Semantic distances for technology landscape visualization

Wei Lee Woon
Stuart Madnick

Working Paper CISL# 2011-10

November 2011

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

Semantic distances for technology landscape visualization

Wei Lee Woon*, Stuart Madnick†

*Masdar Institute of Science and Technology,

Information Technology Program,

P.O. Box 54224, Abu Dhabi, U.A.E.

†Sloan School of Management, M.I.T.,

E62-422, Cambridge MA, 02139, U.S.A.

wwoon@masdar.ac.ae, smadnick@mit.edu

Abstract

This paper presents a novel approach to the visualization 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 actual content. In particular, we explore the use of term co-occurrence frequencies as an indicator of semantic closeness between pairs of terms. To demonstrate the utility of this approach, a number of visualizations are generated for a collection of renewable energy related keywords. As these keywords are regarded as manifestations of the associated research topics, we contend that the proposed visualizations can be interpreted as representations of the underlying technology landscape.

Index Terms

Data Mining, Technology Forecasting, Clustering, Semantic Distance

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]); 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, Woon et al., 2011] 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].

B. Visualization of technology

However, of particular relevance in the current context are studies which are centered on the visualization or clustering of research topics. There are many such studies in the literature and it would be impractical to list them all. However, in general, we have identified the following general approaches:

1) Co-citation patterns [Small, 2006, Saka and Igami, 2007, Igami, 2008, Porter and Rafols, 2009, Upham and Small, 2010], in which individual articles are visualized, where similarity between a pair of article is quantified by the number of time they are cited in the same paper.
2) Citation networks [Kajikawa et al., 2007, Takeda et al., 2009, Takeda and Kajikawa, 2009, Kajikawa and Takeda, 2008], where inter-citation between any pair of documents is used to create a link. This frequently results in very densely interconnected networks of articles.
3) Term co-occurrence statistics [Ding et al., 2001, Porter, 2005, Janssens et al., 2006, Woon and Madnick, 2009] - the use of term co-occurrence in abstracts and/or full text to generate visualizations.
4) Co-authorship [Börner et al., 2005, Glänzel and Schubert, 2005, Morel et al., 2009] - this is yet another type of bibliometrics-based visualization technique. However, in general it has a slightly different purpose, which is to study collaboration patterns or to identify highly productive authors and groups.

While some of these techniques, in particular co-citation analysis, are known to be very effective at capturing the flow of information, they suffer from a number of disadvantages. Firstly, bibliometric records, which may include abstracts, author names and bibliographies for all of the articles to be visualized have to be downloaded and processed locally (sometimes even the full-texts of articles are required, for example in the case of [Janssens et al., 2006] in which the authors used optical character recognition techniques). This can be a time-consuming and expensive process if large collections of records need to be retrieved. For example, in [Kajikawa and Takeda, 2008], the bibliographic records of 79,705 articles were collected and analyzed (in the end, the authors only used a subset consisting of 62,745 “maximally-connected” articles). An alternative would be to use smaller document collections (for e.g., in [Ding et al., 2001], where 2,012 articles were analyzed, while 938 full-text articles were used in [Janssens et al., 2006]). This will help to relieve some problems, but on the other hand would mean that ultimately, less information is taken into account. In addition, using locally or internally compiled records might also result in non-standard results as a result of variations in the way in which the records are parsed and analyzed.

Secondly, for methods that process individual articles (such co-citation and inter-citation based methods), the resulting maps can contain a very large number of nodes, which results in issues of computational complexity as well as difficulty in interpreting the maps. For example, in some cases, a further stage is required where general topics or themes still need to be assigned to the clusters after they have been constructed. There are different ways of doing this, for example in [Saka and Igami, 2007], experts are consulted, while in [Takeda et al., 2009, Kajikawa and Takeda, 2008], the cluster names are manually assigned based on the article abstracts

and titles contained in each paper). Each of these issues mean that these techniques can be computationally expensive, and scale very poorly to larger document collections.

While this brief overview may not be comprehensive, we can already make a number of pertinent observations. One is that the visualization and clustering of technological domains is certainly not a new topic; in fact it is one which has been attracting quite a lot of attention for some time now, which is indicative of the importance of the problem. At the same time, it would appear that conventional methods each have their respective advantages and disadvantages and that, while it is difficult to conceive of a single "perfect" solution, there is still scope for further development in this area. We believe that the method proposed in this paper will help to address some of the issues which still exist in this domain.

C. Contributions and scope

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 activities. 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.

D. 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 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-occurrence 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.

The Normalized Google Distance

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?

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

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:

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

where NGD stands for the *Normalized Google Distance*, t_x and t_y are the two terms to be compared, n_x and n_y are the number of results returned by a Google search for each of the terms individually and $n_{x,y}$ 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-occurrences which are within the context of renewable energy; as such, to calculate the co-occurrence frequency $n_{i,j}$ between terms t_x and t_y, the search term "'renewable+energy"+ "t_x"+"t_y"' 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-occurrence 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.

Jaccard distance

While the NGD certainly seemed to be a good choice for the purposes of this research, it is only one particular interpretation of "distance". Given the somewhat abstract nature of distances between keywords, and to try to reduce the arbitrariness of this choice, it was decided to conduct some brief experiments on one other distance measure.

The method which was chosen was the "Jaccard distance". This is basically a version of the well known Cosine similarity measure which is commonly used in information retrieval [Manning et al., 2008], and which has been specialized to deal with distances between sets of objects. It is defined as:

$$J_\delta(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}, \tag{6}$$

where X and Y are the two sets to be compared. In the context of search results, this is easily adapted by setting:

$$\begin{aligned} |X \cap Y| &= n_{x,y} \\ |X \cup Y| &= n_x + n_y - n_{x,y} \end{aligned}$$

where n_x, n_y and $n_{x,y}$ are the hit counts obtained when searching for terms t_x, t_y, and (t_x AND t_y) respectively.

Having determined appropriate measures for inter-keyword distance, 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.

