Technology Forecasting Using Data Mining and Semantics: Third & Final Annual Report

Stuart Madnick, Wei Lee Woon, Andreas Henschel,
Erik Casagrande, Ayse Firat, Steven Camina, Blaine Ziegler,
Satwik Seshasai, Georgeta Vidican, Hatem Zeineldin, Toufic Mezher,
Isam Janajreh, Gihan Dawelbait

Working Paper CISL# 2011-01

May 2011

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

TECHNOLOGY FORECASTING
USING DATA MINING AND SEMANTICS

FINAL REPORT

MIT/MIST Collaborative Research Progress Report for Period 1/1/2010 to 31/12/2010

Principal Investigator at MIT

Professor Stuart Madnick

Principal Investigator at MIST

Dr. Wei Lee Woon

Team-members

Andreas Henschel (MIST Postdoctoral Researcher)

Erik Casagrande (MIST Postdoctoral Researcher)

Ayse Firat (MIT M.Sc Candidate/Research Assistant)

Steven Camina (MIT M.Eng Candidate/Research Assistant)

Blaine Ziegler (MIT M.Eng Candidate/Research Assistant)

Collaborators

Satwik Seshasai (MIT Ph.D Candidate/Collaborator)

Dr. Georgeta Vidican (MIST Assistant Professor, Collaborator)

Dr. Hatem Zeineldin (MIST Assistant Professor, Collaborator)

Dr. Toufic Mezher (MIST Professor, Collaborator)

Dr. Isam Janajreh (MIST Associate Professor, Collaborator)

Dr. Gihan Dawelbait (MIST Postdoctoral research assistant, Collaborator)

Research Project Start Date

10/01/2007

Table of Contents

EXECUTIVE SUMMARY

The planning and management of research and development is a challenging process, and one that is further compounded by the increasing rate of technological progress. Nevertheless, the importance of properly distributing limited resources and on focusing only on research directions with the potential for high growth means that this process is of critical importance.

To keep up-to-date with these developments, experts often make use of the scientific literature but the acceleration of technological progress has also resulted in a corresponding increase in the volume of relevant information that is available. Even an expert is unlikely to know all the new developments in his field.

Another concern is the inherent subjectivity of existing decision-making processes. Scientists, research strategists and other decision makers often rely on intuition and domain knowledge to arrive at management or investment decisions. While expert decisions can and often do produce well-informed and effective decisions, the problem is that these are still subject to individual perspectives and human biases. Furthermore, it can be difficult to document the reasons and contexts for decisions which depend heavily on internalized knowledge and experience.

These facts provide strong motivation for the use of automated methods of processing and analyzing large datasets, often known as "data mining." The overall goal of this project is to mine science and technology publication databases for patterns and trends which can facilitate the formation of research strategies. Examples of the types of information sources that are available are very diverse and could include academic journals, patents, blogs and news stories. However, to keep the scope of our research activities manageable, we have focused our efforts on academic publications and, to a lesser extent, patents and blogs.

It must be emphasized that the goal of this research is not to replace the important role of experts in making research and investment decisions, but to help them to improve the quality of their decisions by calling to their attention emerging technologies that they might have otherwise overlooked.

The original proposed outputs of the project were:

1. A **case study** of the renewable energy domain, including tentative forecasts of areas with high potential for future growth. A secondary element of these reports is the identification of influential researchers or research groups.

2. An improved **understanding** of the underlying research landscape, represented in a suitable form such as a renewable energy ontology.
3. Scholarly **publications** in respected and peer-reviewed journals and conferences.
4. Software **tools** to automate the developed techniques.

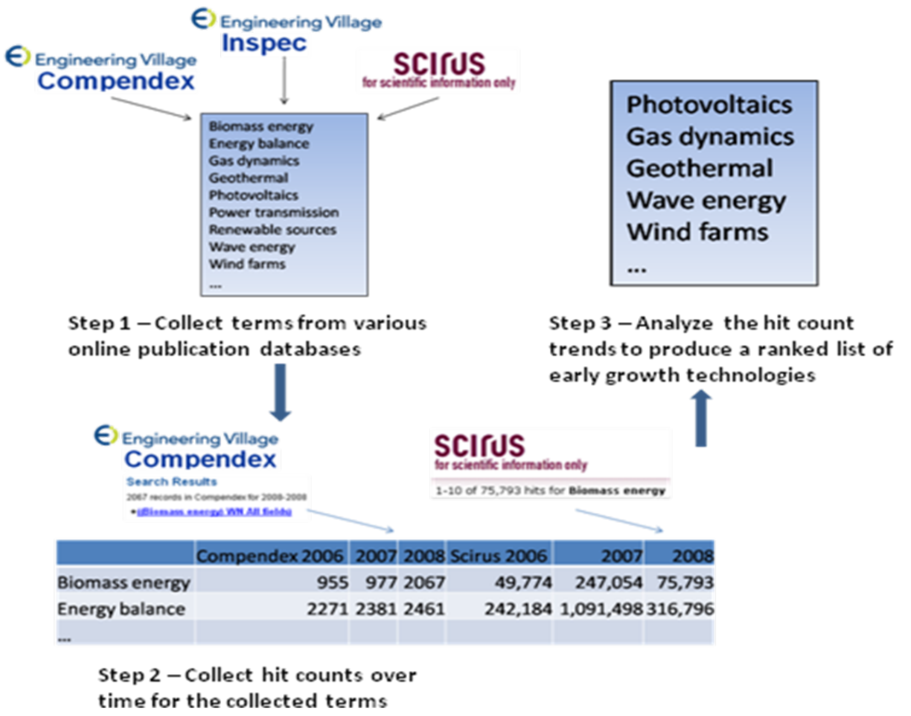
To achieve these goals, we have developed a comprehensive framework which incorporates all of the required stages of technology forecasting. The main components of this framework are:

1. **Keyword discovery** – this important component of the framework forms the keystone for successful completion of the other tasks. Extracting keywords which are relevant to the domain of interest is of particular importance because in later stages these keywords will be used to represent the individual technologies or research directions which constitute of the broader domain being studied.
2. **Organizing** these terms using semantic distances and self-organizing methods of data visualization; tools drawn from the fields of data mining and pattern recognition are used to automatically create structures for organizing and visualizing the so-called "research landscape" of the domain. These are useful in their own right (as a summary of the overall area of research) but are also used in a later stage (step 5 in this list) of the framework to incorporate semantic information into the technological forecasting process.
3. Extraction of **numerical features** for measuring the *prevalence* of areas of research. Ideally we would like to be able to directly measure the amount of research activity that is relevant to each research domain. However as this is not possible the alternative is to find suitable numerical features or indices which can serve as proxy variables for the level of research activity. We find that the frequency at which a particular term is observed in the literature presents one such feature, subsequently referred to as the *term frequency* (TF).
4. **Detecting and highlighting** research areas which appear to be highly promising. Once the prevalence measures described above have been described, their temporal evolution patterns can be used to identify technologies which are growing quickly, or which are on the verge of rapid growth. In this report, such technologies will be referred to as "*early growth*" technologies, or are said to be in the "early growth" phase of development.

5. Enriching these measures via the **semantic distance** measures. Early growth technologies are, almost by definition, relatively unknown and have comparatively low publication volumes; as such it can be difficult to accurately measure their levels. To help counter this problem, we explore the use of keyword taxonomies developed as part of the framework to aggregate and smooth the growth measures derived from the prevalence measures.

6. **Presenting** the results of this analysis in an intuitive and visually manner.

Some of the key steps, and how they relate to each other, are illustrated in the figure below:



	Compendex 2006	2007	2008	Scirus 2006	2007	2008
Biomass energy	955	977	2067	49,774	247,054	75,793
Energy balance	2271	2381	2461	242,184	1,091,498	316,796
...						

Step 1 – Collect terms from various online publication databases

Step 2 – Collect hit counts over time for the collected terms

Step 3 – Analyze the hit count trends to produce a ranked list of early growth technologies

Finally, it is worth noting that, as this project has been in progress for over three years now, the original proposed outputs have expanded and evolved somewhat. The following section, which reviews the main accomplishments of the project, provides a more comprehensive reflection of the project scope.

Review of Objectives and Accomplishments

All of the stated goals of this project have either been accomplished or are still in progress in view of unforeseen expansions in the scope of the research. In addition, several additional outcomes have emerged during the execution of the research project. The following is a review of the main achievements and progress made during the course of this research.

1. **Data collection and term extraction** – Various tools and techniques were developed to support the automated collection of data from online sources of data, and the extraction of relevant keywords from these collections. We've also worked on a technique where by data from Wikipedia can be used to enrich an existing core set of keywords.

2. **Taxonomy generation** – We studied a number of approaches for automatically organizing and visualizing our collection of relevant keywords. Primarily these were in the form of taxonomies which organized the keywords in a hierarchical manner. Three main approaches were investigated:

   - A Genetic Algorithms based approach.
   - A variation on the approach described in Heymann and Garcia-Molina, 2006, for which we developed a number of important modifications to support the application on technology forecasting.
   - A semi-automatic approach which would allow for expert opinion to be integrated into the taxonomy generation process.

3. **Early growth indicators** – A set of numerical indices for evaluating the growth potential of individual keywords were identified and tested. These are fairly simple statistics but can be quickly applied to obtain "scores" for a large number of early-growth candidates. We also developed a technique by which the scores for individual keywords can be aggregated via the above-mentioned taxonomies to obtain more reliable results.

4. **Renewable energy case study** – Several case studies focusing on individual domains of renewable energy have been completed. These have included: (i) Solar PV (ii) Solar desalination (iii) Geothermal (iv) Distributed Generation (v) Waste to Energy, and (vi) Cybersecurity. While topic (vi) is not directly related to renewable energy, we felt that it was still indirectly relevant, and was also a useful and interesting test case for our methods.

   Our main finding was that it was somewhat difficult to produce a single overall case study that covers all aspects of renewable energy, but that smaller studies that targeted specialized domains within renewable energy seemed to be just as useful. Also, please note that this portion of the project is still active. Dr. Andreas Henschel, the postdoctoral researcher on this project, is contracted until the end of May, and is currently conducting a solar PV related case study amongst other activities.

5. **Development of software tools** – A very large body of software was developed and continues to be refined, as part of this project. In addition, two user-friendly GUI tools were developed in earlier iterations of the project – the "Cameleon Scheduler" and the "Early Growth Analysis tool". In the latest reporting period, we also created a tool for visualizing topic maps.
6. **Collaborations with other MIST faculty -** Collaborations were developed with four other MIST faculty members namely Dr. Georgeta Vidican, Dr. Hatem Zeineldin, Dr. Isam Janajreh and Dr. Toufic Mezher and MIT PhD student Satwik Seshasai. Most of these collaborations have resulted in research results in the form of papers submitted to conferences, journals or as working papers.
7. **Completed 25 research papers -** of which 7 have been published in peer-reviewed publications, 1 has been accepted subject to revisions, 2 are under review, 3 Master's theses and 1 Doctoral Dissertation at MIT, and 11 distributed as MIT Working Papers on the Social Sciences Research Network (SSRN). These are currently being expanded and will be submitted for consideration in reputable journals and conferences.

Report overview

The rest of this report is structured as follows.

The **Introduction** section (*pg.10*) reviews key ideas and motivations for the project.

The **Research Tasks** section (*pg.12*) comprises the bulk of the report, and is organized into several subsections. As this is the final report in this project, this section will seek to sum up the overall activities and achievements of the research project.

For clarity, this section will be further divided into two main subsections – in the first subsection, the *Technology Forecasting Framework* (*pg.14*) mentioned above will be described. In this subsection, we also provide a detailed description of the *reference implementation* of the framework, which is a set of techniques that instantiate each component in the framework. Having this reference implementation is important to provide stability as novel ideas are constantly being proposed and tweaked for each of the components of the framework.

That said, a lot of additional work has gone on outside this primary implementation (this is similar to "stable release" and "experimental features" branches in a software development project). The second subsection focuses on the latter, where three particularly interesting directions are described in detail: the first is in the area of *probabilistic topic modeling* (*pg.28*), which is an alternative way of analyzing underlying trends and concepts in document corpora. The second direction is the *bibliometric analysis of blogs* (*pg*.46), which was a research activity undertaken in collaboration with MIT PhD student Satwik Seshasai. The final effort is an *evaluation of taxonomy generation algorithms* (*pg.46*).

The **Project Summary** (*pg.104*) reviews and discusses the main findings of the project and specifies the relative divisions of labor between the teams at MIT and MIST.

Finally, while we have already reached the official end date of the project, the related research work will continue for a while more as we have been able to extend the stays of the postdoctoral researchers on this project. The section on **Future Work** (*pg.106*) will describe the research activities which will be assigned to him for this period as well as directions for future research.

Finally, there is a series of **Appendices** (*pg.121*) which provides supplementary materials.

INTRODUCTION

Background

For decision makers and researchers working in a technical domain, understanding the state of their area of interest is of the highest importance. Typically, these fields of research are not homogeneous but rather can be thought of as consisting of many subfields and underlying technologies which are related in intricate ways. In addition, this composition, or *research landscape*, is not static as new technologies are constantly being developed while existing ones become obsolete, often over very short periods of time. This constant process of reorganization means that researchers working in presently unrelated domains might one day become dependent on each other's findings and expertise.

Against this scenario, 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 deciding NSF and NIH grant awards [Porter, 07], while many countries spend huge sums on technology foresight programs which consist of elaborate surveys and opinion aggregation [Eto, 03][Bengisu and Nekhili, 2006]. While expert opinion is a hugely important component in the decision making process, it can have a number of shortcomings when used on its own. 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 they 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" (or sometimes "Tech-Mining") [Porter and Newman, 2011][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].

In particular, our research has addressed the challenge of *technology forecasting*. In contrast to the large body of work already present in the literature, there is currently very little research that attempts to provide concrete, actionable results on which researchers and other stakeholders can base their actions.

General Approach

The high-level aim of the project is to create improved methods for conducting technology mining - i.e.: a combination of technology related activities which includes forecasting, mapping and visualization (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 include various modeling 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.

In response to these principles, we describe a novel framework for automatically visualizing and predicting the future evolution of domains of research. Our framework incorporates the following three key contributions:

1. A methodology for automatically creating taxonomies from bibliometric data. A number of approaches have been tested where the basic principle is to assign terms that co-occur frequently to common sub-trees of the taxonomy.
2. A set of numerical indicators for identifying technologies of interest. In particular, we are interested in developing a set of simple growth indicators, similar to technical indicators used in finance, which may be easily calculated but which can be applied to hundreds or thousands of candidate technologies at a time. This is in contrast to more traditional curve fitting techniques which require relatively larger quantities of data.

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

RESEARCH TASKS

As this is the final report for the project, this report will attempt to summarize the overall achievements of the project, though extra emphasis will be given to the progress and activities since the last progress report (i.e. for the last year of the project).

As the research activities undertaken as part of the project are extremely broad in scope, it has been helpful to divide the main research tasks into three functional "blocks", which are illustrated in Figure 1. The nature and purpose of each of the three blocks are:

Block (a): Encapsulates all the techniques for collecting and storing data from generic sources (which will frequently be online and unstructured), and in transforming them in a way which facilitates the subsequent stages.

Block (b): Methods for detecting keywords or technologies which are promising, or more generally, for sorting keywords according to their degree of promise.

Block (c): Techniques for visualizing and organizing the technologies being studied. Taxonomies are one way of achieving this but in recent times we have also reviewed other possibilities. Note that this block receives inputs from both the other blocks – this is because the final output of the system is frequently captured in the overlay of the growth indicators on the visualizations.

However, it is important to differentiate between the framework depicted in Figure 1, and the specific set of techniques implemented by our research team, which instantiates each of these blocks (the "reference implementation" mentioned previously). The former is a conceptual model which abstracts the entire process of extracting data from generic, textual databases, while the latter is provided only as an concrete example of this framework. An important implication of this distinction is that alternative implementations can easily be created and customized to meet the requirements of specific applications.

A further issue that should be noted is that this reference implementation provides stability to the results and analysis produced in this research. Over the three years during which this project has been in progress, we have designed, implemented and tested a very wide array of techniques addressing each of the three blocks above. Some of the techniques work best with certain configurations of the system while in other cases, individual pieces of this system can be used in isolation to produce results which are interesting in their own right.

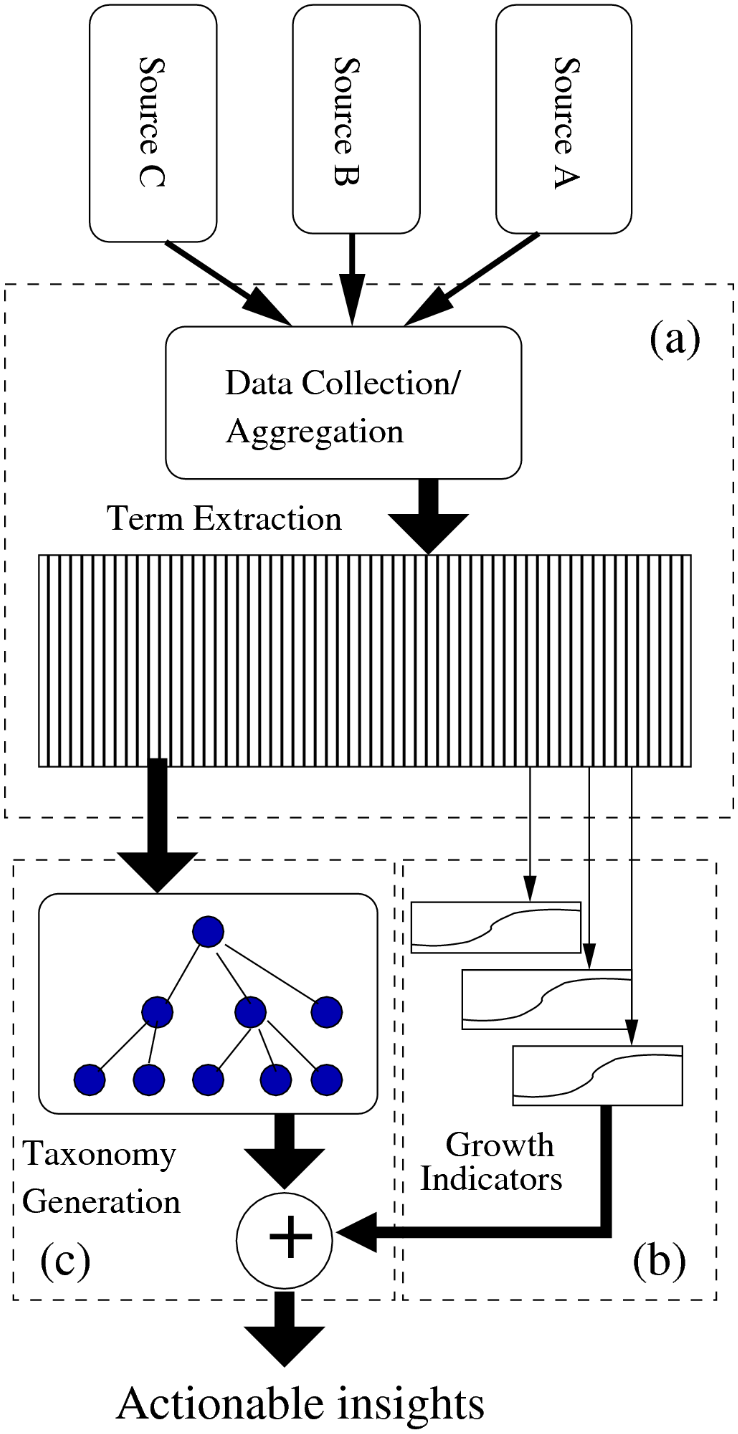


Figure 1 Schematic of tech-forecasting system

The reference implementation hence represents the “official version” of the framework but, as mentioned previously, a lot of work has been done outside of this main branch to each of the components. While a lot of this work is very promising, integrating every proposed technique and approach into the full reference implementation is a large task and one which will have to be the subject of future research efforts.

In the rest of this section on Research Tasks, each of these blocks will be discussed in a separate section. There will also be two additional sections: one will be on probabilistic topic modeling, which is a new research direction which has emerged largely over the last year of the project, while the other will be on a study of methods used to evaluate the technology visualizations.

Block (a): Data collection and term extraction

Data collection

The type of data source, collection mechanism and number of sources used can be modified as required but for the reference implementation, information extracted from the Scopus[1] database was used (in addition, the team has MIT has also used a variety of other sources including Engineering Village, Scirus, Inspec and Google Blog Search). Scopus is a subscription-based, professionally curated citations database provided by Elsevier. Other possibilities, such as Google’s scholar search engine and ISI’s Web of Science database were also considered and tested (the results of these tests were reported on in previous progress reports) but Scopus proved to be a good initial choice as it returned results which were generally of a high quality, both in terms of the publications covered and relevance to search terms, and was normally able to retrieve a reasonable number of documents. In addition, the Scopus database included information regarding relevant publications as well as patents.

The database used in this project was a combination of a number of subject specific databases and was constructed by collecting records resulting from searches for the following keywords:

{biofuel, biodiesel, distributed generation, dispersed generation, distributed resources, embedded generation, decentralized generation, decentralized energy, distributed energy, on-site generation, fuel cell, geothermal, photovoltaic*, solar cell*, renewable energy, solar AND cooling, waste AND*

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

(gasification OR pyrolysis OR syngas OR fermentation OR esterification OR bioreactor OR biomass OR devulcanization OR depolymerization OR reforming), wind power, wind energy}

This list of terms was generated based on discussions with relevant subject matter experts from amongst the Masdar Institute faculty.

While the Scopus dataset provided a useful amount of data on which to test our methods, one of the conclusions of our study was that a full industrial Technology Forecasting study would require a much larger collection of data. Such datasets can be purchased from many of the academic indexing services currently in operation.

Term Extraction

Term extraction is the process of automatically generating a list of keywords on which the technology forecasting efforts will be focused. There are a many ways in which this can be achieved and during the course of this project we have experimented with a number of these (our experiences have been documented in [Ziegler et al., 2009]). Overall, our observation was that there wasn't a single "optimum" method for keyword extraction, but that, depending on the field of research, a given technique might be more (or less) appropriate; alternatively, keywords could be collected using multiple techniques and merged to form a comprehensive term list. In earlier reports, we had described two broad approaches, which were as follows:

1. Extraction using built-in extensions of search operations. For example, the Scirus search engine provides a "refine your search" option which lists relevant search terms. We have developed software tools for automating this process as well as incorporating additional relevance checks to ensure the quality of the retrieved corpus of keywords.
2. Collection of keywords from document abstracts and databases. Academic papers are often associated with a set of relevant keywords to facilitate indexing and categorization. These keywords can be collected and filtered to provide a list of subject-specific phrases for use with technology forecasting.
3. A third approach was to use Wikipedia. Categories relevant to the research landscape being studied were first identified. The titles for articles associated to each of these are then collected and are used to seed the keyword search. In addition, Wikipedia provides a separate database with

metadata such as association of articles to categories, abbreviations and alternative spellings. All of this information is taken into account when generating the augmented keyword list.

For the reference implementation, the second approach is used. For each document retrieved, a set of relevant keywords is provided. These are collected and, after word stemming and removal of punctuation marks, sorted according to number of occurrences in the text. To help ensure relevance, the following measures were adopted:

1. To remove overly generic terms like "priority journal" and "international conference", a search for non-energy related keywords is conducted in parallel and top-ranking keywords are extracted and used as a stopword list.
2. We also noted the presence of a large number of geographical terms like "America" and "Southern Europe". While these are undoubtedly relevant in the context of the individual papers they again are simply too generic to be of use in a broader domain-level context. As such, a list of such terms was compiled semi-automatically (for example by parsing a list of the countries of the world) and merged with the previous list of stopwords.
3. As a final pass, a manual sweep of the remaining keywords was performed and an additional set of terms which were too generic or context-dependent were added to the list of stopword. The terms added in this stage were as follows:

   {*elsevier (co), surveys, marketing, technologies, light, reliable, products, reviews, speed, humans, comparative studies, probable, test, 21st centuries, innovation, air, case study, lead, vegetable, matlab, customer satisfaction, engineering research, extended abstract, sales, probability distribution, surveys, future prospect, usa., greece, solid, exhibitions, students, renewable resource, electric powers, electric supplies, applications, manager, international (co), low-cost, solid wastes*}

Since the last update, we have also studied a range of more advanced term extraction algorithms. In particular, we note that the state of the art of automated techniques includes three different approaches:

- **Linguistic features for keyword extraction** have been proposed, such as part of speech tags and part of speech tag patterns (for phrases), have been proposed (Hulth, 2003).

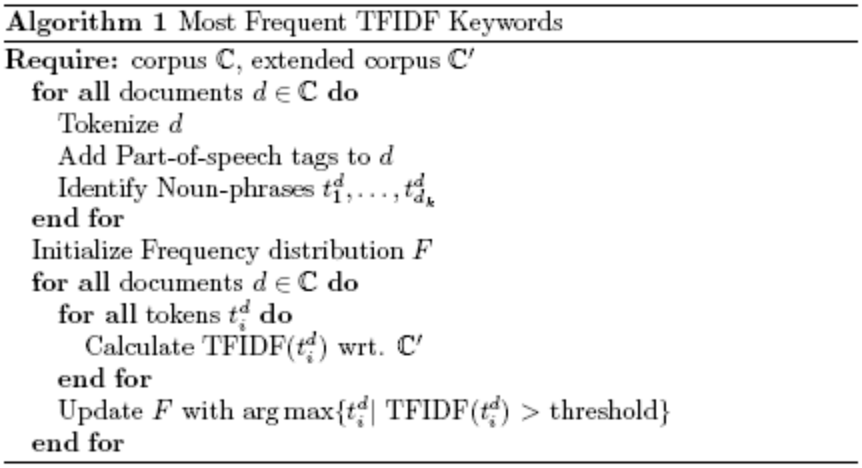
- **Supervised Machine Learning techniques:** they take as training data a set of documents for which keywords have been assigned manually. Documents are represented with features using the abovementioned techniques (Turney, 2000; Hulth, 2003).
- **Statistical Natural language processing (NLP) based techniques:** In order to detect potential candidates, N-gram (often unigram, bigram and trigram) models have been suggested (Manning, 1999). They provide a mean to estimate the probability of observing a phrase in a text using conditional probabilities of the n-1 preceding words. Further estimates for the significance of a term in a document are term frequency (TF), inverse document frequency (IDF), their product (TFIDF) and the position of a term in a document.

The third technique in particular came in especially useful when conducting the waste-to-energy case study [Henschel et al, 2010a], and will now be described in slightly more detail.

NLP based term extraction

Multiple-word noun phrases are essential for Technology Forecasting, since many technology descriptions in the English language are composed of more than one word. Noun phrases can be detected with reasonable accuracy using a chain of state-of-the-art tools from Natural Language Processing, typically sentence and word tokenization, part-of-speech tagging and noun phrase chunking [Loper and Bird, 2002]. Noun phrases can be compounds of nouns (“waste combustion”), adjective noun phrases (“thermal treatment”), prepositional and noun phrases (“board of directors”). Statistical significance of words or word phrases can then be estimated using information retrieval measures, in order to avoid irrelevant terms which are not specific to a certain technology, e.g. “review” or “approach”. These terms cause confusion in the downstream data processing such as the taxonomy creation. Ideally the extracted terms should be of low ambiguity and high specificity. These properties are preconditions for the conceptualization of a knowledge domain and the creation of a domain taxonomy/ontology. One advantage of using an automated term extraction tool in addition to a manual one is that emerging technology terms and research field names can potentially be detected before they are recognized and consistently attached as keywords to articles. It also allows for working with data sources having poor or no keyword annotations, such as a variety of blogs.

It is important to emphasize that simple features as employed in KEA [Witten, 1999] using TFIDF and Naive Bayesian Classifiers perform reasonably well in comparison to sophisticated Machine Learning approaches. Moreover, statistical methods do not require any training data, are straightforward to implement and are very efficient. For a thorough review of keyword extraction methods, the reader is referred to [Pazienza et al., 2005].



```
Algorithm 1 Most Frequent TFIDF Keywords
Require: corpus $\mathbb{C}$, extended corpus $\mathbb{C}'$
  for all documents $d \in \mathbb{C}$ do
    Tokenize $d$
    Add Part-of-speech tags to $d$
    Identify Noun-phrases $t_1^d, \ldots, t_{d_k}^d$
  end for
  Initialize Frequency distribution $F$
  for all documents $d \in \mathbb{C}$ do
    for all tokens $t_i^d$ do
      Calculate TFIDF($t_i^d$) wrt. $\mathbb{C}'$
    end for
    Update $F$ with $\arg\max\{t_i^d \mid \text{TFIDF}(t_i^d) > \text{threshold}\}$
  end for
```

Figure 2 Keyword Extraction Algorithm

Figure 2 summarizes the approach described above. Basic NLP functionality such as tokenizers and word stemming was provided by the NLTK toolbox [Loper and Bird, 2002].

The algorithm generates keywords for each abstract by identifying noun phrases. These are then scored by the TFIDF scheme, resulting in a collection of TFIDF-keywords. The TFIDF of a keyword *ti* is given by the product of the term frequency TF[*ti,d*], i.e. the number of times a term ti occurs in a document d divided by the number of words in that document, and the inverse document frequency IDF[*ti*] i.e. the logarithm of the number of all documents divided by the number of documents where the term occurs. The former quantifies the emphasis of a word in terms of number of repetitions in a document while the latter makes sure that words occurring almost everywhere are downgraded. We then select the most frequent of these TFIDF-keywords. As a result, the most frequently occurring words of the Waste to energy corpus such as results, affect, study and paper are not present in the list of most frequent TFIDF-keywords.

Unfortunately, general or more abstract terms that are still useful for taxonomy creation such as process, temperature or biomass are also eliminated because they are abundant in the selected corpus. This effect can be mitigated by extending the corpus via the addition of unrelated scientific documents (denoted C’ in the algorithm); this allowed terms that were specific to a certain research field to be distinguished from terms which are more broadly used.

Block (b): Identification of early growth technologies

To identify promising technologies, we first need a suitable measure for the “prevalence” of a given technology as a function of time, as this would then allow us to measure its growth over time. It is difficult to achieve this directly but an indirect means of doing so would be to observe the occurrence statistics of terms relevant to the domain of interest. To allow for the overall growth in publication numbers over time (given the emergence of new journals, conferences, etc.), we choose to use the *term frequency* instead of the raw occurrence counts.

Once the term frequencies for all terms have been extracted and saved, they can be used to calculate growth indicators for each of the terms. These, in turn, are used to rank the list of terms. As stated previously, we are particularly interested in keywords with term frequencies that are relatively low at present but that have been rapidly increasing; in [Ziegler et al., 2009][Ziegler et al, 2009b] this is first referred to as the “early growth” phase of technological development, and represents the fields to which an expert would wish to be alerted. Existing techniques are often based on fitting growth curves (see [Bengisu and Nekhili, 2006] for example) to the data. This can be difficult as the curve-fitting operation can be very senstive to noise. Also, data collected over a relatively large number of years (approximately ≥10 years) is required, whereas the emergence of novel technological trends can occur over much shorter time-scales.

The search for suitable early growth indicators is currently still an area of active research. A sampling of the more promising indicators were (more documentation on our efforts in this area is provided in [Ziegler et al., 2009,Ziegler et al, 2009b]):

Mean publication year – calculated over the range of years for which the growth indicator is estimated, this statistic measures the currency of a given topic over the range of years studied; i.e. a more recent mean publicaion year could be an indicator of a research topic that is trending upwards.

$$\theta_i = \frac{\sum_{y \in Y} y.TF_y[t_i]}{\sum_{y \in Y} TF_y[t_i]} \quad ..... \quad (1)$$

Log growth rate – this is essentially a coarse measure of the rate at which a given technology has grown during the analysis period. The idea is that technologies in the early growth phase could have very low initial TF values but could still be growing rapidly and hence worthy of attention.

$$\theta_{0,i} = \log TF(t_i) - \log TF(t_0) \quad . . . . . \quad (2)$$

Second order growth indicator - based on the shape of the graph in Figure 2, it might be possible to detect technologies which are in the early growth phase using an approximation of the second order derivative of :

$$\dot{\theta}_i = \theta_{n/2,n} - \theta_{0,n/2} \qquad . . . . . \quad (3)$$

where are the three growth indicators measured over the range of years from t_0 to t_n for keyword i, and $TF_i[t]$ is the term frequency for term i and year t. As can be seen, this gives the average publication year for articles appearing in the five year period from 2003 until 2008, and which are relevant to term i (a more recent year indicates greater currency of the topic).

At this point in time, the choice of the best early growth indicator is under study and is the subject of a M.Sc. thesis project that is being undertaken at Masdar Institute.

Block (c): Keyword taxonomies and semantics enriched indicators

One of the problems encountered in earlier experiments involving technology forecasting is that there is a lot of noise when measuring technology prevalence using simple term occurrence frequencies.

This is a fundamental problem when attempting to infer an underlying property (in this case, the size of the relevant body of literature) using indirect measurements (hit counts generated using a simple keyword search), and cannot be entirely eliminated. However, as part of our framework we propose an approach through which these effects may be reduced; the basic idea is that hit counts associated with a single search term will invariably be noisy as the contexts in which this term appear will be extremely diverse and will contain a large number of extraneous mentions (and will also include papers which are critical of the technology it represents). However, if we can find collections of related terms and use aggregate statistics instead of working with individual terms, we might reasonably expect that a lot of this randomness will cancel out.

We concretize this intuition in the form of a *predictive taxonomy*; i.e. a hierarchical organization of keywords relevant to a particular domain of research, where the growth indicators of terms lower down in the taxonomy contribute to the overall growth potential of higher-up "concepts" or categories.

In earlier reports we had already developed the theory behind our proposed methods for automatic taxonomy generation, so these will not be described in detail in this report. However, in the period since, we have experimented with a number of new techniques for taxonomy generation, one of which was the use of expert knowledge to support and enhance the taxonomy generation process.

One area in which this appeared to be of particular benefit was in the domain of Waste to Energy (W2E) conversion. We here examine the differences between the two approaches to taxonomy generation.

Taxonomy creation: Fully automated taxonomies

Firstly, the fully automated taxonomy creation process was used to analyze the data. As mentioned previously, our approach is based on the Heymann-Algorithm as described in Section "Taxonomy Creation". To demonstrate the applicability of this approach to technology forecasting, we apply it to a number of different domains. In general, the taxonomies resulting from these analyses are quite large, so what we show in the following pages is a sampling of interesting sub-trees which have been extracted from the corresponding publication corpora.

It is also important to note that the algorithms demonstrated here have a number of variations or "settings" which control the execution of the algorithm. Examples of these include the number and selection of keywords used, the type of centrality measure and the type of similarity metric used to compare the tags. We concede that varying these settings can significantly alter the resulting taxonomy. However it is not within the scope of this chapter to systematically investigate the effect each of these settings has on the taxonomy generation process; instead, readers are referred to (Henschel et al., 2009, Camina, 2010), which provide a much more detailed treatment of this issue. The subtrees presented here are chosen to be typical of the kinds of results that were obtained, and are aimed at providing the reader with an idea of the capabilities of our approach.

All taxonomies were generated using the Heymann algorithm, and the Sine distance was used to create the distance matrices (this is the distance-based analog of the Cosine distance). The number of keywords used for each taxonomy ranged from 100 to 400.

