U.S. DEPARTMENT OF ENERGY

PNNL- 19493

Prepared for the National Biodefense Analysis and Countermeasures Center (NBACC)

Morphing Terminology Study

SJ Rose
FJ Brockman
ML Hart
DW Engel
NB Valentine
AJ Calapristi

June 2010

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL- 19493

Morphing Terminology Study

SJ Rose
FJ Brockman
ML Hart
DW Engel
NB Valentine
AJ Calapristi

June 2010

Prepared for the National Biodefense Analysis and Countermeasures Center (NBACC)

Pacific Northwest National Laboratory
Richland, Washington 99352

Abstract

This study investigates methods of automatically identifying and characterizing significant transitions in term usage over time. Within scientific literature, the occurrence of terms reflects the use of technologies and techniques as well as the study of specific species and materials. Transitions in terminology usage may be a result of vocabulary standardization or specialization in which terms are replaced with their shorter form. They may also be a result of new applications, combinations, alternatives, or interests that result in the appearance of new or existing terminology in unexpected contexts.

Summary

Observations of term usage over time in scientific literature have shown clear emergence of certain terms and phrases. However, in many cases these "emergent terms" eventually lose their emergent signatures (i.e., occurrences of the term diminish) even though domain experts identified that the technology actually continued to advance. The hypothesis is that when a new technology emerges, the vocabulary associated with that technology becomes more specialized and changes (morphs) into new, but related, terminology. This study reports research on methods to automatically capture these phenomena and incorporate them into emergence detection algorithms and visual representations.

We demonstrate application of several methods that support the identification of morphing terminology and terminology transitions over time that aid users in identifying new technologies and their precursors, as well as new applications and combinations of science and technology. We expect that users will have an interest in exploring terminology transitions in relation to a focused context, defined through a set of documents, search terms, or computed themes. Temporal analysis of terms' document occurrences, co-occurrences, and associations within the user-defined context will reveal terms that are transitioning into or out of that context. The comparison of terms' feature associations and document associations provides a means to evaluate whether terms are being used as expected, or whether they are dropping out of (or into) a given context unexpectedly. This method also enables identification of terms that should appear in documents together based on their feature associations, regardless of whether they do in fact co-occur within documents.

Acronyms and Abbreviations

PCR	Polymerase Chain Reaction
RAKE	Rapid Automatic Keyword Extraction
SME	Subject Matter Expert

Glossary

Emergence	Algorithms to detect increasing or decreasing trend in topic
Feature	A term or attribute within a document
Term	Single word or multi-word noun or phrase
Term Association	Measurement of a term's association with a document, feature, or term
Term Document Occurrences	Number of documents in which a term occurs
Term Document Co-occurrences	Number of documents in which two terms co-occur
Term Document Association	Number of occurrences of a term in a document
Term Feature Association	Number of occurrences of a term in documents with a given feature
Term-pair	Two terms that are being compared
Surprise	Algorithms for detecting sudden event occurrence
Root Term	A term of particular interest for a user and that provides a context for analysis

Contents

Figures

Tables

1.0 Introduction

This study investigates methods of automatically identifying and characterizing significant transitions in term usage over time. Within scientific literature, the occurrence of terms reflects the use of technologies and techniques as well as the study of specific species and materials. Transitions in terminology usage may be a result of vocabulary standardization or specialization in which terms are replaced with their shorter form, such as ***Escherichia coli → E. coli***, or ***polymerase chain reaction → pcr***. They may also be a result of new applications, combinations, alternatives, or interests that result in the appearance of new or existing terminology in unexpected contexts. As expertise within a particular area develops, established topical areas may generate specialized sub-topics that focus on further developing specific scientific or operational aspects of a technology or research field.

Our initial tests focus on scientific literature that has been subjected to peer review prior to acceptance for publication. Peer review affects scientific literature by standardizing terminology and by reducing the redundancy of published material (in contrast with news sources, which frequently republish material). Transitions in terminology therefore likely reflect real changes in the sciences.

While the methods presented here may identify significant morphing terminology, the real significance of these transitions as insight to actual developments can only be evaluated by an informed and interested user. Therefore, this study will focus on defining the underlying information, assumptions, and utility for each method and on effective ways of presenting each method's results to a user.

2.0 Terms of Interest

We generally define a term as a single word or multi-word noun or phrase that conventionally labels a specific or general subject. Typical terms of interest may be highly specific, such as *Escherichia coli*, or very general, as *synthesis*, and their appearance in text documents may be very common, such as *dna*, or infrequent, as *mutagenic primers*. Statistics associated with the usage of these terms not only provide insight into the topical content but also can provide insight into how topics change over time.

In literature, these terms of interest belong to a larger class of content-bearing words and are distinct from function words in a language, such as the English function words *and*, *of*, *the*, and *for*, which occur in many documents. Because function words appear in documents independent of their content, they will have distribution patterns that are distinct from content-bearing words, which typically occur in clumps, or clusters, of related documents [Bookstein 1998].

For the purposes of this study, terms of interest occurring in the source documents are selected through the application of the Rapid Automatic Keyword Extraction (RAKE) method. RAKE automatically extracts keywords from individual documents and aggregates statistics across a set of documents to identify content-bearing keywords and keyphrases [Rose 2010]. These are our "terms of interest." We also refer to these terms of interest as *features* of the documents in which they occur. Appendix A lists the top 200 terms of the 2614 identified by RAKE for the BioTechniques dataset. This dataset comprises 5693 abstracts from the journal *BioTechniques*[1] from 1988 through 2008.

