Technology Forecasting Using Data Mining and Semantics: Second Annual Report

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

Working Paper CISL# 2009-15

December 2009

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

Technology Forecasting Using Data Mining and Semantics

MIT/MIST Collaborative Research Progress Report for Period 10/1/07 to 9/30/09

Principal Investigator at MIT

Professor Stuart Madnick

Principal Investigator at MIST

Dr. Wei Lee Woon

Team-members

Andreas Henschel (MIST Postdoctoral Researcher)

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

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

Steven Camina (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, Colalborator)

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

Research Project Start Date

10/01/2007

EXECUTIVE SUMMARY

The planning and management of research and development is a challenging process, further compounded by the ever increasing rate of technological progress. 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 corresponding increases in the volume of relevant information that is available.

Also of concern is the 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 databases for patterns and trends which can facilitate the formation of research strategies. Examples of the types of information sources which are available are very diverse and could include academic journals, patents, blogs and news stories. However, at the current stage of the research project, the focus has been on academic publications though we have also begun to consider the case of blog postings.

It must be emphasized that the goal of this research is not to replace the important role of experts in making technology 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 proposed outputs of the project were:

1. A detailed **case study** of the renewable energy domain, including tentative forecasts of areas of future growth potential and 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, the research activities in this project have centered on the development of a comprehensive framework which incorporates all the required stages of technology forecasting, starting with

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 little known and as such it is often difficult to measure their prevalence accurately due to the comparatively low publication volumes. 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.

Review of Objectives and Accomplishments

All of the original goals of this project have either been accomplished or are well on schedule. In addition, several additional outcomes have emerged during the execution of the research project:

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.
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. Two 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.
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 modest sized case studies have been completed, a more comprehensive case study is still underway and will be the focus of the remaining three months of the project.
5. **Development of software tools** - Two user-friendly main tools were developed – the "Cameleon Scheduler" and the "Early Growth Analysis tool." In addition, numerous programs and sub-programs were developed, for use by researchers, to test various algorithms under study.
6. **Collaborations with other MIST faculty** - Collaborations were developed with Dr. Georgeta Vidican, Dr. Hatem Zeineldin, and Dr. Toufic Mezher and MIT PhD student Satwik Seshasai. Most of which have resulted in research results in the form of papers submitted to conferences or journals or Working Papers.
7. **Completed 13 research papers** - of which 2 have been published, 2 are under review, 2 have been amended and are to be re-submitted in the very near future and 8 distributed as MIT Working Papers on the Social Sciences Research Network (SSRN.)

Work to be accomplished between now and 31 December 2009, the remaining months of this second phase of the joint MIST-MIT research effort

From now until the formal end date of this second phase of the project we will be focusing on applying the tools described above to the creation of a more comprehensive renewable energy case study. While we have already conducted a number of limited pilot studies, as well as tested the proposed methods on data relevant to renewable energy, these efforts have hitherto been largely experimental. For almost every stage of the technology forecasting framework, we have tested more than one alternative technique which, in most cases, is further customizable via a variety of parameters and settings. As such, the main challenge left will be to select the best possible set of parameters and to apply these towards the compilation of the case study. A further task will be the collection of extra data where appropriate in cases where there is insufficient coverage of promising keywords and technologies. We have every confidence that this undertaking will be a success and look forward to the successful achievement of all of the project goals.

We hope that this report will convey the elegance of the system, particularly in terms of the highly integrated manner in which it achieves its aims. The techniques and methods for visualizing and understanding domains of research are closely tied in with the subsequent derivation of numerical features for technology prevalence and finally the detection of early growth technologies. It will also be seen that the various components of the system work together closely to help achieve a more reliable prediction of technological growth potential.

At this point it is also important to differentiate between the ***framework*** as a generic system based on a set of interchangeable software or mathematical techniques, and the ***implementation*** of the framework, which is a particular instance of the framework implemented in a specific manner and with an appropriate set of tools, data sources, and platforms. The contribution of this project is both the overall concept of the ***framework***, as well as a ***reference implementation*** which demonstrates the effectiveness of this direction of research.

Report overview

The rest of this report will be structured as follows.

The ***Introduction*** section reviews the key ideas and motivations for the project.

The ***Research Tasks*** section comprises the bulk of the report, and is organized into several subsections. However, unlike in the previous research reports, which reviewed the progress based on the project schedule, this section will seek to sum up the overall activities and directions of the research project, the relevance of each of these to the stated goals and deliverables specified in the project proposal and finally the degree to which these have been achieved.

In particular, the focus will be on the technology forecasting *framework* mentioned above

The ***Current Reporting Period Summary*** section reviews and discusses the findings presented in this report and specifies the relative divisions of labor between the teams at MIT and MIST.

Finally, there is a section on ***Future Work***, which in the current report refers to the final activities between now and the conclusion of the project duration.

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. Any given research field is composed of many subfields and underlying technologies which are related in intricate ways. This composition, or *research landscape* is not static as new technologies are constantly developed while existing ones become obsolete, often over very short periods of time. Fields that are presently unrelated may one day become dependent on each other's findings.

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 dreciding NSF and NIH grant awards [Porter, 07], while many countries spend huge sums on technology foresight programs [Eto, 03][Bengisu and Nekhili, 2006]. Expert opinion is a hugely important component in the decision making process; however when used on its own, it can have a number of shortcomings. In particular, expert decisions are subjective and can be influenced by personal perspectives or biases. In addition, it is difficult to systematically record the reasons for such decisions, or the contexts in which decisions were made. Finally, it can also be difficult and expensive to obtain the help of suitably qualified experts.

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

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 which 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 "tech-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 including various modelling activities.

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

In response to this apparent shortcoming, 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 subtrees of the taxonomy.
2. A set of numerical indicators for identifying technologies of interest. In particular, we are interested in developing a set of simple growth indicators, similar to technical indicators used in finance, 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

A framework for technology forecasting

As mentioned in the executive summary, this report will focus on our efforts to create a comprehensive framework which integrates all of the previous research efforts in this project. In previous reports, we described various techniques which could be used for extracting keywords which were representative of a domain of research, generating representations of the research landscape for these domains, extracting and modelling the publication counts for research terms and for estimating the "early growth" potentials of particular keywords.

The emphasis now is to show how each of these project components are in fact part of a larger framework which will use their collective capabilities to produce a more reliable forecast of the future evolution and growth of a domain of research.

Before proceeding further, it is important to define the form of forecasting that is intended. In particular, it must be stressed that it is not "forecasting" in the sense of a weather forecast, where specific future

outcomes are intended to be predicted with a reasonably high degree of certainty. It is also worth noting that certain tasks remain better suited to human experts; in particular, where a technology of interest has already been identified or is well known, we believe that a traditional review of the literature and of the technical merits of the technology would prove superior to an automated approach.

Instead, the proposed framework targets the preliminary stages of the research planning exercise by focussing on what computational approaches excel at: i.e. scanning and digesting large collections of data, detecting promising but less obvious trends and bringing these to the attention of a human expert. This overall goal should be borne in mind as, in the following subsections, we present and describe the individual components which constitute the framework.

Overview

Figure 1 depicts the high-level organization of the system. As can be seen, the aim is to build a comprehensive technology analysis tool which will collect data, extract relevant terms and statistics, calculate growth indicators and finally integrating these with the keyword taxonomies to produce actionable outcomes.

To facilitate discussion, the system has been divided into three segments:

1. Data collection and term extraction (labelled **(a)** in the figure)
2. Prevalence estimation and calculation of growth indicators (labelled **(b)**)
3. Taxonomy generation and integration with growth indicators (labelled **(c)**)

These components are explained in the following three subsections.

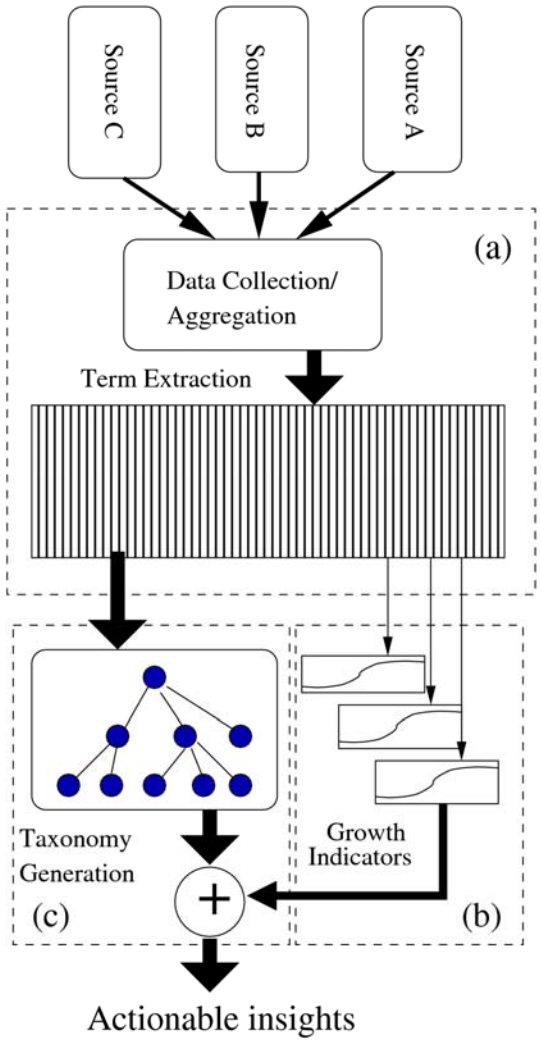


Figure 1: Proposed framework: overall structure

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. 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.

1http://www.scopus.com

Term extraction

Term extraction is the process of automatically generating a list of keywords on which the technology forecasting efforts will be focussed. Again, there are a variety of ways in which this can be achieved; we have experimented with a number of these and our experiences have been documented in [Ziegler et al., 2009], but in brief, the two main approaches which were tested were:

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 ensur e 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.

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 wasts*}

For the example results shown later in the report, a total of 900 keywords were extracted and used to build the taxonomy which will be displayed.

Case study

As stated previously, one of the aims of the project is to conduct a case study on the domain of renewable energy technology. The incredible diversity of renewable energy research, the obvious importance to the

well-being of society and of course its central position in the vision of the Masdar Initiative provide clear motivations for conducting this case study.

Besides high-profile topics like solar cells and nuclear energy, renewable energy related research is also being conducted in fields like molecular genetics and nanotechnology, making it a rich and challenging domain on which our methods may be aptly tested.

To collect the data for use in this study, the following renewable-energy related keywords (organized into a number of themes) were submitted to Scopus:

Themes	**Keywords**	**Search type (#hits)**
Renewable energy (general)	renewable energy	Title, Abstract, Keyword (18,213)
Wind	wind energy, wind power	Title, Abstract, Keyword (17,656)
Geothermal	geothermal	Title, Abstract, Keyword (20,552)
Distributed generation	distributed generation, dispersed generation, distributed resources, embedded generation, decentralized generation, decentral ized energy, distributed energy, on-site generation	Title, Abstract, Keyword (5,436)
Biofuels	biofuel, biodiesel	Title, Abstract, Keyword (11,216)
Energy policy	Energy policy	Title, Abstract, Keyword (15,724)
Fuel cell	fuel cell	Title (20,206)
PV	photovoltaic, solar cell	Title, Keyword (48,245)

Table 1: Keywords used for data collection

Records for each category were downloaded separately, stored in local databases and finally merged to create a genera l database consisting of a total of 153,537 records.

Block (b): Identification of early growth technologies

There are actually two steps to this activity. The first is to find a suitable measure for the “prevalence” of a given technology as a function of time. In the context of a database of academic publications, this would be some means of measuring the size of the body of relevant publications appearing each year. It is difficult to achieve this directly but an alternative would be to search for 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. This is defined as:

$$TF_i = \frac{n_i}{\sum_{j \in \chi} n_j} \quad (1)$$

where n_i is the number of occurrences of keyword i, and I is the set of terms appearing in all article abstracts (this statistic is calculated for each year of publication to obtain a time-indexed value). Once the term frequencies for all terms have been extracted and saved, they can be used to calculate growth indicators for each of the keywords (and, by extension, the associated technologies). These, in turn, are used to rank the list of terms.

As stated previously, we are most 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, depicted in Figure 2, 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.

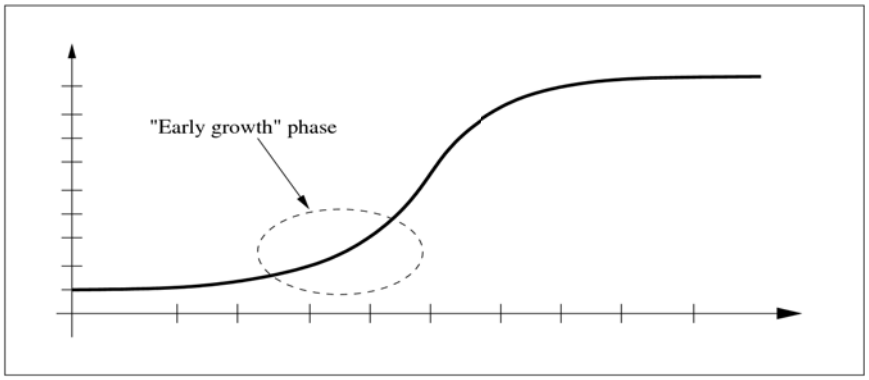


Figure 2: Early growth of technological development

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]):

1. ***Mean publication year –*** calculated ove r 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.

$$\mu_i(t_0, t_n) = \frac{\sum_{t=t_0}^{t_n} t.TF_i[t]}{\sum_{t=t_0}^{t_n} TF_i[t]} \quad (2)$$

2. ***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.

$$\theta_i(t_0, t_n) = \log(TF_i[t_n]) - \log(TF_i[t_0]) \quad (3)$$

3. ***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 TF_i :

$$\dot{\theta}_i(t_0, t_n) = \theta_i(t_{n/2}, t_n) - \theta_i(t_0, t_{n/2}) \quad (4)$$

where $\mu_i(t_0, t_n), \theta_i(t_0, t_n)$ and $\dot{\theta}_i(t_0, t_n)$ 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 last five yeard (the year 2009 is exluded), and which are relevant to term *i* (a more recent year indicates greater currency of the topic).

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.

Taxonomy generation

The question remains, how do we obtain such a taxonomy? In a limited number of cases, these taxonomies may be available from external sources such as government agencies and other manually curated sources. However, in many cases, a suitable taxonomy is either unavailable, or is available but is not sufficiently updated to be of use for the application at hand. As such, to make our framework broadly applicable, an important research direction is the *automated* creation of keyword taxonomies based on the statistics of term occurrences.

As indicated in a previous section, the basic idea is to group together terms which tend to co-occur frequently. Again, we have tested a number of different ways of achieving this (two earlier attempts are

described in [Woon and Madnick, 2009, Woon and Madnick, 2008]) though we will not be comparing these in depth here. Instead, we present one particular method which was found to produce reasonable results while being scalable to large collections of keywords. This is based on the algorithm described in [Heymann and Garcia-Molina, 2006] which was originally intended for social networks where users annotate documents or images with keywords. Each keyword or tag is associated with a vector that contains the annotation frequencies for all documents, and which is then comparable, for e.g. by using the cosine similarity measure.

We adapt the algorithm to general taxonomy creation by adopting two important modifications; firstly, instead of using the cosine similarity function, the *asymmetric* distance function proposed in [Woon and Madnick, 2009] is used (this is based on the "Google distance" proposed in [Cilibrasi and Vitányi, 2007]):

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

where and are the two terms being considered, and , and are the occurrence counts for the two terms occurring individually, then together in the same document respectively. Note that the above expression is "asymmetric" in that $\overrightarrow{NGD}(t_x, t_y)$ refers to the associated cost if t_x is classified as a subclass of t_y, while $\overrightarrow{NGD}(t_y, t_x)$, corresponds to the inverse relationship between the terms.

The algorithm consists of two stages: the first is to create a similarity graph of keywords, from which a measure of "centrality" is derived for each node. Next, the taxonomy is grown by inserting the keywords in order of decreasing centrality. In this order, each unassigned node, , is attached to one of the existing nodes such that:

$$j = \underset{j \in \tau}{argmin}\, \overrightarrow{NGD}(t_i, t_j) \qquad (6)$$

(where $\mathcal{T}$ is the set of terms which have already been incorporated into the taxonomy.)

Given the critical importance of the taxonomy generation procedure, a small study was conducted to evaluate the accuracy of the automatically generated taxonomies. This will described in the following section.

An evaluation of automatically-generated taxonomies

To assess the quality of automated taxonomy generation methods some gold standard taxonomy is required. Medical Subject Headings (MeSH) is a human-curated ontology for medical terms proposed in [U.S. Dept. of Health,], and is well suited for use as such a gold standard. We focus on several diverse branches in order to avoid overfitting. For the automatic comparison of a manually and automatically generated taxonomies, the input terms are taken from the MeSH benchmarks (admittedly this poses a simplification of the overall taxonomy creation process, where terms are selected using various methods, see e.g.[Frantzi et al., 2000, Alexopoulou et al., 2008] for instance.)

We then measure the accuracy by counting how many direct links of the original taxonomy are reproduced by the algorithm. Further we consider those links that are not only direct parent-child related but also grandchildren or great-grandchildren (dotted lines in Figure 3) in the original benchmark.

Occurrences are detected in the abstracts of 18 Million articles (a literature database for the life sciences), using GoPubmed annotations ([Doms and Schroeder, 2005]).

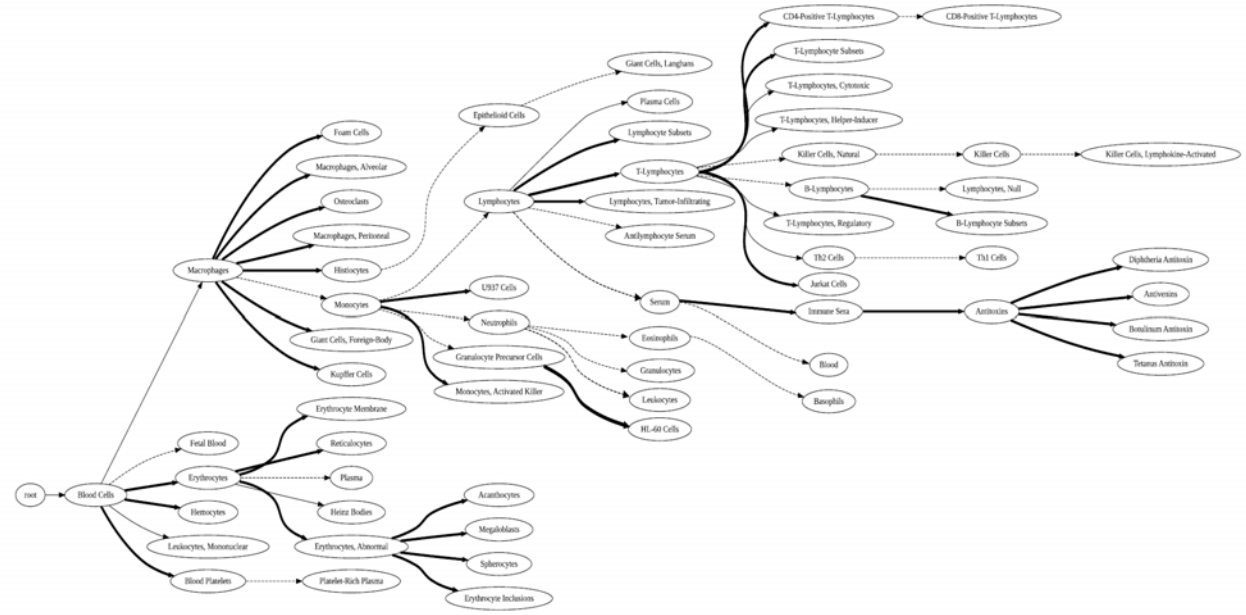


Figure 3: A generated taxonomy for "Blood". The fat links are correct wrt. the MeSH benchmark, semi-fat links are in grand- or great-grandchild relation in MeSH.

Heymann-Algorithm

The taxonomy creation algorithm presented in [Heymann and Garcia-Molina, 2006] (Heymann-Algorithm) was originally intended for social networks where users annotate documents or images with keywords. The algorithm is fast, deterministic and easily extensible. Each keyword or "tag" is associated with a vector that contains the frequencies of annotations for all documents. These tag vectors are then comparable, for e.g. by using cosine similarity. We adapt the algorithm to general taxonomy creation from scientific literature using binary tag vectors.

The algorithm consists of two stages: the first creates a similarity graph of tags, from which an order of centrality for the tags is derived. Obeying this order and starting from the most general tag, the tags are inserted to a growing taxonomy by attaching tags to either the most similar tag or to the taxonomy root.

Several aspects can be modified to boost the algorithm: generality ordering, measures, similarity measures and weight functions insertion of new nodes. Two thresholds are used in the algorithm: first, the value above which an edge is permitted to the similarity graph (τ_s) filters very small similarities that might have occured by chance during the generality calculation. Second, the similarity above which a node is attached to its most similar non-root node rather than the root (τ_R) influences the topology of the taxonomy.

Term generality derived from centrality in similarity graphs

A set of *n* terms gives rise to a similarity graph $G=(V,E)$ where the nodes represent terms and the edges E are similarities as provided by the similarity measure. A variety of centrality measures exist. Amongst them betweenness and closeness centrality are elaborate, global measures and therefore subject to further scrutiny.

The ***betweenness centrality*** c_b for a node v is defined as:

$$c_b = \frac{\sum_{v \in V} \sigma_{st}(n)}{\sigma_{st}} \qquad (7)$$

where is the number of shortest paths from *s* to *t*, and is the number of shortest paths from s to t that pass through a vertex v. A fast algorithm () for unweighted graphs is given in [Brandes, 2001] and implemented e.g. in [Hagberg et al., 2008].

The ***closeness centrality*** c_b for a node v is given as:

$$c_C = \frac{1}{\sum_{t \in V \setminus \{v\}} 1 - similarity(v,t)} \qquad (8)$$

If and only the final order of generality is relevant, it is sufficient to calculate the centrality as sum of all similarities. The complexity is $O(n^2)$.

Considering graph-theoretical aspects: Edge weights and disconnected graphs

Betweenness and closeness centrality can be calculated using weighted or unweighted graphs. We investigate both types.

Since centrality is well defined for connected graphs and a high τ_S yields disconnected graphs, centrality measures can yield unexpected results (high centrality values for two-node components). We therefore order centrality lexicographically by first considering the membership to the size of the component (ranking the members of the largest component highest) and second its actual centrality value.

Figure shows the comparison of the Heymann-Algorithm with several centrality calculations in dependence. Various values for τ_R are probed and shown as seperate curves with different symbols but are of less influence.

Distance measures

Originally [Heymann and Garcia-Molina, 2006] used vectors of length equal to the number of documents *N*, where describes, how many times a numbered document *i* in a user community has been annotated with term *t*. We adapt this to binary term-vectors (or set representations) indicating whether a term occurs in a document (1) or not (0).

$$s_{\cos}(x,y) = \frac{x.y}{\|x\|.\|y\|} \qquad (9)$$

where **x** and **x** are binary term vectors. Hence, the similarity between two terms is simply the dot product of its normalized term vectors.

Reranking

A further modification to the Heymann algorithm is the intermediate reranking of the remaining terms w.r.t. their centrality after inserting a term into the taxonomy. Note, that this step increases the algorithm complexity since the centrality calculation is run for every inserted term (and for closeness and betweenness centrality, resp.). Depending on the similarity graph threshold , the intermediate reranking improves performance in 46% of the cases, decreases accuracy in 23% and achieves equal accuracy in 30% of the cases.

Entropy of similarities

According to [Widdows, 2003] taxonomy generation algorithms usually achieve only 40-50% precision on general benchmarks, while [Velardi et al., 2007] therefore suggests a semi-automatic approach which includes systematic human validation steps. We note that the basic Heymann algorithm attaches nodes to the most similar node in the growing taxonomy; in many cases, non-specific or ambiguous terms exhibit similarities to many subjects yet the highest-similarity condition imposed by the Heymann algorithm forces the assignment of such nodes to only a single parent. This is a problem but also presents an opportunity to incorporate some form of confidence-based filtering mechanism into the taxonomy generation procedure.

Entropy is an information theoretical concept that can be used to quantify the above intuition by providing a numerical indication of the uncertainty of adding a node. This can then be used to annotate individual edges and for quality assessment and semi-automatic curation. Entropy is defined as:

$$E_S(j) = -\sum_{i \in T}^{n} s_{ij} \log_b s_{ij} \; for \; s_{ij} > 0 \qquad (10)$$

where s_{ij} are the similarities between nodes *i* and *j* and *T* is the set of nodes which have already been incorporated into the taxnomy. Similarities are normalised such that they sum to 1. Thus, a node that is similar to exactly one node but with 0 similarity to all other nodes would have an entropy of 0, whereas a maximal entropy of 1 is reached when all nodes in the taxonomy are equally similar to the node to be inserted.

Results

The benchmark sets were scrutinised with respect to algorithm variants (centrality, rooting threshold τ_R , similarity graph threshold τ_S). One example is given in Figure 4, which depicts the relationship between precision (*y-axis*) and τ_S (*x-axis*) for the "Carbohydrates" subtree of the MeSH taxonomy, while similar graphs have been produced for other subtrees such as "Fungi" (Figure 5) and "Sense organs" (Figure 6). While there is some variability between individual subtrees, the results are broadly consistent and indicate that betweenness centrality yields the best results for $0 < \tau_S < 0.1$.

The threshold for attaching a term to the root has been systematically probed and best results were commonly achieved with a very small value, i.e. avoiding to attach to the root as much as possible. For

most benchmarks the best results were achieved with betweenness centrality around that threshold. Since the precision-graphs are commonly smooth, a hill climbing approach or simulated annealing around that value seems most promising.

The threshold for attaching a term to the root R has been systematically probed in the range of 0 - 0.06 with a stepsize of 0.005, and the best results were consistently achieved with a very small value, i.e. avoiding node-attachments to the root as much as possible (in the fact in the actual tests for which results are presented later in this report, we make the assumption that $\tau_R = 0$, thus ensuring that all nodes originate from a common super-node). Note that a histogram of all similarities revealed that most similarities are below 0.01.

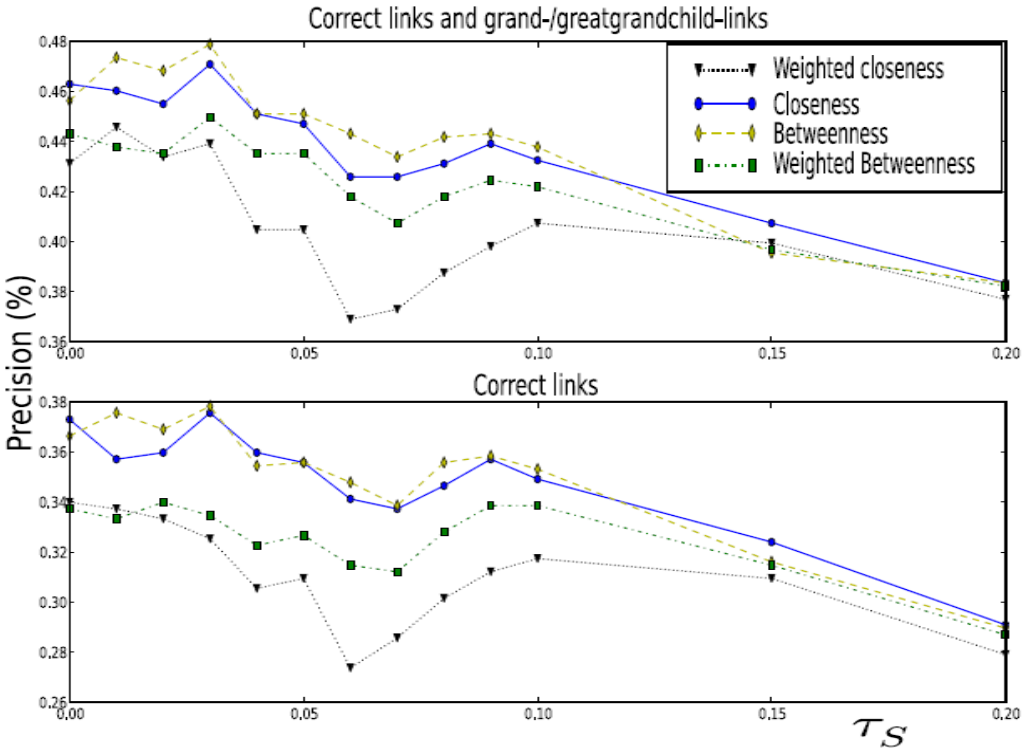


Figure 4: Precision for the MeSH-benchmark "Carbohydrates".

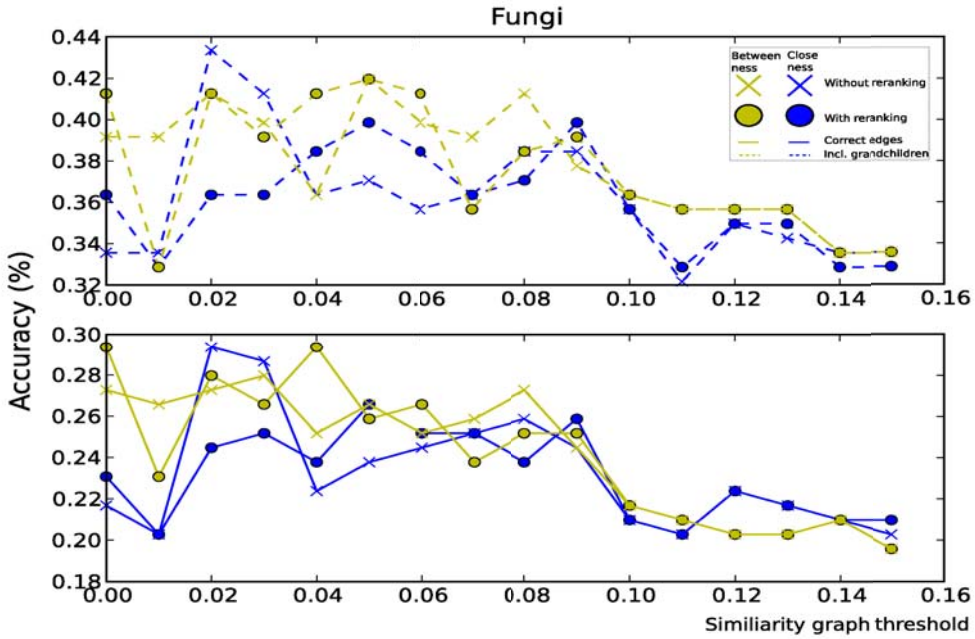


Figure 5: Benchmark: Fungi, with and without generality reranking between insertions

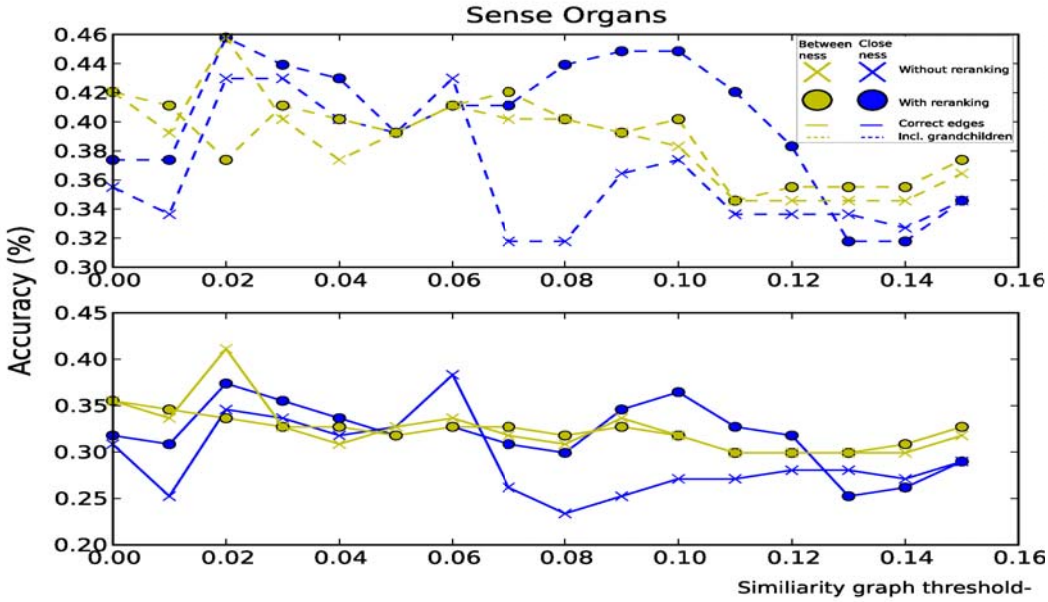


Figure 6: Benchmark: Sense Organs, with and without generality reranking between insertions

Entropy based filtering improves precision

According to [Widdows, 2003] taxonomy generation algorithms usually achieve only 40-50% precision on general benchmarks. [Velardi et al., 2007] therefore suggests a semi-automatic approach which includes systematic human validation steps. In order to serve as a basis for hand-curated taxonomies, the precision (as compared to just the recall) of automatically generated draft taxonomies is of paramount importance. In addition, we also monitor the *F-measure*, which is frequently used in information theory, as a means of tracking the precision-recall trade-off. This is defined as:

$$F_\beta = (1+\beta)^2 \frac{(precision.recall)}{\beta^2(precision+recall)} \qquad (11)$$

To emphasize the importance of precision, the F-measure for example values precision as twice as important as recall. Omitting links comes to the expense of decreasing the recall.

