MIT Sloan School of Management

MIT Sloan School Working Paper 4711-08
8/25/2008

Semantic Distances for Technology Landscape Visualization

Wei Lee Woon, Stuart E. Madnick

Semantic Distances for Technology Landscape Visualization

Wei Lee Woon
Stuart E. Madnick

Working Paper CISL# 2008-04

August 2008

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

Semantic distances for technology landscape visualization

Wei Lee Woon*, Stuart Madnick†
*Masdar Institute of Science and Technology,
(Visiting Scholar) Technology and Development Program,
M.I.T., 1-175, Cambridge MA, 02139, U.S.A.
†Sloan School of Management, M.I.T.,
E53-321, Cambridge MA, 02139, U.S.A.
wwoon@mist.ac.ae, *smadnick@mit.edu*

***Abstract*— This paper presents a novel approach to the visualization and subsequent elucidation of research domains in science and technology. The proposed methodology is based on the use of *bibliometrics*; i.e., analysis is conducted using information regarding trends and patterns of publication rather than the contents of these publications. In particular, we explore the use of term co-occurence frequencies as an indicator of the semantic closeness between pairs of words or phrases. To demonstrate the utility of this approach, a case study on renewable energy technologies is conducted, where the above techniques are used to visualize the inter-relationships within a collection of energy-related keywords. As these are regarded as manifestations of the underlying research topics, we contend that the proposed visualizations can be interpreted as representations of the underlying technology landscape. These techniques have many potential applications, but one interesting challenge in which we are particularly interested is the mapping and subsequent prediction of future developments in the technological fields being studied.**

I. Introduction

A. Technology mining

The planning and management of research and development activities is a challenging task that is further compounded by the large amounts of information which researchers and decision-makers are required to sift through. One difficult problem is the need to gain a broad understanding of the current state of research, future scenarios and the identification of technologies with potential for growth and which hence need to be emphasized. Information regarding past and current research is available from a variety of channels (examples of which include publication and patent databases); the task of extracting useable information from these sources, known as “tech-mining”[Porter, 2005], presents both a difficult challenge and 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 research addressing this problem (for a good review, the reader is referred to [Porter, 2005], [Porter, 2007], [Losiewicz et al., 2000], [Martino, 1993]); interesting examples include visualizing the inter-relationships between research topics [Porter, 2005], [Small, 2006], identification of important researchers or research groups [Kostoff, 2001], [Losiewicz et al., 2000], the study of research performance by country [de Miranda et al., 2006], [Kim and Mee-Jean, 2007] the study of collaboration patterns [Anuradha et al., 2007], [Chiu and Ho, 2007], [Braun et al., 2000] and the prediction of future trends and developments [Smalheiser, 2001], [Daim et al., 2005], [Daim et al., 2006], [Small, 2006]. Nevertheless, given the many difficulties inherent to these undertakings, there is still much scope for further development in many of these areas.

B. Novelty and motivations

An important motivation for attempting technology-mining is the possibility of gaining a better understanding of future developments and trends in a given field of research. This is a complex task that is composed of a number of closely inter-related components or activiites. While there is no single authoritative classification, we present the following scheme, proposed in [Porter et al., 1991], to help focus our discussion:

- *Monitoring* - Observing and keeping up with developments occurring in the environment, and which are relevant to the field of study [Kim and Mee-Jean, 2007], [King, 2004].
- *Expert opinion* - An important method for forecasting technological development is via intensive consultation with subject matter experts [Van Der Heijden, 2000].
- *Trend extrapolation* - This involves the extrapolation of quantitative historical data into the future, often by fitting appropriate mathematical functions [Bengisu and Nekhili, 2006].
- *Modeling* - It is sometimes possible to build causal models which not only allow future developments to be known, but also allow the interactions between these forecasts and the underlying variables or determinants to be better understood [Daim et al., 2005], [Daim et al., 2006].
- *Scenarios* - Forecasting via scenarios involves the identification of key events or occurrences which may determine the future evolution of technology [Mcdowall and Eames, 2006], [Van Der Heijden, 2000].

In this context, the emphasis of the current study is on the first item, *viz* technology *monitoring*, as the primary objective is to devise methods for monitoring, understanding and mapping the current state of technology. In particular, our aim is to develop novel approaches to visualize and understand the relationships between connected areas of science and technology. Towards this end, this paper will address the following objectives:

1) To devise a method for quantifying the degree of similarity between research areas.
2) To use the distance measure to study the structure of the research “landscape” of the target domain. We are also interested in detecting and exploiting clusters of closely related topics.
3) To conduct a preliminary case study in renewable energy as a demonstration of the proposed approach.

C. Case study

To provide a suitable example on which to conduct our experiments and to anchor our discussions, a preliminary case study was conducted in the field of renewable energy.

The importance of energy to the continued well-being of society cannot be understated, yet 87%[1] of the world’s energy requirements

[1] year 2005. Source: Energy Information Administration, DOE, US Government

are fulfilled via the unsustainable burning of fossil fuels. A combination of environmental, supply and security problems compounded the problem further, making renewable energies such as wind power and solar energy one of the most important topics of research today.

An additional consideration was the incredible diversity of renewable energy research, which promises to be a rich and challenging problem domain on which to test our methods. Besides high-profile topics like solar cells and nuclear energy, renewable energy related research is also conducted in fields like molecular genetics and nanotechnology. It was this valuable combination of social importance and technical richness that motivated the choice of renewable energy as the subject of our case study.

II. Methods and data

In the following subsections, the methods used for both data collection and analysis will be discussed in some detail. The overall process will be based on the following two stages:

1) Identification of an appropriate indicator of closeness (or distance) between terms which can be used to characterize the relationships between areas of research,
2) Use of this indicator to perform feature extraction on the data, which could be in the form of intuitive visualizations or clusters.

A. Keyword distances

The key requirement for stage one is a method of evaluating the similarity or distance between two areas of research, represented by appropriate keyword pairs. Existing studies have used methods such as citation analysis [Saka and Igami, 2007], [Small, 2006] and author/affiliation-based collaboration patterns [Zhu and Porter, 2002], [Anuradha et al., 2007] to extract the relationships between researchers and research topics. However, these approaches only utilize information from a limited number of publications at a time, and often require that the text of relevant publications be stored locally (see [Zhu and Porter, 2002], for example). As such, extending their use to massive collections of hundreds of thousands or millions of documents would be computationally unfeasible.

Instead, we choose to explore an alternative approach which is to define the relationship between research areas in terms of correlations between the occurrences of related keywords in the academic literature. Simply stated, the appearance of a particular keyword pair in a large number of scientific publications implies a close relationship between the two keywords. Accordingly, by utilizing the co-occurence frequencies between a representative collection of keywords, we seek to demonstrate that it is possible to infer the overall research "landscape" for a particular domain of research.