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

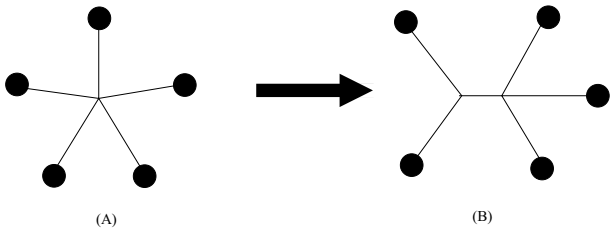


Fig. 1: Creation Of Hierarchical Tree Using The Neighour Joining Method

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*($\mathcal{T}, \mathbf{D}$)

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:

$$S_{i,j} = \frac{1}{2(|\mathcal{T}|-2)} \sum_{k=3}^{N} (d_{1,k} - d_{2,k}) + \frac{1}{2} d_{i,j} + \ldots \\ \ldots + \frac{1}{|\mathcal{T}|-2} \sum_{3 \le i \le j} d_{i,j}. \quad (7)$$

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 structures 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

[2]http://scholar.google.com

[3]http://www.scirus.org

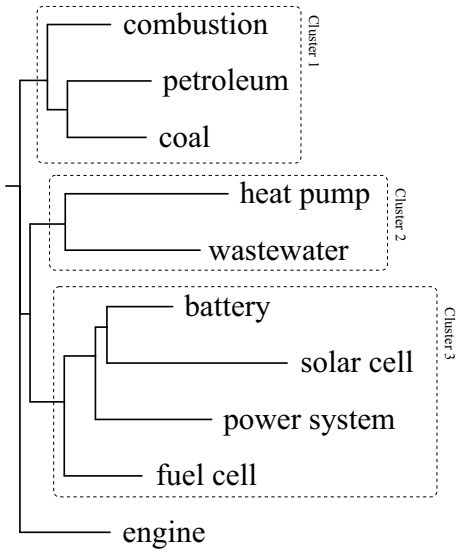


Fig. 2: Visualization tree for Kajikawa data (the three clusters referenced in the text are clearly labelled)

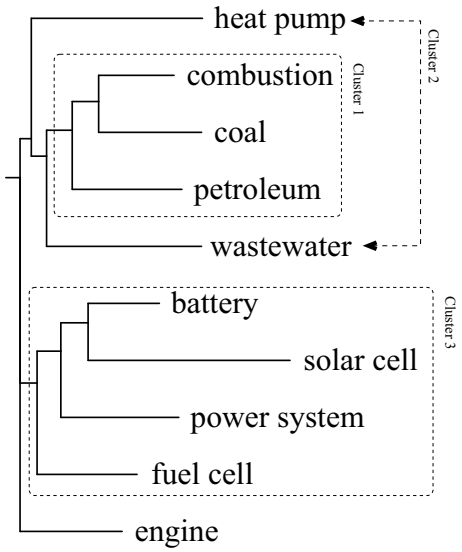


Fig. 3: Visualization tree for Kajikawa data, generated using the Scirus database (Clusters are labelled)

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 larger 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.

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 $\bigcap_{i=i}^{k} \mathcal{C}_i = \emptyset$

1. Select random centroids $t_1^* \ldots t_k^* \in \mathcal{T}$
2. $\mathbf{t} \leftarrow [t_1^*, \ldots, t_k^*]$
3. $\mathbf{c} \leftarrow [\{t_1^*\}, \ldots, \{t_k^*\}]\,(= [\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k])$
4. **repeat**
5. $\quad \mathcal{T}' \leftarrow \mathcal{T} - \{\mathbf{t}\}$
6. $\quad$ **for** $t_i \in \mathcal{T}'$
7. $\quad\quad$ **do** $l \leftarrow \text{argmin}_j \{d_{i,j} : t_j \in \{\mathbf{t}\}\}$
8. $\quad\quad\quad \mathcal{C}_l \leftarrow \mathcal{C}_l + \{t_i\}$
9. $\quad$ **for** $i \leftarrow [1, k]$
10. $\quad\quad$ **do** $j \leftarrow \text{argmin}_j \left\{\sum_{t_l \in \mathcal{C}_j} d_{j,l}\right\}$
11. $\quad\quad\quad t_i^* \leftarrow t_j$
12. **until** termination criterion met

The complexity of the algorithm is easy to determine as it is a subset of the standard k-means algorithm, where it is basically the assignment phase (no re-calculation is required since the distances between all points are pre-calculated and only need to be looked-up during execution of the algorithm). As such, the overall

complexity is $\mathcal{O}(ikn)$, where i is the number of iterations, k is the number of clusters, and n is the number of points in the collection [Manning et al., 2008]. This is a modest requirement as was borne out during our experiments.

However, we note that this is a Greedy algorithm and that hence, 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:

$$D = \min_{\{i,j:i,j\in\mathbf{c},i\neq j\}} \left\{ \frac{\mathbf{d}_{i,j}}{\max_{1\leq k\leq n} \delta_k} \right\}, \tag{8}$$

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.

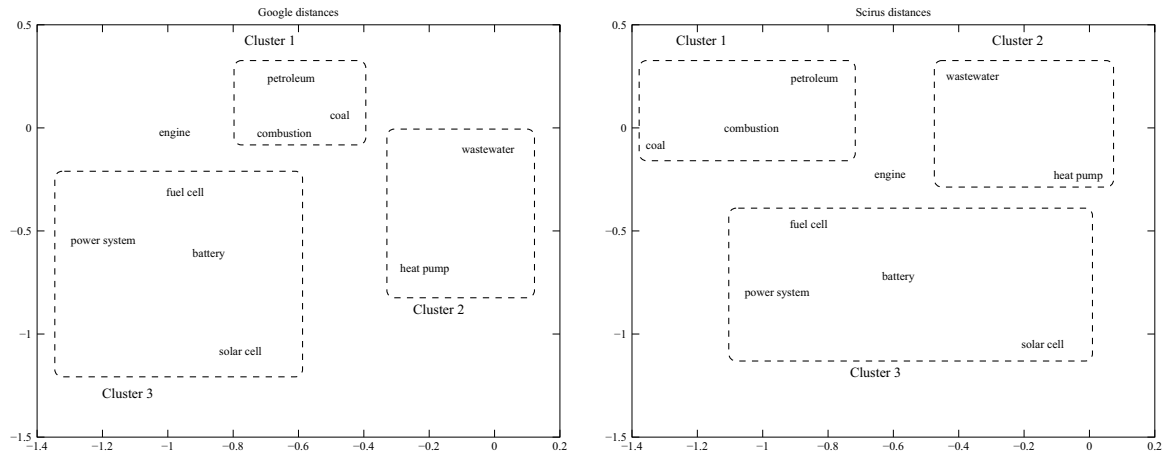
Topographic Mapping

As discussed in the preceding two subsections, the main visualization and clustering techniques used in this paper are the hierarchical clustering technique proposed in [Saitou and Nei, 1987], and the k-means algorithm. However, it is important to note that these are not the only methods which can be used for this purpose, and it is likely that other visualization techniques, for example, could be fruitfully deployed to analyze or emphasize other aspects of the data.

