# Forecasting and Visualization of Renewable Energy Technologies using Keyword Taxonomies

Wei Lee Woon[1], Zeyar Aung[1], Stuart Madnick[2]

[1]Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE.
[2]Massachusetts Institute of Technology
77 Mass. Ave.,Building E53-321
Cambridge, MA 02139-4307, U.S.A.

**Abstract.** Interest in renewable energy has grown rapidly, driven by widely held concerns about energy sustainbility and security. At present, no single mode of renewable energy generation dominates and consideration tends to center on finding optimal combinations of different energy sources and generation technologies. In this context, it is very important that decision makers, investors and other stakeholders are able to keep up to date with the latest developments, comparative advantages and future prospects of the relevant technologies. This paper discusses the application of bibliometrics techniques for forecasting and integrating renewable energy technologies. Bibliometrics is the analysis of textual data, in this case scientific publications, using the statistics and trends in the text rather than the actual content. The proposed framework focuses on a number of important capabilities. Firstly, we are particularly interested in the detection of technologies that are in the *early growth* phase, characterized by rapid increases in the number of relevant publications. Secondly, there is a strong emphasis on visualization rather than just the generation of ranked lists of the various technologies. This is done via the use of automatically generated keyword taxonomies, which increase reliability by allowing the growth potentials of subordinate technologies to be aggregated into the overall potential of larger categories. Finally, by combining the keyword taxonomies with a colour-coding scheme, we obtain a very useful method for visualizing the technology "landscape", allowing for rapidly evolving branches of technology to be easily detected and studied.

## 1 Introduction

### 1.1 Motivation

The generation and integration of Renewable Energy is the subject of an increasing amount of research. This trend is largely driven by widely held concerns about the energy sustainability and security and climate change. However, the relevant technical issues are extremely diverse and cover the entire gamut of challenges ranging from the extraction and/or generation of the energy, integration of the energy with existing grid infrastructure and the coordination of energy generation and load profiles via appropriate demand response strategies. For decision makers, investors and other stakeholders,

the sheer number and variety of the relevant technologies can be overwhelming. In addition this is an area which is evolving rapidly and a huge effort is required just to stay abreast with current development.

All research fields are invariably 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.

### 1.2 Theoretical background

There is already a significant body of research on the topic of technology forecasting, planning and bibliometrics. An in-depth review is beyond the scope of this article but the interested reader is referred to [1–4].

In terms of the methodologies employed, interesting examples include visualizing interrelationships between research topics [5, 6], identification of important researchers or research groups [7, 8], the study of research performance by country [9, 10], the study of collaboration patterns [11–13] and the analysis of future trends and developments [14–16, 6]. It is also noteworthy that bibliometric techniques have been deployed on a wide array of research domains, including ones which are related to renewable energies. Some examples include thin film solar cells [17], distributed generation [18], hydrogen energy and fuel cell technology [19, 20] and many others.

Our own research efforts have centered on the challenge of *technology forecasting* [21, 22], on which this paper is focussed. However, in contrast to the large body of work already present in the literature as indicated above, there is currently very little research which attempts to combine the elements of technology forecasting and visualization.

In response to this apparent shortcoming, in [23] we described a novel framework for automatically visualizing and predicting the future evolution of domains of research. Our framework incorporated the following three key characteristics:

1. A system for automatically creating taxonomies from bibliometric data. We have attempted a number of approaches for achieving this but the basic principle is to create a hierarchical representation of keyword representations where terms that co-occur frequently with one another are assigned to common subtrees of the taxonomy.
2. A set of numerical indicators for identifying technologies of interest. In particular, we are interested in developing a set of simple growth indicators, similar to technical indicators used in finance. These growth indicators are specially chosen to be easily calculated so that they can be readily applied to hundreds or thousands of candidate technologies at a time. In contrast, traditional curve fitting techniques are more complex and tend to incorporate certain assumptions about the shape in which the growth curve of a technology should take. In addition, more complex growh models require relatively larger quantities of data to properly fit.

3. A technique whereby the afore-mentioned taxonomies can be combined with the growth indicators to incorporate semantic distance information into the technology forecasting process. This is an important step as the individual growth indicators are quite noisy. However, by aggregating growth indicators from semantically related terms spurious components in the data can be averaged out.

In this paper we present further investigations into the use and effectiveness of this framework, particularly in terms of the growth indicators used as well as a more intuitive method of visualizing the scores corresponding to each technology.