Our main observations are as follows:

1. The quality of the results varied significantly between domains and between subtrees within the same domain.
2. The two W2E subtrees (“biomass” and “wastewater”) are significantly larger and more complex than the other subtrees shown here, and helped to highlight the performance of the algorithm with respect to very complex taxonomies. Broadly speaking, the two taxonomies seemed to provide a good illustration of their respective subject areas. However, upon closer inspection, we see that there are a number of irregularities, which would merit further study.
3. In Figure 3, the series of nodes from “heavy metal” to “mercury” represent compounds which are related but which are clearly not subclasses of each other. A similar situation is encountered with the “granulation” to “diameter” path in Figure 8, where we see that each of the three intervening nodes contain some variant of the term “granule”. In this example, the similarity function would appear to be picking up semantic relationships rather than actual technological dependencies.
4. While a traditional taxonomy is commonly defined by “is-a” relationships, it is clear that the automatically generated taxonomies do not necessarily follow this rule.

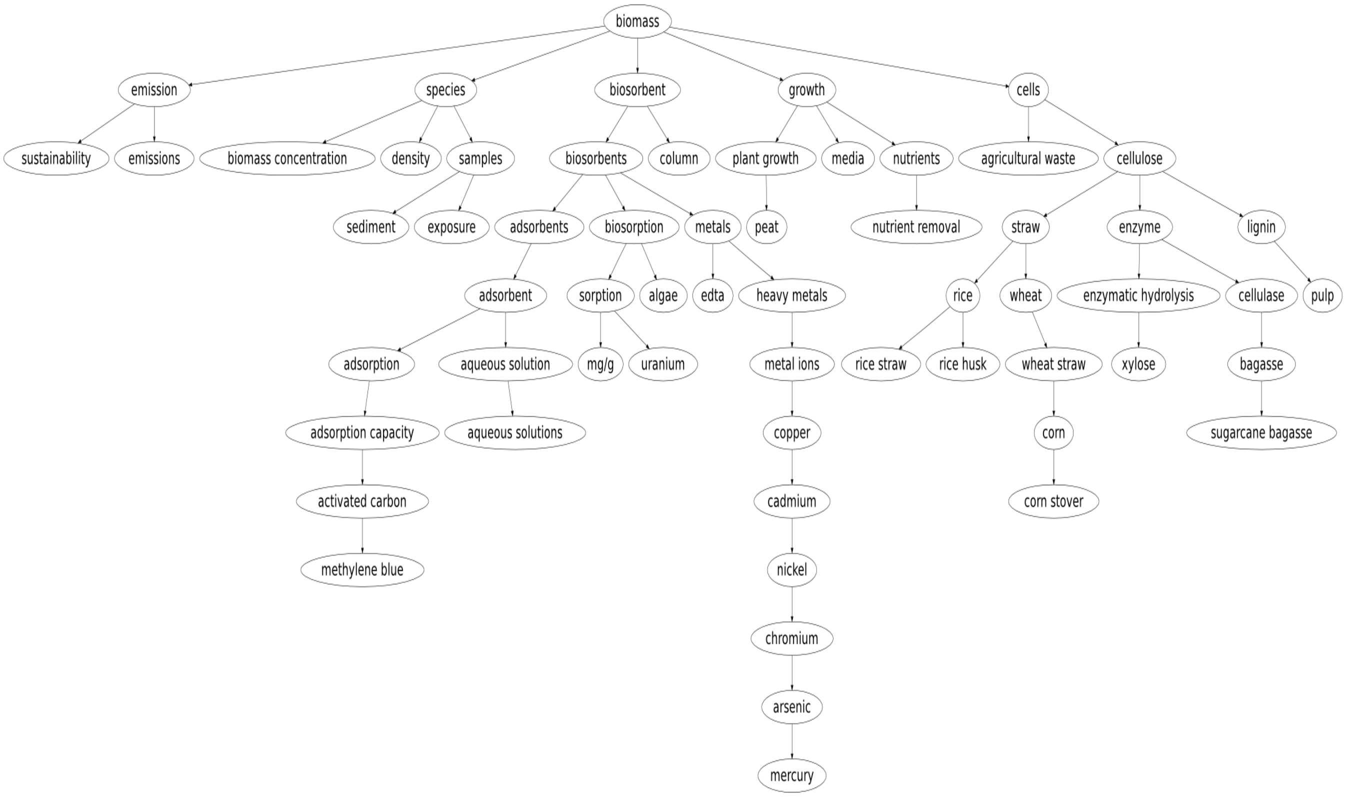


Figure 3 Taxonomy Subtree for "Biomass"

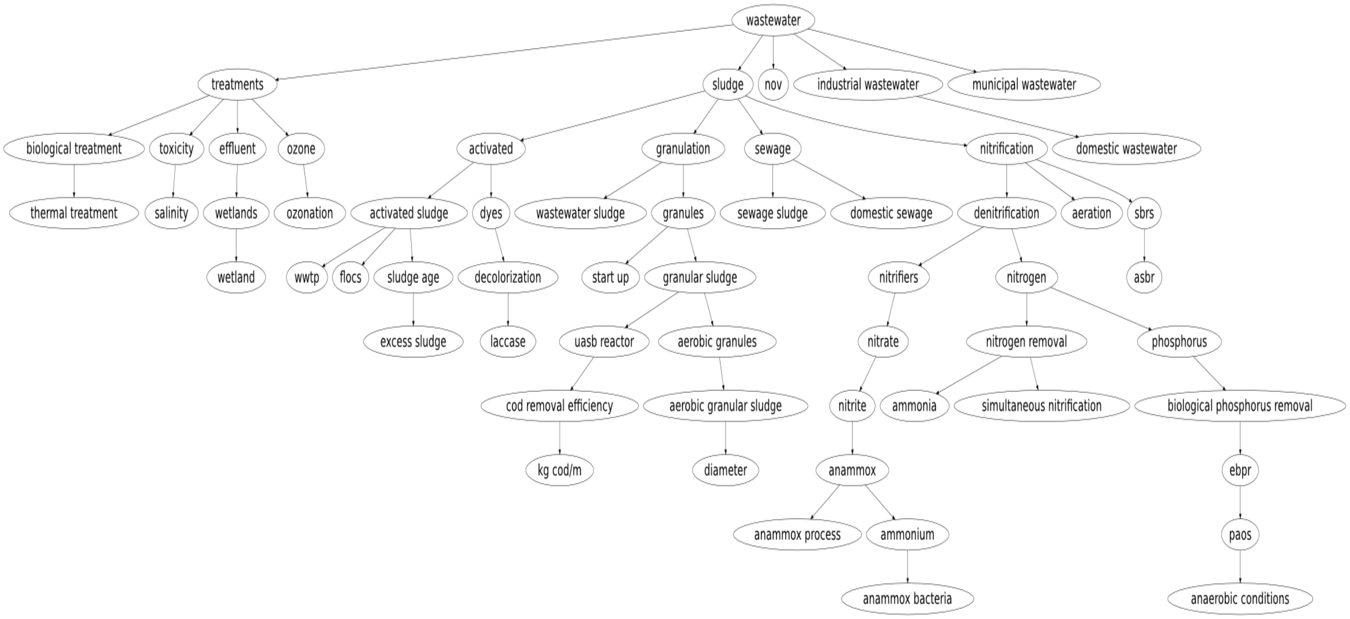


Figure 4 Taxonomy Subtree for "wastewater"

Taxonomy creation: Incorporation of Expert knowledge

As such, it would appear that in the case of waste energy generation, while fully automated taxonomy generation techniques are able to produce results that are interesting, there are also a variety of problems. Firstly, automated taxonomy generation is a somewhat inconsistent process that can, under unfavorable conditions, result in inaccurate or noisy results. Secondly, the choice of algorithm settings is also a difficult problem for which there is no straightforward solution.

A viable alternative might be to opt for a semi-automatic process that would allow some prior knowledge to be incorporated into the taxonomy generation process. This allows for the best of both worlds to be enjoyed. On the one hand, we benefit from the advantages of the automatic approach, namely the ability to quickly incorporate the latest developments as well as to efficiently utilize very large quantities of data; on the other hand, taking a semi-automatic approach allows for valuable input from experts and other manually curate sources to be taken into account. By providing a scaffold or framework with which the taxonomies may be initialized, this approach helps to significantly reduce the uncertainty and inconsistency experienced when purely automatic approaches are used.

Depending on the desired accuracy and the final purpose of taxonomies, their fully automated creation remains a very ambitious endeavor. Many researchers have therefore suggested semi-automated protocols, in which experts have manual influence during various stages of the taxonomy/ontology creation process. The field of Ontology Engineering deals with these aspects. Cimiano points out that automatic extension of existing ontologies have been shown to work successfully (Cimiano, 2006). As a consequence, tools have been created which help to extend ontologies by suggesting terms and their location in the ontology, e.g. within the context of the Gene Ontology project. We therefore investigate the possibility to capitalize on available expert knowledge as an initial guidance to the taxonomy creation process. Note that this approach is an appropriate alternative to the fully automated procedure where expert knowledge is available. We emphasize that we can easily extend the formalism of the Heymann-Algorithm to accommodate initial expert knowledge. The precondition is that terms -at least the expandable nodes- of the expert guidelines must occur frequently in the corpus in order to provide compatibility in terms of the similarity measure. In that case, expert knowledge can be formalized as an initial taxonomy, which is then extensible in the same way, the automatic Heymann algorithm extends a growing taxonomy.

Figure 5 shows a taxonomy which has been constructed in collaboration with an expert of W2E technologies. The taxonomy largely consists of predefined taxonomic relations, which are subsequently extended with 100 TFIDF keywords (larger taxonomies can be found in the Supplementary material). In addition, as mentioned in section "Growth indicator accumulation and visualization", the growth indicators were also incorporated into the figure by modulating the font sizes. As mentioned previously, we have used both font sizes and color-codings to convey growth indicators; however using font sizes has an important advantage in that these are preserved when the document is printed in black and white, which is the reason for its use here.

If such a semi-automatically created taxonomy is embedded in the general framework (Figure 1), it is interesting to inspect the growth indicators, i.e. the recency and the volume of the research bodies associated to each node (shown respectively in all nodes). In particular we note that the font-size modulation allows the growth potential of the nodes to be very clearly visualized. It can for example be seen that the top level categories at taxonomy level 1 and 2 are all balanced out in terms of recency (all are within 1999-2001) due to the average of their associated subtopics. “Plasma Gasification” is the most recent topic (2006). Moreover, it becomes apparent that recently “Biodiesel production” is frequently discussed in the context of "Transesterification". This is evidenced by the presence of a body of 105 publications with an average publication year of 2007. The findings for Biodiesel are consistent with earlier results reported in (Dawelbait et al, 2010) even though a different corpus and a different keyword set were used. The taxonomy created with frequent keywords (Supplementary material S1) unravels that “Removal experiments” have been mentioned in 272 documents with an average date of 2007. A further inspection into the corpus reveals that, indeed a large number of recent papers mention different kinds of removal experiments, such as nitrogen removal. Another term that grew to recent popularity is “Wastewater reclamation”, which was mentioned in 433 papers.

In general, it must be said that the recency of subordinate terms are generally independent from each other, i.e. the W2E research landscape developed rather heterogeneous. This is in contrast to the related study on Renewable Energy (Supplementary material, Figure 2). There, complete branches including subordinate terms could be identified as hot topics, for example most subordinate terms of “Biofuels”.

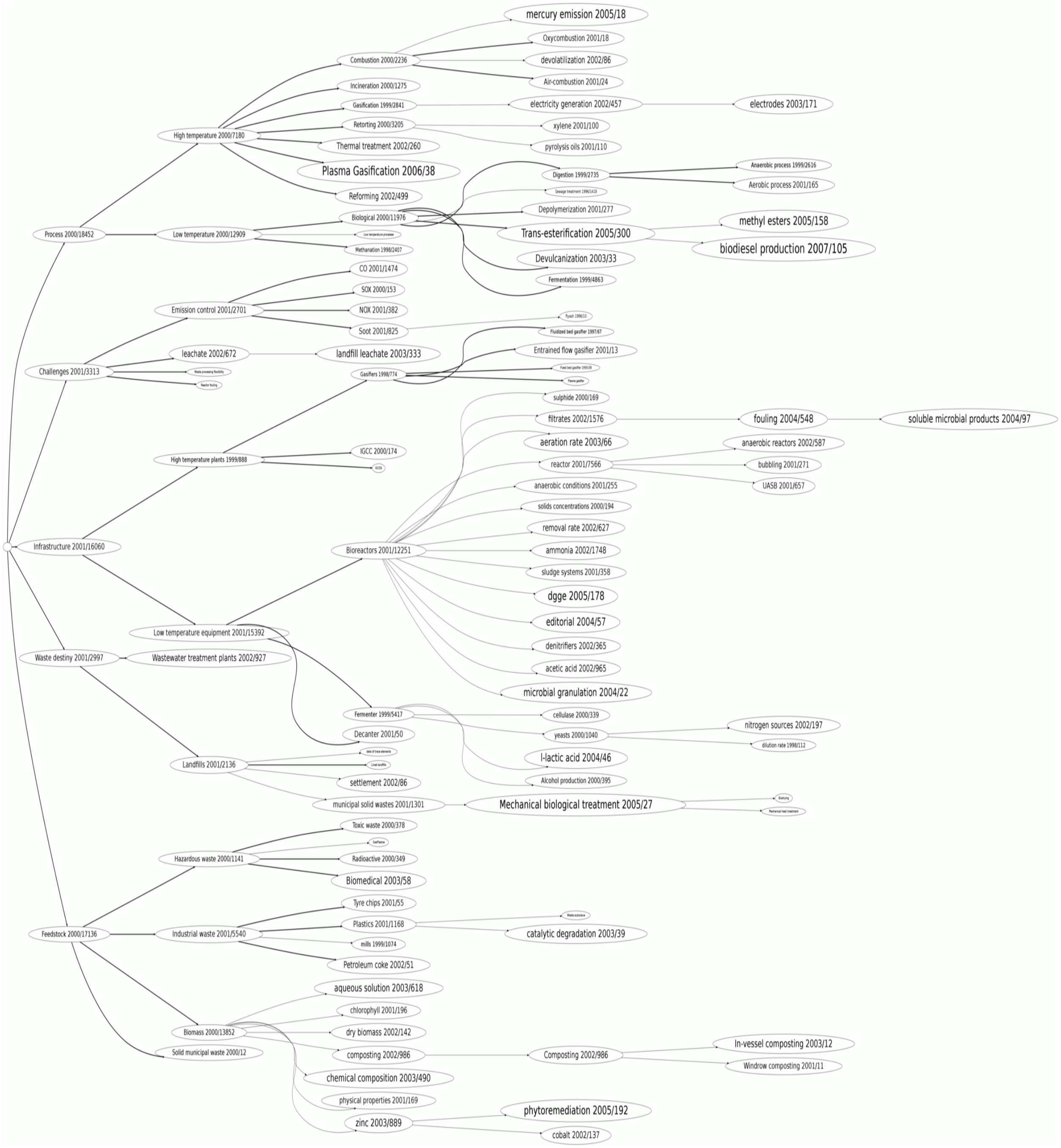


Figure 5 Semi-automatic taxonomy for W2E, incorporating Expert knowledge and 100 TFIDF keywords. Average publication year and associated research body is provided for each term. Large fonts indicate strong recent growth.

Topic Modeling using Latent Dirichlet Allocation

In addition to extending the existing taxonomy generation framework and applying it to different renewable energy domains, significant effort was dedicated to exploring a new research direction, which was the use of a probabilistic modeling approach to technology visualization. The technique that was used is known as *Probabilistic Topic Models*, and is well known within the machine learning and statistical modeling community, where it forms part of the wider field of “Topic Modeling” (TM).

While many highly effective automated tools have recently emerged, practitioners of TM still rely on manual procedures based largely on human expertise. Recent developments in machine learning have provided us with tools which promise to reduce this manual effort. Topic models are a class of text mining methods which represent the content of a document or more generally a set of documents, i.e. a corpus, as a mixture of latent topics. A topic is an abstract entity which represents the concept or idea conveyed by the author of a document. In particular, the aim here is to associate these topics to scientific and technological paradigms and to discover the related relationships and trends embedded in the S&T literature.

In particular, we consider here one particular algorithmic implementation of a topic model which has gained significant popularity in recent times: the *Latent Dirichlet Allocation* or LDA [Blei et al., 2003][Heinrich, 2008][Griffiths and Steyvers, 2004]. LDA is a generative causal topic model that has been shown to be an effective method for performing inference on these latent semantic information, and which has been shown to produce superior performance in comparison to pre-existing text mining methods (as discussed for instance in [Griffiths et al, 2007]). Besides, LDA is also very flexible and can be easily extended to accommodate linguistic constraints, authorship relations, citations, multimedia and other form of meta-information typically present in S&T documents. With this recent extension, LDA was successfully shown to improve the modeling performance when describing text corpora, and is hence likely to be very suitable for TM applications.

In this section, we provide an overview of our TM approach that includes a generic LDA implementation as the basis for learning the latent semantic structure of S&T corpora. In particular, this framework is developed around specific key ideas which we consider to be especially important in the context of Tech-Mining:

- We seek to understand the latent "research landscape" of a technological field of interest. In our previous work we considered a methodology for creating, managing and visualizing taxonomies using bibliometric data (for instance in [Henschel et al., 2009][Woon and Madnick, 2009]). Taxonomies are a particular form of knowledge representation which capture the relationships between semantic concepts in a hierarchical format. However, while hierarchies are known to be useful for organizing, classifying and providing access to information contained in technical-scientific corpora, they are by no means the only applicable way of doing so. Indeed, one of the findings of this research is that topic modeling can provide an alternative methodology for mining and visualizing the research landscape.
- For decision-makers it is not only important to understand the state of art of the S&T field but also to be able to track and predict its future outcome, i.e. Technology Forecasting [Daim et al., 2006][Porter, 2007][Porter and Newman, 2005][Martino, 2003. In the present context, temporal information can be added to topic models in order to track the semantic evolution of the research landscape and in this way produce useful indicators of innovation.
- We have built a prototype of a Graphical User Interface (GUI) which can be used by TM practitioners in the process of mining, knowledge discovery and decision making. Several GUIs have been developed in the last few years, especially in the patent community (see [Moehrle et al., 2010] for a good review)

The main goal of this research effort was to extract patterns from the data that would reveal the specific technologies currently being discussed in the literature, how they are relate to each other and finally how these technologies and their relationships are growing in time. Understanding these subfields, linking their results and planning ahead to find new solutions in the field of renewable energy is a challenging task in tech-mining as well as a socio-economically important problem for a sustainable and clean world.

Methodology

A TM application can be organized into a set of separate processing stages, as is often presented in the specialized literature (i.e. in [Tseng et al., 2007]). These stages parallel the overall framework that we have been using in our research; indeed the same guidelines are used where the overall analysis is divided into three conceptual parts: information retrieval and preprocessing, topic modeling and technology forecasting. The purpose of the first stage is to interface with a chosen electronic database, extracting relevant documents as dictated by a query of interest and identifying and structuring their basic internal contents. Text mining algorithms are applied to this basic (unstructured) information in the second stage in order to provide a suitable model for the underlying global database. As stated in the introduction, this part describes the LDA algorithm. The final task is to address the problem of technology forecasting which is the main interest of our research. In this third stage we utilize the information extracted during the LDA stage to generate effective visualizations of the relationship between technologies in the research landscape as well as to perform time domain analysis, i.e. the extraction of trends and innovation indicators embedded within the global database.

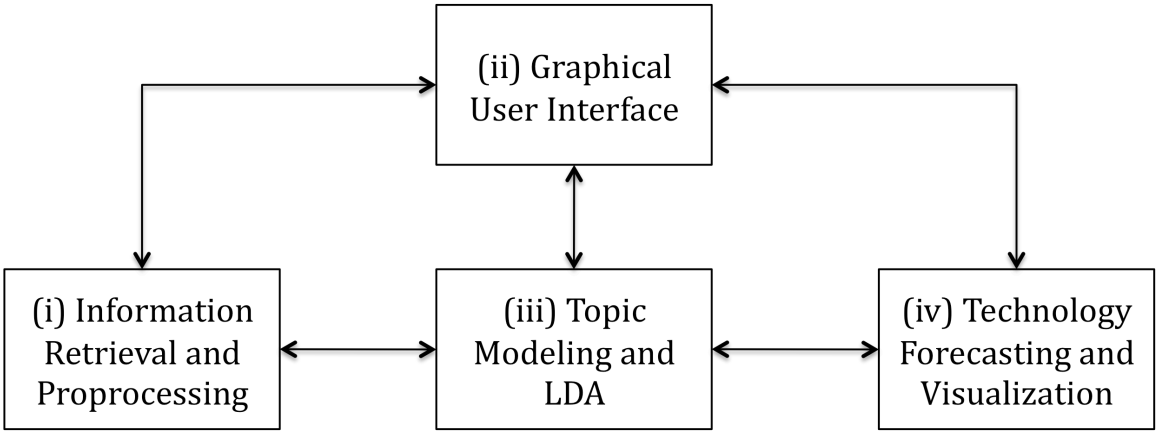

(ii) Graphical User Interface
(i) Information Retrieval and Proprocessing
(iii) Topic Modeling and LDA
(iv) Technology Forecasting and Visualization

Figure 6 Schematic of a TM application

Finally, to facilitate experimentation with topic modeling and to provide a concrete "product" from this research direction, the different analytical components of the framework were interconnected by means of a graphical user interface. The conceptual layout of this framework is illustrated in Figure 6 above.

Next, each of the blocks depicted will be discussed in greater detail.

Information retrieval and preprocessing - block (i)

In formulating the initial process of data acquisition and corpus generation, several important questions need to be answered before our framework can be adapted to the corpus modeling stage. What is the source of data? How do I query this source in order to extract relevant information? What are the preprocessing steps to be taken in order to structure the corpus S&T before the topic modeling stage?

News feed, blogs, conference papers, journal papers and patents are examples of documents which are accessible from online repositories, and which contain a variety of forms of information; these can be organized in order of increasing accuracy and decreasing time scope. Patents, for instance, hold intellectual property information that fulfills a certain set of requirements (utility, novelty, non-obviousness and enablement), where conformity is established via a lengthy review process (~18 months) [Kasravi and Risov, 2007][Bonino et al., 2010]. Therefore, a patent discloses more mature and detailed information than a paper in a journal or conference. For this reasons, patents plays an important role in the S&T process when this is conducted in the context of business or industry, where the quality of the data is critical to the decision making process. The analysis of patents is a rapidly growing sub-domain of Tech-mining which has even earned its own name: "Patinformatics" [Trippe, 2003]. On the other end of this spectrum, blogs contain information which is usually unstructured and does not undergo any review or censorship process (with the exception of "official" blogs run by companies or government entities); conversely, blogs allow information to be distributed over very short periods of time. While the quality of the information contained in blogs may not be as reliable as other sources, the quantity of data that can potentially be collected from blogs is far greater and in some cases this can compensate for the greater levels of noise present in data generated from mining blogs. The additional advantage of blogs is that rapidly emerging trends and patterns can be detected long before they appear in patents or even academic publications.

A broad variety of online repositories are available, examples of which include Scopus, Google Scholar, Scirus, Compendex, PubMed, ISI Web of Science, the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), and the Derwent World Patents Index. They have the advantage of providing complex query capabilities that are accessible via the web. On the other hand, these repositories typically do not provide facilities for conducting queries in an automated or programmatic fashion, meaning that the quantity of data that is publicly accessibly is often limited.

When seeking to mine one of these repositories, there are two general ways of acquiring relevant documents [Porter and Cunningham, 2005]. The first approach, which we have been using, is to simply submit relevant keywords to the repository's internal search engine and to consider the most relevant documents retrieved. The second is to use an appropriate document classification system to categorize all documents beforehand. This is frequently done internally by the operator of the repository, the output of which is data that is structured as a taxonomy, for instance in the case of patent databases; our taxonomy construction algorithms can be used to either validate these schemes or to provide new structure (or substructure) to the particular area of interest. This analysis is left as a future research direction.

When the corpus relevant to a technological field has been build, we consider structuring its raw documents by extracting relevant elements which can be used for LDA and for the following TM analysis. This process is known as *term extraction*, where the exact approach taken can vary depending on the kinds of text features focused on in the analysis. A variety of different linguistic features have been proposed in the literature, prominent examples of which include manually selected keywords, TFIDF keywords, n-grams and community-generated tags. In our work, as a preprocessing step, we firstly collect every single words in the text (tokenization) and secondly, we remove stop-words from standard lists available in the web (articles or other grammatical particles with little or no information). All the other words are taken as the vocabulary $\mathcal{W}$ used by LDA without using any pre-stemming algorithm 20 (). We assume that LDA, being a clustering algorithm, would mix words with similar roots together in a single topic.

Graphical User Interface – block (ii)

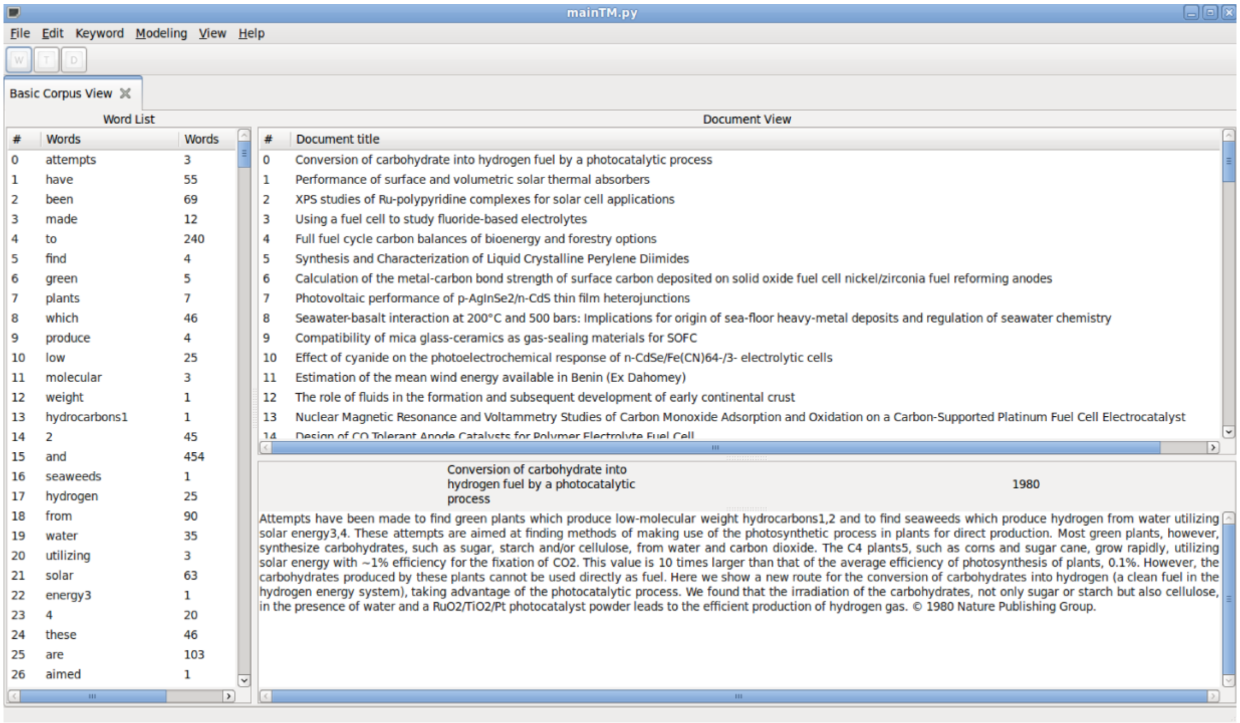
During the course of the project, a number of simple GUI components have been developed, notably the early growth analysis tool. These have been described in earlier reports and are not described in detail here.

In the latest reporting period, we have worked to build a new GUI for embedding the LDA technology visualization framework. In this report some of the screenshots of this latest GUI are presented, and are based on all articles relevant to renewable energy and some of its sub fields, i.e. wind power, solar energy, waste-to-energy. This corpus was downloaded manually by querying the Scopus database and was used in our previous investigations (see in [Woon et al., 2010]). The field of renewable energy is an excellent example where tech-mining techniques could be applicable. This domain offers a diverse literature landscape which can be divided into many scientific disciplines, each of which may contain independently evolving threads of research. Understanding these subfields, determining the linkages between them and planning ahead in the search for new solutions in the field of renewable energy is both a challenging task in tech-mining and a socio-economically important problem for a sustainable and clean world.

We develop the software part of our framework in Python. The GTK (**G**imp **T**ool-**K**it) library was used to generate the interface elements required in the program, the Matplotlib package for the graphical visualizations and Numpy/Scipy for numerical computation. The GUI interfaces allows the user to load the selected Corpus from a local SQL database (currently only the Renewable Energy corpus is available).

The various user menus provide a friendly interface to the text analysis and LDA functionality (other methods such as LSA can be “plugged” into the system if required). Finally, the results of the analysis are presented to the user by using different views to represent different types of exploratory analysis conducted on the data.

Figure 7 shows the main screen when the corpus is first loaded by the program. The main view of Figure 7 includes the vocabulary list of the corpus on the left hand side as a table which can include some of the most common word statistics, i.e. frequencies, TFIDF statistics, etc. In this view the user can interact in order to eliminate words from the corpus by activating the corresponding “stop-word” flag. The program is already hard-wired to use a number of standard stop-word lists, while the user has the option of incorporating additional stop-words by selecting them in the vocabulary list. In the right hand side of Figure 7, the list of documents present in the database is shown by the title. The user can select any of the documents in this list, upon which further information can be displayed conveniently in the central part of Figure 7, i.e. abstract and year of publication. Documents can also be discarded by interacting with the list if the content is found to be irrelevant.



#	Words	Words
0	attempts	3
1	have	55
2	been	69
3	made	12
4	to	240
5	find	4
6	green	5
7	plants	7
8	which	46
9	produce	4
10	low	25
11	molecular	3
12	weight	1
13	hydrocarbons1	1
14	2	45
15	and	454
16	seaweeds	1
17	hydrogen	25
18	from	90
19	water	35
20	utilizing	3
21	solar	63
22	energy3	1
23	4	20
24	these	46
25	are	103
26	aimed	1

#	Document title
0	Conversion of carbohydrate into hydrogen fuel by a photocatalytic process
1	Performance of surface and volumetric solar thermal absorbers
2	XPS studies of Ru-polypyridine complexes for solar cell applications
3	Using a fuel cell to study fluoride-based electrolytes
4	Full fuel cycle carbon balances of bioenergy and forestry options
5	Synthesis and Characterization of Liquid Crystalline Perylene Diimides
6	Calculation of the metal-carbon bond strength of surface carbon deposited on solid oxide fuel cell nickel/zirconia fuel reforming anodes
7	Photovoltaic performance of p-AgInSe2/n-CdS thin film heterojunctions
8	Seawater-basalt interaction at 200°C and 500 bars: Implications for origin of sea-floor heavy-metal deposits and regulation of seawater chemistry
9	Compatibility of mica glass-ceramics as gas-sealing materials for SOFC
10	Effect of cyanide on the photoelectrochemical response of n-CdSe/Fe(CN)64-/3- electrolytic cells
11	Estimation of the mean wind energy available in Benin (Ex Dahomey)
12	The role of fluids in the formation and subsequent development of early continental crust
13	Nuclear Magnetic Resonance and Voltammetry Studies of Carbon Monoxide Adsorption and Oxidation on a Carbon-Supported Platinum Fuel Cell Electrocatalyst
14	Design of CO Tolerant Anode Catalysts for Polymer Electrolyte Fuel Cell

Conversion of carbohydrate into hydrogen fuel by a photocatalytic process 1980

Attempts have been made to find green plants which produce low-molecular weight hydrocarbons1,2 and to find seaweeds which produce hydrogen from water utilizing solar energy3,4. These attempts are aimed at finding methods of making use of the photosynthetic process in plants for direct production. Most green plants, however, synthesize carbohydrates, such as sugar, starch and/or cellulose, from water and carbon dioxide. The C4 plants5, such as corns and sugar cane, grow rapidly, utilizing solar energy with ~1% efficiency for the fixation of CO2. This value is 10 times larger than that of the average efficiency of photosynthesis of plants, 0.1%. However, the carbohydrates produced by these plants cannot be used directly as fuel. Here we show a new route for the conversion of carbohydrates into hydrogen (a clean fuel in the hydrogen energy system), taking advantage of the photocatalytic process. We found that the irradiation of the carbohydrates, not only sugar or starch but also cellulose, in the presence of water and a RuO2/TiO2/Pt photocatalyst powder leads to the efficient production of hydrogen gas. © 1980 Nature Publishing Group.

Figure 7   Main screen of the TM GUI application

The GUI is considered as a work-in-progress and it is the main topic of further phases of development of our approach of Tech-mining. A complete version of the GUI has not been released yet.

LDA and Topic Modeling – block (iii)

A topic model is a mathematical representation of a text corpus defined as $\mathcal{C}$, i.e. a collection of documents $\left\{d_i\right\}_{i=1}^{D}$. In our research, each document is an S&T document which can be defined as a sequence of N_{d_i} words (documents can have different length) taken from a vocabulary $\left\{w_i\right\}_{i=1}^{W}$. In order to simplify the computational load and the modeling problem, firstly, only the abstract is considered during the analysis. The abstract contains a concise description of the content of the article and therefore it should be sufficient to extract its most important message. Secondly, the sequential order of the words in the abstract is ignored. The latter assumption is known in literature as the "Bag of Word" representation of a document: the grammatical structure of documents are not taken into account and the corpus can be rephrased mathematically using a co-occurrence matrix $\mathbf{C}=(n(d_i,w_j))_{ij}$ where $n(d,w)$ represent the number of times the word w is included in the document.

While the bag of word assumption discards the grammatical structure of documents, it is the most commonly used text representation; other representations or modeling strategies have been proposed but these tend to be variations on this same basic theme. In [Griffiths et al, 2007], for instance, the authors review three of the most common approaches taken by the scientific community, i.e. the semantic network, the semantic vector space and topic modeling approaches. From the second category, it is important to mention a popular matrix algebra technique known as Latent Semantic Analysis (LSA). LSA decompose the matrix **C** using the Singular Value Decomposition (SVD) in order to find the directions in the *semantic vector space* where the data have maximum variance. These orthogonal directions, which are a linear combination of the words in the vocabulary can be taken as a possible representation of some hidden concepts $\left\{z_s\right\}_{s=1}^{K}$. The latter are abstract entities which represent the underlying process of knowledge generation in the database.

Topic models are statistical method which conceptual framework is based on the semantic space as a mixture of topics. As in the case of LSA, the observations, i.e. the words w_j in the corpus, are considered convex combinations of z_s. However, as described for instance in [Griffiths et al, 2007], they have several other advantages respect to the traditional vector space and network approaches since topic models is base on a more general and solid probabilistic framework which can be used for problem of model complexity, comparison and fitting. In particular we have that:

$$p(w_i) = \sum_{s=1}^{K} p(w_i \mid z_s) p(z_s) \qquad (4)$$

which is the probability of each of the words within a document *d*. In particular, following the notation in [Griffiths and Steyvers, 2004], we can define $\phi^s = p(w \mid z_s)$ as a multinomial distribution of words given the topic *s* and $\theta^d = p(z)$ the multinomial distribution of the topics in a document. This can be expressed also as $p(z|d_i)$, i.e. the proportion of a particular topic in a single document. Extended to the entire corpus $\mathcal{C}$, Φ is a *V*×*K* matrix while Θ is a *K*×*D* matrix. As discussed in [Steyvers and Griffiths, 2007], a topic model can be seen as a probabilistic matrix factorization process similar to the SVD algorithm in the LSA (see Figure 8).