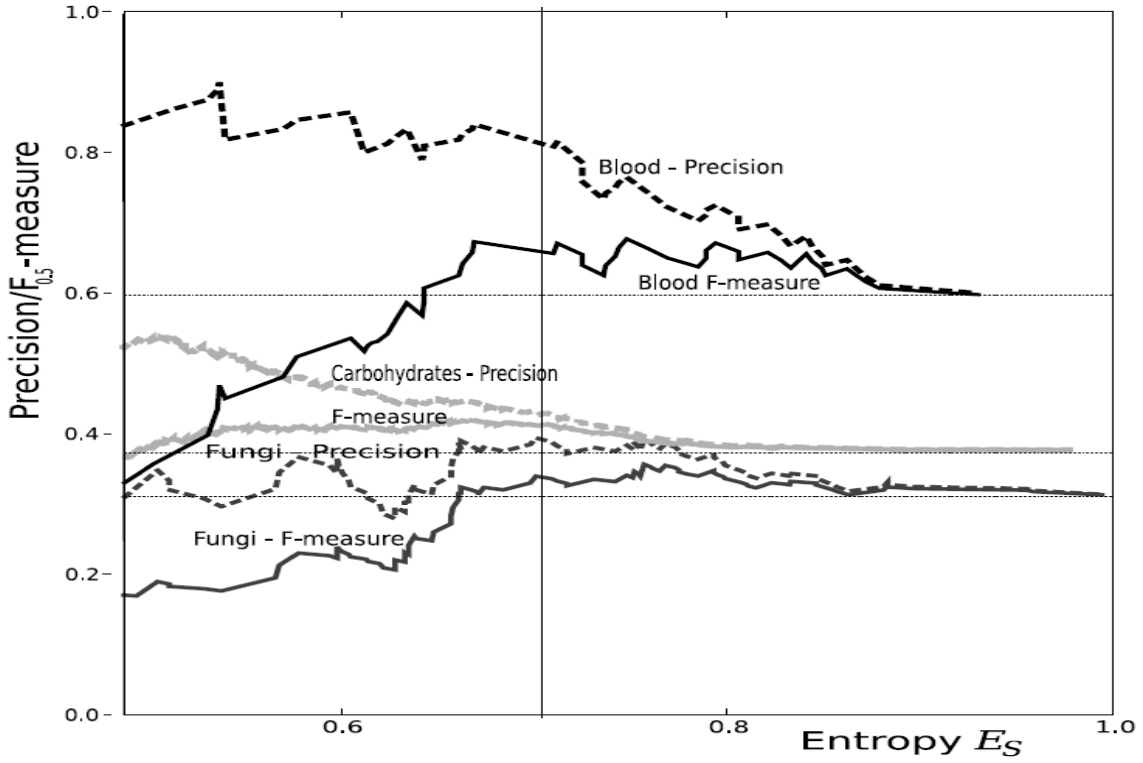


Figure 7: The figure shows the F-measure (solid) and the precision (dotted) for the seven MeSH benchmarks. Higher entropy of similarities expresses lower confidenence in a taxonomy-link. Not filtering by entropy at all yield in precision and F-measure equal to the rightmost data point of each curve, indicated by horizontal lines. The figure shows that indeed high entropy links are often wrong and precision decreases for all benchmark sets. Therefore, by filtering these low confidence links, the algorithm improves in terms of precision, while maintaining or slightly improving the F-measure. Any threshold above 0.7 increases precision without worsening the F-measure.

By filtering high entropy links with , precision increases most notably for benchmark “Blood” (from 60% to 81%), “Carbohydrates” (from 38% to 43%) and “Fungi” (from 31% to 39%).

The precision of all other benchmarks improves as well, but to a smaller extend. Larger margins are possible with other thresholds but might yield in overfitting to the given benchmarks. A particular example where taxonomy creation and improvement is difficult is the case of “Cardiovascular Systems”. This benchmark emphasises the intrinsic difficulty of co-occurrence based similarity measures:

Observations

To conclude this section, we note that taxonomies inferred using term co-occurrence statistics can provide a useful alternative to manually-curated taxonomies, or could at least serve to create “scaffolds” for more extensive taxonomy creation efforts as well as being easily extractable from literature databases. While the accuracies obtained may not seem particularly high they are comparable to the levels reported in [Widdows, 2003] and in any case due consideration must be given to the inherent subjectivity of taxonomies.

In more detail, it was found that unweighted betweenness centrality generally performs best but often only marginally better than the faster unweighted closeness centrality. Neither method strictly dominates the other and both are dependent on fine-tuning of the similarity graph threshold. A good choice for τ_S is not obvious but should be a value between 0 and 0.1. Both methods are complementary in the sense that their highest scoring taxonomies are not identical. A consensus-based meta-algorithm can benefit from this fact by only including the links both methods agree on.

Using weighted similarity graphs rarely improved the performance and hence did not justify the higher computational cost. Moreover, they fluctuate stronger wrt. τ_S . Reranking the centrality often improves the algorithm performance but increases the computational expense. Finally the proposed entropy-based filter for edges allows to shift focus towards more precise (but less complete) taxonomies which arguably facilitates manual post-processing.

Co-occurrence based similarity measures of terms are easily extractable from literature databases and can provide a scaffold for taxonomy creation. However, they also limit the success of taxonomy creation when dealing with semantically related terms that can not be ordered by generality: High-level terms such as ``wind power" or ``solar energy", or terms that somehow interact (e.g., ``hammer" and ``nail") frequently co-occur and hence exhibit a misleadingly high co-occurrence similarity.

Yet neither are subsumable in the strict sense (``is-a" or ``part-of" relations) of standard taxonomies. As a result, the semantics of taxonomy sub- and superconcepts merely allows the interpretation as ``is-related-to" relation. Such a property is not transitive and hence less useful for purposes, where complete semantic subtrees of the taxonomy are required. As a remedy, it would be beneficial to incorporate more sophisticated similarity and generality measures using Natural language processing techniques as proposed in [Ryu and Choi, 2006]. To this end it seems most promising to devise a meta-algorithm, for which the Heymann algorithm is a suitable platform.

Enhancing early growth indicators using keyword taxonomies

The taxonomies created in the previous section are already useful for visualization of research domains. However, even more importantly they also provide a straightforward method for enhancing the early growth indicators using information regarding the co-occurrence statistics of keywords within the document corpus. As with almost the other aspects of the proposed framework, there are a number of

ways in which this may be implemented but the basic idea is to re-calculate the early growth scores for each keyword based on the aggregate scores of each of the keywords contained in the subtree rooted in the corresponding node in the taxonomy.

For the reference implementation described in this report, the aggregation operation used was a straight average, though other more elaborate schemes are clearly possible.

Results and discussions

We present results for a simple pilot study in renewable energy. As described in section Data collection, the Scopus database was used to collect a total of 500 keywords which were relevant to the renewable energy domain, along with 153,537 document abstracts. These keywords were then used to construct a taxonomy as described in section Taxonomy generation, and the three early-growth indicators described in Equations (2) to (4) were used to evaluate each node. These were then re-calcaluted via aggregation based on the keyword taxonomy. Finally, the list of keywords was sorted according to order of decreasing publication year. Using this method of evaluation, the top 20 keywords for each growth measure are shown in Table 2 below.

A few observations from this table were:

1. One interesting observation is the number of biology and biochemistry related keywords in this list. This indicates that biological aspects of renewable energy are amongst the most rapidly growing areas of research. The term “concentration (composition)” is rated #1 in both the mean-publishing year and log-growth categories
2. The highest-rated non-biological term on the list was “seminconducting zinc compounds” (positions #2, #5 and #3 on the three ranking lists respectively), which are related to the field of thin-film photovoltaics.
3. There were also a number of terms related to wireless sensor networks and communications such as “telecommunication equipment”, “telecommunication systems”, “wireless sensor networks” and “wireless network” but these seemed to *only* be present in the first two rankings, indicating that these areas could be growing well but are possibly not in the *early growth* stage .
4. In general, the mean publishing year and log-growth indicators showed a reasonable degree of consistency (13 terms in common in the top 20), while the second order growth indicator appeared to give somewhat different results from the other two indicators. This is to be expected since we would expect this to measure a different aspect of growth (i.e. the second order derivative). However, we might reason that terms that show both high overall growth rates and high second order growths are potential “early growth” candidates. From the table above, we note that the only term which appears in all three top-20 lists is ***semiconducting zinc compounds***.
5. By studying the taxonomy, we see that this technology is classified under the broad category of *photovoltaic cell*. Following the sequence of terms from *photovoltaic cell* to *semiconducting zinc compounds*, other interesting terms were *thick film, nanostructure*, *absorption*, *heterojunction* and *zinc oxides*, all of which are materials science topics relevant to photovoltaic cells. In this way, individuals unfamiliar to a domain of research can quickly become familiar with the key research topics within a relatively short period of time.

#	Mean publication year	Log growth rate	Second order growth rate
1)	concentration (composition)	concentration (composition)	eirev
2)	semiconducting zinc compounds	concentration process	energy conversion
3)	chlorine compound	chlorine compound	semiconducting zinc compounds
4)	lignin	hybrid sensors	fluidized beds
5)	hybrid sensors	semiconducting zinc compounds	fluid dynamics
6)	telecommunication equipment	international symposium	sorption
7)	sorption	industrial electronics	rate constants
8)	wireless sensor networks	telecommunication equipment	reverse osmosis
9)	wireless network	gas fuel purification	waste disposal
10)	architecture design	wireless sensor networks	climatology
11)	data storage equipment	lignin	semiconductor quantum wells
12)	zea mays	wireless network	potential energy
13)	ecosystems	sensor-networking	acetic acid
14)	biofilms	architecture design	seawater
15)	sensor-networking	thermochemistry	biochemical engineering
16)	biochemical engineering	detectors	filling factors
17)	computer networks	hydraulic machinery	semiconducting cadmium telluride
18)	gas fuel purification	telecommunication systems	chemically bonded
19)	sewage	zea mays	gas generation
20)	sugars	ecosystems	induction motors

Table 2: Top 20-ranked keywords based on each of the early-growth indicators

Next the actual keyword taxonomies generated will be presented. However, in practice, taxonomies which are sufficiently large to be interesting will also be too large to inspect in great detail. What is needed is instead a means of presenting these taxonomies and the growth information contained therein ina form that is easily grasped and which quickly reveals areas that are experiencing higher growth or at least the potential for growth.

To achieve this, it was decided to experiment with the use of color-codes which represent the early growth score of a given node. Using the concept of black-body radiation as a basis, we tested a number of different color-maps which corresponded to everyday notions of "hotness". After trying out a variety of combinations, the color-map shown in Figure 8 was used.

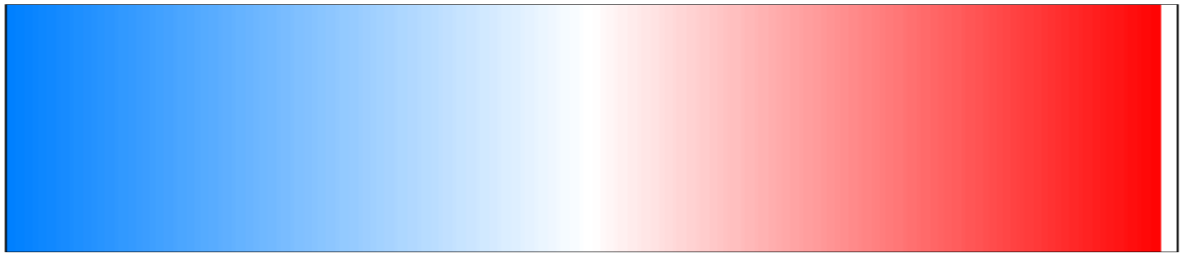


Figure 8: Color-map for encoding the "temperature" of nodes. Nodes with low growth rates are represented by colors in the left side of the spectrum, while increasing growth rates will be reflected by moving towards the right end of the spectrum.

In order to match the early growth indicators to one of the colors in the color-map, the values of the indicators were first "flattened" by mapping them to a uniform distribution. This was done by first ranking each node based on the value of its aggregated early growth score. This resulted in a transformed "score" of 1 to 900 for each nodes. This is then mapped linearly to one of the 255 colors in the color-map.

The full taxonomy (900 terms) is presented in Figure 9, where the color-coding is based on the mean publishing year, and again in Figure 10, where the colors reflect the second order growth rate indicator (only two taxonomies are presented because the taxonomy for the log growth indicator is very similar to the mean publishing year.

Based on the mean publishing year, we note that two regions in particular, corresponding to the subtrees rooted in *biomass* and in *wind power*, are predominantly red in color, indicating that these are areas of high relative growth. However, looking at the taxonomy in Figure 10, we note that the *wind power* subtree is now almost entirely blue/light colored, while the *solar energy* subtree has become a lot "hotter" in appearance. The *biomass* subtree is still quite "hot" though it appears to have cooled slightly.

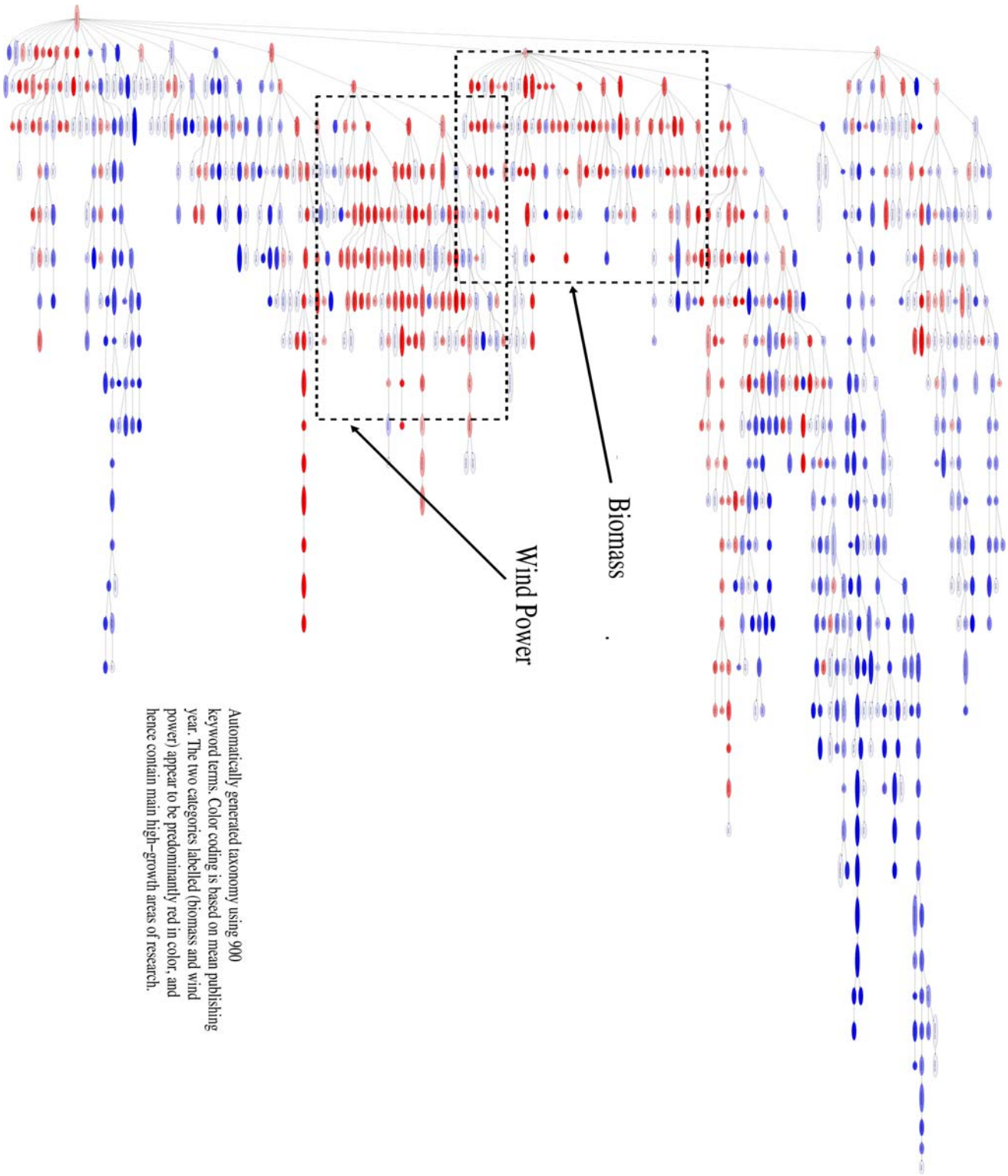


Figure 9: Keyword taxonomy: renewable energy domain (Mean publishing year)

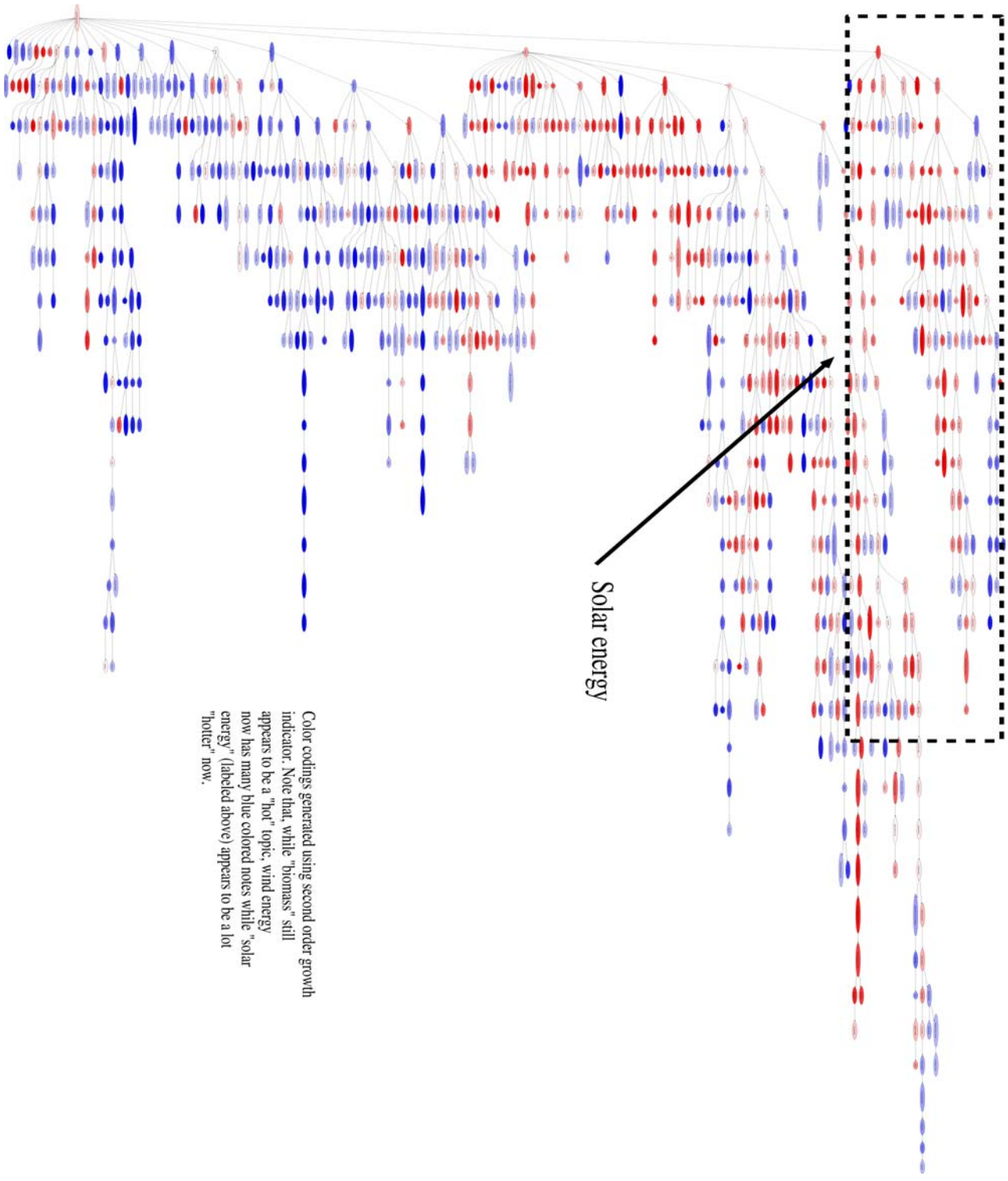


Figure 10:Keyword taxonomy: renewable energy domain Second order growth)

From these two figures, it would appear that topics in Biology are of particular importance in terms of growth potential. To investigate this further, and also to provide a more detailed look at a paticular segment of the taxonomy, the subtrees rooted in *biological materials* are presented in Figure 11 and. This is a subset of the *biomass* tree, which is still too large to clearly reproduce here.

Some observations from these two subtrees are:

1. In both cases, the biological materials subtree is still clearly a "hot" area, which would imply that many of the subsumed topics are vibrant areas of research.
2. When viewed using the second-order growth colorings, the subtree is even hotter when compared to the colorings associataed with the mean-publishing year. This shows that not only is this is an active area of research, it is also *getting more active*. One example which was "cool" in terms of mean publishing year but which scored very high with the second order growth indicator was the branch:

   [adsorption→*infrared spectroscopy*→*fourier transform infrared spectroscopy*→*spectroscopic analysis*]

   A quick search revealed that Fourier Transform Infrared Spectroscopy (FTIR) is a powerful technique which can be used to study the nature of chemical bonds within biological molecules and possesses many applications in biomass, for example in determining the lignin content in forage grasses [Allison et al, 2009].
3. Note that all studies in our database are related to renewable energy due to the original search terms used (refer to earlier section on data collection and keyword extraction). So, for example, while *sugars* may not sound like a particularly exciting research topic, the high score obtained using both growth indicators only refers to the research in sugar *in the context of renewable energy*, which is likely going to be on topics like biodiesel and ethanol production.

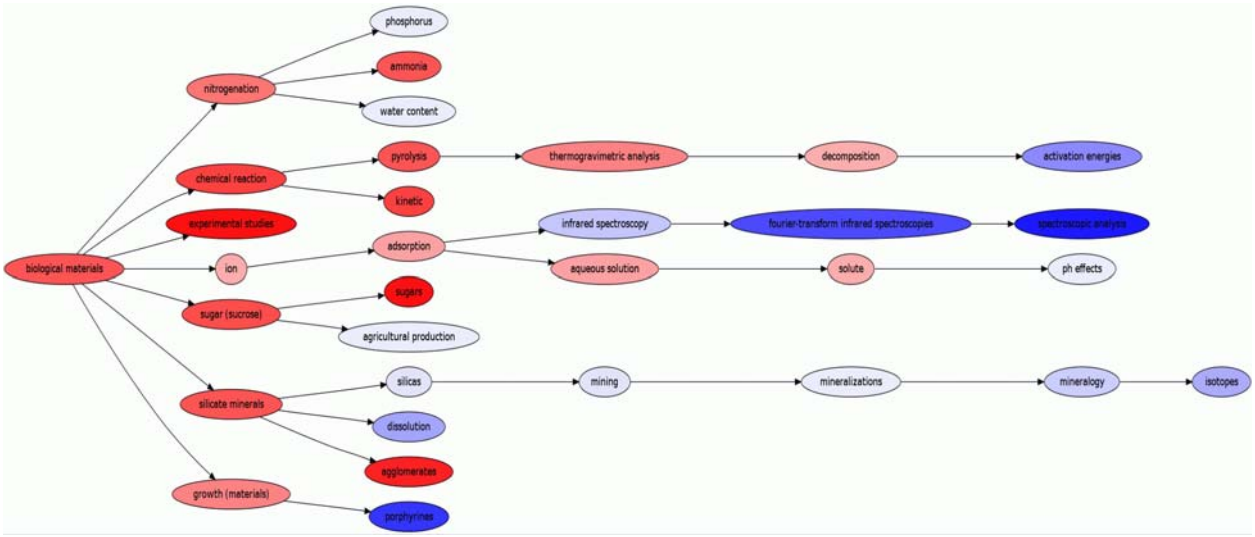


Figure 11: Subtree for *keyword "biological materials". Nodes are colored according to the mean publishing year*

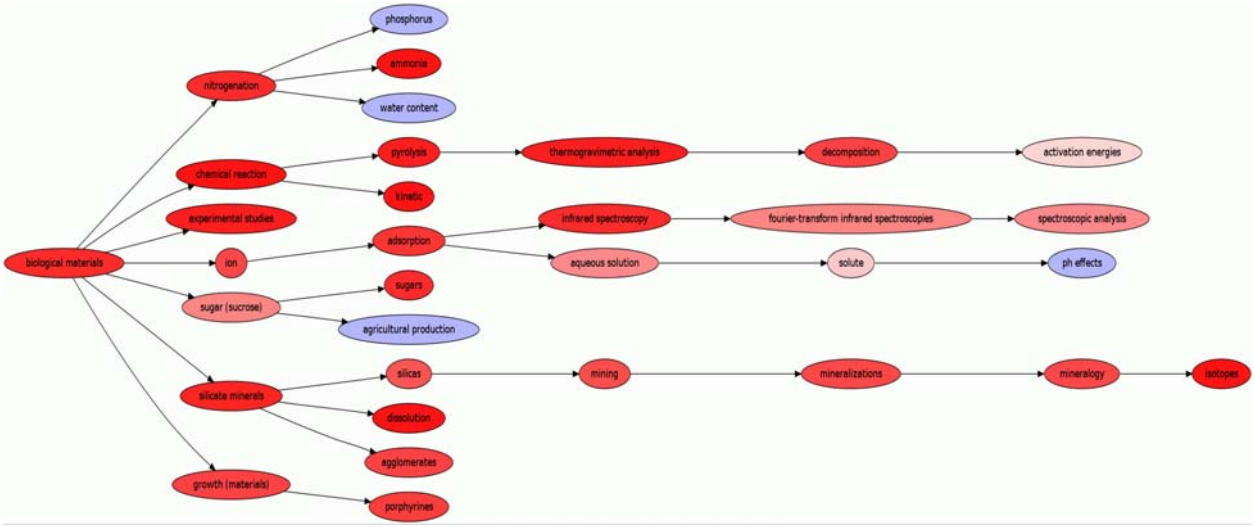


Figure 12: Subtree for keyword "biological materials". Nodes are colored acording to the second-order growth indicator

Reference implementation: software and technical details

As mentioned at the start of the project, the proposed framework is a generic platform for technology forecasting and can be instantiated in a number of ways. As a demonstration of its effectiveness a reference implementation of this framework was created. We now briefly describe the implementation details of each of the blocks depicted in Figure 1

(a) Data collection and term extraction

For the results presented in this report, data was mainly collected from the Scopus database. This was done in a semi-automated manner, where the initial download of the records had to be done manually as there is a limit to the number of citations that can be downloaded at a time. As such, the keyword searches had to be structured manually so as to limit the number of results to under 2000 to avoid losing any data.

Once this was done, the records could be downloaded in CSV (Comma-Separate-Values) format, which were then processed using software that we had developed in the Python programming language. These automated the process of filtering and converting the data before it can be stored locally. This process involved two processing steps:

1. Checking the CSV files for inconsistencies , particularly in the number of fields for each record (typically due to illegal characters in the text records, etc). Where possible these are resolved automatically but in a number of cases "irretrievable" records would be deleted. This typically happened to only a very small number of record (less than 1 for every 1000 records).
2. Converting the extracted CSV fields into SQL commands for storage in a local database to facilitate analysis at a later stage.

While any relational database software can be used for this purpose the SQLite "database in a file" package was selected because it is lightweight, has been ported a large number of platforms and provides adequate performance for small datasets. More recently we have been considering migrating our data to

an MySQL database as the size of our database has been increasing and to support the possibility of hosting the data and the implementation on separate machines.

To store the Scopus data a single table is used.

(b) Growth indicators

As described earlier in this report, a number of the growth indicators were tested, with the results obtained using three particular examples presented in this report. These were all implemented in the Python programming language which allows for fast implementation times, instant portability to a number of platforms and acceptable performance. The software was designed in a modular manner thus allowing additional growth indicators to be added and test easily.

The the keyword prevalences were found by first loading the document abstracts into main memory then using regular expressions to detect occurrences of each of the search terms.

(c) Taxonomy generation

Again, Python was used to implement the automated creation of taxonomies and subsequent integration with the early growth parameters. Each keyword taxonomy was represented using $n \times n$ connection matrix **T**, where a given element $\mathbf{T}_{i,j}$ is set to 1 to denote a "is-subset-of" relationship between terms t_i and t_j, and 0 otherwise.

Finally the generation of graphics depicting the taxonomies was performed using *GraphViz*[2]. This is an open-source package yet provides a great deal of flexibility when generating graphs of various types.

Wikipedia for ontology generation

Finally, as an introduction to some of the directions which we are currently pursuing, we describe an alternative approach to the keyword collection and taxonomy generation stages of the framework, which was to mine this information from Wikipedia.

The Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project and is one of the most popular websites worldwide[3]. As of June 2009, it contains five million articles in English. These are written in a semi-structured way, containing a general multiple inheritance categorisation, info-boxes, references, links to similar pages and links to articles in other languages.

Recently, the Wikipedia has been recognised as a powerful resource for Artificial Intelligence tasks (e.g. inference of semantic relatedness) as well as Semantic Web and Information Extraction applications ([Wu et al., 2008, Weld et al., 2008]). The Wikipedia has been previously used to create general ontologies using info-boxes ([Wu and Weld, 2008]) and to extract taxonomy relations based on syntactic patterns [Ryu and Choi, 2006]. A notable Semantic Web application is the organisation of large knowledge database from media-related web-contents ([Kobilarov et al., 2009]) using comprehensive sets of parsed Wikipedia data (dbpedia,[Auer et al., 2007]).

2 http://www.graphviz.org

3Rank 8 according to the comScore Metrix, January 2008

Generation of a Renewable Energy taxonomy

The Wikipedia provides a rich terminology for Renewable Energy. The current, pruned data structure contains 155 categories and 794 Wikipedia terms, 2875 terms when including Wikipedia redirects which represent alternative descriptions. Since the Wikipedia provides a hierarchical structure for categories, the categories are arranged in a directed acyclic graph (DAG).

The term-generation tool TerMine [Frantzi et al., 2000] has been used to generate 5000 terms from 18200 abstracts of Renewable energy related papers. It was observed that 50-80% of the TerMine generated list are already Wikipedia terms or contained in Wikipedia terms, in particular terms top ranked by TerMine. As a consequence, it is possible to build a taxonomy of these terms by means of Wikipedia Categories.

In order to perform technology forecasting, it was necessary to bridge the gap between the scientific literature and Wikipedia it is possible to add terms extracted from scientific literature, e.g. keywords. This approach requires a reliable similarity measure between terms. As described earlier, frequently occurring terms can be compared via their occurrence vectors in scientific literature. The Heymann algorithm can then be applied to insert new terms into the existing Wikipedia taxonomy.

To investigate the feasibility of this approach, we count how often Wikipedia terms appear in our corpus of scientific literature. From the 2875 terms, 555 are mentioned at least once, 229 terms appear at least in 10 articles and 80 terms appear at least in 100 articles. We expect to increase the number of Wikipedia terms in the scientific literature corpus by specifically searching for identified Wikipedia terms in scientific literature databases and include the found articles in the current corpus. It is highly likely that papers containing Wikipedia terms exist in databases like SCOPUS. We therefore generate recursive queries for a category *c* as shown in the following equation:

$$Q(c) = \bigvee_{c' \prec c} Q(c') \vee \bigvee_{t \in N(c)} t \vee \bigvee_{t \in A(c)} (t \wedge (name(c)) \qquad (12)$$

N and *A* are the nonambiguous and ambiguous Wikipedia terms associated to a category resp., and *name* provides the name of a category. This procedure provides a powerful search and largely increases the recall of retrieved documents. In order to increase the precision, ambiguous terms are identified and required to be in the context of the category name.

Note that the Wikipedia provides the potential for further improvements: more synonyms can often be extracted from the article header ([Ryu and Choi, 2006]). Furthermore, definitions/glossary, related links (internal/external), literature references and other languages can contribute to determine term similarity.

Relevance filters for terms

In order to enhance the quality of a term selection, we identify non-specific and ambiguous Wikipedia terms. Terms like “Water” or “Vehicle” or “Review” appear ubiquitously and are unsuited as category indicators. Note that the need for filtering applies to termlists from term extraction tools as well as to Wikipedia-terms. The following techniques were used:

1. **Using Wikipedia’s category hierarchy -** Due to Wikipedia’s multiple taxonomies it is possible to filter Wikipedia terms like “Hawaii” (belonging to category “Ethanol fuel”, but also to “States

of the United States") by identifying them as geographic places. Likewise, companies or people can be filtered that way.

