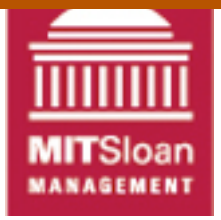

MIT Sloan School of Management

MIT Sloan School Working Paper 4726-09
9/1/2008

Latent Semantic Analysis Applied to Tech Mining

Blaine Ziegler, Wei Lee Woon, Stuart Madnick

Latent Semantic Analysis Applied to Tech Mining

Blaine Ziegler
Wei Lee Woon
Stuart Madnick

Working Paper CISL# 2008-12

September 2008

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

Latent Semantic Analysis Applied to Tech Mining

Blaine Ziegler, Wei Lee Woon, Stuart Madnick

September 2008

Abstract – This paper presents an approach to bibliometric analysis in the context of technology mining. Bibliometric analysis refers to the use of publication database statistics, e.g., hit counts relevant to a topic of interest. Technology mining facilitates the identification of a technology's research landscape. Our contribution to bibliometrics in this context is the use of a technique known as Latent Semantic Analysis (LSA) to reveal the concepts that underlie the terms relevant to a field. Using this technique, we can analyze coherent concepts, rather than individual terms. This can lead to more useful results from our bibliometric analysis. We present results that demonstrate the ability of Latent Semantic Analysis to uncover the concepts underlying sets of key terms, used in a case study on the technologies of renewable energy.

1 Introduction

1.1 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 available to researchers and decision-makers. One difficult problem is the need to gain a broad understanding of the current state of research, future scenarios and the identification of technologies with potential for growth and which hence need to be emphasized. Information regarding past and current research is available from a wide variety of channels (examples of which include publication and patent databases); the task of extracting useable information from these sources, known as "tech-mining" [Porter, 2005], presents both a difficult challenge and a rich source of possibilities; on the one hand, sifting through these databases is time consuming and subjective, while on the other, they provide a rich source of data with which a well-informed and comprehensive research strategy may be formed.

There is already a significant body of research addressing this problem (for a good review, the reader is referred to [Porter, 2005, Porter, 2007, Losiewicz et al., 2000, Martino, 1993]); 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 [Smallheiser, 2001, Daim et al., 2005, Daim et al., 2006, Small, 2006]. Nevertheless, given the many difficulties inherent to these undertakings, there is still much scope for further development in many of these areas.

1.2 Novelty and motivations

An important motivation for technology-mining is the possibility of gaining a better understanding of future developments and trends in a given field of research. This is a complex task that is composed of a number of closely inter-related components or activities. While there is no single authoritative classification, we present the following scheme, proposed in [Porter et al., 1991], to help focus our discussion:

- Monitoring - Observing and keeping up with developments occurring in the environment, and which are relevant to the field of study [Kim and Mee-Jean, 2007, King, 2004].
- Expert opinion - An important method for forecasting technological development is via intensive consultation with subject matter experts [Van Der Heijden, 2000].
- Trend extrapolation - This involves the extrapolation of quantitative historical data into the future, often by fitting appropriate mathematical functions [Bengsiu and Nekhili, 2006].
- Modeling - It is sometimes possible to build causal models which not only allow future developments to be known, but also allow the interactions between these forecasts and the underlying variables or determinants to be better understood [Daim et al., 2005, Daim et al., 2006].
- Scenarios - Forecasting via scenarios involves the identification of key events or occurrences which may determine the future evolution of technology [Mcdowall and Eames, 2006, Van Der Heijden, 2000].

In this context, the emphasis of the current study is on the first item, technology monitoring, as the primary objective is to devise methods for monitoring, understanding and mapping the current state of technology. In particular, our aim is to develop novel approaches to visualize and understand the concepts that underlie terms in a field of science and technology. Towards this end, this paper will address the following objectives:

1. To utilize a technique known as Latent Semantic Analysis (LSA) to quantitatively identify such underlying concepts in areas of research.
2. To conduct a preliminary case study in renewable energy by using the results from LSA, as a demonstration of the proposed approach.

1.3 Case study

To provide a suitable example on which to conduct our experiments and to anchor our discussions, a preliminary case study was conducted in the field of renewable energy.