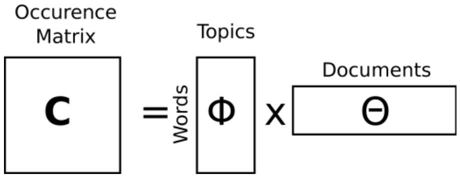


Figure 8

The final goal of the modeling stage is to learn the previous probabilities distribution in Equation 4 which is the entries of the Φ and Θ matrices as shown in Figure 8. In the literature, we mention two closely approaches for learning the mixture of topics in Equation 4: *Probabilistic Latent Semantic Analysis* (pLSA) and the LDA. PLSA is an algorithm based on the popular EM minimization algorithm, often used for estimating a mixture of models. LDA is going further than the PLSA because it offers a fully generative approach to Equation 4. In this respect, PLSA have been shown to be a particular case of LDA since it is consider the MAP estimator of the Bayesian model which is shown in Figure 9 by a graphical model. Figure 9 describes Equation 4 including the two extra Dirichlet priors α and β [Griffiths and Steyvers, 2004].

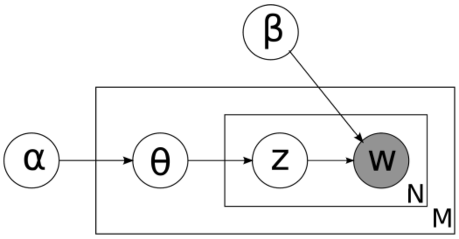
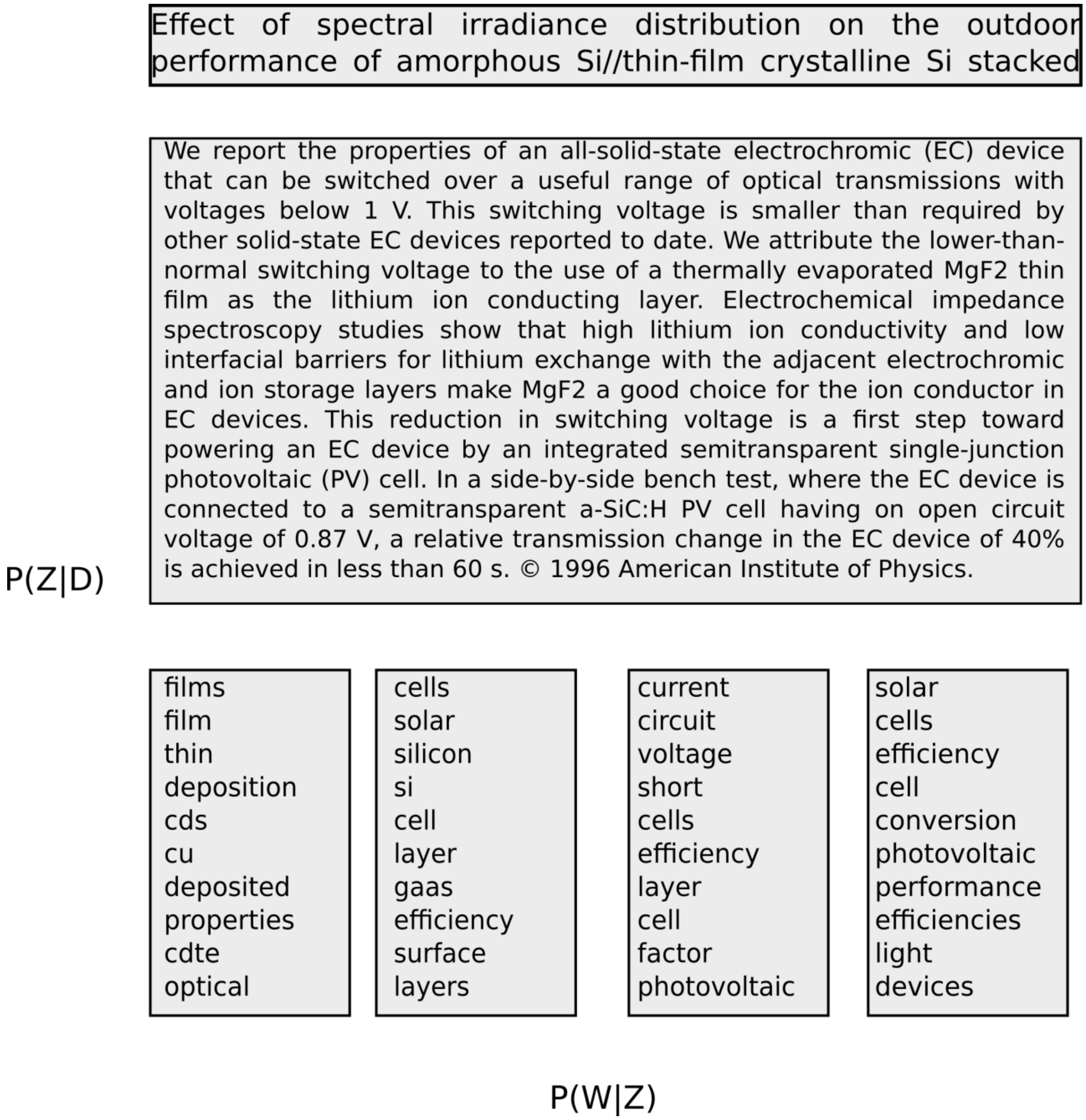


Figure 9

Taking Dirichlet distributions for priors has mathematical benefits, i.e. in calculating posterior probabilities in a Bayesian setting, since it is conjugate to the multinomial distributions of Equation 4. There are two methods in the literature to perform the Bayesian learning with LDA. The original method, proposed in [Blei et al., 2003], is based in a variational learning scheme of learning while in [Griffiths and Steyvers, 2004], the authors proposed a method based on Gibbs sampling. In our research we used the latter one since easy to implement and because it has shown better results in comparative studies [Griffiths and Steyvers, 2004].

Summarizing, the result of an LDA computation is the estimation of the two probability distributions $p(w \mid z_s)$ and $p(z \mid d_j)$ (in Equation 4 $p(z)$ is relative to one document only). $p(w \mid z_s)$ and thus the columns of Φ contain the probabilities of words belonging to a determined topic. $p(z \mid d_j)$ tells the proportion of topics in a particular document of the corpus. In Figure 10, we show a practical example of these distributions. An exemplar document of the renewable energy corpus is considered. We run the LDA for a number of topics K=50. At the bottom of the figure we show the most relevant topics for the document (i.e. the topic having the largest $p(z_s \mid d_j)$ value for each document respectively). Each topic is represented by a collection of the 10 most relevant words (sorted by the probability $p(w \mid z_s)$). In practice, the first few ranked words of a specific topic z_j would typically provide a reasonable description of the topic itself, i.e. each topic would be explained by a certain mixture of words. As can be seen, the mixture of words contains terms that can be related to an abstract underlying concept.

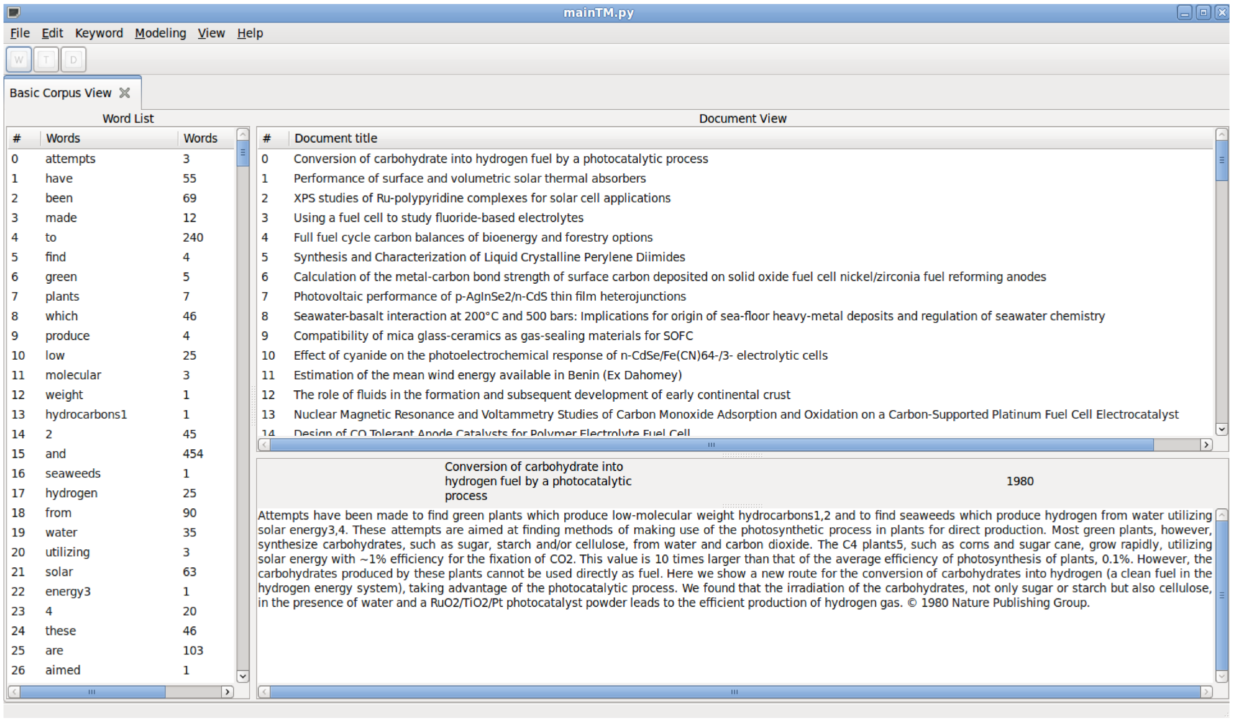


Effect of spectral irradiance distribution on the outdoor performance of amorphous Si//thin-film crystalline Si stacked

We report the properties of an all-solid-state electrochromic (EC) device that can be switched over a useful range of optical transmissions with voltages below 1 V. This switching voltage is smaller than required by other solid-state EC devices reported to date. We attribute the lower-than-normal switching voltage to the use of a thermally evaporated MgF2 thin film as the lithium ion conducting layer. Electrochemical impedance spectroscopy studies show that high lithium ion conductivity and low interfacial barriers for lithium exchange with the adjacent electrochromic and ion storage layers make MgF2 a good choice for the ion conductor in EC devices. This reduction in switching voltage is a first step toward powering an EC device by an integrated semitransparent single-junction photovoltaic (PV) cell. In a side-by-side bench test, where the EC device is connected to a semitransparent a-SiC:H PV cell having on open circuit voltage of 0.87 V, a relative transmission change in the EC device of 40% is achieved in less than 60 s. © 1996 American Institute of Physics.

Figure 10 Example of topics extracted from a document on Renewable Energy.

After using the GUI to perform LDA, the main view shown in Figure 11 is augmented as shown in Figure 12 with the addition of a further notebook tab; clicking this brings up the list of topics as a mixture of words. At this point, the user going back to the principal view can choose to display the abstract with a basic 'tagging mode' (as depicted in Figure 12), where each word in the abstract is colored by the topic for which the word contributes the most. In this way, the user will be able to inspect the topic signature of the document and thus gain important insight into the overall content. Finding the best visualization technique for exploring this tagging is not currently the main aim of this research and its implementation is considered as a future research effort.



Word List

#	Words	Words
0	attempts	3
1	have	55
2	been	69
3	made	12
4	to	240
5	find	4
6	green	5
7	plants	7
8	which	46
9	produce	4
10	low	25
11	molecular	3
12	weight	1
13	hydrocarbons1	1
14	2	45
15	and	454
16	seaweeds	1
17	hydrogen	25
18	from	90
19	water	35
20	utilizing	3
21	solar	63
22	energy3	1
23	4	20
24	these	46
25	are	103
26	aimed	1

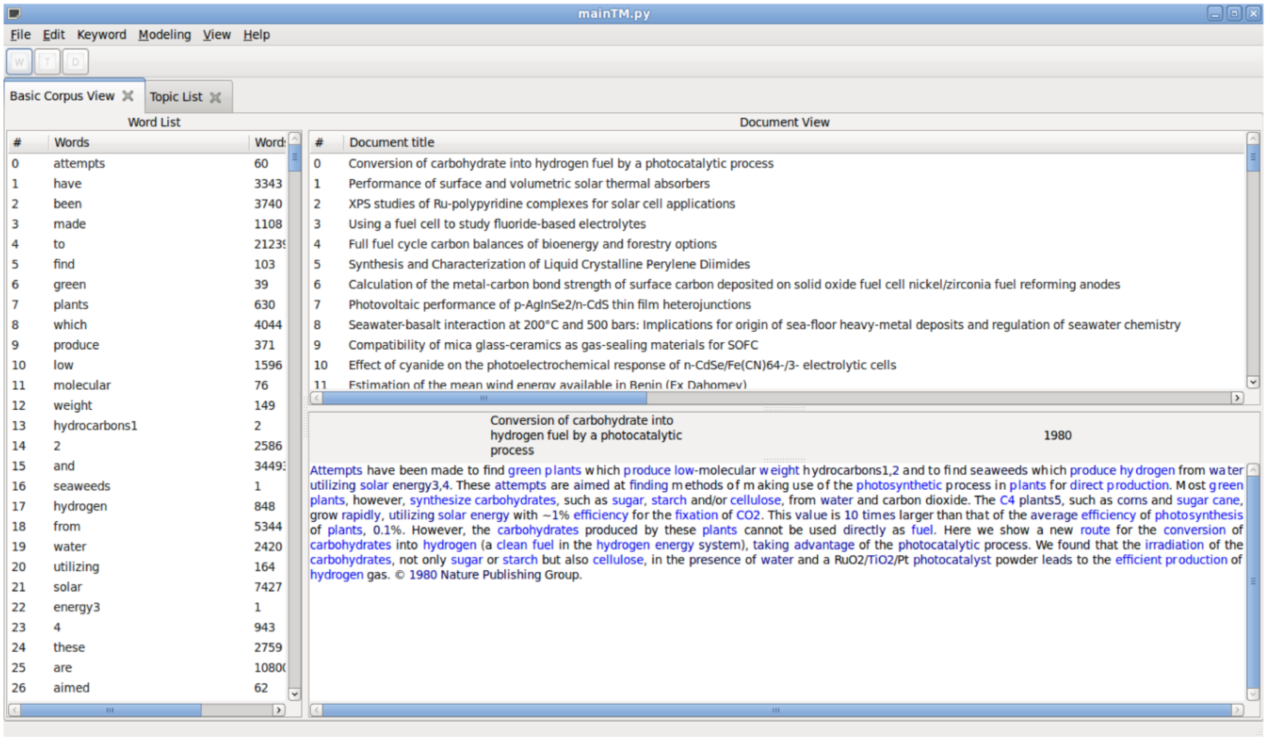
Document View

#	Document title
0	Conversion of carbohydrate into hydrogen fuel by a photocatalytic process
1	Performance of surface and volumetric solar thermal absorbers
2	XPS studies of Ru-polypyridine complexes for solar cell applications
3	Using a fuel cell to study fluoride-based electrolytes
4	Full fuel cycle carbon balances of bioenergy and forestry options
5	Synthesis and Characterization of Liquid Crystalline Perylene Diimides
6	Calculation of the metal-carbon bond strength of surface carbon deposited on solid oxide fuel cell nickel/zirconia fuel reforming anodes
7	Photovoltaic performance of p-AgInSe2/n-CdS thin film heterojunctions
8	Seawater-basalt interaction at 200°C and 500 bars: Implications for origin of sea-floor heavy-metal deposits and regulation of seawater chemistry
9	Compatibility of mica glass-ceramics as gas-sealing materials for SOFC
10	Effect of cyanide on the photoelectrochemical response of n-CdSe/Fe(CN)64-/3- electrolytic cells
11	Estimation of the mean wind energy available in Benin (Ex Dahomey)
12	The role of fluids in the formation and subsequent development of early continental crust
13	Nuclear Magnetic Resonance and Voltammetry Studies of Carbon Monoxide Adsorption and Oxidation on a Carbon-Supported Platinum Fuel Cell Electrocatalyst
14	Design of CO Tolerant Anode Catalysts for Polymer Electrolyte Fuel Cell

Conversion of carbohydrate into
hydrogen fuel by a photocatalytic
process
1980

Attempts have been made to find green plants which produce low-molecular weight hydrocarbons1,2 and to find seaweeds which produce hydrogen from water utilizing solar energy3,4. These attempts are aimed at finding methods of making use of the photosynthetic process in plants for direct production. Most green plants, however, synthesize carbohydrates, such as sugar, starch and/or cellulose, from water and carbon dioxide. The C4 plants5, such as corns and sugar cane, grow rapidly, utilizing solar energy with ~1% efficiency for the fixation of CO2. This value is 10 times larger than that of the average efficiency of photosynthesis of plants, 0.1%. However, the carbohydrates produced by these plants cannot be used directly as fuel. Here we show a new route for the conversion of carbohydrates into hydrogen (a clean fuel in the hydrogen energy system), taking advantage of the photocatalytic process. We found that the irradiation of the carbohydrates, not only sugar or starch but also cellulose, in the presence of water and a RuO2/TiO2/Pt photocatalyst powder leads to the efficient production of hydrogen gas. © 1980 Nature Publishing Group.

Figure 11 TM application after performing LDA



#	Words	Word:
0	attempts	60
1	have	3343
2	been	3740
3	made	1108
4	to	2123
5	find	103
6	green	39
7	plants	630
8	which	4044
9	produce	371
10	low	1596
11	molecular	76
12	weight	149
13	hydrocarbons1	2
14	2	2586
15	and	3449
16	seaweeds	1
17	hydrogen	848
18	from	5344
19	water	2420
20	utilizing	164
21	solar	7427
22	energy3	1
23	4	943
24	these	2759
25	are	1080
26	aimed	62

#	Document title
0	Conversion of carbohydrate into hydrogen fuel by a photocatalytic process
1	Performance of surface and volumetric solar thermal absorbers
2	XPS studies of Ru-polypyridine complexes for solar cell applications
3	Using a fuel cell to study fluoride-based electrolytes
4	Full fuel cycle carbon balances of bioenergy and forestry options
5	Synthesis and Characterization of Liquid Crystalline Perylene Diimides
6	Calculation of the metal-carbon bond strength of surface carbon deposited on solid oxide fuel cell nickel/zirconia fuel reforming anodes
7	Photovoltaic performance of p-AgInSe2/n-CdS thin film heterojunctions
8	Seawater-basalt interaction at 200°C and 500 bars: Implications for origin of sea-floor heavy-metal deposits and regulation of seawater chemistry
9	Compatibility of mica glass-ceramics as gas-sealing materials for SOFC
10	Effect of cyanide on the photoelectrochemical response of n-CdSe/Fe(CN)64-/3- electrolytic cells
11	Estimation of the mean wind energy available in Benin (Ex Dahomey)

Conversion of carbohydrate into hydrogen fuel by a photocatalytic process

1980

Attempts have been made to find green plants which produce low-molecular weight hydrocarbons1,2 and to find seaweeds which produce hydrogen from water utilizing solar energy3,4. These attempts are aimed at finding methods of making use of the photosynthetic process in plants for direct production. Most green plants, however, synthesize carbohydrates, such as sugar, starch and/or cellulose, from water and carbon dioxide. The C4 plants5, such as corns and sugar cane, grow rapidly, utilizing solar energy with ~1% efficiency for the fixation of CO2. This value is 10 times larger than that of the average efficiency of photosynthesis of plants, 0.1%. However, the carbohydrates produced by these plants cannot be used directly as fuel. Here we show a new route for the conversion of carbohydrates into hydrogen (a clean fuel in the hydrogen energy system), taking advantage of the photocatalytic process. We found that the irradiation of the carbohydrates, not only sugar or starch but also cellulose, in the presence of water and a RuO2/TiO2/Pt photocatalyst powder leads to the efficient production of hydrogen gas. © 1980 Nature Publishing Group.

Figure 12 TM application in "Tagging Mode"

Technology Visualization and Forecasting using Topic Modelling – block (iv)

Once we have generated the topic models using LDA, these can be directly plugged into the technology forecasting effort by finding trends, smoothed growth indicators or by creating topic model-based visualizations in order to find inter-linkages between topics. For the latter case, we have explored two alternative techniques by which topic models can be used to generate visualizations:

- Taxonomy generation (similar to our previous approach)
- Generation of topic maps.

Each of these three potential applications (Trend detection, Taxonomy Generation and Topic Maps) will now be briefly described.

<u>Trend detection</u>

Extraction of trends for words, keywords or topics requires the topic model to take into account the year of publication of a document. This information can be embedded into our framework but the original LDA algorithm, as is presented here, does not consider the temporal domain of a corpus. At the same time, it is worth noting that the literature contains studies where the Bayesian model of LDA has been extended with the inclusion of a further random variable $\{y_t\}_{t=1}^{D}$, i.e. publication years for each document. This approach can be found, for instance, in the Dynamic Topic Model [Blei, 2006], and in the Topics over Time Model [Wang and McCallum, 2006] approaches.

In our research, we have considered simpler approaches for the first stage of the development of our software. We consider the occurrence frequencies of keywords of interests. The variable $\{y_t\}$ is taken into account after the application of the LDA algorithm by,

$$p(z_s \mid y) = \sum_{d:t_d = y} p(z_s \mid d) p(d \mid y) = \frac{1}{C} \sum_{d:t_d = y} \sum_{z'_s \in d} I(z'_s = z) \quad (7)$$

where $I(\cdot)$ is the indicator function, t_d is the year of the document d and $p(d|y_t)$ is set to the constant $\frac{1}{C}$. In practice the normalization constant is the count of the topic z_s in the documents d in the years y. Similarly we can compute a similar quantity if we consider words per year:

$$p(w_j \mid y) = \sum_{d:t_{d=y}} \sum_{w_j' \in d} I(w_j' = w) \qquad (8)$$

An example of trend analysis is depicted in Figure 13. A few prominent keywords were selected from the database and their trends are plotted after the LDA algorithm computes Φ and Θ. The user can choose what information to visualize by selecting topics or words to be included in the plot by using a dedicated menu. Figure 13 represents the evolution of certain words of the database. The user can tag the curves that he finds more interesting to emphasize. As discussed previously for the map-like visualization, in the next development of the software we are targeting keywords and noun phrases instead of single words or topics.

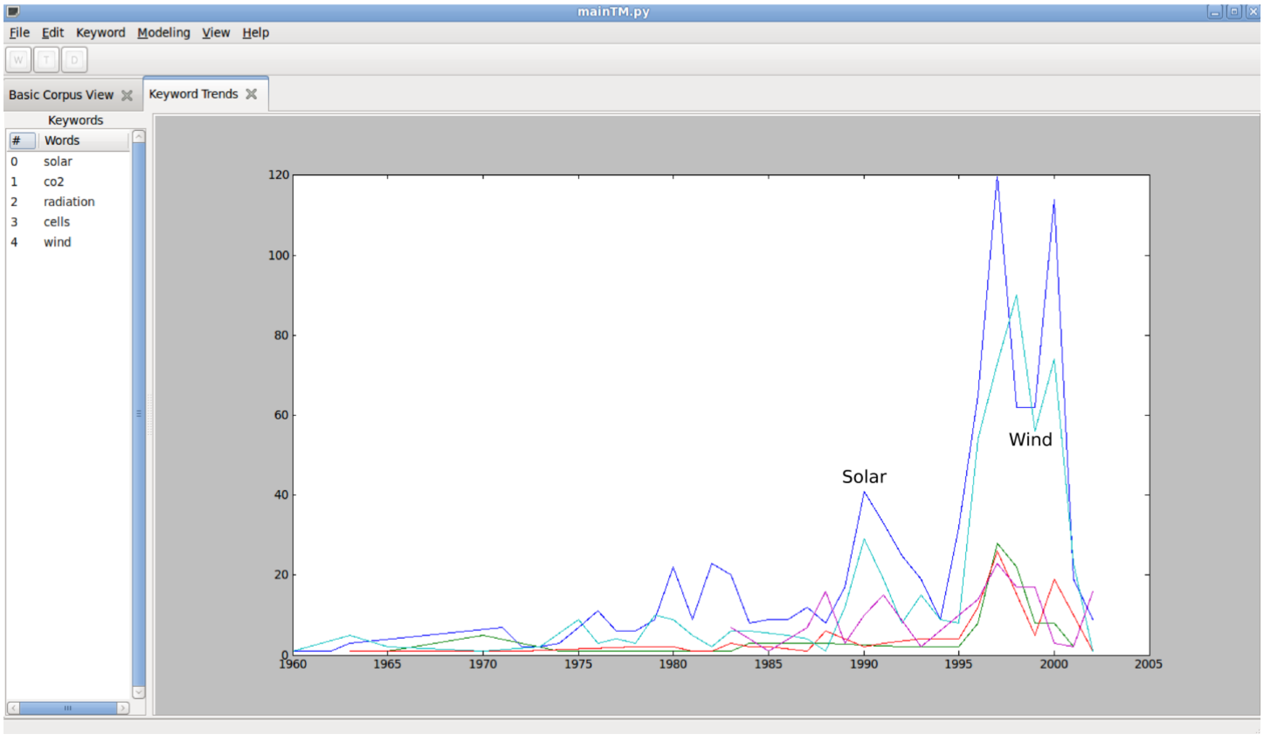


Figure 13 Performing trend analysis using the TM application

Taxonomy Generation

As in our previous approaches, the taxonomy generation process will be achieved via the following two stage process:

1. A distance measure is applied to terms in the text producing a similarity matrix and a weighted graph. A measure of centrality is computed for each of the nodes of this graph. In graph theory, centrality is a measure of the connectedness of a node in a graph (see [Newman, 2010] for a good description)

2. The nodes are then ranked according to their respective centralities, and are inserted into a growing taxonomy in accordance to this ranking; the attachment of these text elements is also determined by the similarity measure described above, where each nodes is attached to either the most similar nodes or to the taxonomy root.

The first part of the algorithm requires a matrix of distances between the nodes of the taxonomy, i.e. words or topics. A common choice of distance function in text analysis in case of the words is the cosine vector similarity of the matrix **C** or its binary representation [Henschel et al, 2009]. In [Woon and Madnick, 2009], an asymmetric distance function is discussed in order to consider the distances of node in the taxonomy from a "is-a" relationship. In our work, since we use a probabilistic framework, we considered instead the entries of the matrices Φ and Θ using an information-like measure. Thus in case of the topic distribution we have the following Kullback–Leibler divergence using Θ:

$$KL(z_n \parallel z_m) = \sum_k p(z_n \mid d_k) \frac{p(z_n \mid d_k)}{p(z_m \mid d_k)} \qquad (5)$$

Since $KL(z_n \| z_m)$ is asymmetric, we use the symmetric version, i.e. $KL_s(z_n \| z_m) = KL(z_n \| z_m) + KL(z_m \| z_n)$. If the taxonomy is based on single words or keywords the same formula is applied by using the entries of Φ:

$$KL(w_n \parallel w_m) = \sum_k p(w_n \mid d_k) \frac{p(w_n \mid d_k)}{p(w_m \mid d_k)} \qquad (6)$$

As in the previous topic case, we consider the symmetric KL version, i.e. $KL_s(w_n \| w_m) = KL(w_n \| w_m) + KL(w_m \| w_n)$

Topic maps

A topic map is a 2D representation of a similarity matrix which is constructed using the KL divergence but for which the Heymann algorithm is replaced with the Multi Dimensional Scaling algorithm (MDS). The MDS is a well-known method of visualization and dimensionality reduction that has often been used in data mining (technical details are left to the specialized literature, for e.g. in [Bishop, 2006]). It has been applied also in the tech-mining literature to build up technology maps [Porter, 2007]. In the present context, we are interested in visualizing the linkages between the semantic structure of a corpus and to use a tool alternative to taxonomy. As in the previous section, this map can be built using topics or words. In the case of a topic map, the MDS is simply run in the similarity KL matrix of topics. In the case of word map, we consider the following different strategy

- We first run the MDS algorithms on the topics.
- We considered a weight interpolation between words in the each topic in the 2D space. This operation of interpolation would produce the coordinate of each word in the 2D space giving more importance to the position of relevant topics. In practice we compute coordinate as the weighted mean of a words given the topics map.

- We set the dimension of the font of a word proportional to the inverse of its variance in the topic map.

The main idea of using this approach is to reduce the computational load of the MDS algorithm since the dimension of word similarity matrix is large. An example of a topic map thus generated is provided in Figure 14, below.

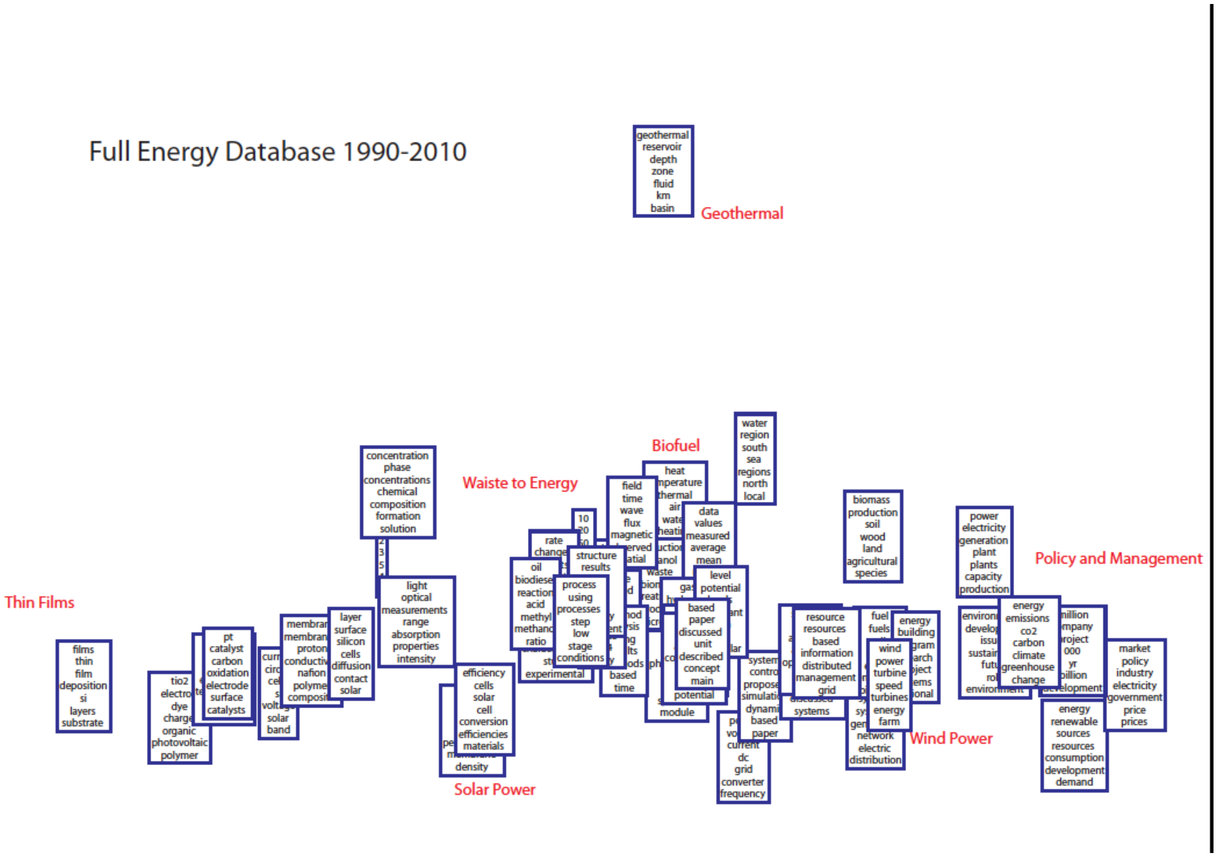


Figure 14 Topic map generated using LDA analysis

In Figure 15 an example of a topic map of our database is shown after applying the MDS algorithm in the GUI similar to Figure 14. The plot allowed zooming in and out to better inspect the map if the elements are overlapping or too close to each other. From the plot one can discern the distinction of topics in the renewable energy field: In the left hand side of the graph topics related to photovoltaic, thin films, solar energy are placed. In the right hand side the mixture of words are more related to policy and economy arguments. Alone in the bottom part, few topics related to the geothermal industry.

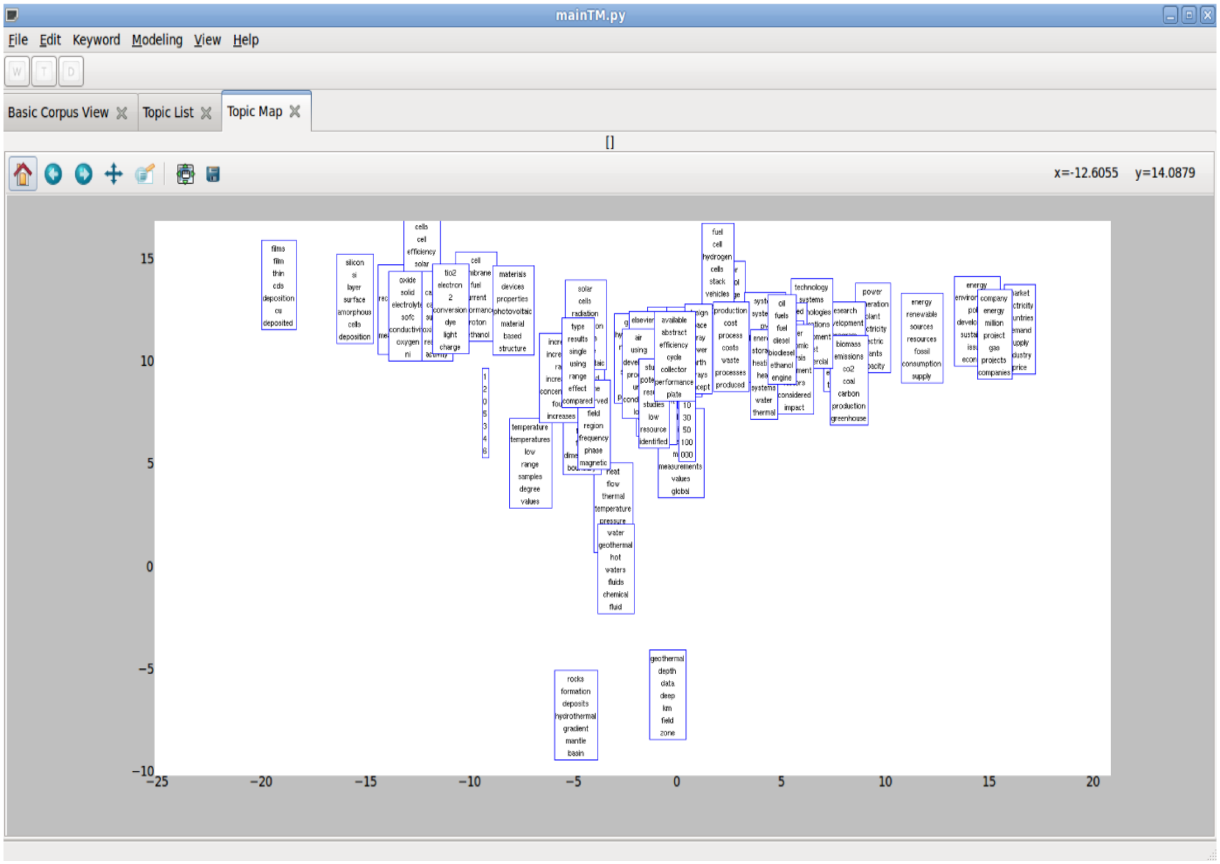

mainTM.py
File
Edit
Keyword
Modeling
View
Help
W
T
D
Basic Corpus View
Topic List
Topic Map
[]
x=-12.6055 y=14.0879
15
10
5
0
−5
−10
−25
−20
−15
−10
−5
0
5
10
15
20
films film thin cds deposition cu deposited
silicon si layer surface amorphous cells deposition
cells cell efficiency solar
oxide solid electrolyte sofc conductivity oxygen ni
tio2 electron 2 conversion dye light charge
materials devices properties photovoltaic material based structure
solar cells radiation
type results single using range effect compared
temperature temperatures low range samples degree values
field region frequency phase magnetic
heat flow thermal temperature pressure
water geothermal hot waters fluids chemical fluid
available abstract efficiency cycle collector performance plate
fuel cell hydrogen cells stack vehicles
production cost process costs waste processes produced
oil fuels fuel diesel biodiesel ethanol engine
technology systems
power generation plant electricity
biomass emissions co2 coal carbon production greenhouse
energy renewable sources resources fossil consumption supply
company energy million project gas projects companies
market electricity countries demand supply industry price
10 30 50 100 000 measurements values global
studies low resource identified
considered impact
water thermal
rocks formation deposits hydrothermal gradient mantle basin
geothermal depth data deep km field zone

Figure 15 Topic map generated using LDA analysis

In Figure 16 an example of a word map populated by single words is presented. The use of this kind of map has the advantage that single words are more directly interpretable than topics, i.e. mixtures of words. On the other end, this will creates additional challenges in the understanding of the corresponding visualization when single words from the corpus are allowed to populate the map. Our targets for this type of problem is to perform LDA using only keywords or noun phrases with “relevant” meanings, i.e. solar energy, biomass, wind power and so on, as doing so will allow the resulting visualizations to be directly interpreted by the user. A manual method is to provide them as a list to the program. Another possibility is to use a automatic/semiautomatic technique, discussed in the literature, which performs the labeling of the topics using noun phrases [Mei et al., 2007].

