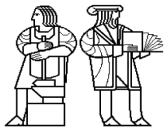


Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2009-04

TECHNOLOGY FORECASTING USING DATA MINING AND SEMANTICS: FIRST ANNUAL REPORT

Wei Lee Woon[1], Stuart Madnick[2], Ayse Firat[3], Blaine Ziegler[4], and Satwik Seshasai[5]

[1]Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
wlwoon@gmail.com

[2]Sloan School of Management and Engineering Systems Division, MIT
smadnick@mit.edu

[3]Doctoral Candidate, Engineering Systems, MIT
ayshe@mit.edu

[4]Doctoral Candidate, Electrical Eng. & Computer Sci., MIT
bziegler@mit.edu

[5]Doctoral Candidate, Engineering Systems, MIT
satwik@mit.edu

April 2009

Technology Forecasting Using Data Mining and Semantics

MIT/MIST Collaborative Research Progress Report for Period 4/1/07 to 9/30/08

Principal Investigator at MIT

Professor Stuart Madnick

Principal Investigator at MIST

Dr. Wei Lee Woon

Collaborators/Team-members

Satwik Seshasai (Ph.D Candidate/Collaborator)

Ayse Firat (M.Sc Candidate/Research Assistant)

Blaine Ziegler (M.Eng Candidate/Research Assistant)

Research Project Start Date

10/01/2007

EXECUTIVE SUMMARY

The planning and management of research and development is a challenging process which is compounded by the large amounts of information which is available. Research managers and other decision makers often rely on intuition and domain knowledge to arrive at management decisions. The overall long-term goal of this project is to mine science and technology databases for patterns and trends which facilitate the formation of research strategies. Examples of the types of information sources which we hope to exploit are very diverse and include academic journals, patents, blogs and news stories, while the intended outputs of the project include growth forecasts for various technological sectors (but with an emphasis on sustainable energy), an improved understanding of the underlying research landscape, as well as the identification of influential researchers or research groups.

In this second phase of the project our main focus was on developing methods for visualizing and organizing the structure of technology landscapes. In particular, the direction of the project during the past five months has centered on the development of techniques to both organize and visualize the data in a way which reflects the semantic relationships between keywords. We studied the use of the *joint term frequencies* of pairs of keywords, as a means of characterizing this semantic relationship – this is based on the intuition that terms which frequently appear together are more likely to be closely related.

In more detail, for the five month period following the submission of the first progress report, the focus of our research has been on:

- The development of methods and techniques for **visualizing** and **organizing** the structure of the "research landscape" underlying observed publication patterns. In particular, the methodologies which we have devised are aimed at:
    1. Providing researchers and planners with a clearer view of the inter-relationships between the many components of the renewable energy research domain.
    2. Supporting the creation of semantically enhanced feature-sets through which more reliable technology growth forecasts may be derived.
    3. Facilitating the renewable energy case study.
- The creation of **computer programs/tools** to extract and analyze online data (though development will continue throughout the project). These tools provide much of the plumbing through which data is obtained and studied. In particular, we have started investigating a number of modalities through which the results of our research may be made more accessible to end users.

To test and fine-tune the performance of programs and tools created, systematic procedures for data collection and retrieval have been established. Consistent with the aims of the Masdar Initiative, our experiments have largely dealt with themes and topics relevant to renewable energy and sustainability.

Initial results indicate that:

- Using appropriate tools and methods, exploitable patterns and information can certainly be extracted from publicly available databases.
- Adaptation of the Normalized Google Distance (NGD) formalism can provide measures of keyword distances that facilitate keyword clustering and hierarchical visualization.
- Further adaptation of the NGD formalism can be used to provide an asymmetric measure of keyword distances to allow the automatic creation of a keyword taxonomy.
- Adaptation of the Latent Semantic Approach (LSA) can be used to identify concepts underlying collections of keywords.

A series of research papers (both for publication and for internal use) have been produced. Research directions for the immediate future will evolve around translating the findings of the present research efforts into improved techniques and features for monitoring and forecasting the growth of technologies.

Report overview

The rest of this report will be structured as follows.

The ***Introduction*** section reviews the key ideas and motivations for the project. In addition, the key goals as well as the high-level approach taken are described here.

The ***Research Tasks*** section comprises the bulk of the report, and is organized into several subsections based on the original research components outlined in the project proposal. As will be seen, during the current reporting period, six such components were deemed to be "active", and the activities corresponding to each component are described in detail. Briefly, this section contains the following subsections (two of the tasks have been combined into a single subsection, hence there are only five of them):

- *Survey of databases* – an ongoing review of sources of technology-related information which may be utilized by our methods. This activity also encompasses the literature review process, which is also viewed as an ongoing activity.
- *Tool development* – the creation and extension of software tools which both facilitate project-related research, and which may also be regarded as project deliverables in their own right. Two main classes of software are discussed in this report: programs for facilitating the data collection and management process, and programs which allow non-technical user to access and exploit this data. As will be described in greater details, significant progress has been made on both fronts.
- *Base indicators* – in data mining terminology, these are the "features" which summarize key trends and patterns in the data being analyzed. As will be explained in greater detail, progress on this activity is currently on hold.

- *Base/Enhanced visualizations* – this is actually a combination of two closely-related activites from the project proposal → "base visualizations" and "enhanced visualizations". These are techniques for representing and displaying complex data sets in such a way as to facilitate human comprehension. For the current reporting period we have focussed our efforts on detecting and organizing clusters of keywords related to renewable energy.
- *Contextual/semantic enhancements* – the incorporation of semantic and contextual enhancements to existing features and visualizations. Two approaches are proposed: one is the use of an asymmetric distance measure to infer the underlying class-inheritance structure of a collection of keywords, while the other approach is based on linear-algebraic methods to extract technology "concepts", which reflect underlying research trends and themes.

The ***Current Reporting Period Summary*** section reviews and discusses the findings presented in this report. Also included in this section is a description of a number of complementary activities which were undertaken.

Finally, there is a section on ***Future Work***, which looks ahead to the research directions in the following reporting period.

INTRODUCTION

Background

The planning and management of research and development is a challenging process which is compounded by the large amounts of information which is available. Research managers and other decision makers often rely on intuition and domain knowledge to arrive at management decisions. For example, peer review is still the primary mechanism for dreciding NSF and NIH grant awards [Porter, 07], while many countries spend huge sums on technology foresight programs [Eto, 03][Bengisu and Nekhili, 2006]. Expert opinion is a hugely important component in the decision making process; however when used on its own, it can have a number of shortcomings. In particular, expert decisions are subjective and can be influenced by personal perspectives or biases. In addition, it is difficult to systematically record the reasons for such decisions, or the contexts in which decisions were made. Finally, it can also be difficult and expensive to obtain the help of suitably qualified experts.

These issues motivate the development of tools and techniques for conducting "technology-mining" [Porter, 2007][Porter, 2005]. This is loosely defined as the application of computational tools for collecting empirical information from R&D information resources, and subsequently using this information to enrich R&D decision making. Two aspects of tech-mining are of particular interest: the prediction of future technological developments [Smalheiser, 2001], [Daim et al., 2005], [Daim et al., 2006], [Small, 06] and the visualization of the technology "landscape" [Porter, 2005], [Small, 2006].

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, 2001], 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].

Rather than being alternative approaches, these five aspects are complementary, and a comprehensive technology forecasting effort should incorporate all of the elements above. While the activities falling within the scope of this project are fairly broad, the emphasis during the current reporting period 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.

Objective and Approach

The high-level aim of the project is to create improved methods for conducting so-called “tech-mining” - i.e.: a combination of technology related activities which includes forecasting, mapping and visualization (this is defined in greater detail in Section 1.3 of the project proposal).

The general approach and methodologies adopted in this project are guided by the following principles:

- To adopt a *data-driven* approach to understanding the evolution of technology. This means that model driven techniques will not be used, even though these have also proved to be very useful. An alternative view is that data-driven methods operate on a different level from, rather than as

an alternative to, causative models. A more appropriate perspective is that the techniques developed in this project could eventually serve as inputs to later stages which could certainly including various modelling activities.

- The use of *bibliometric* techniques as a means of deriving empirical information regarding the state of technological development. These are methods which emphasize publishing patterns and trends over the actual content of the publications.
- As far as possible, to adopt methods which are *generalizable* to a variety of databases – in particular, we seek to avoid techniques which are customized to the particular capabilities of any single database or information resource.

RESEARCH TASKS

In the project schedule, the following key activities were listed (for future reference these have been labelled T1→T7):

(T1) Survey of databases

(T2) Tool development

(T3) Base indicators

(T4) Base visualizations

(T5) Enhanced visualizations

(T6) Contextual/semantic extensions

(T7) Data analysis

For the period covered by this report, six of the seven tasks above (in bold), are active for at least part of the current reporting period, while those that are underlined are active for the full duration, and are the focus of this report. The diagram below provides a high-level view of the interrelationships between the different components in the project, and links each component to the relevant task from the list above.

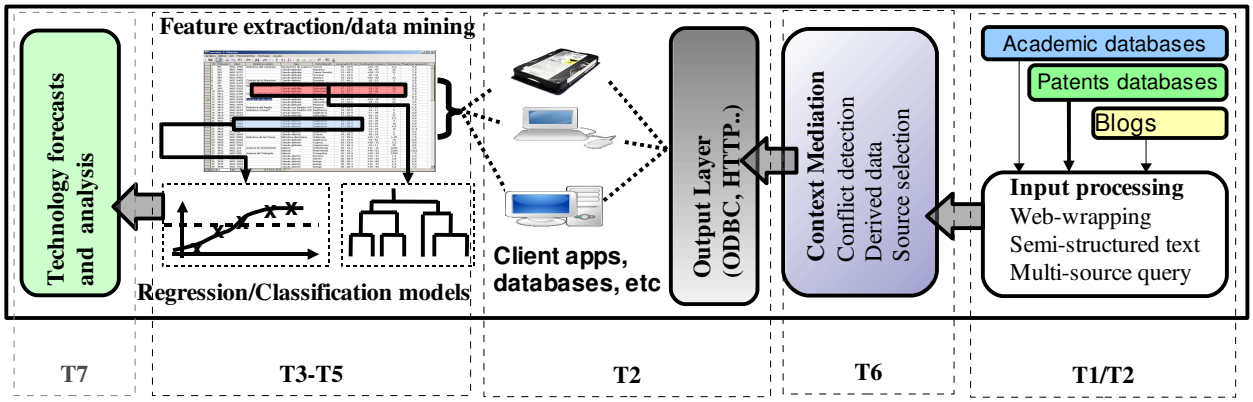


Figure 1.System flow for information extraction from technology databases

Each active item will now be reviewed, and the activities during the current period will be described and discussed.

T1: Survey of databases and approaches to technology forecasting

As indicated before, there is no provision for a formal literature review in the plan, so this component includes not only the survey of text databases and other sources of information, but also includes all the required literature and tool surveys which were carried out as part of this project. Our activities in this regard have largely been covered in the previous progress report but this activity is also regarded as an ongoing "background process" which will continue to receive attention throughout the duration of the project.

During the current five month period, we have added a number of wrappers to those available before. For example, wrappers for the following databases were added to the data collection system:

1. Web of Science
2. Compendex
3. Engineering Village
4. Google blogsearch
5. Free Patents online search
6. Scirus Patents search
7. Google News archive search

In addition, the project team has had two new members and this has allowed us to initiate a much more thorough survey of the field of Technology Mining than we have been able to attempt in the past. This

survey is available as working paper CISL #2008-15 (please refer to the section "Publications/Presentations").

T2: Tools development

In the current reporting period, significant progress was made under this category – in particular, a number of tools were built which allowed for a more user-friendly interface to the various tools and scripts developed as part of this project, and to support the collection and reliable storage of the data used in this project. We will now describe the main development activities which are currently ongoing.

Automated data collection and storage

Building on some of the work initiated in the first reporting period, we have been developing tools for facilitating the collection of bibliometric data from much larger collections of keywords – currently the largest collection which we have been working on contains 155 keywords, requiring a total of 23870 joint frequency terms in all (more on this later).

In the earlier implementation, data collection was performed using ad-hoc python scripts and the data was stored in the native python file formats. While this approach was sufficient for the initial investigations, they were unsuitable for more extensive data collection activities for a number of reasons:

1. As data was essentially stored in large matrices with annotation restricted to the file names, keeping track of data which had been collected got progressively more difficult as larger sets of keywords were introduced.
2. It was desirable to store meta-data regarding the stored information so that data cleaning scripts could be designed in the future for checking data which could be out-of-date or which had been wrapped using older collection protocols.
3. Previously, it was not possible to collect the data in an incremental fashion as all searches had to be conducted in a single run. To avoid exceeding limits imposed by certain search engines, as well as to guard against dropped connections, it was necessary to stagger the processing of large keyword collections so that it runs over multiple days. Further, in this way it was possible for additional keywords to be added in phases.
4. It was also helpful to be able to conduct multiple searches at the same time, as in general the main bottleneck in the data collection process was the response time of the search engine and the latency in the network connection.

To address these issues, a more robust data collection system was developed which included the following features:

1. Storage was based on a relational-database engine, which allows for more reliable and transparent storage of data via a set of named fields. In addition, metadata like the extraction dates and the

exact search terms used could be stored together with the data; it was hoped that storing this information would allow much larger collections of data to be effectively managed. In addition, using a proper database engine introduces additional benefits such as atomic data writes and accesses (necessary for multi-threaded access to the database, for example). For the present implementation we used the SQLite database engine as this is a very lightweight system, yet is completely sufficient for our current needs. However, generic database API calls[1] were used so upgrading to a "heavier" database engine in the future should be straightforward.

2. A distributed, multi-threaded mechanism has also been implemented which would allow for multiple searches to be conducted. In addition, communications was done using a combination of XML-RPC web-services and SOCKS proxies, so that web searches could be directed through a number of machines at the same time. This system allowed searches to be conducted at a much higher rate. Also, allowing searches to be distributed over a number of hosts would hopefully decrease the likelihood of violating limits set by search engines.