2. **Using WordNet's synsets -** In order to increase the precision, ambiguous terms are identified using WordNet's synsets ([Fellbaum, 1998]). If a term has several synsets, i.e. distinguishable meanings, it is considered ambigouos. If a term is not registered in WordNet then it technically can still be ambiguous but is assumed to be specific enough to cause few or no False Positive hits.
3. **Using stopwords from unrelated scientific literature -** To further filter irrelevant terms we created a corpus of abstracts, that is supposed to reflect the common, but unspecific scientific terminology such as "quality assessment" or "survey" or "review". The literature is taken from Pubmed.org (a publicly accessible server for biomedical literature), querying for cancer related articles (in order to avoid biofuels). This way we retrieved 81.000 abstracts. Terms from this corpus can be extracted using general term extraction tools like TerMine. These terms help to identify non-specific terms in Renewable Energy termlists.

Implementation

All tools are implemented in Python. Wikipedia categories are arranged as a multiple taxonomy or directed acyclic graph, i.e. a category can inherit from several parents (multiple inheritance). The categories are therefore arranged in a DAG data structure. Many tasks, such as finding sub-categories require comprehensive, traversals of the graph. The multiple inheritance aspect is taken into account by memorising previously traversed branches such that multiple traversals of branches are avoided during all recursive top-down traversals.

To obtain the data, XML dump of the all articles of the June-2009 Wikipedia was downloaded. The file is split into many small files and Xapian is used for indexing. An SQL database has been created for a table that links Wikipedia pages to a category and for a table that links terms and their redirects. A corpus of abstracts and their associated meta-data was extracted from SCOPUS using the query term "Renewable".

We identified renewable energy categories using Wikipedia's CategoryTree tool. For each category, the super-categories are looked up using the off-line Wikipedia. Each branch is included in the DAG-structure. Since the Wikipedia is the result of human efforts, violations to the directed acyclic graph assumptions may occur. Therefore, we provide a method to detect and remove cycles. For each category we look up the associated Wikipedia articles and their respective redirects. Occurrences of Wikipedia terms in the SCOPUS corpus were identified by generating word-stemmed abstracts for the complete corpus and second matching stemmed Wikipedia articles

Results

As in the previous sections, an intuitive visualization of current trends in the Renewable Energy domain was generated by coloring the categories as well as terms by their mean publication year, one of the growth identifiers discussed previously. A "hot" research topic is one with the predominat share of its publications in very recent years. As before, categories recursively aggregate all subcategories and their corresponding terms, which improves the evidence for a trend. The resulting color-coded taxonomy is presented in Figure 13

Some observations were:

1. It can be seen that Biofuels and Hydrogen technology are very active fields of research in general. In particular, notable topics with very recent publications are Biohydrogen, Biomass to liquid and Biodiesel. This particularly interesting because it corroborates well with the results obtained using the automatically generated taxonomies presented earlier in this report.

   Unfortunately due to the size of the taxonomy the nodes are very small but the two top level categories concerned have been clearly labelled.

2. While other top level categories like Solar Energy could well contain some very active topics (e.g. Building-integrated photovoltaics (BIPV), Organic Photovoltaics and Dye sensitised solar cells), these are often balanced by less active fields (silicon photovoltaics), so the overall trend is less clear.

3. It is apparent that within categories very different trends prevail. Whereas energy crops as a whole is a topic that has been investigated for a long period of time (mean publication year is 2003), very active fields are Triticum, Brassica napus and Switchgrass.

Note that currently a number of Wikipedia categories have low coverage in our literature corpus. This is because this direction of research has yet to be merged with the "main" branch in which we are currently utilizing a fairly large database of abstracts. We intend to improve this very soon by merging the two branches (auto-generated and wikipedia based taxonomies).

One benefit of this will be the use of a common pool of abstracts and citations, but it should also be noted that the Wikipedia ontology is largely the result of manually linked categories and terms and hence will have significantly different characteristics from the automatically generated taxonomies. On the one hand it should better reflect intuitive notions about what a subset is, and may be easier to interpret or to use an a visualization tool. However, on the other hand, the automatically generated taxonomy captures the actual structure of the academic literature as is manifested in the term co-occurrences. This means that terms that are correlated will tend to be grouped in the same subtree, which would allow underlying trends to reinforce each other and to be more easily detectable. In addition, an automatically generated taxonomy could be quickly updated in response to new information and data becoming available, making it more adaptive and up-to-date than a manually curated alternative.

In short, it is expected that some means of combining the two methods, rather than choosing one over the other, might eventually turn out to be the best approach to take.

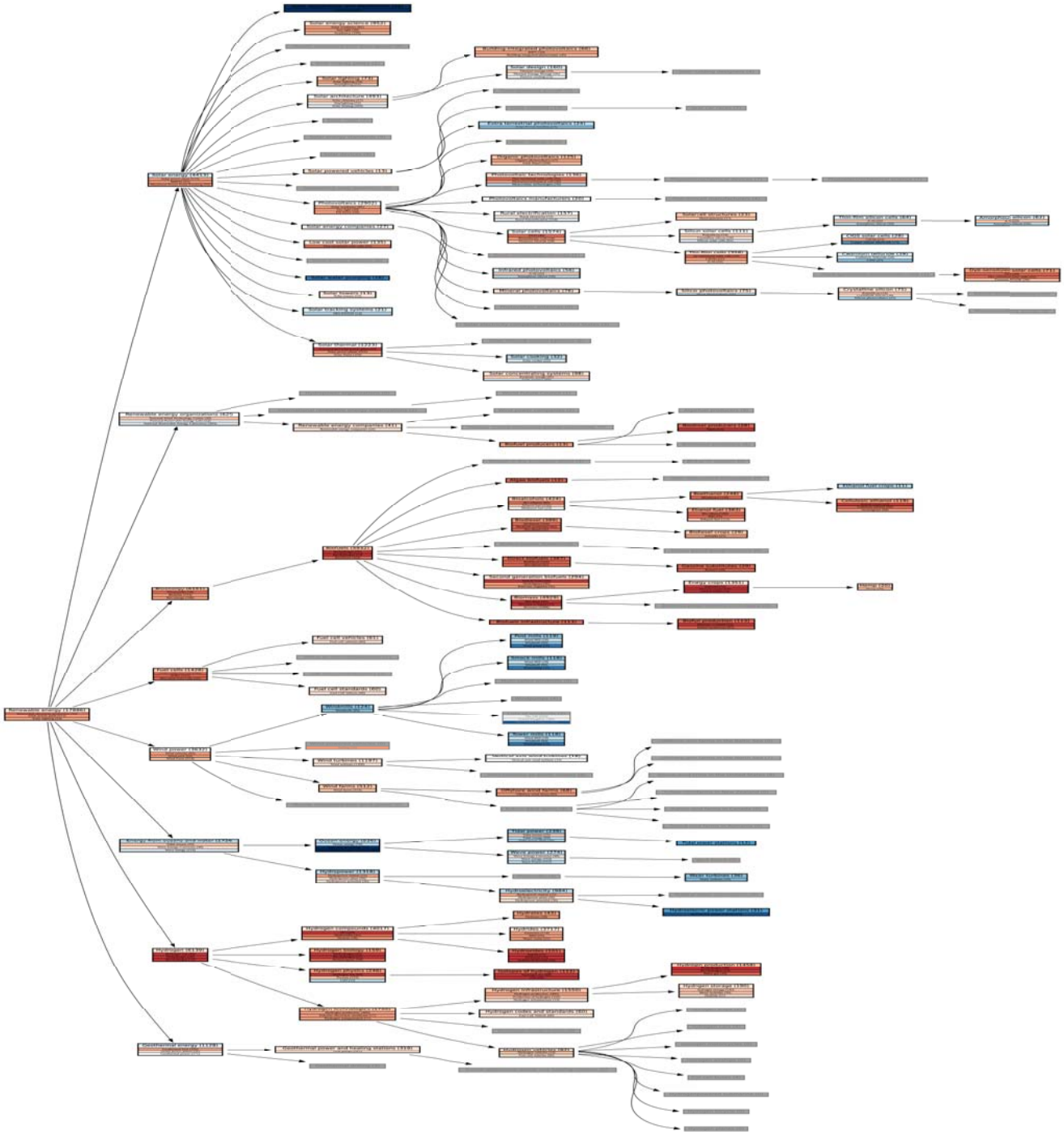


Figure 13: The Wikipedia Renewable Energy categories, coloured by average publication year, red refers to most recent and blue to the oldest reserch fields. Displayed with each category are also the top three terms. Categories with less than 10 publications are greyed out.

CURRENT REPORTING PERIOD SUMMARY (10/01/2008-03/31/2009)

Review of objectives

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.

   MIST/MIT division: Joint

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. Two 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.

   MIST/MIT division: Primarily MIST

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 proposed a technique by which the scores for individual keywords can be aggregated via the above-mentioned taxonomies to obtain more reliable results.

   MIST/MIT division: Joint

4. *Renewable energy case study* – This activity is still underway and will be the focus of the remaining three months of the project.

   MIST/MIT division: Joint

5. *Development of software tools* - Two main tools were worked on during this reporting period – the "Cameleon Scheduler", and the "Long Tail analysis tool", both of which have been described in the preceding sections.

   MIST/MIT division: Primarily MIT

FUTURE WORK

From now until the formal end date of the project we wil be focussing on applying the tools described here to the creation of the renewable energy case study. While we have already conducted a number of

limited pilot studies, as well as tested the proposed methods on data relevant to renewable energy, these efforts have hitherto been largely experimental.

Note that for almost every stage of the technology forecasting framework, we have tested more than one alternative technique which, in most cases, is further customizable via a variety of parameters and settings. As such, the main challenge for the remaining three months of the project will be to select the best possible set of parameters and to apply these towards the compilation of the case study. A further task will be the collection of extra data where appropriate in cases where there is insufficient coverage of promising keywords and technologies.

We have every confidence that this undertaking will be a success and look forward to the successful achievement of the project goals.

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.

[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.

[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.

[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.

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

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

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

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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*, Online first.

[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

[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.

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

PUBLICATIONS/PRESENTATIONS

The following is an updated list of publications produced as part of this project. Note that items labelled as "*in preparation*" refer to articles which are almost complete (>95%), or which are undergoing modifications prior to re-submission, and will be submitted in the very near future.

Journal articles

"Asymmetric information distances for automated taxonomy construction " (2009), W.L Woon and S.E. Madnick (in press) *Knowledge and Information Systems.*

"Visualizing technology domains using inter-keyword distances" (2009), W.L. Woon and S.E. Madnick, (in preparation).

"Bibliometric analysis of distributed generation" (2009), W.L. Woon, H.Zeineldin and S.E. Madnick (in preparation).

Conference publications

"A Framework for Technology Forecasting and Visualization", W.L. Woon, A. Henschel and S.E. Madnick, (under review) IEEE International Conference on *Innovattions in IT*, Al Ain, UAE, 2009

"Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation", W.L. Woon, A. Henschel and S.E. Madnick, (under review) IEEE International Conference on Innovattions in IT, Al Ain, UAE, 2009

"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.

Presentations

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

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

Research Thesis

"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

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

APPENDICES

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

Here, we provide copies of the most relevant research papers produced as part of this project, which should provide further details on most of the activities described in this report.

APPENDIX

Copies of Selected Papers

1] “A Framework for Technology Forecasting and Visualization”, W.L. Woon, A. Henschel and S.E. Madnick, (under review) *IEEE International Conference on Innovations in IT*, Al Ain, UAE, 2009

2] “Asymmetric information distances for automated taxonomy construction,” W.L Woon and S.E. Madnick (in press) *Knowledge and Information Systems* (KAIS), Vol. 21, No. 1, October 2009, pp. 91 - 111

3] “Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation”, W.L. Woon, A. Henschel and S.E. Madnick, (under review) *IEEE International Conference on Innovations in IT*, Al Ain, UAE, 2009

4] “Bibliometric analysis of distributed generation,” W.L. Woon, H.Zeineldin and S.E. Madnick (in preparation), 2009.

5] “Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry" G. Vidican, W.L. Woon and S.E. Madnick, *Proceedings of the Advancing the Study of Innovation and Globalization in Organizations* (ASIGO) Conference, Nurnberg, Germany, May 29-30, 2009.

A Framework for Technology Forecasting and Visualization

Wei Lee Woon
Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE.

Andreas Henschel
Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE.

Stuart Madnick
Massachusetts Institute of Technology
77 Mass. Ave.,Building E53-321
Cambridge, MA 02139-4307, U.S.A.

Abstract

This paper presents a novel framework for supporting the development of well-informed research policies and plans. The proposed methodology is based on the use of bibliometrics*; i.e., analysis is conducted using information regarding trends and patterns of publication. Information thus obtained is analyzed to predict probable future developments in the technological fields being studied. While using bibliometric techniques to study science and technology is not a new idea, the proposed approach extends previous studies in a number of important ways. Firstly, instead of being purely exploratory, the focus of our research has been on developing techniques for detecting technologies that are in the* early growth *phase, characterized by a rapid increase in the number of relevant publications. Secondly, to increase the reliability of the forecasting effort, we propose the use of automatically generated keyword taxonomies, allowing the growth potentials of subordinate technologies to aggregated into the overall potential of larger technology categories. As a demonstration, a proof-of-concept implementation of each component of the framework is presented, and is used to study the domain of renewable energy technologies. Results from this analysis are presented and discussed.*

1 Introduction

For decision makers and researchers working in a technical domain, understanding the state of their area of interest is of the highest importance. Any given research field is composed of many subfields and underlying technologies which are related in intricate ways. This composition, or research landscape, is not static as new technologies are constantly developed while existing ones become obsolete, often over very short periods of time. Fields that are presently unrelated may one day become dependent on each others findings.

Information regarding past and current research is available from a variety of channels, providing both a difficult challenge as well as a rich source of possibilities. On the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data with which a well-informed and comprehensive research strategy may be formed.

There is already a significant body of related research, and for a good review, the reader is referred to [16, 13, 14]. Interesting examples include visualizing interrelationships between research topics [15, 18], identification of important researchers or research groups [11, 12], the study of research performance by country [8, 10], the study of collaboration patterns [1, 4, 3] and the analysis of future trends and developments [17, 7, 6, 18].

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

In response to this apparent shortcoming, 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 subtrees of the taxonomy.

2. A set of numerical indicators for identifying technologies of interest. In particular, we are interested in de-

veloping 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.

2 A framework for technology forecasting

It is important to define the form of forecasting that is intended. In particular, it must be stressed that it is not "forecasting" in the sense of a weather forecast, where specific future outcomes are intended to be predicted with a reasonably high degree of certainty. It is also worth noting that certain tasks remain better suited to human experts; in particular, where a technology of interest has already been identified or is well known, we believe that a traditional review of the literature and of the technical merits of the technology would prove superior to an automated approach.

Instead, the proposed framework targets the preliminary stages of the research planning exercise by focussing on what computational approaches excel at: i.e. scanning and digesting large collections of data, detecting promising but less obvious trends and bringing these to the attention of a human expert. This overall goal should be borne in mind as, in the following subsections, we present and describe the individual components which constitute the framework.

2.1 Overview

Figure 1 depicts the high-level organization of the system. As can be seen, the aim is to build a comprehensive technology analysis tool which will collect data, extract relevant terms and statistics, calculate growth indicators and finally integrating these with the keyword taxonomies to produce actionable outcomes. To facilitate discussion, the system has been divided into three segments:

1. Data collection and term extraction (labelled **(a)** in the figure)
2. Prevalence estimation and calculation of growth indicators (labelled **(b)**)
3. Taxonomy generation and integration with growth indicators (labelled **(c)**)

These components are explained in the following three subsections.

2.2 Data collection and term extraction

2.2.1 Data collection

The type of data source, collection mechanism and number of sources used can be modified as required but for the proof-of-concept implementation, information extracted from the Scopus[1] database was used. 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 but Scopus proved to be a good initial choice as it returned results which were

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

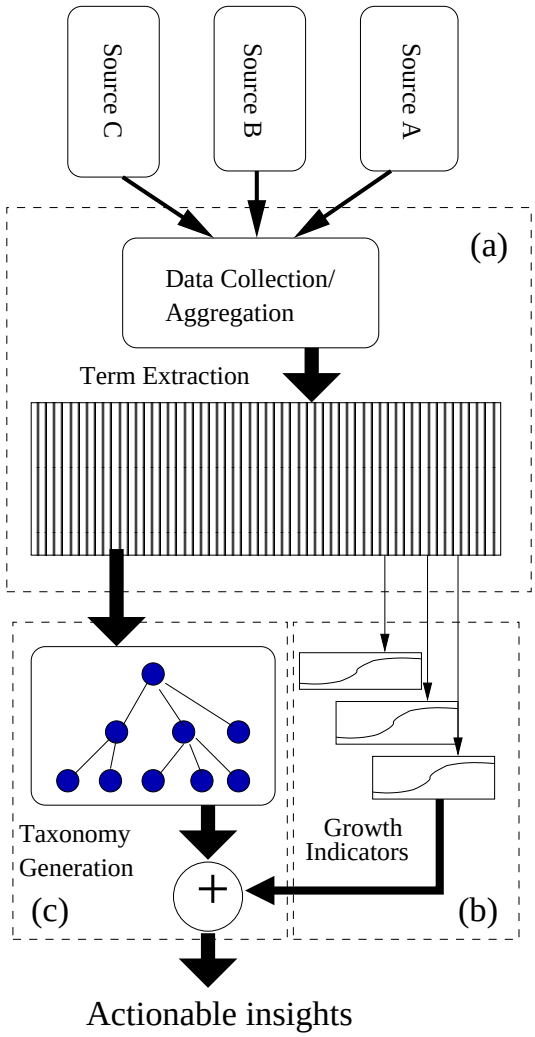


Figure 1. Proposed framework: overall structure

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.

2.2.2 Term extraction

Term extraction is the process of automatically generating a list of keywords on which the technology forecasting efforts will be focussed. Again, there are a variety of ways in which this can be achieved; we have experimented with a number of these and our experiences have been thoroughly documented in [21]. For the present demonstration the following simple but effective technique is used: for each document retrieved, a set of relevant keywords is provided. These are collected and, after word-stemming and removal of punctuation marks, sorted according to number of occurrences in the text. For the example results shown later in this paper, a total of 500 keywords have been extracted and used to build the taxonomy.

2.2.3 Pilot study

To provide a suitable example on which to conduct our experiments and to anchor our discussions, a pilot study was conducted in the field of renewable energy. The incredible diversity of renewable energy research offers a rich and challenging problem domain on which we can test our methods. Besides high-profile topics like solar cells and nuclear energy, renewable energy related research is also being conducted in fields like molecular genetics and nanotechnology.

To collect the data for use in this pilot study, a variety of high-level keywords related to renewable energy (please see Appendix A) were submitted to Scopus, and the abstracts of the retrieved documents were collected and used. In total, $119,393$ abstracts were retrieved and subsequently ordered by year of publication.

2.3 Identification of early growth technologies

There are actually two steps to this activity. The first is to find a suitable measure for the "prevalence" of a given technology as a function of time. In terms of a database of academic publications, this would be some means of measuring the size of the body of relevant publications appearing each year. It is difficult to achieve this directly but an alternative would be to search for 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.

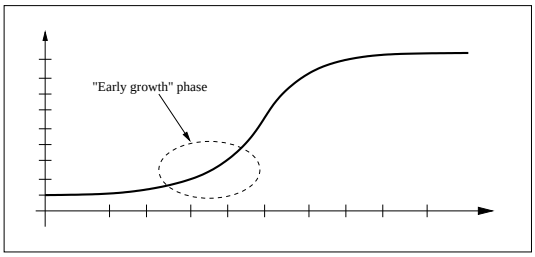


Figure 2. Early growth of technological development

This is defined as:

$$TF_i = \frac{n_i}{\sum_{j \in \mathcal{I}} n_j} \tag{1}$$

where, n_i is the number of occurrences of keywords i, and $\mathcal{I}$ is the set of terms appearing in all article abstracts (this statistic is calculated for each year of publication to obtain a time-indexed value). Once the term frequencies for all terms have been extracted and saved, they can be used to calculate growth indicators for each of the keywords (and, by extension, the associated technologies). These, in turn, are used to rank the list of terms.

As stated previously, we are most interested in keywords with term frequencies that are relatively low at present but that have been rapidly increasing; this will be referred to as the "early growth" phase of technological development, depicted in figure 2, and represents the fields to which an expert would wish to be alerted. Existing techniques are often based on fitting growth curves (see [2] 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 an area of active research but for this paper the following indicator will serve as an example:

$$\theta_i = \frac{\sum_{t \in [2004,2008]} t.TF_i[t]}{\sum_{t \in [2004,2008]} TF_i[t]}, \tag{2}$$

where, θ_i is the growth potential 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 last five year (excluding 2009), and which are relevant to term i (a more recent year indicates greater currency of the topic).

2.4 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.

2.4.1 Taxonomy generation

The question remains, how do we obtain such a taxonomy? In a limited number of cases, these taxonomies may be available from external sources such as government agencies and other manually curated sources. However, in many cases, a suitable taxonomy is either unavailable, or is available but is not sufficiently updated to be of use for the application at hand. As such, to make our framework broadly applicable, an important research direction is the *automated* creation of keyword taxonomies based on the statistics of term occurrences.

The basic idea, as indicated in section 1 is to group together terms which tend to co-occur frequently. Again, we have tested a number of different ways of achieving this (two earlier attempts are described in [20, 19]) but it is not possible in the present scope to discuss and compare these in depth. Instead, we present one particular method which was found to produce reasonable results while being scalable to large collections of keywords. This is based on the algorithm described in [9] which was originally intended for social networks where users annotate documents or images with keywords. Each keyword or tag is associated with a vector that contains the annotation frequencies for all documents, and which is then comparable, for e.g. by using the cosine similarity measure. We adapt the algorithm to general taxonomy creation by using two important modifications; firstly, instead of using the cosine similarity function, the *asymmetric* distance function proposed in [20] is used (this is based on the "Google distance" proposed in [5]):

$$\overrightarrow{\mathrm{NGD}}(t_x, t_y) = \frac{\log n_y - \log n_{x,y}}{\log N - \log n_x}, \tag{3}$$

where t_x and t_y are the two terms being considered, and n_x, n_y and $n_{x,y}$ are the occurrence counts for the two terms occurring individually, then together in the same document respectively. Note that the above expression is "asymmetric" in that $\overrightarrow{\mathrm{NGD}}(t_x, t_y)$ refers to the associated cost if t_x is classified as a subclass of t_y, while $\overrightarrow{\mathrm{NGD}}(t_y, t_x)$, corresponds to the inverse relationship between the terms.

The algorithm consists of two stages: the first is to create a similarity graph of keywords, from which a measure of "centrality" is derived for each node. Next, the taxonomy is grown by inserting the keywords in order of decreasing centrality. In this order, each unassigned node, t_i, is attached to one of the existing nodes t_j such that:

$$j = \arg\min_{j \in \mathcal{T}} \overrightarrow{\mathrm{NGD}}(t_i, t_j), \tag{4}$$

(where $\mathcal{T}$ is the set of terms which have already been incorporated into the taxonomy.)

2.4.2 Enhanced early growth indicators

Once the keyword taxonomies have been constructed, they provide a straightforward method of enhancing the early growth indicators using information regarding the co-occurrence statistics of keywords within the document corpus. As with almost all aspects of the proposed framework, a number of variants are possible but the basic idea is to recalculate the early growth scores for each keyword based on the aggregate scores of each of the keywords contained in the subtree descended from the corresponding node in the taxonomy.

For the results presented in this paper, the aggregation operation used was a straight average, though other more elaborate schemes are clearly possible.

3 Results and discussions

We present results for a simple pilot study in renewable energy. As described in section 2.2.1, the Scopus database was used to collect a total of 500 keywords which were relevant to the renewable energy domain, along with $119,393$ document abstracts. These keywords were then used to construct a taxonomy as described in section 2.4.1, and the average publication year for each keyword was calculated as

shown in equation (2). Finally, these were aggregated using the keyword taxonomy and the list of keywords was sorted according to order of decreasing publication year. Using this method of evaluation, the top 30 keywords were (numbers in brackets are the taxonomy-aggregated average publication years):

1. cytology (2007.31)
2. nonmetal (2007.24)
3. semiconducting zinc compounds (2007.19)
4. alga (2006.94)
5. hydraulic machinery (2006.91)
6. hydraulic motor (2006.91)
7. bioreactors (2006.81)
8. concentration process (2006.77)
9. metabolism (2006.73)
10. sugars (2006.69)
11. computer networks (2006.66)
12. experimental studies (2006.63)
13. ecosystems (2006.58)
14. direct energy conversion (2006.57)
15. lignin (2006.56)
16. zea mays (2006.56)
17. bioelectric energy sources (2006.56)
18. phosphorus (2006.55)
19. biological materials (2006.53)
20. cellulose (2006.52)
21. nitrogenation (2006.52)
22. bacteria (microorganisms) (2006.52)
23. adsorption (2006.52)
24. soil (2006.52)
25. hydrolysis (2006.51)
26. glycerol (2006.51)
27. fermenter (2006.51)
28. glucose (2006.50)
29. potential energy (2006.50)
30. biodegradable (2006.43)

Some quick observations:

1. One of the most striking observations is the number of biotechnology related keywords in this list. This indicates that biological aspects of renewable energy are amongst the most rapidly growing areas of research.

2. Amongst the highly-rated non-biological terms on the list were “nonmetal” (#2) and “seminconducting zinc compounds” (#3), both of which are related to the field of thin-film photovoltaics.

3. However, the top-30 list contained a large number of keywords which were actually associated with leaves in the taxonomy, so the confidence in the scores were lower.

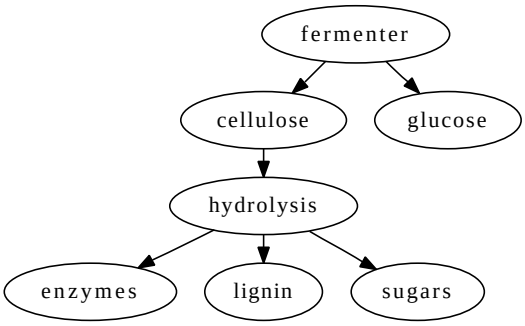


Figure 4. Subtree for node “fermenter”

4. Looking at the terms with relatively large associated subtrees, we see that three of the largest in the top 30 were “biological materials” (15 nodes), “fermenter” (7 nodes) and “hydrolysis” (4 nodes). The subtrees for the first two terms are shown in figures 3 and 4 respectively, while the hydrolysis subtree is actually part of the “fermenter” subtree and as such is not displayed.

5. It can be seen that the fermenter subtree is clearly devoted to biofuel related technologies (in fact, two major categories of these technologies are represented - “glucose”-related or first generation biofuels, and “cellulosic” biofuels which are second generation fuels.

6. The biological materials subtree is less focussed but it does emphasize the importance of biology to renewable energy research. The “soil” branch of this subtree is devoted to ecological issues, while the “chemical reaction” branch is associated with gasification (waste-to-energy, etc.) research.

4 Conclusion

In this paper, a novel framework for facilitating research planning and decision-making has been presented. The proposed system covers the entire chain of activities starting with the collection of data from generic information sources (online or otherwise), the extraction of keywords of interest from these sources and finally the calculation of semantically-enhanced “early growth indicators”.

In addition, a simple proof-of-concept implementation of this framework is described and is applied to the domain of renewable energy. Results of this study are presented and discussed, and are already quite encouraging though currently the process is still a little too noisy to pick out “very early growth” technologies. However, we are investigating numerous avenues for enhancing the basic implementation referenced here, and are confident of presenting improved findings in upcoming publications.

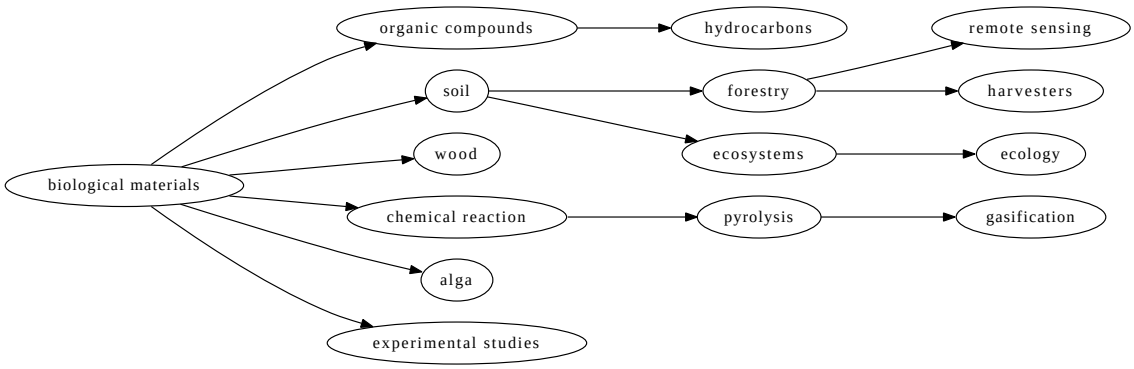


Figure 3. Subtree for node "Biological materials"

A Seed terms for data collection

The search terms used for extraction of renewable energy related abstracts and keywords from Scopus were: *"renewable energy", "biodiesel", "biofuel", "photovoltaic", "solar cell", "distributed generation", "dispersed generation", "distributed resources", "embedded generation", "decentralized generation", "decentralized energy", "distributed energy", "on-site generation", "geothermal", "wind power", "wind energy"*.

References

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

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

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

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

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

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

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

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

[9] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, Technical report 2006-10. http://dbpubs. stanford. edu: 8090/pub/2006-10, 2006.

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

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

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

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

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

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

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

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

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

[19] W. Woon and S. Madnick. 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, 2008.

[20] W. Woon and S. Madnick. Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, Online first, 2009.

[21] B. Ziegler, A. Firat, C. Li, S. Madnick, and W. Woon. 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, 2009.

Asymmetric Information Distances for Automated Taxonomy Construction

Wei Lee Woon*, Stuart Madnick†

*Masdar Institute of Science and Technology,

(Visiting Scholar) Technology and Development Program,

M.I.T., 1-175, Cambridge MA, 02139, U.S.A.

†Sloan School of Management, M.I.T.,

E53-321, Cambridge MA, 02139, U.S.A.

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

Abstract

A novel method for automatically constructing taxonomies for specific research domains is presented. The proposed methodology uses term co-occurrence frequencies as an indicator of the semantic closeness between terms. To support the automated creation of taxonomies or subject classifications we present a simple modification to the basic distance measure, and describe a set of procedures by which these measures may be converted into estimates of the desired taxonomy. To demonstrate the viability of this approach, a pilot study on renewable energy technologies is conducted, where the proposed method is used to construct a hierarchy of terms related to alternative energy. These techniques have many potential applications, but one activity in which we are particularly interested is the mapping and subsequent prediction of future developments in the technology and research.

I. Introduction

A. Technology mining

The planning and management of research and development activities is a challenging task that is further compounded by the large amounts of information which researchers and decision-makers have at their disposal. Information regarding past and current research is available from a variety of channels, examples of which include publication and patent databases. The task of extracting useable information from these sources, known as "tech-mining"[Porter, 2005], presents both a difficult challenge and a rich source of possibilities; on the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data which, if effectively utilized, will allow a well-informed and comprehensive research strategy to be formed.

There is already a significant body of research addressing this problem (for a good review, the reader is referred to [Porter, 2005], [Porter, 2007], [Losiewicz et al., 2000], [Martino, 1993]); interesting examples include visualizing the inter-relationships between research topics [Porter, 2005], [Small, 2006], [Kandylas et al., 2008], identification of important researchers or research groups [Kostoff, 2001], [Losiewicz et al., 2000], the study of research performance by country [de Miranda et al., 2006], [Kim and Mee-Jean, 2007] the study of collaboration patterns [Anuradha et al., 2007], [Chiu and Ho, 2007], [Braun et al., 2000], analysis of citation patterns

[Lu et al., 2007], [An et al., 2004] and the prediction of future trends and developments [Smalheiser, 2001], [Daim et al., 2005], [Daim et al., 2006], [Small, 2006].

We also note that taxonomy creation has been addressed before in the literature though the approaches used have been somewhat different. For example, the study described in [Blaschke and Valencia, 2002] uses a distance measure based on the number of shared keywords between clusters of abstracts; a hierarchical clustering scheme is subsequently used to group these abstracts, and the resulting representation is used to create the taxonomies. The study described in [Makrehchi and Kamel, 2007] is more similar to our approach in that it is based on Google search terms - however, the creation of the taxonomy tree is based on a greedy algorithm while the approach proposed here attempts to take the global structure of the taxonomy into consideration. Certainly, in view of the many difficulties inherent to these undertakings, there is still much scope for further development in many of these areas.