The importance of energy to the continued well-being of society cannot be understated, yet 87%[1] of the world's energy requirements are fulfilled via the unsustainable burning of fossil fuels. A combination of environmental, supply and security problems compounded the problem further, making renewable energies such as wind power and solar energy one of the most important topics of research today.

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

2 Latent Semantic Analysis

Towards our goal of technology mining, we use bibliometric techniques; i.e., analysis of hit counts[2] returned by publication databases[3] for technologies of interest. However, directly querying just the name of the technology can give misleading results. This is because a one-word search query tends not to be representative of the field as a whole. For example, if we are interested in the state of the renewable energy field, we would want to include in our search applicable technologies such as hydroelectric power and wind power, in order to get a representative number of hits. Similarly, if we are interested in "oil," we might also like to include "petroleum," since some documents may use that word instead.

We want to automate the process of collecting related terms. It is not enough to make educated guesses as to which terms should be combined in search queries. Instead, we seek a mathematically sound algorithm for generating groupings of terms based on data from publication databases. One such technique, clustering, is presented in [Woon and Madnick, 2008]. In this section, we propose to use a different approach, known as Latent Semantic Analysis, which is used to reveal the underlying concepts that relate terms to one another.

2.1 Latent Semantic Analysis background

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

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

[2] The "hit count" is the number of documents that were found that contain the specified term(s).

[3] Examples of publication databases includes Google Scholar, Scirus, etc.

(literally orthogonal vectors in a space, shown later), but the terms that make up each concept are found, with some weighting, in all of the concepts.

LSA is a matrix algebra process. The procedure takes as input a *term-document matrix* [Berry et al., 1995]. This matrix is a representation of the frequency of occurrence of each term in each document in the database. Terms are listed along the rows of the matrix, and documents are listed along the columns.

The frequency values in the term-document matrix can be obtained from a few different metrics. One straightforward approach is to simply use the count of the number of times the given term occurs in the given document. A second method normalizes these counts by the total number of words in the document. If a term appears 5 times in a document with 500 words, its value in the term-document matrix would be 0.01. A third method, known as Term Frequency-Inverse Document Frequency (tf-idf), further normalizes this term frequency by the fraction of documents in the entire database that contain the given term [Landauer et al., 1998]. In the example of the term that appears 5 times in a 500-word document, if that term also appears in all other documents in the database, its document frequency will be 1, and the value that goes in the term-document matrix will be 5 divided by 1, or 5. If, however, the term occurs in only 1% of all documents, the term-document value will be 5 divided by 0.01, or 500. Any of the above methods produce suitable term-document matrices for LSA.

Given a suitable term-document matrix, the next step in LSA is to calculate the Singular Value Decomposition, SVD, of the matrix. In general, the SVD is defined as follows: Given an $m \times n$ matrix A, the SVD factors it into the form

$$U \sum V^T.$$

U is $m \times m$ and contains the eigenvectors of AA^T. V is $n \times n$ and contains the eigenvectors of $A^T A$. $\sum$ is an $m \times n$ diagonal matrix that contains the square roots of the eigenvalues of AA^T along its diagonal. It is also important to note that U and V are orthonormal matrices. Their columns are unit-magnitude vectors orthogonal to one another.

In the context of LSA, the columns of the U matrix contain our "concepts" [Berry et al., 1995]. Recall that the $m \times n$ term-document matrix (m terms and n documents) produces an $m \times \mathrm{m}$ U matrix. Each of the m columns of U, or concepts, is then a vector in term-space. The concept vectors can be thought of as an orthonormal basis of the term space. This is intuitively appealing as it implies independence across the set of concepts; i.e., knowledge that a document contains one particular concept gives no information about whether that document also contains a different concept. This is as opposed to the use of terms or concepts that are not independent, say "car" and "automobile." The likelihood of a document containing "automobile" increases if one knows that the document also contains the word "car."

2.2 Intuitive appeal

As a hypothetical example of LSA, consider the following set of terms: Storm, Lightning, Bolt, Nut, Muffin, Whale.