It is not within the scope of this paper to review all such techniques, but one class of methods which thus far have not been mentioned are topographic mapping techniques, which aim to find a low-dimensional representations of the data which preserve the topographic ordering of objects (i.e. a distance-preserving transform).

One technique which falls into this category is Sammon Mapping [Sammon, 1969], which solves this by minimizing the differences between the actual inter-object distances, and the corresponding distances in the visualization space. This quantity is represented by a nonlinear cost function known as the Sammon Stress, defined as:

$$E_{ss} = \frac{\sum_{ij} d^*_{ij} - d_{ij}}{\sum_{ij} d_{ij}}, \tag{9}$$



(a) Google distance (b) Scirus distance

Fig. 4: Topographic mapping of Kajikawa data, using both Google and Scirus keywords. As this is a visualization space, the x and y axes have no real physical interpretations, which is why they are not labeled.

where E_{ss} is the Sammon stress, d_{ij} is the distance between documents i and j, and d^*_{ij} is the distance between two corresponding points in the visualization space. The derivative of E_{ss} w.r.t. the positions of the points in the vizualisation space can be calculated, allowing the stress function to be optimized via any non-linear optimization algorithm (examples presented here were generated using the Scaled Conjugate Gradients algorithm; see [Bishop, 1995] for more details).

To demonstrate how this could be used, we used Sammon Mapping to generate low dimensional representations of the ten keywords from [Kajikawa et al., 2007], and the resulting maps are presented in Figures 4 (a) and (b) for the Google and Scirus distances respectively. Clusters 1 to 3 from the previous sections have been clearly marked on these figures and it can be seen that the inter-relationships between these keywords have been preserved in this alternative representation.

Finally, to facilitate comparisons and to provide some degree of validation of the hierarchical tree figures, annotated Sammon Maps of all the datasets have been included in Appendix B.

C. Data collection

As mentioned in section I-D, 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 A 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 highly 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 A).

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. 5.

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. 5, 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.
- **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*

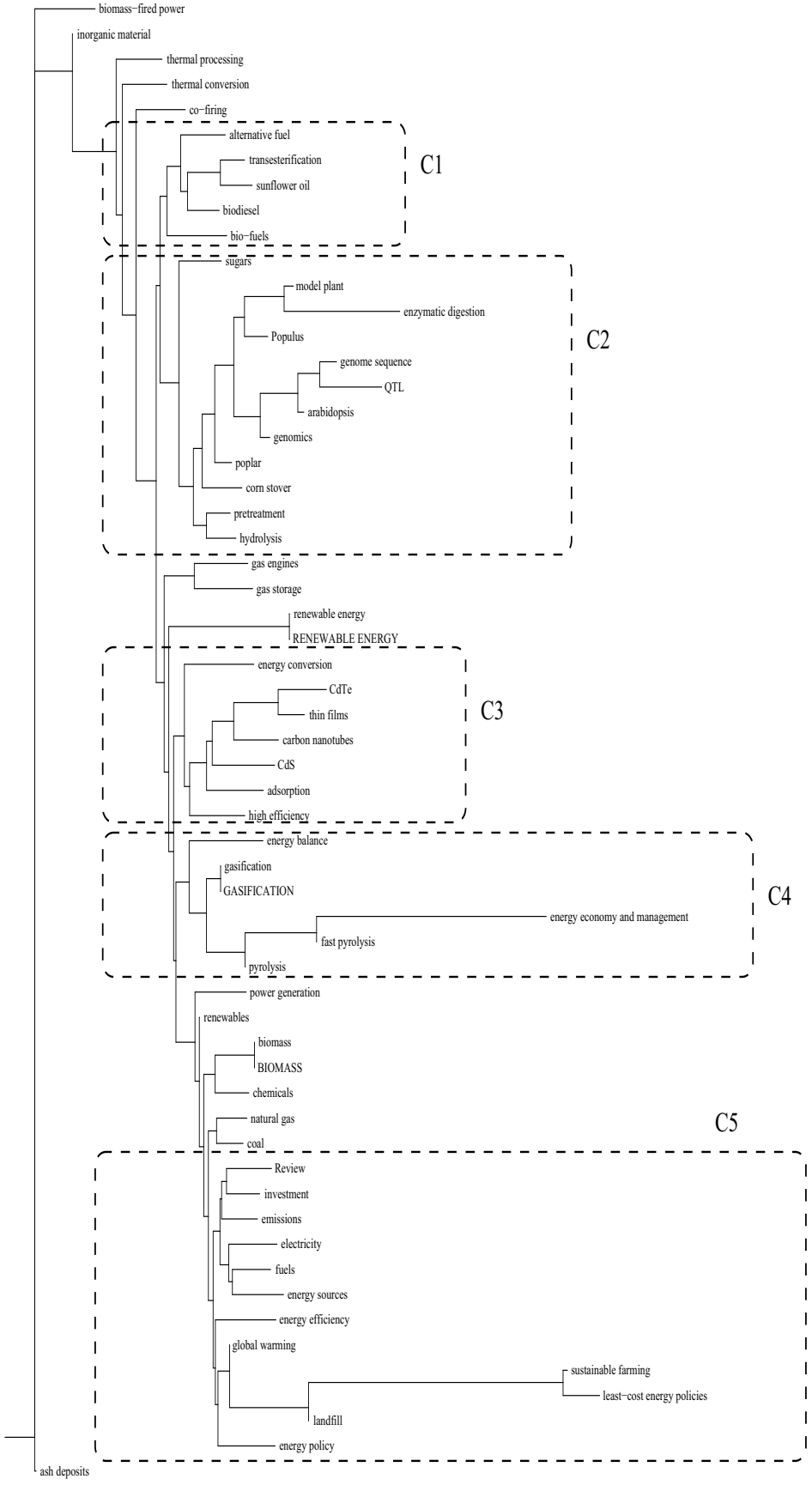


Fig. 5: Visualization tree for Author keyword data

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

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 8) 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$\rightarrow$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.

As mentioned previously, alternative visualizations were also generated using the Sammon map and the Jaccard distance. To help maintain the readability and flow of this paper, these results have been included in Appendix B rather than in the main text.