For researchers and managers new to a field, it is critical to quickly gain a broad understanding of the current state of research, future scenarios and the identification of technologies with potential for growth and which hence need to be prioritized. The work described in this paper targets this important aspect of technology-mining. Specifically, we seek to answer the following research question: given a collection of keywords relevant to a research area of interest, is it possible to automatically organize these keywords into a taxonomy which reflects the structure of the research domain? In seeking an answer to this question, the following issues will also be addressed:

1) Derivation of an asymmetric measure of distance between keywords which indicates the degree to which one keyword is a subclass of the other.
2) Investigation of methods for converting these distance measurements into an estimate of the underlying topic taxonomy.
3) A pilot study in renewable energy as a demonstration of the proposed approach.

B. Pilot study

To provide a suitable example on which to conduct our experiments and to anchor our discussions, a pilot study was conducted in the field of renewable energy. The incredible diversity of renewable energy research offers a rich and challenging problem domain on which we can test our methods. Besides high-profile topics like solar cells and nuclear energy, renewable energy related research is also being conducted in fields like molecular genetics and nanotechnology.

II. Keyword distances for taxonomy creation

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

1) Identification of an appropriate indicator of closeness (or distance) between a collection of terms which can be used to quantify the relationships between areas of research,
2) Use of this indicator to automatically construct a subject area hierarchy or taxonomy which accurately captures the inter-relationships between these terms.

A. Keyword distances

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

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

In practice, exploiting this intuition is more complicated than might be expected, particularly because an appropriate normalization scheme must be devised. It is certainly not clear what the exact form of this distance expression should be; even more importantly, can it be grounded in a rigorous theoretical framework such as probability or information theory? As it turns out, there is already a closely-related technique which provides this solid theoretical foundation, and which exploits the same intuition; known as the *Google Distance* [Cilibrasi and Vitanyi, 2006], [Cilibrasi and Vitányi, 2007], this method utilizes the term co-occurrence frequencies as an indication of the extent to which two terms are related to each other. This is defined as:

$$\mathrm{NGD}(t_1, t_2) = \frac{\max\{\log n_1, \log n_2\} - \log n_{1,2}}{\log N - \min\{\log n_1, \log n_2\}}, \tag{1}$$

where NGD stands for the *Normalized Google Distance*, t_1 and t_2 are the two terms to be compared, n_1 and n_2 are the number of results returned by a Google search for each of the terms individually and $n_{1,2}$ is the number of results returned by a Google search for both of the terms. Finally, N is the size of the sample space for the "google distribution", and can be approximated by the total number of documents indexed by Google or the search engine being used, if this is not Google.

While a detailed discussion of the theoretical underpinnings of this method is beyond the scope of the present discussion, the general reasoning behind expression in eq. (1) is quite intuitive, and is based on the normalized information distance, given by:

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

where x and y are two strings (or other data objects such as sequences, program source code, etc.) which are to be compared. $K(x)$ and $K(y)$ are the Kolmogorov complexities of the two strings individually, while $K(x, y)$ is the complexity of the combination of the two strings. The distance is hence a measure of the additional information which would be required to encode both strings x and y given that an encoding of the shorter of

the strings is already available. The division by $\max\{K(x), K(y)\}$ is a normalization term which ensures that the final value of the distance lies in the interval [0,1].

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

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

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

Substituting (3),(4) $\rightarrow$ (2) leads to the expression in eq. (1).

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

1) Instead of a general Web search engine, the prefix code length will be measured using hit counts obtained from a scientific database such as Google Scholar or Web of Science.
2) N is set to the number of hits returned in response to a search for "renewable+energy", as this represents the size of the body of literature dealing with renewable energy technologies.
3) We are only interested in term co-occurrences which are within the context of renewable energy; as such, to calculate the co-occurrence frequency $n_{i,j}$ between terms t_1 and t_2, the search term "'renewable+energy"+"t_1"+"t_2"' was submitted to the search engine. Admittedly this measure may result in some under-reporting of hit counts as the term "renewable+energy" may not explicitly appear in all relevant documents. However, overall it was deemed necessary as many of the keywords such as *solar*, *turkey* and *wind* are very broad and would admit many studies which are not associated with renewable energy.

As explained in [Cilibrasi and Vitányi, 2007], the motivation for devising the Google distance was to create an index which quantifies the degree of semantic dissimilarity between objects (words or phrases) which reflects their usage patterns in society at large. By exploiting the same intuition, it would be logical to assume that a similar measure which utilizes term co-occurrence patterns in the academic literature, instead of a general Web search engine, would be able to more appropriately characterize the similarity between technology related keywords in terms of their usage patterns in the scientific and technical community.

B. Asymmetric distances for detection of subclassing

One of the important properties of a distance measure is that it should be symmetric, i.e.: for a given distance function $d(,)$:

$$d(i,j) = d(j,i) \quad \forall i,j.$$

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

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

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

$$\overrightarrow{\text{NID}}(x,y) = \frac{K(x,y) - K(y)}{K(x)}. \tag{5}$$

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

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

$$\textit{information}(\text{elephant}) \subset \textit{information}(\text{circus elephant}),$$

$$\therefore K(\text{circus elephant}, \text{elephant}) - K(\text{circus elephant}) \approx 0.$$

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

Finally, as before, we can obtain a form of this equation suitable for use with search engines by substituting eqs. (3) and (4) into eq. (5), which yields the corresponding directional version of the NGD:

$$\overrightarrow{\text{NGD}}(t_x, t_y) = \frac{\log n_y - \log n_{x,y}}{\log N - \log n_x}, \tag{6}$$

It is now easy to check the validity of this intuition. Through the appropriate Google searches, we have:

Example 1

$$\begin{aligned}
n_{\text{elephant}} &= 80,300,000 \\
n_{\text{circus elephant}} &= 106,000 \\
n_{\text{circus elephant,elephant}} &= 91,800 \\
\therefore \overrightarrow{\text{NGD}}(\text{circus elephant,elephant}) &= \frac{\log 106,000 - \log 91,800}{\log 10^{10} - \log 80,300,000} \\
&= 0.03, \\
\overrightarrow{\text{NGD}}(\text{elephant,circus elephant}) &= \frac{\log 80,300,000 - \log 91,800}{\log 10^{10} - \log 106,000} \\
&= 0.59,
\end{aligned}$$

where, as suggested in [Cilibrasi and Vitányi, 2007], N can be approximated by any suitably large number. As can be seen, these figures correctly indicate that "circus elephant" is indeed a subclass of "elephant".

It should be noted, however, that this distance provides a measure of subclassing strictly in the context of term occurrences in academic texts; i.e. if the majority of texts in which term A occurs also contain term B (but not vice versa), this is an indication that term A is frequently researched in the context of term B, and is a "subclass" of term B only in this sense. In many cases this would correlate to other notions of subclassing, though there will definitely be exceptions.

To concretize this further, we look at two further examples, one of which is on classes of technology, and the other from biology (document counts in these examples are from Google Scholar):

Example 2

$$
\begin{aligned}
n_{\text{computer science}} &= 2,590,000 \\
n_{\text{artificial intelligence}} &= 977,000 \\
n_{\text{computer science,artificial intelligence}} &= 539,000 \\
\therefore \overrightarrow{\text{NGD}}(\text{artificial intelligence,computer science}) &= \frac{\log 977,000 - \log 539,000}{\log 10^{10} - \log 2,590,000} \\
&= 0.07, \\
\overrightarrow{\text{NGD}}(\text{computer science,artificial intelligence}) &= \frac{\log 2,590,000 - \log 539,000}{\log 10^{10} - \log 977,000} \\
&= 0.17,
\end{aligned}
$$

Example 3

$$
\begin{aligned}
n_{\text{mammal}} &= 560,000 \\
n_{\text{dog}} &= 1,290,000 \\
n_{\text{mammal,dog}} &= 69,800 \\
\therefore \overrightarrow{\text{NGD}}(\text{dog,mammal}) &= \frac{\log 1,290,000 - \log 69,800}{\log 10^{10} - \log 560,000} \\
&= 0.30, \\
\overrightarrow{\text{NGD}}(\text{mammal,dog}) &= \frac{\log 560,000 - \log 69,800}{\log 10^{10} - \log 1,290,000} \\
&= 0.23,
\end{aligned}
$$

In Example 2 (*computer science*, *artificial intelligence*), the $\overrightarrow{NGD}$ correctly indicates that "artificial intelligence" is a subclass of "computer science". However, Example 3 (*mammal*, *dog*) is an exception to this rule in that the $\overrightarrow{NGD}$ values now indicate that "dog" is a subclass of "mammal". Two issues seem to be at the root of this problem; firstly, it is clear that "dog" is a much more common topic of research than "mammal", even though biologically it is a member of the class of mammals. More critically, articles about dogs frequently do

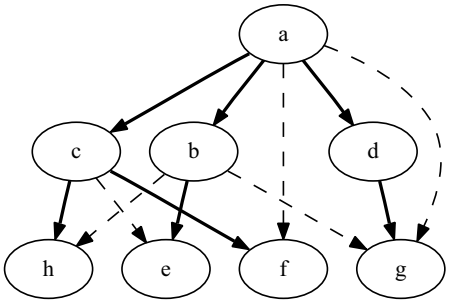


Fig. 1. Directed graph. The solid lines show one of a number of arborescences in the graph

not even contain the term "mammal" since this is already well known, and may not be relevant to the issue being studied. We concede that this is an inherent shortcoming of this measure of distance, and one which would be difficult to overcome without significantly extending the proposed approach, which is beyond the current scope. On the other hand, our method does produce useful results in many cases, as illustrated in Examples 1 and 2 above, as well as in our pilot study, described later.

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

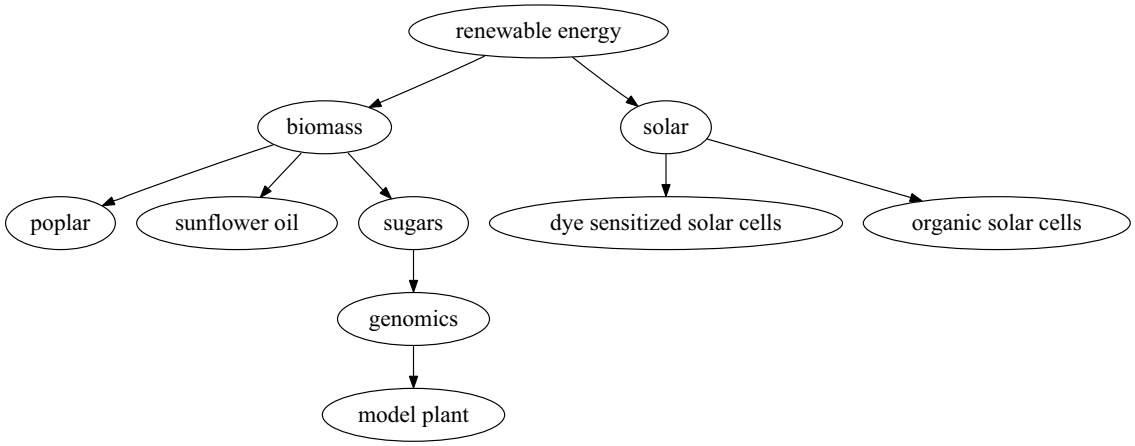
$$w\left[e(v,w)\right] = \overrightarrow{\text{NGD}}(v,w). \tag{7}$$

In this context, a keyword taxonomy is represented by a subgraph $(\mathcal{V}, \mathcal{E}^*)$, where:

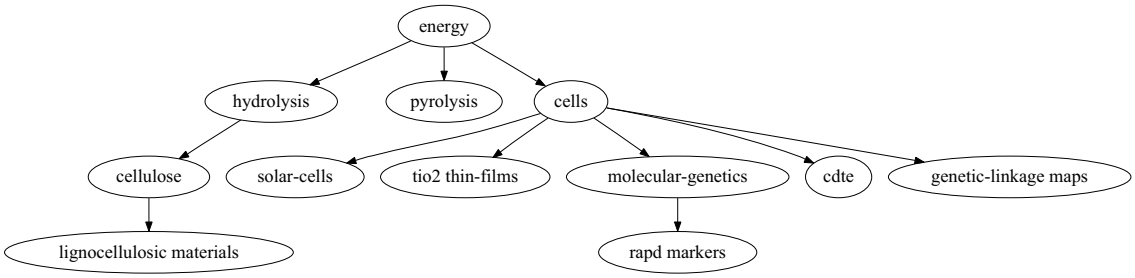
1) $\mathcal{E}^* \subset \mathcal{E}$, $|\mathcal{E}^*| = |\mathcal{V}| - 1$
2) All nodes except one have exactly one incoming edge.
3) $(\mathcal{V}, \mathcal{E}^*)$ is connected, and there are no cycles.

In graph theory this construct is known as an *arborescence*, which is basically the directed equivalent of a spanning tree (fig.1). However, for any digraph there could be a very large number of such arborescences, any one of which could potentially be a valid keyword taxonomy. To solve this, we choose to follow the principle of parsimony in suggesting that the arborescence with the *minimum total edge weight* provides the best possible organization of the terms. In graph theory the problem of finding this arborescence is referred to as the minimum arborescence problem.

To demonstrate that this principle works, it is used to automatically infer the taxonomic structure of two small selections of renewable energy related keywords, and these are shown in fig. 2. The resulting topic trees show that the terms have been organized into hierarchies that approximately reflect the inter-dependencies between the terms.



(a) Example 1



(b) Example 2

Fig. 2. Sample renewable energy taxonomies

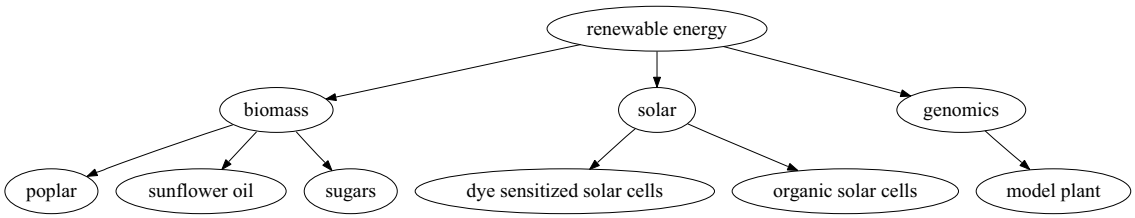
C. Weighted cost functions

As mentioned above, when searching for the most likely taxonomy of keyword terms, the selection criteria is the total weight (i.e. distance values) of the edges in the corresponding arborescence.

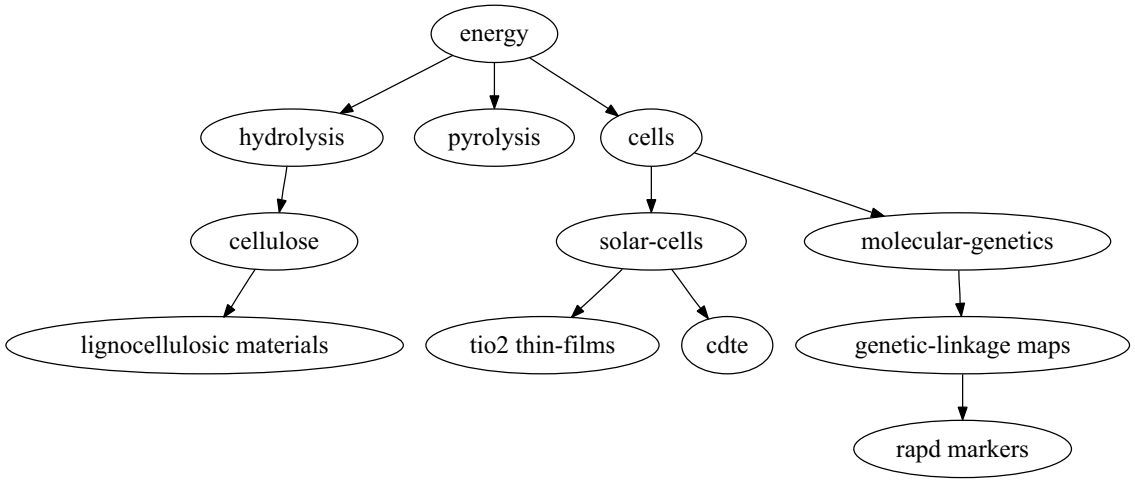
Using the cost function derived from eq. (6) often resulted in local structure which did not reflect the actual inheritance structure. In a noiseless environment this would not be a problem but in practice there are a number of situations where this reduces the accuracy of the results.

For example, consider the taxonomy in fig. 2(a). We see that *sugars* has been classified under the biomass subtree. However, *genomics* and *model plant* have subsequently been placed as subclasses of sugars. However, it would appear that the aspect of genomics research related to sugars may be separate from the subset of research in sugars related to biomass. We can check this by studying the directional distances: $\overrightarrow{\mathrm{NGD}}(sugars, biomass) = 0.237$, while $\overrightarrow{\mathrm{NGD}}(genomics, sugars) = 0.336$, both of which are the smallest values in the respective rows of the distance matrix. However, $\overrightarrow{\mathrm{NGD}}(genomics, biomass) = 0.462$ which is somewhat greater than $\overrightarrow{\mathrm{NGD}}(genomics, renewable\ energy) = 0.395$, suggesting that perhaps the genomics subtree might be better portrayed as a separate branch of research from biomass.

Another example is shown in fig. 2(b), where the term *cell* has attracted a large number of direct descendants: *solar-cells, TiO_2 thin films, molecular genetics, CdTe, genetic-linkage maps*. This is a problem which is



(a) Example 1



(b) Example 2

Fig. 3. Sample taxonomies generated using the weighted cost function

frequently encountered, in which very broad terms (such as *cells*) tend to dominate the subclassing process, resulting in extremely flat hierarchies. A further complication is that the keyword *cells* has two senses: solar "cells", and biological "cells".

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

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

$$f_{\mathcal{V}}(\mathcal{E}^*) = \sum_{v \in \mathcal{V}} \frac{\sum_{i=1}^{n} \alpha_i \overrightarrow{\mathrm{NGD}}(v, v^i_{\mathcal{E}^*})}{\sum_{i=1}^{n} \alpha_i}, \tag{8}$$

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

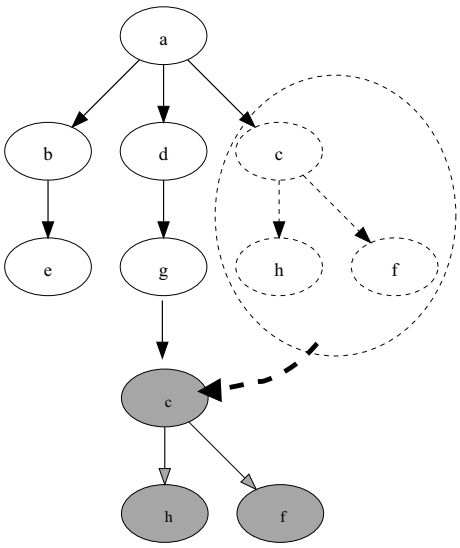


Fig. 4. Taxonomy tree mutation operator. The dashed lines denote nodes and edges which are to be removed.

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

1) **Uniform weighting** $\alpha_{1\ldots n} = 1$
2) **Linear weighting** $\alpha_i = n - i$
3) **Exponential weigthing** $\alpha_i = \frac{1}{2}^{i-1}$

As an example, taxonomies containing the same keywords have been generated by optimizing the linear weighted cost function, and are shown in fig. 3 (optimization was done using a genetic algorithm, which is discussed in the following section). As can be seen from these two figures, the use of the weighted cost function produces some noticeable improvements in the resulting taxonomies. In particular, the sub-tree *genomics→model plant* in fig.3(a) has been directly connected to the root node, while in In fig.3(b), the sub-tree descending from *cells* is now more structured (in fig.2(b), this subtree was mainly a flat hierarchy. Accordingly, the two sense of *cells* have now been appropriately divided into two separate subtrees, each of which shows a reasonable inheritance structure.

III. Methods and data

A. Genetic algorithms for taxonomy optimization

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

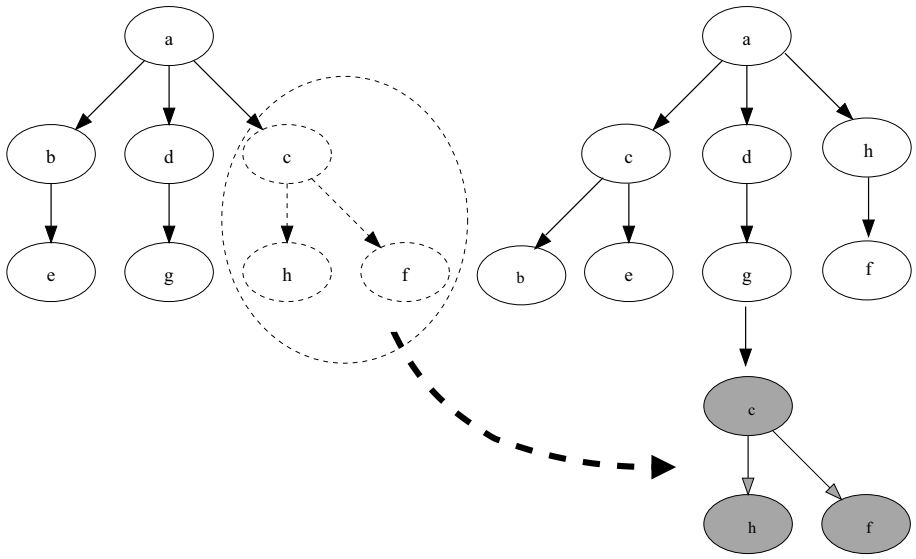


Fig. 5. Taxonomy tree crossover operator (stage 1). The dashed lines denote nodes and edges which are to be removed.

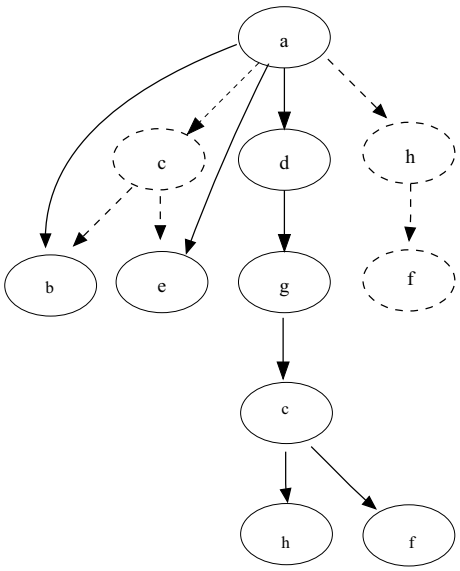


Fig. 6. Chromosome repair process. The dashed lines denote nodes and edges which are to be removed.

In particular, Edmond's algorithm is inapplicable for the cost function in eq. (8), nor does there appear to be any efficient algorithm for finding the global optimum of this function. As the number of possible taxonomies grows exponentially with the number of nodes, exhaustive searches quickly become computationally infeasible.

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

with some success, though in these previous studies the GAs were applied to problems involving undirected trees.

The basic components of any GA are:

1) A method for encoding a full set of the parameters to be optimized, where each encoded parameter set is called a "chromosome". For this study, the chromosomes were simply the connection matrices representing the digraphs. A connection matrix is a matrix with elements $c_{i,j}$ where $c_{i,j} = 1$ indicates that there is an edge linking node i to node j, while $c_{i,j} = 0$ means that there is no connection between the two nodes. In GA terminology, each chromosome is sometimes associated to an "individual".
2) A fitness function for evaluating each chromosome. As discussed previously, in this study the GAs will be used to test the weighted subclassing cost functions.
3) A set of *cross-over* and *mutation* operations on the chromosomes. Traditionally, GAs have been based on linear, binary chromosomes but this would be inappropriate in the current application where the natural representation of parameters is as a tree structure. In [Li and Bouchebaba, 2000], [Li, 2001], the mutation and crossover operations were designed to preserve *paths* in the trees, while in [Raidl, 2000], operators based on individual edges were used instead. These operations were certainly better suited for use with trees but in the present context it was deemed more appropriate to use *subtree*-preserving operations as these represent specific subdomains. In more detail, the chromosome transformations adopted were as follows:
   - *Mutation* - the mutation procedure operates on individual trees. A random subtree is moved from one point of the hierarchy to another randomly selected point in the same tree (fig.4).
   - The *Cross-over* procedure accepts pairs of trees at a time. The operation comprises two stages: in the first stage, a random subtree is selected from each of the original trees and is transplanted onto a random point in the other tree (fig. 5). However, this process invalidates the original taxonomies as the transplanted nodes would now appear twice in the same taxonomy. To resolve this, the transplantation stage is immediately followed by a chromosome repair process (fig. 6) where the *originals* from the duplicated nodes are removed and all descendants thereof promoted to the ancestor nodes at the next level in the hierarchy.

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

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

B. Data collection

To conduct the pilot study on renewable energy, energy related keywords were extracted using ISI Web of Science's database in the following manner: a search for "renewable+energy" was submitted, and the matching

publications were sorted according to citation frequency, then the top 35 hits were used. In total, 72 "Author Keywords", i.e. keywords specified by the authors were extracted (the complete lists of keywords are provided in Appendix II of this paper).

Once the keywords were collected, the distances discussed in II-A could be calculated where, as discussed, hit counts obtained from the Google scholar search engine were used. A number of other alternatives were considered including the Web of Science, Inspec, Ingenta, Springer and IEEE databases. However, our preliminary survey of these databases indicated that zero hits were returned for a large number of keyword pairs. There appeared to be two main reasons for this observation: Firstly, most of these search engines simply did not index a large enough collection to provide ample coverage of the more specialized of the keywords that were in the list; furthermore, not all of the search engines allowed full text searches (the Web of Science database, for example, only allows searching by keywords or topics) - while sufficient for literature searches and reviews, keyword searches simply did not provide sufficient data for our purposes.

Even when using Google scholar, there were also a number of keyword pairs for which there were no hits at all. This can cause serious problems it will cause the logarithms of $n_{i,j}$ in eq.5 to be undefined. This can be viewed as a type of round-off error as $n_{i,j}$ is used to estimate the probability of co-occurrence of the terms t_i and t_j - as hit counts can only take integer values, small values of this probability could very possibly result in $n_{i,j} = 0$. To resolve this, we set $n'_{i,j} = \max\{\epsilon, n_{i,j}\}$, where ϵ is the machine precision (in our implementation $\epsilon = 2.22 \times 10^{-16}$), $n'_{i,j}$ is then used in place of $n_{i,j}$.

IV. Results

Having described the proposed approach as well as the specific procedures and techniques adopted, we are now in a position to discuss the experimental results and subsequently to evaluate the effectiveness of these methods.

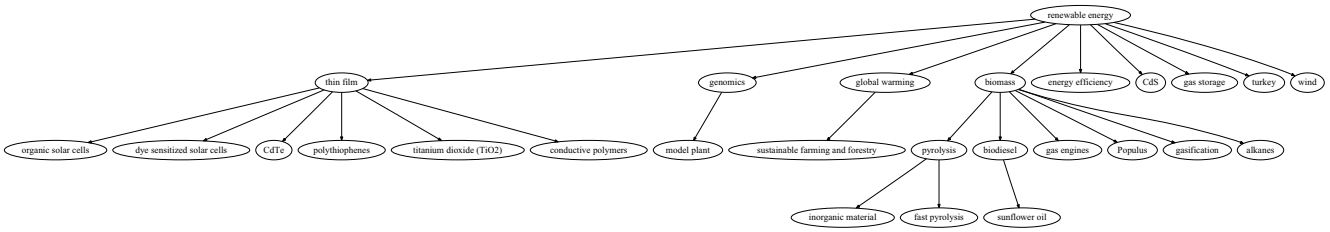
As discussed in the previous section, 72 keywords were collected using the ISI Web of Science "Author Keywords" feature. To facilitate presentation and analysis of the results, this collection was then randomly divided into two subsets - Set 1 contains 35 keywords, and Set 2 contained the remaining 37 keywords. In addition, any occurrences of the stop-words described in section III-B were also removed before analysis was carried out. In the following subsections the observations obtained which each of the sets are discussed in greater detail.

A. Set 1

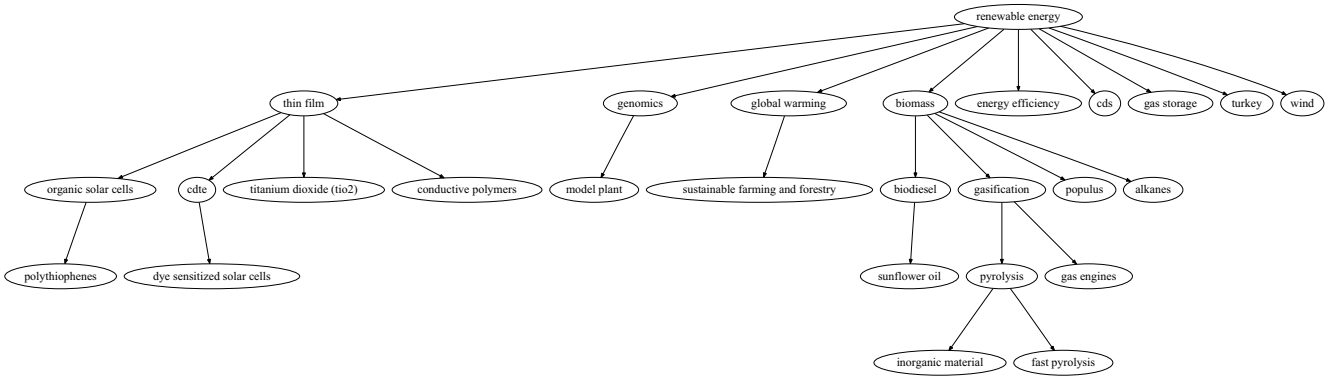
The proposed methods were first applied to the keywords in Set 1. Taxonomies were generated using Edmond's algorithm and GA optimization using first the uniform weighting then the exponential weighting functions; these are presented in fig. 7.