Note that the words "Bolt" and "Nut" have (at least) two different meanings. If one were interested in searching for documents related to storms, the word "Bolt" might be a poor choice of search term because unwanted documents on machining would also be returned. Similarly, if one were interested in food with nuts in it, a direct search for "Nut" would also return irrelevant results. For the search for storm documents, "Storm" or "Lightning" would also return non-ideal search results, but for the opposite reason: some documents may contain one term but not the other, and a simple search for only one of the terms would leave out some relevant documents. The semantic similarity between terms such as "Storm," "Lightning," and "Bolt," as well as the multiple meanings of terms such as "Bolt" and "Nut" suggest that we would do better by first breaking the terms down into independent concepts. These concepts are, intuitively:

[Storm, Lightning, Bolt]

[Bolt, Nut]

[Nut, Muffin]

[Whale]

The results returned from actual use of LSA would be similar, but in a slightly different form. The two key differences are that: (1) each concept contains a combination of *every* term, and (2) there are as many concepts as there are terms (six in this case). The first difference is usually alleviated by the fact that most concepts contain terms with weightings that are virtually zero. If very small values are rounded to zero, then concepts can be considered to contain only a subset of the terms rather than all of them. The second difference, that there must be as many concepts as there are terms, can be harder to grasp intuitively. In some cases, the "extra" concepts can be thought of as concepts that do in fact exist in the set of documents, but only actually occur a small number of times. In other cases, a concept may appear to be an "inverted" version of another. For example, we may see [Bolt: 0.8, Nut: 0.6] as one concept, and [Nut: 0.8, Bolt: 0.6] in another.

While it may be difficult in some cases to interpret the results from LSA, it is important to keep in mind that the results are simply a representation of the data given to it; they are neither right nor wrong. In cases where the results seem puzzling, it could mean either that the documents have an unusual bias to them, or, more interestingly, that we did not previously have an accurate mental picture of the interrelationships among the terms being analyzed.

2.3 Modification for practical implementation

For our purposes, we use a novel modification to the standard LSA algorithm. The sheer number of documents to be searched renders the first step of standard LSA – generating the term-document matrix – highly impractical. Such a matrix could potentially have millions of columns. Additionally,

determining each value in the matrix would require searching the complete text of each document to count term occurrences. Furthermore, creating this matrix would only begin to be feasible if all documents were available on a local machine, which is an unacceptable limitation for our purposes given the sources that we are using (e.g., Scirus, Google Scholar), which are not under our control or available for local usage.

Our approach instead allows us to easily use the billions of published papers available on the Internet as our document set. We only have to be willing to accept a simplifying assumption about our documents. The key assumption is that any given term occurs in any given document either 0 times or 1 time, i.e. that all document vectors in the term-document matrix are binary encoded and cannot contain values other than 0 or 1. While this is clearly not completely accurate, the assumption gives us the ability to use enormous online collections of documents which makes this approach far superior to standard LSA for our purposes.

Instead of counting term occurrences in documents, we use only the hit count returned from a search engine to generate a term *covariance matrix*. If we are analyzing m terms, the covariance matrix A is $m \times \mathrm{m}$, with each value $A_{i,j}$ representing the covariance between terms i and j. Such a matrix could have been derived from the term-document matrix, but for our purposes, our simplifying assumption allows us to construct an approximation of the covariance matrix from hit counts alone.

The derivation of the covariance matrix is as follows. We treat the numbers of occurrences of a particular term in each document as realizations of a random variable. Then, by definition, the covariance of two of these random variables (the covariance between two terms i and j in a term-document matrix A) is

$$cov(i,j) = E[(A_i - E[A_i]) \times (A_j - E[A_j])]$$

The expectations of A_i and A_j are calculated by taking the average of each value across all n documents for the given term. For example, $E[A_i]$, the expectation of the number of occurrences of term i, is $\frac{1}{n}\sum_{k=1}^{n} A_{i,k}$.