## 2 Analytical framework

We now describe the framework which will be used to conduct the technology forecasting. However, it is important to first define the form of forecasting that is intended in the present context. It should be pointed out 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 framework proposed in [21] 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.

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.

### 2.1 Data collection and term extraction

This consists of the the following two stages:

– **Data collection** The exact collection mechanism, type and number of data sources used are all design parameters that be modified based on user requirements and

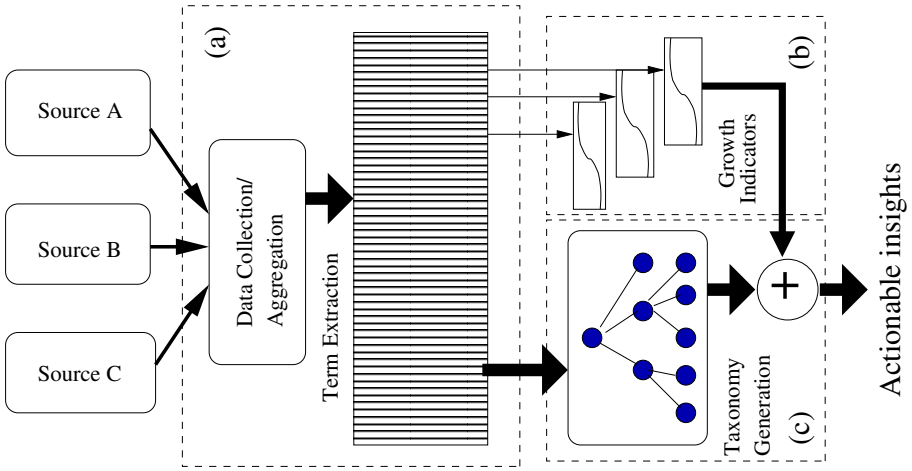


**Fig. 1.** Analytical framework[23]

available resource. However, for the implementation presented in this paper this information was extracted from the Scopus[1] publication database. Scopus is a subscription-based, professionally curated citations database provided by Elsevier. For the results described in this paper, a total of 119,393 document abstracts were collected and processed for subsequent analysis.
Other possible sources of bibliometrics data that were considered include Google's scholar search engine and ISI's Web of Science database. However, Scopus proved to be a good initial choice as it returned results which were of a generally high quality both in terms of the publications covered and relevance to search terms. In addition, the coverage was sufficiently broad such that searches submitted to Scopus were normally able to retrieve a reasonable number of documents.

– **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 [24]. 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 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 within the particular context being looked at. In the

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

context of an academic publications database, this would refer to the size of the body of relevant publications appearing each year which in turn would serve as an indicator of the amount of attention that the technology in question receives from the academic community.

For a variety of reasons achieving this directly is not straightforward but a workable alternative would be to search for the occurrence statistics of terms relevant to the domain of interest. To allow for changes (mainly growth) in overall publication numbers over time, the *term frequency* is used instead of the raw occurrence counts. 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 hence the associated technologies.

As stated previously, we are most interested in keywords with term frequencies that are relatively low at present but that have been rapidly increasing, which will henceforth be referred to as the "early growth" phase of technological development. Focusing on this stage of technological development is particularly important because we believe that it represents the fields to which an expert would most wish to be alerted since he or she would most likely already be aware of more established research areas while techologies with little observable growth can be deemed to be of lower potentil.

Existing techniques are often based on fitting growth curves (see [25] 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 $\geq 10$ 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 an ongoing area of research but for this paper we consider the following two indicators as illustrative examples:

$$\eta_i = \frac{[TF_i[y_2] + TF_i[(y_2 + 1)]]}{[TF_i[y_1] + TF_i[(y_1 + 1)]]} \tag{2}$$

$$\theta_i = \frac{\sum_{t \in [y_1, y_2]} t.TF_i[t]}{\sum_{t \in [y_1, y_2]} TF_i[t]}, \tag{3}$$

where, $\eta_i$ and $\theta_i$ is the two different measures of growth for keyword $i$, $TF_i[t]$ is the term frequency for term $i$ and year $t$ while $y_1$ and $y_2$ are the first and last years in the study period.