The main observations were:

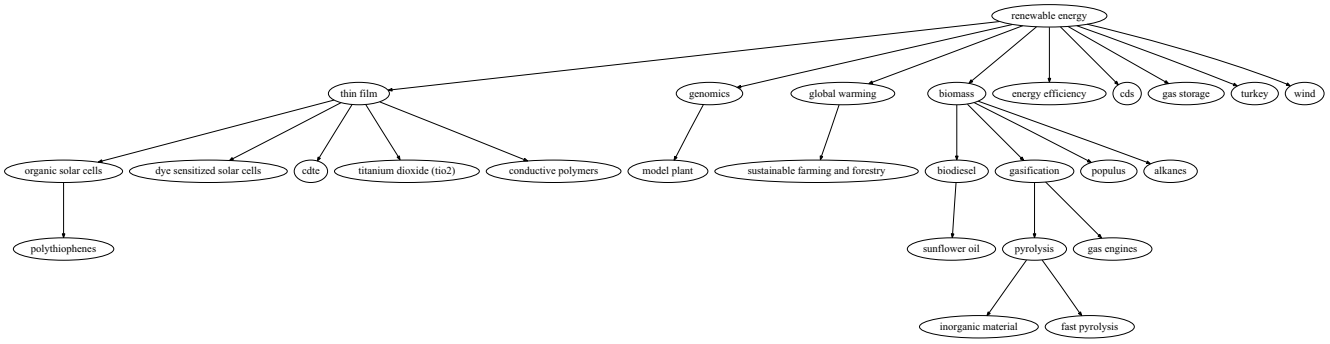
1) In general, the generated taxonomies appear to capture the high level orderings of the terms in the collection, at least to a reasonable degree of accuracy. In particular, there were two big clusters: one dedicated to biomass-related technologies and the other to technologies associated with thin-film solar



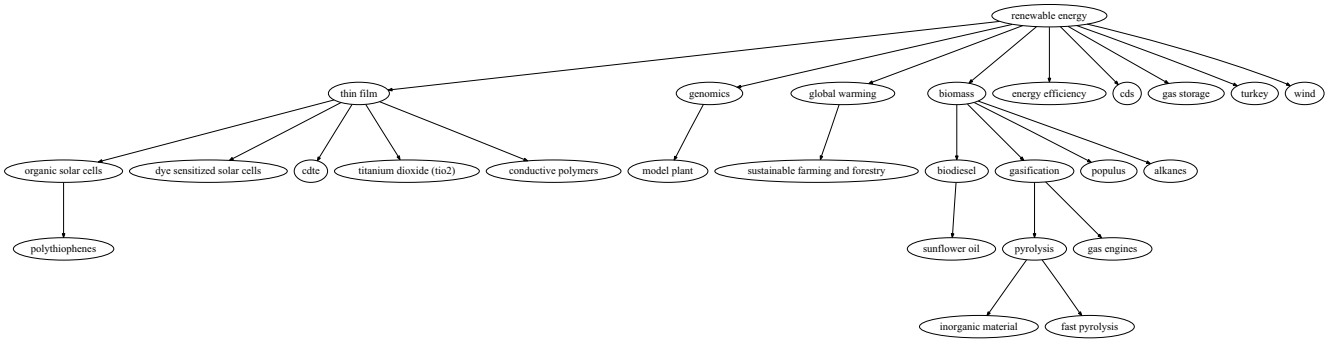
(a) Edmonds algorithm



(b) Uniform weights

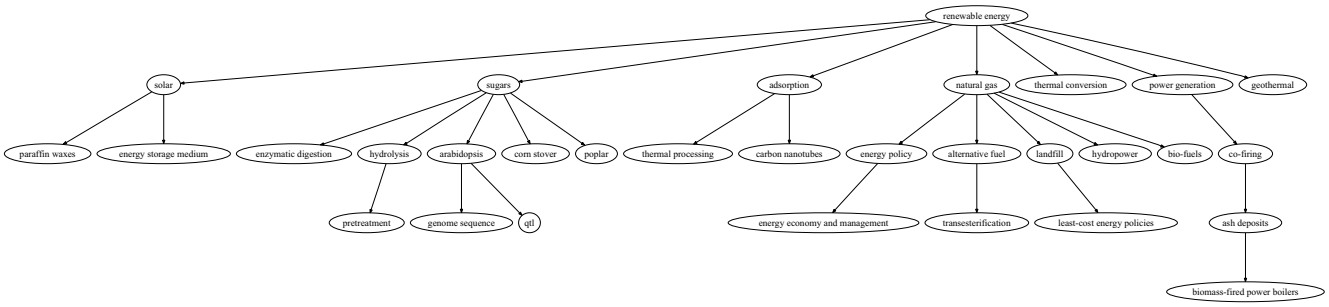


(c) Linearly decaying weights

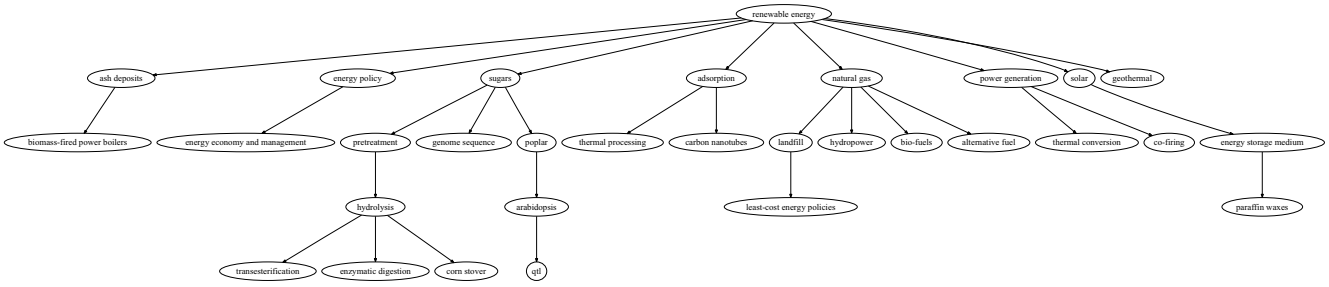


(d) Exponentially decaying weights

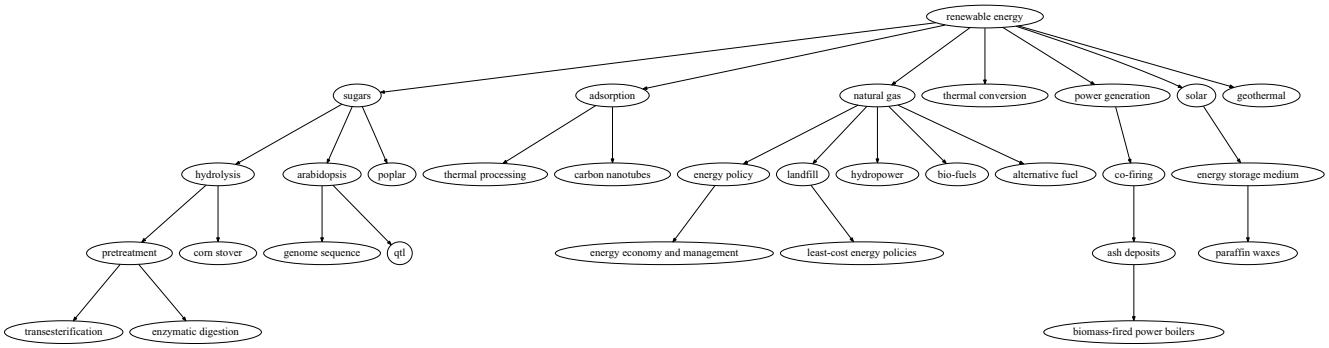
Fig. 7. Automatically generated taxonomies: **Set 1**



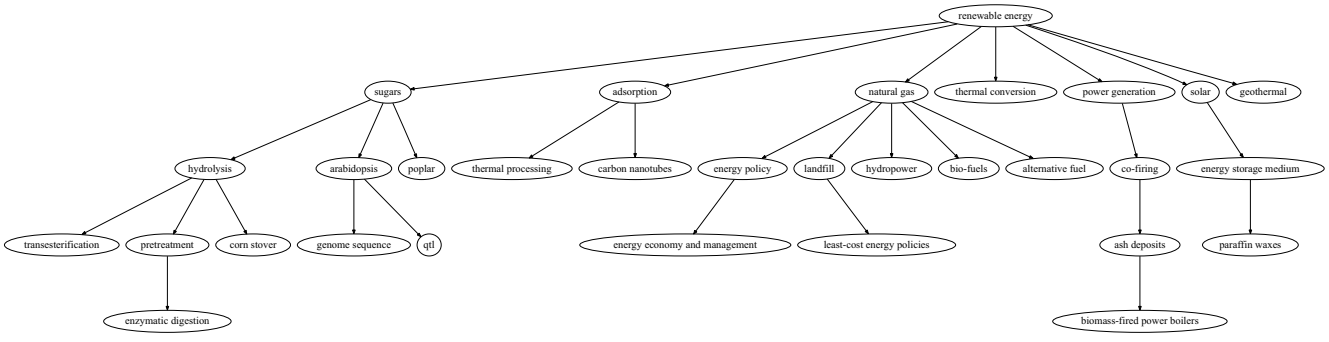
(a) Edmonds algorithm



(b) Uniform weights



(c) Linearly decaying weights



(d) Exponentially decaying weights

Fig. 8. Automatically generated taxonomies: **Set 2**

cells. There were also other nodes and "micro-clusters" which descended directly from the root, notably the pairs {*genomics→model plant*} (molecular genetics related) and {*global warming→sustainable farming and forestry*} (policy related).

2) The results obtained using the weighted schemes were almost identical - when α_i was set to linearly and exponentially decaying values, identical results were obtained. When using uniform weights, the results were still similar but there was a change in the *thin film* subtree, where *dye sensitized solar cells* was classified as a subclass of CdTe instead of being a direct subclass of *thin film*.
3) However, there is a bigger difference between the taxonomy generated using Edmond's algorithms (fig.7(a)) and those generated using the genetic algorithm. While the overall structure remained the same, the former had a flatter hierarchy, with much less subtree formation.

   Consider, especially, the *biomass* subtree; in fig.7(a), six edges originate from this node, only two of which have any further descendants. In contrast, in fig.7(b) (uniform weights), four nodes descend directly from *biomass*, namely *biodiesel, gasification, populus* and *alkanes*. Of these, *biodiesel* is further linked to *sunflower oil*, which can be used to create biodiesel via transesterification. Similarly, *gasification* is joined to a pair of related concepts - *pyrolysis* and *gas engines*.

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

B. Set 2

Next, the second set of keywords (Set 2) was organized into a taxonomy using the proposed approach. The resulting graphs are shown in fig. 8.

Our observations on these graphs are:

1) As before, the taxonomies show a number of significant clusters, which include *solar*, *sugars*, *adsorption*, *natural gas* and *power generation*.
2) However, it was observed that there is much less consistency amongst the four taxonomies.
3) As before, the results using Edmond's algorithm produced a slightly flatter hierarchy than when using the weighted cost functions; however, this difference was less pronounced than in the case of Set 1.
4) The taxonomies created when α_i was linearly and exponentially decreasing were very similar, though this time there was one very minor difference between them.
5) The *natural gas* subtree is somewhat mixed in its composition (which also changes significantly in the four taxonomies for Set 2), and appears to be a kind of "catch-all" cluster for a number of orphaned terms. While a more reliable analysis would require further domain knowledge, an informal scan of the academic literature on this subject suggests that this problem occurred as a result of a number of factors: firstly, *natural gas* is an extremely common term in renewable energy, while technical research that focusses specifically on natural gas is relatively less common. Instead, we notice that this term frequently appears in articles that are broader in scope, such as review papers and papers on various strategic issues such as global warming, energy markets and the like. This allows the term to attract a broad range of "subclasses"

which may not easily fit into other sections of these taxonomies. In particular, note that many of the terms descended from *natural gas* are themselves fairly broad in nature - and would likely appear in similar publications.

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

V. Discussions

This paper presented a novel approach for automatically organizing selections of keyword into taxonomies. In addition to being an important step in the ontology creation process, these techniques can be hugely useful to researchers seeking a better understanding of the overall research landscape associated with the collection being studied.

On the other hand, the results obtained indicate that there are many technical problems which need to be overcome before this methodology can be used in a fully-automated manner. The main issues include:

1) Complexity - as with many other inverse problems, inferring the underlying taxonomy of a collection of keywords is ill-posed: even ontologies created by subject matter experts can show significant variability. This is because the exact structure and organization of a taxonomy is very sujective and depends heavily on the perspective and motivations of the developer.
2) Inconsistent quality of data; data obtained from publicly available sources are unregulated and are frequently noisy; this further underscores the need for appropriate filtering and data cleaning mechanisms.
3) Non-uniform coverage - the number of hits returned for very general or high-profile keywords such as “energy” or “efficiency” was a lot greater than for more specialized topics. This is unfortunate as it is often these topics which are of the greater interest to researchers. One way in which we hope to overcome this problem is by aggregating information from a larger variety of sources, examples of which include technical report and patent databases and possibly even mainstream media and blogs.

That said, the methods described in this paper were only intended as an early demonstration of the proposed approach, and in spite of the above-mentioned problems, we believe that the results described here already demonstrate the potential of the approach.

It must also be conceded that while promising, the results were still far from perfect and contained a number of irregularities as described in the paper. These may be viewed from a number of perspectives; on the one hand, they could be manifestations of hitherto unknown relationships or underlying correlations which may

only be understood after a more in-depth study of these results. On the other hand, it is difficult to think of these results as either "right" or "wrong" - the $\overrightarrow{NGD}$ is a numerical index derived from the term co-occurrence frequencies, which in turn depend on the data available to the algorithm - nothing more, nothing less; under the correct circumstances and provided that our assumptions are sufficiently met, it can be very useful as a means of detecting subclassing. Certainly, from the results obtained so far it would appear that these requirements are satisfied for at least a reasonable proportion of the time. However, under less favourable conditions, it can return values which are difficult to understand or to explain, as has also been observed in some of the examples presented here.

Avenues for future research include working more closely with domain experts to improve and validate the results produced using the proposed methodology. We also plan to screen alternative distance measures, which might provide a more robust indication of subclassing. An additional issue would be to test a broader range of optimization techniques and genetic operators.

APPENDIX I
EDMOND'S ALGORITHM

Edmond's algorithm [Korte and Vygen, 2006] provides an efficient means of finding the minimum arborescence. Briefly, this is as follows:

Algorithm *Edmonds*$(\mathcal{V}, \mathcal{E})$

Input: A digraph consisting of vertices $\mathcal{V}$ and edges $\mathcal{E}$

Output: Minimum weight arborescence $\mathcal{E}^*$

1. $\mathcal{E}^* \leftarrow \emptyset$, $\mathcal{V}^* \leftarrow \mathcal{V}$
2. **for** $v \in \mathcal{V}^*$
3. **do**
4. Identify $u = \text{argmin}_u \{w[e(u,v)] : u \in \mathcal{V}, u \neq v\}$
5. $\mathcal{E}^* \leftarrow \mathcal{E}^* + \{e(u,v)\}$
6. **if** no cycles formed,
7. Expand pseudo-nodes (if any), and return $\mathcal{E}^*$
8. **else**
9. Contract the nodes $\mathcal{V}' \subseteq \mathcal{V}$ in each cycle into a pseudo-node v'
10. $\mathcal{V}^* \leftarrow \mathcal{V}^* - \mathcal{V}'$, $\mathcal{V}^* \leftarrow \mathcal{V}^* + \{v'\}$
11. Replace all *incoming* edges with:

$$w[e(u,v')] = w[e(u,v)] - w[e(x(v),v)] + \sum_{\{e:e\in\mathcal{E}', e\neq x(v)\}} w[e],$$

where, $x(v)$ is the immediate ancestor of node v and $\mathcal{E}'$ is the set of edges in pseudonode v'.

12. For each *outgoing* edge, set:

$$w[e(v',u)] = min_{v\in\mathcal{V}'} w[e(v,u)]$$

13. Repeat from step 2 until all cycles have been eliminated

APPENDIX II

RENEWABLE ENERGY RELATED KEYWORDS

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

REFERENCES

[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.

[An et al., 2004] An, Y., Janssen, J., and Milios, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6):664–678.

[Blaschke and Valencia, 2002] Blaschke, C. and Valencia, A. (2002). Automatic ontology construction from the literature. *Genome informatics.*, 13:201–213.

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

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

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

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

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

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

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

[Kandylas et al., 2008] Kandylas, V., Upham, S., and Ungar, L. (2008). Finding cohesive clusters for analyzing knowledge communities. *Knowledge and Information Systems*, 17(3):335–354.

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

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

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

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

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

[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.

[Lu et al., 2007] Lu, W., Janssen, J., Milios, E., Japkowicz, N., and Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge and Information Systems*, 11(1):105–129.

[Makrehchi and Kamel, 2007] Makrehchi, M. and Kamel, M. S. (2007). Automatic taxonomy extraction using google and term dependency. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 321–325, Washington, DC, USA. IEEE Computer Society.

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

[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.

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

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

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

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

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

Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation

Andreas Henschel
Wei Lee Woon
Thomas Wachter
Stuart Madnick

Working Paper CISL# 2009-12

September 2009

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

Comparison of generality based algorithm variants for automatic taxonomy generation

Andreas Henschel
Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE
ahenschel@masdar.ac.ae

Wei Lee Woon
Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE
wwoon@masdar.ac.ae

Thomas Wächter
Technische Universität Dresden
Tatzberg 47-52, 01307 Dresden, Germany
thomas.waechter@biotec.tu-dresden.de

Stuart Madnick
Massachusetts Institute of Technology
77 Mass. Ave.,Building E53-321
Cambridge, MA 02139-4307, U.S.A.
smadnick@mit.edu

Abstract

We compare a family of algorithms for the automatic generation of taxonomies by adapting the Heymann-algorithm in various ways. The core algorithm determines the generality of terms and iteratively inserts them in a growing taxonomy. Variants of the algorithm are created by altering the way and the frequency, generality of terms is calculated. We analyse the performance and the complexity of the variants combined with a systematic threshold evaluation on a set of seven manually created benchmark sets. As a result, betweenness centrality calculated on unweighted similarity graphs often performs best but requires threshold fine-tuning and is computationally more expensive than closeness centrality. Finally, we show how an entropy-based filter can lead to more precise taxonomies.

1. Introduction

Taxonomies for scientific research bodies facilitate the organisation of knowledge. They are used in Information Retrieval and Text Mining where it is beneficial to abstract from plain words to hierarchical concepts, which allows to structure document databases semantically. Immediate applications are Ontology based searching [10], a successfully applied search engine for biomedical literature [4] and emerging trend detection [3].

Manual taxonomy construction is accurate but is unsuitable for many resources that contain vast amounts of text documents. Further, it is desirable to deterministically and objectively develop taxonomies in order to provide consistent maintenance, which is not guaranteed with nondeterministic algorithms or subjective curators.

To extract subsumption (taxonomic) relationships from text, there are two classes of approaches described in the literature: syntactic patterns such as *'A' is a 'B'* ([8]) and statistical methods (e.g. [13]). Both classes rely on the distributional hypothesis introduced by Harris [7], which defines that two words which appear in many similar linguistic contexts are semantically similar. A promising approach among the latter class is the algorithm developed in [9] which is simple, fast and extensible, and hence can include ideas from various approaches. Although it was originally designed for tagging systems in social web communities, it can be adapted to general literature databases using co-occurrence of terms as the base for expressing term similarity. In [18] it was shown that by utilising the co-occurrence frequencies between a collection of representative keywords, it is possible to infer the overall taxonomy of a given domain of research. A similar approach is presented by [12], where the authors propose a subsumption criterion for terms based on conditional probabilities for their co-occurrences. Other term distance measures employed are citation based, collaboration pattern based as well as more elaborate techniques of context similarity.

The remainder of this document is organised as follows: we elucidate several techniques originated from the Heymann algorithm, including generality ordering methods, various distance measures, weighting schemes and reranking. The algorithms are systematically compared using seven benchmarks derived from a manually created ontology of medical terms. Finally we show, how cautious insertion into a taxonomy can improve the precision without

worsening the F-measure.

2. Systematic comparison of algorithms

2.1. Creation of MeSH benchmark sets

Quality assessment of taxonomy generation methods is preferably carried out using gold standard taxonomies. Medical Subject Headings (MeSH) is a man-curated ontology for medical terms [14]. It is well suited as a benchmark to test the ability of an algorithm to reproduce a gold standard. We focus on several diverse branches in order to avoid over-fitting. For the automatic comparison of a manually and automatically generated taxonomies, the input terms are taken from the MeSH benchmarks. This poses a simplification of the overall taxonomy creation, where terms are selected using various methods (see e.g.[5, 1])

We then measure the precision by counting how many direct links of the original taxonomy are reproduced by the algorithm. Further we consider those links that are not only direct parent-child related but also grandchildren or great-grandchildren (upper part in Fig. 2) in the original benchmark.

Occurrences are detected in the abstracts of 18 Million articles from Pubmed (a literature database for the life sciences), using stemming and term alignment ([4]).

2.2. Heymann-Algorithm

The taxonomy creation algorithm presented in [9] (Heymann-Algorithm) was originally intended for social networks where users annotate documents or images with keywords. The algorithm is fast, deterministic and easily extensible. Each keyword or "tag" is associated with a vector that contains the frequencies of annotations for all documents. These tag vectors are then comparable, e.g. using cosine similarity. We adapt the algorithm to general taxonomy creation from scientific literature using binary tag vectors.

The algorithm consists of two stages: the first creates a similarity graph of tags, from which an order of centrality for the tags is derived. Obeying this order and starting from the most general tag, the tags are inserted to a growing taxonomy by attaching tags to either the most similar tag or to the taxonomy root.

Two thresholds are used in the algorithm: first, the value above which an edge is permitted to the similarity graph (τ_S) filters very small similarities that might have occurred by chance during the generality calculation. Second, the similarity above which a node is attached to its most similar non-root node rather than the root (τ_R) influences the topology of the taxonomy. An example of a generated taxonomy is shown in Figure 1.

2.3. Algorithm modifications

2.3.1 Term generality derived from centrality in similarity graphs

A set of n terms gives rise to a similarity graph $G = (V, E)$ where the nodes V represent terms and the edges E are similarities as provided by the similarity measure, see section 2.3.2. Generality can be deducted from a terms' centrality in such a similarity graph. A variety of centrality measures exists. Amongst them betweenness and closeness centrality are elaborate, global measures and therefore subject to further scrutiny.

Betweenness centrality c_B for a node v is defined as:

$$c_B(v) = \sum_{\substack{v \in V \setminus \{s,t\} \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a vertex v. The complexity of the betweenness centrality is $O(n^3)$. A fast algorithm for unweighted graphs of complexity $O(ne)$ (e is the number of edges, which could be $O(n^2)$ in fully connected graphs, but can be less in other graph types) is given in [2] and implemented e.g. in [6], which we use in our benchmark system.

Closeness centrality c_C for a node v is given as :

$$c_C(v) = \frac{1}{\sum_{t \in V \setminus \{v\}} 1 - \mathrm{sim}(v,t)}. \tag{2}$$

with $\mathrm{sim}(v,t)$ being the similarity between nodes t and v. The complexity is $O(n^2)$.

Considering graph-theoretical aspects: Edge weights and disconnected graphs Betweenness and closeness centrality can be calculated using weighted or unweighted graphs. We investigate both types.

Figure 2 compares the precision of the Heymann-Algorithm variants with several centrality calculations in dependence of τ_S. Various values for τ_R are probed (see Supplementary Material) but are of less influence.

2.3.2 Vector based term similarity

Originally Heymann et. al used vectors $\mathbf{x}_t = [x_1, \ldots, x_N]$ of length equal to the number of documents N, where x_i describes, how many times a numbered document i in a user community has been annotated with term t. We adapt this to binary term-vectors (or set representations) indicating whether a term occurs in a document (1) or not (0). Standard cosine vector similarity is therefore applicable.

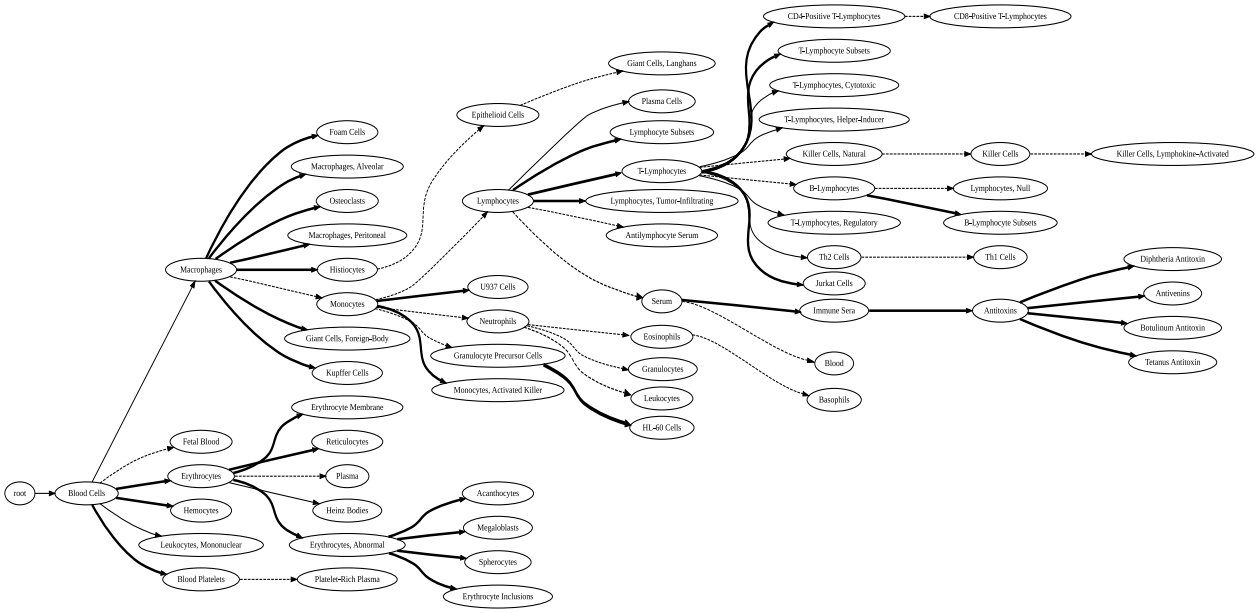


Figure 1. A generated taxonomy for "Blood". The fat links are correct wrt. the MeSH benchmark, semi-fat links are in grand- or great-grandchild relation in MeSH.

2.3.3 Reranking

A further modification to the Heymann algorithm is the intermediate reranking of the remaining terms wrt. their centrality after inserting a term to the taxonomy. Note, that this step increases the algorithm complexity since the centrality calculation is run for every inserted term ($O(n^3)$ and $O(n^4)$ for closeness and betweenness centrality, resp.).

2.3.4 Entropy of similarities

The basic Heymann algorithm attaches nodes to the most similar node in the growing taxonomy. Often terms, in particular non-specific or ambiguous terms, exhibit similarities to many subjects. The Entropy E_S, given in equation 3, is an information theoretical concept that can be used to quantify that intuition and hence accounts for the uncertainty of adding a node. This edge annotation can later be used for quality assessment and semi-automatic curation.

$$E_S(j) = -\sum_{i \in T}^{n} s_{ij} \log_b s_{ij} \text{ for } s_{ij} > 0 \qquad (3)$$

where s_{ij} are the similarities of the node to be inserted j and the nodes i that are already in the taxonomy T. Similarities are normalised such that their sum yields 1. Thus a node j_0 being similar to exactly one node but having 0 similarity to all other nodes leads to a minimal entropy of 0, whereas a maximal entropy of 1 is reached when all nodes are equally similar to the node to be inserted.

3. Results

3.1. Term generality and systematic threshold evaluation

The benchmark sets were scrutinised with respect to algorithm variants (centrality, rooting threshold τ_R, similarity graph threshold τ_S). One example is given in Figure 2. It shows that unweighted closeness and betweenness centrality yield the best results for $0 < \tau_S \leq 0.1$. This finding was consistent with most benchmarks. Exceptions occurred for the "Blood" and "Cardiovascular system" benchmarks, where single peaks of weighted closeness scored highest (Supplementary material).

The threshold for attaching a term to the root τ_R has been systematically probed in the range of $0 - 0.06$ with stepsize 0.005 and best results were consistently achieved with a very small value, i.e. avoiding node-attachments to the root as much as possible. Note that a histogram of all similarities revealed that most similarities are below 0.01.

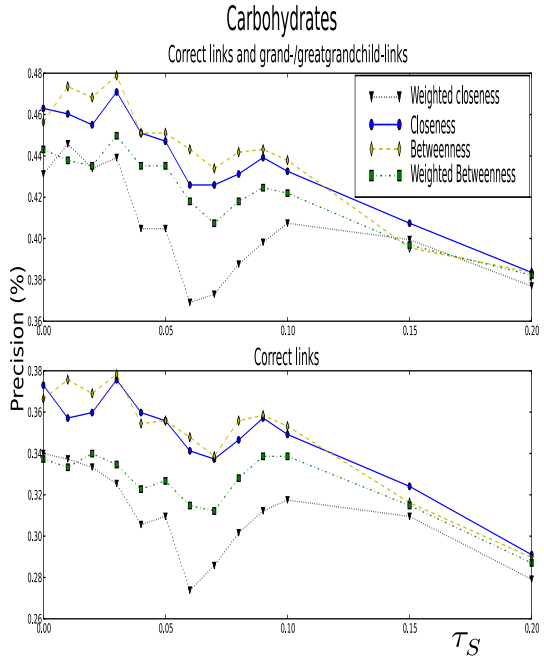


Figure 2. Precision curves for centrality variants for the MeSH-benchmark "Carbohydrates".

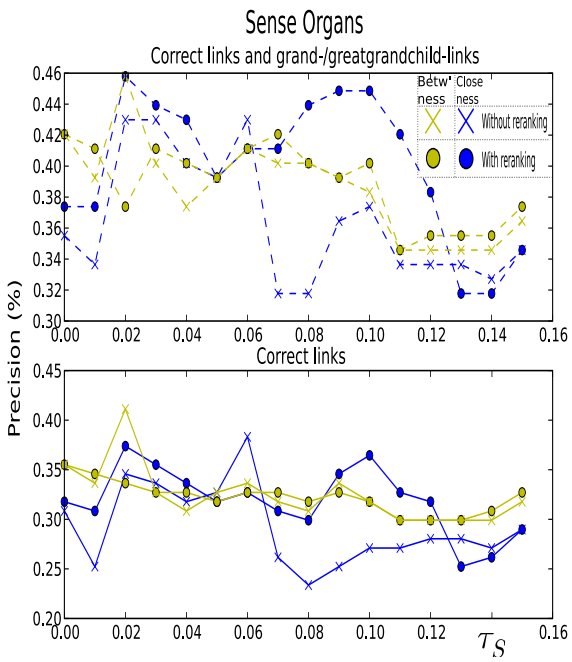


Figure 3. Benchmark: Sense Organs, with and without generality reranking

3.2. Intermediate reranking of term generality

Depending on the similarity graph threshold τ_S, the intermediate reranking improves precision in 46% of the cases, decreases precision in 23% and achieves equal precision in 30% of the cases.

3.3. Entropy based filtering improves precision

According to [17], taxonomy generation algorithms usually achieve only 40-50% precision on general benchmarks. Velardi et al. therefore suggests in [16] to follow a semi-automatic approach including systematic human validation steps. As a basis for hand-curated taxonomies, precision becomes paramount when automatically generating draft taxonomies. We therefore monitor the F-measure, which trades off precision vs. recall and is frequently used in information theory to evaluate performance based on a single number [15].

$$F_\beta = (1+\beta^2) \cdot \frac{(\text{precision} \cdot \text{recall})}{\beta^2 \cdot \text{precision} + \text{recall}} \qquad (4)$$

In order to appreciate precision, the $F_{0.5}$-measure for example values precision twice as important as recall. Omitting links comes to the expense of decreasing the recall. Yet, we argue that the omissions are justified as long as the F-measure improves.

By filtering high entropy links with $E_S > 0.7$, precision increases most notably for benchmark "Blood" (from 60% to 81%), "Carbohydrates" (from 38% to 43%) and "Fungi" (from 31% to 39%), see Figure 4.

The precision of all other benchmarks improves as well, but to a smaller extend. Larger margins are possible with other thresholds but might yield in over-fitting to the given benchmarks.

4. Conclusion

Unweighted betweenness centrality generally performs best but often only marginally better than the faster unweighted closeness centrality. Neither method strictly dominates the other and both are dependent on fine-tuning of the similarity graph threshold. A good choice for τ_S is not obvious but should be a value between 0 and 0.1. Both methods are complementary in the sense that their highest scoring taxonomies are not identical. A consensus-based meta-algorithm can benefit from this fact by only including the links both methods agree on.

Using weighted similarity graphs rarely improved the performance and hence did not justify the higher compu-

tational cost. Moreover, they fluctuate stronger wrt. τ_S.

Reranking the centrality often improves the algorithm performance but increases the computational expense. Finally the proposed entropy-based filter for edges allows to shift focus towards more precise (but less complete) taxonomies which arguably facilitates manual post-processing.

Co-occurrence based similarity measures of terms are easily extractable from literature databases and can provide a scaffold for taxonomy creation. However, they also limit the success of taxonomy creation when dealing with semantically related terms that can not be ordered by generality: High-level terms such as "wind power" or "solar energy", or terms that somehow interact (e.g., "hammer" and "nail") frequently co-occur and hence exhibit a misleadingly high co-occurrence similarity. Yet neither are subsumable in the strict sense ("is-a" or "part-of" relations) of standard taxonomies. As a result, the semantics of taxonomy sub- and superconcepts merely allows the interpretation as "is-related-to" relation. Such a property is not transitive and hence less useful for purposes, where complete semantic subtrees of the taxonomy are required. As a remedy, it would be beneficial to incorporate more sophisticated similarity and generality measures using Natural language processing techniques as proposed in [11]. To this end it seems most promising to device a meta-algorithm, for which the Heymann algorithm is a suitable platform.

References

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

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

[3] C. Chen. 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, 2006.

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

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

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

[7] Z. Harris. *Mathematical Structures of Language*. Wiley, 1968.

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

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

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

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

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

[13] R. Snow, D. Jurafsky, and A. Y. Ng. 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, 2006. Association for Computational Linguistics.

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

[15] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

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

[17] D. Widdows. 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, 2003. Association for Computational Linguistics.

[18] W. Woon and S. Madnick. Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, 2009.