[1] BioTechniques, The International Journal of Life Science Methods, http://biotechniques.com

In order to apply the methods outlined in this study, a set of documents, such as technical abstracts, that includes each document's publication date is needed. This unstructured source information is processed to generate a list of terms of interest, a concordance (or inverted index) that provides access to the frequency of each term within each document, and a mapping of each document to its publication date. Taken together, this information can be applied to measure each term's document frequency, document associations, and feature associations across the entire time range or within arbitrarily defined time intervals.

3.0 Measuring Term Associations

Term associations provide a means to characterize and compare terms across documents within a collection, or dataset, and are measured as term document occurrences, term document co-occurrences, term document associations, and term feature associations. All term measurements are fundamentally based on counts of term document occurrences and co-occurrences with other terms (features).

Term Document Occurrences

Term document occurrences represent the number of documents in which a given term occurs. Occurrences can be aggregated into histograms or raw counts for sets of documents to indicate counts for specific time intervals and groups within the dataset. The Facets and Time tools within IN-SPIRE™ v5.0[2] provide examples of the analytic utility of this straightforward approach. While relative proportions to the dataset are not calculated, overlapping histograms can provide a means for the user to infer proportions relative to the larger context. As an example, Figure 1 shows the top 5 frequently used terms from the BioTechniques dataset. The left plot represents the number of documents for each term, where each term profile is scaled (0, 1). The right plots show the same terms, but they are scaled to all of the documents. The top profile of each plot shows the temporal profile of all the documents scaled (0, 1).

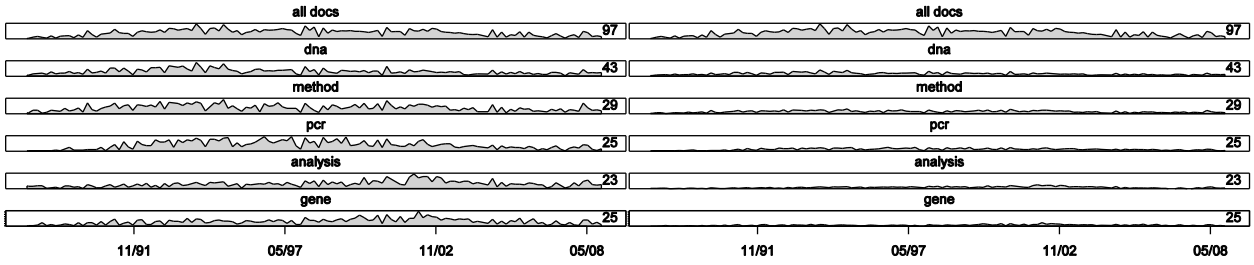

Figure 1 Histogram of document frequency for all documents (top profile) and the top 5 occurring terms normalized by maximum profile frequency (left plot) and normalized by maximum of all document profile (right plot), for the BioTechniques dataset.

[2] IN-SPIRE™ Visual Document Analysis, http://in-spire.pnl.gov

Term Document Co-occurrences

Term document co-occurrences reflect the number of documents in which two terms co-occur. Terms that co-occur within documents are considered to be associated. For a set of documents, this association is characterized as strong if the terms frequently co-occur. If the set of documents is filtered to only include documents containing a defined root term, then term document occurrences are identical to term document co-occurrences with the root term.

Term Document Associations

Term document associations represent the association of a term with individual documents. These may be calculated as the count of term occurrences within a document or as a normalized score. For this evaluation, we calculated each term's document association as the count of its occurrences within the document.

Term Feature Associations

Within this study, a feature is a term of interest. Term feature associations represent the association of a term with other terms. This association is calculated as the count of the term's document co-occurrences with another term. A term's feature association vector represents its related terms and is similar to terms that are related to the same terms even if they occur in different documents.

Temporal Information

Each of the methods for evaluating term characteristics can be further applied to specific time intervals as well as across the entire time range. This enables finer-grained insight to terminology dynamics as term associations will change over time.

4.0 Analyzing Term Transitions

Because transitions of term usage occur over time, we have investigated several methods for the temporal analysis of term usage statistics. Regardless of the method selected, many analyses will focus on a specific set of documents within which to investigate transitions within a larger dataset. This document context may be obtained by a specified root term, search criteria, selected theme, or documents from a particular source. Each of the methods described for measuring term associations will effectively count term-occurrence statistics solely from documents within the current document context. A user is therefore able to define a document context and identify terms whose usage changes within those documents over time. When the document context is defined by a specific root term, such as ***pcr***, the change over time of an individual term's occurrences within the document context reflects increasing or decreasing usage of the term with the root term.

Each of the methods is applied to the known transition from ***polymerase chain reaction*** to ***pcr*** as this is a known example upon which we can more effectively evaluate and compare these methods.

Temporal Analysis of Term Document Occurrences

Surprise and Emergence

To find candidate morphing terms, our *Surprise* algorithms for detecting trends can be used [Engel 2010 and Engel 2009]. In these algorithms, the first step is to bin each term's document occurrences into time intervals. Then a comparison of the document occurrence counts in two adjacent time windows for a specific term is performed in order to identify a significant change in document counts between the two windows. The *Emergence* statistic for decreasing trends can be used to order the most significant candidate terms by using the following hypothesis test:

$$H_o : \frac{1}{nc} \sum_{j=i}^{i+nc} x_j = \frac{1}{nc} \sum_{j=i-np}^{i-1} x_j$$

$$H_a : \frac{1}{nc} \sum_{j=i}^{i+nc} x_j < \frac{1}{nc} \sum_{j=i-np}^{i-1} x_j$$

The symbol *np* represents the number of time bins in the first time window. The symbol *nc* represents the number of time bins in the second time window. A Gaussian algorithm can then be used to compare counts in the two adjacent windows, as shown in Equation (1).

$$G = \frac{\frac{1}{nc} \sum_{j=i}^{i+nc} x_j - \frac{1}{np} \sum_{j=i-np}^{i-1} x_j}{\sqrt{\frac{s_i}{nc} + \frac{s_j}{np}}}$$

The symbol s_j represents the standard deviation of counts in the first time window and the symbol s_i represents the standard deviation of counts in the second time window.