Hit Aggregator

This was a simple interface which allows non-technical users to automate and view the results of web searches to online academic databases such as Web of Science and Google scholar. Queries can be specified in a number of ways including as SQL queries and as values passed into an excel spreadsheet (a screenshot of the hit aggregator in action is provided in figure 2, below).

In its present form, this "Hit Aggregator" provides a userfriendly interface to a subset of the web-wrapping operations which we have been using throughout the project and is primarily a proof-of-concept design; however, it is hoped that this will provide the basis for a more comprehensive tool which may be used by non-technical users to conduct a variety of technology-mining activities.

1 The Python DB-API standard was used [Kuchling, 98]

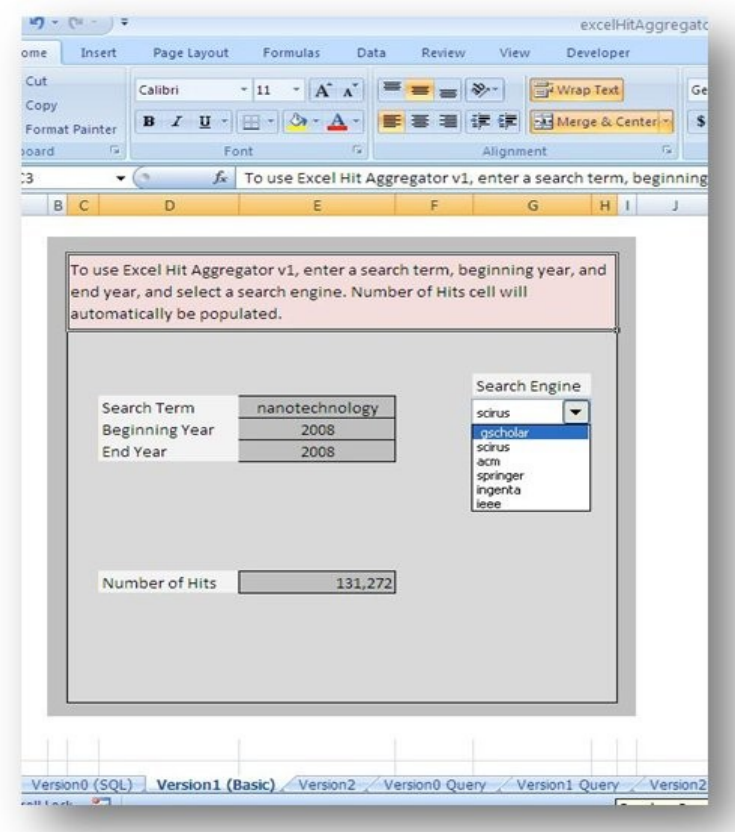


Figure 2. Hit aggregator (Excel) interface

PyCameleon

Cameleon is the technology developed in the COIN group for facilitating the extraction of data from heterogeneous web sources, allowing web data to be treated as ordinary relational data sources with some capability restrictions. One of the main advantages of the Cameleon system is that it allows all information regarding the specifics of the extraction process, such as the URLs to be visited, regular expressions and other extraction rules, to be separated from the actual implementation of the wrapper itself (whether in Perl, Java, or C#). This is achieved by expressing the former in a specification file (spec file), which is written in XML.

While the current data collection activities in the project are conducted via customized scripts written in Python, we believe that the Cameleon system will be critical for ensuring continued progress as it will allow future users of the system to incorporate support for additional databases and other sources without having to modify the underlying software. As such, one of the activities of the current reporting period has been the creation of a Python port of the Cameleon web-wrapping engine, which we will call "PyCameleon". This is a ground up implementation of Cameleon which accepts the same specification files (presently compatibility is not 100% yet but that will be the target).

The motivations for creating another port of the system are twofold. Firstly, in addition to web wrapping, we hope that the system will eventually be expanded to allow the use of a variety of other sources of data, some of which may reside on databases or which could be accessed programmatically. Creating a Python port of the wrapping engine will provide the project team with greater flexibility in experimenting and expanding the Cameleon system, particularly as most of the other software components of the project are currently implemented in Python. A second reason was to leverage all the benefits of better cross-platform compatibility as indicated above. While we have also successfully deployed the existing Cameleon implementation (Cameleon#) on a linux platform using Mono, it is felt that Python provides better cross platform support as it was designed from the start to work seamlessly across a number of operating systems and architectures. Mono, on the other hand, is a third-party .Net implementation with no formal support from Microsoft and which is known to only support a subset of the .Net libraries.

The figure below shows a screen capture of the web-interface to PyCameleon.

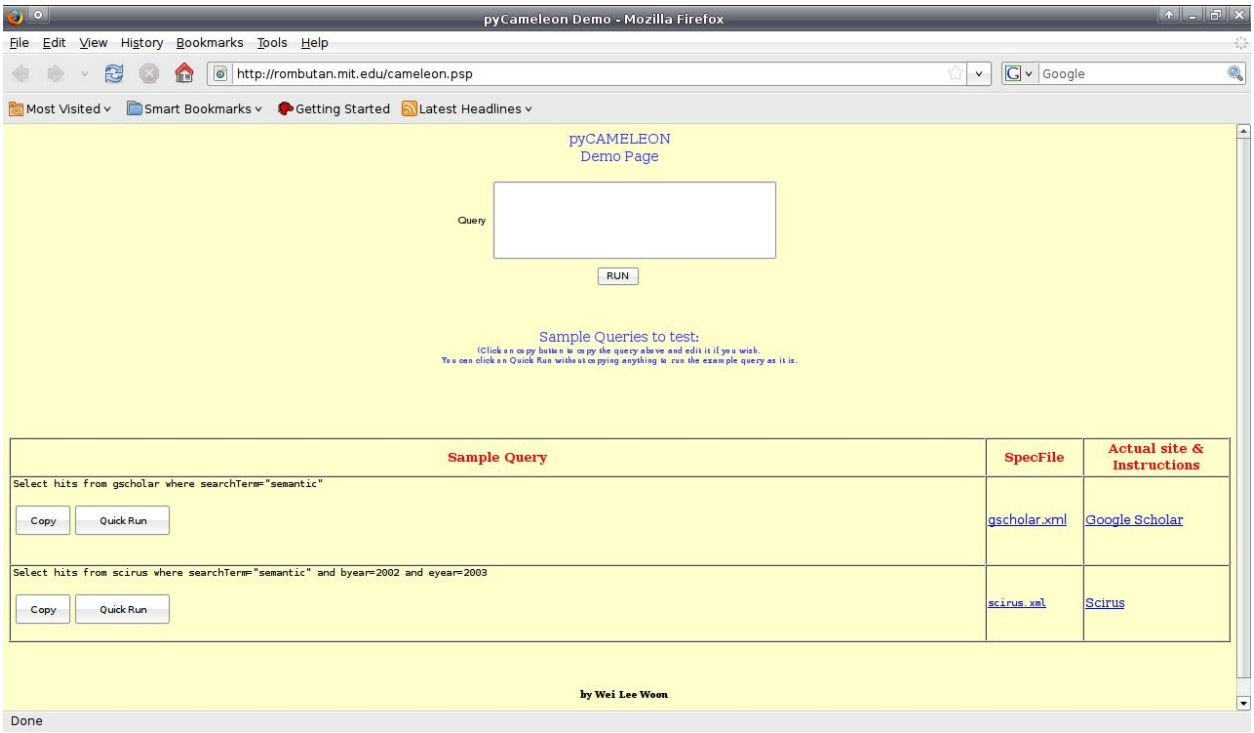


Figure 3. Screenshot of PyCameleon demo page – modeled after a simple search engine styled page but which accepts SQL queries.

T3: Base indicators

The base indicators in this case refer both to the bibliometric features use to monitor the growth of technology. In this project, this has been achieved through term frequencies of keywords of interest, where the term frequencies act as a proxy for studying the level of interest shown by the academic community in a particular topic. This is because academic publications in general are the primary means by which new discoveries and theories are presented to the broader scientific community – as such, the generation of publications and technical papers is in a sense a manifestation of the creation of knowledge by the scientific community.

However, using these term frequencies by themselves are unlikely to produce reliable results as the number of publications containing a particular term is affected by a variety of factors. The above reasoning is only reasonable if we are able to obtain an accurate estimate of the number of papers which address the topic of interest. In practice we are of course unable to do so directly; using the "hit count" returned by a search to an academic database is an attempt estimate this figure.

Also there is tendency for terms which are extremely generic to return a large number of hits. For example, a search for "fossil fuel" would clearly return a far greater number of publications than a search for "trichoderma reesei" would. This is because "fossil fuel" is an extremely broad concept which will no doubt occur in a large number of publications involving renewable energy. This problem is partially addressed in the later section on "contextual/semantic extensions", but it is nonetheless a difficult one to resolve fully.

In the previous reporting period, we had demonstrated how the sequence of publication counts for a number of consecutive years produced a time series which could be modelled by growth functions such as the logistic and Gompertz curves. However, further examination of the curvature parameters of a number of these growth functions, as well as over a number of different databases, showed that results obtained using different conditions were often inconsistent; this indicated that the use of raw keyword counts as an indicator was not fully capturing the underlying research themes and concepts.

Our response to this problem has been to try and move away from the use of raw hit counts as an indicator of technological progress. The perspective taken is that the "true" state of technology is a latent variable that is not directly observable. While term frequency is closely related, it appears that it is too noisy and unreliable in its raw state and some form of more advanced feature extraction procedure is required to extract the desired information.

One way in which this can be approached is to use the combined results from groups of related keywords rather than individual terms, in the hope that some of these errors will be removed when taking the average of a large number of observations. The research efforts in the visualization and semantic extensions sections are intended to address this problem, amongst other things. The following section will focus on the former.

T4, T5: Data visualization

Background and motivations

This section combines the base and enhanced visualization components of the project plan, as the latter is largely a continuation of the former. While there were some preliminary attempts at visualization during the previous reporting period, this is one of the main research activities for the current period. In particular, much effort has been dedicate to extending the techniques to larger and more complex data sets.

The main motivations for seeking visualization techniques are twofold. Firstly, being able to visualize and organize the data will help to increase the accuracy and reliability of the base indicators discussed above by allowing relevant terms to be grouped together such that underlying research concepts and trends are combined in a more natural manner. Secondly, effective visualization techniques are useful in their own right as they can provide researchers and managers with a useful and intuitive tool for understanding the technological "landscape".

In the following subsections, the methods used for both data collection and analysis will be discussed in some detail. Bearing in mind that the idea was to use the semantic distances between collections of keywords to construct intuitive visualizations of the underlying research landscape. 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 measure to form intuitive representations and to organize or cluster the data.

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 Vitanyi, 2007], and is defined as:

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

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:

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

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].

To adapt the NGD 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 a means of representing the size of the body of literature dealing with renewable energy technologies. We acknowledge that this is not comprehensive as many relevant studies may not explicitly mention "renewable energy", or may use equivalent terms such as "sustainable energy" or "green energy". Smaller scale experiments using such alternative terms will be conducted in the future to confirm the consistency of the results.

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_i and t_j, the search term "renewable+energy"+ t_i+ t_j was submitted to the search engine. Again, the limitations and issues described above will apply.

As explained in [Cilibrasi and Vitanyi, 06], 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**:

$$D = \begin{bmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & d_{n,n} \end{bmatrix} \qquad (3)$$

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. We have attempted a number of different approaches, and some have been touched on in the previous progress report. However, to provide a suitable context we discuss two in particular in this report: **hierarchical visualization** and **clustering**.

Hierarchical representation

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, as defined by the tree with the smallest total branch length.

To demonstrate the basic feasibility of the approach, we briefly revisit the example first presented in the previou progress report, though a much more extensive set of results will be described in a later section. For the initial demonstration, we had used 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 figure 4.

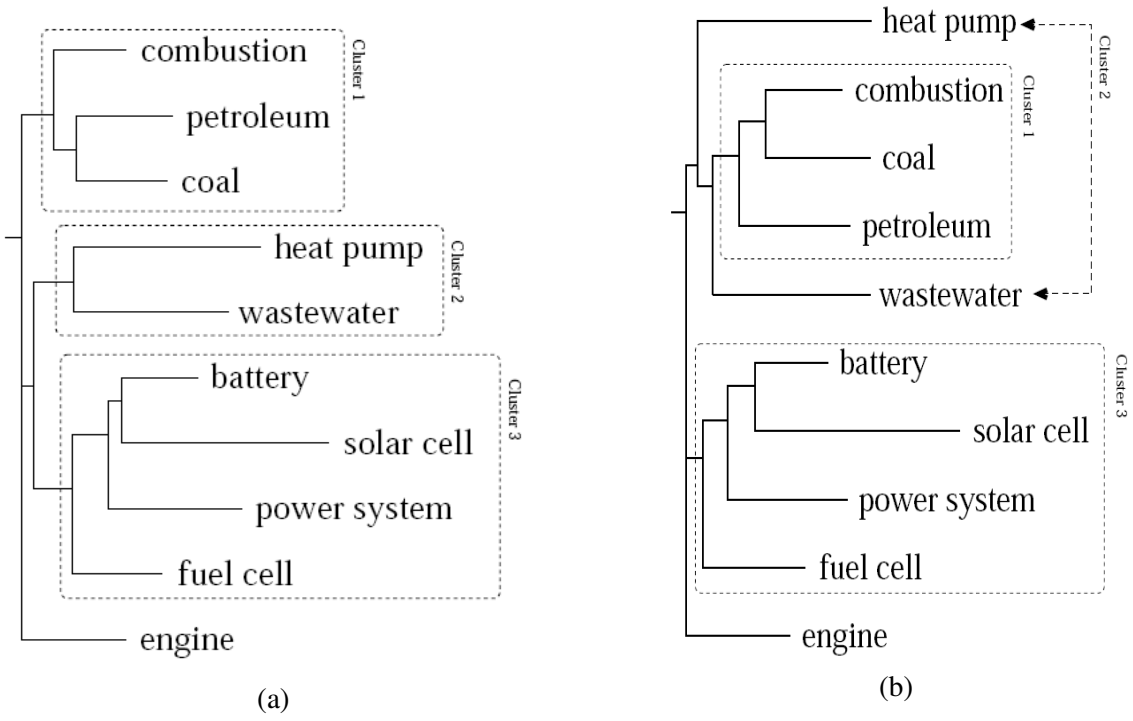


Figure 4. Visualization trees generated using Google scholar and Scirus statistics

Even in this very simple example, we 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 figure 4(a) while in figure 4(b) *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.