In practice, exploiting this intuition is more complicated than might be expected as it is not clear what the exact expression for this distance should be. Rather than screen a number of alternatives on an ad-hoc basis, can this distance be derived using a rigorous theoretical framework such as probability or information theory? As it turns out, there is already a method which provides this solid theoretical foundation, and which exploits the same intuition. This method is known as the *Google Distance* [Cilibrasi and Vitanyi, 2006], [Cilibrasi and Vitányi, 2007], and is defined as:

$$\mathrm{NGD}(t_1, t_2) = \frac{\max\{\log n_x, \log n_y\} - \log n_{x,y}}{\log N - \min\{\log n_x, \log n_y\}}, \tag{1}$$

where NGD stands for the *Normalized Google Distance*, t_1 and t_2 are the two terms to be compared, n_1 and n_2 are the number of results returned by a Google search for each of the terms individually and $n_{1,2}$ is the number of results returned by a Google search for both of the terms. A detailed discussion of the theoretical underpinnings of this method is beyond the present scope but the general reasoning behind eq.(1) is quite intuitive, and is based on the normalized information distance:

$$\mathrm{NID}(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}, \tag{2}$$

where x and y are the two strings (or other data objects such as sequences, program source code, etc.) which are to be compared. $K(x)$ and $K(y)$ are the Kolmogorov complexities of the two strings individually, while $K(x, y)$ is the complexity of the combination of the two strings. The distance is hence a measure of the additional information which would be required to encode both strings x and y given an encoding of the shorter of the strings. The division by $\max\{K(x), K(y)\}$ serves as a normalization term which ensures that the final distance lies in the interval [0,1].

In the present context, the Kolmogorov complexity is substituted with the prefix code length, which is given by:

$$K(x, y) \Rightarrow G(x, y) = \log\left(\frac{N}{n_{x,y}}\right), \tag{3}$$

$$K(x) \Rightarrow G(x) = G(x, x). \tag{4}$$

In the above, N is the size of the sample space for the "google distribution", and can be approximated by the total number of documents indexed by the search engine being used. Substituting (3),(4) $\rightarrow$ (2) then leads to eq. (1).

To adapt the framework above for use in technology mapping and visualization, we introduce these simple modifications:

1) Instead of a general Web search engine, the prefix code length will be measured using hit counts obtained from a scientific database such as Google Scholar or Web of Science.
2) N is set to the number of hits returned in response to a search for "renewable+energy", as this represents the size of the body of literature dealing with renewable energy technologies.
3) We are only interested in term co-occurences which are within the context of renewable energy; as such, to calculate the co-occurence frequency $n_{i,j}$ between terms t_1 and t_2, the search term "'renewable+energy"+"t_1"+"t_2"' was submitted to the search engine.

As explained in [Cilibrasi and Vitányi, 2007], the motivation for the Google distance was to create an index which quantifies the semantic similarity between objects (words or phrases) which reflected their usage patterns in society at large. By following the same line of reasoning, we can assume that term co-occurence patterns in the academic literature would characterize the similarity between technology related keywords in terms of their usage patterns in the scientific and technical community.

This distance measure can now be used to calculate the distances between all pairs of keywords in the corpus, resulting in the following distance matrix **D**:

$$\mathbf{D} = \begin{bmatrix} d_{1,1} & \dots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \dots & d_{n,n} \end{bmatrix}, \tag{5}$$

where $d_{i,j}$ denotes the distance between keywords or terms t_i or t_j.

Given this matrix, the next challenge is to investigate methods for converting matrix **D** into useful representations of the data. This can be done in a variety of ways but for now the focus will be on *clustering* and *visualization*; these will be described briefly in the following section.

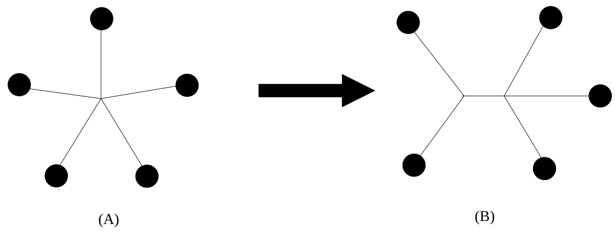

Fig. 1. Creation Of Hierarchical Tree Using The Neighour Joining Method

B. Data representations

Visualization

When dealing with high-dimensional or complex datasets, algorithms for visualizing the data in an intuitive way are extremely useful, serving as a source of valuable insight into the general structure of the data.

For our experiments, we used the popular hierarchical visualization algorithm proposed in [Saitou and Nei, 1987]. The algorithm produces the keyword hierarchy which provides the simplest explanation for the distances observed between the keywords. Simplest here is achieved via finding the tree with the smallest total branch length. Briefly, the algorithm proceeds in iterative fashion as follows:

Algorithm *Neighbor-Join*($\mathcal{T}$, **D**)
Input: A term-set $\mathcal{T}$ with the elements $t_1, \ldots, t_N$; a matrix **D** with elements $d_{i,j}$ representing the distances between terms $t_i, t_j \in \mathcal{T}$
Output: An unrooted tree visualization
1. Initialize the tree in a star topology as illustrated in fig. 1(a) (example depicted is of a five-keyword collection)
2. **for** $t_i, t_j \in \mathcal{T}$
3. **do**
4. Identify $i, j = \text{argmin}_{i,j}\,(S_{i,j})$, where:

$$S_{i,j} = \frac{1}{2(|\mathcal{T}|-2)} \sum_{k=3}^{N} (d_{1,k} - d_{2,k}) + \frac{1}{2} d_{i,j} + \ldots$$
$$\ldots + \frac{1}{|\mathcal{T}|-2} \sum_{3 \le i \le j} d_{i,j}. \quad (6)$$

5. Combine nodes t_i and t_j as shown in fig. 1.
6. **until** no node has more than three branches emanating from it.

As an example, we consider the following collection of ten keywords which were highlighted as being high-growth areas in renewable energy [Kajikawa et al., 2007]: *combustion, coal, battery, petroleum, fuel cell, wastewater, heat pump, engine, solar cell, power system*.

Distance matrices generated using the Google Scholar[2] search engine were used to create a hierachical visualization tree as described above. These are shown in fig. 2. For comparison, the visualization tree generated using the Scirus search engine [3] has also been included in fig.3. Though only intended as a preliminary demonstration, we already see some interesting patterns:

1) Broadly speaking, the structure of the keyword trees seem logical in that keywords which seem related to similar areas of research have been placed in related branches.
2) Also, it can be seen that the two trees have almost identical structures. In both cases there are three main clusters; the first consists of {*combustion, coal, petroleum*}, the second {*wastewater, heat pump*}, while the third cluster consists of {*battery, solar cell, power system, fuel cell*}. The only real difference is that *heat pump* and *wastewater* are paired up in fig.2 while in 3 *heat pump* is an immediate "ancestor" of *wastewater*.
3) This is an important observation, as it supports the notion that the distance measure proposed has at least a certain degree of independence from the databases which were used to calculate it. This is not a given fact as our observations have been that the results returned by these two search engines can vary a lot - In general Google scholar returns a very much large number of hits, and also includes patents in its searches. Manual inspection of the actual publications returned by the two search engines also indicated that the techniques used to index and sort these publications are likely to be very different, though detailed information about the ranking and selection procedures used is not available.