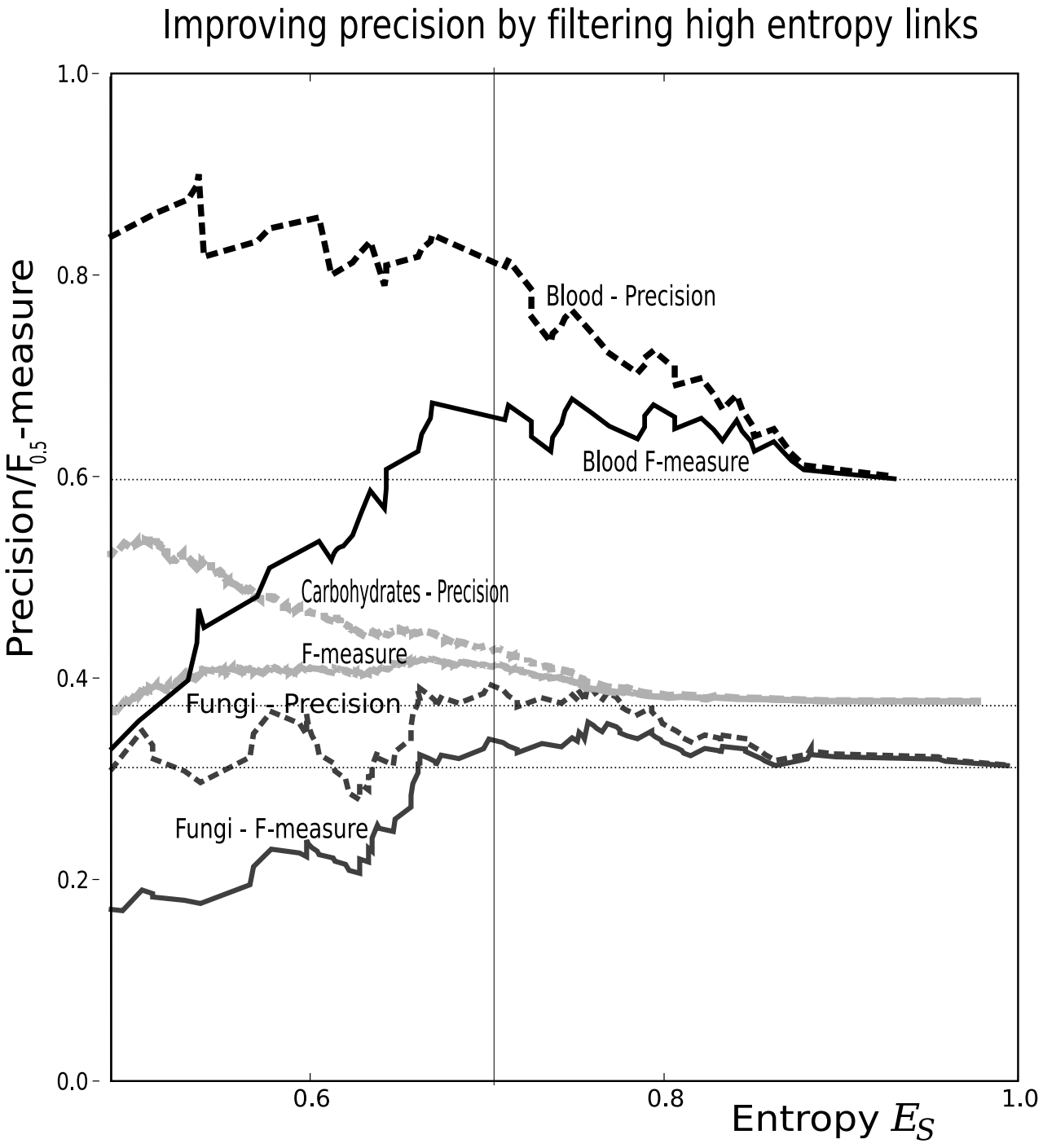


Figure 4. The figure shows the $F_{0.5}$-measure (solid) and the precision (dashed lines) for three MeSH benchmarks. Higher entropy of similarities expresses lower confidence in a taxonomy-link. Not filtering by entropy at all yield in precision and F-measure equal to the rightmost data point of each curve, indicated by horizontal dotted lines. The figure shows that indeed high entropy links are often wrong and precision decreases for all benchmark sets. Therefore, by filtering these low confidence links, the algorithm improves in terms of precision, while maintaining or slightly improving the F-measure. Any threshold above 0.7 increases precision without worsening the F-measure.

Bibliometric Analysis of Distributed Generation

Wei Lee Woon
Hatem Zeineldin
Stuart E. Madnick

Working Paper CISL# 2009-03

April 2009

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

Bibliometric Analysis of Distributed Generation

Wei Lee Woon*, Hatem Zeineldin*, Stuart Madnick†
*Masdar Institute of Science and Technology,
MASDAR, PO Box 54224,
Abu Dhabi, U.A.E.
†Sloan School of Management, M.I.T.,
E53-321, Cambridge MA, 02139, U.S.A.
wwoon@mist.ac.ae, hzainaldin@mist.ac.ae, smadnick@mit.edu

***Abstract*— This paper describes the application of data mining techniques for eludicating patterns and trends in technological innovation. Specifically, we focus on the use of bibliometric methods, *viz* techniques which focus on trends in the publication of text documents rather than the content of these documents. Of particular interest is the relationship between publication patterns, as characterized by term occurrence frequencies, and the underlying technological trends and developments which drive these trends. To focus the discussions and to provide a concrete example of their applicability, a detailed case study focussing on research in the area of Distributed Generation (DG) is also presented; however, the techniques and general approach devised here will be applicable to a broad range of industries, situations, and locations. Our results are promising and indicate that interesting information and conclusions can be derived from this line of analysis. The results obtained using data extraction techniques highlight and present the evolution of DG-related technology focus areas, and their relative importance within this field.**

I. Introduction

The planning and management of research and development activities is a challenging task that is further compounded by the large amounts of information which researchers and decision-makers are required to sift through. Information regarding past and current research is available from a wide variety of channels (important examples include publication and patent databases), providing both a difficult challenge and a rich source of possibilities. On the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data with which a well-informed and comprehensive research strategy may be formed.

Using bibliometrics to study the progression of research and technological development is not a new idea and there is already a significant body of research addressing this problem (for a good review, the reader is referred to [1], [2], [3], [4]. Interesting examples include visualizing the inter-relationships between research topics [1], [5], identification of important researchers or research groups [6], [3], the study of research performance by country [7], [8], the study of collaboration patterns [9], [10], [11] and the prediction of future trends and developments [12], [13], [14], [5]. Nevertheless, given the many difficulties inherent to these undertakings, there is still scope for further development, as well as for testing and fine-tuning these methodologies via appropriately scoped pilot studies. The research described in this paper was motivated by, and seeks to address this need.

To focus the discussions and to provide a suitable domain on which to test the capabilities of the approach, a pilot study was conducted on the domain of distributed generation (DG). In general, DG are generation sources that are connected to the distribution system close to the load point. Distributed generation could be classified into two main types; renewable such as wind and solar or non-renewable such as diesel generators and micro-turbines. The increasing penetration of DG in distribution systems coupled with the growing interest among the electrical power engineering community to meet environmental and energy efficiency constraints resulted in an increasing interest in DG. Thus, distribution systems worldwide are experiencing significant changes and challenges due to the increasing penetration of DG including renewable and non-Renewable Energy (RE) sources.

In general, distribution systems are radial in nature, meaning that power flows in one direction from the main substation to the load point. The addition of DG on the distribution system transforms the distribution system into a multi-source system. Topics such as distribution system protection, control and stability, which were not considered when designing distribution systems, have now become of major importance [15], [16], [17]. The planning and reliability of the distribution system will be affected with the addition of DG sources [18], [19], [20], [21]. In addition, new concepts and operational issues have emerged from this new distribution system structure which includes smart grids, micro-grids and islanding detection.

As such, while DG research provides a well-defined scope in which to conduct our analysis, it still spans a wide range of technical areas and promises to be a rich and challenging problem domain on which our methods can be suitably tested. In addition, we note that bibliometric analyses have been conducted on a variety of energy related issues (for e.g.: energy research in general [22], power sources [23], biofuels [24]), but not in the DG domain. As will be demonstrated, the application of bibliometric techniques can be extremely helpful in highlighting the main directions among the power engineering community in the area of DG.

The rest of the paper is structured as follows. The current section presented the background and motivations for the research, while Section II details the data collection and research

methodology used. In Section III the main results along with preliminary observations are described. Section IV analyzes these results in the context of specific developments and trends within the DG domain. Section V concludes the paper by summarizing the main findings and presenting suggestions for future areas of research.

II. Methods and data

The main aim of this study is to investigate the use of bibliometric techniques for studying technological innovation relevant to DG. These are techniques which focus on patterns and trends of textual information, rather than on the actual content of the text to be analyzed. In particular, we would like to test the usefulness of term usage statistics as a measure of the level of research activity or interest in that field.

To conduct the pilot study, we first collect a set of records consisting of journal and conference publications relevant to the field of distributed generation. To do this, the following keywords were submitted to the Scopus database through its web interface (as a title/abstract search):

- **"distributed generation"**
- **"dispersed generation"**
- **"distributed resources"**
- **"embedded generation"**
- **"decentralized generation"**
- **"decentralized energy"**
- **"distributed energy"**
- **"on-site generation"**

Searching for these keywords returned a total of 4734 records, which were saved to a database and used to represent the body of research in DG.

As our objective is to study the evolution of individual sub-fields within the broad context of DG, the next task was to identify subsets of this body of research, linked to representative keywords or search terms. This mechanism is then used to study the general growth and direction of the DG domain. We discuss this in terms of the following three-stage process:

1) Identification of comparable topics or technologies
2) Extraction of relevant/related studies
3) Normalization and preprocessing.

A. Distributed Generation: topics and themes

In order to obtain results that are interesting and useful, it is important that our analysis of the publication data is appropriately framed. For example, comparing the relative growths of two sub-areas of DG would only make sense if the two areas were somehow competitive with each other, or were at least similar in some sense. At the same time, the field of DG research is quite broad and involves a variety of interesting problems and challenges. To allow for this diversity, we define three separate dimensions in which to conduct our analysis:

1) **Energy generation** - i.e. part of the attraction of using a DG system is that it allows energy from a variety of localized sources to be integrated into the grid. As such, one way of decomposing the field of DG research is in terms of the technologies for generating this electricity. Towards this end, we choose to study the following four topics: (1) *Wind* (2) *Microturbines* (3) *Solar PVs* (4) *Fuel Cells*
2) **DG interfaces and technologies** - another important component of a DG system is the interface to the utility grid or to the customer. Publication patterns for the following three classes of DG interfaces will be analyzed: (1) *Inverter based* (2) *Synchronous based* (3) *Induction based*
3) **General studies** - besides the above two categories, research in DG also spans a range of other challenges and issues. The third dimension in our analysis groups a number of these topics: (1) *Control* (2) *Reliability* (3) *Islanding detection* (4) *Planning* (5) *Stability* (6) *Power quality* (7) *Electricity markets/economics* (8) *Protection* (9) *Forecasting* (10) *Microgrids*.

To provide additional depth and breadth to our study, we supplement our results with two further forms of analysis:

1) To elucidate more detailed trends in the DG domain, the database is further "sliced and diced" by searching for occurrences of terms in the second dimension (inverter, synchronous and induction based) within the context of the topics listed in the third dimension.
2) High level statistics regarding the distribution of publications amongst countries and journal titles are also presented to give a broader perspective of the current state of DG research.

B. Data extraction

Once the initial database is constructed, patterns and trends of individual research topics can be easily extracted and graphed using a combination of appropriate SQL queries and regular expressions.

In general, we are interested in the level of research activity or interest in a particular topic, as reflected in a representative keyword. Of course, the true research activity is unobservable, but we postulate that the *frequency* at which a particular term appears in the academic literature, henceforth referred to as the *Term Frequency* (TF), can serve as a proxy variable for this interest. Given that the academic literature is the primary channel through which research results are disseminated, it is reasonable that the TF approximates the rate at which research on a particular topic is generated.

To illustrate this process, we briefly consider the three topics: {"solar PV","microturbine" and "wind"}. To monitor development in these three sub-fields, the abstracts for all of the records in our database are compiled and grouped according to year of publication. Regular expressions corresponding to the three terms above are created and used to search the annualized abstract collections. To obtain a measure of TF for a term $\mathcal{T}_i$, the number of occurrences of this term for a given year are counted then normalized by the total number of words in all of the abstracts in the year as follows:

$$TF_i(t) = \frac{n_i}{|\mathcal{A}_t|}, \tag{1}$$

where n_i is the number of occurrences of term $\mathcal{T}_i$ in all abstracts published in the year t, and $\mathcal{A}_{\sqcup}$ is the string formed by concatenating all abstracts in that same year. These terms are graphed and are shown in figure 1(a).

C. Normalization and post-processing

One problem with using the raw frequencies in this way is that it tends to favour terms which are very general in nature, such as "wind" in this case, over terms which are very specific like "microturbine". In Fig. 1(a), it can be seen that the line for "wind" is a lot higher than the other two lines. However, this could simply be because wind energy is indeed a common form of renewable energy (which in the US and EU exceeds solar power in terms of its share of electricity generation [25]). The larger term frequencies for "wind" are hence a straightforward reflection of this fact. In addition "wind" is also a very general term which occurs frequently in common usage while "microturbine", for example, is a very specific term used exclusively to refer to a particular device.

In many cases we are less concerned with the absolute value of TF than with the overall *trend* observed in these figures - a low absolute value for TF may be secondary to the fact that these values are doubling every year, for example. To better observe these trends, we define TF^*, a normalized form of TF:

$$TF_i^*(t) = \frac{TF_i(t)}{\sum_{j=1992}^{2008} TF_i(j)}. \tag{2}$$

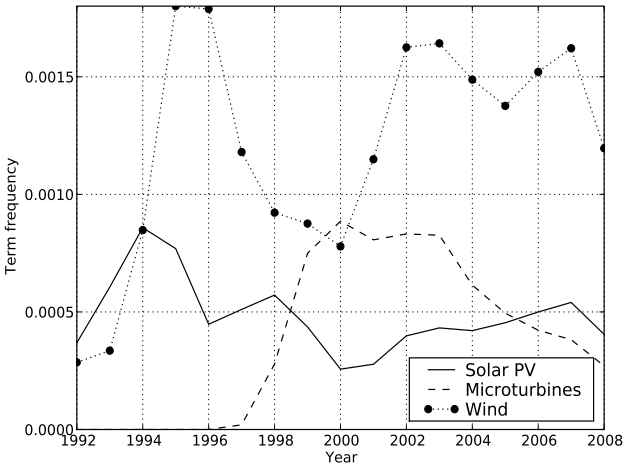
Doing this for the three topics listed above results in the graph in Fig. 1(b). From this subfigure, it can be seen that, if we disregard the absolute values, it would appear that "solar PV" and "wind" have very similar trends over the last 16 years, while research in microturbines experienced a very sharp increase in research activity between 1996 and 2000, remained relatively stable for a few years, then started dropping again from 2003 onwards (while this same information was present in the first graph it is evident how appropriate normalization allows different aspects of the data to be more easily noticed).

Finally, an alternative form of normalization was considered:

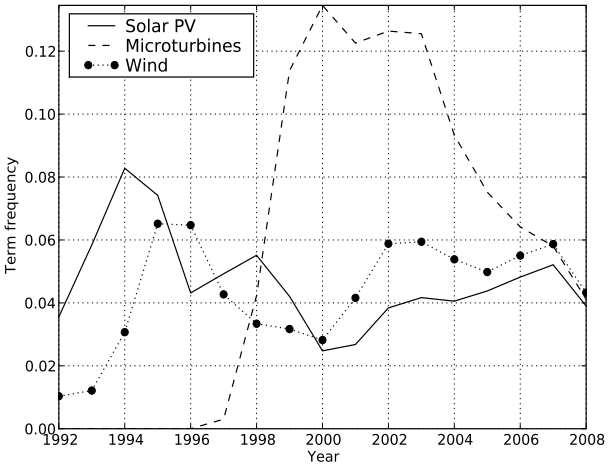
$$\widehat{TF}_i(t) = \frac{TF_i^*(t)}{\sum_{j=i}^{n} TF_i^*(t)}, \tag{3}$$

where n is the number of topics being studied concurrently. This would hopefully allow the growth in the different research areas to be more easily compared. When applied to the three topics from above, the graph in Fig. 1(c) is obtained. Note that, while this graph is quite similar to Fig. 1(b), there is an important difference early on in the graph where, in Fig. 1(c), the plot for solar PV related research is seen to be starting from a high value and slowly decreasing while for wind the curve shows an early rise in $\widehat{TF}$. This highlights the fact that, while both topics can be seen gaining in popularity early on, the $\widehat{TF}$ for "wind" gradually gains on solar PV and exceeds it by around 1995.

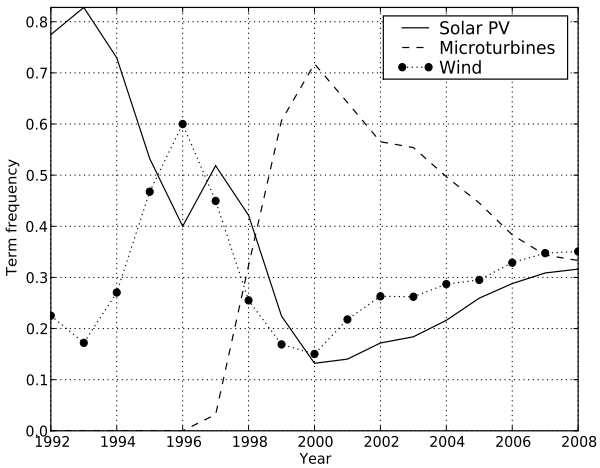
In brief, what we seek to demonstrate is not that any one of these graphs are necessarily "correct" or "wrong", but that different forms of normalization highlight different aspects of



(a) Unnormalized



(b) Topic-wise normalization



(c) Topic-wise & cross-sectional normalization

Fig. 1. Term frequencies for "solar PV", "microturbine" and "wind"

the data. As such, there is value in considering all the different forms of normalization when conducting our analysis.

D. Implementation

The required computational tools were implemented using the Python programming language and the SQLite database engine, as these facilitated faster development. The former also includes a broad selection of libraries, including those useful for the analysis of text and for data collection from the WWW. Both products are also cross-platform and open-sourced, which allowed applications to be deployed on a range of different operating systems and environments, and at a very low cost.

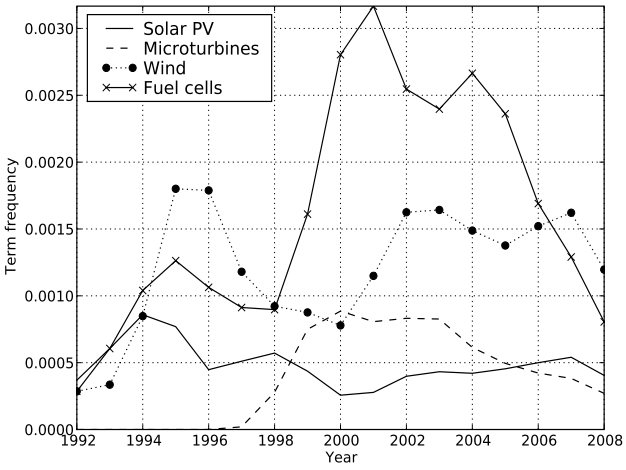
III. Results

The data collected from the Scopus database is used to generate yearly TF values for the research areas listed in section II-A. In addition, a three-tap gaussian filter was used to smooth the resulting time series as this was quite noisy and in some cases the number of publications retrieved were quite low. This is a reasonable pre-processing step because the research which results in a publication would have been carried out over a period of time prior to the appearance of the publication; as these publication counts are in fact a proxy for the underlying research activities, smoothing the raw data in this way also serves as a means of taking this spread into account.

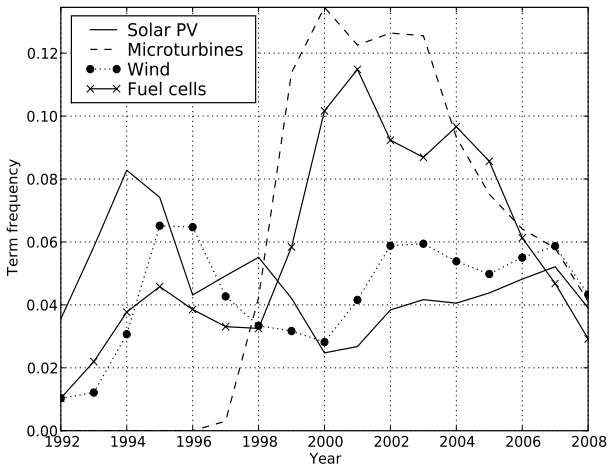
Four different sets of graph are presented here, corresponding to the three "dimensions" described in section II-A (dimension three comprised a large number of topics and was split into two sets of topics); these are presented in Figs. 2 to 5. In the following subsections, the initial observations are noted and discussed. In section IV a more detailed analysis will be presented, which will take into account the underlying drivers within the field of DG.

A. Energy generation (Fig. 2)

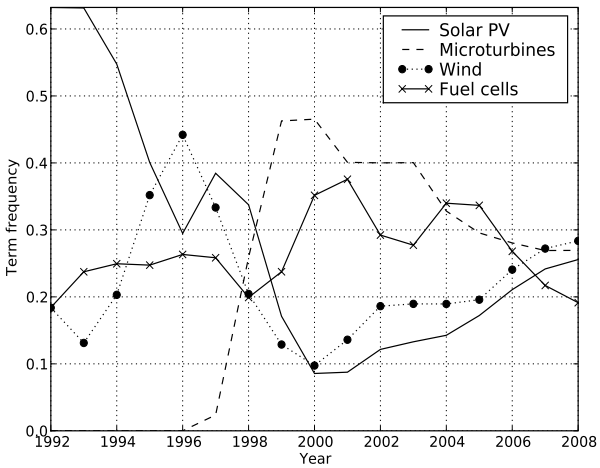
1) As with the example described in section II (in Fig. 1), the results varied significantly when subjected to the different normalization schemes. In general, the normalized data tended to show underlying trends with more clarity. However, using cross-sectional normalization appeared to produce noisier curves.
2) One prominent trend which was consistent across the three versions was with micro-turbine research, which exhibited a marked increase in TF between the years 1996 and 1999, before levelling off and finally declining again. Research in fuel cells also showed a similar trend though the TF in this case started from a higher point compared to the plot for microturbine.
3) In the case of solar PV, a "V" shaped curve centered around the year 2000 was observed. While this is most prominent in Fig. 2(c), it can be discerned in all three graphs and is consistent with general trends in the field of solar PV research [26].



(a) Unnormalized

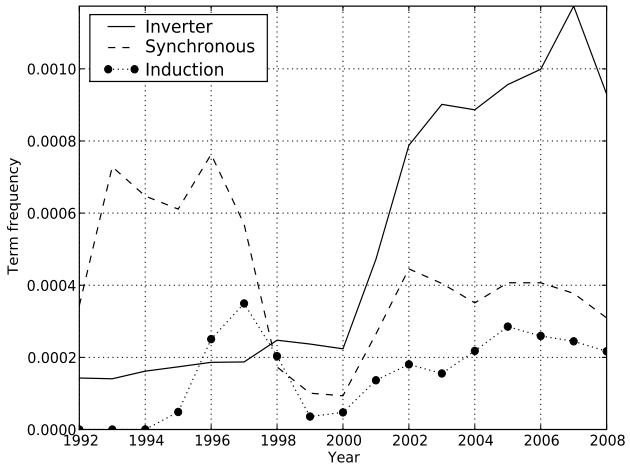


(b) Topic-wise normalization

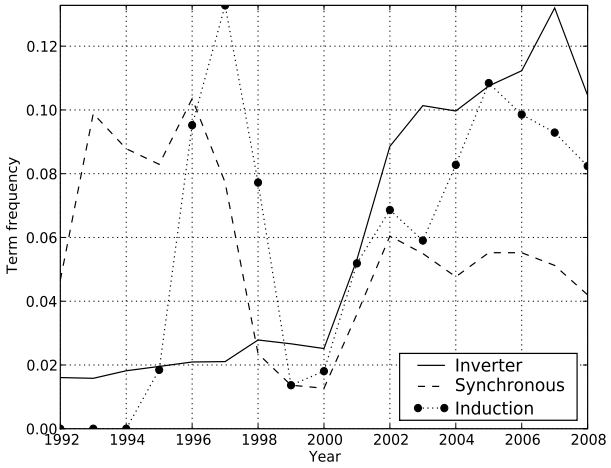


(c) Topic-wise & cross-sectional normalization

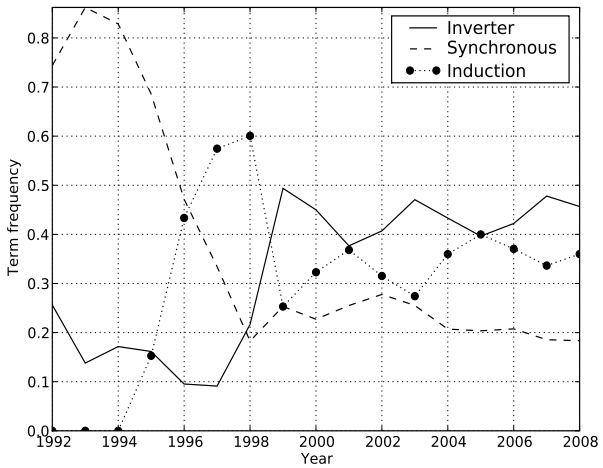
Fig. 2. Energy generation



(a) Unnormalized



(b) Topic-wise normalization



(c) Topic-wise & cross-sectional normalization

Fig. 3. DG interfaces and technologies

B. DG interface technologies (Fig. 3)

Two broad trends which were observed was with the TF plots for Inverter and Synchronous interfaces - specifically, over the period of analysis, research in synchronous interfaces seems to have declined somewhat, with the transition point falling somewhere between 1996 and 1998. In contrast, research in inverter-based interfaces showed the exact opposite trend - the TF for "inverter" started out low but increased sharply at or around the year 1998.

For induction-based interfaces, the trend is less clear but broadly, research in this topic is seen increasing gradually throughout the analysis period, starting from around 1994.

C. General studies (Figs. 4 and 5)

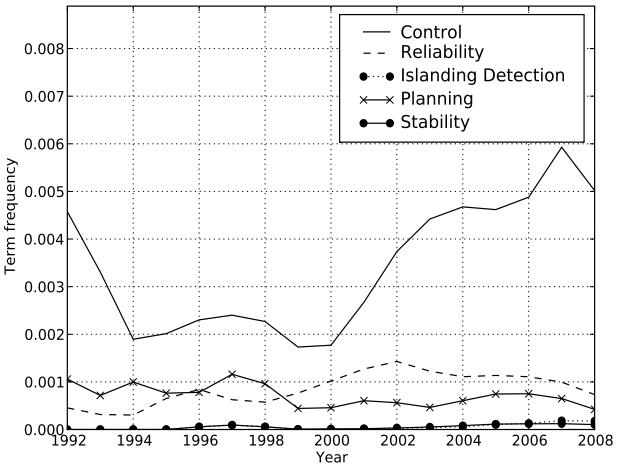
This dimension included a large number of topics, which were further divided into two separate figures (Figs. 4 and 5) to facilitate interpretation and analysis of the results. The division into the two batches was done randomly, where the only aim was to reduce the clutter within individual graphs.

Batch 1 - $\{$**Control, Reliability, Islanding, Planning, Stability**$\}$ (Fig. 4)

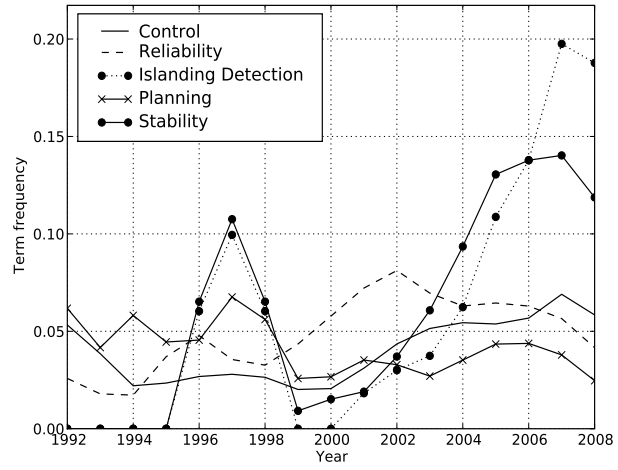
1) There are a number of prominent trends in this first batch of topics. One interesting example is with research in "islanding"; in Fig. 4(a), this plot can hardly be seen as its overall term frequency is quite low. However, when either of the two normalization schemes are applied, it is immediately clear that this is in fact one of the fastest growing topics of research from amongst this batch of topics.
2) A similar pattern is observed with "stability", with the TF curve showing a sharp increase at or around the year 1997, but decreasing somewhat for the next few years before again starting to increase sharply.
3) Research interest in "planning" starts out with a relatively high TF, but gradually decreases over the analysis period.
4) Another interesting example is in the case of "control" - where the values of TF and TF^* for this topic (Figs. 4(a) and 4(b)) are seen increasing quite steadily; however, when normalized with respect to the other topics ($\widehat{TF}$ - Fig. 4(c)), it appears that this area is in fact declining gradually.

Batch 2 - $\{$**Power Quality, Electricity Markets/Economics, Protection, Forecasting, Micro-grids**$\}$ (Fig. 5)

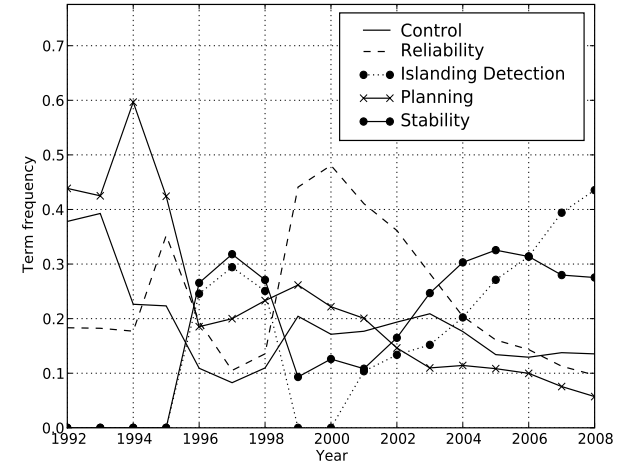
1) The most prominent trend in this second batch of topics was for "micro-grids", which in all three figures can be seen to be increasingly rapidly post-2000.
2) Also, both "Power Quality" and "Protection" gradually increase in popularity throughout the analysis period. The TF for "forecasting" can also be observed to be increasing quite quickly in the initial period, however, after around the year 2000, it starts to drop for around 3 years before staging a moderate "rebound" over the remaining years (though never quite regaining its initial high).



(a) Unnormalized

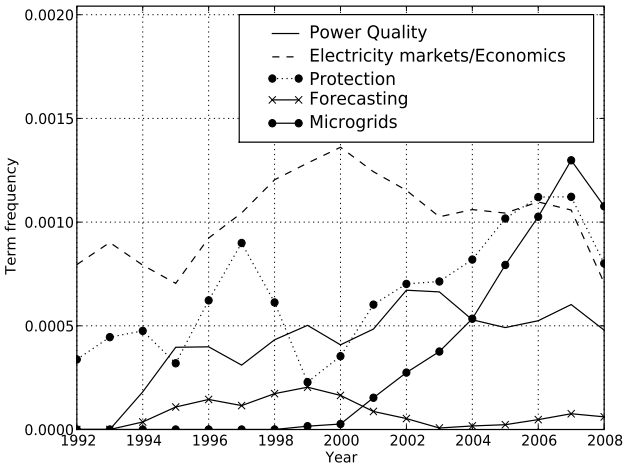


(b) Topic-wise normalization

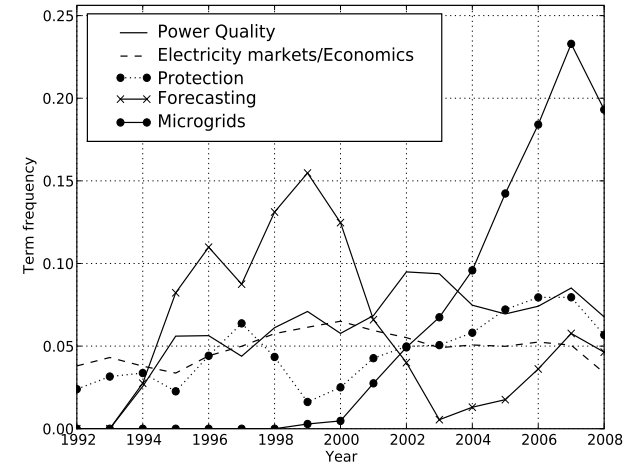


(c) Topic-wise & cross-sectional normalization

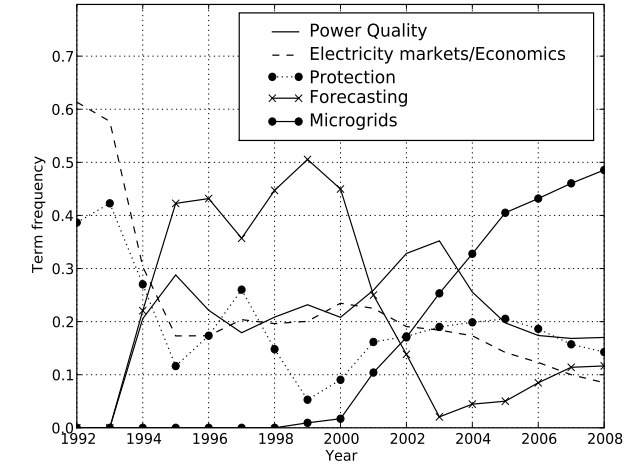
Fig. 4. General studies (batch 1)



(a) Unnormalized



(b) Topic-wise normalization



(c) Topic-wise & cross-sectional normalization

Fig. 5. General studies (batch 2)

3) "Electricity markets/economics" is another topic with inconsistent results - in Figs. 5(a) and 5(b), a mixed trend is observed though broadly speaking the popularity of this topic is seen to have increased on average. However, in contrast, the $\widehat{TF}$ plot (Fig. 5(c)) appears to show that activity in this topic is on the decline.

D. Cross-dimensional analysis

As mentioned in Section II, we would also like to analyse the data using a "slice and dice" approach involving the topics in dimensions 2 and 3. For each of the "general studies" topics, we are interested in the relative growth potential of each of the interface technologies:{Inverter,Synchronous,Induction}.