2 http://scholar.google.com

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, which we hope to use in the construction of enriched keywords queries. One of the most common methods is the *K*-means algorithm [Bishop, 2006], which works by partitioning 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.

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. The k-means algorithm was applied to the ten keywords extracted from [Kajikawa et al, 2007]. Again, the Google and Scirus distances were generated 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 and , 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.

Data collection

To support the renewable energy case study, a set of energy related keywords and a populated distance matrix containing the inter-keyword distances was required.

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 **B** 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 were calculated as discussed. Only hit counts from Google scholar were used this time - the Scirus search engine was also used but the results were less reliable 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 even lower 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.

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.

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 figure 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 figure 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 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.

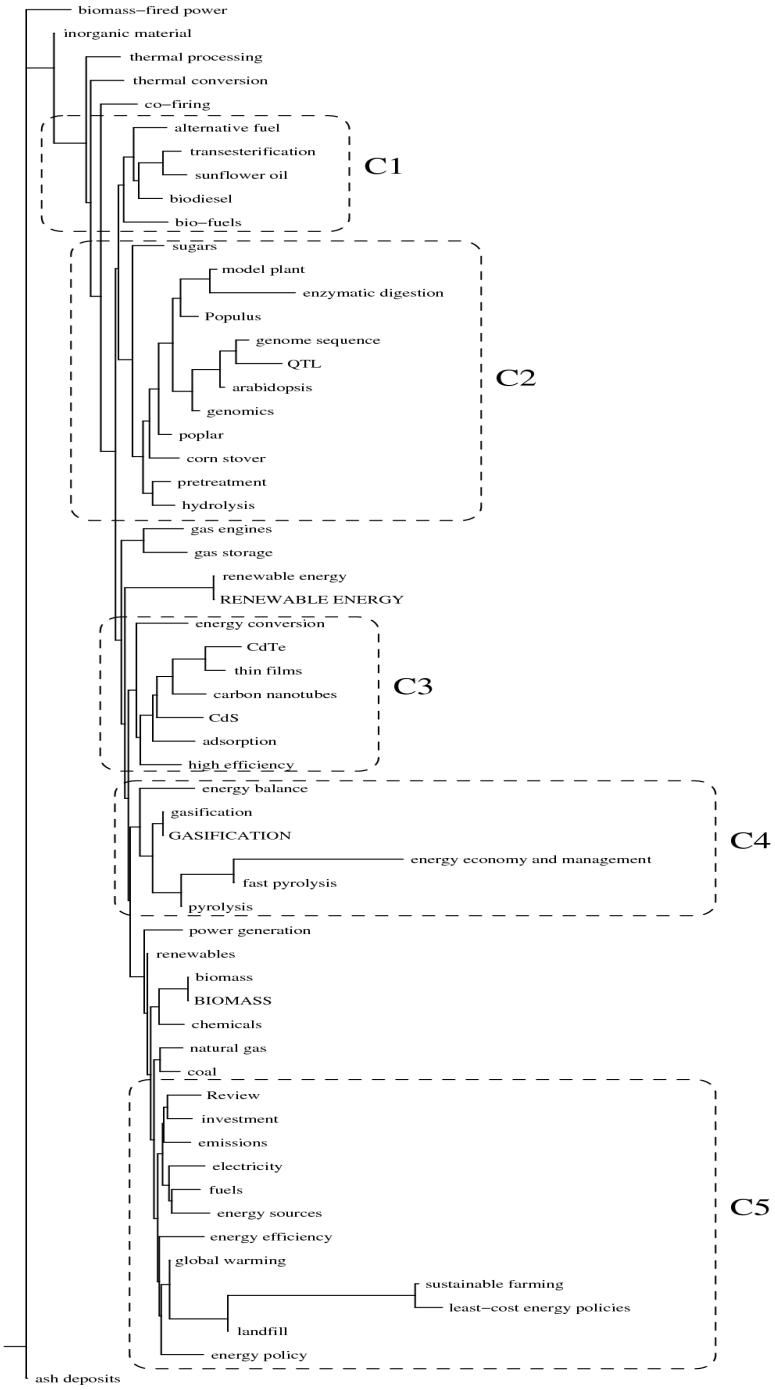


Figure 5.Keyword tree for author keywords collection

Next, we study the keyword clusters generated using the k-means algorithm. The matrix-k-means algorithm was used to automatically partition the author keyword collection into 10 categories. As the clustering operation has an element of randomness in it, 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 the table below. 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 sugars, populus, pretreatment, Arabidopsis, QTL, co-firing,
K3	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 natural gas, renewable energy, review,

	energy efficiency,
K8	investment, electricity, global warming, renewables, fuels, energy sources, energy policy, power generation, coal, emis-sions, renewable energy
K9	alternative fuel, biomass, gasification, biodiesel, gas storage, chemicals, GASIFICATION, BIOMASS
K10	sustainable farming and forestry, least-cost energy policies, landfill

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 TiO2, 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, Ti02 thin films, TiO2 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 ([Antolin 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.

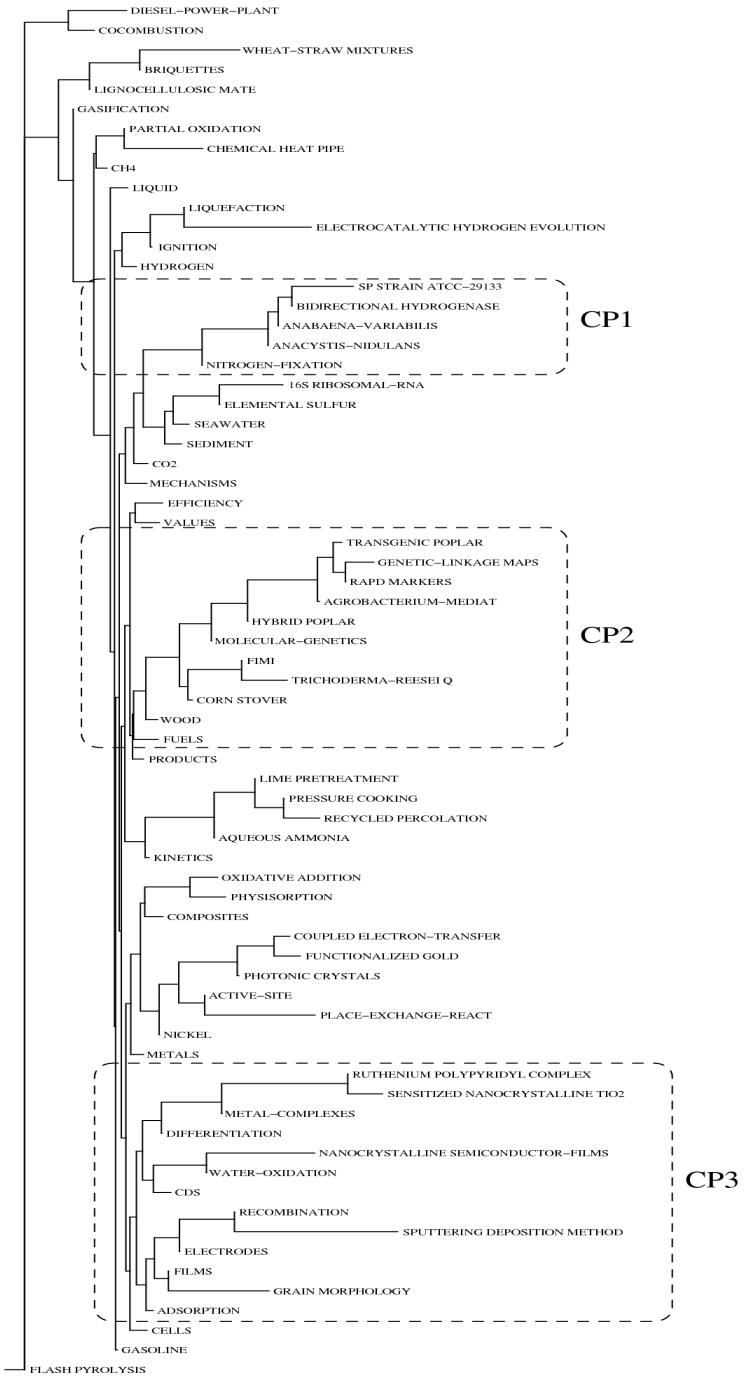


Figure 6.Visualization for keyword plus data (set 1)

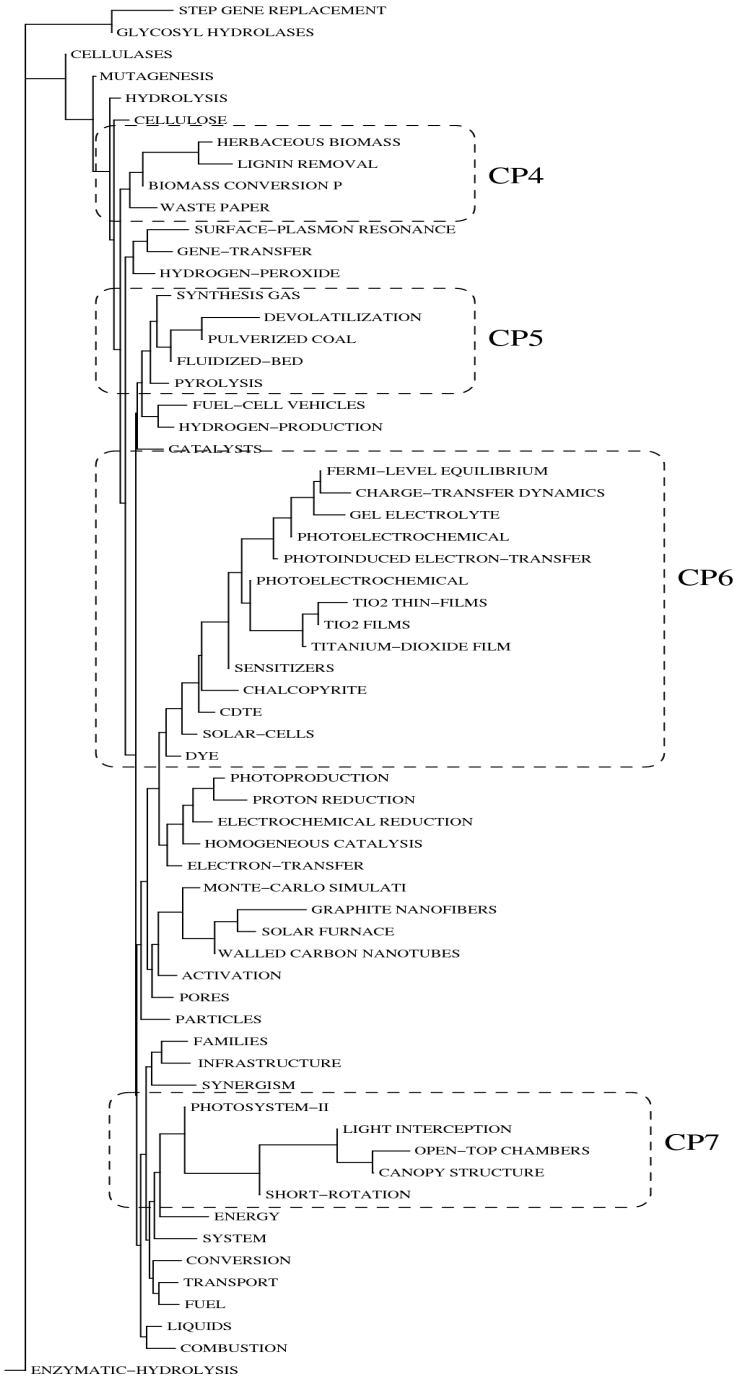


Figure 7. Visualization for keyword plus data (set 2)

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

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:

- KP14 is identical with CP1, which is associated with the production of hydrogen.
- Clusters KP13 and KP16 are both related to solar cells and match the contents of CP3 and CP6 very closely.
- In addition, the terms in KP16 are drawn from the field of nanotechnology, a field with a great many applications in renewable energy.
- 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.
- 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 Values, Products, Cds, Families, Energy, System, Fuel
KP3	Fimi, Trichoderma-Reesei Qm-9414, Diesel-Power-Plant,
KP4	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 Oxidative Addition, Ruthenium Polypyridyl Complex,
KP13	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
KP 15	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

This concludes the update on the data visualization task. The following section will describe techniques and methods for incorporating contextual/semantic enhancements into the tech-mining process. Two main concepts will be discussed. The first evolves around an extension of the basic NGD (utilized in this section) to allow for asymmetries in the distance measures. It will be shown that this allows for the automated generation of taxonomies which reflect the class structure of the keywords or terms being studied.

The second approach is based on the use of linear algebra techniques to analyse term-occurrence statistics. The main aim is to extract underlying concepts which reflect research threads or themes. While in some ways this is similar to clustering, one key difference is that there is no absolute partitioning of the

terms into clear-cut classes or clusters. Rather, the association of terms with particular concepts is indicated by a weighting term, which is automatically calculated.