As can be seen, the Sammon Maps reveal structure which is similar to that obtained with the hierarchical visualizations (though it is quite difficult to see this clearly due to the large number of keywords). Similarly, for the k-means clusters generated using the Jaccard distances, similar themes were picked up though the exact cluster memberships and orderings are different. However, there were some closely matching clusters; in particular, cluster K6 (biomass related) is identical to cluster JK5 in table III. Other similar clusters are K7 and JK7 (related to gasification), K1 and JK4 (related to thin film PV) and K3 and JK10 (biomass/waste to energy related).

It is important to note that the aim of this exercise is not to prove that the results are *identical*, but to show that the results are not completely arbitrary and might change completely when a different distance measure is used, for instance.

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 6 and 7 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.
- **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,*

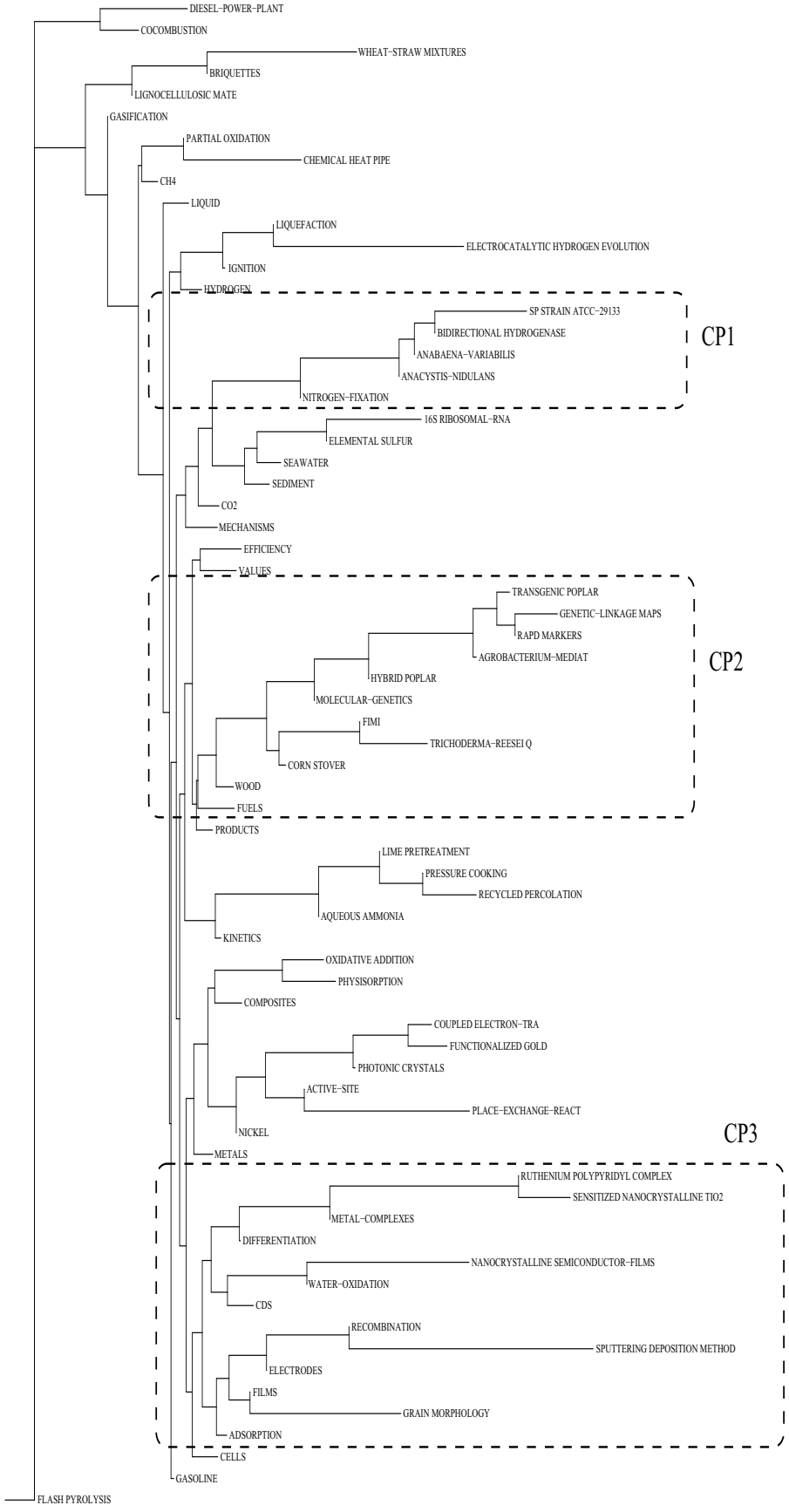


Fig. 6: Visualization tree for the keyword plus data (set 1)