Hence, $\eta_i$ gives the ratio of the TF at the end of the study period to the TF at the start of the period, where two year averages are used for the TF terms for improved noise rejection. In contrast, $\theta_i$ gives the average publication year for articles appearing over the range of years being studied and which are relevant to term $i$ (a more recent year indicates greater currency of the topic). Using these different expressions provides two separate ways of measuring growth "potential" and helps to avoid confounding effects that may be peculiar to either of these measures.

### 2.3 Keyword taxonomies and semantics enriched indicators

One of the problems encountered in earlier experiments involving technology forecasting is that of measuring technology prevalence using term occurrence frequencies. This involves the fundamental problem of inferring an underlying, unobservable 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, one aspect of this problem is a semantic one where individual terms may have two or even more meanings depending on the context. Through our framework an approach was proposed in [23] through which this effect may be reduced. The basic idea is that hit counts associated with a single search term will invariably be unreliable as the contexts in which this term appear will differ. Individual terms may also suffer from the problem of extraneous usage, as in the case of 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 this problem will be mitigated. 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.

– **Taxonomy generation** - Taxonomies can sometimes be obtained from external sources and can either be formally curated or "scraped" from online sources such as Wikipedia [26].
  While many of the taxonomies obtained in this way may be helpful for the technology forecasting process, in other cases a suitable taxonomy may simply not be available, or even if available is either not sufficiently updated or is extremely expensive, thus limiting the wider adoption and use of resulting applications. As such, an important capability that has been a focus of our research is the development of a method to perform *automated* creation of keyword taxonomies based on the statistics of term occurrences.
  A detailed discussion of this topic is beyond the scope of this paper However, it is sufficient to focus on the basic idea which, 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 [27, 23] and we have also conducted a survey into different methods of perform taxonomy construction [22]), but in the present context we discuss results produced using 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 [28] 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 [23] is used (this is based on the "Google

distance" proposed in [29]):

$$\overrightarrow{\text{NGD}}(t_x, t_y) = \frac{\log n_y - \log n_{x,y}}{\log N - \log n_x}, \tag{4}$$

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 $\overrightarrow{\text{NGD}}(t_x, t_y)$ refers to the associated cost if $t_x$ is classified as a subclass of $t_y$, while $\overrightarrow{\text{NGD}}(t_y, t_x)$, 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:

$$j = \arg\min_{j \in \mathcal{T}} \overrightarrow{\text{NGD}}(t_i, t_j), \tag{5}$$

(where $\mathcal{T}$ is the set of terms which have already been incorporated into the taxonomy.)
In addition, two further customizable optimizations were added to the basic algorithm described above to improve stability, these are:

1. Attachment of a node to a parent node is based on a weighted average of the similarities to the parent but also to the grandparents and higher ancestors of that node.
2. On some occasions it was necessary to include a "child penalty" whereby the cost of attaching to a given parent increases once the number of children of that parent exceeds a certain number.

These measures and the associated parameters haven't yet been fully explored and in general are set by ad-hoc experimentation. As such they are not discussed in detail in the present context but are the subject of intense further investigations and will be explained in greater detail in future publications.

– **Enhancement and visualization of 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 re-calculate 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, aggregation was achieved by simply averaging the respective nodes' scores together with the scores of all child nodes. However, other schemes have also been considered, for example ones which emphasize the score of the current node over the child nodes.
– **Visualization** - The final piece of the puzzle is the development of a good method for representing the results of the above analysis in an intuitive and clear way. A common method for presenting information like this is in the form of a ranked list,

which in theory would allow high scoring items to be easily prioritized. However, in practice this can very often produce very misleading results. This is particularly true in our specific application where the target is to study a large numbers of keywords, many of which are closely related. In such a scenario, merely sorting the keywords by their respective scores would most likely result in closely related terms "clumping up" on these lists.

In contrast, the keyword taxonomy provides a natural alternative framework for achieving this. Firstly, the taxonomy itself allows for relations between the different taxonomies to be easily and quickly grasped. For the growth potentials, we have been experimenting with different ways of representing this information directly within the taxonomies. One simple way is to use a colour-coding scheme where "hot" technologies are coded red (for instance), and there is a range of colours leading up to "cold" technologies in blue. This, in combination with the keyword smoothing step from above means that actively growing areas of researching should turn up as red patches within the taxonomies which can then be detected easily and quickly.