[2]http://scholar.google.com
[3]http://www.scirus.org

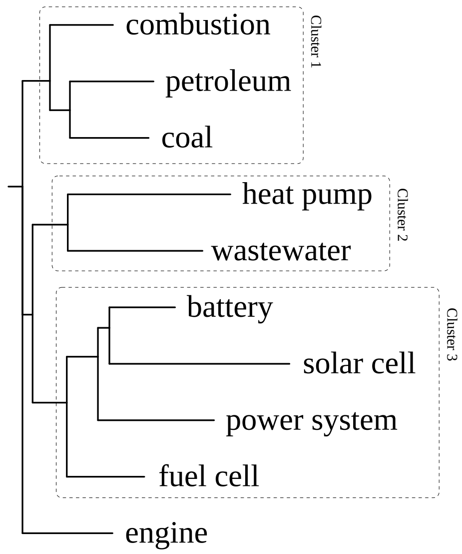

Fig. 2. Visualization tree for Kajikawa data (the three clusters referenced in the text are clearly labelled)

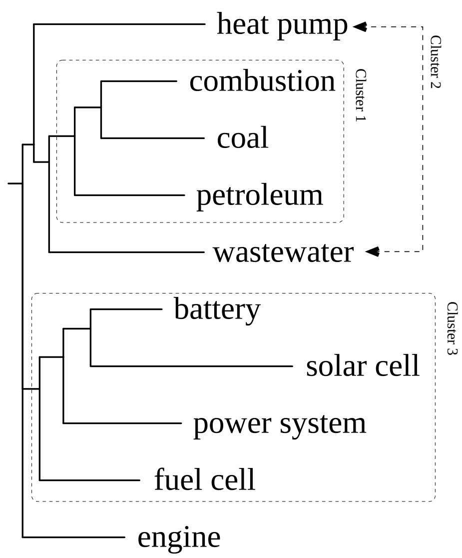

Fig. 3. Visualization tree for Kajikawa data, generated using the Scirus database (Clusters are labelled)

4) All three of these clusters appear to consist of topics which are closely related: clusters 1 and 3 are somewhat self-evident, while cluster 2 also makes sense as there is a significant amount of research in the use of heat pumps to reclaim heat from wastewater [Baek et al., 2005], [Elnekave, 2008].
5) The keyword $\{$*engine*$\}$ is seen to be somewhat isolated from the rest of the group.

Clustering

Clustering is the process of dividing large sets of objects - in this case keywords - into smaller groups containing closely related terms; this is useful as these groupings could then be used to construct enriched keywords queries, organize the objects into topical hierarchies and to perform various classification tasks.

This is an important operation in data mining and can be attempted in a number of ways; one of the most common methods is the k-means algorithm [Bishop, 2006]. This works by dividing the data into k clusters, each anchored by a centroid vector representing the mean position of the cluster. The optimal clustering is found iteratively by alternating between:

1) Re-estimating the position of the centroids (by calculating the mean of the assigned vectors),
2) Revising the groupings by re-assigning data points to the clusters with the closest centroids.

In the present context there is a slight complication in that instead of data vectors, only the distance matrices are available. As such, instead of the regular k-means algorithm, the following modified algorithm, *Matrix-k-means*, is proposed:

Algorithm *Matrix-k-means*($\mathcal{T}$, **D**,k)
Input: A term-set $\mathcal{T}$; a matrix **D** with elements $d_{i,j}$ representing the distances between terms $t_i, t_j \in \mathcal{T}$; k, the number of clusters
Output: A clustering $\mathbf{c} = [\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k]$, where $\bigcup_{i=i}^{k} \mathcal{C}_i = \mathcal{T}$ and $\bigcap_{i=i}^{k} \mathcal{C}_i = \emptyset$
Select random centroids $t_1^* \ldots t_k^* \in \mathcal{T}$
$\mathbf{t} \leftarrow [t_1^*, \ldots, t_k^*]$
$\mathbf{c} \leftarrow [\{t_1^*\}, \ldots, \{t_k^*\}] \; (= [\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k])$
repeat
  $\mathcal{T}' \leftarrow \mathcal{T} - \{\mathbf{t}\}$
  for $t_i \in \mathcal{T}'$
    do $l \leftarrow \operatorname{argmin}_j \{d_{i,j} : t_j \in \{\mathbf{t}\}\}$
      $\mathcal{C}_l \leftarrow \mathcal{C}_l + \{t_i\}$
  for $i \leftarrow [1, k]$
    do $j \leftarrow \operatorname{argmin}_j \left\{\sum_{t_l \in \mathcal{C}_j} d_{j,l}\right\}$
      $t_i^* \leftarrow t_j$
until termination criterion met

As the k-means algorithm is a Greedy algorithm, there is a dependence on the initial choice of cluster centroids which, for larger collections, can make a significant difference in the final outcome of the iterations. As such, in practice, the algorithm above was run for a number of times, then Dunn's validity index was used to select the optimal clustering. This is defined as:

$$D = \min_{\{i,j: i,j \in \mathbf{c}, i \neq j\}} \left\{ \frac{\mathbf{d}_{i,j}}{\max_{1 \leq k \leq n} \delta_k} \right\}, \qquad (7)$$

where **d** is the *inter*-cluster distance, defined as mean distance between elements in clusters i and j, δ_k is the *intra*-cluster distance, defined as the mean distance between all elements within cluster k and n is the number of clusters.

As in the previous section, the modified k-means algorithm described above was applied to the ten keywords extracted from [Kajikawa et al., 2007]. Again, the Google and Scirus distances were generated as explained in section II-A and used to decompose the keywords into a number of smaller sets. The procedure was repeated 10 times and the best clustering was selected based on the Dunn index. The same clusters were obtained in both cases, and were as follows:

- cluster 1: *battery, fuel cell, solar cell, power system*
- cluster 2: *heat pump*
- cluster 3: *engine, combustion, petroleum, coal, wastewater*

Comparing the results obtained here, and the clusters labelled in figures 2 and 3, we see that the divisions of the keywords into categories are extremely similar. The only exceptions are that *engine* and *wastewater* have now been moved into the same cluster with *combustion, petroleum and coal,* while *heat pump* is now in its own cluster.

C. Data collection

As mentioned in section I-C, a more extensive case study on renewable energy technologies was conducted to evaluate the proposed techniques. The main data requirement was for a set of energy related keywords and a populated distance matrix containing the inter-keyword distances.

