

MIT Open Access Articles

A framework for technology forecasting and visualization

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Woon, Wei Lee, Andreas Henschel, and Stuart Madnick. "A Framework for Technology Forecasting and Visualization." IEEE, 2009. 155–159. © Copyright 2009 IEEE

As Published: <http://dx.doi.org/10.1109/IIT.2009.5413768>

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: <http://hdl.handle.net/1721.1/71808>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



A Framework for Technology Forecasting and Visualization

Wei Lee Woon

Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE.

Andreas Henschel

Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE.

Stuart Madnick

Massachusetts Institute of Technology
77 Mass. Ave., Building E53-321
Cambridge, MA 02139-4307, U.S.A.

Abstract

This paper presents a novel framework for supporting the development of well-informed research policies and plans. The proposed methodology is based on the use of bibliometrics; i.e., analysis is conducted using information regarding trends and patterns of publication. While using bibliometric techniques in this way is not a new idea, the proposed approach extends previous studies in a number of important ways. Firstly, instead of being purely exploratory, the focus of our research has been on developing techniques for detecting technologies that are in the early growth phase, characterized by a rapid increase in the number of relevant publications. Secondly, to increase the reliability of the forecasting effort, we propose the use of automatically generated keyword taxonomies, allowing the growth potentials of subordinate technologies to be aggregated into the overall potential of larger technology categories. A proof-of-concept implementation of each component of the framework is presented, and is used to study the domain of renewable energy technologies. Results from this analysis are presented and discussed.

1 Introduction

For decision makers and researchers working in a technical domain, understanding the state of their area of interest is of the highest importance. Any given research field is composed of many subfields and underlying technologies which are related in intricate ways. At the same time, information regarding past and current research is available from a variety of channels, providing a rich source of data with which an effective research strategy may be formed. The challenge is to find ways of effectively mining these sources for useful patterns and trends.

There is already a significant body of related research, and for a good review, the reader is referred to [14, 11, 12]. Interesting examples include visualizing interrelationships between research topics [13, 15], identification of important researchers or research groups [9, 10], the study of research performance by country [6, 8], the study of collaboration patterns [3, 2] and the analysis of future trends and developments [5, 15].

In particular, our research has addressed the challenge of *technology forecasting*, on which this paper is focussed. In contrast to the large body of work already present in the literature, there is currently very little research which attempts to provide concrete, actionable results on which researchers and other stakeholders can base their actions.

In response to this apparent shortcoming, we describe a novel framework for automatically visualizing and predicting the future evolution of domains of research. Our framework incorporates the following three key contributions:

1. A methodology for automatically creating taxonomies from bibliometric data.
2. A set of numerical indicators for identifying technologies of interest. In particular, we have developed a set of simple growth indicators which can be applied to thousands of candidate technologies at a time.
3. A novel approach for using the taxonomies to incorporate semantic distance information into the technology forecasting process.

2 A framework for technology forecasting

It is important to define the form of forecasting that is intended. In particular, it must be stressed that it is not “forecasting” in the sense of a weather forecast, where specific future outcomes are intended to be predicted with a

reasonably high degree of certainty. It is also worth noting that certain tasks remain better suited to human experts; in particular, where a technology of interest has already been identified or is well known, we believe that a traditional review of the literature and of the technical merits of the technology would prove superior to an automated approach.

Instead, the proposed framework targets the preliminary stages of the research planning exercise by focussing on what computational approaches excel at: i.e. scanning and digesting large collections of data, detecting promising but less obvious trends and bringing these to the attention of a human expert. This overall goal should be borne in mind as, in the following subsections, we present and describe the individual components which constitute the framework.

2.1 Overview

Figure 1 depicts the high-level organization of the system. As can be seen, the aim is to build a comprehensive technology analysis tool which will collect data, extract relevant terms and statistics, calculate growth indicators and finally integrating these with the keyword taxonomies to produce actionable outcomes. To facilitate discussion, the system has been divided into three segments:

1. Data collection and term extraction (labelled **(a)** in the

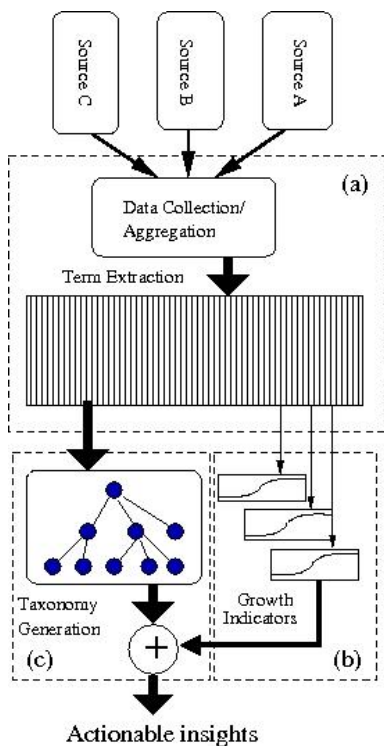


Figure 1. Proposed framework

figure)

2. Prevalence estimation and calculation of growth indicators (labelled **(b)**)
3. Taxonomy generation and integration with growth indicators (labelled **(c)**)

These components are explained in the following three subsections.

