.org/docs/tpp/2014_global_rd_funding_forecast.pdf [Date accessed: 29th Jan 2016]

Gartner: Gartner Hype Cycle, 2016. URL: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp [Date accessed: 29th Jan 2016]