Energy related keywords were extracted using ISI's Web of Science database: a search for "renewable+energy" was submitted, and the matching publications were sorted according to citation frequency. The top 30 records were retained, then two separate groups of keywords were collected for use in our experiments - the first collection was obtained using the "Author Keywords" feature and the second collection was obtained using the "Keyword Plus" feature; the former is composed of keywords specified by the authors, while the latter consists of keywords extracted from the titles of linked publications (the complete lists of keywords are provided in Appendix I of this paper). In total, 59 author keywords were extracted while 133 terms were extracted using the keyword plus feature.

Once the keywords were collected, the distances discussed in section II-A could be calculated. Only hit counts from Google scholar were used this time - the Scirus search engine was not used as there were many specialized terms in the collections for which Scirus returned no hits at all. Similarly, a number of other alternatives were considered including the Web of Science, Inspec, Ingenta, Springer and IEEE databases; again, a preliminary survey indicated that very low numbers of hits, or none at all, were returned for a large proportion of the keyword pairs. There appeared to be two main reasons for this observation: Firstly, most of these search engines simply did not index a large enough collection to provide ample coverage of the more specialized of the keywords that were in the list; Secondly, not all of the search engines allowed full text searches (the Web of Science database, for example, only allows searching by keywords or topics) - while sufficient for literature searches and reviews, keyword searches simply did not provide sufficient data for our purposes.

III. Results

The experiments described in the previous sections were performed on the two keyword collections. Some overall observations were:

1) As expected, an informal inspection of the search results confirmed that terms which were closely related had a large number of joint-hits, while distantly related terms only appeared together in a small number of papers. For example, 14000 papers were found to contain the terms *natural gas* and *power generation*, while only 484 hits were returned when a search for *natural gas* and *genomics* was conducted.

2) However, one problem which was encountered was the large number of largely generic keywords, such as *review*, *chemicals* and *fuels* in the case of the author defined keywords, and *liquid*, *mechanisms*, *metals*, *cells* and *products* in the collection of plus keywords. Problems might arise as these terms tended to have a high degree of intersection with almost all other terms - for example, searching for *Review* and *natural gas* resulted in 21000 joint hits, and *Review* and *genomics* yielded 1610 joint hits. Depending on the type of data analysis technique used, these results could erroneously imply a high degree of similarity between *genomics* and *natural gas*.
3) There were also some problems with data quality and consistency. As the data in the Google scholar database is constantly evolving, it is not possible to ensure consistenty of all the hit counts. In one specific case, we noticed that the number of publications which contained both *Trichoderma Reesei QM-9414* and *System* was actually more than the hit count returned when a search for only *Trichoderma Reesei QM-9414* was conducted. It later turned out that this was due to the two searches being conducted on different days, and that in the intervening time additional publications had already been found containing the two terms.
   Another example is the fact that the hit counts returned by Google scholar are known to be approximations of the total number of relevant publications (as the user clicks through the results pages, the number reported gradually converges to the actual value). For instance, it was observed that the hit counts from searches over a range of years, conducted individually, did not add up to the total number of hits returned when the entire range of years were searched in a single query. Problems such as these arise because of the novel ways in which these databases are being used. It is hoped that because we are using aggregate data over a range of search terms, inconsistencies such as these will be averaged out.

In the following subsections the results obtained from carrying out the proposed analysis on the two sets of keywords will be described in greater detail.

A. Author keywords

As mentioned previously, these are the keywords specified manually by the authors of publications (a full list of the 59 keywords in this collection are provided in Appendix I).

As in section II-B, we start by using the hierarchical visualization to obtain an overall view of the keyword inter-relationships. This is shown in fig. 4.

From the tree diagram, we can see that there is a definite clustered structure in the data. In some cases, it is difficult to judge the validity of the clusterings, in particular in the case of general terms like “chemicals”, “review” and “electricity”. However based on fig. 4, we can identify at least five major clusters. These have been clearly labelled in the figure and are:

- **C1**: This is composed of the terms {*thermal processing, thermal conversion, co-firing, alternative fuels, transesterification, sunflower oil, biodiesels, bio-fuels*}. These terms are definitely closely linked, and are representative of research efforts related to biodiesel processing.
- **C2**: Consisting of the keywords {*Sugars, Model plant, Enzymatic digestion, Populus, Genome sequence, QTL, Arabidopsis, Genomics, Poplar, Corn stover, Pretreatment, Hydrolysis*}, this second cluster spans a selection of renewable energy relevant biotechnology applications, in particular the production of biomass.