2.2 Data collection and term extraction

2.2.1 Data collection

The type of data source, collection mechanism and number of sources used can be modified as required but for the proof-of-concept implementation, information extracted from the Scopus¹ database was used. Scopus is a subscription-based, professionally curated citations database provided by Elsevier. Other possibilities, such as Google's scholar search engine and ISI's Web of Science database were also considered and tested but Scopus proved to be a good initial choice as it returned results which were generally of a high quality, both in terms of the publications covered and relevance to search terms, and was normally able to retrieve a reasonable number of documents.

2.2.2 Term extraction

Term extraction is the process of automatically generating a list of keywords on which the technology forecasting efforts will be focussed. Again, there are a variety of ways in which this can be achieved; we have experimented with a number of these and our experiences have been thoroughly documented in [18]. For the present demonstration the following simple but effective technique is used: for each document retrieved, a set of relevant keywords is provided. These are collected and, after word-stemming and removal of punctuation marks, sorted according to number of occurrences. For the results shown later, a total of 500 keywords were extracted and used to build the taxonomy.

2.2.3 Pilot study

To provide a suitable example on which to conduct our experiments and to anchor our discussions, a pilot study was conducted in the field of renewable energy. The incredible diversity of renewable energy research offers a rich and challenging problem domain on which we can test our methods. Besides high-profile topics like solar cells and nuclear energy, renewable energy related research is also being conducted in fields like molecular genetics and nanotechnology.

¹<http://www.scopus.com>

To collect the data for use in this pilot study, high-level keywords related to renewable energy were submitted to Scopus and the abstracts of the retrieved documents were collected and used. In total, 119,393 abstracts were retrieved and stored.

2.3 Identification of early growth technologies

There are two steps to this activity. The first is to find a measure for the “prevalence” of a given technology as a function of time. In the context of an abstract database, an appropriate measure which seemed to give reasonable results was the *term frequency*, defined as:

$$TF_i = \frac{n_i}{\sum_{j \in \mathcal{I}} n_j} \quad (1)$$

where, n_i is the number of occurrences of keywords i , and \mathcal{I} is the set of terms appearing in all article abstracts (this statistic is calculated for each year of publication to obtain a time-indexed value). Once the term frequencies for all terms have been extracted and saved, they can be used to calculate growth indicators for each of the keywords (and, by extension, the associated technologies). These, in turn, are used to rank the list of terms.

As stated previously, we are most interested in keywords with term frequencies that are relatively low at present but that have been rapidly increasing; this will be referred to as the “early growth” phase of technological development and represents the fields to which an expert would wish to be alerted. Existing techniques are often based on fitting growth curves (see [1] for example) to the data. This can be difficult as the curve-fitting operation can be very sensitive to noise. Also, data collected over a relatively large number of years (approximately ≥ 10 years) is required, whereas the emergence of novel technological trends can occur over much shorter time-scales.

The search for suitable early growth indicators is currently an area of active research but for this paper the following indicator will serve as an example:

$$\theta_i = \frac{\sum_{t \in [2004, 2008]} t \cdot TF_i[t]}{\sum_{t \in [2004, 2008]} TF_i[t]}, \quad (2)$$

where, θ_i is the growth potential for keyword i and $TF_i[t]$ is the term frequency for term i and year t . As can be seen, this gives the average publication year for articles appearing in the last five year (excluding 2009), and which are relevant to term i (a more recent year indicates greater currency of the topic).

2.4 Keyword taxonomies and semantics enriched indicators

One of the problems encountered in earlier experiments involving technology forecasting is that there is a lot of noise when measuring technology prevalence using simple term occurrence frequencies.

This is a fundamental problem when attempting to infer an underlying property (in this case, the size of the relevant body of literature) using indirect measurements (hit counts generated using a simple keyword search), and cannot be entirely eliminated. However, as part of our framework we propose an approach through which these effects may be reduced; the basic idea is that hit counts associated with a single search term will invariably be noisy as the contexts in which this term appear will be extremely diverse and will contain a large number of extraneous mentions (and will also include papers which are critical of the technology it represents). However, if we can find collections of related terms and use aggregate statistics instead of working with individual terms, we might reasonably expect that a lot of this randomness will cancel out.