However, as occurences of each of these terms would be very low given the extremely specific conditions, the TF terms for each of the topic combinations would be averaged over the ranges [1992,2000] and [2000,2008] (inclusive). As an indicator of the growth potential for each combination, we simply use the differences between the mean TF for each of these two periods. These results are presented in Table I. The most prominent observation from this table was that for 7 out of 10 the topics, the inverter based interface technology had the highest rate of growth. Conversely, induction based interface technologies had the lowest growth rate for those 7 topics. However, it had the *highest* growth rate for the remaining three topics (stability, planning and forecasting), though in the case of planning, it had the same growth rate as for synchronous technologies.

E. Publication statistics by country and journals

In addition to the previous set of searches, we also generated publication counts according to the country of origin and journal. Statistics for country, journals and recent (last three years) journals are presented in tables II, III and IV respectively.

The distribution of publications by country of affiliation is presented in table II; this table is divided into two halves - the first provides statistics which take all the papers into account, while the second half only takes "recent" (defined as the last three years) publications into account. The main observations were:

1) As expected the USA had the largest share of the publications. However, its lead over the other countries has been reduced in recent years; while the total number of papers published in the USA is around 2.5 times the number published by it's closest competitor, the ratio for the last three years has reduced to slightly under 1.5 times.
2) Overall, the UK has been ranked second in this list. However, while it appeard to have gained slightly on the USA in the last few years, the largest shift has been in papers published by Chinese researchers - not only has China moved up to second position, the ratio of Chinese to American papers has increased tremendously from around a third to almost three quarters.
3) Besides China, other developing countries also appear to have increased their research output, notably India (from 10th to 7th place), Iran (15th to 10th place).
4) From amongst the developed countries, Canada has made the greatest gains, moving from 6th to 4th place. However, Germany and France suffered significant declines from 7th and 8th to 9th and 11th positions respectively while Australia dropped off the top-fifteen list entirely.
5) In terms of publication density (the ratio of number of papers to population), the UK had the highest overall number, followed by the Netherlands and Canada. However, for the last three years, Canada had the largest publication density, while the UK and Netherlands had dropped to 3nd and 4th position respectively. Belgium, a "newcomer" in the top-fifteen list, now occupied the 2nd position.

Statistics on the distribution of publications by journal titles have also been divided into general and recent tranches; however due to formatting considerations these have been divided into two tables, III and IV. Our observations are:

1) The top journal for DG-related research appears to be the IEEE Transactions on Power Systems; overall it had the largest number of relevant papers as well as the largest number of incoming citations. However, when only papers published after 2006 were considered, it was ranked third in terms of number of papers published, but still had the largest number of incoming citations.
2) Besides the above, the list had a large number of other IEEE Transactions - for both lists five of the top fifteen journals were IEEE Transactions.
3) One surprising observation that was partially addressed was that the journals IET Generation, Transmission and Distribution, IET Renewable Generation and IET Power Electronics, which are extremely prominent in the DG field, were initially missing from the overall list of prolific journals. However, upon closer inspection, it turned out that this was because both IET Renewable Generation and IET Power Electronics are new journals which have only appeared in the last two or three years, and as such would not fare well in a ranking based on total number of papers (in fact, it was this observation which originally motivated us to compile a set of tables which only considered papers published after 2006). In the table using recent papers, we see that IET Renewable Generation has appeared in the top-15 list, however, IET Power Electronics is still off the list - subsequent checks to Scopus' source list confirmed that this journal had yet to be added to their database. For IET Generation, Transmission and Distribution, it was noted that the journals title had previously been "IEE Proceedings: Generation, Transmission and Distribution", which had caused papers from the journal to be allocated to different bins, thus diluting their impact. When the two titles are combined, we find that this journal has also moved into the list of top-15 titles. These are good examples of the limitations of the proposed method - i.e. that results

[1]Population statistics collected in 2009 from http://en.wikipedia.org/wiki/List_of_countries_by_population

Rank	Interface / Topic	$TF_{[1992,2000]}$ - $TF_{[2000,2008]}$			
		Inverter	Synchronous	Induction	Mean
1	Micro/Smartgrids	$\mathbf{2.9 \times 10^{-3}}$	2.8×10^{-4}	1.1×10^{-5}	1.1×10^{-3}
2	Power Quality	$\mathbf{2.2 \times 10^{-3}}$	3.6×10^{-4}	2.8×10^{-4}	9.4×10^{-4}
3	Islanding Detection	$\mathbf{1.8 \times 10^{-3}}$	4.4×10^{-4}	0.0	7.4×10^{-4}
4	Stability	7.1×10^{-5}	8.3×10^{-4}	$\mathbf{1.1 \times 10^{-3}}$	6.8×10^{-4}
5	Control	$\mathbf{1.5 \times 10^{-3}}$	3.4×10^{-4}	1.4×10^{-4}	6.8×10^{-4}
6	Reliability	$\mathbf{6.7 \times 10^{-4}}$	2.6×10^{-4}	7.9×10^{-5}	3.3×10^{-4}
7	Planning	1.4×10^{-4}	$\mathbf{1.7 \times 10^{-4}}$	$\mathbf{1.7 \times 10^{-4}}$	1.6×10^{-4}
8	Forecasting	0.0	5.4×10^{-5}	$\mathbf{4.2 \times 10^{-4}}$	1.6×10^{-4}
9	Electricity markets/Economics	$\mathbf{1.5 \times 10^{-4}}$	1.2×10^{-4}	0.0	9.2×10^{-5}
10	Protection	$\mathbf{3.0 \times 10^{-4}}$	2.2×10^{-5}	-2.4×10^{-4}	2.5×10^{-5}

TABLE I

GROWTH RATES FOR INTERFACE TECHNOLOGIES W.R.T. EACH OF GENERAL STUDIES TOPICS. THE HIGHEST GROWTH RATE PER ROW IS PRINTED IN BOLD, AND THE ROWS ARE SORTED IN ORDER OF DESCENDING AVERAGE GROWTH RATE.

Overall statistics				Papers published after 2006			
Country	No. papers	Population (millions)	No. papers /population	Country	No. papers	Population (millions)	No. papers /population
USA	1572	306	5.1	USA	400	306	1.3
United Kingdom	639	62	10.3	China	308	1336	0.2
China	574	1336	0.4	United Kingdom	215	62	3.5
Japan	323	128	2.5	Canada	142	34	4.2
Italy	315	60	5.2	Italy	119	60	2.0
Canada	298	34	8.8	Japan	118	128	0.9
Germany	264	82	3.2	India	99	1161	0.1
France	203	65	3.1	Spain	96	46	2.1
Spain	199	46	4.3	Germany	90	82	1.1
India	185	1161	0.2	Iran	75	70	1.1
Netherlands	150	16	9.4	France	64	65	1.0
Brazil	130	191	0.7	Netherlands	52	16	3.2
Australia	128	22	5.8	South Korea	51	48	1.1
South Korea	125	48	2.6	Brazil	49	191	0.3
Iran	122	70	1.7	Belgium	45	11	4.1

TABLE II

PROLIFIC COUNTRIES [1]

are only as good as the data extracted from the source database, and is a motivating factor for combining query results from more than one database.

4) Another interesting development is the emergence of a number of Asian journals - Dianli Xitong Zidonghua and Zhongguo Dianji Gongcheng Xuebao from China,

Rank (Weighted)	Journal	No. papers	No. citations	No. citations/ No. papers
1 (1)	IEEE Transactions on Power Systems	93	1132	12
2 (3)	IEEE Transactions on Power Delivery	68	478	7
3 (5)	Electric Power Systems Research	47	247	5
4 (9)	Dianli Xitong Zidonghua/Automation of Electric Power Systems	43	168	3
5 (4)	Energy Policy	37	275	7
6 (6)	IEEE Transactions on Energy Conversion	37	235	6
7 (2)	IEEE Transactions on Power Electronics	32	610	19
8 (7)	Journal of Power Sources	28	234	8
9 (15)	BWK - Energie-Fachmagazin	28	3	0
10 (12)	Renewable Energy	23	72	3
11 (8)	IEEE Transactions on Industry Applications	23	176	7
12 (10)	IET Generation, Transmission and Distribution	22	159	7
13 (11)	Energy	17	88	5
14 (14)	IEEJ Transactions on Power and Energy	17	11	0
15 (13)	International Journal of Electrical Power and Energy Systems	16	37	2

TABLE III

TOP-FIFTEEN JOURNALS (BY NO. PAPERS). RANKS IN BRACKETS ARE BASED ON THE NUMBER OF INCOMING CITATIONS

Rank (Weighted)	Journal	No. papers	No. citations	No. citations/ No. papers
1 (2)	Electric Power Systems Research	37	46	1
2 (9)	IEEE Transactions on Power Delivery	35	26	0
3 (1)	IEEE Transactions on Power Systems	34	55	1
4 (4)	Dianli Xitong Zidonghua/Automation of Electric Power Systems	28	32	1
5 (5)	IEEE Transactions on Energy Conversion	20	32	1
6 (6)	Energy Policy	16	29	1
7 (8)	IEEE Transactions on Power Electronics	14	27	1
8 (15)	Transactions of the Korean Institute of Electrical Engineers	12	0	0
9 (13)	Zhongguo Dianji Gongcheng Xuebao/Proc. Chinese Soc. of Electr. Eng.	10	2	0
10 (11)	IET Renewable Power Generation	10	9	0
11 (10)	Journal of Power Sources	9	14	1
12 (3)	IEEE Transactions on Industrial Electronics	9	37	4
13 (7)	Energy	9	28	3
14 (12)	IET Generation, Transmission and Distribution	9	3	0
15 (14)	Electric Power Components and Systems	9	1	0

TABLE IV

TOP-FIFTEEN JOURNALS W.R.T. PAPERS PUBLISHED AFTER 2006. RANKS IN BRACKETS ARE BASED ON THE NUMBER OF INCOMING CITATIONS

and the Transactions of the Korean Institute of Electrical Engineers. In the overall data, only Dianli Xitong Zidonghua is in the top-15 list but in the list which reflect recent publication trends we find that the other two have appeared in positions 8 and 9 respectively.
However, upon more careful inspection, we find that the majority of the incoming citations for the first two journals originate from within the same journal, while for the Transactions of the Korean Institute of Electrical Engineers, there are no incoming citations at all. It would hence appear that, while the volume of research published in these journals is certainly increasing, the impact of this research is still relatively low, or is restricted to the respective local contexts.

IV. Analysis

In the previous section, the results of the bibliometric analysis were presented, and broad trends and patterns described. In this section a more detailed discussion will be presented, and attempts will be made to link prominent observations to relevant developments in the DG domain.

As highlighted earlier, DG could be interfaced to the distribution system through an inverter, synchronous or induction machine. From Table I it can be seen that most of the interface topics are of relative importance for inverter based and induction based DG. This could be related to the fact that an inverter interface is the best candidate for DG sources that inherently generate DC such as PV and Fuel cells. On the other hand, while existing wind turbine systems can use either synchronous or induction interfaces, the induction machine interface has several advantages which include lower capital cost, lower maintenance cost as well as better transient performance. For this reason, the induction machine interface is widely and extensively used for wind turbines [27]. For brevity, we will focus our discussion of some of the topics presented in Table I.

The main role of an inverter is to convert the DC power generated by the DG source to AC power necessary for feeding loads. Inverters are composed of a group of switches that are controlled in such a way to synthesize the DC waveform into an AC waveform. In this conversion process, harmonics are generated which results in power quality problems [28]. Synchronous and induction machines generally do not suffer from this problem and for this reason power quality is considered one of the main challenges when designing inverter based DG (refer to Table I). The second challenge that is of relative importance is islanding detection. Islanding is a condition where the DG is operating in an isolated mode and not directly connected to the utility. The IEEE Standards necessitate the disconnection of a DG once it is islanded to avoid any power quality and safety problems[2]. This topic is a major challenge when implementing inverter and synchrony based DG because both types can sustain stable operation in the absence of the grid connection. On the other hand, induction machines require a grid connection for stable operation and for this reason an islanding condition could be easily detected and no sophisticated islanding detection method needs to be implemented [29]. This coincide with the results presented in Table I where most of the research work for islanding detection falls under the inverter and synchronous interface area.

Forecasting, stability and planning are two research topics where induction based DG become more dominant. Induction based DG are commonly used for wind generation. Large scale wind installation commonly referred to as wind farms are usually connected on the transmission and sub-transmission levels and their capacity is in the range of a common generation plant. Accurate forecasting techniques and stability analysis becomes an essential requirement to guarantee efficient, reliable and stable power system operation. Taking into account wind generation in planning studies is a result of the increasing reliance on and penetration of wind generation. Two research topics that have been drawing much attention recently are micro-grids and smart grids. DG can form a new type of power system, the so-called micro-grid. Micro-grids can be viewed as a group of DG sources operating either connected or isolated to/from the main grid [30] and [31]. For utilities with high penetration of DG sources, the practice of disconnecting all DG during a grid failure is neither practical nor reliable. There is a pressing need for modifications to these policies with regard to DG disconnection after a disturbance. The current IEEE standards do not address the topic of micro-grids. However, a new IEEE standard is currently being developed to address micro-grids [32]. Inverter based DG provides an attractive option due to its fast transient response and flexibility in interface control design. Smart grids are an extension of the micro-grid concept. Smart grid is a distribution grid with a communication infrastructure with capability to control the different components within the smart grid to achieve efficient, reliable and secure operation.

V. Conclusions

Below we summarize the main findings, discuss limitations of the current analysis, and offer suggestions for future research in the area of bibliometric analysis of DG research.

A. Results

The results presented in this paper demonstrate that the proposed methodology, which is based on a bibliometric approach, is capable of extracting valuable information from semi-structured sources of data. While this study is still preliminary, this information is already useful in helping to improve our understanding about trends and patterns in research, and would already be of great interest to a researcher in the field of DG. As has been discussed in sections III and IV, the trends highlighted by the bibliometric analysis certainly appear to be in agreement with developments in the field of DG, and could support the formulation of well informed and effective research policies. Some recommendations are:

1) Smart grids and micro-grids are considered amongst the hottest research area in the field of DG. This is also evident from the increasing number of conferences that have put those topics as a research track.

[2]IEEE Std. 1547-2003

2) Among all DG interface types, inverter based DG seems to be the dominant type with many interesting and challenging problems.
3) Topics such as protection and stability, commonly not addresed in distribution system design, are now becoming of major importance due to the high penetration of DG.

B. Methods

As with any computational framework which exploits semi-structured data, there were some issues which need to be highlighted. Firstly, it is important to note that success in using this approach to analyse research progress is contingent upon our ability to correctly identify publications which are relevant to the topics of interest. So, for example, the frequency of the term “wind” is used to estimate the level of activity in wind power research. While this successfully identifies many relevant papers, there will certainly be occurrences of false positives or negatives; i.e. there might be publications with abstracts containing the word “wind”, but which are not directly relevant to wind power. Conversely, there might be publications which are in fact relevant to wind power, but which have abstracts which do not explicitly mention the word “wind”.

A further problem is with inconsistent database capabilities. To access a larger body of documents, an obvious measure would be to submit queries to a number of different academic search engines (for example, the ”Scirus” search engine, or Google’s Scholar search engine), or to conduct full text searches of these publications. Unfortunately, many search engines do not allow full-text searches, or exporting of search results for use in bibliometric analysis.

To help counter these problems and to increase the quality and applicability of this approach, we propose the following avenues for future work:

1) **Intelligent feature extraction** - A variety of techniques from the machine learning and semantic technology communities could be brought to bear. In particular, it would be interesting to see the value of incorporating semantically-enabled features into the search process - i.e. instead of using manually generated keyword searches, computational techniques could be used to group together terms which are either synonymous, or which are observed to co-occur frequently, and to combine these terms appropriately when conducting the searches.
2) **Information fusion** from multiple, heterogeneous data sources - as noted above, different databases provide different capabilities, and cover different subsets of the academic literature. Instead of accessing individual databases in isolation, information extracted from different sources could be combined, and in a way which allows for the heterogeneity. So for example, a weighting mechanism could be devised to allow results of full-text searches from one database to be combined with a title-only search from another.
3) **Tools development** - thus far the analysis has been carried out using a collection of python scripts. While these have been very useful for our purposes, we plan to make these methods useable for a broader audience by creating a set of user-friendly software applications. These tools will incorporate the functionality of the scripts but in an intuitive and accessible way.

REFERENCES

[1] Alan Porter. Tech mining. *Compet Int Mag*, 8(1):30–36, 2005.
[2] Alan Porter. How ”tech mining” can enhance r&d management. *Res Tech Manage*, 50(2):15–20, 2007.
[3] Paul Losiewicz, Douglas Oard, and Ronald Kostoff. Textual data mining to support science and technology management. *J Int Inf Syst*, 15(2):99–119, 2000.
[4] Joseph Martino. *Technological Forecasting for Decision Making*. McGraw-Hill Engineering and Technology Management Series, 1993.
[5] Henry Small. Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610, December 2006.
[6] R. N. Kostoff. Text mining using database tomography and bibliometrics: A review. *Technol Forecast Soc*, 68:223–253, November 2001.
[7] de Miranda, Gilda M. Coelho, Dos, and Lelio F. Filho. Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technol Forecast Soc*, 73(8):1013–1027, October 2006.
[8] Kim and Mee-Jean. A bibliometric analysis of the effectiveness of koreas biotechnology stimulation plans, with a comparison with four other asian nations. *Scientometrics*, 72(3):371–388, September 2007.
[9] Anuradha, K., Urs, and Shalini. Bibliometric indicators of indian research collaboration patterns: A correspondence analysis. *Scientometrics*, 71(2):179–189, May 2007.
[10] Wen-Ta Chiu and Yuh-Shan Ho. Bibliometric analysis of tsunami research. *Scientometrics*, 73(1):3–17, October 2007.
[11] Tibor Braun, Andrs P. Schubert, and Ronald N. Kostoff. Growth and trends of fullerene research as reflected in its journal literature. *Chem Rev*, 100(1):23–38, 2000.
[12] N. R. Smalheiser. Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Technovation*, 21(10):689–693, October 2001.
[13] T. U. Daim, G. R. Rueda, and H. T. Martin. Technology forecasting using bibliometric analysis and system dynamics. In *Technology Management: A Unifying Discipline for Melting the Boundaries*, pages 112–122, 2005.
[14] Tugrul U. Daim, Guillermo Rueda, Hilary Martin, and Pisek Gerdsri. Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technol Forecast Soc*, 73(8):981–1012, October 2006.
[15] T. Tran-Quoc, L. Le Thanh, C. Andrieu, N. Hadjsaid, C. Kieny, JC Sabonnadiere, K. Le, O. Devaux, and O. Chilard. Stability Analysis for the Distribution Networks with Distributed Generation. In *Transmission and Distribution Conference and Exhibition, 2005/2006 IEEE PES*, pages 289–294, 2005.
[16] N. Kroutikova, CA Hernandez-Aramburo, and TC Green. State-space model of grid-connected inverters under current control mode. *IET Electric Power Applications*, 1:329–338, 2007.
[17] A. Girgis and S. Brahma. Effect of distributed generation on protective device coordinationin distribution system. In *Power Engineering, 2001. LESCOPE’01. 2001 Large Engineering Systems Conference on*, pages 115–119, 2001.
[18] AA Chowdhury, SK Agarwal, DO Koval, M.A.E. Co, and IA Davenport. Reliability modeling of distributed generation in conventional distribution systems planning and analysis. *IEEE Transactions on Industry Applications*, 39(5):1493–1498, 2003.
[19] RC Dugan, TE McDermott, GJ Ball, E.C. Inc, and TN Knoxville. Distribution planning for distributed generation. In *Rural Electric Power Conference, 2000*, page C4, 2000.
[20] G. Celli and F. Pilo. MV network planning under uncertainties on distributed generationpenetration. In *IEEE Power Engineering Society Summer Meeting, 2001*, volume 1, 2001.
[21] P. Jahangiri and M. Fotuhi-Firuzabad. Reliability assessment of distribution system with distributed generation. In *IEEE 2nd International Power and Energy Conference, 2008. PECon 2008*, pages 1551–1556, 2008.
[22] Y. Kajikawa, Y. Takeda, and K. Matsushima. Computer-assisted roadmapping: A case study in energy research. In *Management of Engineering & Technology, 2008. PICMET 2008. Portland International Conference on*, pages 2159–2164, 2008.

[23] RN Kostoff, R. Tshiteya, KM Pfeil, JA Humenik, and G. Karypis. Power source roadmaps using bibliometrics and database tomography. *Energy*, 30(5):709–730, 2005.

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

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

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

[27] M.R. Patel. *Wind and solar power systems: design, analysis, and operation*. CRC Press, 2006.

[28] J. Wong, P. Baroutis, R. Chadha, R. Iravani, M. Graovac, and X. Wang. 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, 2008.

[29] W. Xu, K. Mauch, and S. Martel. 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*, 2004.

[30] HH Zeineldin, EF El-Saadany, and MMA Salama. 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, 2006.

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

[32] B. Kroposki, T. Basso, and R. DeBlasio. 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, 2008.

Measuring Innovation Using Bibliometric Techniques
The Case of Solar Photovoltaic Industry

Georgeta Vidican[1], *Wei Lee Woon*[1], *Stuart Madnick*[2]

[1]Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[2]Massachusetts Institute of Technology, Cambridge, MA

Paper submitted to the *Advancing the Study of Innovation and Globalization in Organizations* (ASIGO) Conference in Nurnberg, Germany, May 29-30, 2009

March 15, 2009

Abstract

In this paper, we use feature extraction and data analysis techniques for the elucidation of patterns and trends in technological innovation. In studying innovation, we focus on the role of public research institutions (research universities and national laboratories) in the development of new industries. More specifically, we are interested in measuring innovation through research collaborations between these institutions and the private sector.

The proposed methods are primarily drawn from the field of bibliometrics – i.e. the analysis of information and trends in the publication of text documents, rather than the contents of these documents. In particular, we seek to explore the relationship between joint publication patterns and trends, R&D funding, technology development choices, and the viability and effectiveness of industry-university collaborations.

To focus the discussions and to provide concrete examples of their applicability, this study will have an initial emphasis on the solar photovoltaic (PV) sector in the U.S., though the techniques and general approach devised here will be applicable to a broad range of industries, situations, and locations.

Our analysis suggests that interesting information and conclusions can be derived from this line of analysis. The results obtained using our data extraction techniques allow us to identify early technology focus in different areas within solar PV technologies, and to determine potential technology pathways, which is critical for innovation policy in the renewable energy domain.

1. Introduction

1.1 Problem statement

The increasing challenge of international competitiveness driven by knowledge production and innovation, calls for an assessment of the quality and use of indicators for science, technology, and innovation (OECD 2007, Smith 1998). While innovation is difficult to quantify, some aspects related to key dimensions of inputs and outputs can still be measured (Smith 2007). Measuring innovation is even more important for emerging industries, such as the renewable energy sector, receiving large amounts of governmental spending both for research and development (R&D) as well as for market expansion.

This paper explores the role of public research institutions (universities and national laboratories) in the development of new industries, focusing on measuring innovation outcomes, in the form of knowledge creation, using novel bibliometric techniques. A plethora of studies have stressed that close academic-industry collaborations are critical for the formation of industry clusters (Saxenian 1996, Scott 2003). Moreover, the technical revolutions we are experiencing in fields such as renewable energy, involve complex interactions between government, industry, and the academic system. To capture some of these dynamics, our research focuses primarily on measuring joint publications between researchers in the academia, national labs, and the private sector. In order to narrow the discussions and analyses, our research has an emphasis on the solar photovoltaic (PV) sector in the U.S. Nevertheless, the specific methods are applicable to a broad range of industries and contexts.

1.2 Novelty and motivation

Using bibliometrics to study the progression of research and technological development is not a new idea and there is already a significant body of research addressing this problem (for a good review, the reader is referred to Porter (2005, 2007), Losiewicz et al. (2000), Martino (1993)). Interesting examples include visualizing the inter-relationships between research topics (Porter 2005, Small 2006), identification of important researchers or research groups (Kostoff 2001, Losiewicz et al. 2000), the study of research performance by country (de Miranda et al. 2006), (Kim and Mee-Jean 2007), the study of collaboration patterns (Anuradha et al. 2007,Chiu and Ho 2007, Braun et al. 2000) and the prediction of future trends and developments (Smalheiser 2001, Daim et al. 2005, Daim et al. 2006, Small 2006).

As such, it would appear that the applicability of bibliometric techniques to the study of technology development is quite well established. Godin and Gingras (2000) have used bibliometrics to assess the role of universities in the system of knowledge production. Their study concludes that while federal R&D funding declined overtime, universities became more important through increased collaborations with the private sector. Other studies, such as Zimmerman et al. (2009), use bibliometrics to examine national research collaborations. However, there is hardly any study on the progression of science in the renewable energy field. Tsay (2008) traces the evolution of hydrogen energy literature worldwide. Aside from a few exceptions covering national and international collaboration patterns in the fuel cell technology in Norway (Godo et al. 2003), and co-authorship networks in the area of nanostructured solar cells using bibliometric and social network analysis (Larsen 2008), there is much more insight to be gained from examining different types of collaborations in the field of renewable energy technologies, in particular solar energy technologies.

However, despite the high level of activity in the general area of research, there does not appear

to have been a corresponding level of interest in their use for analyzing industry-university collaborative research. Godin and Gingras (2000) have used bibliometrics to assess the role of universities in the system for knowledge production. Their study concludes that while federal R&D funding declined overtime, universities became more important through increased collaborations with the private sector. Other studies, such as Zimmerman et al. (2009) use bibliometrics to examine national research collaborations. However, there is hardly any study on the progression of science in the renewable energy field. Tsay (2008) traces the evolution of hydrogen energy literature worldwide. Aside from a few exceptions covering national and international collaboration patterns in the fuel cell technology in Norway (Godo et al. 2003), and co-authorship networks in the area of nanostructured solar cells using bibliometric and social network analysis (Larsen 2008), there is much more insight to be gained from examining different types of collaborations in the field of renewable energy technologies, in particular solar energy technologies.

As such, there certainly appears to be a high level of activity in this general area of research. However, interestingly there does not appear to have been a similar level of interest in using these tools for analyzing collaborative research amongst players in industry, academia and in the national laboratories. Another important issue is to study how the characteristics and trends in these collaborations reflect the underlying factors of government funding, societal and environmental developments. The research described in this paper was motivated by, and seeks to address these important issues.

The rest of the paper is structured as follows. The current section presented the background and motivations for the research, while Section 2 describes in detail the data sources and the research methodology used to measure knowledge production from academia and industry on areas related to solar PV technologies in the U.S. In Section 3 we describe the main results along with preliminary observations. Section 4 then discusses these results in the larger context of innovation measurement and challenges in the solar photovoltaic sector. Section 5 concludes the paper with main findings and suggestions for future areas of research.

2. Data and methods

2.1 Data

We focus our research primarily on two states in the U.S., California and Massachusetts, for several reasons. First, the origins of the solar photovoltaic industry worldwide emerged from these two regions (the research labs of two large oil companies, Mobil Tyco in Massachusetts and ARCO Solar in California) (Margolis 2002). Second, over the years the solar photovoltaic industry in the U.S. has been concentrated in these two locales, with California hosting the largest share of companies. Third, the regional economies in Boston and San Francisco are the innovation engines for the U.S. economy, clustered around the research universities establishments (Saxenian 1996). Hence, we argue that publications emerging from universities and research laboratories in these two regions define the scope of research for other research establishments in the U.S.[1]

[1] The National Renewable Energy Laboratory (NREL), playing a key role in coordinating and funding research for the solar industry, is however, located outside these two states, in Golden, Colorado.

2.2 Methods

The basic premise for this study is to investigate the use of bibliometric techniques for studying technological innovation relevant to photovoltaics. These are techniques, which focus on patterns and trends of textual information, rather than on the actual content of the text to be analyzed. In particular, we would like to test the usefulness of *hit count*s, i.e.: the number of academic publications relevant to a particular field, as a measure of the level of research activity or interest in that field. These hit counts were collected in yearly bins, allowing the time evolution of research activity over the corresponding periods of time to be visualized and studied.

To conduct the pilot study, the keywords "photovolt*", "solar cell", "solar PV", "solar energy", "solar generation" and "solar power"" were submitted to ISI's Web of Science database. In addition, it was also necessary to include additional search terms so as to specify the types of institutions in which the research was being carried out. Two general approaches were used. In the first approach, generic searches were generated using terms which indicated authors from industry, national laboratories and from academia. The search terms used were:

> University: Address field to include: "univ" or "inst"
> National Laboratory: Address field to include "lab" or "laboratory" or "labs*"
> Industry: address field to include: "inc" or "corp" or "co"

To study the research activity in different states, an additional term was included as follows:

> AD=("inc" SAME "MA"),

which would admit publications where the address fields include "inc" and "MA" in the same line (i.e. this would identify publications originating from industrial researchers located in Massachusetts).

Using this approach gave us a lot more flexibility as we could now generate searches which target specific subsets of the academic literature, which in turn would reflect the level of research being conducted in the corresponding sectors.

In the second approach, two lists were manually compiled for Massachusetts and California: lists of companies known to be involved in solar photovoltaic research, as well as lists of universities and national research laboratories in the two states. To extract the hit counts from the Web of Science web interface, the results were broken up into batches of 10 companies at a time (this was necessary as the lists of companies were too long to be entered into a single search).

Unfortunately, our initial experimentation with the second method revealed that it was too restrictive and retrieved too few papers to be of use in the present study. In addition, this approach was very labor-intensive as a separate list of companies would have to be compiled for any future study. In addition, maintaining and keeping the lists up-to-date would also not be easy. As such, all the results presented here were extracted using the first approach.

The required computational tools were implemented in the Python programming language, as it facilitated faster development and includes a broad selection of libraries, including those useful for the analysis of text and for data collection from the WWW. Python is also a cross-platform environment and allows applications to be deployed on a variety of operating systems and environments.

3. Results

Annual publication counts from the Web of Science database were collected for all the years between 1975 and 2008, inclusive. In addition, a five tap gaussian filter was used to smooth the resulting time series as they were quite noisy and in some cases the number of publications retrieved were very low. This was a reasonable pre-processing step because the research which results in a publication would have been carried out over a period of time prior to the appearance of the publication; as these publication counts are in fact a proxy for the underlying research activities, smoothing the raw data in this way may be viewed as a means of taking this spread into account. Six different sets of graph are presented here, which reflect research activities carried out in universities, national laboratories, industry, and collaborative efforts involving pairings between each of these three sectors. These graphs are presented in Figures 1 to 6. Initial observations are:

1. While the details of individual graphs varied somewhat, the same high-level trend was observed in the majority of the cases: the number of publications started off high with peaks in the early to mid-80s'. However, as we move into the 90s', there was a marked decline in the number of papers which continued until around 1995, after which publication counts were observed to increase again.
2. The number of papers published in Massachusetts were found to be significantly lower than in California. This was particularly true of collaborative research, where in many cases only one or two papers were identified over a period of several years. As this is clearly insufficient, for collaborative research we will focus on results for California and for the entire U.S. only. Besides Massachusetts, this was also a problem in the case of university-industry collaborations in California, where the numbers of papers produced were close to zero for much of the study period.
3. In general, national laboratories in California and Massachusetts appear to have produced more papers in the initial high activity period (ranging approximately from 1975 to 1985), than in the second half of the study period (see Figure 2). For universities, the results are the opposite where the number of publications produced in the second period of high activity (approximately 1995-2008) exceed the publications produced in the first (see Figure 1). However, note that in both cases we still observe the same broad trend where there is a significant drop in the hit counts for the period of time ranging from around 1985 to 1995.
4. The hit counts corresponding to industrial research are a lot higher in the first period, and for the case of California and Massachusetts, there is no apparent recovery post-1995 (see Figure 3).
5. The results for collaborative research were more difficult to analyze as the number of publications found was significantly lower across all the sectors (this is to be expected as the search terms used were more restrictive in these cases). As such, the results were invariably noisier and were frequently unreliable (cases in point being all of the results for Massachusetts, and the results for university-industry collaborations in California). Having said that, the results for collaborative research activities could often be seen to be combinations of the sectors involved. So, for example, the hit counts for laboratory research in California were a lot higher earlier in the study period while for university research, the opposite is observed; accordingly, for laboratory-university research in California the two periods are quite well balanced (see Figure 4).

6. However, an interesting counter-example to the previous observation is found in the case of industry-laboratory research nationwide. As mentioned previously, hit counts for research conducted in industry and in national laboratories start out relatively high, but eventually end up lower, at least on average (see Figure 6). Surprisingly, the publication trend for industry-laboratory research is exactly the opposite - the number of publications produced post 1995 is actually significantly greater than the number of publications in the preceding period. This implies that, prior to 1995, a lot of the research being conducted in the two sectors were carried out in isolation, whereas a much greater proportion of research was being conducted in collaboration in the post-1995 period. A similar observation can be made about university-laboratory collaborations nationwide: prior to 1995, we see that less than a third of research in national laboratories was in collaboration with universities; however, in the post 1995 period, this figure has increased to around two thirds, as reflected in the number of academic publications (see Figure 5).

