Asymmetric Information Distances for Automated Taxonomy Construction

Wei Lee Woon
Stuart E. Madnick

Working Paper CISL# 2008-05

August 2008

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

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-occurence 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], 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]. We also note that taxonomy creation has been addressed before in [Blaschke and Valencia, 2002], [Makrehchi and Kamel, 2007], though different approaches are taken in both cases. Nevertheless, 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 importance of energy to the continued and general well-being of society cannot be understated, yet 87%[1] of the world's energy requirements are currently fulfilled via the unsustainable burning of fossil fuels. A combination of environmental, supply and security problems have made renewable energy technologies such as wind and solar power one of the most important topics of research today.

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

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.

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

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-occurence 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-occurence frequencies as an indication of the extent to which two terms are related to each other. This is defined as:

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

where NGD stands for the *Normalized Google Distance*, t_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. While a detailed discussion of the theoretical underpinnings of this method is beyond the present 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:

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

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

In the above, N is the size of the sample space for the "google distribution", and can be approximated by the total number of documents indexed by Google or the search engine being used, if this is not Google. Substituting (3),(4) → (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-occurences which are within the context of renewable energy; as such, to calculate the co-occurence frequency $n_{i,j}$ between terms t_1 and t_2, the search term "'renewable+energy"+"t_1"+"t_2"' was submitted to the search engine. 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 *arabidopsis* and *wind* are very broad and would admit many irrelvant studies.

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-occurence 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:

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

In this equation, the expression $\overrightarrow{\mathrm{NID}}(x, y)$ 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:

$$\begin{aligned} &\textit{information}(\text{elephant}) \subset \textit{information}(\text{circus elephant}), \\ &\therefore K(\text{circus elephant}, \text{elephant}) - K(\text{circus elephant}) \approx 0. \end{aligned}$$

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 find that: $n_{elephant} = 80,300,000$, $n_{circus\ elephant} = 106,000$ and $n_{circus\ elephant,elephant} = 91,800$ (these values are the hit counts returned by the respective Google searches, and are *estimated values* which may change in the future, though presumably not by much). As such:

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

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".

$\overrightarrow{\text{NGD}}$ 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 $\mathcal{V}$, $\mathcal{E} = \{(u, v) : u \in \mathcal{V}, v \in \mathcal{V}, u \neq v\}$, the set of edges of the graph, and the weighting function $w : \mathcal{E} \rightarrow \mathbb{R}$ is given by:

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

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.

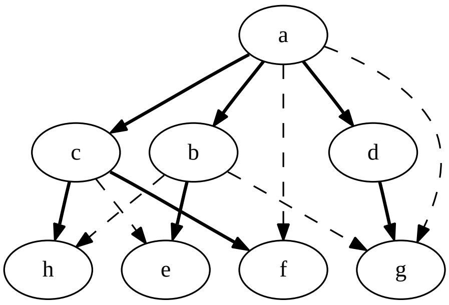


Fig. 1. Directed graph. The solid lines show one of a number of arborescences in the graph

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.

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.