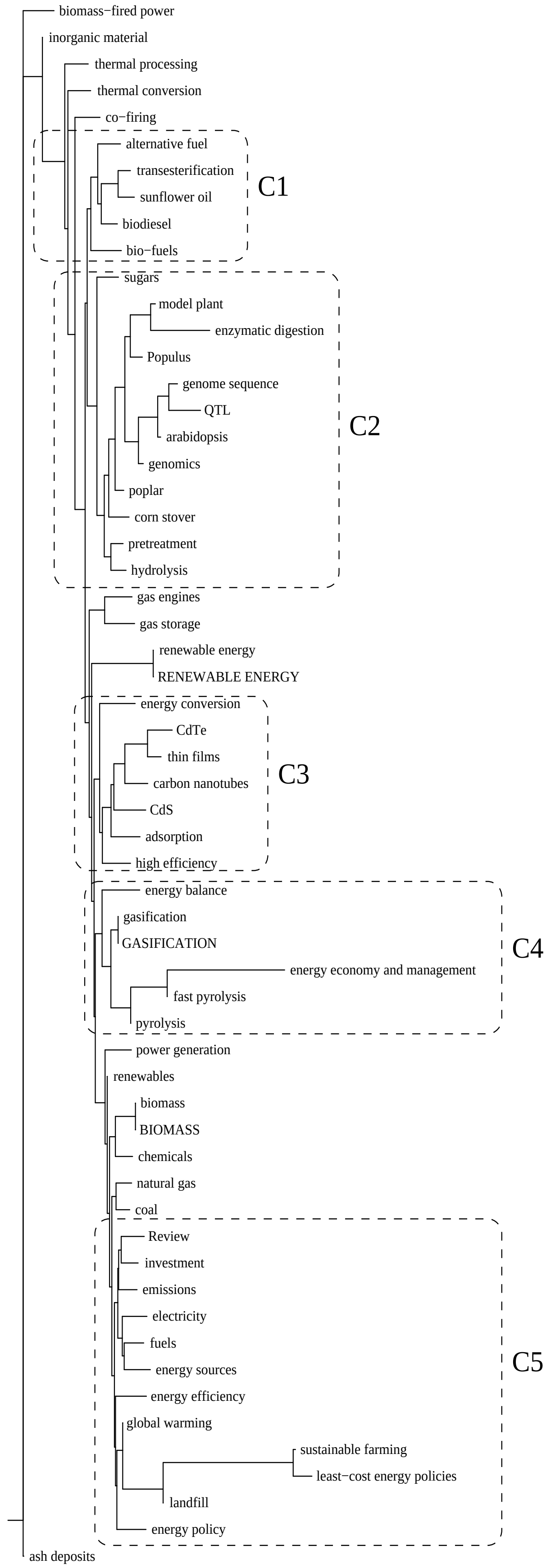

Fig. 4. Visualization tree for Author keyword data

- **C3**: This cluster contains the terms {*CdTe, Thin films, Carbon nanotubes, CdS, Adsorption, High efficiency*}, all of which are associated with the manufacture of thin film solar cells.
- **C4**: This cluster consists of {*Gasification, GASIFICATION, Energy economy and management, Fast pyrolysis, Pyrolysis*}, which are broadly related to the topic of gasification. The exception seems to be the node "energy economy and management", which seems a little out of place (however, it is a very generic term and could be related in a number of indirect ways).
  Note also the occurrence of the terms "gasification" and "GASIFICATION" - both terms were present in the automatically scraped keyword lists and were included as a useful example of "dirty" data, which illustrates the usefulness of grouping semantically similar words together as a means of removing redundancies.
- **C5**: The final cluster consists of the keywords:{*Review, Investment, Emissions, Electricity, Fuels, Energy Sources, Energy efficiency, Global Warming, sustainable farming, least cost energy policies, landfill, energy policy*}, and is a collection of policy related research keywords.

Outside of these five clusters, the remaining terms also form a number of "micro-clusters" consisting of keyword pairs or triplets. The pairs of {*biomass,BIOMASS*} and {*renewable energy, RENEWABLE ENERGY*} are further examples of the semantic matching phenomena observed in cluster 4 earlier. Other keyword collections which also appear reasonable include {*natural gas, coal*}, {*gas engines, gas storage*} and {*review, investment, emissions*}.

Finally, it must also be noted that there are some observations which cannot be explained immediately or in a straightforward manner. For example, there is no clear explanation for the positions of the keywords "Biomass fired power" and "Inorganic material". It is still too early to speculate on the nature of these relationships, except to note that even as we proceed with guarded optimism, some degree of caution must be exercised when dealing with data that is automatically extracted from source over which we have no control.

Next, we study the keyword clusters generated using the k-means algorithm. The *matrix-k-means* algorithm (page 4) was used to automatically partition the author keyword collection into 10 categories. As described in section II-B, the clustering operation has an element of randomness - to reduce this, the operation was repeated a total of 60 times and the best clustering in terms of the Dunn index was selected as the ideal solution. The clusters thus generated are presented in table I. In general we observed the following:

1) Broadly, the clusters generated in this way exhibited a structure that was similar to the groupings observed in the hierarchical tree visualization (to facilitate the following discussions, we have labelled the clusters derived using k-means as K1→K9, to help distinguish the two sets of clusters)
2) Cluster K1 is exactly the same as cluster C3.
3) The combination of clusters K3 and K4 (Biomass related terms) were practically identical to cluster C2, with the only exception being the term "co-firing", which only appeared in K3; however, it is an "ancestor" of C2, which explains its appearance in this group.
4) It appears that a number of keywords relevant to Biodiesel, Biomass and Gasification have become somewhat inter-mingled in clusters K7 and K9, though the emphasis in K7 seems to be on Biodiesel, and K9 seems more focussed on Gasification. This is not surprising given the broad overlaps between these three topics.
5) Finally, the combination of clusters K8 and K10 contains many policy related issues, and closely matches the keywords found in C5.

Cluster#	Keywords
K1	Energy Conversion, Cdte, Adsorption, High Efficiency, Cds, Thin Films
K2	Energy Economy And Management
K3	Sugars, Populus, Pretreatment, Arabidopsis, Qtl, Co-Firing, Genomics, Corn Stover, Poplar, Hydrolysis
K4	Model Plant, Enzymatic Digestion, Genome Sequence
K5	Energy Balance
K6	Ash Deposits, Inorganic Material, Biomass-Fired Power Boilers
K7	Transesterification, Gas Engines, Bio-Fuels, Thermal Conversion, Thermal Processing, Carbon Nanotubes, Sunflower Oil, Pyrolysis, Fast Pyrolysis
K8	Natural Gas, Renewable Energy, Review, Energy Efficiency, Investment, Electricity, Global Warming, Renewables, Fuels, Energy Sources, Energy Policy, Power Generation, Coal, Emissions, Renewable Energy
K9	Alternative Fuel, Biomass, Gasification, Biodiesel, Gas Storage, Chemicals, GASIFICATION, BIOMASS
K10	Sustainable Farming And Forestry, Least-Cost Energy Policies, Landfill

TABLE I
CLUSTERS GENERATED AUTOMATICALLY BY APPLYING THE K-MEANS ALGORITHM TO THE AUTHOR KEYWORDS DATA

B. Keyword plus

Next, the set of key terms extracted using keyword plus of the ISI Web of Science database were studied in the same way. For the hierarchical visualizations, it is not possible to present the entire tree diagram due to the large number of keywords (133 in this collection). Instead, it has been broken into two sub-trees and these are shown in figures 5 and 6 respectively. As in the previous section, the keyword tree indicated a clear clustered structure with a number of prominent, identifiable clusters, labelled as CP1→CP7 (in the interest of brevity, we have been a little more selective this time around due to the larger number of keywords):

- **CP1**: This cluster contained the following terms: {*SP Strain ATCC-29133, Bidirectional Hydrogenase, Anabaena Variabilis, Anacystis Nidulans, Nitrogen Fixation*}; these keywords are associated with bio-production of hydrogen using Cynaobacterial strains.
- **CP2**: Consisting of the following keywords: {*Transgenic Poplar, Genetic Linkage Maps, RAPD Markers, Agrobacterium mediated transformation, Hybrid Poplar, Molecular Genetics, FIMI, Trichoderma Reesei Q, Corn stover, Wood, Fuels*}, this second cluster contained terms related to research on the production of Biomass.
- **CP3**: This next collection of terms included the following: {*Ruthenium Polypyridyl Complex, Sensitized Nanocrystalline* TiO_2, *Metal Complexes, Differentiation, Nanocrystalline seminconductor films, water oxidation, CDS, Recombination, Sputtering deposition, Electrodes, Films, Grain Morphology, Adsorption*}, all of which are relevant to solar cell production.