T6: Contextual/semantic extensions

This is the other major research component of the current reporting period. In general, the aim of this investigation was to discover methods for exploiting the semantic closeness and other similarities between keywords used in the study. As mentioned above, the motivations and intuition behind the activities in this section are similar to those in the visualization section, and could be regarded as a continuation of the same theme. However, it is also hoped that this thread of research will eventually lead to techniques for working with and creating energy ontologies, which in turn can support a variety of further research activities.

Our investigations in this section can be divided into two main directions:

1. The automated creation of keyword taxonomies
2. Latent Semantic analysis

The first item will now be discussed.

Automated Taxonomy Creation

Keyword distances revisited

Resuming from the discussions of the previous section, one of the important properties of a distance measure is that it should be symmetric. However, there are cases where we expect the relationships between objects being mapped to be asymmetric. Indeed, the present situation is one such example where, for two keywords being studied, it is likely that the information attached to one keyword is a subset of the information associated with the other keyword. This can indicate that the field of research linked to one of the keywords is a subtopic of the other. We postulate that these asymmetries can be exploited to build a better representation of the technological landscape being studied.

Firstly, we describe a method by which the NGD can be modified to allow for such asymmetry. Recall that the numerator of the expression in eq. (2) quantifies the amount of information which is needed to produce two objects *x*,*y*, given an encoding of the object with the lesser information content. Choosing the object with less information enforces the symmetry condition but also removes the desired directional property.

Thus, a directional version of this distance can easily be obtained as follows:

$$\overrightarrow{NID} = \frac{K(x,y) - K(y)}{K(x)} \qquad (4)$$

In this equation, the expression denotes the directional version of NID, and can be interpreted as the additional information required to obtain both *x* and *y* given only object *y*. To see how this helps us, consider the scenario where object *y* is a subclass of object *x*; in this case, we expect that *y* would *already incorporate most of the information regarding x.*

Take the example of a circus elephant, which can be considered a subclass of elephant since all circus elephants are elephants while the same does not hold true in reverse. Also, it is clear that any description of a circus elephant must include a definition of what an elephant is, in addition to the fact that this particular elephant lives in a circus. In the present context, we could express this as follows:

$$information(elephant) \subset information(circus\ elephant),$$

$$K(circus\ elephant, elephant) - K(circus\ elephant) \approx 0.$$

Hence, at least in this case, we can see how a small value of $K(x,y)-K(y)$ is an indication of subclassing. $K(x)$ again serves as a helpful normalization term, for example, to guard against the trivial case where $K(x)=0 \Rightarrow K(x,y)=K(y)$.

Finally, eq. (4) can be used to derive the directional version of the NGD

$$\overrightarrow{NGD}(t_x, t_y) = \frac{\log n_y - \log n_x, n_y}{\log N - \log n_x} \qquad (5)$$

This can now be used to analyze collections of technology related keywords from the perspective of graph theory. Given a collection of keywords $\mathcal{V}$, we can construct a *directed graph* or digraph consisting of the pair of $(\mathcal{V},\mathcal{E})$, where the keyword list is mapped to the set of nodes of the graph $\mathcal{V}$, the set of edges of the graph, and the weighting function w: $\mathcal{E} \rightarrow \mathbb{R}$ is given by:

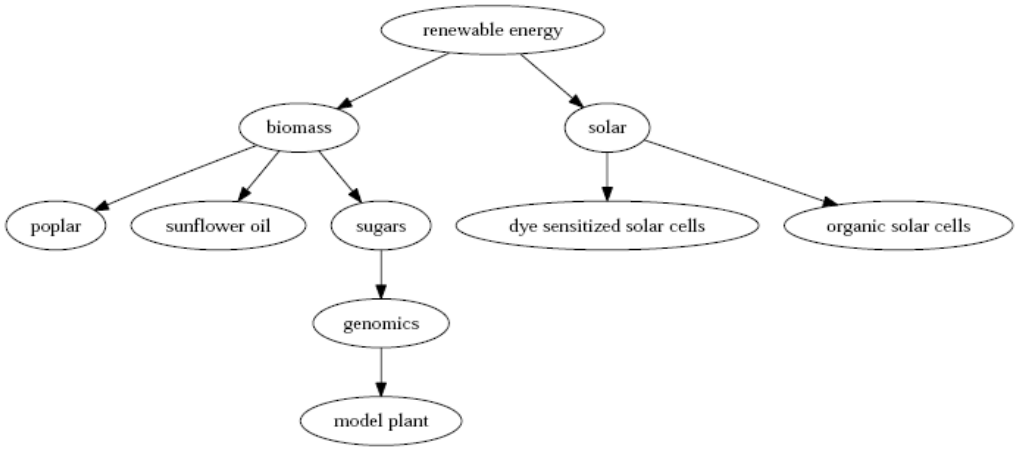
$$w[(u, v)] = \overrightarrow{NGD}(u, v) \qquad (6)$$

In this context, a keyword taxonomy is represented by a subgraph , where:

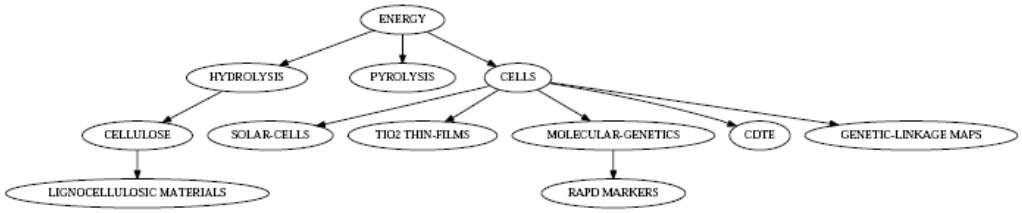
1. $\mathcal{E}^* \subset \mathcal{E}, \ |\mathcal{E}^*| = |V| - 1$
2. All nodes except one have exactly one incoming edge.
3. $(V, \mathcal{E}^*)$ is connected, and there are no cycles.

In graph theory this construct is known as an *arborescence*, which is basically the directed equivalent of a spanning tree. However, for any digraph there could be a very large number of such arborescences, any one of which could potentially be a valid keyword taxonomy. To solve this, we choose to follow the principle of parsimony in suggesting that the arborescence with the *minimum total edge weight* provides the best possible organization of the terms. In graph theory the problem of finding this arborescence is referred to as the minimum arborescence problem.

To demonstrate that this principle works, it is used to automatically infer the taxonomic structure of two small selections of renewable energy related keywords, and these are shown in figure 8 below. The resulting topic trees show that the terms have been organized into hierarchies that approximately reflect the inter-dependencies between the terms.



(a) Example 1



(b) Example 2

Figure 8. Sample taxonomies: renewable energy technologies

Weighted cost functions

As mentioned above, when searching for the most likely taxonomy of keyword terms, the selection criteria is the total weight (i.e. distance values) of the edges in the corresponding arborescence. However, in practice this often resulted in local structure which did not reflect the actual inheritance structure. In a noiseless environment this would not be a problem but in practice there are a number of situations where this reduces the accuracy of the results.

For example, consider the taxonomy in figure 8(a). We see that *sugars* has been classified under the Biomass subtree. However, *genomics* and *model plant* have subsequently been placed as subclasses of sugars. However, it would appear that the aspect of genomics research related to sugars may be separate from the subset of research in sugars related to biomass. We can check this by studying the directional

distances: while $\overrightarrow{NGD}(genomics, sugars)=0.336$, both of which are the smallest values in the respective rows of the distance matrix. However, $\overrightarrow{NGD}(genomics, biomass)=0.462$ which is somewhat greater than $\overrightarrow{NGD}(genomics, renewable energy)=0.395$, suggesting that perhaps the genomics subtree might be better portrayed as a separate branch of research from biomass.

Another example is shown in figure 8(b), where the term *cells* has attracted a large number of direct descendants: *solar-cells, TiO thin films, molecular genetics, CdTe, genetic-linkage maps*. This is a problem which is frequently encountered, in which very broad terms (such as *cells*) tend to dominate the subclassing process, resulting in extremely flat hierarchies. A further complication is that the keyword *cells* has two senses: solar "cells", and biological "cells".

In common with many other inverse problems, the two issues stated above can be linked to the fundamentally ill-posed nature of the problem - not only are we attempting to estimate the underlying taxonomy from indirectly observed and noisy aggregate data, the "truely optimal" structure of the taxonomy itself is also difficult to define - even by human experts.

However, one way in which we can try to improve the situation is by incorporating information regarding global structure into the process, as this will hopefully reduce glaring inconsistencies within the generated taxonomies. As an initial measure, we propose the following weighted cost function for evaluating the quality of generated taxonomies:

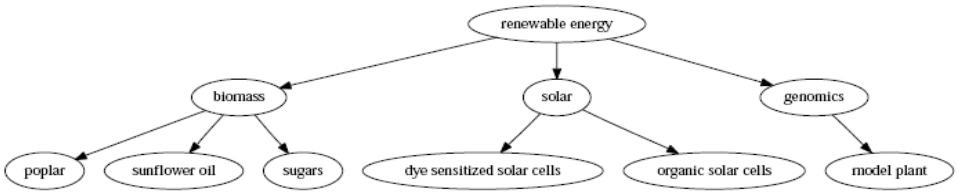
$$f_V(\varepsilon')=\sum_{v\in V}\frac{\sum_{i=1}^{n}\alpha_i\overrightarrow{NGD}(v, v_{\varepsilon'}^{i})}{\sum_{i=1}^{n}\alpha_i} \qquad (7)$$

where ε' is the set of edges in the taxonomy under consideration, $\mathcal{V}$ is the set of nodes, $v^{i}_{\varepsilon'}$ denotes the *i*th ancestor of node *v* given the edge-set and *n* is the number of ancestors for a given node. The co-efficients are weights which determine the extent to which the score of a particular node is affected by its indirect ancestors. Thus, $\alpha_1=1$, $\alpha_{2\,..\,n}=0$ simply results in the total path length objective function (i.e. optimizing this is equivalent to finding the minimum arborescence).

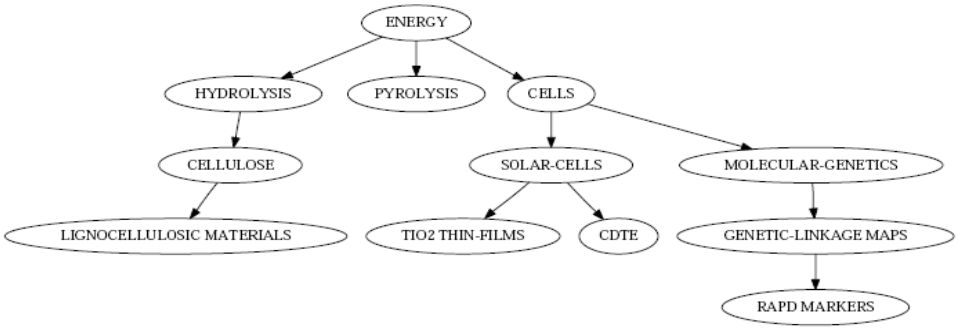
Intuitively, as we traverse the tree from any node *v* towards the root, the distances would be expected to increase as we move away from *v*. As such, a reasonable choice for would be a monotonically decreasing function, i.e. the highest priority is given to the immediate ancestor of a given node, while the influence of subsequent ancestors gradually diminishes. A number of weighting functions were tested and in the following sections we present results generated using three such functions:

1. **Uniform weighting** $\alpha_{1..n}=1$
2. **Linear weighting** $\alpha_1=n$-1
3. **Exponential weigthing** $\alpha_1=(1/2)^{n-1}$

As an example, taxonomies containing the same keywords have been generated by optimizing the linear weighted cost function, and are shown in figure 9 (optimization was done using a genetic algorithm, which is discussed in the following section). As can be seen from these two figures, the use of the weighted cost function produces some noticeable improvements in the resulting taxonomies. In particular, the sub-tree {*genomics* → *model plant*} in figure 9(a) has been directly connected to the root node, while in figure 9(b), the sub-tree descending from *cells* is now more structured (in figure 8(b), this subtree was mainly a flat hierarchy. Accordingly, the two senses of *cells* have now been appropriately divided into two separate subtrees, each of which shows a reasonable inheritance structure.



(a) Example 1



(b) Example 2

Figure 9. Sample taxonomies:Linear weighted cost function

Taxonomy optimization

While efficient algorithms exist for standard problems such as the minimum spanning tree (Kruskal's algorithm, Prim's algorithm), as well as Edmond's algorithm for the minimum arborescence problem [Korte and Vygen, 2006], the situation in cases when the cost function incorporates custom modifications or constraints is less clear.

In particular, Edmond's algorithm is inapplicable for the cost function in eq. (7), nor does there appear to be any efficient algorithm for finding the global optimum of this function. As the number of possible taxonomies grows exponentially with the number of nodes, exhaustive searches quickly become computationally infeasible.

As such, it was decided to use a Genetic Algorithm (GA) to optimize the automatically generated taxonomies. While not the only applicable technique, this approach does provide a very flexible framework in which a variety of different cost functions can be easily tested without having to devise a new optimization algorithm each time. In addition, GAs have been used in similar applications [Li and Bouchebaba, 2000], [Raidl, 2000], [Li, 2001] with some success, though in these previous studies the GAs were applied to problems involving undirected trees.