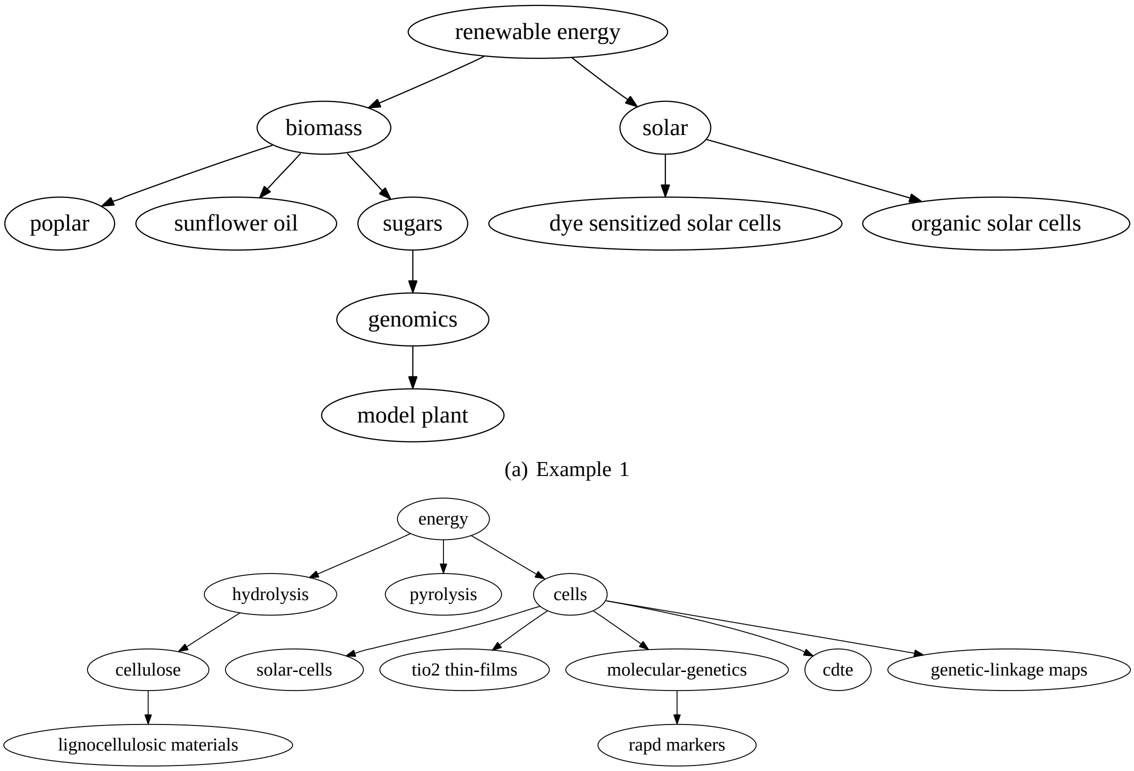
Using the cost function derived from eq. (6) often resulted in local structure which did not reflect the actual inheritance structure. In a noiseless environment this would not be a problem but in practice there are a number of situations where this reduces the accuracy of the results.

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: $\overrightarrow{\text{NGD}}(sugars, biomass) = 0.237$, while $\overrightarrow{\text{NGD}}(genomics, sugars) = 0.336$, both of which are the smallest values in the respective rows of the distance matrix. However, $\overrightarrow{\text{NGD}}(genomics, biomass) = 0.462$ which is somewhat greater than $\overrightarrow{\text{NGD}}(genomics, renewable\ energy) = 0.395$, 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 frequently encountered, in which very broad terms (such as *cells*) tend to dominate the subclassing process, resulting in extremely flat hierarchies. A further complication is that the keyword *cells* has two senses: solar "cells", and biological "cells".