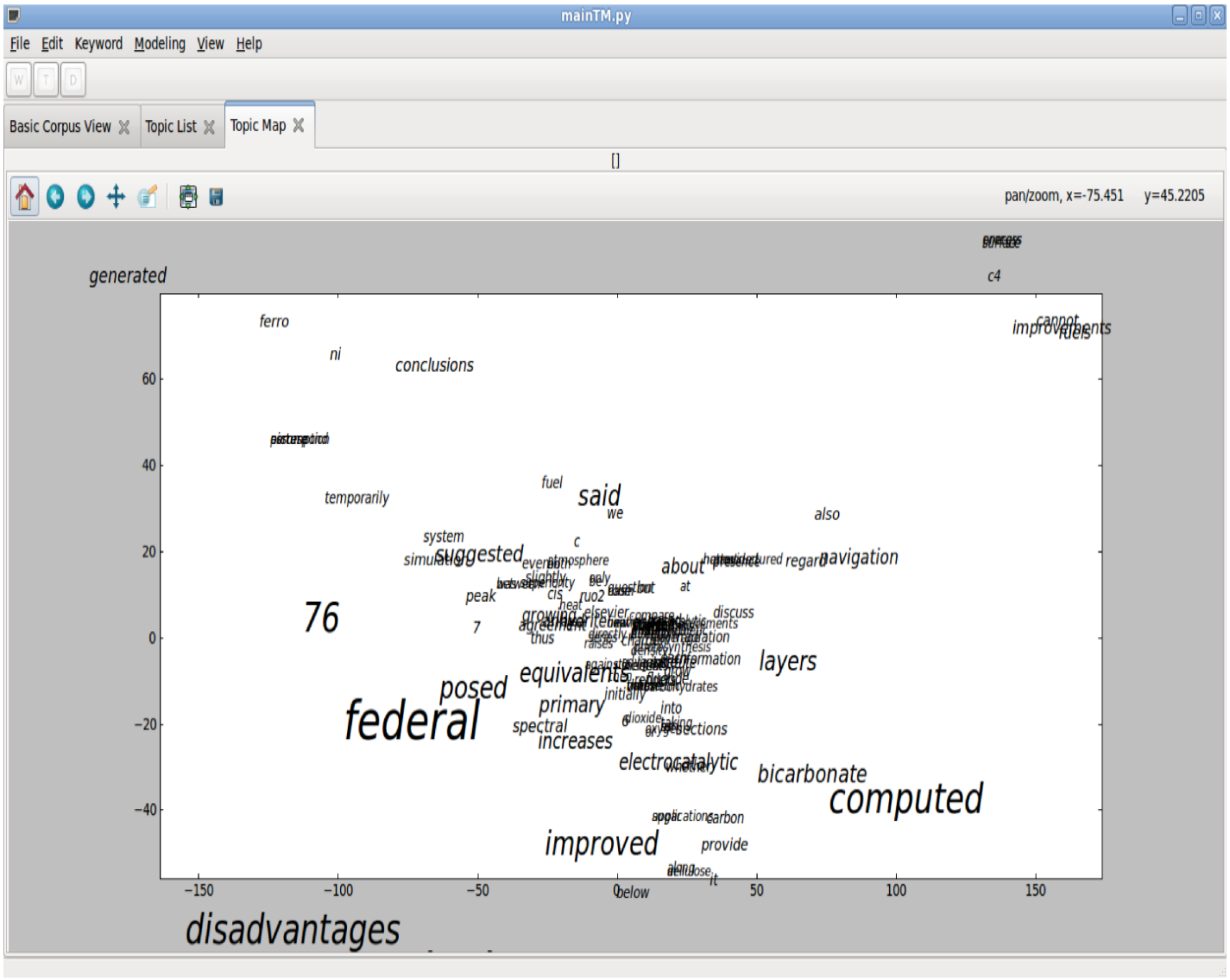


Figure 16 Word map populated using only representative words for each topic

Bibliometric Analysis of Blogs

This section describes the algorithms and process for studying blogs which were developed in the PhD research undertaken by Satwik Seshasai at MIT. The proposed method consists of the following three phases:

Phase 1: study blogs and derive insights,

Phase 2: present the insights in the context of a system representation

Phase 3: interview stakeholders to assess the results.

Background – CLIOS system representation

Firstly, it is important to understand the mechanics of the "CLIOS system representation" process. This is a graphical system visualization tool which captures qualitative information about the structure of a domain.

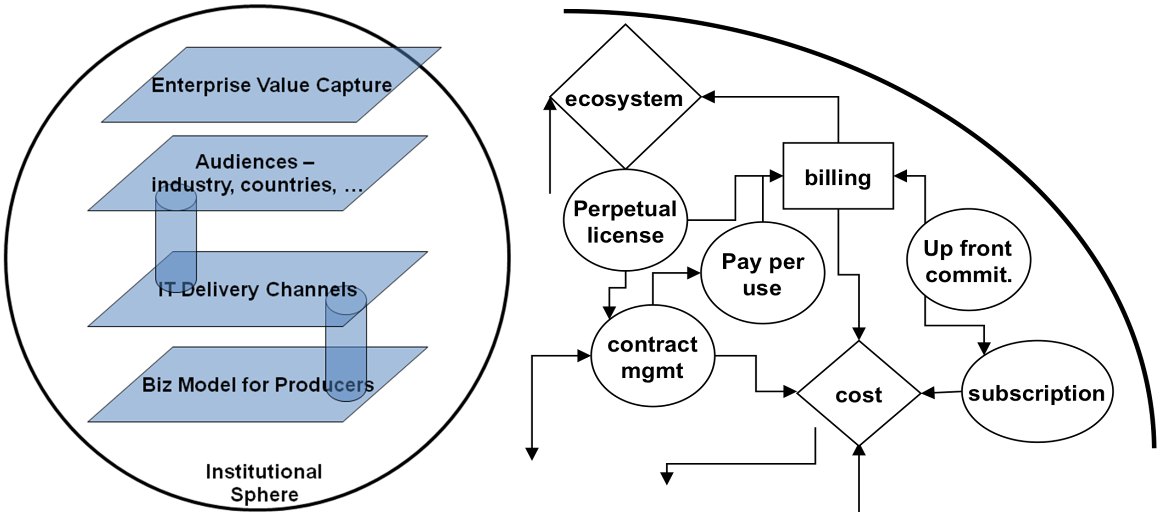


Figure 17 Partial CLIOS system representation diagram

Figure 17 shows an example of a graphical diagram used to show a system representation. Graphical diagrams accompanied by paragraphs and tables of text are a common way to produce the system representation. The subsystems are graphically represented as 2-dimensional planes which are surrounded by a common "institutional sphere" which contains institutions such as government bodies which affect the overall system.

The components within each subsystem are then defined, often through interviews with the various stakeholders. Most components are simply depicted as regular components (with the name of the component within a circle) but there are also three special types of components which are prescribed by the CLIOS process:

1. Performance measures are components which indicate the performance of the system in the eyes of the stakeholder – for example, materials cost. They are depicted with a double line around the component.
2. Policy levers are components which are used by institutions in the institutional sphere to influence the subsystem – for example, an industry regulation. They are depicted with a rectangle.
3. Common drivers are components which are on multiple subsystems and thus show how one subsystem can drive another.

The graphical representation of each subsystem as a plane allows these common drivers to be depicted as cylinders which describe the way in which one subsystem can impact the others. Alongside the graphical representation, a few words of text describing the definition of each component is included.

The final step in the representation is to define the links between components. Links indicate some form of relationship between components and can be drawn as strong or weak and can also be directional in nature. The definition of a link is left purposely broad, and the accompanying text should describe the nature of each link, and what it means to be a strong or weak link.

The work on blogs was conducted within the framework of the CLIOS representation system, where developments in a sector (reflected in blogging patterns) are used to generate candidate changes to the corresponding representation. The output of phase 1 is a set of numerical indicators which are then used in phase 2 to generate candidate changes to the system representation.

For example, bibliometric analysis of blogging patterns might indicate that "term A", "term B", and "term C" have grown closer in relationship in the last 15 months of the 54 month period being studied. An appropriate candidate change which might result from this observation is that "a link should be drawn between "term A" and "term C". This change would be presented back to the expert stakeholder as a candidate change; it is up to the stakeholder to decide whether this change should be accepted.

Phase 1 – Bibliometric Blog Analysis

Requirements of Phase 1 – Extracting Concepts Over Time

The key goal of this phase is to identify links between terms by identifying logical concepts that exist within the overall set of terms and link some subset of the terms together. A concept is a grouping of terms which has some semantic meaning – "cats", "fish" and "dogs" may be grouped together because they are all types of domestic pets. Two key technical requirements have been identified which are used to motivate the proposed approach:

1. **Persistence**. Concepts which are extracted must persist over some reasonable period of time. The specific terms which constitute the concept may change over time, but the underlying concept should stay persistent if it is worthy of identification.
2. **Complexity**: Concepts need to be both interesting and relevant to the experts being interviewed. Simple pairs of terms are not sufficiently interesting or are likely to already be identified and known (for e.g. while the pairing of "energy" and "life" may persistent, it is also an obvious link which is unlikely to be of much interest to a stakeholder). Likewise, a large group of terms will probably be too diffuse to adequately represent a recognizable concept.

Step 1 – Produce Raw Hit Counts

To produce the raw data needed for the analysis, the labels for the components of the subsystem were used as keywords. These were then submitted to the Google blog search engine to obtain the number of "hits" for each of these terms. To help ensure relevancy, the set of blogs to be considered was scoped by including the name of the overall system (ex. "cloud computing") and the subsystem ("business models") in every single search query executed.

The next step is where most of the actual data is collected – by executing periodic searches. This is achieved by performing a comprehensive search for all of the keywords over a certain span of time – for example, the past 5 years – in an effort to see which trends are occurring over recent time. To determine the right frequency to use, searches were conducted on a yearly, quarterly, monthly and weekly basis, and it was found that the monthly search frequency was the most effective for the purposes of this study. This determination was made after searching for terms in the renewable energy domain and assessing the rate of growth of each term. The relative rates of growth were inspected for each frequency, and the results for the monthly searches exhibited the most reasonable fit along the growth curve. The choice of monthly frequency is effectively a design assumption of the study without any rigorous validation beyond this pilot study. The code can be configured to specify which period of frequency is used, so future study can explore further whether there is a more appropriate frequency to use.

For each month, we search for the hit count (a “hit count” is the number of results, or hits, for a particular search) of each keyword (alongside the overall system name and the subsystem name), and we search for the number of hits for every pair of keywords. For example, “install” and “subscription” were two keywords in the Business Models subsystem of the Cloud Computing domain. So, we conducted a search for the number of blog entries in the month of January 2005 which contained {“cloud computing”, “business models”,”install”} and {“cloud computing”, “business models”,”subscription”} and finally {“cloud computing”, “business models”,”install”, “subscription”}. Then, we did the same three searches for February 2005, and March 2005, and so on. We followed this same pattern for every single keyword, and every pair of keywords, in every subsystem. This produces a large data repository of “hit counts” that we can use for the other parts of the study.

Step 2 - Analyzing Hit Counts

Once the hit counts had been collected, various indications of similarity, such as Similarity Distance, Term Frequency or Latent Semantic Analysis (LSA), could be used to determine the merits of a particular “re-representation action” (in the work done so far, LSA and Similarity distances were the primary analytical techniques used and these are described in the following sections). Finally, in the Dynamic System Re-Representation section below, a set of atomic re-representation actions which can be taken against an existing system representation is described. These actions map directly to those specific steps one takes in building the initial system representation – identifying components, links, policy levers, etc. The table shown below maps re-representation actions to analytical method.

Relevant CLIOS re-representation	**Analytical Method**	**Details**
New components identified	Term frequency	Terms which increase in relative frequency over time are inspected to see if they should be new components in the system. Lowest priority among all methods.
New link between existing components	LSA	Since links may not be known, using ‘concepts’ from LSA (where terms can be in more than one concept) will suggest new links between components
Increased strength of link between existing components	Google similarity distance	Build clusters (with exclusive membership) among components, calculated on monthly basis, watch for new entrants into clusters

Existing components become performance measures	Google similarity distance Term freq.	Use keywords which indicate that these terms are being assessed to gauge the performance of the system Increasing frequency also inspected for perf measures
Existing components appear on other subsystems and become common drivers	Taxonomy building	Asymmetric tree of terms identified (subtopics, …). Provides a structured means for identifying whether components are appearing under other topics, over time (suggests they are becoming common drivers). "Fuzzy tree" concept of topic being in multiple nodes is also possible.
Policy levers identified with existing or new institutions	Google distance	Associate certain key terms that indicate policy levers Include institutions in all cluster creation
Re-grouping into new subsystems identified	Taxonomy building	If components are within a subtopic which doesn't align with the selected subsystems

Table 1 *CLIOS re-representation actions mapped to Analytical Methods.*

There are two main analytical methods which were exercised in the actual case studies – latent semantic analysis and similarity distance.

Step 2a - Analytical Method - Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a latent variable technique which detects underlying correlations in word occurrences and which represents these as "concepts". These in term are encoded as word vectors, where the values of the ordinates denote the relevance of each word to the respective concept.

Overall steps:

1. Extract concepts for each month / each subsystem – this step involves actually searching and analyzing the results for each month using LSA. In this step, there is a set of concepts produced for each month, and every concept includes a weighting for each term.

2. Parse "real concepts" at the max gap between values. The goal of "real concepts" are to select the few terms in a given concept which represent the actual underlying meaning of that concept. Since all concepts contain all terms, this step is focused on selecting and keeping only those terms which are part of the "real concept".

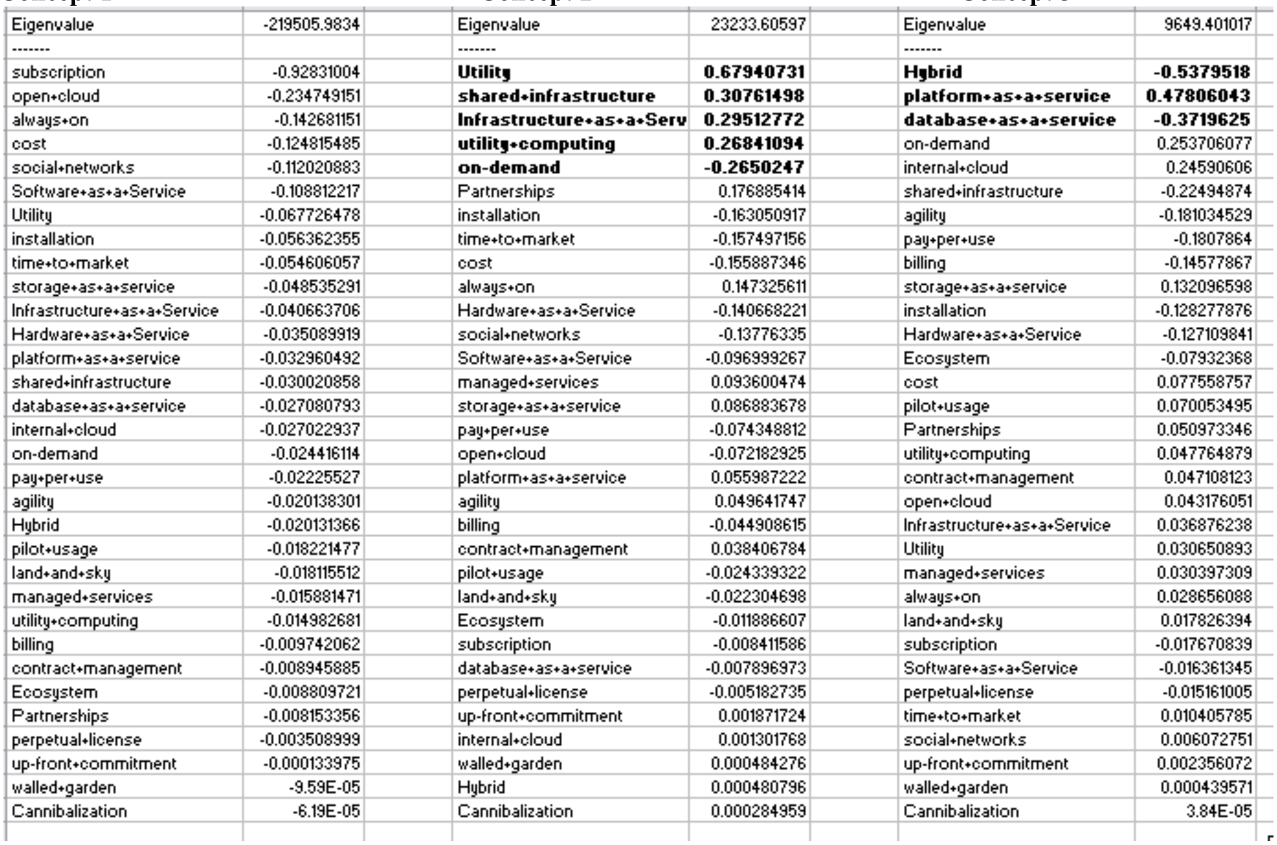
3. Identify concepts which grow/decay over time by indexing each concept by each pair of terms contained in that concept. If a concept has 4 terms in it, the concept is indexed 6 times – once for each possible combination of pairs of terms. The reason for doing this is that it allows pairs to be tracked over time, as this is the only method available for linking concepts over time. Thus, if in February there is a concept with 3 terms, and in March there is a concept with 4 terms, but both

concepts contain a common “anchor pair”, it can be said that the two concepts are the same and the third term in February and the third and fourth terms in March are considered to be associated with the anchor pair.

4. Select every concept with a pair that exists in at least 50 months (or some other threshold). Pairs which occur in 50 months or more are deemed to be worthy of presentation to the stakeholder, regardless of the size of the concepts in which the pairs appear.
5. Select every concept that exists in at least 20 months, but with less than 5 terms in each concept (or some other threshold). By constraining the number of terms in the concept to 5 or less, we can assure that each concept has some true underlying meaning. However, when concepts are constrained to 5 terms or less, the earlier constraint of 50 months has to be relaxed to a much lower threshold because the goal here is to include those concepts which could have only existed for the beginning or end of the overall time period.
6. Prepare results for presentation to stakeholder by replacing term IDs with names of terms and making any other changes required based on the software implementation of the algorithm.

The steps are discussed here in more detail. The goal of these steps is to identify new links between existing components. Our implementation of Latent Semantic Analysis (LSA) is based on an LSA simplification algorithm which relies on whether a term is in the blog post or not, rather than relying on how many times the term appears in the blog post. LSA takes a pre-defined set of keywords and produces a set of “concepts” where each concept contains every keyword, and a weighting for each keyword. For example, analyzing the three terms {dog, cat, bone}, might result in these three concepts: {dog 0.9, cat 0.8, bone 0.2},{dog 0.7, cat 0.3, bone 0.6},{dog 0.2, cat 0.2, bone 0.2}. Each of these concepts includes all three terms but in the first concept, dog and cat are weighted much higher than bone, and in the second concept dog and bone are weighted much higher than cat. This would suggest that there could be a link between dogs and cats (as both are terms that describe an animal), and between dog and bone (since dogs often enjoy chewing on bones). These weightings would not indicate a strong link between cats and bones, which makes intuitive sense. In our actual study, a subsystem often has anywhere from 20 to 30 keywords and, hence,20 to 30 concepts per month. With 54 months studied in our project, each of these concepts would have to be analyzed for each month, and the strongly linked keywords detected and extracted. After the links for each month are identified, the overall set of links is analyzed to see which links emerge over time and which links disappear over time, and this produces the set of candidate links to be proposed to the system re-representation.

From Step 1, we produced a set of hit counts for every keyword in the subsystem as well as every pair of keywords in the subsystem. This data was processed using the LSA simplification algorithm discussed above, to produce a set of concepts for every month. Each concept contained each keyword and also the weighted value of that keyword in that particular concept. So, for example, if there were 30 concepts in a subsystem, and there were 50 months studied, then a total of 1500 concepts were produced for the subsystem.



Concept 1		**Concept 2**		**Concept 3**	
Eigenvalue	-219505.9834	Eigenvalue	23233.60597	Eigenvalue	9649.401017
-------		-------		-------	
subscription	-0.92831004	**Utility**	**0.67940731**	**Hybrid**	**-0.5379518**
open+cloud	-0.234749151	**shared+infrastructure**	**0.30761498**	**platform+as+a+service**	**0.47806043**
always+on	-0.142681151	**Infrastructure+as+a+Serv**	**0.29512772**	**database+as+a+service**	**-0.3719625**
cost	-0.124815485	**utility+computing**	**0.26841094**	on-demand	0.253706077
social+networks	-0.112020883	**on-demand**	**-0.2650247**	internal+cloud	0.24590606
Software+as+a+Service	-0.108812217	Partnerships	0.176885414	shared+infrastructure	-0.22494874
Utility	-0.067726478	installation	-0.163050917	agility	-0.181034529
installation	-0.056362355	time+to+market	-0.157497156	pay+per+use	-0.1807864
time+to+market	-0.054606057	cost	-0.155887346	billing	-0.14577867
storage+as+a+service	-0.048535291	always+on	0.147325611	storage+as+a+service	0.132096598
Infrastructure+as+a+Service	-0.040663706	Hardware+as+a+Service	-0.140668221	installation	-0.128277876
Hardware+as+a+Service	-0.035089919	social+networks	-0.13776335	Hardware+as+a+Service	-0.127109841
platform+as+a+service	-0.032960492	Software+as+a+Service	-0.096999267	Ecosystem	-0.07932368
shared+infrastructure	-0.030020858	managed+services	0.093600474	cost	0.077558757
database+as+a+service	-0.027080793	storage+as+a+service	0.086883678	pilot+usage	0.070053495
internal+cloud	-0.027022937	pay+per+use	-0.074348812	Partnerships	0.050973346
on-demand	-0.024416114	open+cloud	-0.072182925	utility+computing	0.047764879
pay+per+use	-0.02225527	platform+as+a+service	0.055987222	contract+management	0.047108123
agility	-0.020138301	agility	0.049641747	open+cloud	0.043176051
Hybrid	-0.020131366	billing	-0.044908615	Infrastructure+as+a+Service	0.036876238
pilot+usage	-0.018221477	contract+management	0.038406784	Utility	0.030650893
land+and+sky	-0.018115512	pilot+usage	-0.024339322	managed+services	0.030397309
managed+services	-0.015881471	land+and+sky	-0.022304698	always+on	0.028656088
utility+computing	-0.014982681	Ecosystem	-0.011886607	land+and+sky	0.017826394
billing	-0.009742062	subscription	-0.008411586	subscription	-0.017670839
contract+management	-0.008945885	database+as+a+service	-0.007896973	Software+as+a+Service	-0.016361345
Ecosystem	-0.008809721	perpetual+license	-0.005182735	perpetual+license	-0.015161005
Partnerships	-0.008153356	up-front+commitment	0.001871724	time+to+market	0.010405785
perpetual+license	-0.003508999	internal+cloud	0.001301768	social+networks	0.006072751
up-front+commitment	-0.000133975	walled+garden	0.000484276	up-front+commitment	0.002356072
walled+garden	-9.59E-05	Hybrid	0.000480796	walled+garden	0.000439571
Cannibalization	-6.19E-05	Cannibalization	0.000284959	Cannibalization	3.84E-05

Figure 18 Sample LSA Concepts

Figure 18 depicts a sample set of concepts produced by an LSA analysis. The terms are listed alongside their relative weightings within the concept. Terms in bold represent those terms which are actually relevant to the concept.

The next step is to walk through each of these concepts, and pull out those few keywords which were most relevant to the concept – this is essentially executing the 'dog' and 'cat' example above. With approximately 30 words per concept, though, this required an algorithmic approach rather than the intuitive approach used in the 'dog' and 'cat' example. A set of options were considered and the best approach was decided to be first ordering the keywords in a concept by their values, then identifying the largest gap between values and using this as the place to draw the line between those keywords which are in or out of the concept. In figure 3.3, concept 2, the gap between “on-demand” and “partnerships” is the maximum gap, and so this is where the distinction was made between those terms which remained in the concept and those which were removed. This is similar to how professors sometimes put student grades in rank order and decide on “A” vs “B” designations simply looking for natural breaks in the test scores. For most of the concepts, this largest gap actually came after only the first keyword – so these concepts became easy to simply disregard for our study, since they won’t help us find a link between two or more keywords!

The next step is to list all concepts found in each month (now, only including those few keywords in each concept which made it above the cut of the previous step) and then by observation, looking month by month at which concepts stayed consistent over time and which concepts either were introduced or disappeared over time. This step is to meet the requirement of persistence stated earlier. After multiple approaches, the best approach seemed to be to identify those pairs which stay consistent over time (i.e., those pairs which show up for a relatively consistent set of months over a 5 year period) and then look for those other terms which join with the pair to become an n-tuple concept. Then, both the initial pair which is considered, and the occasional terms which enter into a concept with the pair, are all considered as candidate changes to propose back to the stakeholder for inclusion as new links in the CLIOS subsystem. An example of an anchor pair is described in a later paragraph in this section.

Two sets of concepts are collected independently, and the combination of these two sets meets the requirement of complexity stated earlier. One is the set of pairs which occur in some large percentage of the months (ex. 50 of 60 months) but with no limitation on the number of terms in the concept. These are pairs which are so prevalent in the blogs that they deserve to be presented to the stakeholder. The other is the set of pairs which occur in a limited set of months (ex. 20 months out of 60) but occur within a “reasonably” sized concept such as a concept with 5 terms or less. As discussed earlier, by constraining the number of terms in the concept to 5 or less, we can assure that each concept has some true underlying meaning. However, when concepts are constrained to 5 terms or less, the earlier constraint of 50 months has to be relaxed to a much lower threshold because the goal here is to include those concepts which could have only existed for the beginning or end of the overall time period. Each case study includes examples of concepts which emerged from both types of thresholds.

We present here a sample set of results to show an example of what comes out of this particular analytical method. In a test run on “Cloud Computing”, one pair of terms that showed up consistently in the Business Models was “install” and “subscription”. This was a new link which was proposed to the stakeholder and made intuitive sense because with the easy installation process that cloud computing offers (the servers are managed by the vendor and almost no install is actually required), a subscription model can be offered to clients. For this pair of terms, studied over 2005 to 2009, the term "cost" was a third term often found in triples with "install" and "subscription". This made intuitive sense as well, but was already considered in the original representation. Other associated terms which showed up less frequently were:

- "time to market" (which was already linked to "install" in the original subsystem),
- "open cloud" (which was not originally considered but made intuitive sense because an open cloud allows anyone to "install" or get started more easily),
- "hardware as a service" (which was not originally considered, and generated a new business idea for the stakeholder around offering subscription based appliances), and
- "land and sky" (which made no intuitive sense to the stakeholder).

So of the sample results above, we found a handful of new links which actually could make sense on the re-representation, one particular link which was extra-special because it resulted in a new business idea, and a few links which were proposed but either were already known or made no intuitive sense to the stakeholder. This is just one example; the full analysis shown in the later section has many more candidate changes which came from this one subsystem.

These new links (produced in phase 1 of the process), when embedded into the original subsystem (phase 2 of the process), should result in new insights for the expert. The final phase of the methodology is presenting these candidate changes back to the expert (phase 3 of the process), and observing which changes were not previously considered but were accepted – if there are enough new links that fit into this category, then this tool can be considered worthwhile. Then, as we interview the expert to see what they were able to do as a result of identifying these new links, the specific hypotheses around topics such as new market approaches to existing technologies can each be assessed.

Step 2b - Analytical Method – Similarity Distance

This method is used when the focus is on determining the relationship between exactly two terms. These two terms could each be components in a subsystem, or they could be one component and another term such as "cost". The goal is to find the strength of the relationship between two terms by calculating the probability that if one term appears in a blog post, the other term would appear. This is calculated by first calculating the number of hits for each term, and then the number of hits for the pair of terms. The normalized Google distance [Cilibrasi 2007] is defined as:

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}$$

(In this formula, f(x) is the number of hits for term x, f(y) is the number of hits for term y, f(x,y) is the number of hits for the pair of terms, and N is an approximation of the total number of hits).

The CLIOS system representation process includes an indication of strength for every link between two components. This calculation allows us to assess the strength of links in the CLIOS system over time. By calculating the similarity distance on a month to month basis between two components, we can calculate the rate of growth. If the distance between two terms decreases over time, it is an indication that these two terms are growing closer in relationship. By ranking every pair of components in terms of the rate of growth in similarity distance over time, it is possible to identify those links in the system representation which could be described as stronger.

Another aspect of the CLIOS system representation is the identification of performance measures. These performance measures are used to assess the overall performance of the system towards the goals of the stakeholders for whom the representation was created. For example, “broadband speed” may be a performance measure in the Kenya broadband system, if the stakeholders see increasing the available speed as a primary goal of the system. In this study, the concept of “indicator terms” is introduced. An “indicator term” is a term which indicates that related terms may be a performance measure. To test this method, the following indicator terms were chosen: “performance”, “cost”, “important”, “critical”, ”revenue”, ”market share”, ”effectiveness”, ”speed”, ”growth”. In this study, the similarity distance between every component and each of these indicator terms was calculated on a month to month basis. This produced a set of components which showed a stronger correlation to “cost”, to “important”, and to each of the other indicator terms. These components were proposed as performance measure in the system.

The CLIOS process also includes the identification of policy levers. A policy lever is a component which is used by a member of the institutional sphere to affect the subsystem. Thus, a growing relationship between an institution such as the Federal Communications Commission and a component such as broadband speed could suggest that broadband speed were becoming a policy lever. It is possible to use this same method for identifying policy levers, but that is left for future study.

Visualizations were used to present these results back to the stakeholders. They were chosen so that the interview time with the stakeholder could be kept manageable and so that the stakeholder could quickly scan the various results and report on which candidate changes they chose to select.

The algorithm used to implement similarity distance calculation for link strength are:

1. Calculate Google similarity distance for each pair of terms, monthly
2. Calculate similarity growth/decay over time for each pair using rate of growth against linear, exponential, Gompertz curves
3. Rank pairs by the rate their similarity changes, and identify the fastest growing pairs
4. Visualize network graph of fastest growing pairs as candidate changes to the system representation
5. Present fastest growing pairs in a visual fashion to stakeholder to select

The steps used to implement similarity distance calculation for performance measures were the same, except that instead of every term pair, the pairs consisted of every term coupled with every indicator term. A line graph was produced for each indicator term, which showed those terms which were strongly correlated to that indicator term. These terms were presented back to the stakeholder as candidate performance measures

Phase 2 – Dynamic System Re-Representation

The concept of "dynamic system re-representation" is that a CLIOS system representation which normally takes significant stakeholder interaction to produce may be informed hourly or daily by new information sources. The goal is not to actually change the system representation on a dynamic basis, but instead to treat the output of the quantitative analysis as a qualitative stakeholder, similar to how any other qualitative stakeholder would themselves use quantitative data to provide input. Certain domains are moving at such a rapid pace that incorporating new sources of information dynamically will be very beneficial. Information sources could reveal new institutions, new individuals, new geographies, which have interest in the particular CLIOS.

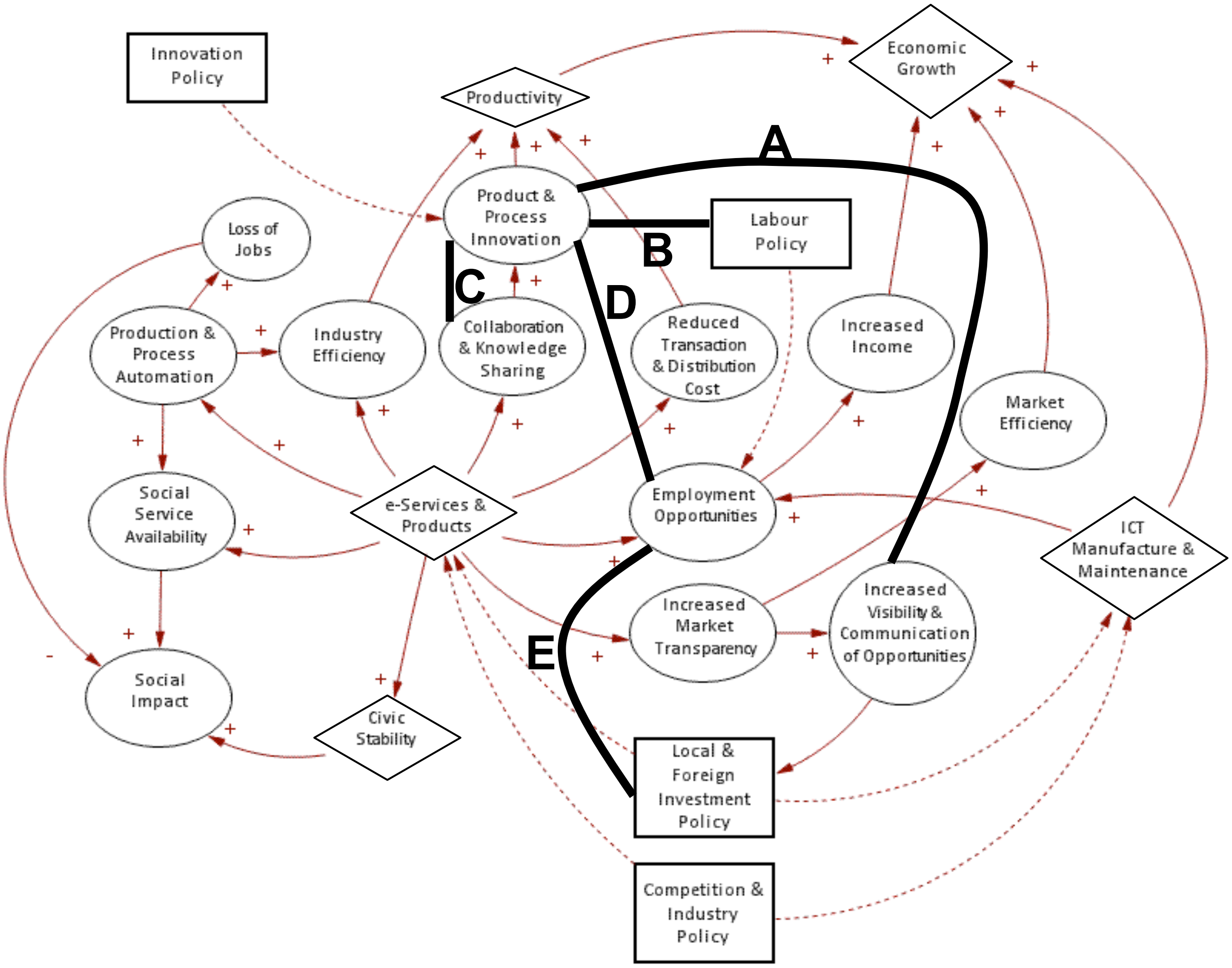

Innovation Policy
Productivity
Economic Growth
A
Product & Process Innovation
Labour Policy
B
Loss of Jobs
C
D
Production & Process Automation
Industry Efficiency
Collaboration & Knowledge Sharing
Reduced Transaction & Distribution Cost
Increased Income
Market Efficiency
Social Service Availability
e-Services & Products
Employment Opportunities
ICT Manufacture & Maintenance
Increased Market Transparency
Increased Visibility & Communication of Opportunities
E
Social Impact
Civic Stability
Local & Foreign Investment Policy
Competition & Industry Policy

Figure 19 Dynamic re-representation of a CLIOS system.