The outer expectation is similarly calculated by averaging the values for terms i and j across all n documents. Then our covariance expression becomes

$$\mathrm{cov}(i,j) = \frac{1}{n}\sum_{k=1}^{n}\left(\left(A_{i,k} - \frac{1}{n}\sum_{l=1}^{n} A_{i,l}\right) \times \left(A_{j,k} - \frac{1}{n}\sum_{l=1}^{n} A_{j,l}\right)\right).$$

This expands to

$$cov(i,j) = \frac{1}{n}\left(\sum_{k=1}^{n} A_{i,k}A_{j,k} - \frac{1}{n}\sum_{l=1}^{n} A_{i,l}\sum_{k=1}^{n} A_{j,k} - \frac{1}{n}\sum_{l=1}^{n} A_{j,l}\sum_{k=1}^{n} A_{i,k} + \frac{1}{n}\sum_{l=1}^{n} A_{i,l}\sum_{l=1}^{n} A_{j,l}\right).$$

Combining the last three terms gives

$\mathrm{cov}(i,j) = \frac{1}{n}\left(\sum_{k=1}^{n} A_{i,k}A_{j,k} - \frac{1}{n}\sum_{k=1}^{n} A_{i,k}\sum_{k=1}^{n} A_{j,k}\right)$.

Simplification of this expression requires us to utilize our assumption that each value in the term-document matrix is either a 0 or a 1. With this simplification in mind, $\sum_{k=1}^{n} A_{i,k}$ represents the number of documents that contain term i, $\sum_{k=1}^{n} A_{j,k}$ represents the number of documents that contain term j, and $\sum_{k=1}^{n} A_{i,k}A_{j,k}$ represents the number of documents that contain both term i and term j. This is represented with this final covariance expression.

$\mathrm{cov}(i,j) = \frac{1}{n}\left(h_{i,j} - \frac{1}{n}h_i h_j\right)$.

In the above expression, $h_{i,j}$ represents the number of hits returned from a search for both terms, while h_i and h_j represent the number of hits returned for terms i and j respectively. n represents the number of documents being searched. We approximate n with the number of hits returned from a search for a large term that subsumes terms i and j. For our purposes, we use the field that is the focus of our case study, renewable energy, as the search term to acquire n.

Using our covariance expression, which was approximated under the assumption that a document can contain a given term zero times or one time, we can construct a term-by-term covariance matrix. Given this matrix, we can then obtain our desired concept vectors by simply calculating its eigenvectors. In other words, the eigenvectors of our covariance matrix are equivalent to the columns of the U matrix obtained from the singular-value decomposition of the term-document matrix.

2.4 Demonstration of basic workability of our LSA approach

In this section we present LSA results for the keyword set from section 2.2 of this paper. The terms, again, are [Storm, Lightning, Bolt, Nut, Muffin, Whale]. These terms were chosen to illustrate two different functions of LSA. The first is to highlight related terms; i.e., terms that are similar in meaning or terms that tend to co-occur due to topic similarity. In this particular set of keywords, [Storm, Lightning, Bolt], [Bolt, Nut], and [Nut, Muffin] all illustrate topical similarity. The other, opposite, function of LSA is to identify terms that have multiple meanings or belong to multiple concepts. In this example, Bolt is a homonym; a bolt of lightning is different from a bolt used for assembly. (The two uses may be derived from the same underlying idea, but LSA still identifies the fact that they are used differently.) Nut is also a homonym; it relates to Bolt and to Muffin in entirely different ways. Finally, the word Whale is added to the set as an example of a lone term unrelated to the others in the set.

Our LSA analysis was performed on this keyword set using the Scirus database. The following concept vectors were obtained:

Concept Vector 1		Concept Vector 2		Concept Vector 3	
Storm	0.9996	Nut	-0.9910	Bolt	-0.9907
Lightning	0.02543	Bolt	-0.1333	Nut	0.1335
Nut	9.696e-3	Storm	0.01054	Lightning	-0.02586
Whale	8.828e-3	Whale	-6.454e-3	Whale	-6.030e-3
Bolt	5.397e-3	Lightning	-5.838e-3	Storm	4.765e-3
Muffin	5.157e-5	Muffin	-1.614e-3	Muffin	2.412e-4

Concept Vector 4		Concept Vector 5		Concept Vector 6	
Whale	-0.9994	Lightning	-0.9988	Muffin	0.999998
Lightning	-0.03310	Whale	0.03354	Nut	-1.629e-3
Storm	9.570e-3	Bolt	0.02631	Whale	-7.107e-4
Bolt	7.769e-3	Storm	0.02495	Lightning	5.253e-4
Nut	5.761e-3	Nut	2.390e-3	Storm	-4.295e-5
Muffin	-6.831e-4	Muffin	5.531e-4	Bolt	1.428e-5

The terms in each vector are sorted by decreasing absolute value. The significance of negative values is still under investigation, but we currently consider only the absolute value of concept vector elements. Given this sorting, we can see that the results roughly correspond to intuition.