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

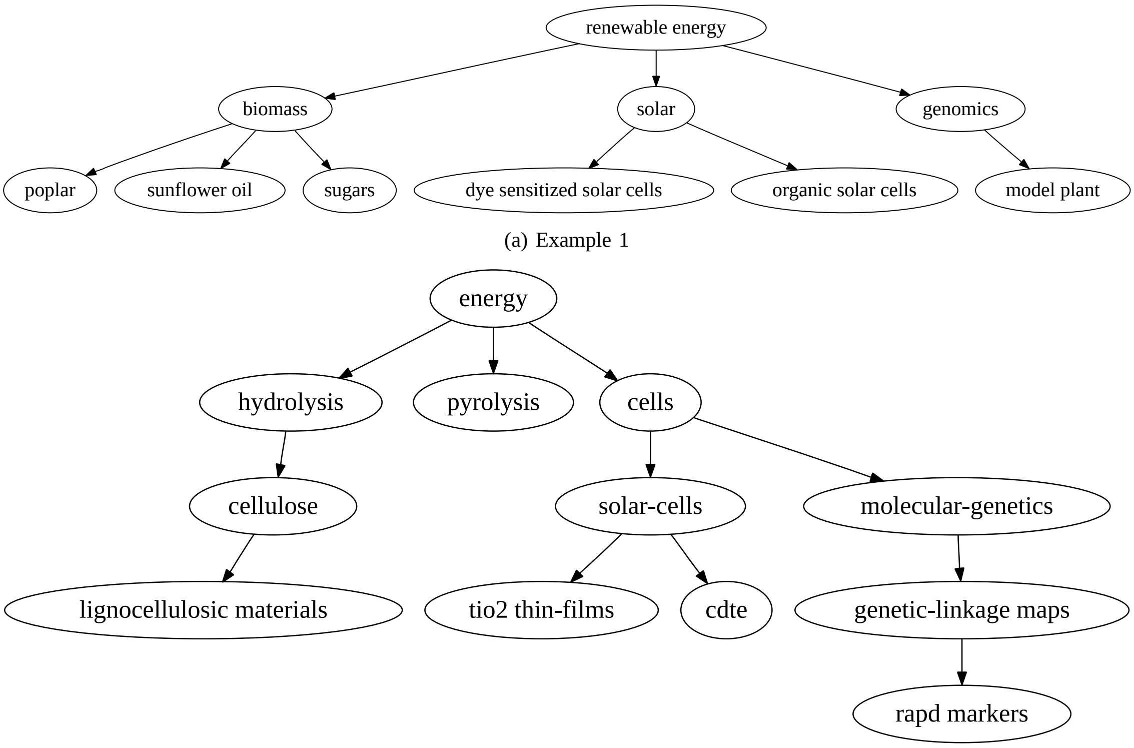
However, one way in which we can try to improve the situation is by incorporating information regarding global structure into the process, as this will hopefully reduce glaring inconsistencies within the generated taxonomies. As an initial measure, we propose the following weighted cost function for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



(a) Example 1

(b) Example 2

Fig. 2. Sample renewable energy taxonomies



(a) Example 1

(b) Example 2

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

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).

Intuitively, as we traverse the tree from any node v towards the root, the distances $\overrightarrow{\mathrm{NGD}}(v, v^i_{\mathcal{E}^*})$ 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. Edmond's algorithm