To illustrate this morphing phenomenon, an example using the BioTechniques dataset is shown in Figures 2 and 3. Figure 2 shows how the documents are distributed over time, while normalized temporal profiles for multiple terms are shown in Figure 3. The profiles represent the number of documents at each time interval that contains the specific term. The number on the right side of the profile represents the maximum number of documents within the profile containing the specific term.

The purple circle identifies the time associated with the greatest decreasing emergence score, which may indicate a possible morphing candidate. The red circles identify the time associated with the greatest increasing emergence score occurring close to the time indicated by the purple circle, which may identify possible transition terms. In fact, the results shown in Figure 3 do actually indicate that the usage of the term ***polymerase chain reaction*** is replaced by the abbreviation ***pcr***.

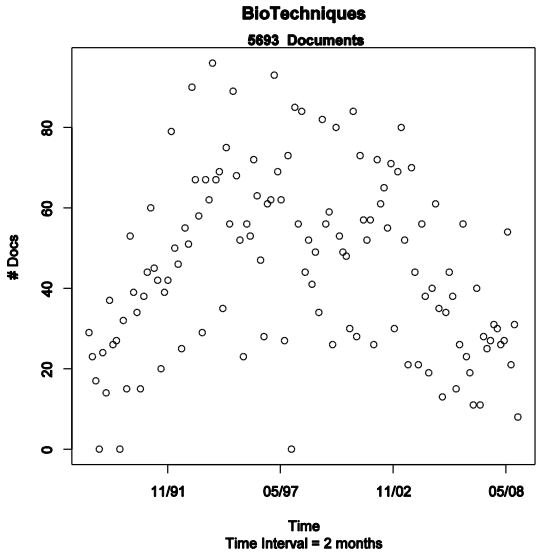

Figure 2 Temporal distribution of documents within the BioTechniques dataset, using two month intervals.

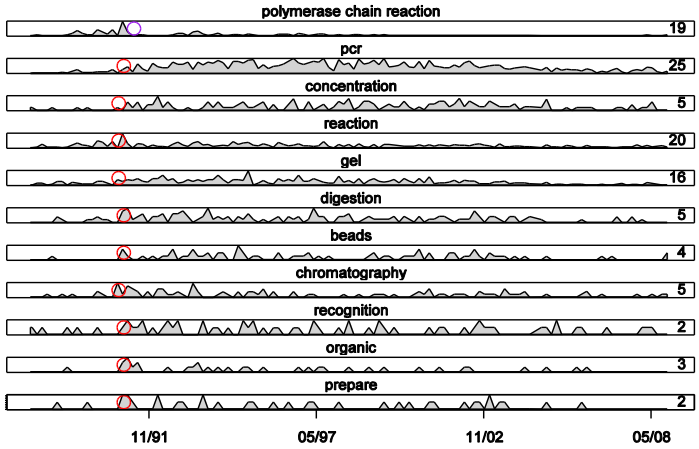

Figure 3 Temporal profiles of document frequencies for the term ***polymerase chain reaction***, along with terms with increasing occurrences close to the same time (BioTechniques dataset).

While temporal analysis of term occurrences can identify temporally related events, additional insight may be gained into surprising combinations of terms by applying a similar analysis to term co-occurrence information.

Temporal Analysis of Term Document Co-occurrences

Evaluating the change over time of term-pair co-occurrences yields insight to whether terms are being used differently. Because it is computationally intensive to calculate all term-document co-occurrences, in practice this is reduced to calculating term document co-occurrences in relation to a particular context, which may be set as a root term of interest. In this case, all co-occurrences with the root term of interest are accumulated.

Surprise and Emergence

The *Surprise* and *Emergence* algorithms can be applied to term-pair co-occurrences in order to identify term transitions. Surprising term co-occurrences may aid identification of new combinations. Emergence of term co-occurrences may aid identification of changing terminology, vocabulary specialization, or novel technology combinations that persist.

We expect that the phenomenon illustrated in Figure 3 occurs much more frequently, albeit in less obvious cases, and to be an indicator of terminology specialization and diversification, which can be an indicator of new technology applications. To help strengthen the "signal" for these transition points, we have added algorithm heuristics for measuring the co-occurrence of terms within documents. The term co-occurrence score can be calculated during times where a temporal nexus exists among terms that are emerging as shown by red circles in Figure 3, while other terms are declining as represented by the purple

circle in the figure. The co-occurrence heuristic will help validate the transition points between related terms and help filter out purely coincidental term co-occurrences.

An example of the morphing phenomenon as identified using term-pair co-occurrences is shown in Figure 4 and Table 1. In this figure, the top profile for each plot shows the number of documents within each time bin that contains the term ***polymerase chain reaction,*** scaled (0,1). The purple circle identifies the time when the decreasing emergence score is maximum. In the right plot, the remaining profiles show the number of documents containing the specific term (e.g., ***chain***, ***pcr***, ***product***), scaled (0,1). The maximum number of documents within a time interval for each term is shown on the right side of each profile. In the left plot, the profiles below the top profile show the number of documents that contain the term ***polymerase chain reaction*** and the specific term (i.e., term-pair co-occurrence), scaled (0,1). The number on the right side of each of these profiles represents the maximum term-pair co-occurrences within a time interval. The order of the term profiles plotted below the ***polymerase chain reaction*** profile is determined based on the proportion of the maximum term-pair co-occurrence (identified by the red circle) to the total number of documents containing the specific term within the same time interval (as shown in Table 1).