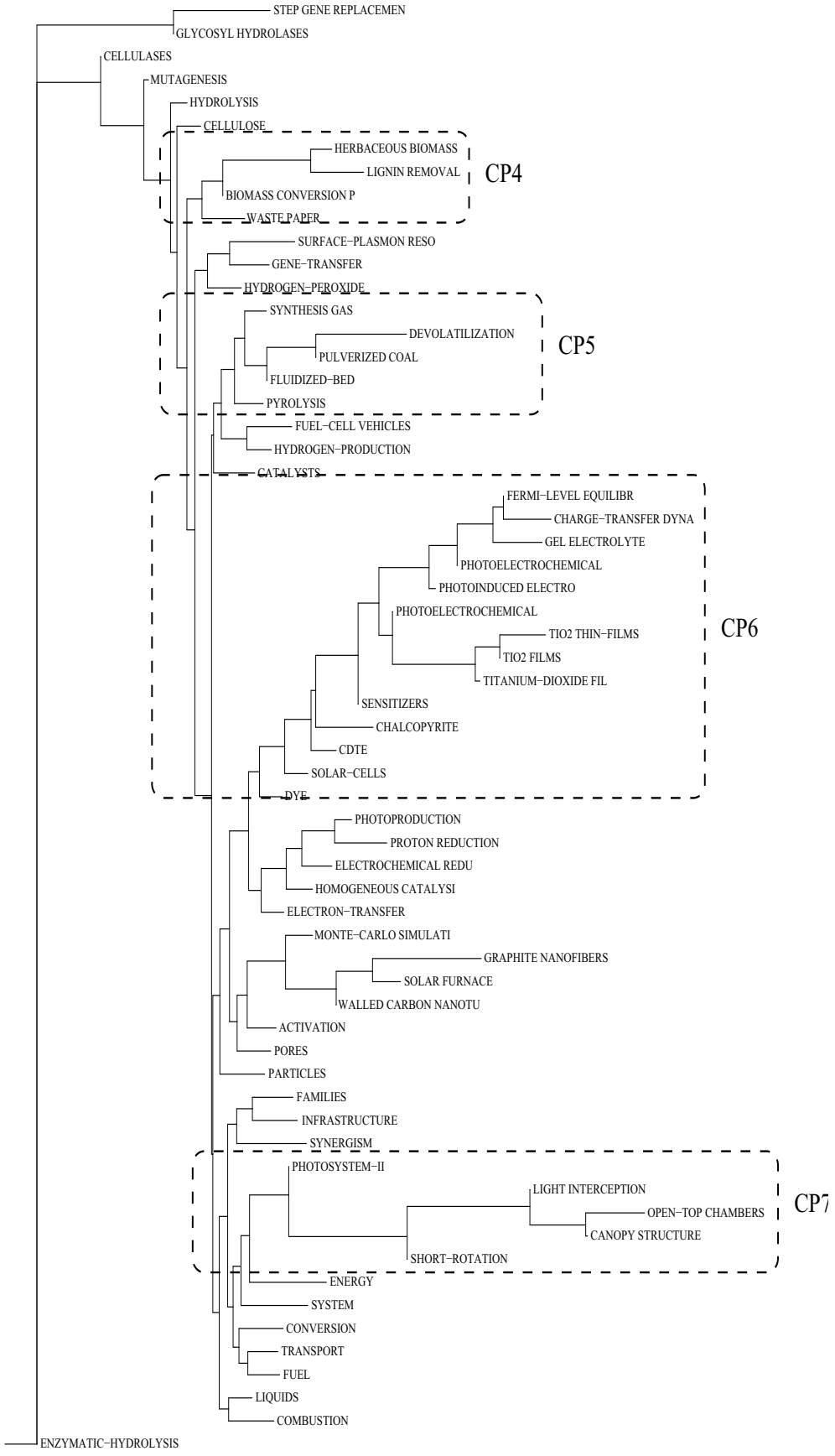


Fig. 7: Visualization tree for the keyword plus data (set 2)

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 cluster 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 nano-technology, 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 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.

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

Finally, as in the case of the Author Keywords data, alternative visualizations have been generated, and these are included in Appendix B as a comparison.

IV. Discussions

This paper presented a novel use of bibliometric techniques in the visualization of technology. It seems clear that 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 are common to many application domains which deal with complex, high dimensional data, and are the subject of much ongoing research besides our own. Problems related to the over-fitting of data, non-unique solutions and information loss resulting from dimensionality reduction, are all symptoms of the inherent difficulty of this problem.

Another important consideration is the scalability of the method. Here, we consider two main aspects, the first of which is scalability with respect to the size of the source database. We believe that our method addresses this aspect very neatly since all that is required is that the database provides a search interface. In most scenarios, such databases are curated by large organizations with very substantial resources (e.g. Google). It is reasonable to expect that these organizations would have put in place appropriate indexing mechanisms to allow the searches to be conducted efficiently. Using only search results to generate the distances shifts a large portion of the computational load to the providers of the database, which in turn allows us to leverage the optimizations or economy-of-scale advantages enjoyed by these providers. Even if locally stored databases are used, the proposed methodology could make use of the efficient search functionality built into these databases, rather than having to manually parse the contents of the entire database.

The second aspect is scalability with respect to the number of keywords to be analyzed. In the current methodology, a search needs to be performed for each pair of terms in the collection to be analyzed, as well as for each term individually. As such, for a collection with n terms, $n.(\frac{n}{2}+1)$ searches will need to be conducted. Depending on the level of availability of the database, this can be time-consuming (also, some databases may only permit a certain number of searches a day). However, as demonstrated here, reasonably sized collections containing over a hundred keywords are quite easily analyzed.

A related issue is the difficulty of updating the distance matrix dynamically when new terms need to be added. With the proposed method, adding each new term to the database will require $n+1$ searches to be

made; i.e. the term will need to be compared to all existing terms, and a search for the term in isolation will also need to be made. For moderately sized collections, this could be alright but if much larger collections are to be analyzed, the basic methodology will need to be extended to incorporate a more efficient updating procedure. This has not been investigated yet but one possibility is to group the keywords into clusters. New keywords could then be compared only to the cluster centroids and not to all existing keywords. Direct comparison with keywords within the most similar cluster could also be conducted to determine the ”local structure”. Assuming that there is a much smaller number of clusters than keywords, this will result in a huge boost in efficiency. However, such a method will need careful tuning to achieve a good trade-off between efficiency and accuracy of the approximated distances. Nevertheless, this would definitely be a very interesting direction for future research.

Finally, a number of additional visualizations were generated using alternative mapping techniques and distance measures. Specifically, the *Sammon Map* was used as an example of a topographic mapping technique, and the *Jaccard distance* was presented as an alternative measure of publication similarity. While these techniques were not comprehensively investigated, the reason for their inclusion was mainly to demonstrate that the keyword relationships observed were not merely artifacts that resulted from using a particular set of visualization techniques, but reflected the true, underlying structure of the research landscape. Here, our observation was that the overall clustering structure and inter-relationships between the keywords were preserved, but that the scaling and distribution of the keywords in the respective visualization spaces experienced some distortion. In particular, we note that the proposed combination of the NGD with the hierarchical visualization or the k-means algorithm seemed to provide results that were somewhat clearer than the alternatives mentioned above.

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 favorable 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

We would like to thank the Masdar Institute of Science and Technology (MIST) and the Masdar Initiative for their support of this work.

APPENDIX A
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.

APPENDIX B
ADDITIONAL VISUALIZATIONS

A. Author keywords

In this appendix, alternative visualizations/clusterings generated using the author-generated keywords are presented. As mentioned in section III-A, two alternative forms are presented here. The first, which is shown in fig. 8, uses Sammon Mapping to generate a topographic representation of the inter-keyword distances. Secondly, k-means clusters were generated using the Jaccard distance, which was described in section II-A. This is presented in table III. We note that:

1) As discussed in section III-A, as well as by the labeling of highly similar clusters found in fig. 8, the different visualizations were broadly in agreement with the earlier results as to the overall structure of the research landscape.
2) However, at the same time, the representations are not identical; this is not surprising since most of these methods are non-linear and would result in distortion and "stretching" of the visualization space.
3) While the results were *consistent* with the earlier results, we felt that the visualizations obtained using the Google distance followed by hierarchical clustering or k-means tended to be clearer.
4) For e.g., the Sammon map shown in fig. 8 is quite difficult to read except under very high magnification. This is a result of the Sammon mapping technique which, by definition, places similar nodes very close to each other in the visualization space; in many cases this resulted in overlapping terms, while on the other hand large sections of the visualization space remained relatively sparse.
5) Also, when using the Jaccard distance the clusters tended to be more unbalanced. For e.g. in table III, we see that many of the keywords have been lumped together in cluster JK1. This implies that the mapping induced by the Jaccard distance was not able to achieve a good uniform "spread" of the keywords.