4. Discussion

In the previous section, data from the Web of Science database was presented, and pertinent numerical trends were discussed. Importantly, instead of a smooth growth curve, as might have been expected, our analysis revealed that innovation in photovoltaics exhibited a rather discontinuous growth pattern from 1975 to 2008. The decade between late 1975-1985, and after 2000 has registered the largest number of publications in the field of solar photovoltaics (PV) in the U.S. (see Figure 7), while the intervening time saw a marked decline in the number of relevant publications.

To better understand our observations, in this section we examine these trends in reference to the federal research and development (R&D) spending on solar PV, collaboration patterns, and institutions involved in solar PV energy research over the years.

4.1. R&D Spending on Solar PV research

The first energy crisis in the late 1970s called for increased attention to renewable energy technologies as alternatives to conventional fuel sources. Drawing on the experience with solar cells for space applications, there were reasons to believe that with sufficient research funding, solar PV can be used to generate cost-competitive energy for residential and commercial use (Margolis 2002). As a result, increased federal R&D funds have been channeled into research for solar technologies (see Figure 8).

The 1980s, however, saw significant reductions in the amount of federal R&D spending for renewable energy technologies, a trend which has been common in most highly industrialized countries (Margolis 2002). Consequently, overtime, the share of funding for solar energy technologies decreased consistently (see Figure 8).

Our results illustrate that there is a correlation between federal R&D funding and the number of scientific publications, until about 2000. More recently, concerns with climate change and energy security, and government support for commercialization of renewable energy technologies, revived the interest in solar energy technologies at public research institutions. Currently, however, a disproportionate amount of funding originates from the private sector, supporting research at universities.

4.2 Institutions involved in solar PV research

Public research institutions (universities and national research laboratories) have been critical for the advance of knowledge in solar PV energy, in terms of both technology development, as well as system integration. Nevertheless, our research suggests that private companies have also been highly involved in solar energy research primarily before 1990s. Below we discuss in greater detail the institutions involved in the solar energy technologies research and their technological focus in different geographical locations.

4.2.1 National Laboratories

Despite the decline in federal R&D funding, national laboratories were instrumental is supporting research interest in solar energy technologies. As Figure 2 shows, while the number of publications from national labs declined, the trend has not been as dramatic as for the level of federal investment. In Massachusetts and California national labs in the respective regions contributed less after 1990s. We argue that the lower regional presence of national labs in the solar PV research could be due to the shift of research competence on solar PV to National Renewable Energy Laboratory (NREL) located in Colorado, created in 1991.

In California, Jet Propulsion Lab (JPL) at California Institute of Technology (Caltech) recorded the largest share of scientific publications until early 1990s (approximately 50% of all publications on solar energy emerging from national labs). The research originating from JPL has been quite diverse in focus, ranging from space solar cells until early 1980s, to improving efficiency and testing reliability of terrestrial solar technology applications in the 1980s, and more recently on third generation solar technologies.

In addition, about 25% of the publications originated at the Lawrence Berkeley National Laboratory (LBNL) associated with the University of California Berkeley (UC Berkeley). While LBNL's involvement in solar research has been very limited early on, more recently, after year 2000, it has concentrated most of national labs research on solar. The spectrum of solar research at LBNL covers different solar technologies.

Other national labs that contributed to the advancement of knowledge in solar energy technologies over the years have been Lawrence Livermore National Laboratory (from 1973-1999), Sandia National Laboratory (1974-2004). Aside from national labs, other research institutions associated with the industry, played an important role in the research landscape of solar technologies in the 70s and 80s, such as Lockheed Missiles and Space Laboratory, Optical Coating Laboratory, US Air Force Laboratory. Interestingly, until 1980s, the research laboratory of a large Silicon Valley semiconductor multinational company, Varian Associates Inc., published 10% of the scientific papers on solar energy technologies in the region.

In Massachusetts, the lower number of national labs is reflected in fewer number of publications emerging from these research institutions. MIT Lincoln Laboratory has been involved in solar energy technologies research only until late 1980s. More than 60% of the research publications from research laboratories originate at MIT Lincoln Lab. Research at Philips Labs, associated with the US Air Force, has been focused entirely on space applications. The MIT Energy Lab was also important in the early stages of research.

4.2.2 Universities

While the share of research on solar PV technologies originating from universities has been lower that from national laboratories, this trend is likely to change with the new focus on advancing knowledge on renewable energy technology and systems at the global level.

In California, Stanford University and UC Berkeley have been the centers of research along with the national labs in the region, until the mid 1980s. UC Berkeley continues to play a key role in the advancement of science, as reflected by a high number of publications throughout the entire period. Nevertheless, in the past five years, research on organic PV technologies has brought Stanford University back to the research landscape.

In Massachusetts, while MIT has usually been the engine and source of innovation for industries such as semiconductors and biotechnology, our results show that this has not been the case for the solar PV industry. While early on MIT's involvement in the solar energy sector has been through the Lincoln Lab, since 2004 MIT has adopted as its mandate to focus on alternative energy technologies. The creation of MIT Energy Initiative and its engagement with multiple energy related private and public partners suggests that MIT's role in energy (and in particular solar) research is likely to expand significantly in the near term.

University of Massachusetts (UMass) Lowell has played a critical role throughout the entire period in advancing knowledge primarily in solar PV energy systems, but also more recently in cutting edge research focused on organic PV (about 40% of university based publications emerged from UMass Lowell). Other universities in the region have also been active in this research area, such as Harvard University, Boston University and Boston College, Northeastern University, UMass Amherst, UMass Boston, Northeastern University, Tufts University, and Clark University.

4.3 Collaboration patterns

Research collaborations have been identified as important for knowledge creation and knowledge transfer. Sharing knowledge and ideas is even more important for an emerging domain of knowledge like solar energy technologies, which builds on interdisciplinary expertise. In the U.S., the Department of Energy (DOE) has initiated several funding schemes to foster research collaborations between public research institutions and the private sector, such as the PV Manufacturing Technology Project (in 1991), the Thin-Film PV Partnership (in 1994), or the Industry Alliance Project (in 2007). Hence, gaining insights into the outcomes of these investment programs over the years, and in the patterns of collaboration, is a valuable exercise.

4.3.1 Collaboration patterns between universities and national laboratories

While a variety of universities are involved in research collaborations with national labs, the share of publications originating from California, and more recently from Colorado (due to NREL's location) is disproportionably higher. Nevertheless, programs such as those initiated by DOE appear to have been successful in stimulated collaborative research efforts since university-national labs collaborations increased significantly after 1992 (see Figure 4).

In California, the proximity of national research labs to the local universities oftentimes leads to blurry boundaries between the two institutions. Hence, intense collaborations are recorded between, for instance, UC Berkeley and LBNL in Northern California, or between Caltech and JPL in Southern California. This type of collaboration is not present in

Massachusetts in the field of solar technologies, although it is strong in other areas, such as biotechnologies.

4.3.2 Collaboration patterns with the industry

In general, we find only a few collaborations between universities or national laboratories with companies (see Figures 5 and 6). At national level, however, such collaborations increased after 1990s. An explanation for this increasing trend is that companies might have realized that solar PV has potential for becoming a market niche following increased government investment in supporting market deployment worldwide (primarily in Germany and Japan).

Collaborations between national labs and companies predominated in the 70s and 80s when large semiconductor and aerospace companies were interested in exploring potential new niche markets. In California, examples of such companies are Spectrolab, Varian Associates Inc, Standard Oil Co., Hughes Aircraft Co., Rockwell Int. Corp., Lockheed Aircraft Corp., and Applied Materials. The focus of research for these collaborations has been on more mainstream solar technologies such the 1st and 2nd generation of solar technologies. In California and Massachusetts very few industry-national labs collaborations were recorded after late 1990s. The increasing trend in such collaborations at national level (see Figure 6) is primarily due to NREL's significant role mainly after 1990 in solar related research.

Collaborations between the industry and academia have been even less frequent. In California IBM has been the main industry collaborator for universities, followed by Solarmer Energy Inc. a start-up in El Monte. Most of the research has been focused on cutting edge solar technologies such as polymer based PV (3rd generation, organic PV solar technologies). University of California (UC) Los Angeles, UC Santa Cruz, and UC San Diego are the universities most engaged in such collaborations. In Massachusetts the few industry collaborations were with universities or research laboratories from outside the state, reflecting a limited solar research agenda at the established institutions.

Given the emerging nature of the solar industry, we did not identify regional clusters of collaboration between companies and universities or national research laboratories. Rather, we find that companies seek research partners with the desired expertise who are not necessarily in their geographical proximity. In California, where the concentration of national labs and universities is higher, collaborations within the geographical cluster are more prevalent.

5. Conclusions

Below we summarize the main findings, discuss limitations of the current analysis, and offer suggestions for future research in the area of bibliometrics and innovation assessment.

5.1 Methods

The results that we have presented in this paper demonstrate that the proposed methodology, which is based on a bibliometric approach, is capable of extracting valuable information from semi-structured sources of data. While this study is still preliminary, it shows that this information is already useful in helping to improve our understanding of trends and patterns in innovation. In the present study, the emphasis has been on innovation in the field of photovoltaics, and more specifically, in the states of Massachusetts and

California in the U.S. However, the described framework is hugely flexible and can be easily generalized to the study of innovation in different fields, or in different geographical locations.

As with any computational framework which exploits semi-structured data, there were certainly some problems. Firstly, we note that success in tracking the progress of innovation in this way is contingent upon our ability to correctly identify publications which are relevant to our study. So, for example, to study photovoltaic research conducted by industry-linked players in California, an appropriate Web of Science search would have to be generated which matches publications resulting from this specific subset of research. For the current study, the following search term was used:

> ***TI=("photovolt*"OR "solar cell" OR "solar PV" OR "solar energy" OR "solar generation" OR "solar power") AND AD=((inc SAME CA) OR (co SAME CA) OR (corp SAME CA))***

In most cases, this successfully extracts the correct results, however we might anticipate a few potential problems:

1. **False positives/negatives:** There might be publications which contain the terms "solar power" or "solar energy" which are not actually relevant to photovoltaic research. Conversely, there might be publications which are relevant to photovoltaic research, but which do not include any of these terms in the titles. Similarly, there might companies or other industrial research entities which do not explicitly state the terms "co", "corp" or "inc" in their address fields.
2. **Inconsistent database coverage:** This is related to the previous problem; in many cases, a much better retrieval rate might have been achieved had we used title/abstract searches instead of simply using title searches - unfortunately, the Web of Science database was only able to conduct abstract searches for publications dating from 1991 and so for uniformity we had to rely exclusively on title searches.
3. **Inconsistent database capabilities:** To search a larger body of documents, an obvious measure would be to submit searches to a number of different academic search engines (for example, the "Scirus" search engine, or Google's Scholar search engine). Unfortunately, many search engines do not permit the searching of address lines explicitly. Even if it might be possible to include terms like "inc" and "MA" in the full text searches, we would still need to be able to specify that "inc" and "CA" be on the same address line.

To help counter these problems and also increase the overall quality and applicability of the approach, we propose the following avenues for future work:

1. **Intelligent feature extraction** - A variety of techniques from the machine learning and semantic technology communities could be brought to bear. In particular, it would be interesting to see the value of incorporating semantically-enabled features into the search process - i.e. instead of using manually generated keyword searches, computational techniques could be used to group together terms which are either synonymous, or which are observed to co-occur frequently, and to combine these terms appropriately when conducting the searches.
2. **Statistical analysis** of the search results - in this study, the data extraction process has largely been automated; however, the analysis of the results is still largely manual. While the final analysis of the results will likely always be manual, we hope to enrich

this process by providing more information to support users of this system. In particular, text mining techniques could be used to process the abstracts of retrieved documents - this can, for example, help users to identify transitions in the emphasis of research projects, and to visualize the evolution of this emphasis.

3. **Tools development** - thus far the analysis has been carried out using a collection of python scripts. While these have been very useful for our purposes, we plan to make these methods useable by a broader audience by creating a set of user-friendly software applications. These tools will incorporate the functionality of the scripts but in an intuitive and accessible way.

5.2 Innovation assessment

Our study suggests that using bibliometrics offers valuable insights for understanding the outcomes of government research expenditures, the institutional players involved in the emergence of an industry, the technological trajectories over the years, and in general the level of interest in a particular domain of knowledge. Especially for the case of renewable energy, such an analysis is important for laying out the foundation for further explorations.

The results from our analysis point to the close association between federal investment in R&D and knowledge production, as measured by number of publications. Especially in the early stages of industry development, 1970s and 80s, R&D funding programs proved to be critical for advancing science in solar photovoltaic technologies. Hence, policy-makers in the domain of science, technology and innovation, should ensure that especially for emergent industries, consistent investment in R&D is being made.

The high geographical concentration of national research labs in California, and the regional presence of space solar research at companies such as Spectrolab, created opportunities for a natural transition in the scientific community towards terrestrial solar applications research. National labs such as JPL and LBNL, and universities such as Caltech, Stanford, UC Berkeley, became the locus of research for U.S. as a whole. More recently, however, NREL in Colorado, concentrates the highest level of research on solar technologies at the national level.

Collaborative research between industry and the scientific community has been higher early on, having decreased more recently. Two reasons could explain this trend. First, until 1990s R&D funding from the federal government emphasized and required partnerships between different institutions (universities, national labs, and private sector) for caring out research activities. Second, collaborations between companies and academia might be easier and more valuable to engage with in the early stages of the industry. When the industry becomes more mature the level of competition increases, shifting the locus of research in companies' research laboratories.

However, the number of collaborative research papers between universities and industry does not reflect the true level of interaction between these institutions. Because of the different nature of these institutions, a large fraction of research outcomes do not end up in the public domain. Hence, these results need to be supplemented with additional information on specific university-industry contracts for research.

As the solar industry becomes more global and knowledge transfer surpasses national borders, it is interesting and relevant to explore the level of international collaborations thorough joint publications. In an extension of this research, we aim to explore who are the main countries and institutions collaborating with U.S. based scientists. Findings from such an

analysis would shed light on existing and growing research potential abroad.

Lastly, while results from our bibliometric analysis allow us to map the intensity of and interest in research over time in different institutions, we are not able to identify the impact that publications are having on the field of solar technologies as a whole. To get to this aspect, future research will need to take into account an analysis of papers citation patterns.

Bibliography

Anuradha, K., and Shalini (2007). Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis. Scientometrics, 71(2):179–189.

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

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

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

de Miranda, 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.

Godin, B. and Gingras, Y. (2000). The Place of Universities in the System of Knowledge Production. Research Policy, 29: 273-278.

Godo, H., Nedrum, L., Rapmund, A. and Nygaard, S. (2003). Innovation in Fuel Cells and Related Hydrogen Technology in Norway-OECD Case Study in the Energy Sector, NIFU report 35/2003.

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.

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

Larsen, K. (2008). Knowledge Network Hubs and Measures of Research Impact, Science Structure, and Publication Output in Nanostructures Solar Cell Research. Scientometrics, 74(1): 123-142.

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.

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

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

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

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

Scott, A. (2003). Flexible Production Systems and Regional Development: The Rise of New Industrial Spaces in North America and Western Europe. In Barnes, T.J. et al. (Eds.), Reading Economic Geography, Wiley-Blackwell.

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

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

Tsay, M. (2008). A Bibliometric Analysis of Hydrogen Energy Literature 1965-2005. Scientometrics, 75(3): 421-438.

Zimmerman, E., Wolfgang, G., and Bar-Ilan, J. (2009). Scholarly Collaboration Between Europe and Israel: A Scientometric Examination of a Changing Landscape. Scientometrics, 78(3): 427-446.

List of Figures

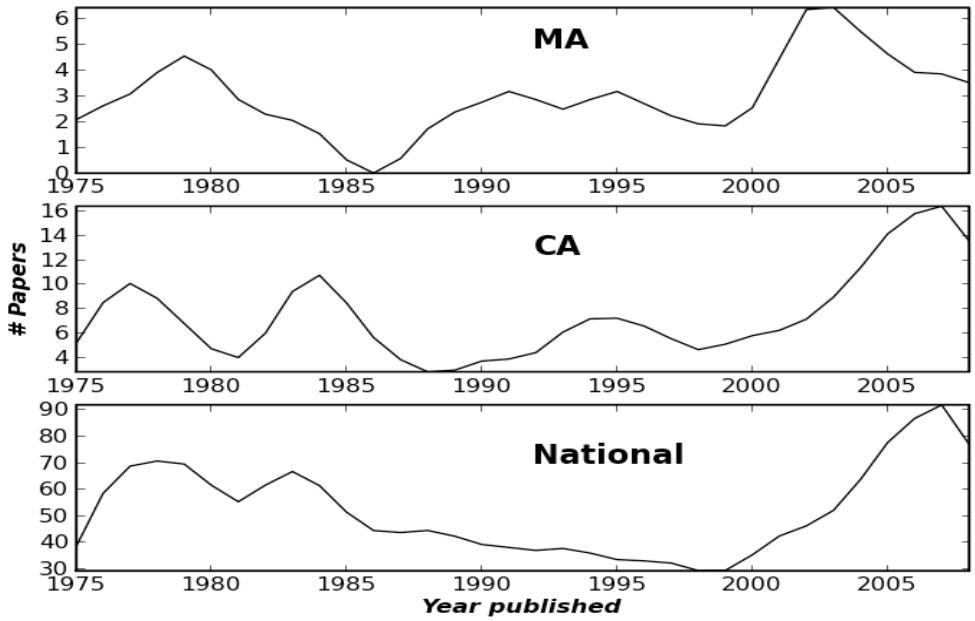


Figure 1: Research at universities

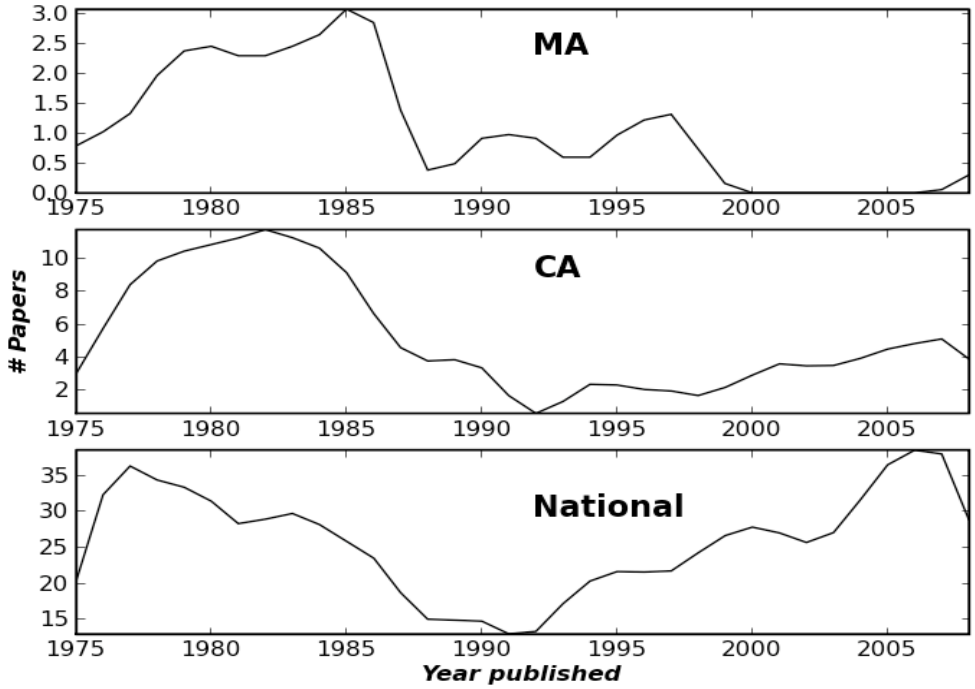


Figure 2:Research at national research laboratories

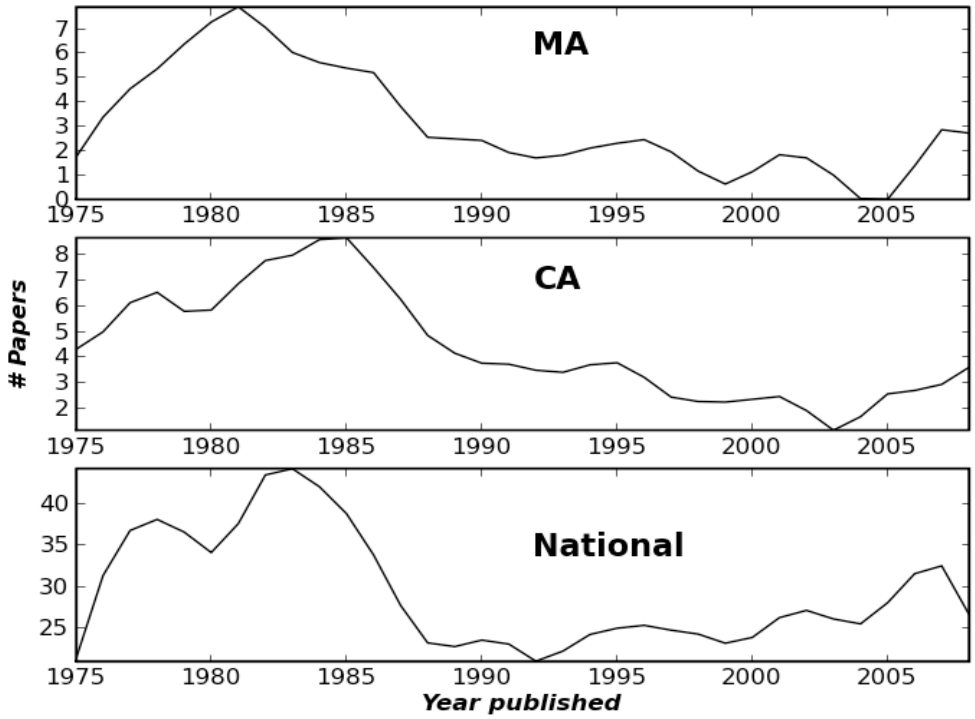


Figure 3: Research by Industry

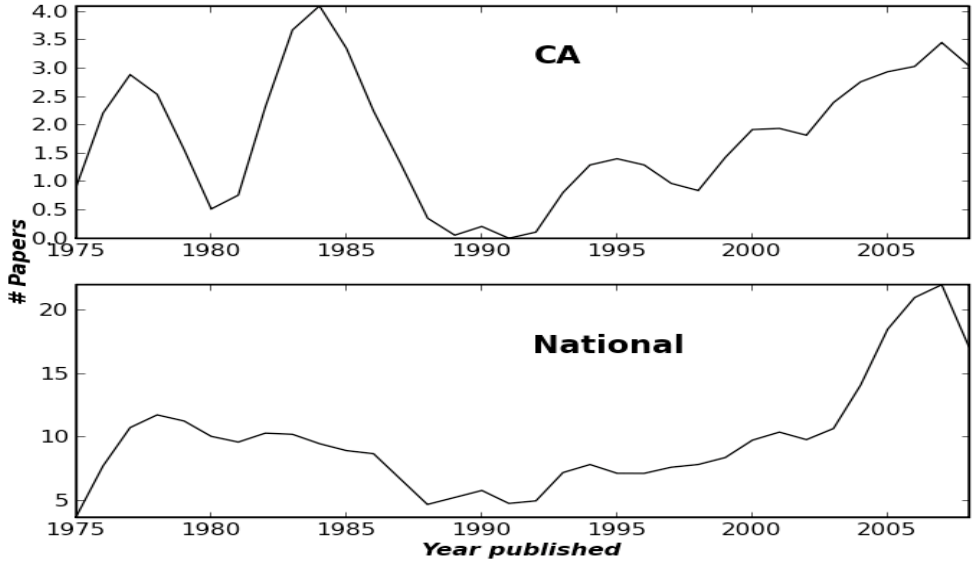


Figure 4: Laboratory-University collaborations

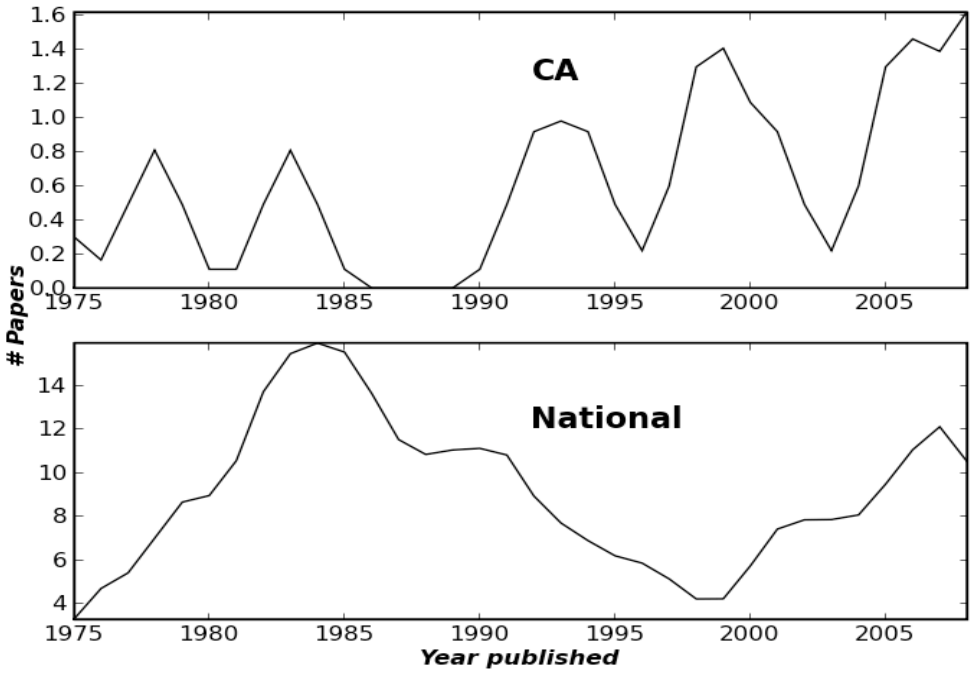


Figure 5: Industry-University collaborations

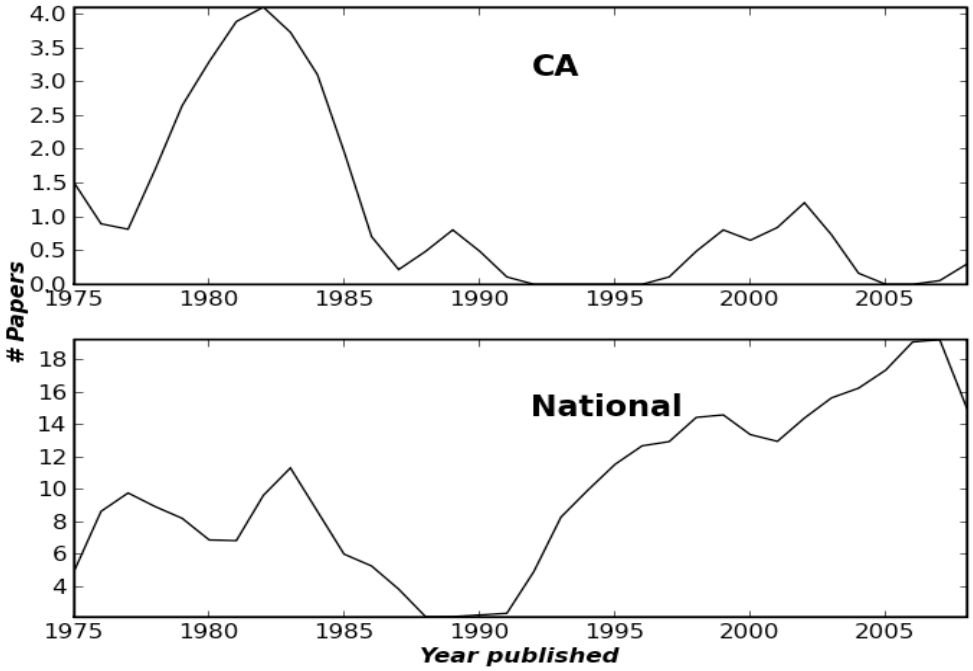


Figure 6: Industry-Laboratory collaborations

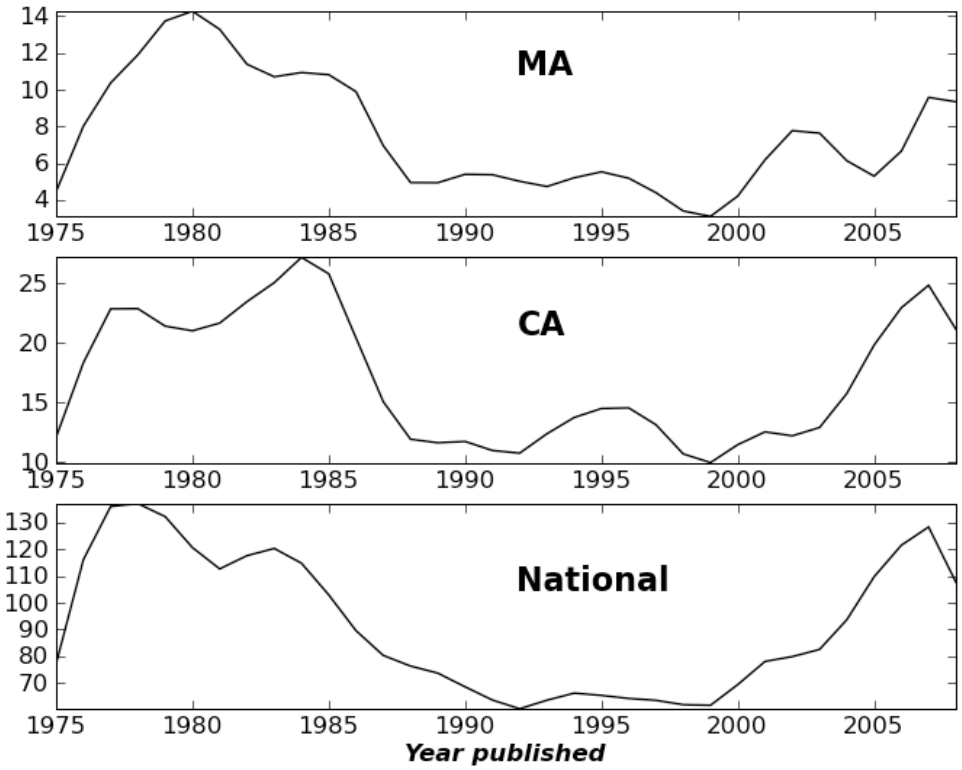
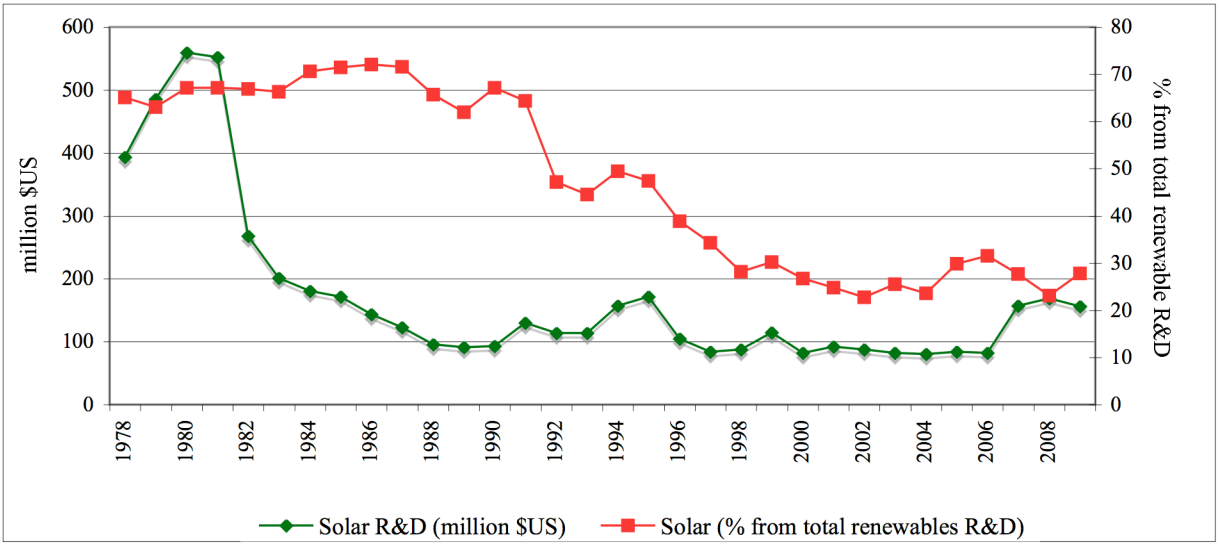


Figure 7: Solar related research in the U.S. (from any type of institution)



Note: Solar includes biofuels, wind, and ocean up to 1998.
Source: Gallagher, K.S., Sagar, A, Segal, D, de Sa, P, and John P. Holdren, "DOE Budget Authority for Energy Research, Development, and Demonstration Database," Energy Technology Innovation Project, John F. Kennedy School of Government, Harvard University, 2006. Database updated by Kelly Gallager, February 2008.

Figure 8: R&D Federal spending on solar related research between 1978 and 2009