A Framework for Technology Forecasting and Visualization

Wei Lee Woon
Andreas Henschel
Stuart Madnick

Working Paper CISL# 2009-11

September 2009

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

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. Information thus obtained is analyzed to predict probable future developments in the technological fields being studied. While using bibliometric techniques to study science and technology 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 aggregated into the overall potential of larger technology categories. As a demonstration, 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. This composition, or research landscape, is not static as new technologies are constantly developed while existing ones become obsolete, often over very short periods of time. Fields that are presently unrelated may one day become dependent on each others findings.

Information regarding past and current research is available from a variety of channels, providing both a difficult challenge as well as a rich source of possibilities. On the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data with which a well-informed and comprehensive research strategy may be formed.

There is already a significant body of related research, and for a good review, the reader is referred to [16, 13, 14]. Interesting examples include visualizing interrelationships between research topics [15, 18], identification of important researchers or research groups [11, 12], the study of research performance by country [8, 10], the study of collaboration patterns [1, 4, 3] and the analysis of future trends and developments [17, 7, 6, 18].

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. A number of approaches have been tested where the basic principle is to assign terms that co-occur frequently to common subtrees of the taxonomy.

2. A set of numerical indicators for identifying technologies of interest. In particular, we are interested in de-

veloping a set of simple growth indicators, similar to technical indicators used in finance, which may be easily calculated but which can be applied to hundreds or thousands of candidate technologies at a time. This is in contrast to more traditional curve fitting techniques which require relatively larger quantities of data.

3. A novel approach for using the taxonomies to incorporate semantic distance information into the technology forecasting process. The individual growth indicators are quite noisy but by aggregating growth indicators from semantically related terms spurious components in the data can be averaged out.

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)
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 Scopus[1] 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

[1] http://www.scopus.com

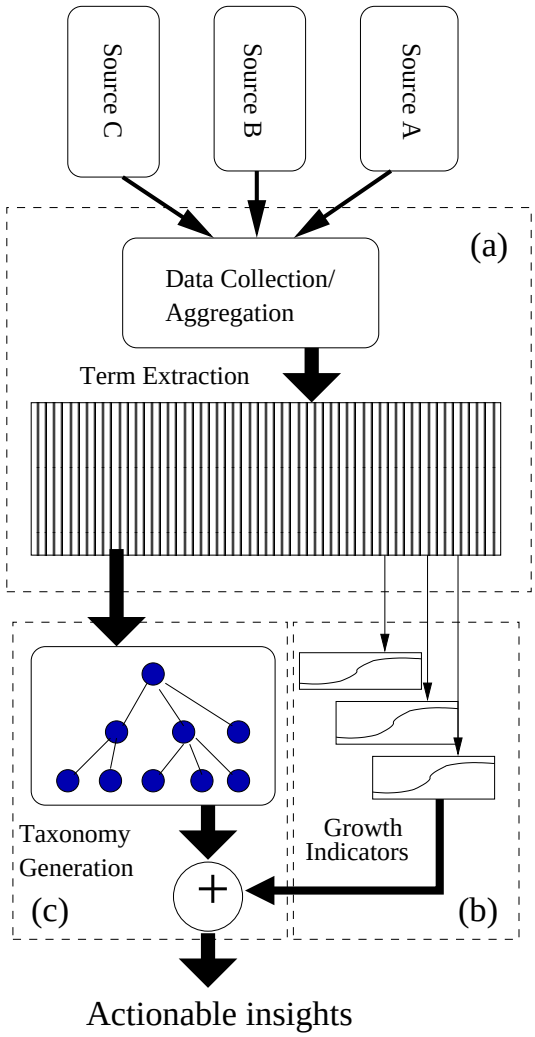

Figure 1. Proposed framework: overall structure

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 [21]. 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 in the text. For the example results shown later in this paper, a total of 500 keywords have been 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.

To collect the data for use in this pilot study, a variety of high-level keywords related to renewable energy (please see Appendix A) were submitted to Scopus, and the abstracts of the retrieved documents were collected and used. In total, $119,393$ abstracts were retrieved and subsequently ordered by year of publication.

2.3 Identification of early growth technologies

There are actually two steps to this activity. The first is to find a suitable measure for the "prevalence" of a given technology as a function of time. In terms of a database of academic publications, this would be some means of measuring the size of the body of relevant publications appearing each year. It is difficult to achieve this directly but an alternative would be to search for the occurrence statistics of terms relevant to the domain of interest. To allow for the overall growth in publication numbers over time (given the emergence of new journals, conferences, etc.), we choose to use the *term frequency* instead of the raw occurrence counts.