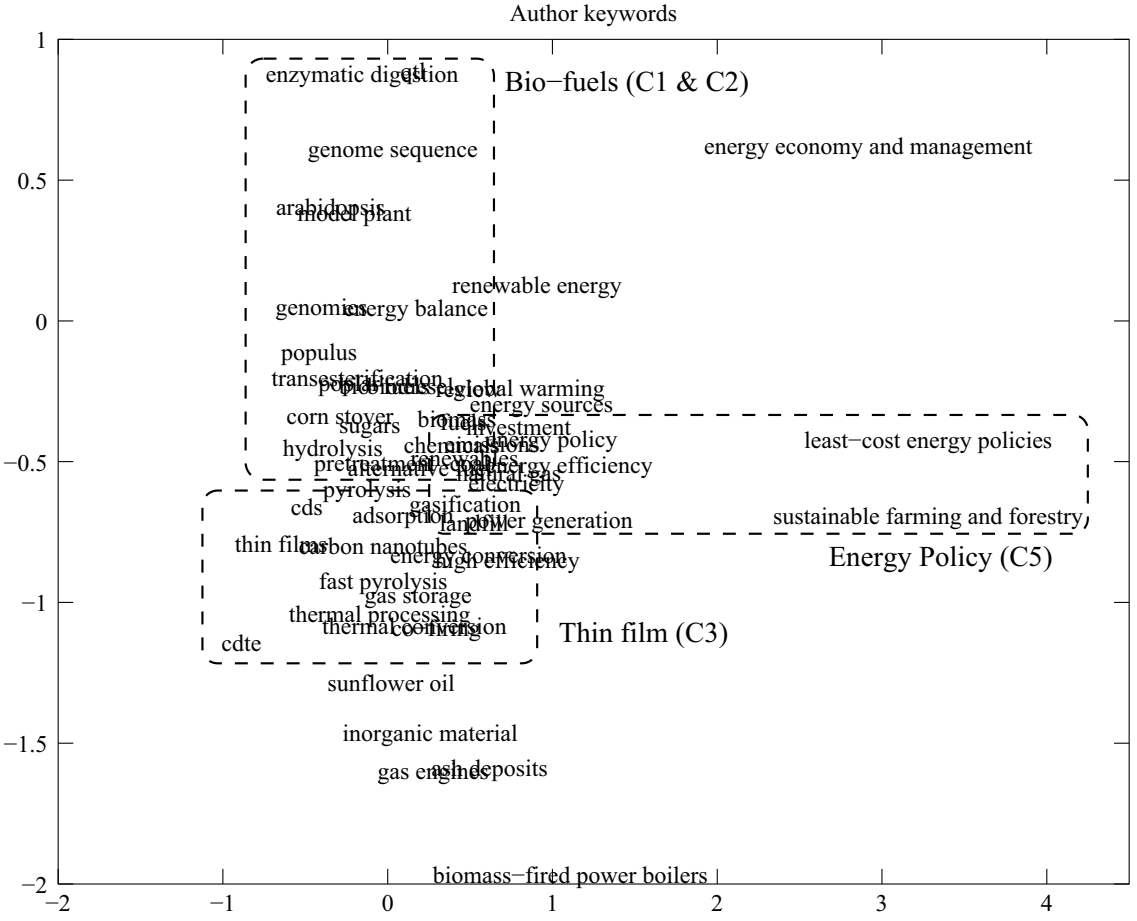


Fig. 8: Sammon Map of author keywords data. Thematic clusters have been highlighted, and where possible have been linked to clusters found in the hierarchical maps

Cluster#	Keywords
JK1	Alternative Fuel, Natural Gas, Renewable Energy, Review, Energy Balance, Energy Efficiency, Investment, Electricity, Global Warming, Renewables, Fuels, Energy Sources, Biodiesel, Sustainable Farming And Forestry, Energy Policy, Power Generation, Least-Cost Energy Policies, Biomass, Landfill, Coal, Emissions, Renewable Energy
JK2	Energy Economy And Management
JK3	Qtl
JK4	Energy Conversion, Cdte, Adsorption, High Efficiency, Chemicals, Carbon Nanotubes, Cds, Thin Films
JK5	Ash Deposits, Inorganic Material, Biomass-Fired Power Boilers
JK6	Model Plant, Genome Sequence, Arabidopsis, Genomics
JK7	Transesterification, Gas Engines, Pretreatment, Gasification, Thermal Conversion, Co-Firing, Thermal Processing, Gasification, Sunflower Oil, Pyrolysis, Fast Pyrolysis, Hydrolysis
JK8	Gas Storage
JK9	Bio-Fuels
JK10	Sugars, Enzymatic Digestion, Populus, Corn Stover, Poplar

TABLE III: Clusters generated automatically by applying the k-means algorithm to the author keywords data (Jaccard distances)

B. Keyword plus

Similar trends were observed here as with the previous sub-section. Again, the broad structure of the research landscape seems to have been preserved. As before, Sammon Mapping (shown in figs. 9 and 10) resulted in a somewhat cluttered representation of the landscape, while application of the k-means algorithm to the Jaccard distances again resulted in a less uniform distribution of keywords amongst clusters, where it can be seen that there are a significant number of clusters with only one keyword or phrase (9 such clusters were found in table IV, as compared to only 1 such cluster in table II).