### 2.4 Renewable energy case study

While this framework can potentially be used on any research domain, we conduct a pilot study on the field of renewable energy to provide a suitable example on which to conduct our experiments and to anchor our discussions. The incredible diversity of renewable energy research as well as the currency and societal importance of this area of research makes it 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 (listed in section 3.1) 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.

A number of discussions were held with subject matter experts to identify domains which were both topically current and of high value in the renewable energy industry. The selected domains were Photovoltaics (Solar Panels), Distributed Generation, Geothermal, Wind Energy and Biofuels. Search terms corresponding to each of these domains were then collected and submitted to Scopus' online search interface. These were:

| | |
|---|---|
| Renewable Energy | Embedded Generation |
| Biodiesel | Decentralized Generation |
| Biofuel | Decentralized Energy |
| Photovoltaic | Distributed Energy |
| Solar Cell | On-site generation |
| Distributed Generation | Geothermal |
| Dispersed Generation | Wind Power |
| Distrubted Resources | Wind Energy |

## 3 Results and discussions

We present results for the renewable energy case study. As described in section 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 2.3, and the growth scores $\eta$ and $\theta$ for each keyword was calculated as shown in equations (2) and (3) respectively.

The two sets of scores thus produced are ranked and the top 30 items from each are presented in 1. These scores were subsequently subjected to the taxonomy based aggregation procedure described in Section 2.3, producing two further ranked lists which are then presented in Table 2.

Based on these results, some observations are:

1. There were significant differences between the scores obtained using the different growth scoring systems, as well as with and without aggregation, as can be seen from the top-30 lists in Tables 1 and 2. However, at the same time there were also broad similarities between the two sets of rankings which pointed to the underlying "ground truth" which these rankings target. as evidenced by a large number of keywords which appeared the top ten items in both lists.
2. In fact, for the aggregated scores, the top six items on both lists are the same (though there were slight differences in the orderings of the terms within this top six set). These were: *cytology, nonmetal, semiconducting zinc compounds, hydraulic machinery, hydraulic motor, alga*. It is interesting to note that these terms correspond to important research topics within three separate sub-domains of renewable energy - biomass, solar cells and wind power.
3. It was interesting to note the number of biotechnology related keywords that were found in all four lists. This reflects the fact that biological aspects of renewable energy are amongst the most rapidly growing areas of research. 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.
4. However, many of the keywords in the lists in Tables 1 and 2 were associated with leaves in the taxonomy; this was a desirable outcome, as these were the less well

known and hence more interesting technologies, but it also meant that the confidence in the scores were lower. Looking at the terms with relatively large associated subtrees, we see that three of the largest 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.

5. 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. 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.
6. As explained in section 2.3, we also tested out a colour-coding scheme where nodes were assigned colours ranging from red through white down to blue, corresponding to the range of high to low growth technologies. This scheme was implemented and it was demonstrated to be capable of representing the results of the analysis in a highly visual and intuitive manner, in particular allowing for high growth "areas" to be identified, as opposed to focusing on individual nodes. The resulting figures are too big to be able to fit into the current format but examples of posters which were created using this technique can be viewed at: *http://www.dnagroup.org/posters/*.

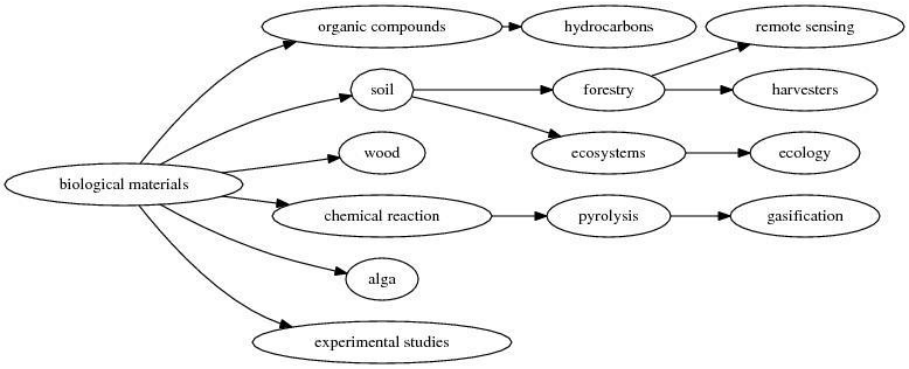


**Fig. 2.** Subtree for node "Biological materials"

### 3.1 Implementation details

The framework described here was implemented using the Python programming language. Data collection was semi-automatic and consisted of two main steps:

| **Growth Ratio ($\eta$)** | **Average publication year ($\theta$)** |
|---|---|
| 1. cytology | 1. cytology |
| 2. biological materials | 2. biological materials |
| 3. nonmetal | 3. nonmetal |
| 4. leakage (fluid) | 4. solar equipment |
| 5. solar equipment | 5. semiconducting zinc compounds |
| 6. semiconducting zinc compounds | 6. leakage (fluid) |
| 7. direct energy conversion | 7. direct energy conversion |
| 8. hydraulic machinery | 8. potential energy |
| 9. hydraulic motor | 9. alga |
| 10. potential energy | 10. hydraulic machinery |
| 11. alga | 11. hydraulic motor |
| 12. computer networks | 12. ecosystems |
| 13. bioreactors | 13. bioelectric energy sources |
| 14. ecosystems | 14. solar power plants |
| 15. bioelectric energy sources | 15. soil |
| 16. solar power plants | 16. bioreactors |
| 17. soil | 17. concentration process |
| 18. metabolism | 18. solar power generation |
| 19. concentration process | 19. metabolism |
| 20. solar power generation | 20. wastewater |
| 21. wastewater | 21. sugars |
| 22. sugars | 22. computer networks |
| 23. nonhuman | 23. nonhuman |
| 24. experimental studies | 24. experimental studies |
| 25. zea mays | 25. organic compounds |
| 26. cellulose | 26. priority journal |
| 27. priority journal | 27. biomass |
| 28. organic compounds | 28. lignin |
| 29. biomass | 29. zea mays |
| 30. lignin | 30. cellulose |

**Table 1.** Top 30 Renewable Energy Related Technology Keywords, based on (left) growth ratio (right) average publication year (raw scores)

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

**Table 2.** Top 30 Renewable Energy Related Technology Keywords, based on (left) growth ratio (right) average publication year (with aggregation)

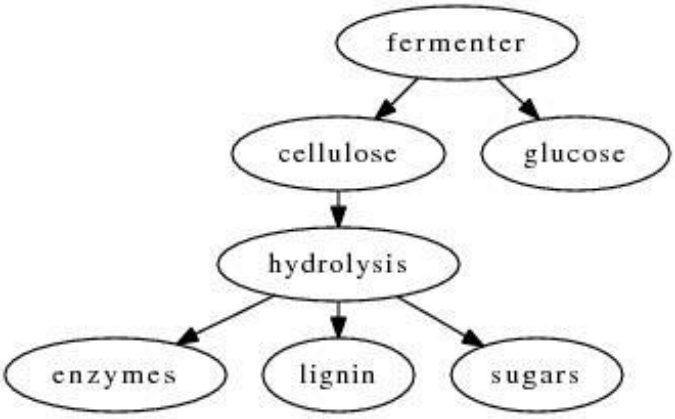


**Fig. 3.** Subtree for node "fermenter"

1. Customized keyword queries were first submitted to the Scopus search portal. The results of these searches were then exported as comma-delimited (*.csv) files and downloaded from the Scopus website.
2. Automated scripts were then created to filter and store the records in a local SQL database (we used the SQLite database system). These which were subsequently accessed using the python SQLite toolkit and appropriate SQL language calls.

Figures were generated using the *pydot* toolkit which provides a Python based interface to the Graphviz Dot language[2].

## 4 Conclusion

In this paper, we present the use of an innovative framework for visualizing the research "landscape" of the domain of renewable energy. Sucha framework will be extremely useful for supporting the relevant research planning and decision-making processes.

The 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". Finally, a colour-coding scheme is used to annotate the resulting taxonomies, allowing rapidly growing areas of research to be easily detected within the overall context of the research domain.

The simple implementation of this framework presented in this paper is used to study developments within the domain of renewable energy. More analysis is required before deeper insights can be gained and these results can be applied "on the field" by investors and other stakeholders. However, we note that the results of the analysis do seem to reflect factors and developments within the field of renewable energy.

[2] code.google.com/p/pydot/

The results of this effort are presented and discussed. While the current implementation still has ample scope for future extensions, the results are already 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.