Finding the minimum arborescence for a digraph can be done efficiently using Edmond's algorithm [Korte and Vygen, 2006]. 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 $u = \mathrm{argmin}_u \{w[e(u,v)] : u \in \mathcal{V}, u \neq v\}$
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:
                $w[e(u,v')] = w[e(u,v)] - w[e(x(v),v)] \ldots + \sum_{\{e: e\in\mathcal{E}', e\neq x(v)\}} w[e]$,
            where, $x(v)$ is the immediate ancester 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 (2) until all cycles have been eliminated
```

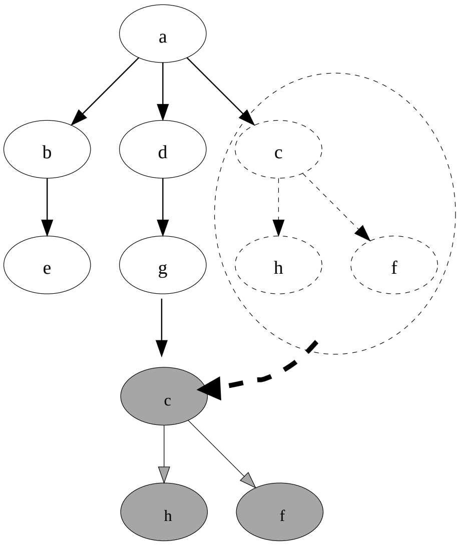


Fig. 4. Taxonomy tree mutation operator. The dashed lines denote nodes and edges which are to be removed.

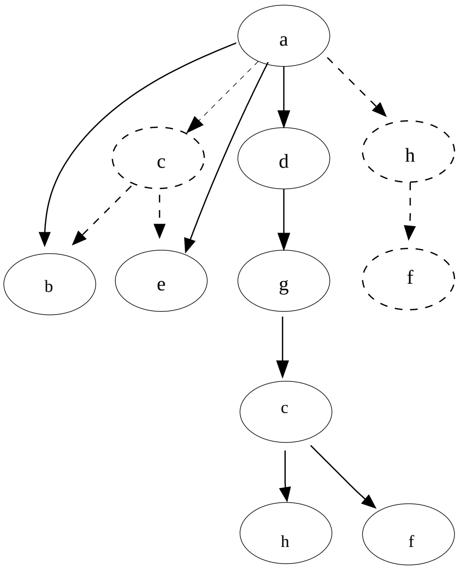


Fig. 6. Chromosome repair process. The dashed lines denote nodes and edges which are to be removed.

B. 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, the situation in cases when the cost function incorporates custom modifications or constraints is less clear.

In particular, Edmond's algorithm is inapplicable for the cost function in eq. (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.

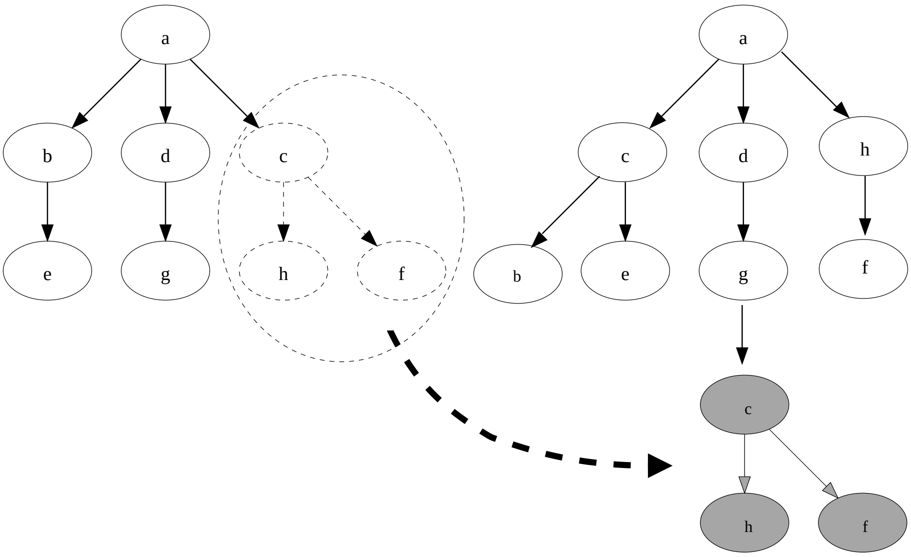


Fig. 5. Taxonomy tree crossover operator (stage 1). The dashed lines denote nodes and edges which are to be removed.

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. Instead, we adopt the following two customized operations for chromosome transformation:
   - *Mutation* - the mutation procedure operates on individual trees. A random subtree is moved from one point of the hierarchy to another randomly selected point in the same tree (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.