Cluster#	Keywords
JKP1	Elemental Sulfur
JKP2	Ruthenium Polypyridyl Complex
JKP3	Oxidative Addition, Coupled Electron-Transfer, Metal-Complexes, Water-Oxidation, Photosystem-Ii, Proton Reduction, Electron-Transfer, Photoinduced Electron-Transfer, Charge-Transfer Dynamics
JKP4	Sensitized Nanocrystalline Tio2
JKP5	Titanium-Dioxide Films
JKP6	Photoelectrochemical Cells, Dye
JKP7	Recombination, Cells, Sputtering Deposition Method, Films, Cdte, Chalcopyrite, Solar-Cells
JKP8	Cocombustion
JKP9	Photonic Crystals, Nickel, Electrodes, Electrocatalytic Hydrogen Evolution, Fuel-Cell Vehicles, Hydrogen-Production, Graphite Nanofibers, Walled Carbon Nanotubes, Electrochemical Reduction
JKP10	Fimi, Lime Pretreatment, Trichoderma-Reesei Qm-9414, Active-Site, Pressure Cooking, Aqueous Ammonia, Recycled Percolation Process, Enzymatic-Hydrolysis, Cellulases, Step Gene Replacement, Mutagenesis, Lignin Removal, Glycosyl Hydrolases
JKP11	Flash Pyrolysis, Transgenic Poplar, Agrobacterium-Mediated Transformation, Molecular-Genetics, Corn Stover, Genetic-Linkage Maps, Hybrid Poplar, Rapd Markers, Lignocellulosic Materials, Herbaceous Biomass, Waste Paper, Light Interception, Hydrolysis, Cellulose, Short-Rotation, Biomass Conversion Processes
JKP12	Sp Strain Atcc-29133, Anacystis-Nidulans, Anabaena-Variabilis, Nitrogen-Fixation, Bidirectional Hydrogenase, Open-Top Chambers, Photoproduction, Gene-Transfer
JKP13	Chemical Heat Pipe, Solar Furnace
JKP14	16S Ribosomal-Rna
JKP15	Efficiency, Gasification, Partial Oxidation, Ch4, Sediment, Physisorption, Co2, Liquefaction, Adsorption, Values, Mechanisms, Diesel-Power-Plant, Hydrogen, Fuels, Kinetics, Products, Ignition, Differentiation, Composites, Functionalized Gold Nanoparticles, Grain Morphology, Cds, Seawater, Metals, Liquid, Wheat-Straw Mixtures, Gasoline, Wood, Briquettes, Synthesis Gas, Families, Infrastructure, Devolatilization, Pulverized Coal, Liquids, Synergism, Pyrolysis, Conversion, Combustion, Monte-Carlo Simulations, Transport, Particles, Homogeneous Catalysis, Energy, Activation, Canopy Structure, Catalysts, Pores, Fluidized-Bed, Hydrogen-Peroxide, System, Fuel
JKP16	Surface-Plasmon Resonance
JKP17	Fermi-Level Equilibration
JKP18	Nanocrystalline Semiconductor-Films, Photoelectrochemical Properties, Gel Electrolyte, Sensitizers
JKP19	Tio2 Thin-Films, Tio2 Films
JKP20	Place-Exchange-Reactions

TABLE IV: Clusters generated using K-means: keyword plus data (Jaccard distances)

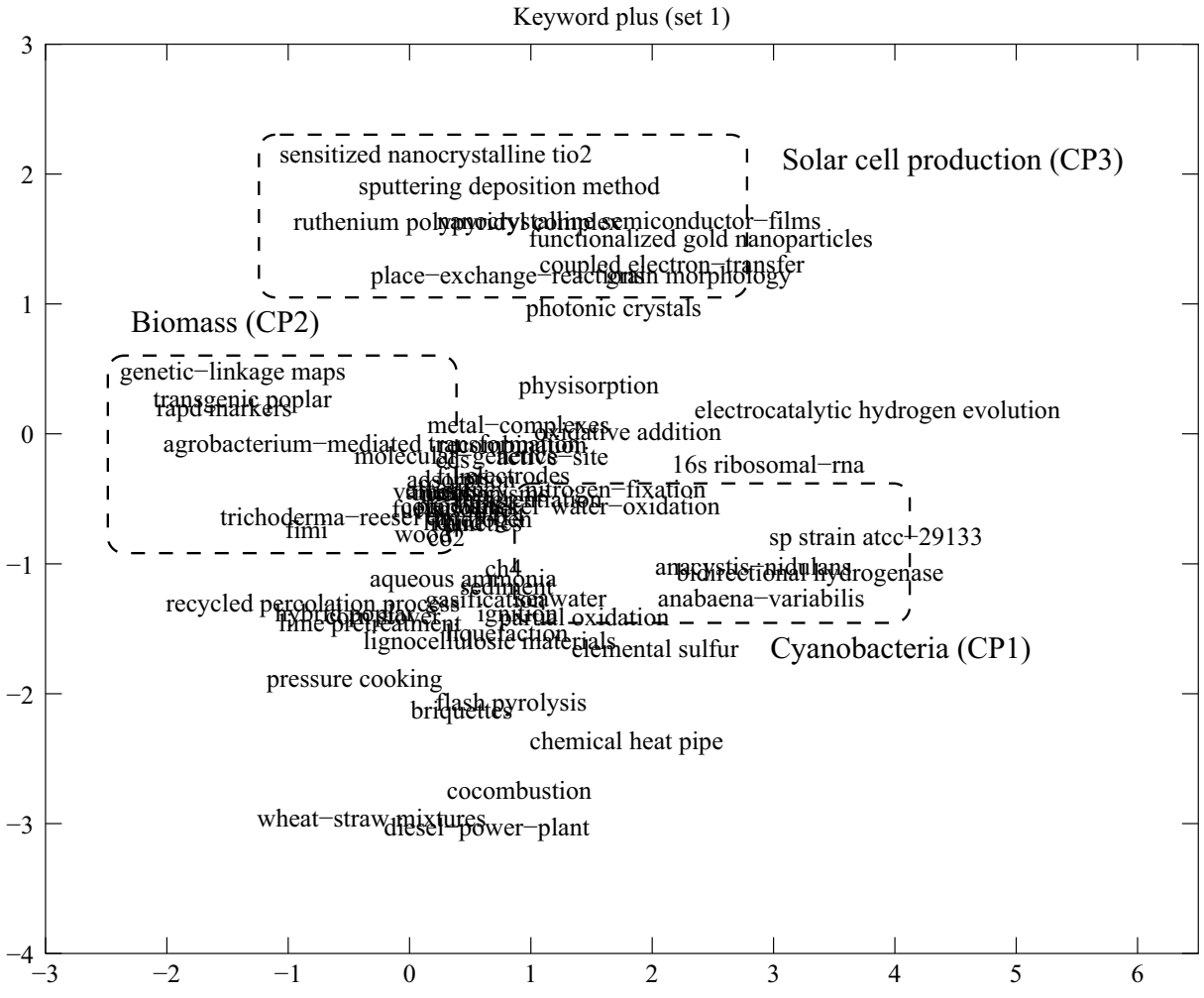


Fig. 9: Sammon Map of keyword plus terms, set 1. Thematic clusters have been highlighted, and where possible have been linked to clusters found in the hierarchical maps