The basic components of any GA are:

1. A method for encoding a full set of the parameters to be optimized, where each encoded parameter set is called a "chromosome". For this study, the chromosomes were simply the connection matrices representing the digraphs. A connection matrix is a matrix with elements where indicates that there is an edge linking node *i* to node *j*, while means that there is no connection between the two nodes. In GA terminology, each chromosome is sometimes associated to an "individual".
2. A fitness function for evaluating each chromosome. As discussed previously, in this study the GAs will be used to test the weighted subclassing cost functions.
3. A set of *cross-over* and *mutation* operations on the chromosomes. Traditionally, GAs have been based on linear, binary chromosomes but this would be inappropriate in the current application where the natural representation of parameters is as a tree structure. Instead, we adopt the following two customized operations for chromosome transformation:
   - *Mutation* - the mutation procedure operates on individual trees. A random subtree is moved from one point of the hierarchy to another randomly selected point in the same tree (figure 10).
   - The *Cross-over* procedure accepts pairs of trees at a time. The operation comprises two stages: in the first stage, a random subtree is selected from each of the original trees and is transplanted onto a random point in the other tree (figure 11). However, this process invalidates the original taxonomies as the transplanted nodes would now appear twice in the same taxonomy. To resolve this, the transplantation stage is immediately followed by a chromosome repair process (figure 12) where the *originals* from the duplicated nodes are removed and all descendants thereof promoted to the ancestor nodes at the next level in the hierarchy.

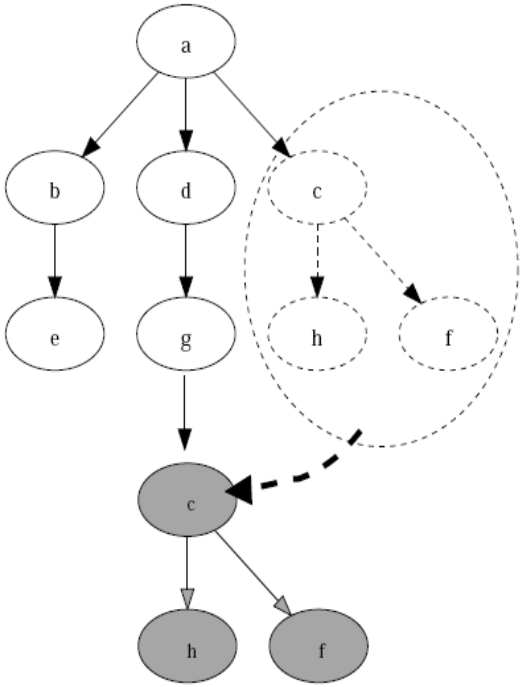


Figure 10. Mutation operation

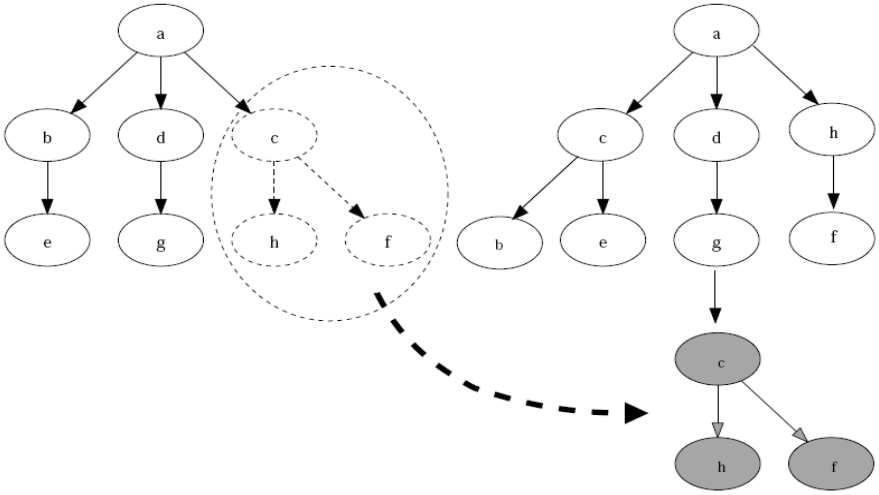


Figure 11. Cross-over operation

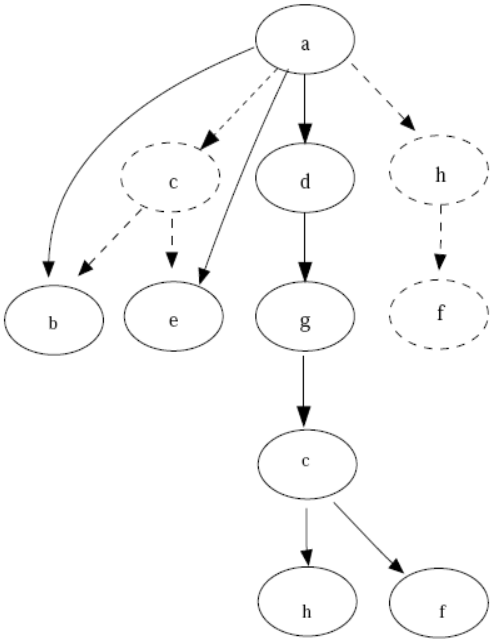


Figure 12. Chromosome repair process

Once all these components have been specified we are ready to attempt the GA optimization. Broadly, this proceeds as follows:

1. Initialization of the GA by creating a population of randomly generated individuals.
2. The fittest amongst these are selected for reproduction and propogation to the next iteration of the algorithm.
3. During this reproduction process, random perturbations are introduced in the form of the mutation and cross-over operations discussed above.

Results

For the experiments on this technique, keywords from the Web of Science database were collected as in the case of the data visualization experiments; however, currently we are stil only working with the "author keywords" set and have yet to start tests using the "keyword plus" set (this research component is still being actively pursued). Keywords extracted from the top 35 cited papers on "renewable+energy" where collected, resulting in 72 keywords in total, and the taxonomy generating process described in the preceeding sections carried out.

To facilitate presentation and analysis of the results, the collection was randomly divided into two subsets - set one contains 35 keywords, and set two contained the remaining 37 keywords. In addition, overly generic terms such as "review" and "high efficiency", were removed using a manually generated stopword list. In the following subsections the observations obtained which each of the sets are discussed in greater detail.

Set 1

The proposed methods were first applied to the keywords in set 1. Taxonomies were generated using Edmond's algorithm and GA optimization using first the uniform weighting then the exponential weighting functions; these are presented in fig 13.

The main observations were:

1. In general, the generated taxonomies appear to capture the high level orderings of the terms in the collection, at least to a reasonable degree of accuracy. In particular, there were two big clusters: one dedicated to Biomass related technologies and the other to technologies associated with thin-film solar cells. There were also other nodes and "micro-clusters" which descended directly from the root, notably the pairs {*genomics, model plant*} (molecular genetics related) and {*global warming, sustainable farming and forestry*} (policy related).
2. The results obtained using the weighted schemes were almost identical - when was set to linearly and exponentially decaying values, identical results were obtained. When using uniform weights, the results were still similar but there was a change in the *thin film* subtree, where *dye sensitized solar cells* was classified as a subclass of CdTe instead of being a direct subclass of *thin film*.

3. However, there is a bigger difference between the taxonomy generated using Edmond's algorithms (figure 13(a)) and those generated using the genetic algorithm. While the overall structure remained the same, the former had a flatter hierarchy, with much less subtree formation.

   Consider, especially, the *biomass* subtree; in figure 13(a), six branches emanate from this node, only two of which have any further descendants. In contrast, in figure 13(b) (uniform weights), four nodes descend directly from *biomass*, namely *biodiesel, gasification, populus* and *alkanes*. Of these, *biodiesel* is further linked to *sunflower oil*, which can be used to create biodiesel via transesterification. Similarly, *gasification* is joined to a pair of related concepts - *pyrolysis* and *gas engines*.

   We note that, while a flatter hierarchy is not necessarily "wrong", the presence of more structure is generally more valuable (provided it is accurate, which it appears to be in this case) as the objective of the whole exercise is to organize and sort the information in a more intuitive way.

Set 2

Next, the second set of keywords (set 2) were organized into a taxonomy using the proposed approach. The resulting graphs are shown in figure 14.

Our observations on these graphs are:

1. As before, the taxonomies show a number of significant clusters, which include *solar*, *sugars*, *adsorption*, *natural gas* and *power generation*. However, it was observed that there is much less consistency amongst the four taxonomies.
2. As before, the results using Edmond's algorithm produced a slightly flatter hierarchy than when using the weighted cost functions; however, this difference was less pronounced than in the case of set 1.
3. The taxonomies created when was linearly and exponentially decreasing were very similar, though this time there was one very minor difference between them.
4. The *natural gas* subtree is somewhat mixed in its composition (which also changes significantly in the four taxonomies for set 2), and appears to be a kind of "catch-all" cluster for a number of orphaned terms. While a more reliable analysis would require further domain knowledge, an informal scan of the academic literature on this subject suggests that this problem occurred as a result of a number of factors: firstly, *natural gas* is an extremely common term in renewable energy, while technical research that focusses specifically on natural gas is relatively less common. Instead, we notice that this terms frequently appears in articles that are broader in scope, such as review papers and papers on various strategic issues such as global warming, energy markets and the like. This allows the term to attract a broad range of "subclasses" which may not easily fit into other sections of these taxonomies. In particular, note that many of the

terms descended from *natural gas* are themselves fairly broad in nature - and would likely appear in similar publications.

5. The other major subtree was *sugars*. Again, there was significant variability across the taxonomies in terms of the nodes classified under this subtree, as well as the intra-tree ordering of these nodes, but in general there appeared to be three main areas of research: one was on the chemical processes used to break down and exploit sugars or related compounds (examples of constituent nodes were *hydrolysis, enzymatic digestion* and *pretreatment*). The second area was molecular genetics, with terms such as *arabidopsis* and *genome sequence*. The final related area of research mainly consisted of a single node, *poplar*. This is a species of tree which is used as a source of pulp and hence cellulose, a complex carbohydrate (the exploitation of cellulosic materials such as pulp as an energy feedstock is now an active area of research as these will not threaten food supplies). While represented by a single node in the present collection of keywords, this appears to be a major area of research in biomass based sources of renewable energy.

In brief, the results obtained from these experiments indicate that the proposed methodology has the potential to be a useful tool for facilitating the ontology creation process, as well as for providing researchers with a high-level summary of the research domain. On the other hand, there are still challenges to be overcome before this approach can be packaged into a fully automated software application. The main issues include:

1. Complexity - as with many other inverse problems, inferring the underlying taxonomy of a collection of keywords is ill- posed: even ontologies created by subject matter experts can show significant variability. This is because the exact structure and organization of a taxonomy is very sujective and depends heavily on the perspective and motivations of the developer.

2. Inconsistent quality of data; data obtained from publicly available sources are unregulated and are frequently noisy; this further underscores the need for appropriate filtering and data cleaning mechanisms.

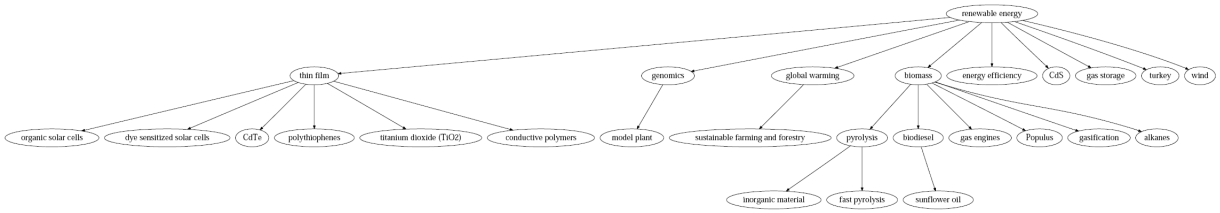
3. Non-uniform coverage - the number of hits returned for very general or high-profile keywords such as “energy” or “efficiency” was a lot greater than for more specialized topics. This is unfortunate as it is often these topics which are of the greater interest to researchers. One way in which we hope to overcome this problem is by aggregating information from a larger variety of sources, examples of which include technical report and patent databases and possibly even mainstream media and blogs.

4. 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

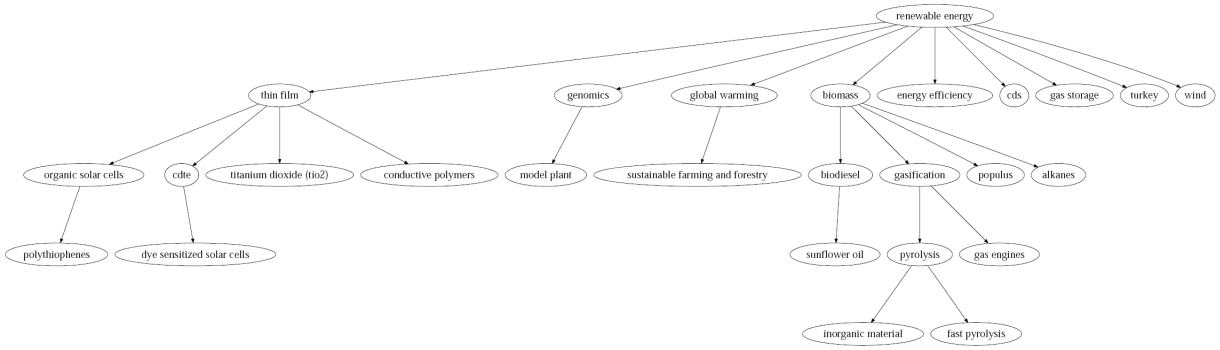
solutions and information loss resulting from dimensionality reduction, are all symptoms of the inherent difficulty of this problem.

While the taxonomy creation technique just discussed allows for explicit organization and structuring of the keyword collection, it still suffers from many of the shortcomings encountered when using clustering (especially hierarchical clustering) schemes. In particular, keywords need to be uniquely assigned to groups or classes, even if the a particular term is extremely generic, or may have more than one meaning.