Figure 19 is another example of a CLIOS system representation. This example is from a study which we conducted on the Kenya broadband subsystem. In this figure, ovals represent regular components, rectangles are policy levers used by institutional actors, and diamonds represent drivers common to multiple physical subsystems. Five new links are shown, which were identified through the implementation of the methodology being described in this section. Each of these links represents a new insight gathered from the bibliometric blog analysis and accepted by the expert interviewee as worthy of inclusion in the system re-representation. The goal of this step is to actually adjust the system representation to one which is useful to the stakeholder in making technology strategy decisions. Thus, the fact that the selected changes are based on personal opinion of the stakeholder is acceptable. When this process is used in a real situation, it is expected that the algorithm in phase 1 will be repeated on a periodic basis, so that the system representation can be re-represented to accommodate new insights being found over time.

Phase 3 - Interviewing Stakeholders to Validate the Results

The candidate changes to the system representation which were produced in phase 1 are then presented back to the stakeholder in phase 3. In this study, interviews were conducted with expert stakeholders where they were asked to select which candidate changes to the system representation they would accept into the system representation. This step is heavily dependent on the particular stakeholders being interviewed. Given the same set of candidate changes, it is expected that different stakeholders would select different candidate changes to accept into the system. This is no different than stakeholder input into the initial representation, where different stakeholders would suggest different components to be added into the representation.

In the case studies, the goal was both to test the candidate changes produced by the algorithm for usefulness as well as to understand what types of technology strategy decisions would be made based on this tool. Based on the interviews which were conducted, certain methods were found to be more useful than others in reaching insights which drove technology strategy decisions.

Face Validity as the Distinguishing Characteristic

The interview protocol detailed below is designed to collect expert opinion on the results which were obtained by the various analytical methods. The interview design discussed below includes two key phases – in the first phase, the experts are asked to classify each insight into one of three categories while the second phase is a more general discussion about the value of this research to the expert. The experts were asked to put insights from the research into these three categories:

1. Insights which have face validity and are already known to the expert. Face validity refers to an expert's assessment on whether an assertion presented to them has the appearance of being valid [Mosier 1947].
2. Insights which are not already known but after further exploration could have face validity.
3. Insights for which the expert is sure they do not have face validity.

By putting insights into these three categories, we were able to both assess the validity of the research method in terms of finding insights which aligned to general expert opinion as well as finding insights which resulted in added value to the experts. In the actual interviews, the explanation of these categories took time, and did not fully sink in to any of the experts interviewed until a few insights were presented and the expert developed a familiarity with the three categories.

Interview Protocol

The interviews consisted of two phases. The first was a quick review of the candidate changes, asking the stakeholder to categorize each change into one of three categories as described above. The second was a more general discussion, focused on the true impact which the stakeholder saw coming from the overall set of results in this project.

Before beginning the main questions, the interviewee would describe this research project, explain the CLIOS process briefly, and go over the whole system representation and ask if there are any obvious gaps. This is done to make sure that the system representation produced by the original stakeholder was not invalid to this stakeholder. Next, each candidate change would be reviewed and categorized. This phase was often cut short because an interesting discussion around impact of the project was more critical.

In the second phase of the interview, the following set of questions was used, until the interviewee led into a particular discussion that was important to them:

- What is your intuition behind why this link showed up in the analysis?
- Does this link suggest a relationship between any existing groups within your organization that should exist but doesn't already?
- Does this link suggest a new audience for any existing products or solutions offered by your firm?
- Does this link suggest any business development activities (acquisitions, partnerships …) for your organization?
- Does this link suggest any new product ideas for your organization?
- Does the CLIOS representation help you understand the problem domain any better? If so, how?
- What is your intuition on the face validity of the new aspects to the system re-representation that were found?
- Are there any high level observations you can make about the domain given the analysis?
- Would you like to repeat this analysis on a monthly or yearly basis, or is the one time sufficient?
- Did this result in too many suggested changes to the CLIOS representation, thus making it too complicated to be useful?

As part of his research, Mr. Seshasai paid a visit to the Masdar Institute and conducted a series of interviews based on the methodology described in the preceding sections. His findings are described in the following subsection.

Using Blogs and Bibliometrics to Analyze Renewable Energy: Summary of Findings

Case Study Structure and Experts Interviewed

This case study focused on renewable energy and academia. Renewable energy is a very broad domain referring to the various efforts to leverage sources of energy that do not face the supply challenges of fossil fuels. The system used for study in this case is an aggregation of two specific technical areas in renewable energy – biofuels and smart grids – and the related aspects of energy consumption and socio-technical factors. These two areas were chosen in part due to the subject matter expertise of the interviewees, because the real goal of this case study was to provide a vehicle for understanding how the research in this dissertation can be applied to the academic community.

All of the interviewees in this study are members of the faculty or staff at the Masdar Institute of Science and Technology in Abu Dhabi. Masdar is a cross disciplinary institute where the faculty is focused on progressing research around renewable energy solutions from a variety of disciplines. The case study was conducted by spending a week at the Masdar Institute, understanding the various domains in which the professors work, and then conducting interviews in the same format as the cloud computing and Kenya broadband case studies. By spending an extended period of time with the faculty at Masdar, it was possible to prepare more focused questions on how the research can apply to the academic community.

The specific fields in which the interviewees work are as follows:

- Biofuels – this professor focuses on the production of fuels based on biological materials. [Hashaikeh 2010]
- Electrical Power and Smart Grids – this professor focuses on the distributed generation of electrical power through "smart grids" which optimize the distribution of electricity [Zeineldin 2010]
- Wave Energy – this professor focuses on harnessing the energy generated from ocean or river waves [Tabaei 2010]
- Materials Science and Solar Energy – this professor focuses on the use of innovative materials to capture and store energy from sunlight [Chiesa 2010]
- Materials Science and Solar Energy – this professor also focuses on the use of innovative materials to capture and store energy from sunlight [Emziane 2010]
- Engineering Systems –this professor focuses on the notion of sustainable energy and the various dynamics involved in large-scale energy systems [Sgouridis 2010]

- Project Management – this individual is a project manager who manages research projects at the Masdar Institute, from funding applications through execution of the project [Prieto 2010]

Renewable energy domain background

The background on renewable energy was collected by discussions with the interviewees on their areas of research and challenges being faced by each of their areas. This background provided a context for the qualitative interviews which examined the overall relevance of this research to the academic field.

Masdar Institute

The case study was conducted by interviewing professors and staff at the Masdar Institute. This research institute was funded in part by the government of the United Arab Emirates, and founded in conjunction with the Massachusetts Institute of Technology. The Institute includes professors from a variety of science, engineering, and other fields, whose research all focuses on some facet of the renewable energy domain. The funding backdrop and the location in Abu Dhabi where energy production is not just a technical issue, makes a systems level analysis of the renewable energy domain particularly useful. The Institute is housed within the Masdar City, which is a 6 square kilometer city that is designed to be carbon neutral by implementation of a number of the technologies being developed at the Institute. The political context of energy, the funding and location in Abu Dhabi, as well as the juxtaposition within a city project, makes public opinion and perception a very important input – this arose in the interviews as a significant value provided by this research.

In the following paragraphs, relevant pieces of background on the different areas covered by the interviewees are briefly discussed.

Solar Energy

The key piece of insight on the field of solar energy is that an entire "solution" is needed for success, which includes awareness of funding sources, physical location of the desired solar energy recipient, building design, and materials access [Chiesa 2010]. For example, Masdar City is built on an open desert which has a high degree of sunlight and open space, and can afford to utilize that open space to house solar panels which collect and store solar energy. In Masdar City, the location also presents cooling as a primary use of solar energy, and this use case has influenced the type of solar technology which has been built. Rather than using photovoltaic panels which convert solar energy to electrical power, mirrors are used to collect solar energy and heat liquids that are collocated near buildings that need to be cooled. This is an example of a solution which was built by combining an understanding of the geographic perspective, technical perspective, and funding perspective. The interview feedback discusses the need to collect these perspectives as a motivation for using this research.

Electric Cars

In the oil based culture of Abu Dhabi, electric car adoption has a variety of technical and non-technical factors which influence it [Chiesa 2010]. Technical factors include the need to support power for air conditioning over long distances. Non-technical factors include the impact of promoting electric cars on the oil-based businesses in Abu Dhabi.

Biofuels

In the area of biofuels, the technical focus is on building multi-carbon based compounds as fuel sources [Hashaikeh 2010]. When non-technical factors such as government funding are included, terminology becomes a key factor for which blog discussion may provide useful insights. The public awareness is largely around ethanol, which is a 2-carbon based fuel source, and the general field of biofuels is often equated in public discourse to ethanol [Hashaikeh 2010].

Wave Energy

Wave energy is a relatively new field compared to solar and biofuels, but is trying to learn from the adoption patterns of solar energy and electric cars [Tabaei 2010]. Issues around storage and transport of wave energy, and the geographic proximity of the energy source (i.e., oceans) and the energy consumers, and the type of technology required for oceans versus rivers, are factors that are very similar to what are faced in the solar area. Additionally, the marketing of wave energy and explanation to consumers has the potential to learn from electric cars.

Electrical Power - Distributed Generation

Distributed generation is the technical term that refers to optimizing the distribution of electrical power across a region. This is being achieved through so-called “smart grids” which seek to monitor the utilization of electrical power and then allocate energy appropriately. A major focus of Information Technology (IT) investment in energy is around this area because of the analytics and algorithms required [Zeineldin 2010]. This field, as with the others, also has regional factors – different regions have different local energy distributors and different needs based on how physically distributed the consumers are. Terminology and the use of terms such as smart grids versus distributed generation is also inconsistent between regions [Zeineldin 2010].

Domain Summary

In summary, the renewable energy domain is made up of parallel areas of technical research, each grounded in a particular set of scientific and engineering disciplines, but with common concerns around energy consumption, regional differences, funding and public perception.

Initial CLIOS Representation

The goal of this CLIOS representation was to illustrate the potential of the research, rather than to reach any formal conclusions around the domain of renewable energy. The system representation was created by examining content provided by the IBM HorizonWatch program and initial interviews with the Masdar faculty. Given the time constraints around the interviews and the true goal of examining output of the bibliometric blog analysis, the focus was on identifying the right subsystems and components to include in the representation, and thus the representation did not include an initial diagram with component types and links.

The four subsystems were chosen to provide some analysis of two specific areas of renewable energy, and then two subsystems which cut across all areas of renewable energy. Smart Grid was chosen due to its relation to the IT industry and the prevalence of IT related topics in blogs. Biofuels were chosen because of the prevalence of this area in public discourse and political discussion, which could translate to increased discussion in blogs. Other areas could very well have been chosen, but the needs of this case study to study illustrative examples were met by choosing these two. A socio-economic factors subsystem was studied to include a set of factors which cut across the entire renewable energy domain. The components selected in this subsystem were taken directly from the Kenya broadband CLIOS representation, because those topics represented a good illustrative set of socio-economic factors. The final subsystem is energy consumption. This subsystem includes a set of components which relate to how any energy source is distributed and consumed by individuals and businesses.

The specific components used in each subsystem are listed in Table 2.

Smart Grid "smart grid","average system availability","automated meter reading","building automation system","Conservation voltage regulation","Customer Average Interruption Duration","demand side management","distributed generation","hydroelectric plant","load management","rolling blackout"
Biofuels "bioenergy","biomass","biofuel","cellulose","bioreactor","glucose","enzymes","ethanol","E10"," E85","solar energy","moisture","total ash“
Socio-economic Factors (taken from Kenya Broadband) "product process innovation","jobs","production process automation","industry efficiency","collaboration knowledge sharing","transaction distribution costs","inice availability","market efficiency","eyment opportunities","social impact","market transparency","visibility communication opportunities","economic activity","productivity","manufacture maintenance","innovation policy","labor policy","local foreign investment policy","anti-trust industry policy"

Energy Consumption
"asset management","deregulation","energy service provider","rate of return", "reliability", "residential", "security","service area“

Table 2 *List of components in the Renewable Energy system.*

Candidate Changes to System Representation

Two forms of candidate changes were included in this case – new links, based on the Latent Semantic Analysis method, and strength of links, based on the Google Similarity Distance method. A sample set of results is provided here to illustrate the types of changes which resulted, and the detailed feedback received on certain changes.

New Links

Sample new links are shown below from the two technical subsystems, smart grid and biofuels. In each, the “anchor pair” is listed in bold, and then the related terms are listed below, followed by comments from the interviewees. Each row has a number following it, which indicates the category which the interviewee placed that specific new link.

The first two sample results are from the Smart Grid subsystem.

hydroelectric plant and distributed generation (9, 8) - 1
{"['load management', 'smart grid', 'demand side management']": ['Feb07'], - 1
"['['demand side management']": ['May07', 'Nov07'], - 1
mart grid', 'load management']": ['Apr07'], - 1
"['smart grid', 'load management', 'demand side management']": ['Jan07'], 1
"['smart grid', 'demand side management', 'building automation system']": ['May06'], 1
'[]': ['Mar06', 'Mar06', 'May06', 'May06', 'Aug06', 'Jan07', 'Feb07', 'Apr07', 'May07', 'May07', 'Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Nov07', 'Nov07', 'Dec07', 'Jan08', 'Apr08'], - 2
"['building automation system', 'demand side management']": ['May06'], - 3
"['load management']": ['Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Jan08', 'Apr08'], - 1
"['smart grid', 'demand side management']": ['Nov07']} – 1
load management and distributed generation (10, 8) - 1
{"['smart grid', 'demand side management', 'hydroelectric plant']": ['Feb07'], - 1
"['hydroelectric plant']": ['Jun07', 'Jun07', 'Sep07', 'Oct07', 'Oct07', 'Jan08', 'Apr08'], - 1
"['smart grid']": ['Feb07'], - 2
'[]': ['Jan06', 'Feb06', 'Feb06', 'Jan07', 'Jan07', 'Feb07', 'Feb07', 'Feb07', 'Mar07', 'Apr07', 'Apr07', 'Jun07', 'Jun07', 'Jun07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Oct07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08', 'Mar08', 'Apr08', 'Apr08', 'May08', 'May08', 'Jun09', 'Jun09', 'Feb10', 'Feb10'], 1
"['automated meter reading', 'Customer Average Interruption Duration']": ['Jan06'], - 1
"['hydroelectric plant', 'smart grid']": ['Apr07'], - 2
"['Conservation voltage regulation']": ['Apr08', 'May08'], - 1
"['smart grid', 'hydroelectric plant', 'demand side management']": ['Jan07'], - 1
"['automated meter reading']": ['Feb06']} – 1

In the above set of results, the expert in Smart Grids indicated that the link anchor pair was already known, as hydroelectric plants are used for distributed generation. Of the related terms under this anchor pair, “building automation system’ and “demand side management” were in the third category of being rejected by the interviewee because of the suggested link to hydroelectric plants [Zeineldin 2010]. The term “smart grid” is the one related term which the interviewee put in category 2 (face validity and worthy of further investigation) because although smart grid is the topic area for the overall subsystem, the technical component smart grid referring to an actual smart grid is not always linked to hydroelectric plants.

building automation system and smart grid (4, 1) - 1

```
{"['distributed generation', 'hydroelectric plant', 'demand side management']": ['May06'], 2
"['average system availability']": ['Aug09', 'Sep09', 'Jan10', 'Mar10', 'May10', 'Jun10'], 2
"['automated meter reading']": ['Jan09', 'Mar09'], 1
'[]': ['May06', 'Jan09', 'Feb09', 'Feb09', 'Mar09', 'Jun09', 'Jun09', 'Aug09', 'Aug09', 'Sep09',
'Sep09', 'Nov09', 'Jan10', 'Jan10', 'Feb10', 'Feb10', 'Mar10', 'Mar10', 'May10', 'May10', 'Jun10',
'Jun10']}, 1
```

The above anchor pair, “building automation system” and “smart grid”, was also listed in category 1 as a known link. The timing of this link is 2009 to 2010, shown by the months listed in the row with the “[]”. The expert indicated that this timing made sense given the recent introduction of smart grids as part of the considerations for building construction [Zeineldin 2010]. The triple of related terms, 'distributed generation', 'hydroelectric plant', and 'demand side management', were put into category 2 primarily because of the introduction of “demand side management”. Another related term, “average system availability”, was put in category 2 only because the expert expected the timing to line up with the timing of the overall anchor pair.

The following result is from the Biofuels subsystem.

E85 and ethanol (10, 8) – 2

```
{"['biofuel', 'bioenergy']": ['Feb06'],
"['bioenergy', 'biofuel']": ['Aug07'],
"['bioenergy', 'biofuel', 'biomass']": ['Apr06', 'Sep08'],
"['cellulose', 'biofuel']": ['Oct07', 'Dec09'],
"['biofuel', 'bioenergy', 'biomass']": ['Jun09'],
"['solar+energy', 'bioenergy', 'biofuel']": ['Jun07'],
'[]': ['Feb06', 'Mar06', 'Mar06', 'Apr06', 'Oct06', 'Feb07', 'Apr07', 'May07', 'Jun07', 'Aug07',
'Aug07', 'Sep07', 'Sep07', 'Oct07', 'Oct07', 'Dec07', 'Jan08', 'Jan08', 'Feb08', 'Feb08', 'Mar08',
'Mar08', 'Apr08', 'May08', 'Jun08', 'Jul08', 'Jul08', 'Sep08', 'Nov08', 'Nov08', 'Dec08', 'Jan09',
'Jun09', 'Jul09', 'Jul09', 'Aug09', 'Sep09', 'Nov09', 'Dec09', 'Dec09', 'Jan10', 'Mar10', 'Apr10',
'May10', 'May10'],
"['biomass']": ['Oct07', 'Aug09', 'Sep09'],
"['cellulose', 'biomass', 'biofuel']": ['Jul08'],
"['biofuel', 'cellulose']": ['May10'],
"['cellulose', 'bioenergy', 'biomass']": ['Jul09'],
"['cellulose', 'bioenergy']": ['Jun08'],
"['bioenergy', 'biomass']": ['Nov08'],
"['cellulose', 'biomass', 'bioenergy']": ['Dec07', 'Jan08', 'Feb08', 'Mar08']}
```

The main feedback from the expert on this set of results is that ethanol was part of the anchor pair, and terms such as biofuel are related terms that only appear in certain months [Hashaikeh 2010]. Ethanol is a two-carbon based compound, and one specific form of biofuel. Although much of the research innovation is in multi-carbon based componts, ethanol is a much more well-known term publicly, and the expert would like to investigate further to understand whether research proposals which seek to study biofuels would be better served to include the word “ethanol” if going to government bodies or other agencies which are influenced by public perception.

The following anchor pair is a sample result from the socio-economic factors subsystem.

visibility+communication+opportunities and production+process+automation (13, 3)

This link was put in category 2 because the expert observed that production process automation in renewable energy depend on visibility and communication. Many of the technologies in renewable energy are new and production processes have not yet been optimized. With communication of opportunities, more opportunities for automation can be identified [Emziane 2010].

Common Drivers

The next set of results collected in this case study were around identification of new common drivers. This was done by comparing the Google Similarity Distance of each pair of terms between each subsystem, for each month. So, for the four subsystems, 6 tables were produced, one for each pair of subsystems. In each table, the term pairs which grew fastest in similarity over time were proposed as indicators of common drivers. As a reminder, a common driver is a component which appears on multiple subsystems and drives behavior from one subsystem into the other. In a system re-representation where a pair of terms is identified as being linked, this could either be achieved by replicating a component from one subsystem to another, or by declaring an independent component as a common driver that appears on both subsystems and is linked to each pair.

Table 3 is the table of fastest growing pairs from the Smart Grid vs Socio-economic Factors.

Cat.	Smart Grid	Socio-economic Factors
3	rolling+blackout	market+transparency
1	Conservation+voltage+regulation	industry+efficiency
1	Customer+Average+Interruption+Duration	industry+efficiency

1	hydroelectric+plant	jobs
3	smart+grid	collaboration+knowledge+sharing
3	smart+grid	social+service+availability
2	rolling+blackout	industry+efficiency
3	smart+grid	production+process+automation
3	building+automation+system	social+service+availability
3	building+automation+system	income
1	smart+grid	employment+opportunities
3	building+automation+system	production+process+automation
3	building+automation+system	transaction+distribution+costs
3	average+system+availability	transaction+distribution+costs
2	automated+meter+reading	transaction+distribution+costs
1	building+automation+system	manufacture+maintenance
2	automated+meter+reading	social+service+availability
3	building+automation+system	economic+activity
3	building+automation+system	employment+opportunities
2	smart+grid	income
3	smart+grid	manufacture+maintenance
2	automated+meter+reading	anti-trust+industry+policy
2	smart+grid	anti-trust+industry+policy
2	building+automation+system	labor+policy

Table 3 *Candidate common drivers between the Smart Grid and Socio-economic Factors subsystems*

In Table 3, the pairs listed each have one term from the Smart Grid subsystem and one term from the Socio-economic Factors subsystem. The interviewee was asked to place each pair in each of the three categories which are being used in all of the quantitative interviews [Emziane 2010]. For this type of candidate change, category 1 implies that the pair is represented in some fashion by an already known link between the subsystems. In other words, each pair in category 1 is not itself tied to an independent common driver. Pairs in category 2 are the ones which are worthy of further investigation, and which the interviewee believes there are potential new common drivers to be identified in the manner described earlier.

Three pairs from Table 3 which all fell into category 2 are discussed here. The components "rolling blackout" and "industry efficiency" might indicate that decisions made in the technical subsystem around how rolling blackouts are enacted could drive industry efficiency [Emziane 2010]. The components "automated meter reading" and "transaction distribution costs" are worthy of more investigation because it may indicate that the generic notion of "automation" is a common driver between the two subsystems

[Emziane 2010]. Finally, “smart grid” and “income” are put in category 2 simply because the relationship is not clear to the interviewee but if there were a relationship, it would be of interest [Emziane 2010].

Quantitative feedback on candidate changes

Six of the experts were each asked to go through the candidate changes and place changes into each of the three categories which have been used in this research. Due to time constraints in the interviews and the desire to have sufficient coverage of the qualitative portion of the interviews, each interviewee did not cover every candidate change. Interviewees with specific subject matter expertise in specific areas started the interviews with their own areas of expertise.

Interviewee - Energy	**Known Link (1)**	**Face Validity and Worthy of Investigation (2)**	**Reject Outright (3)**
Biofuels	100%	0%	0%
Project Manager	31%	69%	0%
Engineering Systems	68%	23%	9%
Wave Energy	24%	28%	48%
Materials Science/Solar	92%	8%	0%
Electrical Power	79%	17%	4%

Table 4 *Categorization of candidate changes in the renewable energy domain*

Table 4 covers the overall categorization of candidate results by the various interviewees. This is similar to the table produced in the Cloud Computing study, where each row represents the categorization for a distinct interviewee. The main result shown in Table 4 is that each interviewee had a different spread of results between the categories. The biofuels expert put all of the candidate changes into category 1, whereas the project manager and the wave energy expert put less than a third of the changes they reviewed into category 1. The biofuels expert only reviewed results in the biofuels subsystem, and indicated that while terminology choice was worthy of investigation, the underlying links between components were all as expected and it was hard to distinguish when a link in the subsystem was warranted because the goal of the subsystem was not clear [Hashaikeh 2010]. The project manager indicated that he placed many of the links in the worthy of investigation category because he did not have the technical expertise to identify many links as category 1 [Prieto 2010]. The wave energy expert indicated that he placed many links in category 3 which were not relevant to his field, as his field is new and focused on technical growth as opposed to the various socio-economic and consumption factors in other subsystems [Tabaei 2010]. The engineering systems, materials science and electrical power experts all fell closer to the expected distribution of results across the categories.

In summary, these quantitative results reinforce the notion that different roles have different perspectives on the utility of this research. It is good that none of the interviewees had a blanket rejection of the research, and all of them found some candidate changes which were worthy of investigation. Half of the interviewees responded with categorization that was in the expected distribution. The qualitative interviews demonstrated that every interviewee had perspectives to share on how they would use this research.

Qualitative feedback on candidate changes

The following sections describe the qualitative feedback received from the academic experts interviewed in the renewable energy case study.

Overall insight on renewable energy domain

The interviews highlighted the fact that each of the different subfields in renewable energy all have a common set of goals around driving adoption of their technology and effective distribution of energy. Thus, even though the technology choices are different, technical patterns can translate from one field to another, and non-technical patterns almost always will translate [Hashaikeh 2010].

The field of renewable energy is also a heavily political field and thus has a number of non-technical actors whose views on technical choices often have funding consequences. Reconciling term usage is one way in which communication between technical and non-technical actors can be bridged [Prieto 2010].

The final observation is that different fields are at different stages of the maturity lifecycle in terms of adoption. For example, wave energy was cited as being very early in the lifecycle as compared to solar energy. However, it is certainly possible for fields to learn from each other as they progress. For example, early investment in electric and hybrid cars has been in small vehicles rather than large SUVs, primarily due to feasibility of powering a large vehicle with electric power. The field may learn from whether adoption is best driven from this market or perhaps could benefit from focusing on SUVs where gasoline usage is highest. Wave energy is similarly focused on oceans, due to feasibility, but may find that looking at rivers may have more impact on consumers [Tabaei 2010].

Term Usage

The use of certain terms and the relative popularity of certain terms came up as a key insight which this research could be used to obtain. For example, in biofuels, the prevalence of the term ethanol as a proxy for the overall field of biofuels was obvious from the results [Hashaikeh 2010]. The linkages between the term ethanol and other terms in the subsystem provides a means for understanding whether researchers should use ethanol as an explanatory term when presenting to non-technical individuals.

In other fields, the same term may have different meanings, and seeing the terms it is linked to help explain the various uses of the same term. This is especially useful in new and changing fields such as renewable energy. In wave energy, the term "security" can have different meanings, it could refer to the security of the energy production (wave energy plants are often on islands), or it could refer to the security of the distribution network [Tabaei 2010]. In distributed generation, the scale factor of terms is an issue, similar to what was reported in the Kenya broadband study. A "small" hydroelectric plant can have a different meaning depending on the region, or can change over time as capabilities increase [Zeineldin 2010].

Finally, understanding term replacement was mentioned by multiple interviewees as a possible use of this research. As the field changes, certain terms will become more popular the others for referring to the same concept. For example, the term biofuel may indeed replace ethanol once public understanding and perception matches scientific understanding. Understanding the changing use of terms, and possibly even mapping it to the use of the same terms in technical papers is another possible application of this research.

Preparing for funding and regulation by understanding public opinion

One of the main reasons that the researchers found interest in understanding the public's use of terms is because the funding sources they utilize are often more in line with the public's understanding of the field. In the ethanol example, funding proposals may benefit from including or avoiding the use of the term "ethanol" depending on public perception of that concept [Hashaikeh 2010]. The same technical research could be proposed with a set of terms that has the most positive connotation at the time. This is very similar to the feedback received in the cloud computing case study around finding the right terms to market the same technology.

Another challenge faced by researchers is the need to explain their work to non-technical audiences, either in the form of introduction sections of papers, or at talks given to more general audiences. With a changing field, even when researchers attend conferences in their own discipline, the topics related to renewable energy may not be familiar and an understanding of public opinion could help attract a broader audience [Tabaei 2010].

There is also a phase delay in public understanding versus technical progress [Emziane 2010]. Often times it takes months or years before the public is familiar and comfortable with topics which have been identified by scientists. This is an interesting result because it is directly opposite the finding in the cloud computing study that the public interest in a topic was seen ahead of when the industry experts delivered products to meet the interest. Understanding the duration of this phase delay could also help focus effort on shortening it. As stated by one of the researchers, "we live in society, and as scientists, researchers, engineers, we have more duty to contribute to the better understanding of the reality of things" [Emziane 2010].

Regional Differences

Regional differences in terms of the requirements for renewable energy as well as the perception of different elements of the field are another potential insight to be gained from this research. The project manager who prepares funding proposals cites the need to observe how certain regions view certain technologies. This study does not explicitly study blogs in a particular region, however the Kenya broadband study did focus on blogs which mention Kenya. Future study could scope blogs to particular regions to study differences between regions. In cases where a major event such as an oil spill adversely affects a region, it may cause a regional difference in how certain technology is perceived in that region [Prieto 2010].

Distributed generation is a term which has synonyms or related terms in various regions, such as "embedded generation", "distributed resources", and "dispersed generation" [Zeineldin 2010]. This is true even though there are IEEE standards for electrical power distribution which mention certain specific terms. There is a desire to capture all of the regional differences as well as track the changing use of terms as regional differences decrease over time.

Regional differences also impact the requirements for certain forms of energy. For example, solar energy is highly dependent on the availability of sunlight in the region, the temperature, and physical space available for solar energy collection and storage materials [Chiesa 2010]. Regional consumption of energy is also different, as certain regions are more urban,and may have more load at certain times. These differences also translate to differences in funding availability and investment.

Filtering Blogs to focus on what to read

When researchers were asked if they already consult blogs as an information source in their regular work, the main comment raised by each was the desire to filter blogs and prepare a subset that were worth actually reading. Even if certain blogs written by respected experts are already selected, it would be good to use this research to identify new blog topics or authors who may be worthy of reading [Sgouridis 2010]. One method of selecting blogs which are likely written by experts is to search for blogs which contain a sufficient number of technical terms. This method only works if the technical terms are not well known by the public and included in all blog posts in the area. One of the benefits of renewable energy being a new field is that the general public has not yet developed a familiarity with all of the technical terms involved [Emziane 2010].

Deeper understanding of technical areas

Understanding common patterns at a high level

The topics covered in this case study were generally at a high level and provided some semblance of patterns between different technical areas. For example, if biofuels and smart grids are considered peer areas, this case study provides some insight on the general discussion of each field and can help identify common patterns such as public understanding of the field or adoption factors related to the field

[Hashaikeh 2010]. It would be more effective, according to two interviewees, to do this level of analysis of a broader set of approximately 10 peer areas, and then use the results to both show how lessons learnt from one field can be applied to the other, as well as how scientific work in one area may be explained in the context of another [Tabaei 2010] [Hashaikeh 2010]. For example, carbon capture is a major area of scientific progress related to biofuels, but if the solar energy field demonstrates an interest, then the same research may be applied to solar cells which capture sunlight.

Understanding deeper technical insights at a low level

Interviewees consistently requested deeper technical coverage of their individual domains as a follow up analysis to the general topics covered in this case study. In certain cases, interviewees suggested drilling down on the specific blogs or specific topics around areas which seemed counterintuitive but still had some face validity [Chiesa 2010]. For example, if a link is suggested which could be plausible but does not seem obvious, then doing an analysis of deeper technical terms in the same area may help make the decision on whether to accept it. One direction to consider here is to use term discovery, which will allow new related terms to be identified without a human guidance. This would allow the many terms which are related to a general topic such as biofuels to be studied.

Growing beyond technical research

Transition from R&D to Commercial Products

All of the interviewees in this case study were in academia but many have an interest in seeing their work translated from research and development to commercialization. To accomplish this, the evolution of certain fields which have seen market success may influence the choices made by other fields. Understanding past behavior may provide a means for predicting the future behavior of another field and driving adoption or regulation [Tabaei 2010]. Further work on this case study could include interviews with researchers and managers from industry who are working on similar areas of renewable energy. Experts from industry may have insights into adoption and product development which could inform the academic researchers interviewed in this study.

Common Drivers between technical and non-technical subsystems

One of the ways to best understand what factors drive success of a certain technical area such as biofuels is to develop a better understanding of common drivers, or components which drive behavior from the technical systems to the non-technical systems. In the field of biofuels, it is interesting to see what factors drive connections between technical aspects to topics such as job creation [Hashaikeh 2010]. For example, blogs in the area of renewable energy which mention job creation also mention hydroelectric plant. This may indicate a public perception that physical plants have the need for employees and may drive job creation. The reality may in fact be that other areas of technical investment have a greater impact on job creation. However, in this example, the perception of what drives job creation may be just as useful as the actual factors which drive job creation, because it helps focus communication to the public as well as regulators who may be influenced by the public.

Process adaptation for the renewable energy case study

In this case, the main process change which was made from prior cases was the simplification in generating the initial CLIOS system representation. Instead of producing a full system representation, the focus was on which terms would be sufficient to illustrate the use of this research. This process adaptation also raised the potential of using this research to do the initial generation of the CLIOS representation for a system. Future study could pursue this further, but would have to include some form of term generation to generate terms which are not initially identified. The other major process point raised in this case study is the time spent understanding the work background of each professor to be interviewed. The system studied in this case was much broader and had a less focused goal than the other two systems. Spending time with the interview subjects allowed the formal interviews to be more focused around potential uses of the research to their work. This suggests that the actual deployment of this research in a real world setting would require some level of service to make it useful to stakeholders in a given domain.

Summary of Renewable Energy case study

This section provided the background and results for the renewable energy case study. This case had the unique aspect of focusing entirely on an academic audience and on the broad area of renewable energy. The actual results of the bibliometric blog analysis were in line with those produced in the other two case studies. The results from the quantitative interviews showed that certain interviewees categorized the results in the expected distribution – with most results marked as already known and a good number of results having face validity – but other interviewees categorized the results in a much different distribution. The feedback provided in the qualitative portion of the interviews had some overlap with the other two case studies – term usage, regional differences, and using this research to justify known insights were examples of overlapping feedback. This case study also produced some unique feedback in the ability of this research to potentially map progress in certain fields to other fields, and in the ability to identify common drivers between technical and non-technical subsystems.

Evaluation of Taxonomy Generation Algorithms

Part of the objective of the latest reporting period has been the evaluation of the technology forecasting and visualization algorithms. Achieving this required a certain degree of reflection on the properties of the taxonomy generation process.