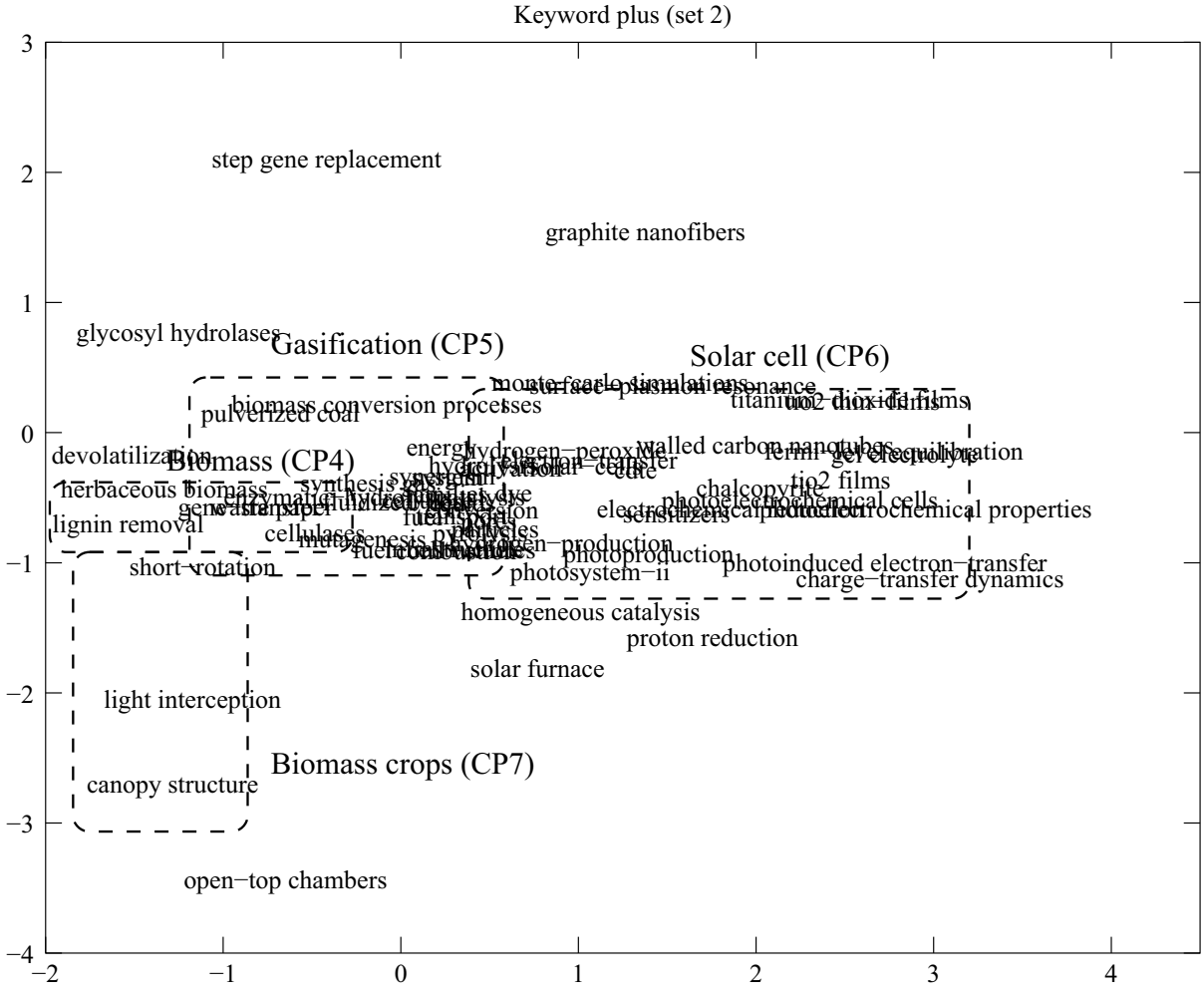


Fig. 10: Sammon Map of author keyword plus terms, set 2. Thematic clusters have been highlighted, and where possible have been linked to clusters found in the hierarchical maps

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, 1995] Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford university press.

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

[Börner et al., 2005] Börner, K., Dall'Asta, L., Ke, W., and Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):57–67.

[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 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.

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

[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.

[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.

[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.

[Ding et al., 2001] Ding, Y., Chowdhury, G., and Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management*, 37(6):817–842.

[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.

[Glänzel and Schubert, 2005] Glänzel, W. and Schubert, A. (2005). Analysing scientific networks through co-authorship. *Handbook of quantitative science and technology research*, pages 257–276.

[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.

[Igami, 2008] Igami, M. (2008). Exploration of the evolution of nanotechnology via mapping of patent applications. *Scientometrics*, 77(2):289–308.

[Janssens et al., 2006] Janssens, F., Leta, J., Gl
"anzel, W., and De Moor, B. (2006). Towards mapping library and information science. *Information processing & management*, 42(6):1614–1642.

[Kajikawa and Takeda, 2008] Kajikawa, Y. and Takeda, Y. (2008). Structure of research on biomass and bio-fuels: A citation-based approach. *Technological Forecasting and Social Change*, 75(9):1349–1359.

[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.

[Manning et al., 2008] Manning, C., Raghavan, P., Sch
"utze, H., and Corporation, E. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, UK.

[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.

[Morel et al., 2009] Morel, C., Serruya, S., Penna, G., and Guimarães, R. (2009). Co-authorship network analysis: A powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. *PLoS neglected tropical diseases*, 3(8):e501.

[Porter and Rafols, 2009] Porter, A. and Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745.

[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.

[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.

[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.

[Sammon, 1969] Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE transactions on Computers*, 100(18):401–409.

[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.

[Takeda and Kajikawa, 2009] Takeda, Y. and Kajikawa, Y. (2009). Optics: a bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics*, 78(3):543–558.

[Takeda et al., 2009] Takeda, Y., Mae, S., Kajikawa, Y., and Matsushima, K. (2009). Nanobiotechnology as an emerging research domain from nanotechnology: A bibliometric approach. *Scientometrics*, 80(1):23–38.

[Upham and Small, 2010] Upham, S. and Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, 83(1):15–38.

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

[Woon et al., 2011] Woon, W. L., Zeineldin, H., and Madnick, S. (2011). Bibliometric analysis of distributed generation. *Technological Forecasting and Social Change*, 78(3):408 – 420.

[Woon and Madnick, 2009] Woon, W. and Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, 21:91–111. 10.1007/s10115-009-0203-5.

[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).