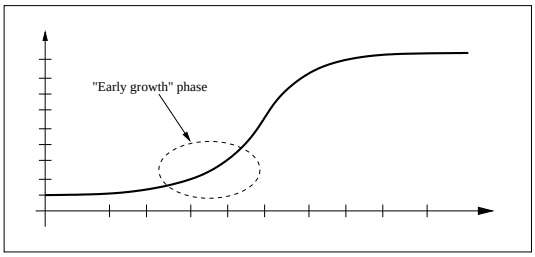

Figure 2. Early growth of technological development

This is defined as:

$$TF_i = \frac{n_i}{\sum_{j \in \mathcal{I}} n_j} \tag{1}$$

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, depicted in figure 2, and represents the fields to which an expert would wish to be alerted. Existing techniques are often based on fitting growth curves (see [2] for example) to the data. This can be difficult as the curve-fitting operation can be very senstive 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:

$$\theta_i = \frac{\sum_{t \in [2004,2008]} t.TF_i[t]}{\sum_{t \in [2004,2008]} TF_i[t]}, \tag{2}$$

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, as indicated in section 1 is to group together terms which tend to co-occur frequently. Again, we have tested a number of different ways of achieving this (two earlier attempts are described in [20, 19]) but it is not possible in the present scope to discuss and compare these in depth. 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 [9] which was originally intended for social networks where users annotate documents or images with keywords. 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 using two important modifications; firstly, instead of using the cosine similarity function, the *asymmetric* distance function proposed in [20] is used (this is based on the "Google distance" proposed in [5]):

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

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. Note that the above expression is "asymmetric" in that $\overrightarrow{\mathrm{NGD}}(t_x, t_y)$ refers to the associated cost if t_x is classified as a subclass of t_y, while $\overrightarrow{\mathrm{NGD}}(t_y, t_x)$, corresponds to the inverse relationship between the terms.

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:

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

(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. As with almost all aspects of the proposed framework, a number of variants are possible but the basic idea is to recalculate the early growth scores for each keyword based on the aggregate scores of each of the keywords contained in the subtree descended from the corresponding node in the taxonomy.

For the results presented in this paper, the aggregation operation used was a straight average, though other more elaborate schemes are clearly possible.

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 keywords which were relevant to the renewable energy domain, along with $119,393$ document abstracts. These keywords were then used to construct a taxonomy as described in section 2.4.1, and the average publication year for each keyword was calculated as

shown in equation (2). Finally, these were aggregated using the keyword taxonomy and the list of keywords was sorted according to order of decreasing publication year. Using this method of evaluation, the top 30 keywords were (numbers in brackets are the taxonomy-aggregated average publication years):

1. cytology (2007.31)
2. nonmetal (2007.24)
3. semiconducting zinc compounds (2007.19)
4. alga (2006.94)
5. hydraulic machinery (2006.91)
6. hydraulic motor (2006.91)
7. bioreactors (2006.81)
8. concentration process (2006.77)
9. metabolism (2006.73)
10. sugars (2006.69)
11. computer networks (2006.66)
12. experimental studies (2006.63)
13. ecosystems (2006.58)
14. direct energy conversion (2006.57)
15. lignin (2006.56)
16. zea mays (2006.56)
17. bioelectric energy sources (2006.56)
18. phosphorus (2006.55)
19. biological materials (2006.53)
20. cellulose (2006.52)
21. nitrogenation (2006.52)
22. bacteria (microorganisms) (2006.52)
23. adsorption (2006.52)
24. soil (2006.52)
25. hydrolysis (2006.51)
26. glycerol (2006.51)
27. fermenter (2006.51)
28. glucose (2006.50)
29. potential energy (2006.50)
30. biodegradable (2006.43)

Some quick observations:

1. One of the most striking observations is the number of biotechnology related keywords in this list. This indicates that biological aspects of renewable energy are amongst the most rapidly growing areas of research.

2. Amongst the highly-rated non-biological terms on the list were “nonmetal” (#2) and “seminconducting zinc compounds” (#3), both of which are related to the field of thin-film photovoltaics.

3. However, the top-30 list contained a large number of keywords which were actually associated with leaves in the taxonomy, so the confidence in the scores were lower.

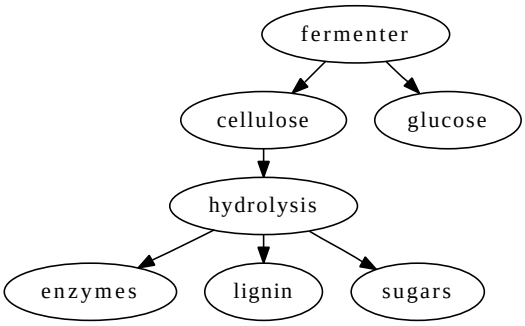

Figure 4. Subtree for node “fermenter”

4. 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 3 and 4 respectively, while the hydrolysis subtree is actually part of the “fermenter” subtree and as such is not displayed.

5. It can be seen that the fermenter subtree is clearly devoted to biofuel related technologies (in fact, two major categories of these technologies are represented - “glucose”-related or first generation biofuels, and “cellulosic” biofuels which are second generation fuels.

6. The biological materials subtree is less focussed but it does emphasize the importance of biology to renewable energy research. The “soil” branch of this subtree is devoted to ecological issues, while the “chemical reaction” branch is associated with gasification (waste-to-energy, etc.) research.

4 Conclusion

In this paper, a novel framework for facilitating research planning and decision-making has been presented. The proposed system covers the entire chain of activities starting with the collection of data from generic information sources (online or otherwise), the extraction of keywords of interest from these sources and finally the calculation of semantically-enhanced “early growth indicators”.