Concept 1 mostly contains Storm, with a contribution from Lightning. **Concepts 2 and 3** both contain Nut and Bolt, with one emphasizing Nut and the other emphasizing Bolt. The **fourth** is mostly about Whale, although Lightning seems to make a small contribution. The inclusion of Lightning may seem counterintuitive, but again, it simply represents the data given to it. Evidently, some articles contain both Whale and Lightning. This surprising result should be viewed in a positive light; after all, the purpose of LSA is to uncover latent associations between terms. **Concept 5** is an inversion of 4, in that Lightning is the focus, with a contribution from Whale. And finally, the **sixth concept** is almost entirely Muffin. Intuitively, we had predicted an association between Muffin and Nut. The discrepancy is likely due to our choice of database. The article source is Scirus, a publication database. It is unlikely that many articles contained Muffin at all, and those that did apparently did not mention Nuts. This is corroborated by the fact that Muffin has practically zero significance in each of the other five concepts. Again, this is an example of LSA showing us something that we did not previously realize; i.e., that Muffin and Nut are not closely associated in the set of articles that were used.

3 Case Study Results

Results are presented here for a set of 59 keywords related to our renewable energy case study, compiled from authors of renewable energy publications. These terms have been used in another related study reported in [Woon and Madnick, 2008]. The keywords are the following:

[ash deposits, alternative fuel, natural gas, renewable energy, Review, sugars, biomass, energy balance, model plant, energy conversion, CdTe, transesterification, enzymatic digestion, ENERGY

EFFICIENCY, investment, gas engines, Populus, electricity, pretreatment, gasification, GLOBAL WARMING, adsorption, high efficiency, genome sequence, arabidopsis, bio-fuels, energy economy and management, QTL, renewables, thermal conversion, co-firing, inorganic material, fuels, energy sources, genomics, thermal processing, biodiesel, SUSTAINABLE FARMING AND FORESTRY, gas storage, chemicals, carbon nanotubes, GASIFICATION, CdS, sunflower oil, energy policy, POWER GENERATION, LEAST-COST ENERGY POLICIES, pyrolysis, biomass-fired power boilers, BIOMASS, thin films, landfill, coal, corn stover, poplar, emissions, RENEWABLE ENERGY, fast pyrolysis, hydrolysis]

Due to the size of the data set, it is infeasible to display complete results, which contain 59 concept vectors of 59 terms each. Instead, we display a subset of vectors, and only the significant terms in each; i.e., those whose values are not vanishingly small.

Typically with eigenvector analysis, the most important vectors are those with the largest associated eigenvalues. Those vectors are considered to be more significant because they represent the greatest variance in the data. However, for our purposes, we are at least as interested, and perhaps more so, in the concepts with lower variance. Methods for choosing good, representative subsets of concept vectors calculated by LSA are still being investigated, but for now, interesting vectors are largely chosen by inspection. Such a sample of vectors is displayed here for the renewable energy keyword set.

Concept Vector 1		**Concept Vector 2**		**Concept Vector 3**	
renewable energy	-0.7073	alternative fuel	0.8811	POWER GENERATION	0.6977
RENEWABLE ENERGY	0.6887	biodiesel	-0.4372	electricity	-0.3787
BIOMASS	-0.09997	thin films	0.09212	energy policy	-0.3271
biomass	0.09202	natural gas	-0.07202	coal	-0.2316
		bio-fuels	-0.05371	Review	0.2183
		sugars	0.05141	fuels	0.2139
				adsorption	-0.2045
				ENERGY EFFICIENCY	0.1311