C. 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 I 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

The experiments described in the previous sections were conducted. The Author keywords extracted from the top 35 cited papers on "renewable+energy" from the Web of Science database where collected and the taxonomy generating process described in the preceeding sections carried out.

To facilitate presentation and analysis of the results, the collection was randomly divided into two subsets - set one contains 35 keywords, and set two contained the remaining 37 keywords. In addition, any occurrences of the stop-words described in section III-C 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 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 branches emanate 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) were organized into a taxonomy using the proposed approach. The resulting graphs are shown in fig. 8.

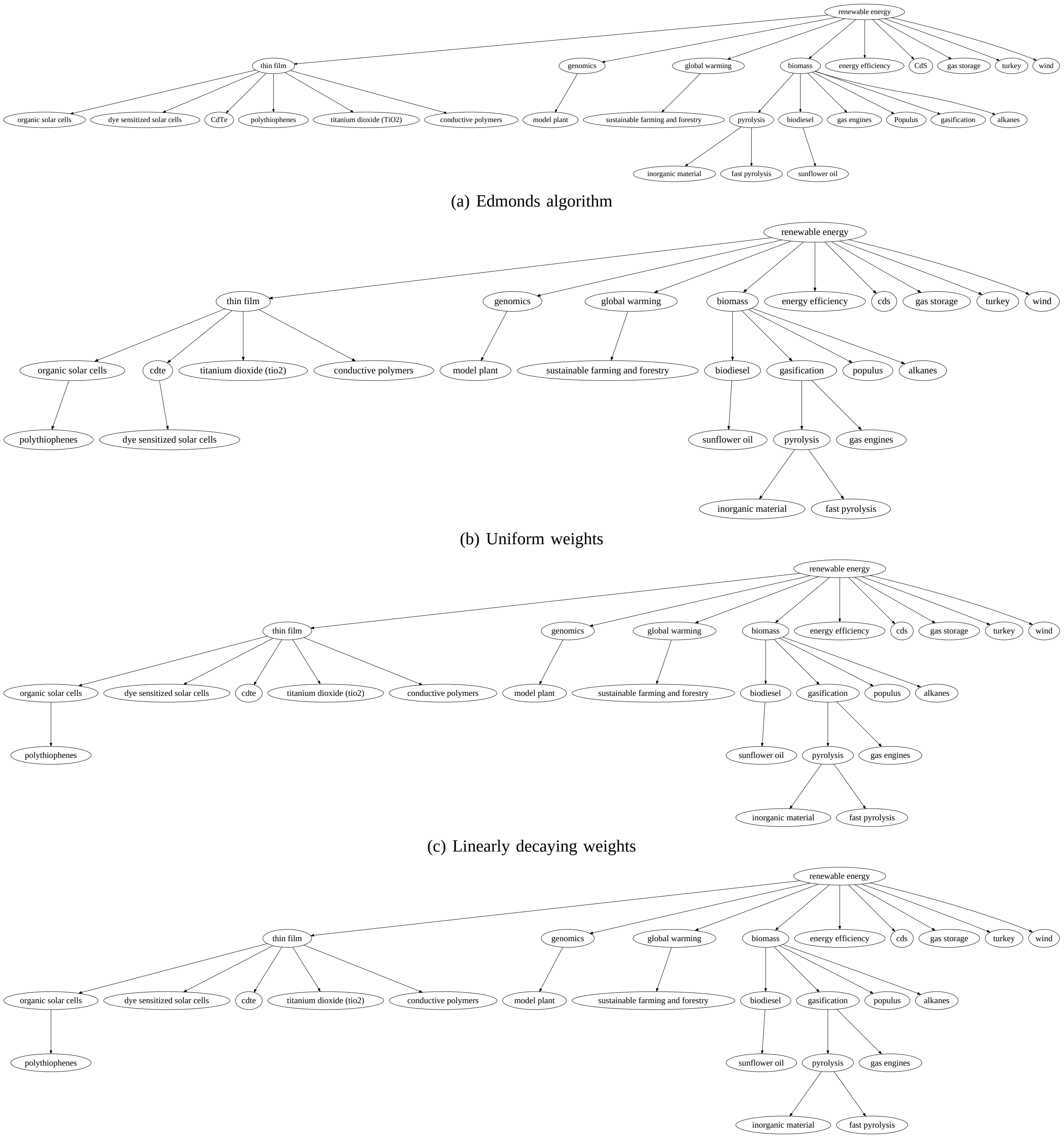
Our observations on these graphs are:

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 terms frequently appears in articles that are broader in scope, such as review papers and papers on various strategic issues such as global warming, energy markets and the like. This allows the term to attract a broad range of "subclasses" which may not easily fit into other sections of these taxonomies. In particular, note that many of the terms descended from *natural gas* are themselves fairly broad in nature - and would likely appear in similar publications.