We concretize this intuition in the form of a *predictive taxonomy*; i.e. a hierarchical organization of keywords relevant to a particular domain of research, where the growth indicators of terms lower down in the taxonomy contribute to the overall growth potential of higher-up “concepts” or categories.

2.4.1 Taxonomy generation

The question remains, how do we obtain such a taxonomy? In a limited number of cases, these taxonomies may be available from external sources such as government agencies and other manually curated sources. However, in many cases, a suitable taxonomy is either unavailable, or is available but is not sufficiently updated to be of use for the application at hand. As such, to make our framework broadly applicable, an important research direction is the *automated* creation of keyword taxonomies based on the statistics of term occurrences.

The basic idea is to group together terms which tend to co-occur frequently. We have tested a number of ways of doing this [17, 16] but it is not possible in the present scope to discuss these here. Instead, we present one particular method which was found to produce reasonable results while being scalable to large collections of keywords. This is based on the algorithm described in [7] where each keyword or tag is associated with a vector that contains the annotation frequencies for all documents, and which is then comparable, for e.g. by using the cosine similarity measure. We adapt the algorithm to general taxonomy creation by substituting the cosine similarity function with the *asym-*

metric distance function proposed in [17] (this is based on the “Google distance” proposed in [4]):

$$\overrightarrow{\text{NGD}}(t_x, t_y) = \frac{\log n_y - \log n_{x,y}}{\log N - \log n_x}, \quad (3)$$

where t_x and t_y are the two terms being considered, and n_x , n_y and $n_{x,y}$ are the occurrence counts for the two terms occurring individually, then together in the same document respectively.

The algorithm consists of two stages: the first is to create a similarity graph of keywords, from which a measure of “centrality” is derived for each node. Next, the taxonomy is grown by inserting the keywords in order of decreasing centrality. In this order, each unassigned node, t_i , is attached to one of the existing nodes t_j such that:

$$j = \arg \min_{j \in \mathcal{T}} \overrightarrow{\text{NGD}}(t_i, t_j), \quad (4)$$

(where \mathcal{T} is the set of terms which have already been incorporated into the taxonomy.)

2.4.2 Enhanced early growth indicators

Once the keyword taxonomies have been constructed, they provide a straightforward method of enhancing the early growth indicators using information regarding the co-occurrence statistics of keywords within the document corpus. This is done by re-calculating the early growth scores for each keyword based on the aggregate scores of the nodes contained in the subtree descended from this keyword.

We might conceive of a number of possible aggregation schemes, but for the results presented in this paper, a straight average was used.

2.5 Implementation details

The framework described here was implemented in the Python programming language. Data collection was semi-automatic where *.csv files were downloaded manually from the Scopus website, while automated scripts had been created to filter and subsequently store the records in a local SQL database (we used the SQLite database system).

3 Results and discussions

We present results for a simple pilot study in renewable energy. As described in section 2.2.1, the Scopus database was used to collect a total of 500 relevant keywords and 119,393 document abstracts. These keywords were organized into a taxonomy (section 2.4.1), and the average publication year for each keyword was calculated as shown in equation (2). Finally, these scores were aggregated using

1. cytology	16. zea mays
2. nonmetal	17. bioelectric energy
3. semiconducting zinc	sources
compounds	18. phosphorus
4. alga	19. biological materials
5. hydraulic machinery	20. cellulose
6. hydraulic motor	21. nitrogenation
7. bioreactors	22. bacteria
8. concentration process	(microorganisms)
9. metabolism	23. adsorption
10. sugars	24. soil
11. computer networks	25. hydrolysis
12. experimental studies	26. glycerol
13. ecosystems	27. fermenter
14. direct energy	28. glucose
conversion	29. potential energy
15. lignin	30. biodegradable

Table 1. Top 30 topics by average publication year

the taxonomy and were sorted according to decreasing publication year. The top 30 keywords are listed in table 1. Some quick observations:

1. From the number of related terms, it can clearly be seen that biological aspects of renewable energy are amongst the most rapidly growing areas of research.
2. Looking at the terms with relatively large associated subtrees, we see that three of the largest in the top 30 were “biological materials” (15 nodes), “fermenter” (7 nodes) and “hydrolysis” (4 nodes). The subtrees for the first two terms are shown in figures 2 and 3 respectively, while the hydrolysis subtree is actually part of the “fermenter” subtree and as such is not displayed.
3. The fermenter subtree is devoted to biofuel related technologies (two major sub-categories are represented - “glucose”-related or 1st generation biofuels, and “cellulosic” biofuels (2nd generation fuels).
4. While less focussed, the biological materials subtree still highlights the importance of biology. The “soil” branch focusses on ecological issues, while the “chemical reaction” branch relates to gasification research.
5. Highly-rated non-biological terms included “non-metal” (#2) and “semiconducting zinc compounds” (#3); both are related to thin-film photovoltaics.
6. However, the top-30 list contained a large number of keywords associated with leaves in the taxonomy, so the confidence in the scores were lower.