We start with the assumption that for a given domain of research, there exists a knowledge "landscape" representing the various concepts, relationships and sub-domains which constitute the domain. Further, given a set of related terms, it should be possible to capture the interrelationships between the above-mentioned concepts in terms of the relationships between these terms. Another objective is a creation of a compact representation of these inter-relationships, for e.g. in the form of a keyword taxonomy. However, while there is only one true landscape, this is likely to be a high-dimensional object and hence, projecting this into a lower dimensional representation such as a taxonomy is unlikely to be a unique process; we can certainly envision that, for any given research domain and the corresponding research landscape, it would be possible to generate multiple keyword-based visualizations, each of which represents a different perspective of the underlying landscape. Nevertheless, in the context of these experiments, we will largely bracket these concerns, and for the sake of brevity will often refer to a single unique taxonomy. This is because we do not feel that the slight inaccuracy introduced by this label would affect the validity of the findings of this study. It is sufficient to be aware that, depending on the specific context, it might actually be the underlying landscape, or the closest matching taxonomy, that is in fact referred to in such situations.

Figure 20 illustrates the underlying model. We believe that every research area consists of several concepts, which are interrelated in some way. However, this underlying structure is not observable but is reflected in the documents and articles produced by researchers in this area (box C). In turn, these documents and articles are collected and subsequently accessed and analyzed to produce the desired bibliometric indicators (box A). Finally, taxonomy generation algorithms are used to analyze this information and to organize them in the form of a taxonomy that reflects the relationships between these terms (box B).

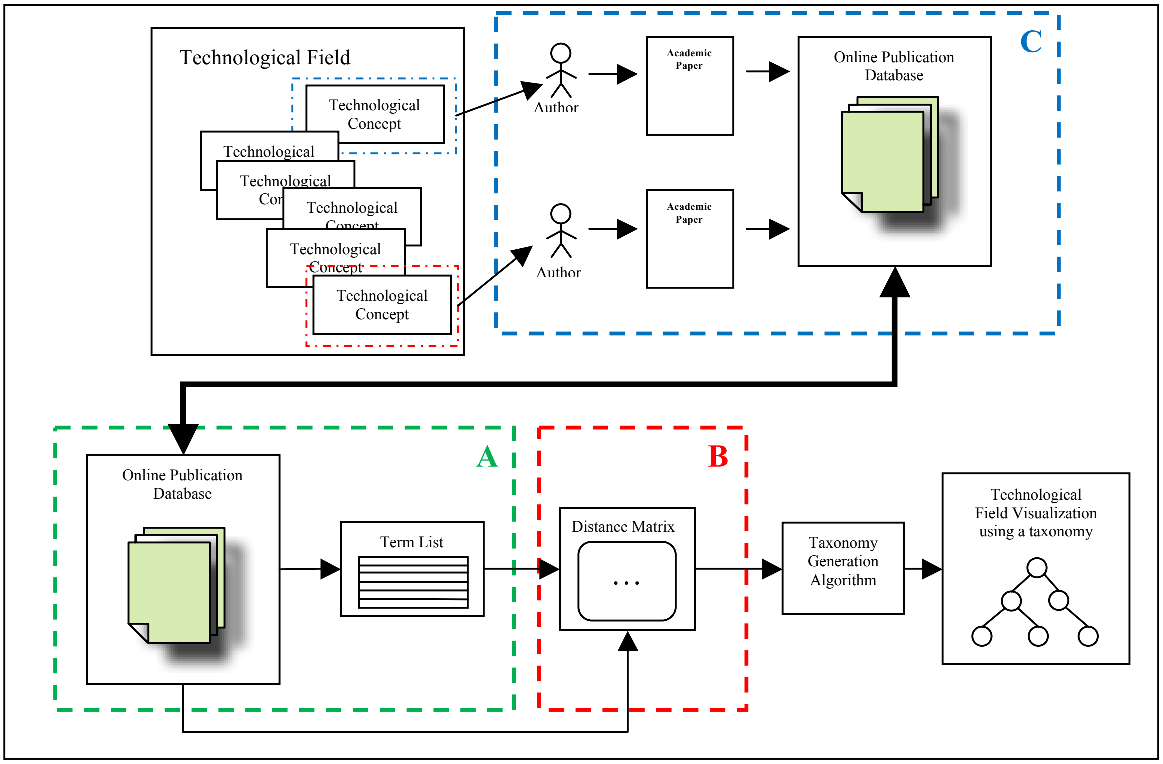


Figure 20 The Model Underlying the Taxonomy Generation Process

In practically all achievable data collection scenarios, the bibliometric information thus gathered will be imperfect for two main reasons: firstly, there could be errors and biases in the documents as well as non-uniform coverage of the underlying area. Secondly, it is almost impossible to collect or analyze all relevant bibliometric data as this is always subject to the quality of the specific database used.

A further concern is that, even if we were to assume complete data, inferring the underlying taxonomy remains a difficult challenge, and is an instance of an inverse problem. Solving these problems require the careful use of effective inference algorithms.

In short, Figure 20 depicts the chain of events that must occur in order for us to obtain information regarding the underlying knowledge domain. Noise or imperfection in any of these events will propagate down the chain and will subsequently affect the quality of our results. However, this formulation also provides us with a very elegant framework in which experiments can be designed to test the taxonomy generation process. These will be described in the following section.

Taxonomy Evaluation Criteria

Evaluating taxonomy generation algorithms is a difficult task, and there do not appear to be any relevant techniques in the literature. However, as mentioned, the model described in the previous sub-section provides a very nice framework for the design of a set of test procedures. By studying this model carefully, we posit that an effective taxonomy-generation algorithm must be one that has the following three characteristics:

1. It must produce consistent taxonomies despite slight perturbations to its backend, or slight changes to the terms in the taxonomy. This is necessary given the issue of imperfect information mentioned in the previous section.
2. It must conform well to the pairwise-relationship-strength matrix (distance matrix) which it is based on; putting this another way, an effective algorithm should maximize the overall similarity of terms in the taxonomy. This is necessary because even if perfect data were to be available, there is still the issue of solving the "inverse problem" – i.e. the identification of the taxonomy most likely to have produced the observed term statistics.
3. It must produce taxonomies that are valid representations of relationships between terms in the underlying landscape. This is necessary because in the end, the deliverable for our team's project is a taxonomy that accurately represents the research landscape. Even if the first two characteristics for a good taxonomy are met, if the final output is a taxonomy that intuitively does not make sense, all our work is invalidated.

To test conformity to these three conditions, the following three tests were formulated:

T1: The consistency of each algorithm is evaluated by attempting to vary either the backend data set or the term list used in the taxonomy. Referring back to Figure 3, this can be seen as trying to perturb box A and seeing its effect on the generated taxonomies.

T2: Taxonomies produced using each algorithm are scored based on conformity to the distance matrix. Referring back to Figure 3, this can be seen as measuring how well the taxonomy generation algorithm can encapsulate the information in box B.

T3: Synthetic data based on a predefined underlying model is generated and compared against the taxonomy generation's output to evaluate each algorithm's effectiveness. Referring back to Figure 3, this can be seen as creating our own documents / publication database, much like box C.

Test 1: Evaluating Algorithm Consistency

We believe that a taxonomy generation algorithm must be *consistent*, where consistency refers to the ability of a taxonomy generation algorithms to produce similarly structured taxonomies in spite of minor variations to its inputs. This is an important requirement since the underlying taxonomy clearly does not change even if only different subsets of the related literature are available at any one time.

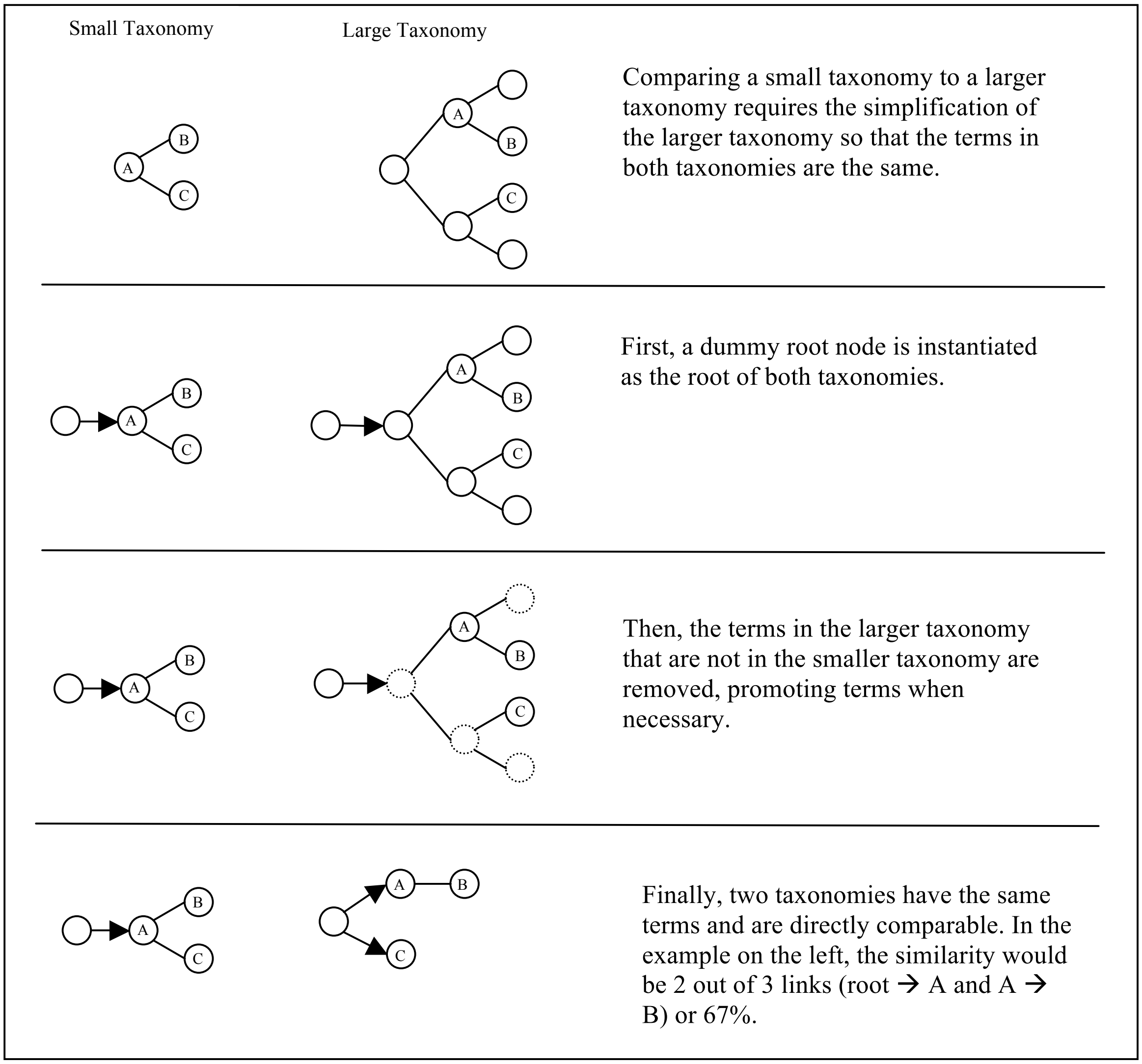
Consistencies against two forms of variation are tested for. Firstly, we test consistency against variations in the size and coverage of the bibliometric data sets that form the basis of every taxonomy generation algorithm. We believe that every good taxonomy generation algorithm must consistently produce the same taxonomies despite slight variations to its backend data set. As mentioned in the previous subsection, we cannot assume that the data contained the backend bibliometric data set is perfect. Thus, a good taxonomy generation algorithm needs to produce similar-looking taxonomies even when the backend data set is altered slightly or is incomplete.

To simulate these effects, modified data sets were created by taking random subsets of the original database. Next, taxonomies were generated for each of these data sets (using the same term list), the structures of which were compared to the taxonomy generated using the original database. The "quality" of the corresponding algorithms is then evaluated based on the degree to which the links within these taxonomies were preserved. For example, if the terms "wind energy" and "turbines" are linked in the original, they should also be linked in a taxonomy produced using only a subset of the backend data set.

The second set of consistency tests were done by varying the terms used. Taxonomies were produced using the different taxonomy generation algorithms but the terms used for each taxonomy were varied while keeping the same backend data set. Specifically, each algorithm was run with a fixed term list using the entire bibliometric data set as backend. Then, some additional terms were added to the term list, to simulate the "noise" that could be introduced in the taxonomy generation process and the entire algorithm was rerun. The outputted taxonomies in both runs were then compared to each other. The taxonomies must, as much as possible, contain the same relative term relationships.

For example, in a "renewable energy" related taxonomy, if the terms "wind energy" and "turbines" are linked directly, adding a few "noise" terms to the taxonomy and re-running the algorithm should still produce a taxonomy where "wind energy" and "turbines" are linked together, although not necessarily directly, provided that the taxonomies were generated based on the same backend data set.

For the term list consistency test, a recurring test involved comparing a small taxonomy to a larger taxonomy that contained a superset of the smaller taxonomy's terms. To do this, the larger taxonomy needed to be simplified so that its links can be directly comparable to the smaller taxonomy. This was accomplished by first creating a new root node, and instantiating it as the root of both taxonomies. Then, the terms in the larger taxonomy were scanned and the terms that did not exist in the smaller taxonomy were removed, promoting the children of the removed terms as children of existing terms. Figure 21 illustrates the process.



Comparing a small taxonomy to a larger taxonomy requires the simplification of the larger taxonomy so that the terms in both taxonomies are the same.

First, a dummy root node is instantiated as the root of both taxonomies.

Then, the terms in the larger taxonomy that are not in the smaller taxonomy are removed, promoting terms when necessary.

Finally, two taxonomies have the same terms and are directly comparable. In the example on the left, the similarity would be 2 out of 3 links (root → A and A → B) or 67%.

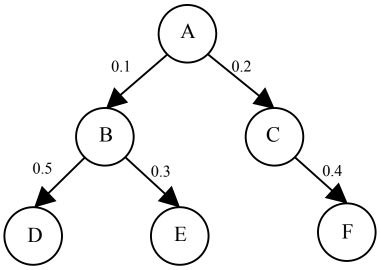
Figure 21 Simplifying a larger taxonomy

Test 2: The Inverse Problem

After checking the consistency of each taxonomy generation algorithm, the next step was to verify the taxonomies produced by each algorithm with respect to the distance matrix on which it was based. To compare the different algorithms, taxonomies were generated using the same term list and backend data set then their conformities to the distance matrix could be evaluated. However, evaluating this quality directly is not straightforward since it is unclear how the branch structure of the taxonomy can be reliably converted into numerical distances between nodes. Instead, we use an approach based on the principle of parsimony: if indeed a taxonomy captures the information embedded in a distance matrix (and, implicitly, the structure of the underlying research landscape), it should be "efficient" in terms of the distances associated with the branches in the taxonomy. i.e. we would expect terms which are highly similar to be grouped together while terms which are unrelated would logically be placed in different regions of the taxonomy. Based on this intuition, we can construct an appropriate scoring metric by first collecting the edge weights contained in the taxonomy, then aggregating these in a weighted fashion to obtain an overall measure of "goodness". However, even this can be done in a number of ways and as it was unclear which was the optimal one, a few reasonable alternatives were attempted. This resulted in the following scoring schemes:

- ***Average -*** This scheme calculates the score by taking the mean of all the direct edge weights in the taxonomy.
- ***Momentum -*** This scheme calculates the score by taking the sum of the means of each node's outgoing edges normalized by its incoming edge. A term node's incoming edge is the edge coming from its parent and its outgoing edges are the edges leading to its direct descendants.
- ***Mean to Root -*** This scheme calculates the score by taking the sum of the means of each term node's edge weights to all its ancestors.
- ***Mean to Grandparent -*** This scheme calculates the score by taking the sum of the means of each term node's edge weights up to two levels above (to its grandparent) node.
- ***Linear -*** This scheme calculates the score by taking the sum of each term node's normalized linearly weighted distance to all its ancestors.
- ***Exponential -*** This scheme calculates the score by taking the sum of each term node's normalized exponentially weighted distance to all its ancestors.

Figure 22 shows the process of scoring a taxonomy using the different weighting schemes.



Average: $\frac{(0.1 + 0.2 + 0.5 + 0.3 + 0.4)}{5} = \mathbf{0.3}$

Momentum: $(A_M + B_M + C_M + D_M + E_M + F_M)$

$= \frac{(0.1+0.2)}{2} + \frac{\left(\frac{(0.5+0.1)}{2} + \frac{(0.3+0.1)}{2}\right)}{2} + \frac{(0.4+ 0.2)}{2} + 0 + 0 + 0 = \mathbf{0.7}$

Mean to Root: $(A_{MTR} + B_{MTR} + C_{MTR} + D_{MTR} + E_{MTR} + F_{MTR})$

$= 0 + 0.1 + 0.2 + \frac{(0.5 + 0.1)}{2} + \frac{(0.3 + 0.1)}{2} + \frac{(0.4 + 0.2)}{2} = \mathbf{1.1}$

Mean to Grandparent: $(A_{MTG} + B_{MTG} + C_{MTG} + D_{MTG} + E_{MTG} + F_{MTG})$

$= 0 + 0.1 + 0.2 + \frac{(0.5 + 0.1)}{2} + \frac{(0.3 + 0.1)}{2} + \frac{(0.4 + 0.2)}{2} = \mathbf{1.1}$

Linear: $(A_L + B_L + C_L + D_L + E_L + F_L)$

$= 0 + 0.1 * 2/3 + 0.2 * 1/3 + (0.5 * 2/3 + 0.1 * 1/3) + (0.3 * 2/3 + 0.1 * 1/3) + (0.4 * 2/3 + 0.2 * 1/3)$

$= \mathbf{1.067}$

Exponential (with exponent 0.5) : $(A_E + B_E + C_E + D_E + E_E + F_E)$

$= 0 + 0.1 * 2/3 + 0.2 * 1/3 + (0.5 * 2/3 + 0.1 * 1/3) + (0.3 * 2/3 + 0.1 * 1/3) + (0.4 * 2/3 + 0.2 * 1/3)$

$= \mathbf{1.067}$

Figure 22 Using Scoring Metrics to Score a Taxonomy

Test 3: Analyzing Synthetic Data

As mentioned, the final set of tests involved testing the algorithms on a synthetic dataset. As will be described, the mechanism for generating this data was designed to simulate a typical database gathered from an online publication index, but for which we have perfect knowledge of the underlying research landscape. The goal of this experiment is to test the ability of the taxonomy generation algorithms to approximate this landscape.

As there is no established method of generating synthetic data given the source taxonomy, a simple probabilistic mechanism was proposed that was based on the related techniques of LSA, PLSA and LDA. The basic principle behind these three methods is that a corpus of documents gives rise to a set of topics or concepts, where each topic or concept is composed not just of a single term, but rather is a weighted sum of terms. However, to facilitate the creation of synthetic data, each topic is linked to a single term, and each term has its own weighted distribution representing the degree to which all the other terms in the taxonomy affect it. In other words, whereas in LSA / PLSA / LDA each of the concepts are distinct from the terms they contain, for our purposes each term is also a concept.

To generate synthetic data, an arbitrary taxonomy is first randomly generated containing a fixed set of terms. Each term in the taxonomy is then assigned its own probability distribution, which is a set of probabilities for each of the terms in the taxonomy to occur in a document whose central term is that term. As a hypothetical example, in a "renewable energy" related taxonomy, a document whose central term is "solar photovoltaics" will most likely have a probability distribution where "solar photovoltaics" has the highest chance of occurring in the document, while related terms like "solar", "renewable energy" and "solar energy" should also have significant probabilities of occurring in the document.

For our purposes, the synthetic data was generated using a fixed set of terms where each term has a probability distribution set to one where each term gets the highest probability of occurring in a document pertaining to itself, its ancestor terms in the taxonomy also get a significant non-zero probability of occurring, and the rest of the nodes get a small non-zero probability of occurring.

Assigning a distribution to each term is a two-phase process. First, each term is given an initial distribution where it is assigned a high probability p and the (1-p) probability is split among the rest of the terms in the distribution. Second, the individual distributions are aggregated by adding each term's distribution to that of its parent and normalizing so that the resulting distribution sums to one.

The following figure illustrates this process:

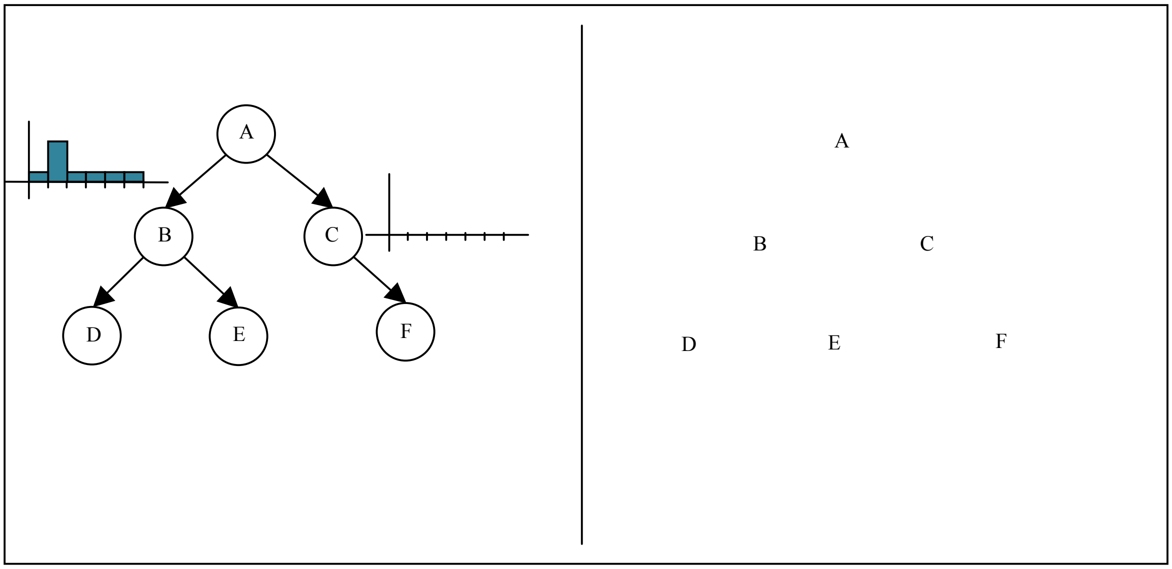


Figure 23 Assigning Probability Distributions for Each of the Terms in a Taxonomy

After each term's distribution is finalized, a set of documents is generated for each term based on the probability distributions for each term. For instance, in Figure 23 (rhs), a document generated for E will have a large probability that terms E, B, and A will be included as terms in the document. Since terms C, D and F only have a small probability within term E's distribution, the chance of them being included in a document relating to term E is slim. For a document generated relating to E, terms C, D and F are the *noise terms.* The probability of these terms being included can be increased by increasing the overall *noise* within the system. We define noise as the total probability associated to the non-related terms. For instance, if the noise was 0, then only terms E, B and A will ever occur in a document relating to term E, however if the noise was 1, then each of terms A, B, C, D, E and F will have an equal probability of occurring in a document relating to term E.

An equal amount of documents relating to each term are generated, and the collection of documents generated is used as the backend data set for the taxonomy generation algorithms. Essentially, this collection of documents is a simulated version of the expected collection from a real publication database.

The beauty of this process is that there is a predetermined underlying taxonomy for the set of documents generated, which is directly comparable to the taxonomies generated using the taxonomy generation algorithms developed in this project. The following figure illustrates a simple example of generating synthetic data:

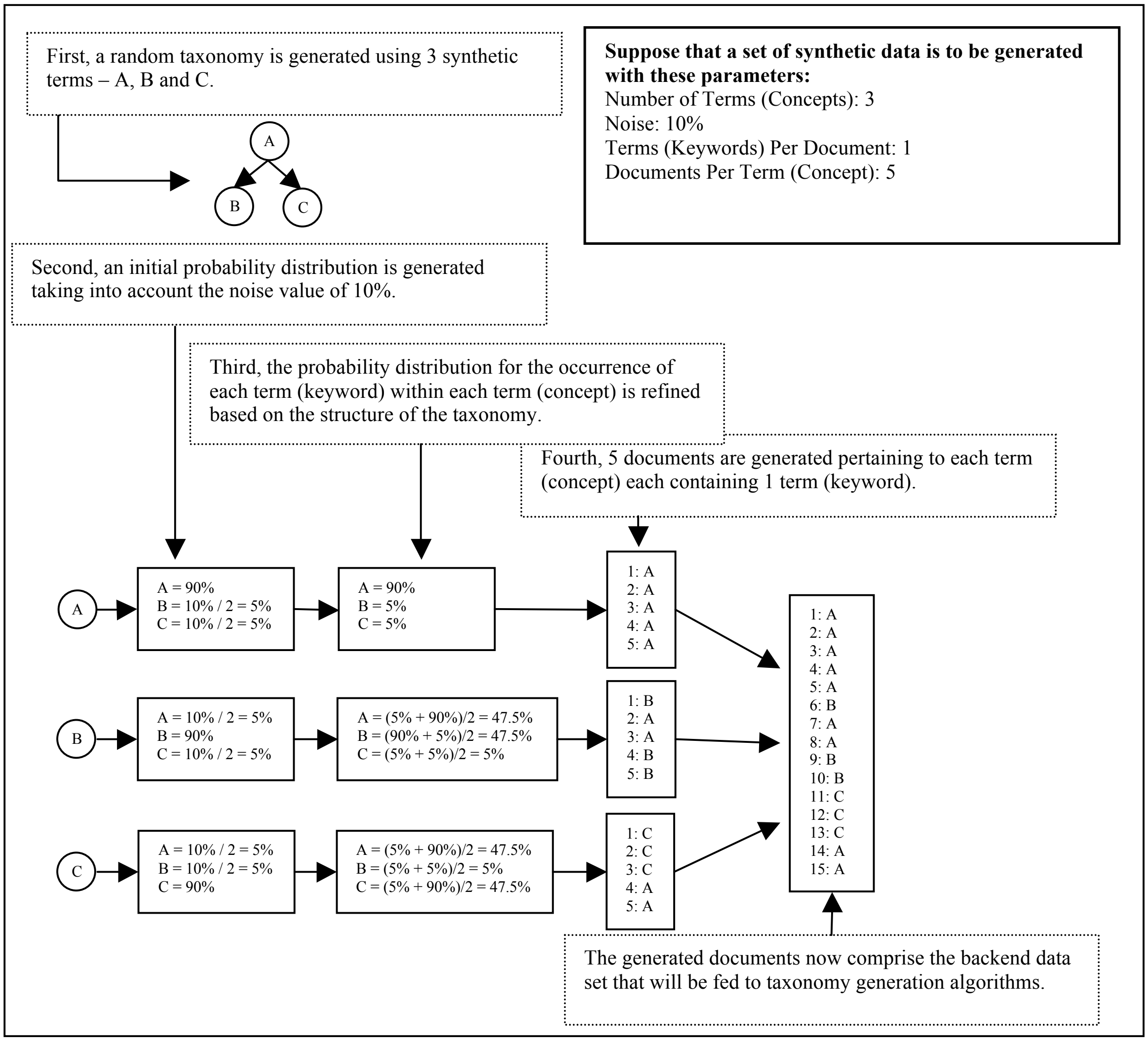

First, a random taxonomy is generated using 3 synthetic terms – A, B and C.
A
B
C
Suppose that a set of synthetic data is to be generated with these parameters:
Number of Terms (Concepts): 3
Noise: 10%
Terms (Keywords) Per Document: 1
Documents Per Term (Concept): 5
Second, an initial probability distribution is generated taking into account the noise value of 10%.
Third, the probability distribution for the occurrence of each term (keyword) within each term (concept) is refined based on the structure of the taxonomy.
Fourth, 5 documents are generated pertaining to each term (concept) each containing 1 term (keyword).
A
A = 90%
B = 10% / 2 = 5%
C = 10% / 2 = 5%
A = 90%
B = 5%
C = 5%
1: A
2: A
3: A
4: A
5: A
B
A = 10% / 2 = 5%
B = 90%
C = 10% / 2 = 5%
A = (5% + 90%)/2 = 47.5%
B = (90% + 5%)/2 = 47.5%
C = (5% + 5%)/2 = 5%
1: B
2: A
3: A
4: B
5: B
C
A = 10% / 2 = 5%
B = 10% / 2 = 5%
C = 90%
A = (5% + 90%)/2 = 47.5%
B = (5% + 5%)/2 = 5%
C = (5% + 90%)/2 = 47.5%
1: C
2: C
3: C
4: A
5: A
1: A
2: A
3: A
4: A
5: A
6: B
7: A
8: A
9: B
10: B
11: C
12: C
13: C
14: A
15: A
The generated documents now comprise the backend data set that will be fed to taxonomy generation algorithms.

Figure 24 Synthetic Data Generation

This section explained the methodology used to analyze the taxonomy generation algorithms we've developed. The next section will discuss the results of the tests we conducted and give recommendations regarding which taxonomy generation algorithm(s) work best.

Results

As described previously, all the tests were run using a backend data set collected from Scopus. The data set consisted of 153,537 terms with 2,326 terms occurring more than 100 times among the entries, and 201 terms occurring more than 1,000 times.

Four main taxonomy generation algorithms were analyzed (the details of these algorithms will not be described here but have been documented in a technical report [Camina, 2010]). The four algorithms were:

1. Dijkstra-Jarnik-Prim's Algorithm
2. Kruskal's Algorithm
3. Edmond's Algorithm
4. Heymann's Algorithm

Each algorithm has several parameters which can be varied, resulting in a number of variants. Different variants could sometimes produce significantly different results and as such, needed to be treated as separate algorithms. Table 5 summarizes the variants tested along with the corresponding settings. For convenience, each algorithm was given an acronym by which it will be referred to for the rest of this section.

Table 5 *List of Taxonomy Generation Variants*

Algorithm Variant Acronym	**Algorithm Type**	**Similarity Metric (used to create the distance matrix)**	**Centrality Metric Used (either to choose root or to decide term centrality at each iteration)**
D-CB	DJP	Cosine	Betweenness
D-CC	DJP	Cosine	Closeness
D-SB	DJP	Symmetric NGD	Betweenness
D-SC	DJP	Symmetric NGD	Closeness
K-CB	Kruskals	Cosine	Betweenness
K-CC	Kruskals	Cosine	Closeness
K-SB	Kruskals	Symmetric NGD	Betweenness
K-SC	Kruskals	Symmetric NGD	Closeness
E-AB	Edmonds	Asymmetric NGD	Betweenness
E-AC	Edmonds	Asymmetric NGD	Closeness
H-AB	Heymann	Asymmetric NGD	Betweenness
H-AC	Heymann	Asymmetric NGD	Closeness
H-CB	Heymann	Cosine	Betweenness
H-CC	Heymann	Cosine	Closeness
H-SB	Heymann	Symmetric NGD	Betweenness
H-SC	Heymann	Symmetric NGD	Closeness

In general, there was not a lot of variation in terms of computational requirements between these variants and as such the run-times or relative speeds of the variants are not emphasized in this report.

Results on Test 1 (Consistency)

The first set of tests was aimed at evaluating the consistency of the taxonomy generation algorithms. To do this, two sets of experiments were conducted to gauge the consistency, in terms of robustness against noise, of the different taxonomy generation algorithms. The first set of experiments measured the consistency of the algorithms when faced with slight perturbations in the backend, while in the second set of experiments, the backend database was fixed and perturbations were introduced to the collection of terms used to form the taxonomy.

Backend Data Set Consistency

For this test, the 153,537-entry bibliometric data set is randomly divided into five separate 100,000-entry subsets. The most popular terms from the entire Scopus "renewable energy" bibliometric data set were then taken and each of the taxonomy generation algorithms were run, keeping constant the term list and varying the backend data set between the five 100,000-entry sets. The percentage similarity of direct links between each of the taxonomies generated was then calculated between each of the 100,000-entry-backend data-set taxonomies and the entire 153,537-entry-backend data-set taxonomy. Table 6 summarizes the mean of the percentage similarities for each algorithm variant.

Table 6 *Backend Data Set Consistency Test Results*

Algorithm Variant Acronym	**25 most frequently occurring terms used as term list**	**50 most frequently occurring terms used as term list**	**100 most frequently occurring terms used as term list**	**200 most frequently occurring terms used as term list**	**500 most frequently occurring terms used as term list**	**Mean of Percentage Similarities**
D-CB	77.60%	98.00%	97.80%	95.80%	94.08%	92.66%
D-CC	94.40%	97.60%	97.60%	95.80%	94.08%	**95.90%**
D-SB	92.80%	96.80%	94.40%	91.70%	91.00%	93.34%
D-SC	93.60%	96.80%	94.40%	91.10%	91.00%	93.38%
K-CB	77.60%	98.00%	97.80%	95.80%	93.88%	92.62%
K-CC	94.40%	97.60%	97.60%	95.80%	93.88%	**95.86%**
K-SB	4.00%	2.00%	1.00%	0.50%	0.20%	1.54%
K-SC	4.00%	2.00%	1.00%	0.30%	0.20%	1.50%
E-AB	90.40%	93.20%	90.60%	88.90%	84.64%	89.55%
E-AC	93.60%	93.20%	90.60%	88.90%	84.52%	90.16%
H-AB	88.00%	92.80%	96.20%	96.30%	97.48%	94.16%
H-AC	95.20%	**98.40%**	98.00%	**97.50%**	**97.68%**	**97.36%**
H-CB	34.40%	36.80%	29.00%	33.30%	29.92%	32.68%
H-CC	96.00%	97.60%	**98.60%**	95.90%	94.96%	**96.61%**
H-SB	78.40%	73.20%	78.60%	83.90%	82.96%	79.41%
H-SC	**96.80%**	96.40%	94.60%	93.80%	91.56%	94.63%

Highlighted in the table above are the top performers for each test run. Based on these results, the best performing algorithm variants (over 95% similarity on average) are:

1. Heymann algorithm, asymmetric NGD metric, closeness centrality (H-AC)

2. DJP algorithm, cosine similarity, closeness centrality for root selection (D-CC)

3. Kruskals algorithm, cosine similarity, closeness centrality for root selection (K-CC)

4. Heymann algorithm, cosine similarity, closeness centrality (H-CC)

Other notable observations are:

1. The use of Kruskals algorithm with symmetric NGD similarity is not a consistent algorithm at all. It was barely able to create a single consistent link between the taxonomies generated using the 100,000-entry-backends and the 153,537-entry-backend.
2. The tests for the Heymann algorithm all show that the use of closeness centrality is a much more consistent metric than using betweenness centrality. Note that the differences between closeness and betweenness centrality are only evident when examining the results of the Heymann algorithm tests because Heymann is the only taxonomy generation algorithm that uses the centrality measures for more than just picking the root node.

Term Consistency

For this test, the backend was kept constant, and consisted of the entire 153,537-entry Scopus "renewable energy" bibliometric data set. However, the term lists were varied by taking the most popular terms in the data set and inserting "noise" terms, which are terms selected randomly from the remaining terms in the data set. We chose to insert an equal number of noise terms to the terms already in the taxonomy. For instance, if a taxonomy was created using the 25 most frequently occurring terms, 25 noise terms were inserted into the taxonomy, then each taxonomy generation algorithm was run using those 50 total terms, and percentage of the number of links consistent in the 25-term noise-free and 50-term noisy taxonomies produced by each algorithm was calculated. Comparing the two taxonomies required simplifying the larger 50-term taxonomy using the method described in the previous subsection. The tests were repeated three times and the mean similarity for each algorithm was taken as the representative score. The results are summarized in Table 7.