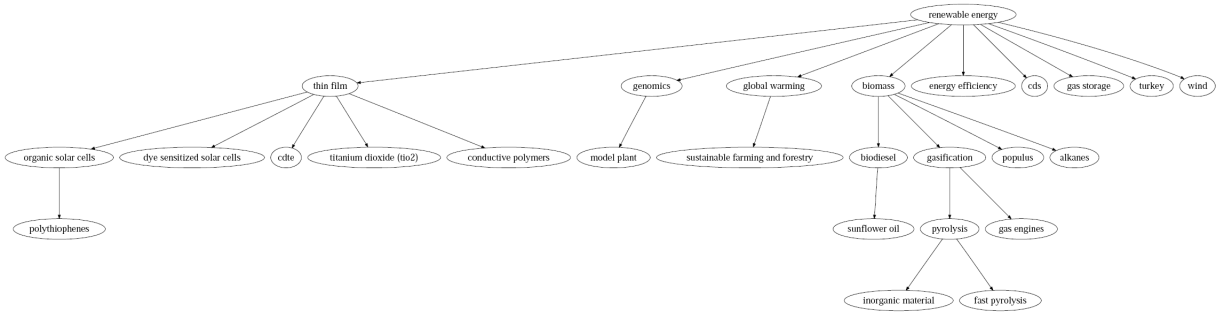
One possible solution is to allow for some degree of “fuzziness” in this assignment process. The following section describes an alternative approach, known as Latent Semantic Analysis, which avoids some of these problems.



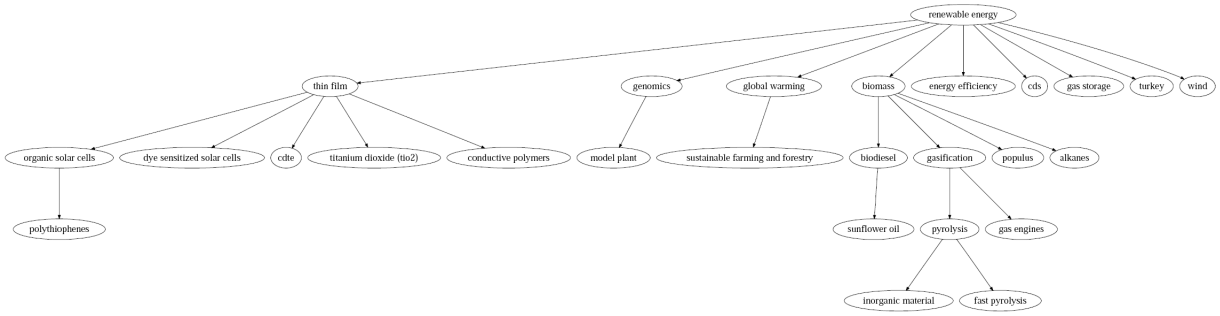
(a) Edmonds algorithm



(b) Uniform weights

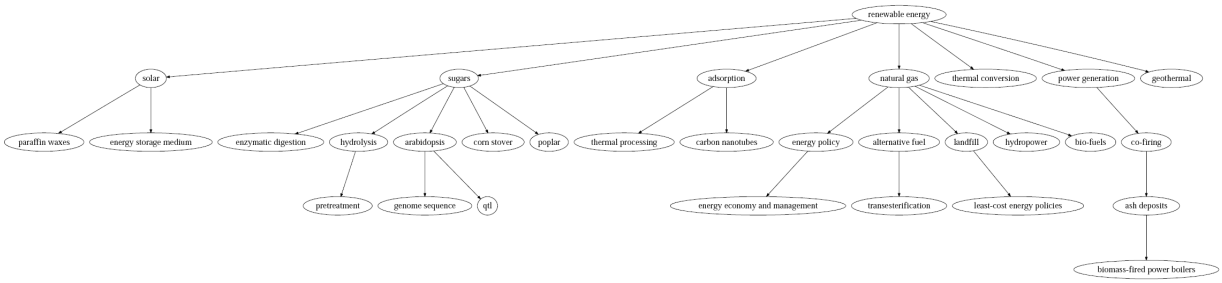


(c) Linearly decaying weights

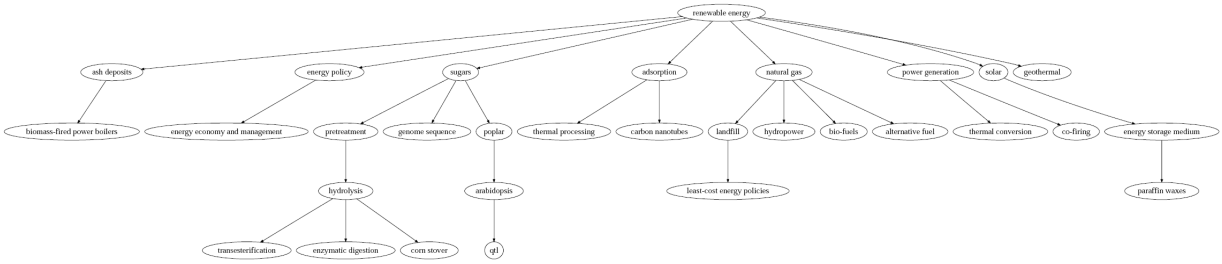


(d) Exponentially decaying weights

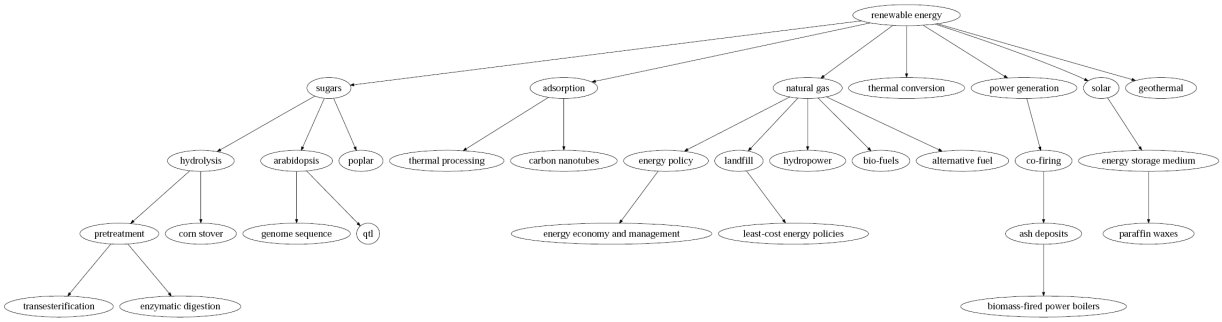
Figure 13. Automatically generated taxonomies: Set 1



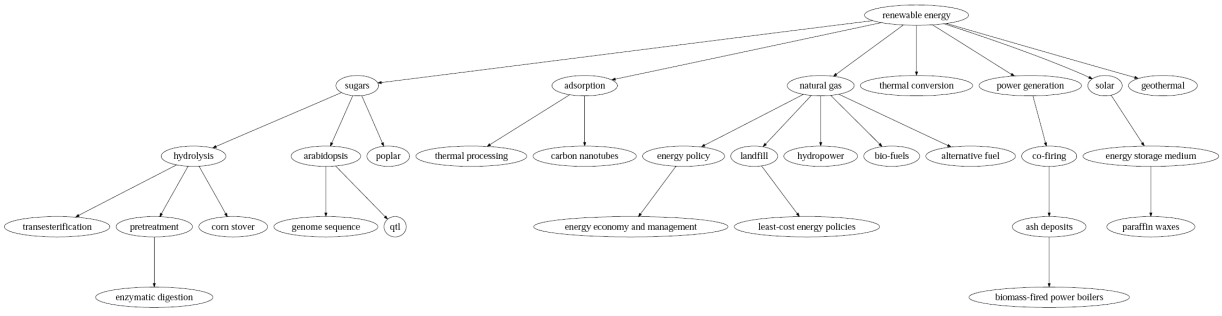
(a) Edmonds algorithm



(b) Uniform weights



(c) Linearly decaying weights



(d) Exponentially decaying weights

Figure 14. Automatically generated taxonomies: Set 2

Latent Semantic Analysis

Background

Latent Semantic Analysis (LSA) is a technique for identifying relationships between key terms in a set of documents. It produces a set of *concepts*, each of which is a different combination of the terms being analyzed. A concept can be thought of as a grouping of terms that relate to one another. However, LSA and the identification of concepts should not be confused with *clustering*. Clusters are disjoint; any given term is in one and only one cluster. Each LSA concept, on the other hand, contains a particular weighted combination of *all* the terms. Each concept, taken as a whole, is independent from all others (literally orthogonal vectors in a space, shown later), but the terms that make up each concept are found, with some weighting, in all of the concepts.

LSA is based on matrix algebra; the basic procedure takes as input a *term-document matrix* [Berry et al., 1995]. This matrix is a representation of the frequency of occurrence of each term in each document in the database. Terms are listed along the rows of the matrix, and documents are listed along the columns. The next step is based on a mathematical techniques known as Singular Value Decomposition or SVD. Briefly, SVN allows the $m \times n$ term-documnet matrix $\boldsymbol{A}$ to be decomposed as follows:

$$A = U \Sigma V^T \quad , \qquad (8)$$

where, $\boldsymbol{U}$ is an $m \times m$ *term-concept* matrix, $\boldsymbol{V}$ is an $n \times n$ *document*-concept matrix and $\sum$ is an $m \times n$ diagonal matrix containing the scale factors for each concept. This is the key concept of LSA, where we would like to move away from hit counts corresponding to individual terms or keywords, and work instead in the *concept-space*, where each concept is specified yb a weighted sum over the entire keyword collection. In the expression above, these weights are then provided by the columns of $\boldsymbol{U}$.

A practical approximation of the keyword covariance matrix

For our purposes, a modification to the standard LSA algorithm is required. The sheer number of documents to be searched prevents us from generating a term-document matrix. Such a matrix could potentially have millions of columns. Furthermore, determining each value in the matrix would require searching the complete text of each document to count term occurrences. These prohibitions necessitate us to adopt a simplification.

A detailed description of the simplification used is provided in the working paper CISL #2008-12 (referenced in the publications section) but the basic idea is that it is possible to rewrite the covariance matrix in a form where all terms are expressed in terms of the *hit counts* returned by a search engine, as follows:

$$cov(i, j) = \frac{1}{N}\left(h_{i,j} - \frac{1}{N} h_i h_j\right) \qquad (9)$$

In the above expression, $h_{i,j}$ denotes the number of hits returned from a search for both terms, while h_i and h_j denote the number of hits returned for terms *i* and *j* respectively. *N* is the total number of documents indexed by the search engine. We approximate this with the number of hits returned from a search for a large term that subsumes terms *i* and *j*. For our purposes, we use the field that is the focus of our case study, "renewable+energy", as the search term to acquire *N*.

Using this simplified expression we can now construct the covariance matrix for the term document matrix containing the entire collection of documents indexed by a particular search engine. Given this matrix, we can now obtain the desired concept vectors by simply calculating the eigenvectors of the covariance matrix. In other words, the eigenvectors of our covariance matrix are equivalent to the columns of the matrix obtained from the singular-value decomposition of the term-document matrix.

Preliminary findings

Concept vectors were generated for the author keyword set (Appendix B (Section B1)). Initial observations indicated that a number of the concept vectors were dominated by very broad terms such as *electricity* and *emissions* (this is expected as the eigenvectors will seek out directions of maximal variance in the data space). However, a number of very interesting vectors were also observed.

Typically with eigenvector analysis, the most important vectors are those with the largest associated eigenvalues. Those vectors are considered to be more significant because they represent the greatest variance in the data. However, for our purposes, we are at least as interested, and perhaps more so, in the concepts with lower variance. Methods for choosing good, representative subsets of concept vectors calculated by LSA are still being investigated, but for now, interesting vectors are largely chosen by inspection. As a topic for our initial discussions, the following tables lists highly weighted terms for five concepts which appeared to be particularly interesting:

Concept Vector 1		**Concept Vector 2**		**Concept Vector 3**	
alternative fuel	0.8811	power generation	0.6977	energy efficiency	0.4777
biodiesel	-0.4372	electricity	-0.3787	electricity	0.4752
thin films	0.09212	energy policy	-0.3271	energy policy	-0.3904
natural gas	-0.07202	coal	-0.2316	renewables	-0.2674
bio-fuels	-0.05371	review	0.2183	global warming	-0.2458
sugars	0.05141	fuels	0.2139	investment	-0.2405
		adsorption	-0.2045		
		energy efficiency	0.1311		

Concept Vector 4		**Concept Vector 5**	
pyrolysis	-0.6103	gas engines	-0.8129

pretreatment	-0.4823	transesterification	-0.3472
hydrolysis	-0.3845	carbon nanotubes	0.2913
sugars	0.2638	corn stover	-0.2785
chemicals	0.1964	thermal conversion	0.1336
landfill	0.1343	sunflower oil	-0.1137
gasification	0.1228	biodiesel	0.07669
fast pyrolysis	-0.1221	thermal processing	0.06852
GASIFICATION	0.1139	alternative fuel	0.05073

Concept 1: as its primary component suggests, is about alternative fuels. Biodiesel, natural gas, bio-fuels, and sugars all fall in this category. Thin films is only slightly off topic. They can be used to increase the efficiency of solar power systems, so while not an alternative fuel per se, they are an important part of an alternative energy.

Concept 2: broadly, this concept consists of power generation topics. Energy policy seems a bit out of place, as the rest of the concept mostly deals with technology, rather than policy. But this tells us that it is common to find information regarding energy policy in the same articles as those about energy technologies. Review also appears to be wildly out of place. However, the explanation here is that review is simply too broad of a term to get clean LSA results from it. Evidently, it appears in articles regarding energy technology, even though its actual meaning is unrelated. In the same way, the word "the" would be a large component of virtually all LSA concept vectors if it were included in the term list.

Concept 3: seems to broadly address energy policy and environmental impact. Energy efficiency, electricity, energy policy, renewables, and global warming are intuitively all components of this concept. Investment, on the other hand, may not be such an obvious member of the set. However, in this particular concept, investment likely refers to investment in alternative energy technologies. After all, the financial aspect of "going green" is often an important issue, so it should be little surprise that LSA has revealed investment as a component of an energy policy concept.