## References


1. Alan L Porter. Technology foresight: types and methods. *International Journal of Foresight and Innovation Policy*, 6(1):36–45, 2010.
2. Alan L Porter. How" tech mining" can enhance r&d management. *Engineering Management Review, IEEE*, 36(3):72–72, 2008.
3. Joseph P. Martino. A review of selected recent advances in technological forecasting. *Technol Forecast Soc*, 70(8):719–733, October 2003.
4. Joseph Martino. *Technological Forecasting for Decision Making*. McGraw-Hill Engineering and Technology Management Series, 1993.
5. Alan Porter. Tech mining. *Compet Intel Mag*, 8(1):30–36, 2005.
6. Henry Small. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610, December 2006.
7. Ronald N Kostoff, Darrell Ray Toothman, Henry J Eberhart, and James A Humenik. Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68(3):223–253, 2001.
8. Paul Losiewicz, Douglas Oard, and Ronald Kostoff. Textual data mining to support science and technology management. *J Intell Inf Syst*, 15(2):99–119, 2000.
9. de Miranda, Gilda M. Coelho, Dos, and Lelio F. Filho. Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technol Forecast Soc*, 73(8):1013–1027, 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. Anuradha, K., Urs, and Shalini. Bibliometric indicators of indian research collaboration patterns: A correspondence analysis. *Scientometrics*, 71(2):179–189, May 2007.
12. Wen-Ta Chiu and Yuh-Shan Ho. Bibliometric analysis of tsunami research. *Scientometrics*, 73(1):3–17, October 2007.
13. Tibor Braun, Andrs P. Schubert, and Ronald N. Kostoff. Growth and trends of fullerene research as reflected in its journal literature. *Chem Rev*, 100(1):23–38, 2000.
14. N. R. Smalheiser. Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Technovation*, 21(10):689–693, October 2001.
15. 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.
16. Tugrul U. Daim, Guillermo Rueda, Hilary Martin, and Pisek Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technol Forecast Soc*, 73(8):981–1012, 2006.
17. Ying Guo, Lu Huang, and Alan L Porter. The research profiling method applied to nano-enhanced, thin-film solar cells. *R&d Management*, 40(2):195–208, 2010.
18. Wei Lee Woon, Hatem Zeineldin, and Stuart Madnick. Bibliometric analysis of distributed generation. *Technological Forecasting and Social Change*, 78(3):408–420, 2011.

19. Ming-Yueh Tsay. A bibliometric analysis of hydrogen energy literature, 1965–2005. *Scientometrics*, 75(3):421–438, 2008.
20. Chen Yu-Heng, Chen Chia-Yon, and Lee Shun-Chung. Technology forecasting of new clean energy: The example of hydrogen energy and fuel cell. *African Journal of Business Management*, 4(7):1372–1380, 2010.
21. Wei Lee Woon, Andreas Henschel, and Stuart Madnick. A framework for technology forecasting and visualization. In *Innovations in Information Technology, 2009. IIT'09. International Conference on*, pages 155–159. IEEE, 2009.
22. Andreas Henschel, Wei Lee Woon, Thomas Wachter, and Stuart Madnick. Comparison of generality based algorithm variants for automatic taxonomy generation. In *Innovations in Information Technology, 2009. IIT'09. International Conference on*, pages 160–164. IEEE, 2009.
23. W.L. Woon and S. Madnick. Asymmetric information distances for automated taxonomy construction. *Knowl Inf Syst*, Online first, 2009.
24. B. Ziegler, A.K. Firat, C. Li, S. Madnick, and W.L. Woon. Preliminary report on early growth technology analysis. Technical Report CISL #2009-04, MIT, http://web.mit.edu/smadnick/www/wp/2009-04.pdf, 2009.
25. Murat Bengisu and Ramzi Nekhili. Forecasting emerging technologies with the aid of science and technology databases. *Technol Forecast Soc*, 73(7):835–844, September 2006.
26. Gihan Dawelbait, Andreas Henschel, Toufic Mezher, and Wei Lee Woon. Forecasting renewable energy technologies in desalination and power generation using taxonomies. *International Journal of Social Ecology and Sustainable Development (IJSESD)*, 2(3):79–93, 2011.
27. Wei Lee Woon and Stuart Madnick. Semantic distances for technology landscape visualization. *Journal of Intelligent Information Systems*, 39(1):29–58, 2012.
28. 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.
29. Rudi L. Cilibrasi and Paul M. B. Vitányi. The google similarity distance. *IEEE T Knowl Data En*, 19(3):370–383, 2007.