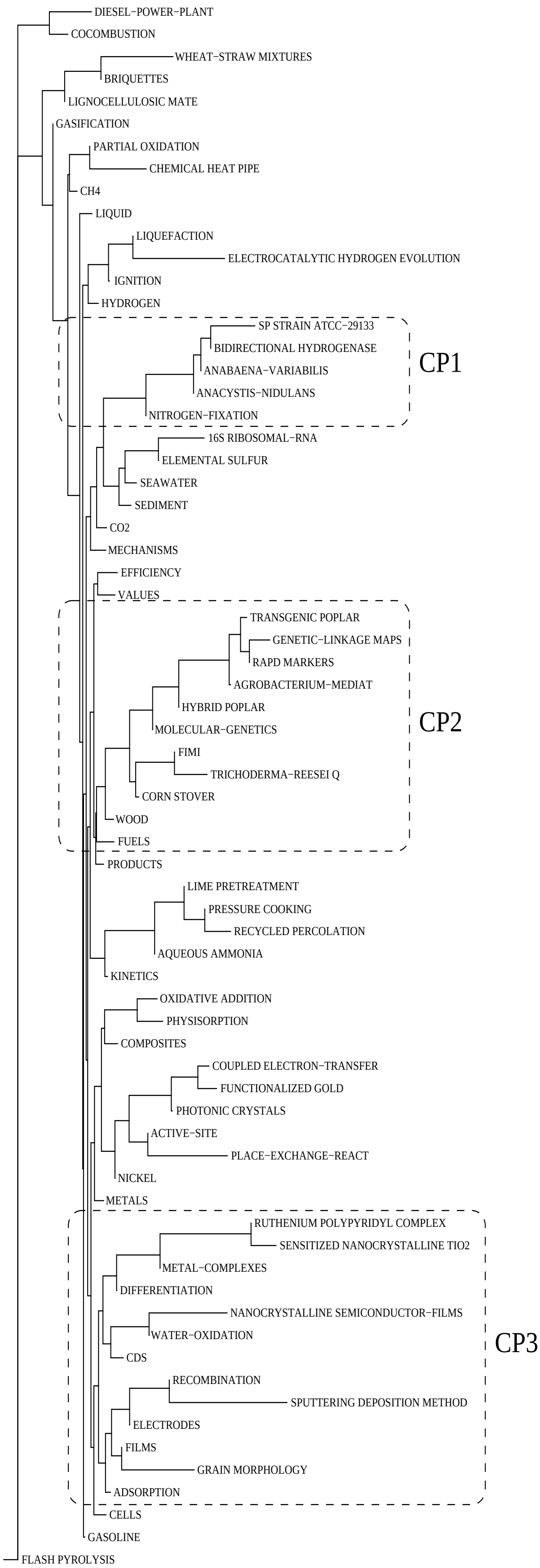

Fig. 5. Visualization tree for the keyword plus data (set 1)

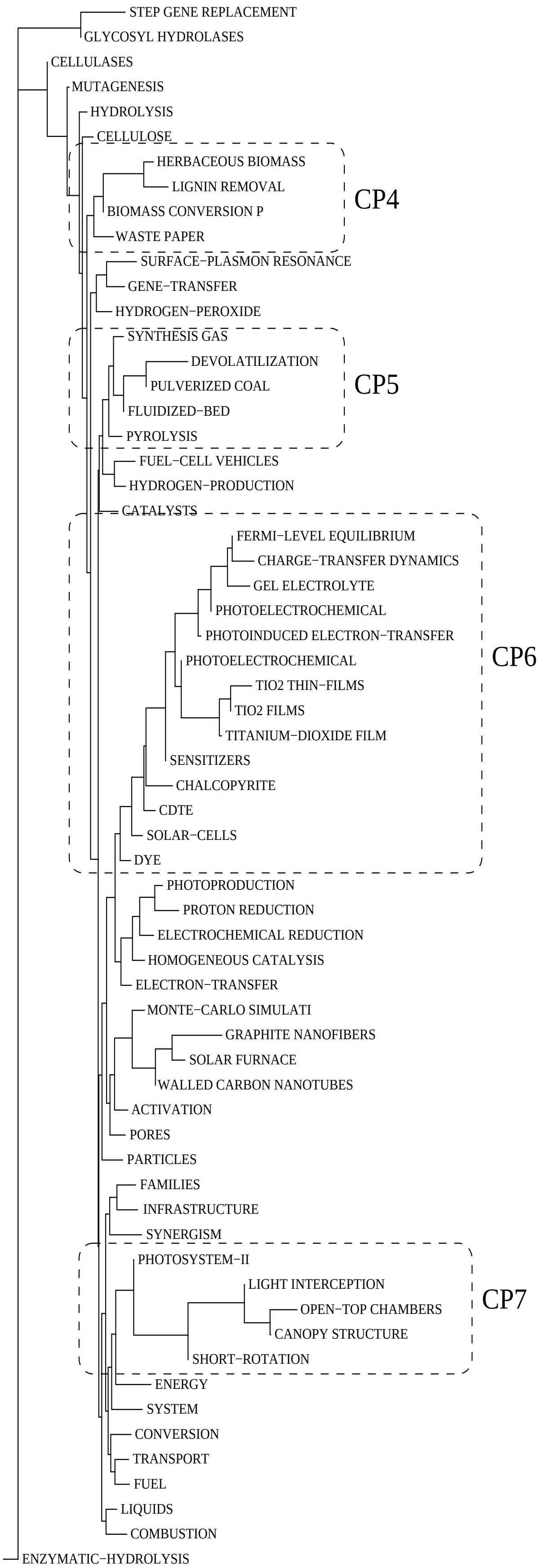

Fig. 6. Visualization tree for the keyword plus data (set 2)

- **CP4**: The fourth cluster comprised the following terms {*Herbaceous biomass, Lignin removal, Biomass conversion processes, Waste paper*}, and is also linked to research on Biomass.
- **CP5**: This cluster consisted of: {*Synthesis gas, Devolatilization, Pulverized coal, Fluidized bed, Pyrolysis*}, all of which are keywords related to gasification.
- **CP6**: This was a very large cluster consisted of the following terms: {*Fermi level equilibrium, Charge-transfer dynamics, Gel electrolyte, Photoelectrochemical properties, Photoelectrochemical cells, Photoinduced electron transfer, $Ti0_2$ thin films, TiO_2 films, Titanium dioxide films, Sensitizers, Chalcopyrite, CdTE, Solar-Cells, Dye*}. All of these keywords are related to research in the field of Solar Cell.
- **CP7**: Finally, the last cluster, which was focused on the area of Biomass crops, contained the following keywords {*Photosystem II, Light interception, Open-top chambers, Canopy structure, Short rotation*}.