From Figure 4, we can see that the number of documents containing the ***polymerase chain reaction*** term in the BioTechniques dataset decreases rapidly in 1991, while other terms (e.g., ***pcr, amplification***) see increasing document occurrences. The goal of our research is to automatically identify those transitioning terms and show the results to a subject matter expert (SME). From the results shown in Figure 4 and Table 1, an SME could easily identify the transition (morphing) of the term ***polymerase chain reaction*** to ***pcr***.

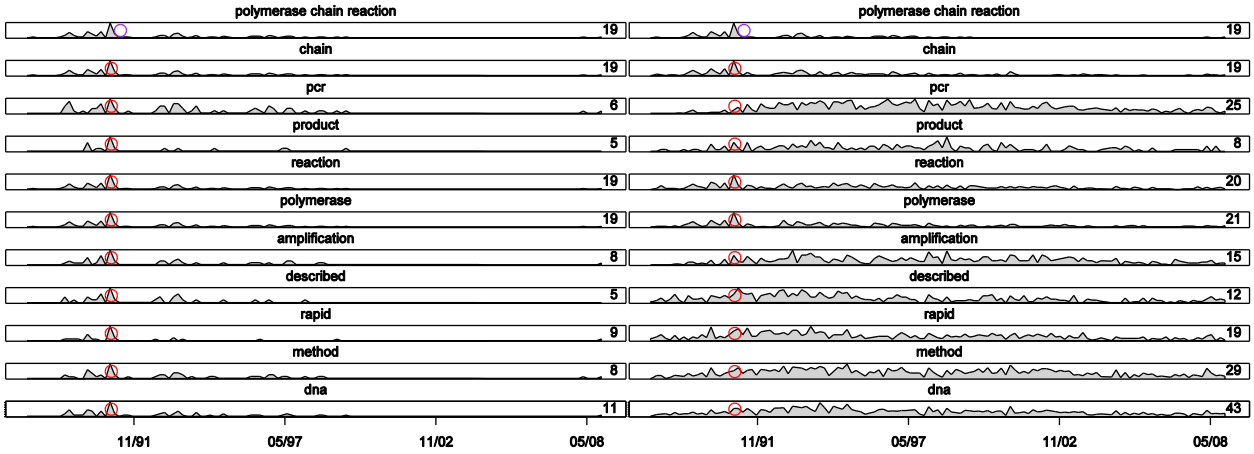

Figure 4 Temporal profiles of term-pair co-occurrences (left plot) and document frequencies (right plot) for the term ***polymerase chain reaction*** and related terms (BioTechniques dataset).

Table 1 Term-pair co-occurrences of ***polymerase chain reaction*** and related terms in late 1991.

Term	Max. Term-Pair Co-occurrences	Number Term Documents	Proportion
chain	19	19	1.00
pcr	6	6	1.00
product	5	5	1.00
reaction	19	20	0.95
polymerase	19	21	0.91
amplification	8	10	0.80
described	5	7	0.71
rapid	9	13	0.69
method	8	16	0.50
dna	11	24	0.46

Ranking Terms by Variance Over Time

Ranking terms by the variance of their document co-occurrences with a root term is an effective means of identifying those terms whose usage with the root term changes significantly over time. Within a defined document context, variance of term document co-occurrences signals change in that term's association with the document context. In the table below, document co-occurrence counts of individual terms with the root term ***polymerase chain reaction*** are shown. Terms are sorted by the sum of their document co-occurrences with ***polymerase chain reaction***. Referring to Table 2, we can see that ***pcr*** co-occurs in 73 documents with ***polymerase chain reaction*** in the BioTechniques dataset.

Table 2 Terms sorted by sum of document co-occurrences with ***polymerase chain reaction***.

	count of document co-occurrences with "polymerase chain reaction" by year																				
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	sum
polymerase chain reaction	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	126
polymerase	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	126
reaction	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	126
pcr	0	10	10	8	5	14	4	2	8	5	3	2	0	0	0	0	0	0	0	0	2	73
dna	1	9	15	15	4	9	4	2	3	4	2	1	0	0	0	0	0	0	0	0	1	70
method	0	4	13	11	4	10	2	0	6	3	3	0	0	0	0	0	0	0	0	0	2	58
amplification	1	7	11	10	3	7	1	1	4	4	2	1	0	0	0	0	0	0	0	0	1	53
primers	0	6	7	7	3	7	2	1	5	2	0	1	0	0	0	0	0	0	0	0	0	41
procedure	0	5	9	5	1	6	3	1	3	1	0	1	0	0	0	0	0	0	0	0	0	35
products	0	3	7	4	0	5	4	0	5	3	1	2	0	0	0	0	0	0	0	0	0	34
analysis	0	9	4	5	2	5	1	0	5	1	2	0	0	0	0	0	0	0	0	0	0	34
polymerase chain reaction pcr	0	10	5	3	2	8	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	33
specific	0	4	7	7	2	4	2	2	2	0	0	1	0	0	0	0	0	0	0	0	2	33
sequence	0	6	6	6	3	2	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0	30
gene	0	5	6	3	2	4	0	0	6	2	0	0	0	0	0	0	0	0	0	0	1	29

Sorting by sum enables easy identification of terms that occur the most frequently within the given context, but it provides little information on terms whose usage with the root term changes over time. Listed in Table 3 below are terms sorted by the standard deviation of their document co-occurrences over

the years with the root term. While useful, the results are dominated by the temporal distribution of the root term, ***polymerase chain reaction***.