Concept Vector 4		**Concept Vector 5**		**Concept Vector 6**	
ENERGY EFFICIENCY	0.4777	pyrolysis	-0.6103	landfill	-0.8820
electricity	0.4752	pretreatment	-0.4823	energy balance	0.2344
energy policy	-0.3904	hydrolysis	-0.3845	RENEWABLE ENERGY	0.1891
renewables	-0.2674	sugars	0.2638	sugars	0.1768
GLOBAL WARMING	-0.2458	chemicals	0.1964	renewable energy	0.1757
investment	-0.2405	landfill	0.1343	chemicals	-0.1079
		GASIFICATION	0.1228	emissions	0.09202
		fast pyrolysis	-0.1221		
		gasification	0.1139		

The **first concept** vector illustrates the ability of LSA to clean up inconsistent data. RENEWABLE ENERGY and renewable energy, along with BIOMASS and biomass become grouped together, just as they should. It is interesting to note that although their (absolute) values are close to one another, they are not equal. Search engines disregard case, so renewable energy and RENEWABLE ENERGY are interpreted exactly the same. However, they are given to the LSA algorithm as two separate (but equal) terms, and are therefore each independently susceptible to noise. In this particular case, not only are the absolute values slightly different, but so are the signs: renewable energy is weighted negatively at -0.7073, while RENEWABLE ENERGY has a positive weight, 0.6887. Similarly, biomass and BIOMASS have positive and negative weightings, respectively. This is a curious observation that lends credence to the idea that only absolute value, and not sign, should be considered when evaluating the importance of a term in a concept.

renewable energy	-0.7073
RENEWABLE ENERGY	0.6887
BIOMASS	-0.09997
biomass	0.09202

Concept 2, as its primary component suggests, is about alternative fuels. Biodiesel, natural gas, bio-fuels, and sugars all fall in this category. Thin films is only slightly off topic. They can be used to increase the efficiency of solar power systems, so while not an alternative fuel per se, they are an important component of an alternative energy.

alternative fuel	0.8811
biodiesel	-0.4372
thin films	0.09212
natural gas	-0.07202
bio-fuels	-0.05371
sugars	0.05141

Concept 3 broadly consists of power generation topics. Energy policy seems a bit out of place, as the rest of the concept mostly deals with technology, rather than policy. But this tells us that it is common to find information regarding energy policy in the same articles as those about energy technologies. Review also appears to be wildly out of place. However, the explanation here is that review is simply too broad of a term to get clean LSA results from it. Evidently, it appears in articles regarding energy technology, even though its actual meaning is unrelated. In the same way, the word "the" would be a large component of virtually all LSA concept vectors if it were included in the term list.

POWER GENERATION	0.6977
electricity	-0.3787
energy policy	-0.3271
Coal	-0.2316
Review	0.2183
fuels	0.2139
adsorption	-0.2045
ENERGY EFFICIENCY	0.1311

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

ENERGY EFFICIENCY	0.4777
electricity	0.4752
energy policy	-0.3904
renewables	-0.2674
GLOBAL WARMING	-0.2458
investment	-0.2405

The terms in **concept 5** represent, for the most part, chemical breakdown of organic compounds. Pyrolysis, the first term in the vector, refers specifically to this process. Fast pyrolysis, found in this vector with a smaller weight, is clearly a related term. Hydrolysis is a process related to pyrolysis that refers to breaking water down into hydrogen and oxygen. Gasification (and GASIFICATION) similarly refers to the decomposition of organic materials into carbon monoxide and hydrogen. Sugars and chemicals belong in this concept as well; hydrolysis of disaccharide sugars produces monosaccaride sugars, and chemicals, while a somewhat general and vague term, clearly applies to the concept as a whole. Pretreatment and landfill seem a bit out of place in this concept, but the purpose of LSA is to reveal unknown term relationships, which is the case here.

pyrolysis	-0.6103
pretreatment	-0.4823
hydrolysis	-0.3845
sugars	0.2638
chemicals	0.1964
landfill	0.1343
GASIFICATION	0.1228
fast pyrolysis	-0.1221
gasification	0.1139

Finally, **concept 6** illustrates that in some cases, one will see vectors that intuitively make little sense. Aside from renewable energy and RENEWABLE ENERGY appearing in the same concept, as expected, the terms in this concept span a large range of fields, loosely linked only by the umbrella field of "energy." Recall that with our keyword set of 59 terms, LSA produces 59 concept vectors. It seems unlikely for each vector to have a clear, unique meaning, and in fact, a large number of the generated vectors, while mathematically meaningful, are not of much use to us. This is why it is important to devise an automated method for choosing useful vectors; this is currently being investigated.

landfill	-0.8820
energy balance	0.2344
RENEWABLE ENERGY	0.1891
sugars	0.1768
renewable energy	0.1757
chemicals	-0.1079
emissions	0.09202

Visualization

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

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

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

Two plots are shown in this section. The first uses concepts 3 and 4 as the axes, and the second uses concepts 4 and 5.