Concept 4: The terms this concept represent, for the most part, chemical breakdown of organic compounds. Pyrolysis, the first term in the vector, refers specifically to this process. Fast pyrolysis, found in this vector with a smaller weight, is clearly a related term. Hydrolysis is a process related to pyrolysis that refers to breaking water down into hydrogen and oxygen. Gasification (and GASIFICATION) similarly refers to the decomposition of organic materials into carbon monoxide and hydrogen. Sugars and chemicals belong in this concept as well; hydrolysis of disaccharide sugars produces monosaccaride sugars, and chemicals, while a somewhat general and vague term, clearly applies

to the concept as a whole. Pretreatment and landfill seem a bit out of place in this concept, but the purpose of LSA is to reveal unknown term relationships, which is the case here.

Concept 5: appears to be related to a number of alternative energy technologies, especially biodiesel. Of the highly weighted terms listed in the table above, the keywords *transesterification, corn stover, thermal conversion, sunflower oil, biodiesel, thermal processing* and *alternative fuel* are directly related to the production of biodiesel; *transesterification, thermal conversion* and *thermal processing* are terms for processes used for creating biodiesel, while *corn stover* and *sunflower oil* are examples of feedstocks.

Visualization

Displaying LSA results in a meaningful way is somewhat of a challenge. The results take the form of high-dimensional vectors, so plotting them is not an option. One simple visualization technique is that which has been used in this paper thus far; simply listing the values of the large vector components. This could be equivalently displayed by plotting terms on a "concept line." In other words, for a given concept vector, each term would be plotted on a number line between -1 to 1, with its location representing that term's weight in the given vector.

The idea of plotting terms on a concept line can be extended to two or three dimensions. In the case of two dimensions, two vectors are chosen at a time (as opposed to one vector at a time, as has been shown thus far). As all concept vectors are orthogonal to one another, the two chosen vectors can be placed on the x- and y-axes of a coordinate plane. Then each term is plotted on the plane, with its x-coordinate representing its weight in one of the two concept vectors, and its y-coordinate representing its weight in the other.

This technique can be used with three concept vectors at a time as well. However, the resulting three-dimensional graph can be difficult to interpret visually, and as such, we present only two-dimensional visualizations in this paper.

Two plots are shown below. The first uses concepts 3 and 4 as the axes, and the second uses concepts 4 and 5. The first was chosen for display because concepts 3 and 4 have some terms weighted reasonably high in each. This makes the graph more interesting because the points are not all located along the axes themselves. The second graph, which plots concepts 4 and 5 together, was chosen because of the clear clusters that it identifies. Four groupings are formed – sugar and chemicals; energy policy, GLOBAL WARMING, renewables, and investment; electricity and ENERGY EFFICIENCY; and hydrolysis, pretreatment, and pyrolysis. By viewing two vectors at a time instead of just one, we can see some new term interrelationships that were not immediately obvious before.

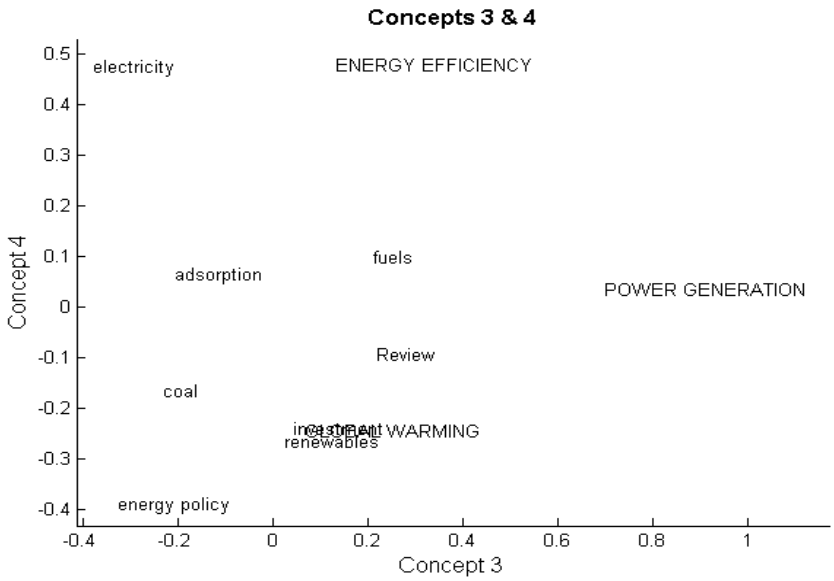


Figure 15. Concepts 3 & 4 – Terms such as energy policy and electricity lie well off of the coordinate axes, showing that they are important components in both concepts

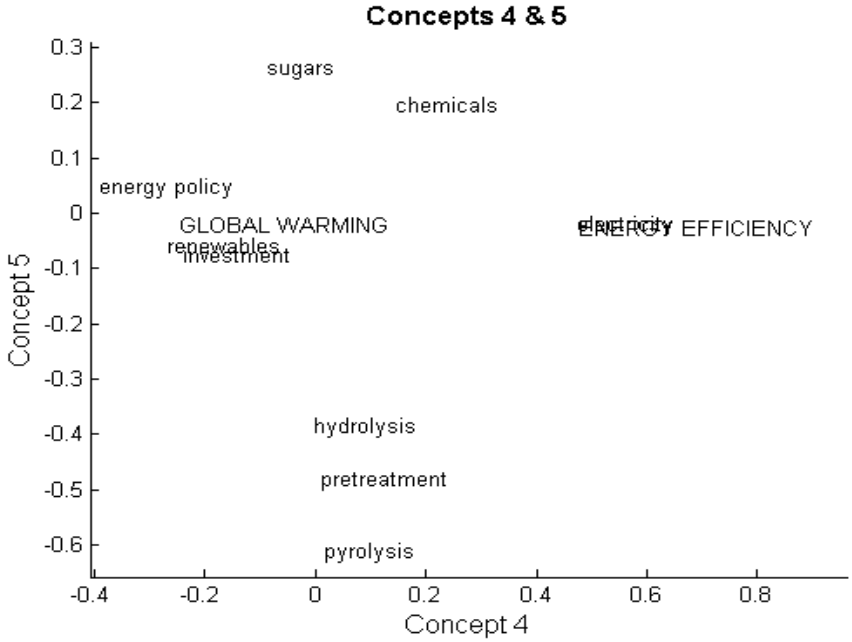


Figure 16. Concepts 4 & 5 – Four term clusters become apparent when these concepts are plotted together

Discussion

Latent Semantic Analysis should prove useful in our ongoing research on technology landscaping and forecasting. Bibliometric trend analysis, for example, involves observing the change in publication counts of a research field over time. With the help of LSA, we can query publication databases for entire concepts, rather than individual terms, which can provide a more accurate picture of the number of publications relevant to a field.

Towards the goal of using LSA for technology landscaping and forecasting, the following issues still need to be investigated:

- The significance of negative values in concept vectors. We have been considering only the absolute value of term weightings up to this point. However, it is likely that there is some significance to the sign of the weighting as well.
- A method for picking a subset of concept vectors from LSA results. One concept vector is produced for every keyword in the data set, so with large numbers of keywords, it could become impractical to use every vector. A method is needed to select important vectors from the set. As was mentioned earlier, this is not as simple as choosing vectors with the largest eigenvalues, as that could potentially cause discard of the most interesting vectors.
- The sensitivity of LSA to change in data source, i.e. publication database. Different databases may produce different concepts entirely, so we may need to explore ways of aggregating the different results.
- Methods for using concept vectors in database querying. A simple "And" query of every term in a concept will result in a very limited set of documents, as not many documents will contain every term. Similarly, an "Or" query will result in far too many results. We are investigating ways to query a database for the concept as a whole.

CURRENT REPORTING PERIOD SUMMARY (04/01/2007-08/31/2008)

Review of objectives

In summary, while the research tasks that were active during the past five months covered a large number of aspects, the direction of the project during this period has centered on the development of techniques to both organize and visualize the data in a way which reflects semantic relationships between keywords. In particular, we study the use of the *joint term frequencies* of pairs of keywords as a means of characterizing this semantic relationship – this is based on the intuition that terms which frequently appear together are more likely to be closely related.

It is also important to note that, while our focus in this reporting period might seem to be somewhat different from that of the previous one, we believe that these visualization and taxonomy generation techniques will play an important role in the broader context of technology forecasting, where they will be useful in the creation of semantically-enhanced technology features.

To summarize the achievements and progress made during this reporting period, listed below is the set of objectives from the earlier section "Description and objectives"; for each objective there is a brief summary of the associated activities. Also, for each activity we have indicated if it was primarily attended to by MIST or MIT researchers, or if it was a joint effort. (However, because all the researchers are currently based at MIT, it was difficult to allocate any one activity exclusively to either team):

1. *Development of software tools* – This was one of the areas emphasized in the first reporting period; however, tool development is regarded as an ongoing activity and improvements are planned throughout the entire duration of the project. For the current period, signficant progress was made in the development of a number of software tools such as the hit aggregator interfaces and a python port of the Cameleon web wrapper engine.

   MIST/MIT division: Joint effort.

2. *Survey of databases* – A substantial review of technological mining was conducted, and the results of this review are provided in a supplementary document attached to this report.

   MIST/MIT division: Primarily MIT.

3. *Development of base indicators* – This activity is currently awaiting outputs of investigations into visualization and semantic techniques.

4. *Development of base/extended visualization techniques* – This was the main focus of the analysis which was carried out during the preceding five months.

   MIST/MIT division: Primarily MIST.

5. Contextual/semantic extensions – This area was also a priority during this reporting period. Two separate techniques were introduced for using the joint term-frequencies to support the creation of semantic or contextually enhanced features.

   MIST/MIT division: Joint effort.

Complementary activities

Collaboration with Dr. Georgeta Vidican

In addition to the regular project related research activities, the project team is constantly striving to seek additional avenues through which scope and impact of the project may be magnified. In particular, we would like to highlight the initiation of a collaborative effort with the project team of Dr. Georgeta Vidican (Masdar Co-PI) and Dr. Richard Lester (MIT Co-PI).

It was noted that the research project currently being pursued by Dr. Vidican was in many ways relevant to our research activities, and that opportunities existed for joint efforts between the two project teams. Titled "*The Development of Innovative Industries and the Role of Public Research Institutions: The Case of Renewable Energy*", this project seeks to elucidate the role of public research institutions (universities and national research laboratories) in driving innovation and structural change, particularly in the renewable energy sector. The key research issues are:

1. What is the relative importance, for innovation and economic development, of public research institutions in knowledge creation, problem-solving, education and training, and in the provision of public space for "interpretive processes"?
2. How can university-industry links evolve to better serve the goals of both kinds of institutions?
3. What types of indicators (metrics) are most useful for evaluating the effectiveness of public spaces for interpretive processes in local innovation systems?

A quick examination reveals that while the two research efforts may differ in their high-level goals, there is definitely a significant degree of overlap between the specific objectives of the projects, which motivated the initiation of the collaboration. For example, one of the key challenges for our project is the need to find suitable procedures for evaluating the accuracy of the various metrics and features developed, and to study the relationship between these numerical indices with "ground-truth" observations. There is also a need for a formal framework in which data or information collected can be interpreted and converted into actionable recommendations.

Dr. Vidican's project, on the other hand, is largely qualitative in nature and seeks to study the processes which influence and lead to the emergence of innovation, especially in the context of industry-university collaborations. The primary means of data collection is via semi-structured interviews and secondary statistics from online sources and other published statistics. The project also seeks to evaluate the outputs from innovation by analyzing patents and other forms of codified knowledge exchange. Nevertheless, this effort would benefit significantly from employing a systematic quantitative method for assessing the intensity and the patterns of joined publications between public research institutions and industry.

As such, there certainly appear to be a variety of synergies between the two projects, which if properly exploited may lead to results and outputs, which were not feasible for either project in isolation. More specifically, it was decided that the collaborative effort would focus on the following key activities:

- To explore quantitative measures of innovation outcomes in the solar PV sector over time. Initially this will be via publication counts, though patents and other text resources could be exploited in future iterations.
- To determine the accuracy and reliability of automatically extracted numerical indices when used to study "fuzzy" phenomena such as the level and nature of university-industry collaborations as well as to establish the relationships between these indices and actual observed trends.

- To identify the kinds of companies (in terms of their positions in the respective value chains), which regularly appear in joint publications with public research institutions, and to understand their motivations in terms of the type of technological problems forming the basis of these collaboration and the patterns and models of knowledge exchange.

The outcomes from this analysis would allow us to better understand and anticipate the evolution of the links between public research institutions and industry, and to highlight how the goals of both kinds of institutions may be best served (i.e. knowledge sharing for universities, knowledge hoarding for industries). There has been a lot of research on patents and licensing, but there is very little understanding of how knowledge is exchanged through joined publications.

Initial results

One way in which the approach discussed above could be adapted for use in the present context is by including additional information in the search terms. For example, with many search engines it is possible to specify the author affiliations or the organizations where the research was carried out; so, searching for "photovoltaics+address:MA+email:*com+email:*edu" could be one way of locating research activity which is relevant to photovoltaics (PV), conducted in MA and which involves authors from both industry and academia. Figure 1 shows the results of queries for joined publications (universities-industry and national laboratories-industry) from the period ranging from 1975 to 2005 for the states of California (CA) and Massachusetts (MA).

One clear observation, which can be seen in both the CA and MA curves, was that there is a significant increase in the number of papers, which occurs approximately after 2003. Despite the fact that since year 2000 there has been a renewed interest in alternative sources of energy both in the policy arena as well as in the academic communities, it is still too early to speculate on the underlying factors which resulted in these figures. Also, for now, we have chosen to look at Massachusetts and California as a starting point for the research, but once the data mining process has been nailed down, this analysis could easily be extended to the national scale for the solar PV sector and possibly for other renewable technologies.