In addition, a simple proof-of-concept implementation of this framework is described and is applied to the domain of renewable energy. Results of this study are presented and discussed, and are already quite encouraging though currently the process is still a little too noisy to pick out “very early growth” technologies. However, we are investigating numerous avenues for enhancing the basic implementation referenced here, and are confident of presenting improved findings in upcoming publications.

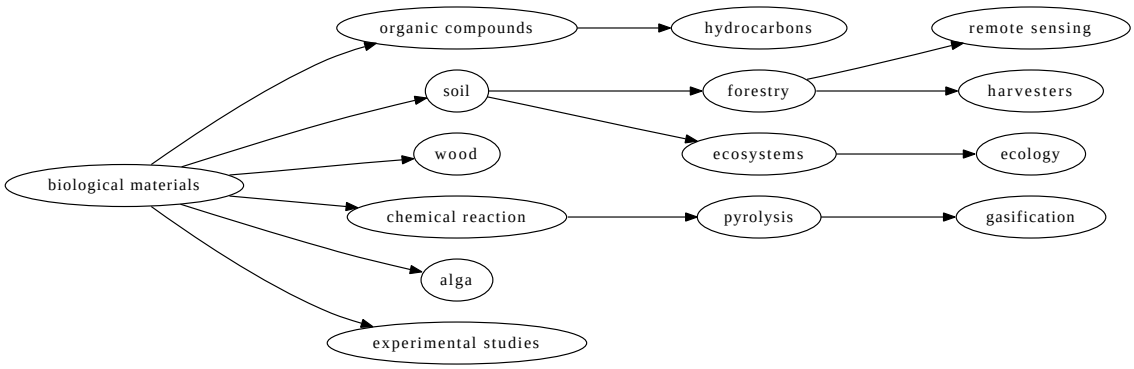

Figure 3. Subtree for node "Biological materials"

A Seed terms for data collection

The search terms used for extraction of renewable energy related abstracts and keywords from Scopus were: *"renewable energy", "biodiesel", "biofuel", "photovoltaic", "solar cell", "distributed generation", "dispersed generation", "distributed resources", "embedded generation", "decentralized generation", "decentralized energy", "distributed energy", "on-site generation", "geothermal", "wind power", "wind energy"*.

References

[1] Anuradha, K., Urs, and Shalini. Bibliometric indicators of indian research collaboration patterns: A correspondence analysis. *Scientometrics*, 71(2):179–189, May 2007.

[2] M. Bengisu and R. Nekhili. Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change*, 73(7):835–844, September 2006.

[3] T. Braun, A. P. Schubert, and R. N. Kostoff. Growth and trends of fullerene research as reflected in its journal literature. *Chemical Reviews*, 100(1):23–38, 2000.

[4] W.-T. Chiu and Y.-S. Ho. Bibliometric analysis of tsunami research. *Scientometrics*, 73(1):3–17, October 2007.

[5] R. L. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE T Knowl Data En*, 19(3):370–383, 2007.

[6] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981–1012, October 2006.

[7] T. U. Daim, G. R. Rueda, and H. T. Martin. Technology forecasting using bibliometric analysis and system dynamics. In *Technology Management: A Unifying Discipline for Melting the Boundaries*, pages 112–122, 2005.

[8] de Miranda, G. M. Coelho, Dos, and L. F. Filho. Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technological Forecasting and Social Change*, 73(8):1013–1027, October 2006.

[9] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, Technical report 2006-10. http://dbpubs. stanford. edu: 8090/pub/2006-10, 2006.

[10] 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.

[11] R. N. Kostoff. Text mining using database tomography and bibliometrics: A review. 68:223–253, November 2001.

[12] P. Losiewicz, D. Oard, and R. Kostoff. Textual data mining to support science and technology management. *Journal of Intelligent Information Systems*, 15(2):99–119, 2000.

[13] J. Martino. *Technological Forecasting for Decision Making*. McGraw-Hill Engineering and Technology Management Series, 1993.

[14] J. P. Martino. A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change*, 70(8):719–733, October 2003.

[15] A. Porter. Tech mining. *Competitive Intelligence Magazine*, 8(1):30–36, 2005.

[16] A. Porter, A. Roper, T. Mason, F. Rossini, and J. Banks. *Forecasting and Management of Technology*. Wiley-Interscience, New York, 1991.

[17] N. R. Smalheiser. Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Technovation*, 21(10):689–693, October 2001.

[18] H. Small. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610, December 2006.

[19] W. Woon and S. Madnick. Semantic distances for technology landscape visualization. Technical Report CISL #2008-04, Massachusetts Institute of Technology, http://web.mit.edu/smadnick/www/wp/2008-04.pdf, 2008.

[20] W. Woon and S. Madnick. Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, Online first, 2009.

[21] B. Ziegler, A. Firat, C. Li, S. Madnick, and W. Woon. Preliminary report on early growth technology analysis. Technical Report CISL #2009-04, Massachusetts Institute of Technology, http://web.mit.edu/smadnick/www/wp/2009-04.pdf, 2009.