The first was chosen for display because concepts 3 and 4 have some terms that are important components of each. If the two concepts were completely unrelated, then all terms would appear along one axis or the other, representing the fact that no terms have a high importance in both concepts. But in this case, we see some off-axis terms, showing that concepts 3 and 4 do appear to be related.

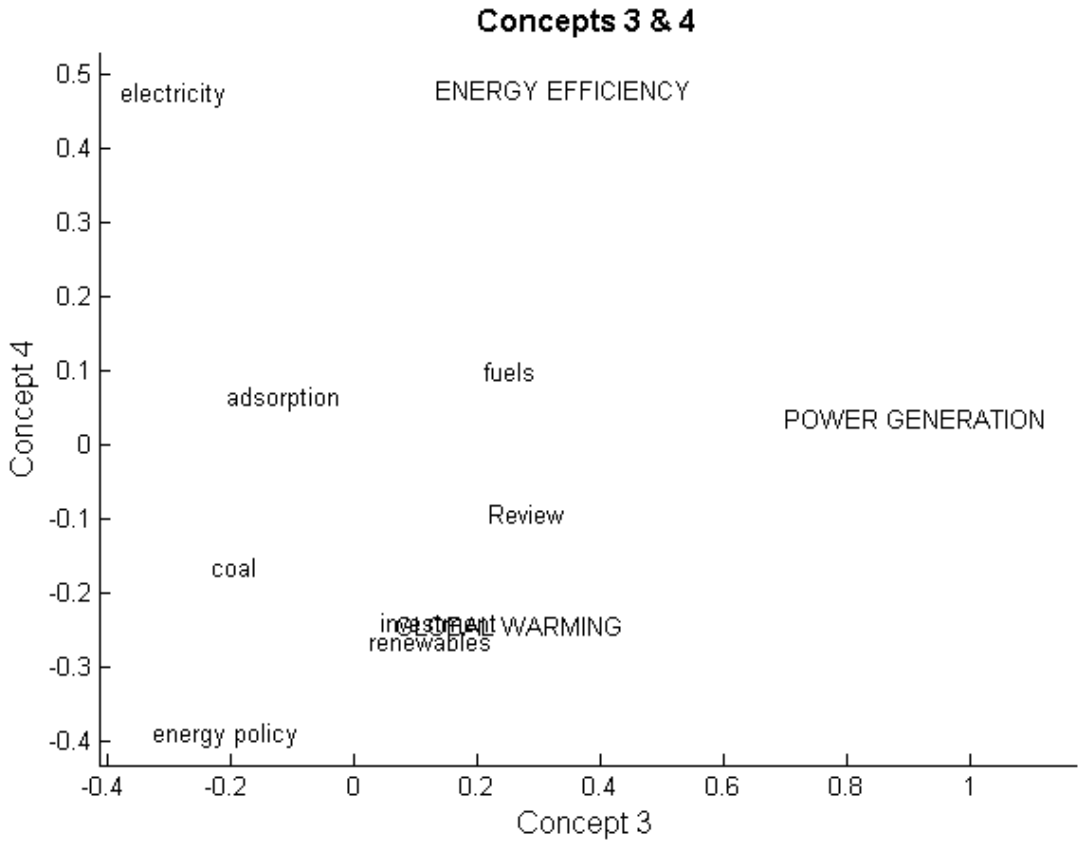

Figure 1: Concepts 3 and 4. Terms such as energy policy and electricity lie well off of the coordinate axes, showing that they are important components of both concepts.

The second graph, which plots concepts 4 and 5 together, was chosen because of the clear clusters that it identifies. Four groupings are formed – sugar and chemicals; energy policy, GLOBAL WARMING, renewables, and investment; electricity and ENERGY EFFICIENCY; and hydrolysis, pretreatment, and pyrolysis.