Table 3 Terms sorted by standard deviation of yearly document co-occurrences with ***polymerase chain reaction***.

	count of document co-occurrences with "polymerase chain reaction" by year																					
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
polymerase chain reaction	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	8.37
polymerase	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	8.37
reaction	2	16	28	25	5	16	5	4	10	6	4	3	0	0	0	0	0	0	0	0	2	8.37
dna	1	9	15	15	4	9	4	2	3	4	2	1	0	0	0	0	0	0	0	0	1	4.73
pcr	0	10	10	8	5	14	4	2	8	5	3	2	0	0	0	0	0	0	0	0	2	4.23
method	0	4	13	11	4	10	2	0	6	3	3	0	0	0	0	0	0	0	0	0	2	4.02
amplification	1	7	11	10	3	7	1	1	4	4	2	1	0	0	0	0	0	0	0	0	1	3.43
polymerase chain reaction pcr	0	10	5	3	2	8	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	2.79
primers	0	6	7	7	3	7	2	1	5	2	0	1	0	0	0	0	0	0	0	0	0	2.71
procedure	0	5	9	5	1	6	3	1	3	1	0	1	0	0	0	0	0	0	0	0	0	2.56
analysis	0	9	4	5	2	5	1	0	5	1	2	0	0	0	0	0	0	0	0	0	0	2.52
sequencing	0	6	6	6	0	5	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	2.24
specific	0	4	7	7	2	4	2	2	2	0	0	1	0	0	0	0	0	0	0	0	2	2.23
products	0	3	7	4	0	5	4	0	5	3	1	2	0	0	0	0	0	0	0	0	0	2.22
gene	0	5	6	3	2	4	0	0	6	2	0	0	0	0	0	0	0	0	0	0	1	2.13

The temporal variance of ***polymerase chain reaction*** can be factored out of each term's standard deviation by calculating the relative proportion of each term's yearly document co-occurrences with ***polymerase chain reaction***. Table 4 lists terms sorted by the standard deviation of their proportion of yearly document co-occurrences relative to the number of documents each year in which ***polymerase chain reaction*** occurs. We can see that ***rt***, ***rt pcr***, and ***reverse transcription*** (the term that the abbreviation ***rt*** refers to) occur with greatest frequency with ***polymerase chain reaction*** from 1994 to 1998, the years that RT-PCR first came into use. This relationship was not evident when sorting by sum or standard deviation of raw term document co-occurrences as shown in Tables 2 and 3.

Table 4 Terms sorted by standard deviation of yearly relative proportion with ***polymerase chain reaction***.

	relative proportion of doc co-occurrences with "polymerase chain reaction" by year																					
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
polymerase chain reaction	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.5
reaction	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.5
polymerase	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.5
pcr	0.33	0.65	0.38	0.35	1.00	0.88	0.83	0.60	0.82	0.86	0.80	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.41
method	0.33	0.29	0.48	0.46	0.83	0.65	0.50	0.20	0.64	0.57	0.80	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.34
dna	0.67	0.59	0.55	0.62	0.83	0.59	0.83	0.60	0.36	0.71	0.60	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.33
rt	0.33	0.06	0.07	0.04	0.17	0.06	0.67	0.40	0.73	0.71	0.80	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.3
rt pcr	0.33	0.06	0.07	0.04	0.17	0.06	0.67	0.40	0.73	0.57	0.80	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.29
reverse transcription	0.33	0.12	0.10	0.08	0.17	0.12	0.67	0.60	0.73	0.71	0.80	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.29
transcription	0.33	0.12	0.14	0.08	0.33	0.12	0.67	0.60	0.73	0.71	0.80	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.29
amplification	0.67	0.47	0.41	0.42	0.67	0.47	0.33	0.40	0.45	0.71	0.60	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.27
specific	0.33	0.29	0.28	0.31	0.50	0.29	0.50	0.60	0.27	0.14	0.20	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.27
products	0.33	0.24	0.28	0.19	0.17	0.35	0.83	0.20	0.55	0.57	0.40	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.26
detection	0.33	0.12	0.10	0.15	0.33	0.41	0.50	0.40	0.45	0.43	0.40	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.26
polymerase chain reaction pcr	0.33	0.65	0.21	0.15	0.50	0.53	0.17	0.20	0.18	0.29	0.40	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.24

Because Tables 2-4 were calculated with the document context set with the root term ***polymerase chain reaction***, only those documents containing that term factor into the rankings. Notably, ***polymerase chain reaction*** did not occur in any documents for the years 2000 – 2007, making it difficult to assess whether related terms such as ***reverse transcription*** or ***rt pcr*** continued to be used or disappeared as well. We can get a better picture of terminology transitions by setting the context to documents containing the more frequent term ***pcr.*** **Table 5** lists terms sorted by the standard deviation of their proportion of yearly document co-occurrences relative to the number of documents each year in which ***pcr*** occurs.