Again, as in the previous set of keywords, the structure of the hierarchy grouped terms which were relevant to particular research issues in renewable energy. Also, there is a good correspondance between the clusters observed here and the clusters created from the "author keywords" collection. This is to be expected since these keywords were obtained from the same corpus of documents. However, that said, there were two notable exceptions:

1) **C1** contains biodiesel related terms, which do not seem to occur in the present clustering. However, on closer inspection, we see that this is because all of the biodiesel terms originated from one publication ([Antolín et al., 2002]), and that the Web of Science entry for this paper does not have any keyword plus terms.
2) Cluster **CP1** is related to hydrogen production using Cyanobacteria, a subject which was not encountered when studying the author keywords. Again, it was discovered that these terms mostly originated from a single document ([Hansel and Lindblad, 1998]); this time, there were no author defined keywords in the Web of Science record for this document.

Next, the k-means algorithm was used to cluster these keywords and the resulting keywords listed in table II.

In general, the results obtained in this second keyword collection have been less conclusive in that it has been harder to find direct mappings between the k-means generated clusters and clusters derived from the tree diagrams.

This was partly because the keyword plus collection was a lot larger. One result of this was that there were invariably more than one clusters devoted to each research topic. Also, having more keywords also meant that there were more degrees of freedom in the clustering process, making the final result a lot more variable. A further complication was that the keyword plus collection has been divided into two sets of terms to allow the visualization trees to fit onto a single page.

Nevertheless, the results still contained a great number of very informative clusters:

1) KP14 is identical with CP1, which is associated with the production of hydrogen.
2) Clusters KP13 and KP16 are both related to solar cells and match the contents of CP3 and CP6 very closely.
3) In addition, the terms in KP16 are drawn from the field of nanotechnology, a field with a great many applications in renewable energy.
4) KP20 contains a collection of closely related keywords which are primarily related to biomass production using cellulosic

Cluster#	Keywords
KP1	Elemental Sulfur, Chalcopyrite
KP2	Grain Morphology
KP3	Values, Products, Cds, Families, Energy, System, Fuel
KP4	Fimi, Trichoderma-Reesei Qm-9414, Diesel-Power-Plant, Active-Site, Cellulases, Synergism, Glycosyl Hydrolases
KP5	Hydrogen, Nickel, Electrodes, Fuel-Cell Vehicles, Hydrogen-Production
KP6	Chemical Heat Pipe, Lime Pretreatment, Corn Stover, Pressure Cooking, Aqueous Ammonia, Lignocellulosic Materials, Recycled Percolation Process, Enzymatic-Hydrolysis, Herbaceous Biomass, Hydrogen-Peroxide, Lignin Removal
KP7	Cocombustion, Pulverized Coal
KP8	Coupled Electron-Transfer, Metal-Complexes, Water-Oxidation, Electrocatalytic Hydrogen Evolution, Photosystem-Ii, Photoproduction, Proton Reduction, Biomass Conversion Processes, Homogeneous Catalysis, Electron-Transfer, Photoinduced Electron-Transfer, Solar Furnace, Electrochemical Reduction
KP9	Composites, Infrastructure, Transport
KP10	Efficiency, Cells, Adsorption, Mechanisms, Films, Metals, Gasoline, Solar-Cells
KP11	Kinetics, Differentiation, Step Gene Replacement, Activation
KP12	16S Ribosomal-Rna, Anacystis-Nidulans, Agrobacterium-Mediated Transformation, Molecular-Genetics, Gene-Transfer, Mutagenesis
KP13	Oxidative Addition, Ruthenium Polypyridyl Complex, Nanocrystalline Semiconductor-Films, Sensitized Nanocrystalline Tio2, Fermi-Level Equilibration, Photoelectrochemical Cells, Tio2 Thin-Films, Tio2 Films, Photoelectrochemical Properties, Gel Electrolyte, Sensitizers, Titanium-Dioxide Films, Charge-Transfer Dynamics
KP14	Sp Strain Atcc-29133, Anabaena-Variabilis, Nitrogen-Fixation, Bidirectional Hydrogenase
KP15	Recombination, Cdte, Dye
KP16	Photonic Crystals, Functionalized Gold Nanoparticles, Place-Exchange-Reactions, Surface-Plasmon Resonance, Graphite Nanofibers, Walled Carbon Nanotubes
KP17	Physisorption, Monte-Carlo Simulations
KP18	Sediment, Sputtering Deposition Method, Seawater, Particles, Pores
KP19	Flash Pyrolysis, Gasification, Partial Oxidation, Ch4, Co2, Liquefaction, Fuels, Ignition, Liquid, Wheat-Straw Mixtures, Wood, Briquettes, Synthesis Gas, Devolatilization, Waste Paper, Hydrolysis, Liquids, Pyrolysis, Conversion, Combustion, Cellulose, Short-Rotation, Catalysts, Fluidized-Bed
KP20	Transgenic Poplar, Genetic-Linkage Maps, Hybrid Poplar, Rapd Markers, Light Interception, Open-Top Chambers, Canopy Structure

TABLE II

CLUSTERS GENERATED USING K-MEANS: KEYWORD PLUS DATA materials (e.g. poplar) - when compared with the hierarchical mappings, the same keywords appear to have been split between clusters CP2 and CP7, which unfortunately appear in separate trees.

5) Besides KP20, there were also a number of other clusters which were devoted to biomass. These included clusters KP4, KP6 and KP19.

IV. Discussions

This paper presented a novel use of bibliometrics techniques in the visualization of technology. It seems clear that bibliometric methods such as the ones demonstrated here will be very useful to researchers seeking a better understanding of the key patterns and trends in research and technology. On the other hand, there are still many problems which will have to be solved before such techniques can be developed into tools useable by end-users in need of "black box" technology visualization solutions. These problems include:

1) Inconsistent quality of data; data obtained from publicly available sources are often unregulated and noisy, and further underscore the need for appropriate filtering and data cleaning mechanisms.
2) Non-uniform coverage - the number of hits returned for very general or high-profile keywords such as "energy" or "efficiency" was a lot greater than for more specialized topics. This is unfortunate as it is often these topics which are of the greater interest to researchers. One way in which we hope to overcome this problem is by aggregating information from a larger variety of sources, examples of which include technical reports, patent databases and even mainstream media and blogs.
3) Inadequacy of existing data analysis tools; while - through the research presented here - we have tried to push the envelope on this front, the problems encountered when dealing with complex, high dimensional data are common to many application domains and are the subject of much ongoing research besides our own. Problems related to the overfitting of data, non-unique solutions and information loss resulting from dimensionality reduction, are all symptoms of the inherent difficulty of this problem.

That said, the methods described in this paper were only intended as an early demonstration of the proposed approach, and in spite of the above-mentioned problems, we believe that the results described here already demonstrate the potential usefulness of the methodology. However, a note of caution would be that it is still not known if it will be possible to fully automate these methods - while very interesting results were obtained, distinguishing these from the background noise was still largely a manual process.