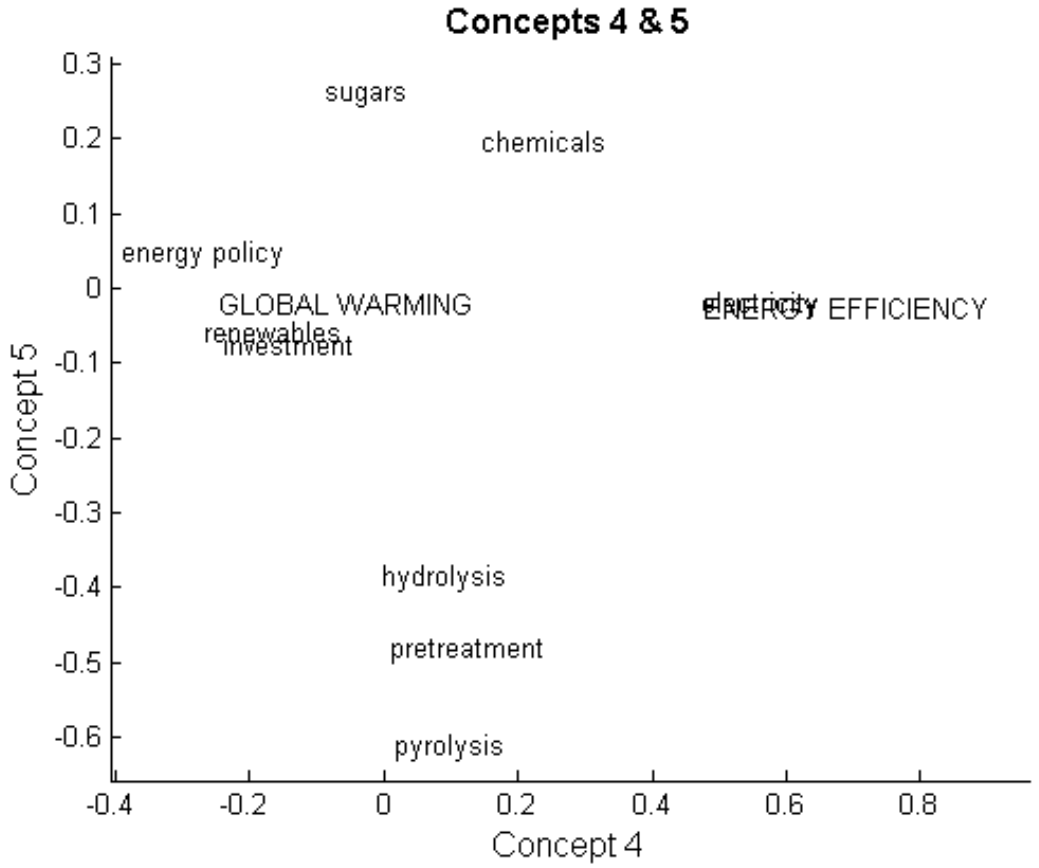

Figure 2: Concepts 4 and 5. Four term clusters become apparent when these concepts are plotted together.

By viewing two vectors at a time instead of just one, we can gain some insight from the different perspective it provides.

4 Discussions

This paper presented the Latent Semantic Analysis technique for revealing the underlying concepts behind a set of terms from an area of research. LSA should prove useful in our ongoing research on technology landscaping and forecasting. With the help of LSA, we can query publication databases for entire concepts, rather than individual terms, which can provide a more accurate picture of the number of publications relevant to a field.

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

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

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.

[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.

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

[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.

[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.

[King, 2004] King, D. A. (2004). The scientific impact of nations. *Nature*, 430(6997):311–316.

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

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

[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.

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

[Mcdowall and Eames, 2006] Mcdowall, W. and Eames, M. (2006). Forecasts, scenarios, visions, backcasts and roadmaps to the hydrogen economy: A review of the hydrogen futures literature. *Energy Policy*, 34(11):1236–1250.

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

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

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

[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.

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

[Woon and Madnick, 2008] Woon, W.L. and Madnick, S. (2008). Semantic distances for technology visualization, MIT Sloan School Working Paper 4711-08, August 2008.