Table 5 Terms sorted by standard deviation of yearly relative proportion with ***pcr***.

	relative proportion of doc co-occurrences with "pcr" by year																					
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
polymerase chain reaction	0.00	0.79	0.73	0.19	0.07	0.15	0.04	0.03	0.08	0.06	0.04	0.03	0.01	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.10	0.22
pcr	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.22
polymerase	0.00	0.79	0.73	0.23	0.15	0.19	0.15	0.07	0.15	0.15	0.10	0.06	0.06	0.09	0.08	0.09	0.02	0.05	0.03	0.03	0.16	0.21
reaction	0.00	0.79	0.73	0.32	0.19	0.24	0.12	0.15	0.16	0.13	0.13	0.10	0.12	0.07	0.13	0.09	0.09	0.12	0.13	0.13	0.26	0.20
polymerase chain reaction pcr	0.00	0.79	0.40	0.09	0.03	0.09	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.06	0.18
dna	0.00	0.71	0.53	0.53	0.66	0.56	0.52	0.41	0.41	0.35	0.49	0.46	0.49	0.53	0.52	0.51	0.46	0.49	0.60	0.53	0.71	0.15
real time	0.00	0.07	0.07	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.03	0.06	0.07	0.11	0.18	0.20	0.35	0.34	0.33	0.30	0.35	0.13
time	0.00	0.14	0.20	0.11	0.01	0.08	0.06	0.04	0.05	0.07	0.09	0.10	0.16	0.16	0.29	0.24	0.37	0.34	0.33	0.33	0.45	0.13
gene	0.00	0.50	0.27	0.17	0.27	0.22	0.17	0.19	0.18	0.17	0.23	0.23	0.29	0.29	0.44	0.40	0.44	0.34	0.50	0.33	0.23	0.13
amplification	0.00	0.50	0.53	0.47	0.35	0.47	0.26	0.25	0.28	0.20	0.32	0.26	0.33	0.29	0.32	0.44	0.22	0.20	0.37	0.37	0.32	0.12
method	0.00	0.43	0.40	0.49	0.52	0.37	0.33	0.30	0.33	0.31	0.38	0.40	0.40	0.43	0.40	0.49	0.35	0.46	0.50	0.43	0.65	0.12
sequencing	0.00	0.57	0.33	0.21	0.19	0.18	0.20	0.15	0.08	0.08	0.12	0.10	0.09	0.09	0.10	0.09	0.09	0.15	0.17	0.23	0.19	0.12
analysis	0.00	0.50	0.20	0.26	0.23	0.22	0.19	0.21	0.23	0.19	0.22	0.24	0.26	0.34	0.42	0.33	0.39	0.41	0.30	0.47	0.32	0.11
real time pcr	0.00	0.07	0.07	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.07	0.14	0.10	0.22	0.29	0.27	0.20	0.23	0.10
sequence	0.00	0.50	0.33	0.30	0.20	0.19	0.20	0.15	0.12	0.16	0.16	0.22	0.13	0.27	0.18	0.33	0.20	0.24	0.23	0.13	0.26	0.10

The term ***pcr*** is initially used in 1989. The term ***polymerase chain reaction*** occurs relatively frequently with ***pcr*** in 1989 and 1990 but then sharply drops off in the following years. Additionally, the terms ***real time***, ***time***, and ***real time pcr*** gradually increase in co-occurrence with ***pcr*** after 2000, as instruments were employed to replace the manual detection assay with quantitative monitoring of each step in the 25- to 40-step automated amplification reaction, enabling simultaneous detection and calculation of the number of starting molecules. Also, in the context of co-occurrence analysis, analysis of the term ***pcr*** enables identification of later morphing terms (***real time pcr***) that the initial term ***polymerase chain reaction*** did not identify.

Ranking terms of interest by the variance of their temporal document occurrences for a given context can help users identify when and to what degree term usage changes within the context. The effectiveness of the ranking is dependent on the validity and accuracy of the measurement of variance. Standard deviation in this case may be affected by the sizes of the time intervals as well as their boundaries. While Tables 4 and 5 show that calculating variance across yearly time intervals can be applied to journal publications, such as *BioTechniques*, a measure of variance that spans the entire time range without defined time intervals may be more broadly applicable to a wider range of datasets.

5.0 Analysis of Term Associations

A particularly interesting challenge for detecting transitions from the use of one term to another term is found when those terms co-occur in only a few, or zero, documents. If we consider that a rapid transition such as the replacement of one term with another, or the combination of previously unrelated terms, may be reflected within very few documents, then methods that depend solely on document associations may miss these transitions. Evaluating term feature associations provides complementary information to term document associations. A term's feature associations comprise the complete set of the term's association with individual features, which are calculated as the sum of the term's occurrences within documents that contain the feature (typically another term). Terms have similar feature associations when they predominantly co-occur with the same terms and can therefore be expected to occur within the same documents. Alternatively, terms that have dissimilar feature associations do not predominantly co-occur with the same terms and can therefore be expected to *not* occur within the same documents. Identifying terms whose usage is unexpectedly increasing or decreasing with a root term is a matter of calculating the difference of the similarity between feature associations for each term and the root term with the similarity between document associations for each term and the root term.

Figure 5 presents a conceptual model and topic of future work for plotting the similarity of a term's document associations and feature associations to those of a given root term. Terms that are similarly associated with the same features as the root term can be expected to co-occur within the same documents (upper right in the plot) as the root term. Alternatively, if terms have dissimilar feature associations, then they can be expected to not co-occur within the same documents (lower left in the plot) as the root term. A term that has dissimilar feature and document associations with the root term is of particular interest and will be plotted off of the main diagonal. Terms that have similar features to the root term but occur in different documents will appear in the red cells in the upper left area of the plot, indicating that they are related by other terms but not necessarily by context of usage. Terms that have dissimilar features to the root term but occur in the same documents will appear in the blue cells in the lower right area of the plot, indicating an unusual or new combination of predominantly unrelated terms.

We have reduced to practice the comparison of terms' feature associations with their document associations over time in order to identify trends as well as unexpected replacements and combinations of terms within distinct time intervals. We discuss several results of applying this analytic approach in the following section.