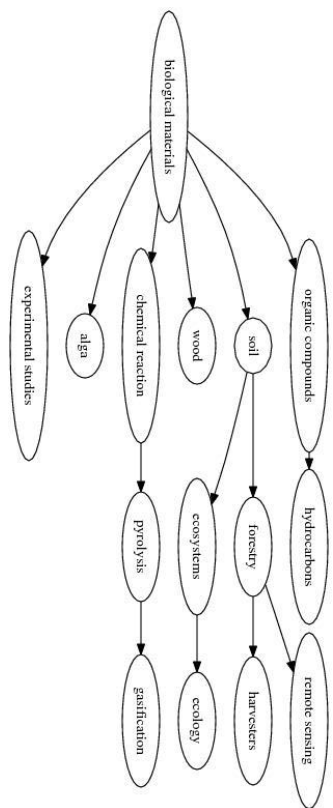


Figure 2. Subtree for “Biological materials”

4 Conclusion

A novel framework for facilitating research planning and decision-making has been presented. The proposed system covers the entire chain of activities including the collection of data, the extraction of keywords of interest and the calculation of semantically-enhanced “early growth indicators”. In addition, a proof-of-concept implementation of this framework is described and is applied to the renewable energy domain. Results of this study have been presented, and are already quite encouraging; however the process is still a little too noisy to pick out “very early growth” technologies. However, we are investigating numerous avenues for enhancing the current implementation, and are confident of presenting improved findings in upcoming publications.

References

- [1] M. Bengisu and R. Nekhili. Forecasting emerging technologies with the aid of science and technology databases. *Technol Forecast Soc*, 73(7):835–844, September 2006.
- [2] T. Braun, A. P. Schubert, and R. N. Kostoff. Growth and trends of fullerene research as reflected in its journal literature. *Chem Rev*, 100(1):23–38, 2000.

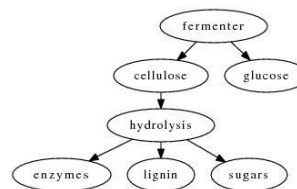


Figure 3. Subtree for node “fermenter”

- [3] W.-T. Chiu and Y.-S. Ho. Bibliometric analysis of tsunami research. *Scientometrics*, 73(1):3–17, October 2007.
- [4] R. L. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE T Knowl Data En*, 19(3):370–383, 2007.
- [5] T. U. Daim, G. Rueda, H. Martin, and P. Gerdtsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technol Forecast Soc*, 73(8):981–1012, 2006.
- [6] de Miranda, G. M. Coelho, Dos, and L. F. Filho. Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technol Forecast Soc*, 73(8):1013–1027, 2006.
- [7] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, #2006-10. <http://dbpubs.stanford.edu:8090/pub/2006-10>, 2006.
- [8] Kim and Mee-Jean. A bibliometric analysis of the effectiveness of koreas biotechnology stimulation plans, with a comparison with four other asian nations. *Scientometrics*, 72(3):371–388, September 2007.
- [9] R. N. Kostoff. Text mining using database tomography and bibliometrics: A review. 68:223–253, November 2001.
- [10] P. Losiewicz, D. Oard, and R. Kostoff. Textual data mining to support science and technology management. *J Intell Inf Syst*, 15(2):99–119, 2000.
- [11] J. Martino. *Technological Forecasting for Decision Making*. McGraw-Hill Engineering and Technology Management Series, 1993.
- [12] J. P. Martino. A review of selected recent advances in technological forecasting. *Technol Forecast Soc*, 70(8):719–733, October 2003.
- [13] A. Porter. Tech mining. *Compet Intel Mag*, 8(1):30–36, 2005.
- [14] A. Porter, A. Roper, T. Mason, F. Rossini, and J. Banks. *Forecasting and Management of Technology*. Wiley-Interscience, New York, 1991.
- [15] H. Small. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610, December 2006.
- [16] W. Woon and S. Madnick. Semantic distances for technology landscape visualization. Technical Report CISEL #2008-04, MIT, <http://web.mit.edu/smadnick/www/wp/2008-04.pdf>, 2008.
- [17] W. Woon and S. Madnick. Asymmetric information distances for automated taxonomy construction. *Knowl Inf Syst*, Online first, 2009.
- [18] B. Ziegler, A. Firat, C. Li, S. Madnick, and W. Woon. Preliminary report on early growth technology analysis. Technical Report CISEL #2009-04, MIT, <http://web.mit.edu/smadnick/www/wp/2009-04.pdf>, 2009.