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



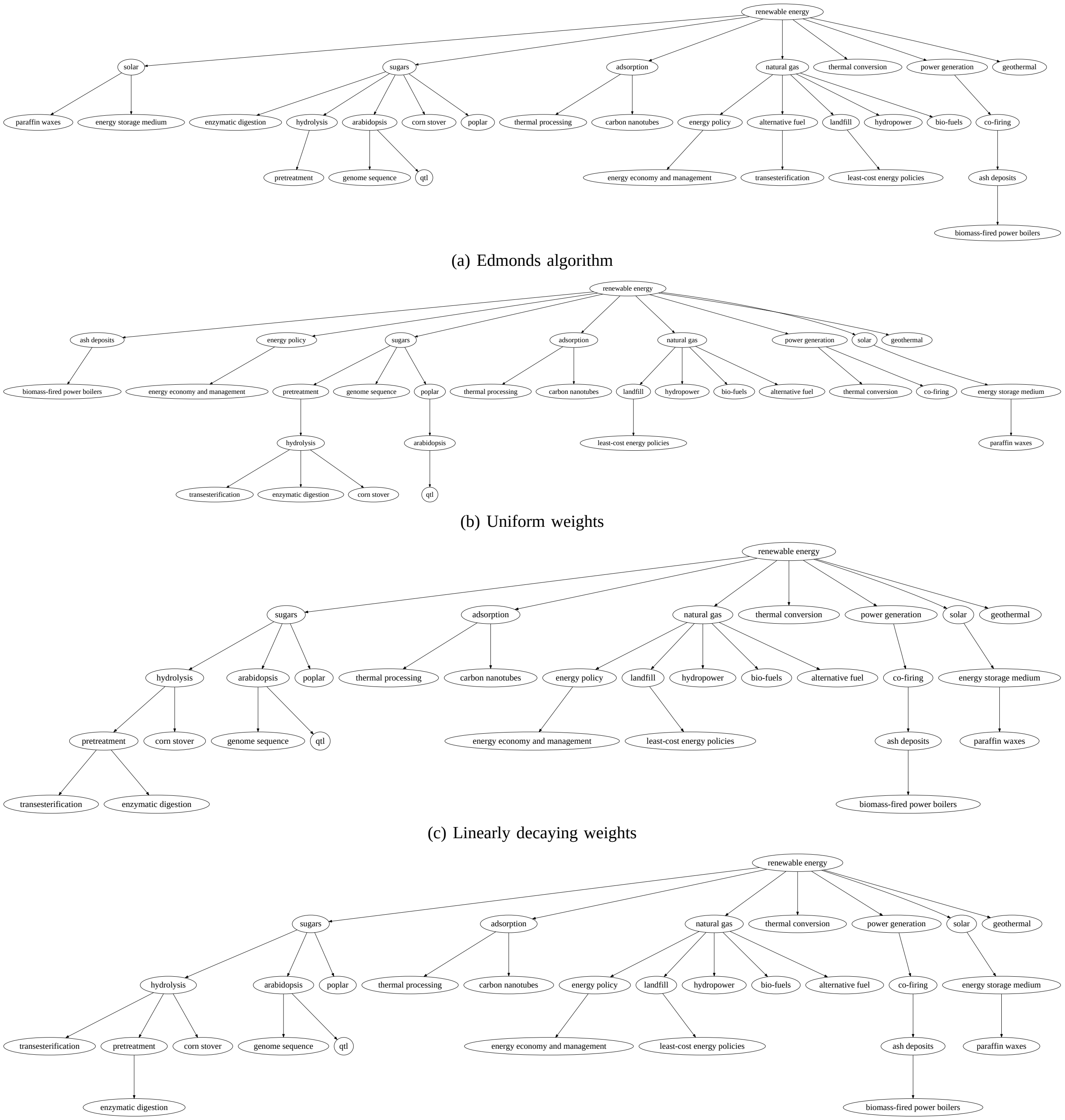
(a) Edmonds algorithm

(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**

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.
4) Inadequacy of existing data analysis tools; while - through the research presented here - we have tried to push the envelope on this front, the problems encountered when dealing with complex, high dimensional data are common to many application domains and are the subject of much ongoing research besides our own. Problems related to the overfitting of data, non-unique solutions and information loss resulting from dimensionality reduction, are all symptoms of the inherent difficulty of this problem.

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-occurence 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.

Our future plans include working more closely with domain experts to improve and validate the results produced using the proposed methodology.

Appendix I
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.

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

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

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

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

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

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

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

[Losiewicz et al., 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.

[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).