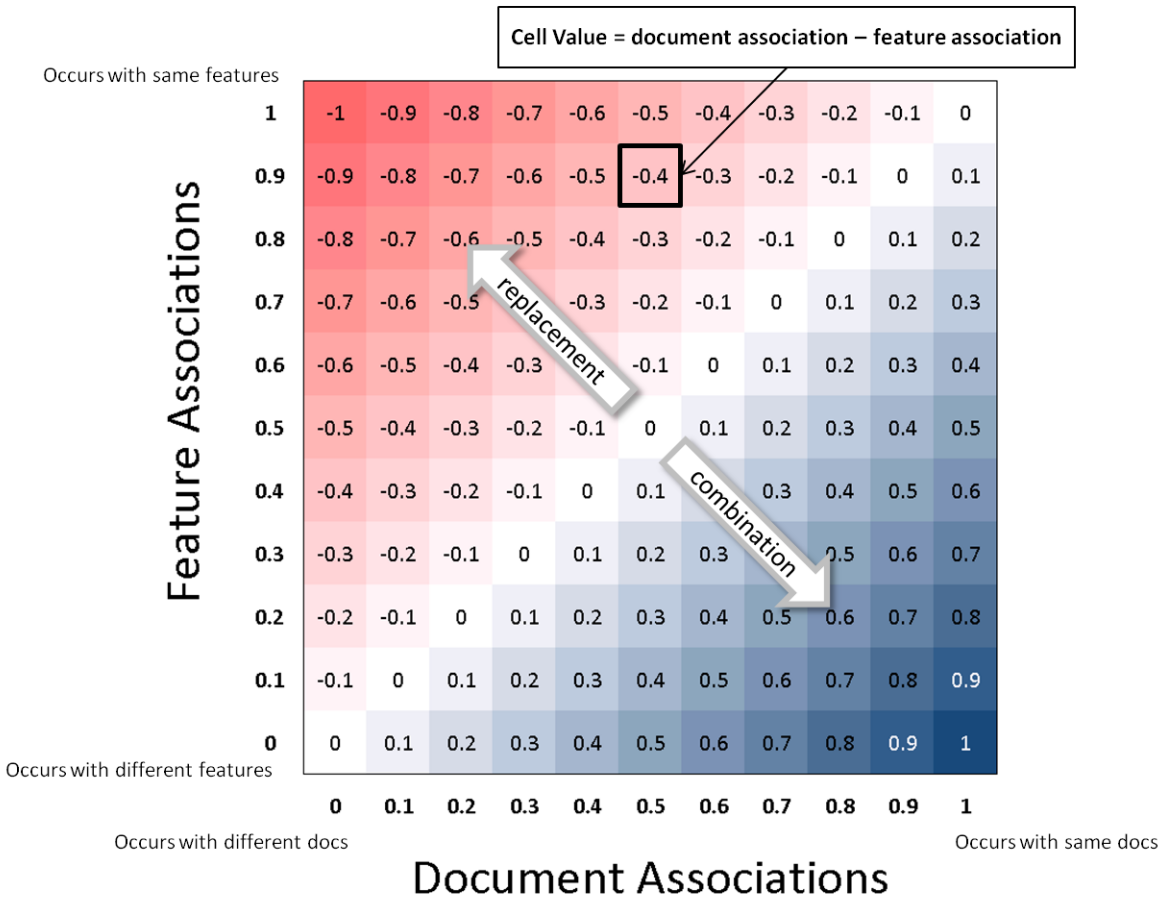

Figure 5 The dimensions of feature association and document association, with representative cells color encoded by the difference of document association and feature association at their respective location in the context of a defined root term.

6.0 Temporal Analysis of Term Associations

Because global statistics across the entire time range can wipe out indications of temporally focused transitions, we explored methods for quantifying how dissimilar a term's document associations are from its feature associations within time intervals. Doing so should enable a user to identify points in time or broad trends where these two associations differ, indicating important transitions in terminology.

Each term's feature associations and document associations are calculated within distinct time intervals and their similarity, as the Czekanowski coefficient, with the root term's associations is calculated. This provides for each year a similarity measure of each term's document associations to that of the defined root term, in this case ***polymerase chain reaction*** (shown in **Table 6**) or ***pcr***, as well as a similarity measure of each term's feature association to that of the defined root term (shown in **Table 7** for ***polymerase chain reaction***).

Table 6 Similarity of terms' document associations by year to the document associations for the term ***polymerase chain reaction***.

czekanowski coefficient of document associations by year in context of "polymerase chain reaction"																						
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
polymerase chain reaction	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5
reaction	0.3	0.8	0.9	0.8	0.3	0.6	0.2	0.2	0.4	0.4	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3
polymerase	0.5	0.7	0.9	0.8	0.3	0.7	0.2	0.2	0.4	0.3	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.3
polymerase chain reaction pcr	0.0	0.7	0.2	0.2	0.6	0.6	0.0	0.0	0.2	0.3	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.3
reverse transcription polymeras	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.4	0.8	0.5	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
amplification	0.7	0.4	0.4	0.2	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2
reverse transcriptase	0.0	0.1	0.1	0.1	0.0	0.0	0.3	0.2	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.2
reverse transcription	0.0	0.1	0.1	0.1	0.0	0.1	0.3	0.3	0.5	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
dna amplification	0.7	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.2	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
cdna synthesis	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
pcr method	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1
polymerase chain reaction produ	0.0	0.1	0.2	0.2	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
reaction products	0.0	0.1	0.2	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.3	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
pcr	0.0	0.5	0.3	0.2	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
pcr technique	0.0	0.1	0.1	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1