Table 7 *Term Consistency Test Results*

Algorithm Variant Acronym	**25 most frequently occurring terms, with 25 more noise terms**	**50 most frequently occurring terms, with 50 more noise terms**	**100 most frequently occurring terms, with 100 more noise terms**	**250 most frequently occurring terms, with 250 more noise terms**	**Mean of Percentage Similarities**
D-CB	76.92%	**97.39%**	87.79%	86.06%	87.04%

D-CC	92.31%	**97.39%**	87.79%	86.06%	**90.88%**
D-SB	**94.87%**	88.24%	82.51%	77.69%	85.83%
D-SC	87.18%	94.12%	84.82%	81.01%	86.78%
K-CB	76.92%	**97.39%**	87.79%	86.06%	87.04%
K-CC	92.31%	**97.39%**	87.79%	86.06%	**90.88%**
K-SB	7.69%	31.37%	0.66%	0.00%	9.93%
K-SC	0.00%	35.29%	1.32%	0.27%	9.22%
E-AB	75.64%	95.42%	83.83%	82.20%	84.27%
E-AC	80.77%	95.42%	79.87%	80.88%	84.23%
H-AB	64.10%	86.93%	86.14%	86.59%	80.94%
H-AC	83.33%	94.77%	83.83%	87.38%	87.33%
H-CB	23.08%	37.25%	27.39%	33.33%	30.26%
H-CC	91.03%	94.12%	**89.44%**	**88.58%**	**90.79%**
H-SB	50.00%	54.90%	60.07%	55.78%	55.19%
H-SC	75.64%	72.55%	78.22%	76.49%	75.73%

Highlighted in the table above are the top performers for each test run. The best performing algorithm variants (over 90% similarity) based on our tests are:

1. DJP algorithm, cosine similarity, closeness centrality for selecting the root (D-CC)
2. Kruskals algorithm, cosine similarity, closeness centrality for selecting the root (K-CC)
3. Heymann algorithm, cosine similarity, closeness centrality (H-CC)

Other notable observations from this test are:

1. The use of Kruskals algorithm with symmetric NGD similarity is not a consistent algorithm at all. It was barely able to create a single consistent link when noise terms were inserted.
2. The tests for the Heymann algorithm all show that the use of closeness centrality is a much more consistent metric than using betweenness centrality.

Consistency Test Summary

The consistency tests were run both by varying the backend data set and the term lists to test for the taxonomy generation algorithms' robustness towards noise. Table 8 combines the information from Table 6 and Table 7 for easier viewing.

Table 8 *Consistency Test Summary*

Algorithm Variant Acronym	**Mean of Percentage Similarities for Backend Data Set Consistency Test**	**Mean of Percentage Similarities for Term List Consistency Test**
D-CB	92.66%	87.04%

D-CC	**95.90%**	**90.88%**
D-SB	93.34%	85.83%
D-SC	93.38%	86.78%
K-CB	92.62%	87.04%
K-CC	**95.86%**	**90.88%**
K-SB	1.54%	9.93%
K-SC	1.50%	9.22%
E-AB	89.55%	84.27%
E-AC	90.16%	84.23%
H-AB	94.16%	80.94%
H-AC	**97.36%**	87.33%
H-CB	32.68%	30.26%
H-CC	**96.61%**	**90.79%**
H-SB	79.41%	55.19%
H-SC	94.63%	75.73%

Based on the results shown in above, the following is clear:

- The use of Kruskals algorithm with symmetric NGD similarity is not a consistent algorithm in any way.
- Closeness centrality seems to be a much better similarity metric compared to betweenness centrality.
- The most consistent algorithms variants are D-CC, K-CC and H-CC, all of which use cosine similarity and closeness centrality to generate taxonomies.

Results on Test 2 (Conformity to Distance Matrices)

Several tests were run which tested each of the taxonomy generation algorithms' outputs individually by taking their outputs and scoring them using the different scoring metrics described in the previous subsection. To recap, the scoring metrics used were (for more information about each of the metrics mentioned above, see [Camina, 2010]):

1. Average
2. Momentum
3. Mean to Root
4. Mean to Grandparent
5. Linear
6. Exponential (0.5)

7. Exponential (0.75)

Note that the scoring algorithms measure each taxonomy's conformity to its distance matrix and as such are only useful when comparing taxonomies generated using the same similarity metric since only one similarity metric characterizes a distance matrix. This means that using a given scoring metric, it is impossible to compare all the taxonomy generation algorithms to each other, however it is possible to compare all the taxonomy generation algorithms that used the cosine similarity metric, symmetric NGD similarity metric, or asymmetric NGD similarity metric to each other.

The top 100, 250 and 500 frequently occurring terms in the Scopus "renewable energy" data set were used in conjunction with the entire bibliometric data set . The results are presented in the following subsections.

Using the top 100 terms

The results summarized in Table 9 are from tests run using the cosine similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 9 *Different Scoring Metrics used on Cosine Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
D-CB	**0.341**	27.632	20.835	27.632	24.066	26.961	23.533
D-CC	**0.341**	27.912	22.701	27.912	25.375	27.884	24.994
K-CB	**0.341**	27.632	20.835	27.632	24.066	26.961	23.533
K-CC	**0.341**	27.912	22.701	27.912	25.375	27.884	24.994
H-CB	0.285	23.340	21.232	23.340	23.391	24.112	22.487
H-CC	0.337	**28.272**	**24.805**	**28.272**	**27.305**	**28.587**	**26.460**

The results summarized in Table 10 are from tests run using the symmetric NGD similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 10 *Different Scoring Metrics used on Symmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
D-SB	**0.110**	13.126	20.918	13.126	18.366	14.085	16.925
D-SC	**0.110**	**13.045**	20.447	**13.045**	18.380	**13.922**	16.665
K-SB	0.323	27.662	27.662	27.662	29.214	29.214	28.327

K-SC	0.323	26.918	26.918	26.918	28.719	28.719	27.690
H-SB	0.119	14.451	17.806	14.451	15.998	14.898	16.471
H-SC	0.114	13.566	**17.307**	13.566	**15.680**	14.183	**15.857**

Finally, the results summarized in Table 11 are from tests run using the asymmetric similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 11 *Different Scoring Metrics used on Asymmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
E-AB	**0.028**	**3.651**	**3.758**	**3.651**	**3.431**	**3.410**	**3.608**
E-AC	**0.028**	**3.651**	**3.758**	**3.651**	**3.431**	**3.410**	**3.608**
H-AB	0.028	3.705	3.791	3.705	3.470	3.449	3.644
H-AC	**0.028**	**3.651**	**3.758**	**3.651**	**3.431**	**3.410**	**3.608**

Using the top 250 terms

The results summarized in Table 12 are from tests run using the cosine similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 12 *Different Scoring Metrics used on Cosine Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
D-CB	**0.333**	67.507	52.780	67.507	60.950	66.572	58.883
D-CC	**0.333**	**67.765**	53.575	**67.765**	61.082	**66.864**	59.388
K-CB	**0.333**	67.507	52.780	67.507	60.950	66.572	58.883
K-CC	**0.333**	**67.765**	53.575	**67.765**	61.082	**66.864**	59.388
H-CB	0.286	57.262	47.932	57.262	54.345	57.773	52.211
H-CC	0.327	67.260	**54.777**	67.260	**62.308**	66.696	**60.048**

The results summarized in Table 13 are from tests run using the symmetric NGD similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 13 *Different Scoring Metrics used on Symmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm	**Average**	**Momentum**	**Mean**	**Mean To**	**Linear**	**Exponential**	**Exponential**

Variant Acronym			To Root	Grandparent		(0.5)	(0.75)
D-SB	**0.112**	34.048	55.832	34.048	48.896	36.804	45.164
D-SC	**0.112**	**33.882**	52.860	**33.882**	47.140	**36.515**	44.071
K-SB	0.335	73.046	73.046	73.046	76.584	76.584	74.562
K-SC	0.335	72.651	72.651	72.651	76.321	76.321	74.224
H-SB	0.122	38.392	51.256	38.392	45.995	40.584	46.358
H-SC	0.117	35.266	**49.411**	35.266	**44.013**	37.741	**43.784**

Finally, the results summarized in Table 14 are from tests run using the asymmetric NGD similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 14 *Different Scoring Metrics used on Asymmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	Average	Momentum	Mean To Root	Mean To Grandparent	Linear	Exponential (0.5)	Exponential (0.75)
E-AB	**0.013**	**7.233**	**8.579**	**7.233**	**6.917**	**6.578**	**7.719**
E-AC	**0.013**	**7.233**	**8.579**	**7.233**	**6.917**	**6.578**	**7.719**
H-AB	0.013	7.820	8.680	7.820	7.067	6.758	7.861
H-AC	**0.013**	7.733	8.654	7.733	7.039	6.722	7.830

Using the top 500 terms

The results summarized in Table 15are from tests run using the cosine similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 15 *Different Scoring Metrics used on Cosine Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	Average	Momentum	Mean To Root	Mean To Grandparent	Linear	Exponential (0.5)	Exponential (0.75)
D-CB	**0.316**	124.705	89.254	124.705	105.961	121.457	103.552
D-CC	**0.316**	**124.998**	90.791	**124.998**	106.101	122.087	104.567
K-CB	**0.316**	124.705	89.254	124.705	105.961	121.457	103.552
K-CC	**0.316**	**124.998**	90.791	**124.998**	106.101	122.087	104.567
H-CB	0.260	100.394	76.155	100.394	88.615	100.057	86.514
H-CC	0.309	124.609	**95.556**	124.609	**111.186**	**122.363**	**107.433**

The results summarized in Table 16 are from tests run using the symmetric NGD similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 16 *Different Scoring Metrics used on Symmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
D-SB	**0.116**	71.257	138.022	71.257	118.328	78.324	100.547
D-SC	**0.116**	**71.123**	115.215	**71.123**	103.762	**77.345**	94.929
K-SB	0.361	170.407	170.407	170.407	173.822	173.822	171.871
K-SC	0.361	158.177	158.177	158.177	165.669	165.669	161.388
H-SB	0.126	79.727	119.995	79.727	104.705	85.882	102.732
H-SC	0.121	74.192	**105.994**	74.192	**93.791**	79.285	**92.860**

Finally, the results summarized in Table 17 are from tests run using the asymmetric NGD similarity metric to generate the distance matrix. Highlighted are the best taxonomy generation algorithms for each scoring metric.

Table 17 *Different Scoring Metrics used on Asymmetric NGD Similarity based Taxonomy Generation Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grandparent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
E-AB	0.015	15.659	**18.957**	15.659	15.374	14.441	17.005
E-AC	0.015	14.375	19.668	14.375	16.034	14.428	17.373
H-AB	0.015	15.737	19.013	15.737	15.444	14.516	17.076
H-AC	**0.015**	**14.165**	19.108	**14.165**	**15.275**	**14.059**	**16.869**

Evaluating Individual Taxonomies

The consistently top-scoring algorithms among the 100, 250 and 500 term list tests are summarized in Table 18 The shaded cells represent the consistently top-scoring algorithm variants for each of the scoring metrics.

Table 18 *Consistently Top Scoring Algorithm Variants*

Algorithm Variant Acronym	**Average**	**Momentum**	**Mean To Root**	**Mean To Grand-parent**	**Linear**	**Exponential (0.5)**	**Exponential (0.75)**
D-CB	▨						

D-CC	▨						
D-SB	▨						
D-SC	▨	▨		▨		▨	
K-CB	▨						
K-CC	▨						
K-SB							
K-SC							
E-AB			▨				
E-AC							
H-AB							
H-AC	▨						
H-CB							
H-CC			▨		▨		▨
H-SB							
H-SC			▨		▨		▨

Based on the data in the table above, the algorithm variant that performed the best is:

- DJP algorithm, symmetric NGD similarity, cosine centrality for root selection (D-SC)

Other notable top performing algorithm variants were:

- Heymann algorithm, symmetric NGD similarity, closeness centrality (H-SC)
- Heymann algorithm, cosine similarity, closeness centrality (H-CC)

Aside from the results summarized in Table 18 it is also clear through the Heymann algorithm tests that closeness centrality seems to be a better centrality metric to use compared to betweenness centrality. This is consistent with our observations in the previous consistency tests.

Results on Test 3 (Synthetic Data)

Finally, synthetic data sets were generated to allow the taxonomy generation algorithms to be tested on data with known characteristics. Using synthetic data provided us with a great deal of flexibility with respect to the kinds of experiments that could be conducted.

In this study, we conducted two different sets of tests. Firstly, it allowed us to vary the size of the data sets considerably. This was used to estimate the optimal range of sizes for bibliometric data sets from which taxonomies could be accurately inferred. By using synthetic data, we could precisely control the amount of noise that was present in a data set; this is exploited in the second set of tests, which was designed to study the performance of taxonomy generation algorithms when faced with different noise levels.

Estimating the Optimal Bibliometric Dataset Size

The first challenge was to estimate the optimal size for a synthetically produced taxonomy as a function of the size of the dataset. This was done by creating underlying taxonomies of different sizes then generating a varying number of documents for each term in these taxonomies. These were then presented to the taxonomy generation algorithms, the outputs of which were then compared to the valid, predetermined underlying taxonomies.

For these tests, a noise level of 0.2 was assumed within the documents. As mentioned in the previous subsection, this is defined as the probabilities of the "off-topic" terms relative to the probability of the relevant term. 0.2 was our subjective reasonable estimate for the noise level to be expected from a real publication database.

The two tables in the following pages summarize the results – note that all experiments were repeated three times to take into account the variance of generating random initial synthetic taxonomies, and the average scores reported. Table 19 lists the percentage similarity of the synthetic "true" underlying taxonomies to the taxonomies generated using algorithms that use betweenness centrality while Table 20 lists the ones for closeness centrality. Highlighted are the best performing data set sizes for each taxonomy generation algorithm.

Table 19 *Accuracy of Taxonomy Generation Algorithms Using Betweenness Centrality's Outputs for Replicating Underlyi Taxonomies*

Total Entries in Data Set	**Number of Terms**	**D-CB**	**D-SB**	**K-CB**	**K-SB**	**E-AB**	**H-AB**
1000	20	78.33%	78.33%	78.33%	21.67%	45.00%	33.33%
2000	20	83.33%	86.67%	83.33%	21.67%	36.67%	26.67%
2,500	50	92.67%	89.33%	92.67%	10.00%	76.00%	60.67%
5,000	50	94.67%	95.33%	94.67%	10.00%	64.67%	48.00%
5,000	100	95.00%	91.33%	95.00%	8.33%	77.00%	62.00%
10,000	100	94.67%	96.33%	94.67%	5.00%	**89.67%**	**73.67%**
20,000	20	81.67%	91.67%	81.67%	20.00%	35.00%	26.67%
50,000	50	91.33%	93.33%	91.33%	13.33%	74.67%	47.33%
100,000	20	80.00%	83.33%	80.00%	23.33%	48.33%	28.33%
100,000	100	**95.67%**	**97.33%**	**95.67%**	8.00%	76.67%	51.00%
200,000	20	85.00%	86.67%	85.00%	**30.00%**	40.00%	33.33%
250,000	50	87.33%	92.67%	87.33%	8.00%	84.00%	57.33%
500,000	50	92.67%	96.00%	92.67%	9.33%	74.67%	52.00%
500,000	100	94.33%	96.67%	94.33%	6.67%	81.00%	55.33%
1,000,000	100	95.00%	97.00%	95.00%	4.67%	88.00%	59.67%

Table 20 *Accuracy of Taxonomy Generation Algorithms Using Closeness Centrality's Outputs for Replicating Underlying Synthetically Generated Taxonomies*

Total Entries in Data Set	**Number of Terms**	**D-CC**	**D-SC**	**K-CC**	**K-SC**	**E-AC**	**H-AC**	**H-CC**	**H-SC**
1000	20	86.67%	90.00%	86.67%	18.33%	45.00%	36.67%	73.33%	68.33%
2000	20	93.33%	95.00%	93.33%	16.67%	36.67%	35.00%	80.00%	76.67%
2,500	50	97.33%	95.33%	97.33%	8.67%	76.00%	70.67%	86.67%	72.67%
5,000	50	98.00%	96.67%	98.00%	8.67%	64.67%	66.00%	83.33%	77.33%
5,000	100	98.67%	94.67%	98.67%	7.33%	77.00%	73.67%	82.00%	61.00%
10,000	100	98.67%	**99.00%**	98.67%	4.00%	**89.67%**	88.00%	86.67%	67.00%
20,000	20	95.00%	95.00%	95.00%	16.67%	35.00%	35.00%	83.33%	**81.67%**
50,000	50	98.00%	98.00%	98.00%	11.33%	74.67%	71.33%	86.67%	73.33%
100,000	20	95.00%	95.00%	95.00%	18.33%	48.33%	46.67%	81.67%	80.00%
100,000	100	**99.00%**	**99.00%**	**99.00%**	7.33%	76.67%	75.33%	88.33%	49.00%
200,000	20	95.00%	95.00%	95.00%	**26.67%**	40.00%	36.67%	88.33%	**81.67%**
250,000	50	98.00%	98.00%	98.00%	6.00%	85.33%	82.67%	87.33%	80.67%
500,000	50	98.00%	98.00%	98.00%	8.00%	74.67%	75.33%	90.00%	72.67%
500,000	100	**99.00%**	**99.00%**	**99.00%**	5.67%	81.33%	82.33%	90.00%	44.33%
1,000,000	100	**99.00%**	**99.00%**	**99.00%**	3.67%	88.00%	**90.00%**	**90.67%**	52.33%

Based on the tables above, we note several key observations:

1. DJP and Kruskals algorithm variants have the general trend where the more terms in the taxonomy or the more entries in the bibliometric data set exist, the more accurate the replication of the underlying taxonomy is.
2. The algorithms that use cosine similarity perform much better than the other algorithm variants.
3. Using the closeness centrality metric produces much more accurate results than using the betweenness centrality. The disparity between the two is evident in the tests run using the Heymann algorithm.

The results of the tests that used closeness similarity are summarized in Table 21.

Table 21 *Average of Closeness Centrality Algorithms Accuracy Results*

Total Entries in Data Set	**Number of Terms**	**Average of Percentage Similarities for all Taxonomy Generation Algorithms**
1000	20	63.13%
2000	20	65.83%
2,500	50	75.58%
5,000	50	74.08%
5,000	100	74.13%
10,000	100	78.96%
20,000	20	67.08%
50,000	50	76.42%
100,000	20	70.00%
100,000	100	74.21%
200,000	20	69.79%
250,000	50	**79.50%**
500,000	50	76.83%
500,000	100	75.08%
1,000,000	100	72.71%

As shown in the table above, taxonomy generation algorithms on average are most accurate (best replicate the underlying taxonomy) when there are 50 terms and 250,000 bibliometric entries in the data set. Past this value the mean of the accuracy of the taxonomy generation algorithms decreases. As such, for the tests in the next section where we varied the noise, we considered the scenario where there were 250,000 total entries in the data set.

Measuring Algorithm Variant Consistency Using Synthetic Data

Using a predetermined, underlying taxonomy with size based on the findings in the previous section, the "noise" values within the data set were varied to calculate the robustness vs. noise, or consistency of each taxonomy generation algorithm.

A data set was created consisting of 50 terms with 5,000 entries generated for each term, totaling to 250,000 entries in the synthetic data set. The test was run three times and the percentage similarity values were averaged. The results of this test are summarized in Table 22 below. Note that the percentage values represent the degree of similarity of the outputs of each taxonomy generation algorithm to the underlying taxonomy. Highlighted values represent the best performing algorithms for every noise value.

Table 22 *Accuracy of Taxonomy Generation Algorithms for Replicating Underlying Synthetically Generated Taxonomies with 50 Terms with Varying Noise*

Algorithm Variant Acronym	**Noise = 0**	**Noise = 0.2**	**Noise = 0.5**	**Noise = 0.8**	**Noise = 1**	**Average**	**Std Dev (does not count case where noise = 1)**
D-CB	94.00%	90.00%	84.67%	13.33%	2.00%	56.80%	38.30%
D-CC	**98.00%**	**98.00%**	90.67%	11.33%	2.00%	60.00%	42.25%
D-SB	**98.00%**	95.33%	92.00%	89.33%	2.00%	75.33%	3.79%
D-SC	**98.00%**	**98.00%**	**98.00%**	**96.67%**	2.00%	**78.53%**	**0.67%**
K-CB	94.00%	90.00%	84.67%	13.33%	**15.33%**	59.47%	38.30%
K-CC	**98.00%**	**98.00%**	90.67%	11.33%	**15.33%**	62.67%	42.25%
K-SB	8.67%	11.33%	10.67%	8.00%	**15.33%**	10.80%	1.59%
K-SC	8.67%	9.33%	8.67%	6.67%	**15.33%**	9.73%	1.15%
E-AB	96.67%	80.67%	21.33%	8.00%	**15.33%**	44.40%	43.56%
E-AC	96.67%	80.67%	21.33%	6.00%	**15.33%**	44.00%	44.24%
H-AB	80.67%	56.67%	2.00%	2.00%	2.00%	28.67%	39.72%
H-AC	96.67%	78.67%	19.33%	6.00%	2.00%	40.53%	44.26%
H-CB	2.00%	2.00%	2.00%	2.00%	2.00%	2.00%	**0.00%**
H-CC	90.67%	91.33%	86.00%	9.33%	2.00%	55.87%	40.07%
H-SB	64.00%	34.00%	2.00%	2.00%	2.00%	20.80%	29.77%
H-SC	**98.00%**	75.33%	86.00%	89.33%	2.00%	70.13%	9.37%

Based on the data in the table above, the best performing and most robust algorithm vs noise was the DJP algorithm with symmetric NGD similarity and closeness centrality for choosing the root term (D-SC). It consistently managed to replicate most of the links in the underlying taxonomy and had a low variance in its percentage accuracy as the noise values were varied.

Analysis of Results

The tests shown above rigorously tested each taxonomy generation algorithm variant. The first set of tests conducted measured each algorithm variant's consistency, or robustness vs noise. The consistency tests were further subdivided into backend data consistency tests and term list consistency tests. From the first set of tests, it was discovered that the most consistent algorithm variants were D-CC, K-CC and H-CC, all of which use cosine similarity and closeness centrality to generate taxonomies. The fact that these three variants were the most consistent seem to show that the cosine similarity metric and closeness centrality are effective algorithm parameters as well.

The second set of tests conducted evaluated individual taxonomies based on several scoring metrics that measured each taxonomy generation algorithm variant's conformity to its distance matrix. Each distance matrix is built using a particular similarity metric, so one downside of this test was that it was impossible to compare algorithms that used different similarity metrics to generate their distance matrices. Among the algorithm variants, the consistent top performer was D-SC, followed by H-CC and H-SC. Once again, closeness centrality was the metric all the efficient algorithm variants used to generate their taxonomies, however this time around the symmetric NGD metric was used by the top performer to generate its distance matrix. Similar to the cosine similarity, Symmetric NGD similarity is another similarity metric that produces an undirected graph. This seems to indicate that the most consistent and top-scoring algorithm variants use similarity metrics that are undirected.

Finally, synthetic data sets were generated based on known, randomly generated taxonomies and were used to measure the respective performances of each of the taxonomy generation algorithm variants in replicating the underlying taxonomy. The first set of synthetic data tests showed that the ideal data set size for which our algorithms can accurately produce valid taxonomies consistently is 250,000 entries. Then, using this data set size, the noise within the data was varied and each taxonomy generation algorithm variant's robustness vs noise was measured. From these tests the best performing algorithm variant was found to be D-SC.

Based on all the tests conducted, there is now convincing evidence that the best algorithm variants are H-CC and D-SC, since these were the two algorithm variants that performed well in multiple (two out of three) tests. As a reality check on the results of the analysis, each high-scoring algorithm variant's output was manually inspected to determine the taxonomy generation algorithm which resulted in the most convincing taxonomy. Specifically, we inspected the taxonomies generated by H-CC and D-SC using the entire “renewable energy” Scopus data set as a backend, and with the top 500 frequently occurring terms as the node-list.

The figures on the succeeding pages show the taxonomies generated by both the algorithm variants. Specifically Figure 25 shows the H-CC taxonomy and Figure 26 shows the D-SC taxonomy.

One main observation that is immediately clear upon inspection of the D-SC taxonomy is that it is very deep, going as far as 25 levels in. Note that the root of the taxonomy in the figure is on the left-hand-side and as such a deeper taxonomy would be a wider / broader figure. In contrast to the D-SC taxonomy, the H-CC taxonomy is not very deep, though it still goes 5-9 levels in.

Upon a more granular inspection of both taxonomies, it seems that the taxonomy generated using the H-CC algorithm makes a little more sense. Both taxonomies generated using H-CC and D-SC used the same term list, but the taxonomy generated using D-SC did not have any clear clustering of terms that represented the same idea, whereas the one generated using H-CC had clear term clusters, which are indicated in Figure 25. The lack of clustering in the D-SC taxonomy is also a by-product of its depth. Since it is very deep, it isn’t very broad, hence each term only has on average 3 children, and hence it’s harder to immediately notice term clusters.

Even though the H-CC taxonomy looks more sensible than the D-SC taxonomy, it is by no means perfect. For instance, there are clusters in the taxonomy that grouped seemingly unrelated terms together. An example of this was a cluster of terms where the parent node was “ph”, a chemistry-related term referring to the acidity of a solution, however its children were “solar” related like “photovoltaic”, “spectrum analysis” and “photoconduction” as well as chemistry-related terms like “solute”.

On the other hand, the taxonomy generated using D-SC also had its advantages. Within the taxonomy, certain logical paths could be traced. For instance, starting from “power generation”, we can trace the following path by going deeper in the taxonomy: “power generation” → “electric powers” → “power system” → “electrical power system” → “power transmission” → “electric power transmission” → “electric power transmission networks” → “electric network analysis.

As such, it can be concluded that both the H-CC and the D-SC produced taxonomies provide information that is useful in different ways. Ultimately, the choice of which taxonomy generation algorithm variant to use is dependent on the viewer’s preferences; for example, if a taxonomy that separates distinct term concepts is required, then H-CC should be used. If, on the other hand, a taxonomy for tracing long paths between related terms is desired, then D-SC would be the preferred option.

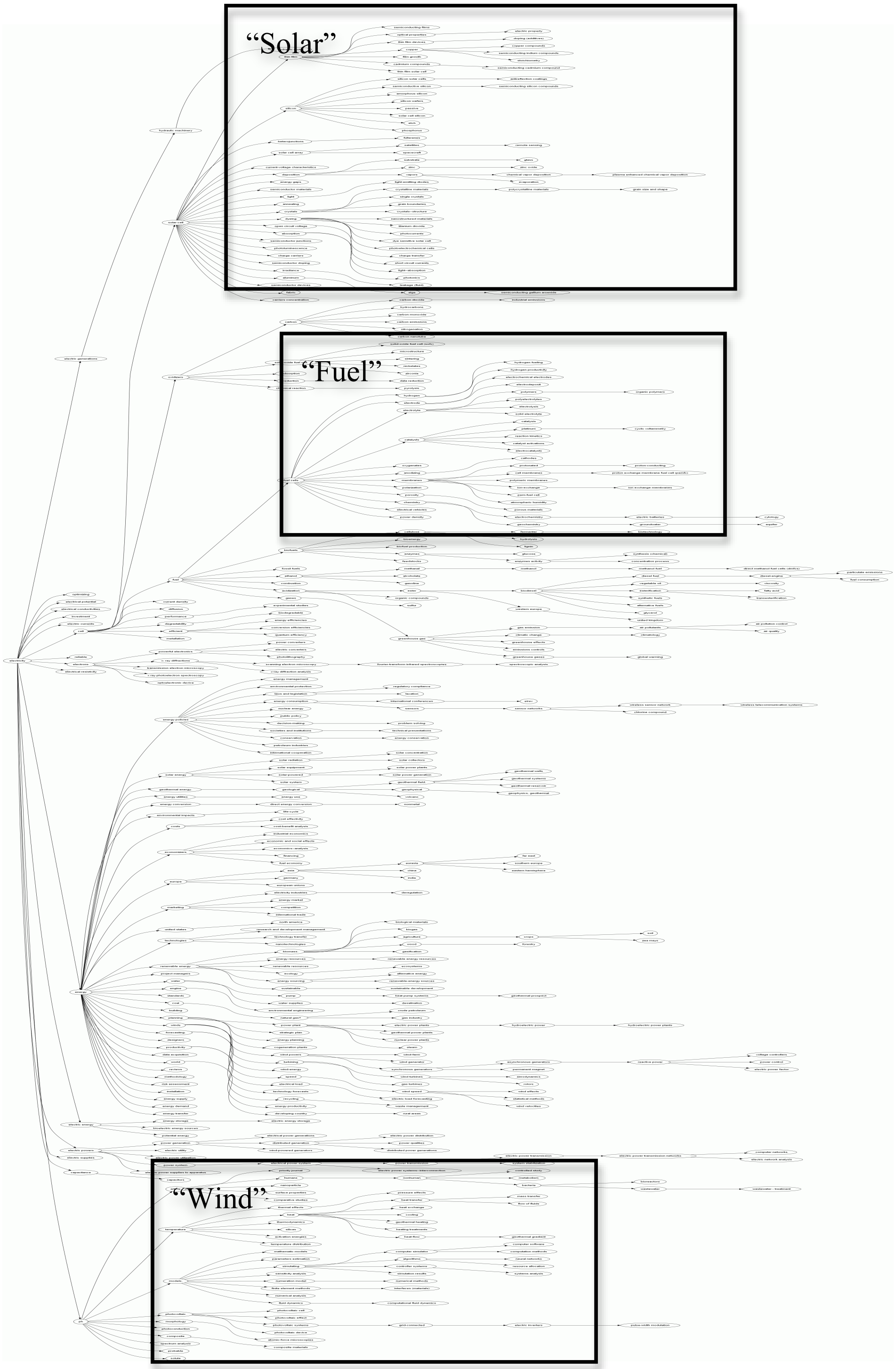


Figure 25 H-CC generated Taxonomy

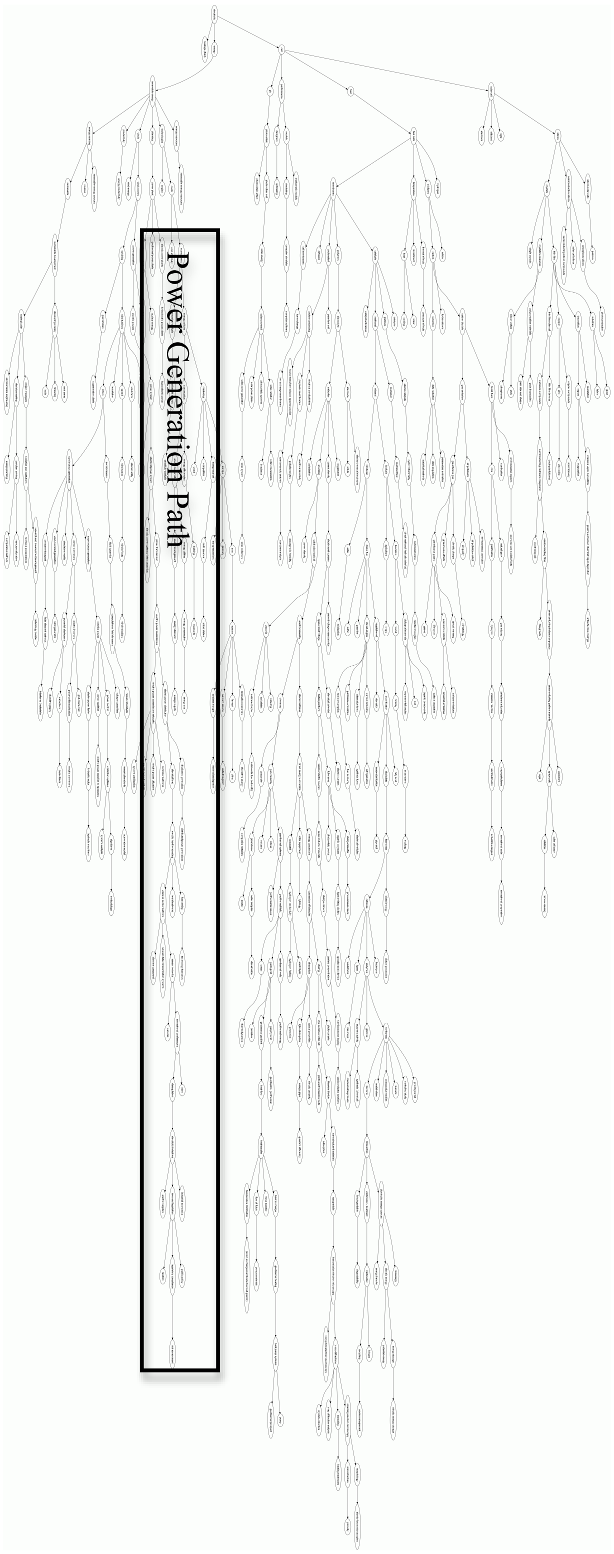


Figure 26 D-SC generated taxonomy

PROJECT SUMMARY

To summarize the achievements and progress made during this reporting period, listed below are the key components of the project based both on the technology forecasting framework described here, and on the stated deliverables on the project; for each there is a brief summary of the associated activities and an indication if it was primarily attended to by MIST or MIT researchers, or if it was a joint effort:

1. *Data collection and term extraction* – Various tools and techniques were developed to support the automated collection of data from online sources of data, and the extraction of relevant keywords from these collections. More recent developments included the use of NLP based tools (specifically noun phrase extraction) to augment this process.

   MIST/MIT division: Joint

2. *Taxonomy generation* – We studied a number of approaches for automatically organizing and visualizing our collection of relevant keywords. These were primarily in the form of taxonomies which organized the keywords in a hierarchical format. Three main approaches were investigated:

   - A Genetic Algorithms based approach.
   - The approach described in [Heymann and Garcia Molina, 2006], for which we also proposed a number of important modifications to support the application on technology forecasting.
   - A semi-automated approach which allowed input from domain experts to be considered

   Another new approach which was explored was the use of “Topic Maps” as an alternative approach to research landscape visualization.

   MIST/MIT division: Primarily MIST

3. *Early growth indicators* – During the course of the project we have developed and screened a number of early growth indicators, and these have been used to generate the annotated taxonomies seen in other areas of the project. Currently Hanan Shemaili, one of our M.Sc students, is finishing the thesis which deals with numerically evaluating the indicators.

   MIST/MIT division: Joint

4. *Bibliometric analysis of blogs* – A set of techniques for extracting key concepts from blog posts was created. These were combined with the CLIOS system representation technique to create a

system for "re-representation" of a domain; i.e.: a method where blog posts could be used to propose candidate modifications to existing CLIOS representations, thereby ensuring their continued relevance and/or correctness. To validate the proposed approach, a number of case studies were conducted including one focusing on renewable energy.

MIST/MIT division: Primarily MIT

5. *Evaluation of taxonomy generation algorithms* – Given the importance of the taxonomy generation process to our framework, a detailed study was conducted on the various algorithm variants available. In this study, a systematic process for taxonomy evaluation was also proposed.

   MIST/MIT division: Primarily MIT

6. *Renewable energy case study* – We have realized that it is difficult to apply the methodology uniformly across the entire spectrum of renewable energy technologies, simply because the structures of the individual research landscapes differ hugely across the various domains. As such, we have shifted our focus to conducting individualized case studies customized to specific domains. This has been captured in a variety of forms, either as academic publications, a set of technology "posters", or as surveys/technical papers which will be submitted for publication in the future. Links to these resources have been provided in Appendix A.