It must also be conceded that while promising, there were also many observations which were difficult to explain. These may be viewed from a number of perspectives; on the one hand, they could be manifestations of hitherto unknown relationships or underlying correlations, and may only be understood after further analysis of these results. On the other hand, it should be realized that the Google distance is a numerical index derived from the term co-occurrence frequencies - nothing more, nothing less. Under the correct circumstances and provided that our assumptions are adequately met, it serves as a useful indicator of the similarity between keywords. Certainly, from the results obtained so far it would appear that these requirements are satisfied for at least a reasonable proportion of the time. However, under less favourable conditions, these numbers can be misleading and yield artifactual results, as has also been observed in some of the examples presented in this paper.

Acknowledgement

The research described in this paper was funded by the Masdar Institute of Science and Technology (MIST). We would like to thank MIST and the Masdar Initiative for their support of this work.

Appendix I
Renewable energy related keywords

A. *Keywords from* Kajikawa et al

combustion, coal, battery, petroleum, fuel cell, wastewater, heat pump, engine, solar cell, power system

B. *Author keywords*

biomass, CDS, CDTE, energy efficiency, gasification, global warming, least-cost energy policies, power generation, populus, qtl, renewable energy, review, sustainable farming and forestry, adsorption, alternative fuel, arabidopsis, ash deposits, bio-fuels, biodiesel, biomass, biomass-fired power boilers, carbon nanotubes, chemicals, co-firing, coal, corn stover, electricity, emissions, energy balance, energy conversion, energy economy and management, energy policy, energy sources, enzymatic digestion, fast pyrolysis, fuels, gas engines, gas storage, gasification, genome sequence, genomics, high efficiency, hydrolysis, inorganic material, investment, landfill, model plant, natural gas, poplar, pretreatment, pyrolysis, renewable energy, renewables, sugars, sunflower oil, thermal conversion, thermal processing, thin films, transesterification.

C. *Keyword Plus*

16s ribosomal-rna, activation, active-site, adsorption, agrobacterium-mediated transformation, anabaena-variabilis, anacystis-nidulans, aqueous ammonia, bidirectional hydrogenase, biomass conversion processes, briquettes, canopy structure, catalysts, cds, cdte, cells, cellulases, cellulose, ch4, chalcopyrite, charge-transfer dynamics, chemical heat pipe, co2, cocombustion, combustion, composites, conversion, corn stover, coupled electron-transfer, devolatilization, diesel-power-plant, differentiation, dye, efficiency, electrocatalytic hydrogen evolution, electrochemical reduction, electrodes, electron-transfer, elemental sulfur, energy, enzymatic-hydrolysis, families, fermi-level equilibration, films, fimi, flash pyrolysis, fluidized-bed, fuel, fuel-cell vehicles, fuels, functionalized gold nanoparticles, gasification, gasoline, gel electrolyte, gene-transfer, genetic-linkage maps, glycosyl hydrolases, grain morphology, graphite nanofibers, herbaceous biomass, homogeneous catalysis, hybrid poplar, hydrogen, hydrogen-peroxide, hydrogen-production, hydrolysis, ignition, infrastructure, kinetics, light interception, lignin removal, lignocellulosic materials, lime pretreatment, liquefaction, liquid, liquids, mechanisms, metal-complexes, metals, molecular-genetics, monte-carlo simulations, mutagenesis, nanocrystalline semiconductor-films, nickel, nitrogen-fixation, open-top chambers, oxidative addition, partial oxidation, particles, photoelectrochemical cells, photoelectrochemical properties, photoinduced electron-transfer, photonic crystals, photoproduction, photosystem-ii, physisorption, place-exchange-reactions, pores, pressure cooking, products, proton reduction, pulverized coal, pyrolysis, rapd markers, recombination, recycled percolation process, ruthenium polypyridyl complex, seawater, sediment, sensitized nanocrystalline TiO_2, sensitizers, short-rotation, solar furnace, solar-cells, sp strain atcc-29133, sputtering deposition method, step gene replacement, surface-plasmon resonance, synergism, synthesis gas, system, TiO_2 films, TiO_2 thin-films, titanium-dioxide films, transgenic poplar, transport, trichoderma-reesei qm-9414, values, walled carbon nanotubes, waste paper, water-oxidation, wheat-straw mixtures, wood.

References

[Antolín et al., 2002] Antolín, G., Tinaut, F. V., Briceño, Y., Castaño, V., Pérez, C., and Ramírez, A. I. (2002). Optimisation of biodiesel production by sunflower oil transesterification. *Bioresource Technology*, 83(2):111–114.

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

[Baek et al., 2005] Baek, N. C., Shin, U. C., and Yoon, J. H. (2005). A study on the design and analysis of a heat pump heating system using wastewater as a heat source. *Solar Energy*, 78(3):427–440.

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

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, Singapore.

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

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

[Cilibrasi and Vitanyi, 2006] Cilibrasi, R. and Vitanyi, P. (2006). Automatic extraction of meaning from the web. In *IEEE International Symp. Information Theory*.

[Cilibrasi and Vitányi, 2007] Cilibrasi, R. L. and Vitányi, P. M. B. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

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

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

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

[Elnekave, 2008] Elnekave, M. (2008). Adsorption heat pumps for providing coupled heating and cooling effects in olive oil mills. *International Journal of Energy Research*, 32(6):559–568.

[Hansel and Lindblad, 1998] Hansel, A. and Lindblad, P. (1998). Towards optimization of cyanobacteria as biotechnologically relevant producers of molecular hydrogen, a clean and renewable energy source. *Applied Microbiology and Biotechnology*, 50(2):153–160.

[Kajikawa et al., 2007] Kajikawa, Y., Yoshikawa, J., Takeda, Y., and Matsushima, K. (2007). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, In Press, Corrected Proof.

[Kim and Mee-Jean, 2007] Kim and Mee-Jean (2007). A bibliometric analysis of the effectiveness of koreas biotechnology stimulation plans, with a comparison with four other asian nations. *Scientometrics*, 72(3):371–388.

[King, 2004] King, D. A. (2004). The scientific impact of nations. *Nature*, 430(6997):311–316.

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

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

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

[Mcdowall and Eames, 2006] Mcdowall, W. and Eames, M. (2006). Forecasts, scenarios, visions, backcasts and roadmaps to the hydrogen economy: A review of the hydrogen futures literature. *Energy Policy*, 34(11):1236–1250.

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

[Porter, 2007] Porter, A. (2007). How ”tech mining” can enhance r&d management. *Research Technology Management*, 50(2):15–20.

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

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

[Saka and Igami, 2007] Saka, A. and Igami, M. (2007). Mapping modern science using co-citation analysis. In *IV ’07: Proceedings of the 11th International Conference Information Visualization*, pages 453–458, Washington, DC, USA. IEEE Computer Society.

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

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

[Van Der Heijden, 2000] Van Der Heijden, K. (2000). Scenarios and forecasting - two perspectives. *Technological forecasting and social change*, 65:31–36.

[Zhu and Porter, 2002] Zhu, D. and Porter, A. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change*, 69(5).