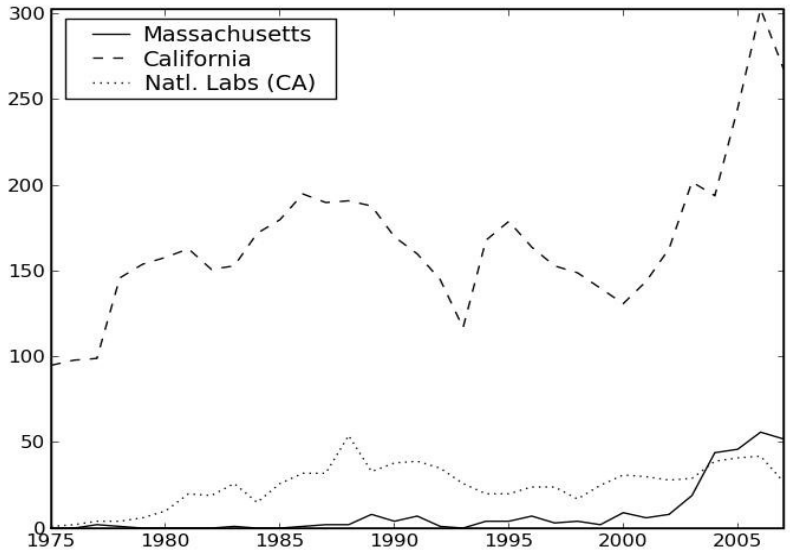


Figure 17.Publication counts reflecting the number of papers resulting from university-industry collaborations

Tech-mining using social-technical information systems

Collaborative research efforts are also underway with Mr.Seshasai (PhD candidate, ESD program), who is investigating the use of social information resources, currently focussed on blogs. In particular, Mr.Seshasai has been able to leverage his access to internal blogs from within the IBM intranet, as a data resource on which to run his experiments. The details of his work are provided in working paper CISL #2008-07 (please refer to the section on "publications/presentations" at the end of this report) but briefly, the benefits of pursuing this joint effort are as follows:

1. Validation and investigation of the effectiveness of the proposed bibliometric techniques on data extracted from blogs and other social databases. Initial results indicated that directly applying the tools on blog data resulted in highly inconsistent data (for further details please refer to the working paper) but this not surprising in view of the many quantitative and qualitative differences between data from blogs and from academic databases.

2. As a Phd candidate in the Engineering Systems Division, Mr. Seshasai can be expected to bring a systems perspective to the project; in this context, blogs may also be viewed as complete socio-technical systems, where the actual content is only one component and is enriched by the social context of authors, readers, commentators and participating organizations.

3. An additional benefit of this collaboration also provides us with a degree of access to the resources and perspective of IBM. On one level, this means thatMr. Seshasai will be able to utilize information gathered from IBM's intranet, which is already of huge interest and value to the research. However, we also hope that this will eventually lead to further ties being developed with other interested parties within IBM. One example is the "Horizon Watch" program, which is a loosely couple group of individuals within IBM whose main purview is to review and discuss emerging trends in technology. As most of these efforts are conducted manually, one of the ideas discussed was to survey Horizon Watch members to obtain their requirements and viewpoints regarding the kinds of tools that we are working on, and perhaps to identify individuals who might be willing to serve as beta users.

Exploring external funding opportunities

In addition to the research work described in this report, the project team has been considering a number of options for extending the scope, duration and depth of the collaboration between the members of the team. One avenue has been via the **DataSpace** NSF proposal which was described briefly in the first progress report. Unfortunately, that proposal was not funded in the first cycle but encouraging feedback was received and these will be taken into account in preparing a re-submission for the second funding cycle.

In addition, a pre-proposal was also submitted to the **BP-MIT Major Projects Research Program**, which is part of a broader research agreement between MIT and BP to conduct research that is related to major energy related projects. One of the stated specifications of the RFP was for projects which "employ integrative, multidisciplinary approaches to address the complex, multifaceted challenges in the selection, design, development and delivery of major projects"; in particular, it was felt that the course of research which we are currently pursuing addresses one of BP's key leadership problems as specified by the RFP: "*In seeking to balance the known challenges of today and the medium to long term challenges of future energy, what is the best way of distributing limited research funds?*" . While the details of the research to be carried out as part of this proposed project will ony be concretized if a full proposal is requested, the pre-preposal is based on the same research themes as the current project but will allow the scope of the current project to be broadened significantly, to magnify its potential impact and to engage BP as a valuable collaborator in our efforts.

FUTURE WORK

Activities for the next reporting period

From the project schedule, it can be seen that apart from "base indicators", "base visusalizations" and "tool development," all the project activities that were current for the present period are still active for the coming period. This sets the tone for the following six months, where our efforts will focus largely on consolidating and extending the research activities discussed in this report. In addition, the "data analysis"

component of the project will become active for the first time in the second quarter of the following reporting period, and will obviously gain in importance; particularly, we will be placing greater importance on the identification of means of interpreting and evaluating the results and methods which have been created thus far. One important direction is to work more closely with researchers and other domain experts from the field of renewable energy, with the aim of obtaining feedback and further insight into the validity of the results as well as the utility of the tools.

Finally, while no longer "officially" active, it is apparent that the development of tools and the discovery of new sources of information will be an ongoing effort which will continue throughout the duration of the project. This includes administrative tasks such as debugging and updating of regular expressions to cope with slight changes in the results pages of search engines, as well as the creation of new tools and programs which will allow the tools developed to be utilized by non-technical users.

REFERENCES

[Antol in et al., 2002] Antol in, G., Tinaut, F. V., Briceno, Y., Castano, V., Perez, C., and Ramirez, 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, 06] 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.

[Berry et al., 1995] Berry, M. W., Dumais, S.T., and Letsche, T. A. (1995). Computational Methods for Intelligent Information Access. In Proceedings of the 1995 ACM/IEEE Supercomputing Conference.

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

[Chiu and Ho, 07] Chiu, W.-T. and Ho, Y.-S. (2007). Bibliometric analysis of tsunami research. Scientometrics, 73(1):3–17.

[Cilibrasi and Vitanyi, 06] Cilibrasi, R. and Vitanyi, P. (2006). Automatic extraction of meaning from the web. In IEEE International Symp. Information Theory.

[Cilibrasi and Vitányi, 07] Cilibrasi, R.L. and Vitányi, P.M.B. (2007). The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering,

[Daim et al., 06] 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., 05] 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., 06] 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.

[Debecker and Modis, 94] Debecker, A. and Modis, T. (1994). Determination of the uncertainties in s-curve logistics. Technological Forecasting and Social Change, 46(2):153–173.

[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.

[Eto, 03] Eto, H. (2003). The suitability of technology forecasting/foresight methods for decision systems and strategy: A Japanese view. Technological Forecasting and Social Change, 70(3):231-249.

[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.

[Jones et al, 01] Jones, E., Oliphant, T. , Peterson, P. and others. (2001-) SciPy: Open Source Scientific Tools for Python, http://www.scipy.org.

[Jordan, 07] Jordan, S. W. (2007). Technology forecasting with science indicators: The case of laptop battery futures. In Management of Engineering and Technology, Portland International Center for, pages 1643–1650.

[Kajikawa et al., 07] Kajikawa, Y., Yoshikawa, J., Takeda, Y., and Matsushima, K. 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 Korea's 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.

[Korte and Vygen, 2006] Korte, B. and Vygen, J. (2006). Combinatorial Optimization: Theory and Algorithms. Springer, Germany, 3rd edition.

[Kostoff et al, 07] Kostoff, R.N., Koytcheff, R.G. and Lau, C.G.Y. (2007). Global nanotechnology research metrics. Scientometrics, 70(3):565-601.

[Kostoff, 01] Kostoff, R. N. (2001). Text mining using database tomography and bibliometrics: A review. 68:223–253.

[Kuchling, 98] Kuchling, A (1998). The Python DB-API. Linux Journal, 1998, Issue 49es.

[Landauer et al., 1998] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis.

[Lee et al, 08] Lee, C.Y., Lee, J.D. and Kim, Y. (2008). Demand forecasting for new technology with a short history in a competitive environment: the case of the home networking market in South Korea. Technological Forecasting and Social Change, 75:91-106.

[Li, 2001] Li, Y. (2001). An effective implementation of a direct spanning tree representation in gas. pages 11–19.

[Li and Bouchebaba, 2000] Li, Y. and Bouchebaba, Y. (2000). A new genetic algorithm for the optimal communication spanning tree problem. pages 162–173.

[Losiewicz et al., 00] 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, 93] Martino, J. (1993). Technological Forecasting for Decision Making. McGraw-Hill Engineering and Technology Management Series

[Martino, 03] Martino, J. P. (2003). A review of selected recent advances in technological forecasting. Technological Forecasting and Social Change, 70(8):719–733.

[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.

[Michener and Sokal, 57] Michener, C.D. and Sokal, R.R. (1957). A quantitative approach to a problem in classification. Evolution, 11:130-162.

[Modis and Debecker, 88] Modis, T. and Debecker, A. (1988). Innovation in the computer industry. Technological Forecasting and Social Change, 33(3):267–278.

[Porter, 07] Porter, A. (2007). How “Tech Mining” can enhance R&D management, Research Technology Management, 50(2):15-20, 2007.

[Porter, 05] Porter, A. (05). Tech mining. Competitive Intelligence Magazine, 8(1):30–36.

[Porter et al., 91] Porter, A., Roper, A., Mason, T., Rossini, F., and Banks, J. (1991). Forecasting and Management of Technology. Wiley-Interscience, New York.

[Raidl, 2000] Raidl, G. R. (2000). An efficient evolutionary algorithm for the degree-constrained minimum spanning tree problem. In Evolutionary Computation, 2000. Proceedings of the 2000 Congress on, volume 1, pages 104–111 vol.1.

[Saitou and Nei, 87] 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, 69] Sammon, J. (1969). A Non-linear Mapping for Data Structure Analysis, IEEE Transactions on Computers, C-18:401-409.

[Smalheiser, 01] Smalheiser, N. R. (2001). Predicting emerging technologies with the aid of text-based data mining: the micro approach. Technovation, 21(10):689–693.

[Small, 06] Small, H. (2006). Tracking and predicting growth areas in science. Scientometrics, 68(3):595–610.

[Trappey and Wu, 07] Trappey, C. and Wu, H.-Y. (2007). An evaluation of the extended logistic, simple logistic, and Gompertz models for forecasting short lifecycle products and services. Complex Systems and Concurrent Engineering, 793–800.

[van der Heijden, 00] van der Heijden, K. (2000). Scenarios and Forecasting: Two Perspectives. Technological Forecasting and Social Change, 65:31-36.

[Zhu and Porter, 02] Zhu, D. and Porter, A. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change, 69:495-506.

PUBLICATIONS/PRESENTATIONS

The following is a list of publications, presentations and working papers which describe and record the work carried out in this project.

As can be seen, two papers are currently under review in peer-reviewed, ISI-indexed journals, but in the upcoming stages of the project it is expected that at least some of the working papers listed below will be expanded and enhanced, leading to submission to relevant journals.

Journal submissions

"Semantic distances for technology landscape visualization" (2008), W.L. Woon and S.E. Madnick, (under review) *Scientometrics*.

"Asymmetric information distances for automated taxonomy construction " (2008), W.L Woon and S.E. Madnick (under review) *Knowledge and Information Systems.*

Presentations

"Data Mining and Semantics: An application in Technology Forecasting" (2008), W.L. Woon and S.E. Madnick, MIT Center for Digital Business Annual Sponsors' Conference, poster presentation.

"Technology Forecasting using Data Mining and Semantics" (2008), W.L. Woon and S.E. Madnick, MIT-Masdar Symposium, poster presentation.

MIT working papers

"Semantic distances for technology landscape visualization" (2008), W.L. Woon and S.E. Madnick, working paper CISL #2008-04 (Sloan School of Management, MIT).
http://web.mit.edu/smadnick/www/wp/2008-04.pdf

"Asymmetric information distances for automated taxonomy construction" (2008), W.L. Woon and S.E. Madnick, working paper CISL #2008-05 (Sloan School of Management, MIT).
http://web.mit.edu/smadnick/www/wp/2008-05.pdf

"Technological Forecasting - a Review" (2008), A.K. Firat, S.E. Madnick and W.L. Woon, working paper CISL #2008-15 (Sloan School of Management, MIT).
http://web.mit.edu/smadnick/www/wp/2008-15.pdf

"Comparison of Approaches for Gathering Data from the Web for Technology Trend Analysis" (2008), A.K. Firat, S.E. Madnick and W.L. Woon, working paper CISL #2008-14 (Sloan School of Management, MIT). **http://web.mit.edu/smadnick/www/wp/2008-14.pdf**

"Latent Semantic Analysis applied to tech mining" (2008), B. Ziegler, W.L. Woon and S.E. Madnick, working paper CISL #2008-12 (Sloan School of Management, MIT).
http://web.mit.edu/smadnick/www/wp/2008-12.pdf

"Research Plan for Leveraging Social Information Systems: Using Blogs to Inform Technology Strategy Decisions", S. Seshasai, working paper CISL #2008-07 (Sloan School of Management, MIT).
http://web.mit.edu/smadnick/www/wp/2008-07.pdf

APPENDICES

A: Project Schedule (unmodified copy from the project proposal)

Tasks \ Time period	10/07-12/07	01/08-03/08	04/08-06/08	07/08-09/08	10/08-12/08	01/09-03/09	04/09-06/09	07/09-09/09
Survey of databases								
Tool development								
Base indicators								
Base visualizations								
Contextual/semantic extensions								
Enhanced visualizations								
Data analysis								
Milestones		**M1**		**M2**		**M3**		**M4**

Figure 6. *Project schedule*

B: Keyword collections

B1: Author keywords 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

B2: Keyword plus

16s ribosomal-rna, activation, active-site, adsorption, agrobacterium-mediated transformation, anabaena-ariabilis, anacystis-nidulans, aqueous ammonia, bidirectional hydrogenase, biomass conversion processes, briquettes, canopy structure, catalysts, cds, cdte, cells, cellulases, cellulose, ch4, chalcopyrite, charge-ransfer 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.