   Also, while we are technically at the end of the project duration, the research is still ongoing on this area and we have funding remaining to retain the services of our Post-Doctoral researcher, Dr. Andreas Henschel, for around two more months. He is currently working on a case study for PV technologies in collaboration with Dr. Matteo Chiesa.

   MIST/MIT division: Joint

7. *Development of software tools* – In previous reporting periods, two main tools were worked on – the "Cameleon Scheduler", and the "Long Tail analysis tool". In more recent work a topic map visualization tool was also developed. In addition, a variety of command-line (text-based) taxonomy construction and early growth indication tools were developed as part of the reference implementation tool-chain.

   MIST/MIT division: Joint

FUTURE WORK

Though we have reached the formal conclusion of the project, the research of course does not stop here; as mentioned, we have extended the stay of Dr. Andreas Henschel, the postdoctoral researcher who has been working on this project. During his remaining time he will be working on additional case studies, and also in completing and debugging the substantial amount of code that he has developed in the last one and a half years. In addition, we are currently collaborating with Dr. Hatem Zeineldin on a further case study on distributed generation and smart grids.

In addition to further case studies, there is also potential for many future improvements to the proposed methodologies. These include:

1. Incorporation of the newer research directions described in this report into the reference implementation (or at least creation of an integrated system that allows these techniques to be plugged-in effortlessly). These include the threads on blog analysis, topic modeling and semi-supervised taxonomy creation.
2. Use of patents as a further source of data. While in principle our techniques are “data-neutral”, in practice it has to be acknowledged that different forms of data present different advantages and constraints. For example, the analysis that was conducted using blogs was of a very different format to that conducted using publications. This was both a result of the nature of the data itself, as well as the format in which the data was made available to us (there is no easy way to collect all blog entries for a range of years, for example). Instead of simply ruling out potentially valuable sources of information, we chose to be flexible in terms of the types of analysis that can be conducted. As such, we expect that the analysis of patents will lead to many innovations and the investigation of new ways in which the evolution of technology can be studied.
3. Another important research area is the discovery of more appropriate distance measures. Co-occurrence statistics have already provided us with many interesting insights; however they are also very noisy and can be very difficult to interpret. The simultaneous appearance of two terms in the same article could be indicative of a variety of relationships (for e.g., that technology A is preferable to B, or that it is an enabling technology for B, of that it is often mistaken for technology B (but is different), or that it is an alternative to technology B, or that technology B is a subset of technology A, etc.). Alternative ways of quantifying relationships might take into account collaborations between authors, link structure (for blogs) and citation patterns.

4. Extension of visualization techniques to include semantic networks (taxonomies can be considered a subset of semantic networks but in some cases can be a little too restrictive). Instead of imposing a specific kind of structure (general → specific in the case of taxonomies), most semantic networks simply require that links capture relationships between similar or related technologies (alternatively, we would like to build structures where links represent *related-to* connections as opposed to *is-a* relationships).

5. Finally, it would also be very interesting to explore the possibility of using *crowd sourcing* techniques to enhance/validate taxonomies. Many of the taxonomies that we generated looked "approximately correct" from a high level point of view, but upon closer inspection contained a number of links which were clearly incorrect. This is difficult to avoid when using fully automated methods, particularly given the inherent difficulty of the process and the noisiness of the data. However, one could envision some mechanism whereby links created with low confidence levels are submitted for human verification or where the co-occurrence statistics are supplemented with "social" or voting patterns collected from a large group of human scorers.

REFERENCES

[Alexopoulou et al., 2008] Alexopoulou, D., Wächter, T., Pickersgill, L., Eyre, C., and Schroeder, M. (2008). Terminologies for text-mining; an experiment in the lipoprotein metabolism domain. BMC Bioinformatics, 9 Suppl 4:S2.

[Allison et al, 2009] Allison, G., Thain S., Morris P., Morris C., Hawkins S., Hauck B., Barraclough T., Yates N., Shield I., Bridgwater A. and Donnison I. (2009) Quantification of hydroxycinnamic acids and lignin in perennial forage and energy grasses by Fourier-transform infrared spectroscopy and partial least squares regression, Bioresource Technology, 100(3).

[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.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G. (2007). DBpedia: A Nucleus for a Web of Open Data. volume 4825 of Lecture Notes in Computer Science, pages 722–735, Berlin, Germany. Springer.

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

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). Natural language process- ing with Python. O'Reilly Media.

[Bishop, 2006] Bishop, C. (2006). Pattern Recognition and Machine Learning. Infor- mation Science and Statistics. Springer-Verlag New York Inc.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

[Bonino et al., 2010] Bonino, D., Ciaramella, A., and Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information, 32(1):30 – 38.

[Brandes, 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25:163–177.

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

[Brown, 2008] Brown, R. (2008). Impact of Smart Grid on distribution system design. In 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pages 1–4.

[Cai and Hofmann, 2004] Cai, L. and Hofmann, T. (2004). Hierarchical doc- ument categorization with support vector machines. In Proceedings of the thirteenth ACM international conference on Information and knowledge man- agement, CIKM ’04, pages 78–87, New York, NY, USA. ACM.

[Chen, 2006] Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57:359–377.

[Chiesa, 2010] Chiesa, M. (2010). Masdar Institute of Science and Technology. Personal Interview.

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

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

[Coll-Mayor et al., 2007] Coll-Mayor, D., Paget, M., and Lightner, E. (2007). Future intelligent power grids: Analysis of the vision in the European Union and the United States. Energy Policy, 35(4):2453–2465.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. In Machine Learning, pages 273–297.

[Cunningham et al., 2006] Cunningham, S. W., Porter, A. L., and Newman, N. C. (2006). Special issue on tech mining. Technological Forecasting and Social Change, 73(8):915–922.

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

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

[Daim et al., 2006] Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Fore- casting emerging technologies: Use of bibliometrics and patent analysis. Techno- logical Forecasting and Social Change, 73(8):981 – 1012. Tech Mining: Exploiting Science and Technology Information Resources.

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

[Dawelbait et al., 2010] Dawelbait, G., Mezher, T., Woon, W. L., and Hen- schel*, A. (2010). Taxonomy based trend discovery of renewable energy tech- nologies in desalination and power generation. In Portland International Cen- ter for Management of Engineering and Technology (PICMET), Technology Management for Global Economic Growth. *Corresponding Author.

[Dawelbait et al., 2011] Dawelbait, G., Henschel*, A., Mezher, T., and Woon, W. L. (2011). Taxonomy based trend discovery of renewable energy technolo- gies in desalination and power generation. International Journal of Social Ecology and Sustainable Development. *Shared first author. Invited paper, (Conditionally accepted).

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

[Doms and Schroeder, 2005] Doms, A. and Schroeder, M. (2005). GoPubMed: Exploring PubMed with the Gene Ontology. Nucleic Acids Res, 33(Web Server issue):783–786.

[Dumais and Chen, 2000] Dumais, S. T. and Chen, H. (2000). Hierarchical clas- sification of Web content. In Belkin, N. J., Ingwersen, P., and Leong, M. K., editors, Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, pages 256–263, Athens, GR. ACM Press, New York, US.

[Emziane, 2010] Emziane, M. (2010). Masdar Institute of Science and Technology. Personal Interview.

[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.

[Fall et al., 2003] Fall, C., T □orcsv ́ari, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. ACM SIGIR Forum, 37(1).

[Fellbaum, 1998] Fellbaum, C., editor (1998). WordNet: an electronic lexical database. MIT Press.

[Frantzi et al., 2000] Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries, V3(2):115–130.

[Giles, 2005] Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438(7070):900–901.

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scien- tific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235.

[Griffithsetal.,2007] Griffiths,T.L.,Steyvers,M.,andTenenbaum,J.B.(2007).Top- ics in semantic representation. Psychological Review, 114(2):211–244.

[Hagberg et al., 2008] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008), pages 11–15, Pasadena, CA USA.

[Halletal.,2008] Hall,D.,Jurafsky,D.,andManning,C.(2008).Studyingthehistory of ideas using topic models. In Proceedings from the EMNLP 2008: Conference on Empirical Methods in Natural Language Processing.

[Harris, 1968] Harris, Z. (1968). Mathematical Structures of Language. Wiley.

[Hashaikeh, 2010] Hashaikeh, R. (2010). Masdar Institute of Science and Technology. Personal Interview.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France.

[Heinrich, 2008] Heinrich, G. (2008). Parameter estimation for text analysis. Techni- cal report, University of Leipzig.

[Henschel et al., 2009] Henschel, A., Woon, W. L., Wä chter, T., and Madnick, S. (2009). Comparison of generality based algorithm variants for automatic taxonomy generation. In Proceedings of the 6th international conference on Innovations in information technology, IIT'09, pages 206–210, Piscataway, NJ, USA. IEEE Press.

[Henschel et al., 2010a] Henschel, A., Casagrande, E., Woon, W. L., Janajreh, I., and Madnick, S. (2010a). Business Intelligence Applications and the Web: Models, Systems and Technologies, chapter ”A unified approach for Taxonomy-based Technology Forecasting”. IGI Global, Universidad de Ali- cante, Spain. Accepted.

[Henschel et al., 2010b] Henschel, A., Wachter, T., Woon, W. L., and Mad- nick, S. (2010b). Renewable Energy Technology Forecasting using validated Taxonomy Generation Algorithms. Expert Systems with Applications. (under review).

[Heymann and Garcia-Molina, 2006] Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University.

[Hotho et al., 2003a] Hotho, A., Staab, S., and Stumme, G. (2003a). Ontologies improve text document clustering. In ICDM ’03: Proceedings of the Third IEEE International Conference on Data Mining, page 541, Washington, DC, USA. IEEE Computer Society.

[Hotho et al., 2003b] Hotho, A., Staab, S., and Stumme, G. (2003b). Wordnet improves text document clustering. In In Proc. of the SIGIR 2003 Semantic Web Workshop, pages 541–544.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European Conference on Machine Learning (ECML), pages 137–142, Berlin. Springer.

[Kasravi and Risov, 2007] Kasravi, K. and Risov, M. (2007). Patent mining - discov- ery of business value from patent repositories. Hawaii International Conference on System Sciences, 0:54b.

[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.

[Klein and Bernstein, 2001] Klein, M. and Bernstein, A. (2001). Searching for services on the semantic web using process ontologies. In In Proceedings of the International Semantic Web Working Symposium (SWWS, pages 159–172. IOS press.

[Kobilarov et al., 2009] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R. (2009). Media meets semantic web - how the bbc uses dbpedia and linked data to make connections. In ESWC, pages 723–737.

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

[Kostoff, 2006] Kostoff, R. N. (2006). Systematic acceleration of radical discovery and innovation in science and technology. Technological Forecasting and Social Change, 73(8):923–936.

[Kroposki et al., 2008] Kroposki, B., Basso, T., and DeBlasio, R. (2008). Microgrid standards and technologies. In 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pages 1–4.

[Liu et al., 2005] Liu, T. Y., Yang, Y., Wan, H., Zeng, H. J., Chen, Z., and Ma, W. Y. (2005). Support vector machines classification with a very large-scale taxonomy. SIGKDD Explor. Newsl., 7(1):36–43.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). NLTK: The natural language toolkit. In In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural

Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

[Loper and Bird, 2002] Loper, E. and Bird, S. (2002). NLTK: The natural language toolkit. In In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

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

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schu □tze, H. (2008). Introduction to Information Retrieval. Cambridge University Press, 1 edition.

[Margolis, 2002] Margolis, R.K. (2002). Understanding Technological Innovation in the Energy Sector: The Case of Photovoltaics. Doctoral Dissertation, Woodrow Wilson School of Public and International Affairs, Princeton University.

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

[Martino, 2003] 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.

[Meietal.,2007] Mei,Q.,Shen,X.,andZhai,C.(2007).Automaticlabelingofmulti- nomial topic models. In KDD '07: Proceedings of the 13th ACM SIGKDD interna- tional conference on Knowledge discovery and data mining, pages 490–499, New York, NY, USA. ACM.

[Moehrle et al., 2010] Moehrle, M., Walter, L., Bergmann, I., Bobe, S., and Skrzipale, S. (2010). Patinformatics as a business process: A guideline through patent research tasks and tools. World Patent Information.

[Newman, 2010] Newman, M. E. J. (2010). Networks: An Introduction. Oxford Uni- versity Press.

[Novosel, 2008] Novosel, D. (2008). Emerging technologies in support of smart grids. In 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pages 1–2.

[Patel, 2006] Patel, M. (2006). Wind and solar power systems: design, analysis, and operation. CRC Press.

[Porter and Cunningham, 2005] Porter, A. and Cunningham, S. (2005). Tech mining: Exploiting new technologies for competitive advantage. Wiley-Interscience.

[Porter and Newman, 2005] Porter, A. and Newman, N. (2005). Patent Profiling for Competitive Advantage, chapter 27, pages 587–612. Springer-Verlag.

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

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

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

[Porter, 2007] Porter, A. (2007). How "tech mining" can enhance R&D management. Research Technology Management, 50(2):15 – 20.

[Porter and Newman, 2011] Porter, A.L. and Newman, N.C. (2011). Mining External R&D. Technovation, 31(2011):171-176.

[Prieto, 2010] Prieto, R. (2010). Masdar Institute of Science and Technology. Personal Interview.

[Ryu and Choi, 2006] Ryu, P.-M. and Choi, K.-S. (2006). Taxonomy learning using term specificity and similarity. In Proceedings of the 2nd Work shop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pages 41–48, Sydney, Australia, July 2006. Association for Computational Linguistics.

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. W. (1999). Deriving concept hierarchies from text. In Research and Development in Information Retrieval, pages 206–213.

[Sao and Lehn, 2008] Sao, C. and Lehn, P. (2008). Control and Power Management of Converter Fed Microgrids. IEEE Transactions on Power Systems, 23(3):1088–1098.

[Saxenian, 1996] Saxenian, A. (1996). Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Cambridge, Harvard University Press.

[Schenk et al., 2008] Schenk, P., Thomas-Hall, S., Stephens, E., Marx, U., Mussgnug, J., Posten, C., Kruse, O., and Hankamer, B. (2008). Second gener- ation biofuels: High-efficiency microalgae for biodiesel production. BioEnergy Research, 1(1):20–43.

[Sgouridis, 2010] Sgouridis, S (2010). Masdar Institute of Science and Technology. Personal Interview.

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

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

[Snow et al., 2006] Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

[Steyvers and Griffiths, 2007] Steyvers, M. and Griffiths, T. (2007). Probabilistic Topic Models, chapter 21, pages 427–446. Lawrence Erlbaum Associates.

[Suchanek et al., 2008] Suchanek, F., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. Web Semantics: Sci- ence, Services and Agents on the World Wide Web, 6(3):203 – 217. World Wide Web Conference 2007Semantic Web Track.

[Tabaei, 2010] Tabaei, A. (2010). Masdar Institute of Science and Technology. Personal Interview.

[Trippe, 2003] Trippe, A. J. (2003). Patinformatics: Tasks to tools. World Patent Information, 25(3):211–221.

[Tseng et al., 2007] Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining tech- niques for patent analysis. Information Processing & Management, 43(5):1216– 1247.

[U.S. Dept. of Health,] U.S. Dept. of Health. Medical subject headings.

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

[Velardi et al., 2007] Velardi, P., Cucchiarelli, A., and Petit, M. (2007). A taxonomy learning method and its application to characterize a scientific web community. IEEE Trans. on Knowl. and Data Eng., 19(2):180–191.

[Vidican et al., 2009] Vidican, G., Woon, W., and Madnick, S. (2009). Measuring innovation using bibliometrics: the case of solar photovoltaic industry. In Advancing the Study of Innovation and Globalization in Organizations (ASIGO), Nuremberg, Germany.

[Wang and McCallum, 2006] Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 424–433, New York, NY, USA. ACM.

[Watts et al., 1998] Watts, R. J., Porter, A. L., and Newman, N. C. (1998). Innovation forecasting using bibliometrics. Competitive Intelligence Review, 9(4):11–19.

[Weld et al., 2008] Weld, D. S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffmann, R., Patel, K., and Skinner, M. (2008). Intelligence in wikipedia. In AAAI, pages 1609–1614.

[Widdows, 2003] Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 197–204, Morristown, NJ, USA. Association for Computational Linguistics.

[Wong et al., 2008] Wong, J., Baroutis, P., Chadha, R., Iravani, R., Graovac, M., and Wang, X. (2008). A methodology for evaluation of permissible depth of penetration of distributed generation in urban distribution systems. In 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pages 1–8.

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

[Woon and Madnick, 2009] Woon, W. and Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. Knowledge and Information Systems, 21(1):91–111.

[Woon et al., 2010] Woon, W., Henschel, A., and Madnick, S. (2010). A framework for technology forecasting and visualization. In Innovations in Information Tech- nology, 2009. IIT'09. International Conference on, pages 155–159. IEEE.

[Wu and Weld, 2008] Wu, F. and Weld, D. S. (2008). Automatically refining the Wikipedia infobox ontology. pages 635–644, New York, NY, USA. ACM.

[Wu et al., 2008] Wu, F., Hoffmann, R., and Weld, D. S. (2008). Information extraction from wikipedia: moving down the long tail. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 731–739, New York, NY, USA. ACM.

[Xu et al., 2004] Xu, W., Mauch, K., and Martel, S. (2004). An Assessment of DG Islanding Detection Methods and Issues for Canada, report# CETC-Varennes 2004-074 (TR), CANMET Energy Technology Centre–Varennes. Natural Resources Canada.

[Zeineldin et al., 2006] Zeineldin, H., El-Saadany, E., and Salama, M. (2006). Distributed Generation Micro-Grid Operation: Control and Protection. In Power Systems Conference: Advanced Metering, Protection, Control, Communication, and Distributed Resources, 2006. PS'06, pages 105–111.

[Zeineldin, 2010] Zeineldin, H. (2010). Masdar Institute of Science and Technology. Personal Interview.

[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.

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

[Ziegler et al, 2009b] Approach and Preliminary Results for Early Growth Technology Analysis (2009), Blaine Ziegler, Ayse Kaya Firat, Stuart Madnick, Wei Lee Woon, Steven Camina, Clare Li, Erik Fogg, MIT Sloan Research Paper No. 4756-09, http://ssrn.com/abstract=1478001

[Ziegler et al, 2009b] Approach and Preliminary Results for Early Growth Technology Analysis (2009), Blaine Ziegler, Ayse Kaya Firat, Stuart Madnick, Wei Lee Woon, Steven Camina, Clare Li, Erik Fogg, MIT Sloan Research Paper No. 4756-09, http://ssrn.com/abstract=1478001

PUBLICATIONS/PRESENTATIONS

The following is an updated list of publications produced as part of this project.

Journal articles and Book Chapters

"Forecasting Renewable Energy technologies in Desalination and Power Generation using Taxonomies", Dawelbait G., Henschel A., Mezher T. and Woon W.L. (2011) *International Journal of Social Ecology and Sustainable Development* (Accepted subject to revisions).

"Bibliometric analysis of distributed generation" (2010), W.L. Woon, H.Zeineldin and S.E. Madnick, *Technology Forecasting and Social Change*, 78(3):408-420. http://dx.doi.org/10.1016/j.techfore.2010.08.009

"A unified approach for taxonomy-based technology forecasting", (2010) A. Henschel, E. Casagrande, W.L. Woon, I. Janajreh and S.E. Madnick, *Business Intelligence And The Web* (Book Chapter – in press).

"Asymmetric information distances for automated taxonomy construction " (2009), W.L Woon and S.E. Madnick, *Knowledge and Information Systems*, 21(1):91-111. http://dx.doi.org/ 10.1007/s10115-009-0203-5

"Renewable Energy Technology Forecasting using validated Taxonomy Generation Algorithms" (2010), A. Henschel, W.L. Woon, T. Wachter and S.E. Madnick, *Expert Systems with Applications*, (under review).

"Semantic Distances for Research Landscape Visualization" (2011), W.L. Woon and S.E. Madnick, *Journal of Intelligent Information Systems* (under review).

Conference publications

"Taxonomy based trend discovery of Renewable Energy technologies in Desalination and Power Generation", G. Dawelbait, T. Mezher, T., W.L. Woon and A. Henschel, in Proc. *Portland International Conference on Management of Engineering and Technology (PICMET '10)*, Phuket, Thailand, pp. 2768-2775, July 2010. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5603370

"A Framework for Technology Forecasting and Visualization", W.L. Woon, A. Henschel and S.E. Madnick, in Proc. IEEE International Conference on *Innovations in IT,* Al Ain, UAE, pp. 155-159, 2009 http://dx.doi.org/10.1109/IIT.2009.5413768

"Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation", W.L. Woon, A. Henschel and S.E. Madnick, IEEE International Conference on *Innovations in IT*, Al Ain, UAE, pp. 160-164, 2009. http://dx.doi.org/10.1109/IIT.2009.5413365

“Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry" G. Vidican, W.L. Woon and S.E. Madnick, *Advancing the Study of Innovation and Globalization in Organizations (ASIGO),* 2009. Alternative version available online from SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1388222

Research Thesis

“Early Growth Technology Analysis: Case Studies in Solar Energy and Geothermal Energy”, Ayse Firat, EECS Thesis. http://web.mit.edu/smadnick/www/wp/2010-02.pdf

“A Comparison of Taxonomy Generation Techniques Using Bibliometric Methods: Applied to Research Strategy Formulation”, Steven Camina (2010) EECS Thesis. http://web.mit.edu/smadnick/www/wp/2010-01.pdf

“Methods for Bibliometric Analysis of Research: Renewable Energy Case Study” Blaine E. Ziegler (2009) EECS Thesis. http://web.mit.edu/smadnick/www/wp/2009-10.pdf

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.
http://www.techforecast.org/posters/DB_poster_v1.pdf

“Technology Forecasting using Data Mining and Semantics” (2008), W.L. Woon and S.E. Madnick, MIT-Masdar Symposium, poster presentation. http://www.techforecast.org/posters/ilp_poster_v2.pdf

MIT research and working papers

“Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation” (2009), Andreas Henschel, Wei Lee Woon, Thomas Wachter, Stuart Madnick, MIT Sloan Research Paper No. 4758-09 http://ssrn.com/abstract=1478201

“A Framework for Technology Forecasting and Visualization” (2009), Wei Lee Woon, Andreas Henschel, Stuart Madnick, MIT Sloan Research Paper No. 4757-09 http://ssrn.com/abstract=1478054

“Approach and Preliminary Results for Early Growth Technology Analysis” (2009), Blaine Ziegler, Ayse Kaya Firat, Stuart Madnick, Wei Lee Woon, Steven Camina, Clare Li, Erik Fogg, MIT Sloan Research Paper No. 4756-09 http://ssrn.com/abstract=1478001

“Early Growth Technology Analysis” (2009), Blaine Ziegler , Ayse Kaya Firat, Clare Li, Stuart Madnick, Wei Lee Woon, working paper CISL #2009-4 (CISL, Sloan School of Management, MIT). http://web.mit.edu/smadnick/www/wp/2009-04.pdf

“Bibliometric analysis of distributed generation” (2009), W.L. Woon, H. Zeineldin and S.E. Madnick, MIT Sloan Research Paper No. 4730-09. http://ssrn.com/abstract=1373889

“Semantic distances for technology landscape visualization” (2008), W.L. Woon and S.E. Madnick, MIT Sloan Research Paper No. 4711-08 http://ssrn.com/abstract=1256482

“Asymmetric information distances for automated taxonomy construction” (2008), W.L. Woon and S.E. Madnick, MIT Sloan Research Paper No. 4712-08 http://ssrn.com/abstract=1256562

“Technological Forecasting - a Review” (2008), A.K. Firat, S.E. Madnick and W.L. Woon, working paper CISL #2008-15 (CISL, 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, MIT Sloan Research Paper No. 4727-09. http://ssrn.com/abstract=1356047

“Latent Semantic Analysis applied to tech mining” (2008), B. Ziegler, W.L. Woon and S.E. Madnick, MIT Sloan Research Paper No. 4726-09. http://ssrn.com/abstract=1356011

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

APPENDIX A: SUPPLEMENTARY MATERIALS

To help keep the size of this report down, The main sections describe only the overall framework and main findings of the project.

It was originally intended that copies of the relevant case studies would be included in this appendix but given the size of this document, these have now been collected and made available for download at the following web-page:

http://www.techforecast.org/studies

APPENDIX B: ONLINE PUBLICATION DATABASES

This appendix describes the databases used in this project. The sources that will be analyzed are ACM, Compendex, Google Scholar, IEEE Explore, IngentaConnect, Inspec, Scirus, Scopus, SpringerLink, and Web of Science.

ACM

The ACM Digital Library is an extensive collection of all of ACM's journals, magazines, peer-reviewed articles, conference proceedings, ACM SIG Newsletters, and multimedia. It contains the largest full-text archive of articles on computing. This archive contains over two million pages of text, with full-text articles from ACM publications dating back to the1950s, and third-party content with selected archives. 20,000 New full-text articles added each year with 34 Special Interest Groups contributing content. It currently has:

2.0+ Million pages of full-text articles

260,000 Articles

45+ High-impact journals

270+ Conference proceeding titles

2,000+ Conference proceedings volumes

6 ACM magazines (including the flagship Communications of the ACM)

800+ Multimedia files containing audio,video, and more[2].

Compendex

Scope of coverage: Compendex contains a compilation of comprehensive engineering literature databases available for engineers. It currently has 11.3 million records, with over 650,000 new ones added annually, across 190 engineering disciplines gathered from 1970 to the present. 98% of the top 50 U.S. engineering schools currently subscribe to Compendex. New information is gathered weekly from engineering conferences, journals and trade magazines from over 55 countries. Every entry is indexed according to the Engineering Index Thesaurus and indexed according to the precise engineering discipline.

[2] Source: http://portal.acm.org/dl.cfm

Compendex covers topics from several engineering disciplines, including:

Chemical Engineering (15% of Compendex content)

Civil Engineering (15% of Compendex content)

Mining Engineering (12% of Compendex content)

Mechanical Engineering (12% of Compendex content)

Electrical Engineering (35% of Compendex content)

General Engineering (12% of Compendex content)[3]

Keyword/indexing system:

- Controlled terms: keywords related to the article coming from a list of controlled vocabulary composed of agreed-upon technical terms made by the compilers of the database.
- Uncontrolled terms: uncontrolled vocabulary indexing containing terms not in the controlled vocabulary list

Search Customization Options:

- Document type: journal articles, conference articles/proceedings, monograph chapters/reviews, report chapters/reviews, and dissertations.
- Treatment type: application, biographical, economic, experimental, general review, historical, literature review, management aspects, numerical, or theoretical
- Language
- Publication Year

Google Scholar

Scope of coverage: Google Scholar provides peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Note however that Google Scholar is specifically focused on scholarly documents, and everything covered by Google Scholar is also covered by Google[4].

Search Customization Options:

3 Source: http://www.ei.org/compendex

[4] Source: http://scholar.google.com/

- Author
- Publication Field
- Date of Publication
- Subject Area

Keyword/indexing system: Google full-text algorithm

IEEE Xplore

IEEE Xplore is an online resource for accessing scientific and technical publications produced by the Institute of Electrical and Electronics Engineers (IEEE) and its publishing partners. IEEE Xplore provides access to a comprehensive collection of full-text PDF documents comprising the world's most highly cited journals in electrical engineering, computer science, and electronics. The content repository supporting IEEE Xplore contains more than 2 million articles from over 12,000 publications that encompass journals, conference proceedings, and technical standards, with select content dating back to 1893.

IEEE Xplore provides access to content from other publishers through the CrossRef Search feature. IEEE Xplore also facilitates federated searching of major science and technology society digital libraries via its integrated Scitopia® search feature.In addition to journals, conference proceedings, and technical standards content, IEEE Xplore provides access to the IEEE Press book collection[5].

IngentaConnect

IngentaConnect is a website that hosts scholarly books and journals from a range of different publishers. IngentaConnect provides researchers with a comprehensive collection of citation data - some 4 million articles from 11,000 publications and online access to the full text of electronic articles, through online purchase of individual articles, or through subscriptions to publications.[6]

Feature-rich online content offers:

- Reference linking
- Forward citation linking

[5] Source: http://ieeexplore.ieee.org/Xplorehelp/Help_Welcome_to_IEEE_Xplore.html
[6] Source: http://www.ingentaconnect.com/

- Supplementary Data
- FastTrack articles (pre-publication)

Inspec

Scope of coverage: Inspec is an abstracting and indexing database for physics, electrical engineering, electronics and computer science information. Updated weekly, it currently has 11 million specially-selected records gathered from 1969 to the present that are precise, targeted and relevant, with over 600,000 new records added annually[7]. Content is available for:

Physics (47% of Inspec content)

Electronics & Electrical Engineering (26% of Inspec content)

Computing & Control (20% of Inspec content)

Manufacturing & Production Engineering (5% of Inspec content)

Information Technology, Networking and Security (2% of Inspec content)

Keyword/indexing system:

- Controlled terms: keywords related to the article coming from a list of controlled vocabulary composed of agreed-upon technical terms made by the compilers of the database.
- Uncontrolled terms: uncontrolled vocabulary indexing containing terms not in the controlled vocabulary list

Search Customization Options:

- Document type: journal articles, conference articles/proceedings, monograph chapters/reviews, report chapters/reviews, and dissertations.
- Treatment type: application, biographical, economic, experimental, general review, historical, literature review, management aspects, numerical, or theoretical
- Language
- Publication Year

[7] Source: http://www.ei.org/inspec_inspecarchive

Scirus

Scope of coverage: Scirus contains scientific topics in found web sites, news, journals, web articles and academic papers. It searches over 485 million science-specific web pages, filtering out non-scientific sites, and finds peer-reviewed articles such as pdf and postscript files. Scirus searches the most comprehensive combination of web information, preprint servers, digital archives, repositories and patent and journal databases.

Scirus currently covers over web pages including156 million .edu sites, 54 million .org sites, 9 million .ac.uk sites, 52 million .com sites, 36 million .gov sites, and over 143 million other relevant STM and University sites from around the world[8].

Scirus also indexes these sources (the numbers are approximate):

447,000 articles from American Physical Society

536,000 e-prints from ArXiv.org

42,000 full-text articles from BioMed Central

19,000 documents from Caltech Coda

3,300 e-prints from Cogprints

81,800 full-text articles from Crystallography Journals Online

24,000 documents from CURATOR

2.1 million documents from Digital Archives

24,000 documents from DiVa

98,500 full-text articles from Project Euclid

3,200 documents from HKUST Institutional Repository

56,000 documents from The University of Hong Kong

12,700 full-text documents available from IISc

11,000 full-text documents available from Humboldt Universität

[8] Source: http://www.scirus.com/srsapp/aboutus/

284,000 full-text articles from Institute of Physics Publishing

23.1 million patent data from LexisNexis

16,000 full-text articles from Maney Publishing

40,000 full-text documents from MD Consult

585,000 full-text documents from Nature Publishing Group

18.1 million Medline citations via PubMed

72,000 documents from MIT OpenCourseWare

24,700 technical reports from NASA

792,000 full-text theses and dissertations via NDLTD

8,900 documents from Organic Eprints

1,690 documents from PsyDok

1.5 million articles from PubMed Central

738,000 documents from RePEc

63,000 full-text articles from Royal Society Publishing

619,000 full-text articles from SAGE Publications

8.0 million full-text articles from ScienceDirect

463,000 full-text journal articles from Scitation

9,100 articles from SIAM

16,600 documents from University of Toronto T-Space

21,800 full-text documents from WaY.

Keyword/indexing system: 'Keywords' can be seen in the 'refine your search' box in the lower left side of the website. Scirus uses an automated extraction algorithm to calculate ranking by relevance. This ranking is determined by two basic values:

- Words - the location and frequency of a search term within a result account for one half of the algorithm. This is known as static ranking.

- Links - the number of links to a page account for the second half of the algorithm - the more often a page is referred to by other pages, the higher it is ranked. This is known as dynamic ranking. Overall ranking is the weighted sum of the static and dynamic rank values. Scirus does not use metatags, as these are subject to ranking-tweaking by users.

Search Customization Options:

- Information Types: abstracts, articles, books, company homepages, conferences, patents, preprints, scientist homepages, theses/dissertations
- Content Sources
- Subject Areas

Scopus

Scope of coverage: Scopus is a large abstract and citation database consisting of both peer-reviewed research literature, patents and quality web sources. It is part of the "SciVerse" suite (by Elsevier), and covers over 18,000 titles from more than 5,000 international publishers, spanning topics in the health, physical, social and life sciences. The approximate distribution of titles amongst these subject areas are as follows ():

- Health sciences ~33%
- Physical sciences ~31%
- Social sciences ~20.5%
- Life sciences ~15.5%

Keyword/indexing system:

- Author keywords: Keywords specified by the authors (for e.g. as part of the submission process)
- Index terms: These are keywords selected from a controlled vocabulary and assigned to the document

Search Customization Options: Scopus permits a very large number of customization options to be specified when its database is searched. Available search keys include:

Authors, Title, Year, Source title, Volume, Issue,Art. No.,Page start, Page end, Page count, Cited by, Link, Affiliations, Authors with affiliations, Abstract, Author Keywords, Index Keywords, Molecular Sequence Numbers, Chemicals/CAS, Tradenames, Manufacturers, Funding Details, References, Correspondence Address, Editors, Sponsors, Publisher, Conference name, Conference date, Conference location, Conference code, ISSN, ISBN, CODEN, DOI, PubMed ID, Language of Original Document, Abbreviated Source Title, Document Type, Source

Springerlink

Springerlink covers topics in Architecture, Life Science, Behavior Science, Business/Econ, Chemistry/Materials, Computer Science, Environmental Science, Engineering, Humanities, Social Science, Law, Math/Statistics, Medicine, Physics, Astronomy, and Applied Computing from journals, books, reference works, protocols, academic publications. It contains over 1,750 peer reviewed journals and 27,000 eBooks online. 3,500 eBooks, eReference Works and eBook Series titles are scheduled to be added each year.

Keywords are pulled from publishers, some articles have none. Search uses frequency analysis as well as keywords. Search customization options include the title, the author, the editor, ISSN / ISBN / DOI and the date of publication[9].

Web of Science

Web of Science has authoritative, multidisciplinary content that covers over 10,000 of the highest impact journals worldwide, including Open Access journals and over 110,000 conference proceedings. Topics in agriculture, biological sciences, engineering, medical and life sciences, physical and chemical sciences, anthropology, law, library sciences, architecture, dance, music, film, and theater with coverage available to 1900. It contains articles, proceedings, papers, reviews, editorials, news[10].

Web of Science offers access to six comprehensive citation databases:

- Science Citation Index Expanded: Over 7,100 major journals across 150 disciplines, to 1900.
- Social Sciences Citation Index: Over 2,474 journals across 50 social science disciplines, as well as 3,500 of the world's leading scientific and technical journals, to 1956.
- Arts & Humanities Citation Index: Over 1,395 arts and humanities journals, as well as selected items from over 6,000 scientific and social sciences journals.
- Conference Proceedings Citation Index: Over 110,000 journals and book-based proceedings in two editions: Science and Social Science and Humanities, across 256 disciplines.

[9] Source: http://www.springer.com/e-content?SGWID=0-113-12-286799-0

[10] Source: http://wos.isitrial.com/help/helpdefs.html

- Index Chemicus: Over 2.6 million compounds, to 1993.
- Current Chemical Reactions: Over one million reactions, to 1986, plus INPI archives from 1840 to 1985.

Search Customization Options:

- Topic
- Title
- Author
- Publication Name
- Year Published
- Address
- Time past since publication