Table 6 shows the similarity of terms' yearly document associations to the ***polymerase chain reaction***'s document associations. As expected, the terms ***reaction*** and ***polymerase***, words that are part of ***polymerase chain reaction***, retain high coefficients for a period after 1988 (1989 - 1991); even though these terms occur in contexts other than ***polymerase chain reaction***, they primarily occur with ***polymerase chain reaction***. These coefficients decrease after 1991 to 1993 as ***reaction*** and ***polymerase*** are increasingly used in other contexts. Looking at terms that are not part of ***polymerase chain reaction***, the document associations for ***amplification*** and ***dna amplification*** are very similar to those for ***polymerase chain reaction*** in 1988 but not in the following years, indicating that they were frequently used in the same documents as ***polymerase chain reaction*** in 1988 but infrequently from 1989 to 2008. This is expected because the first DNA amplification method was PCR, and after 1988 ***amplification*** and ***dna amplification*** would be expected to have high coefficients with ***pcr*** (what the language morphed to) instead of ***polymerase chain reaction*** (what the language morphed away from). A third point is the appearance of moderate coefficients for the term ***reverse transcription polymerase chain reaction rt pcr*** (truncated to ***reverse transcription polymerase*** in Table 6 and Table 8) in 1994 – 1998, and ***cdna synthesis*** in 1995. This represents the period when PCR technology was first widely applied for analyzing RNA molecules, by coupling (a) the use of the enzyme reverse transcriptase to produce cDNA from RNA with (b) the DNA polymerase used to amplify DNA. This is also why there is a high coefficient for ***cdna synthesis*** in 1995. These examples demonstrate that statistical analysis of term usage can be applied to automatically identify temporally related scientific developments described in journal publications.

Table 7 shows the similarity of terms' yearly feature associations to ***polymerase chain reaction***'s feature associations. We can see that the feature associations for ***pcr*** have the strongest overall similarity with the feature associations for the context, ***polymerase chain reaction***. Referring back to **Table 6**, it is worth noting that ***pcr*** only has similar document associations with ***polymerase chain reaction*** in 1989, indicating that these two terms were not used often together in documents in the following years.

Table 7 Similarity of terms' feature associations by year to the feature associations for the term ***polymerase chain reaction***.

czekanowski coefficient of feature associations by year in context of "polymerase chain reaction"																						
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
pcr	0.0	0.2	0.2	0.5	0.6	0.7	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.4	0.4	0.4	0.4	0.2
rt	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.4	0.2	0.4	0.3	0.2	0.1	0.3	0.1	0.2	0.1	0.1	0.1	0.1	0.1
amplification	0.0	0.1	0.2	0.3	0.3	0.5	0.3	0.3	0.3	0.3	0.4	0.3	0.4	0.4	0.5	0.4	0.3	0.2	0.2	0.2	0.2	0.1
products	0.0	0.1	0.2	0.2	0.2	0.4	0.4	0.3	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1	0.1
sequencing	0.1	0.3	0.3	0.4	0.4	0.4	0.5	0.4	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.3	0.0	0.1	0.1	0.1	0.2	0.1
polymerase	0.0	0.3	0.4	0.3	0.3	0.3	0.3	0.2	0.3	0.2	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.0	0.0	0.1	0.1
rt pcr	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.4	0.2	0.3	0.3	0.2	0.1	0.3	0.1	0.2	0.1	0.1	0.1	0.0	0.1
rna	0.0	0.1	0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.3	0.4	0.3	0.3	0.3	0.4	0.3	0.4	0.2	0.2	0.2	0.2	0.1
detection	0.0	0.1	0.2	0.3	0.3	0.3	0.4	0.3	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.2	0.1	0.3	0.1
primers	0.0	0.1	0.2	0.2	0.3	0.4	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.1
reaction	0.1	0.2	0.4	0.4	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
polymerase chain reaction	0.0	0.2	0.4	0.3	0.1	0.2	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
genes	0.0	0.1	0.1	0.1	0.2	0.2	0.4	0.2	0.2	0.2	0.3	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.2	0.2	0.1	0.1
time	0.0	0.1	0.2	0.2	0.1	0.2	0.2	0.1	0.1	0.1	0.3	0.2	0.3	0.2	0.3	0.3	0.4	0.2	0.2	0.2	0.3	0.1
primer	0.0	0.1	0.1	0.1	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.2	0.1	0.2	0.1	0.1	0.1	0.1

Calculating the difference between terms' document and feature similarity with ***polymerase chain reaction***, shown in **Table 8** enables identification of terms that have changed the most in their associations with ***polymerase chain reaction***. A blue cell indicates a year in which a term's document associations with ***polymerase chain reaction*** are more similar than its feature associations with ***polymerase chain reaction***, indicating that the term is co-occurring with ***polymerase chain reaction*** in more documents than expected. A red cell indicates a year in which a term's feature associations are more similar than its document associations with ***polymerase chain reaction***, indicating that the term is co-occurring with ***polymerase chain reaction*** in fewer documents than expected. We can see that after 1990, ***pcr*** occurs in fewer documents with ***polymerase chain reaction*** than would be expected given the feature associations of the two terms, corresponding with the replacement vector in Figure 5. We can also see notable spikes in document association for ***dna amplification*** in 1988 and for ***cdna synthesis*** in 1995, corresponding with the combination vector in Figure 5, which are not evident in **Table 2**.

Table 8 Difference of similarities (**Table 6**-**Table 7**) for document associations and feature associations by year to the term ***polymerase chain reaction***.

difference of czekanowski coefficients (doc - feature) in context of "polymerase chain reaction"																						
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
polymerase chain reaction	1.0	0.8	0.7	0.7	0.9	0.8	0.9	1.0	0.9	0.9	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5
amplification	0.7	0.3	0.1	-0.1	-0.2	-0.4	-0.3	-0.2	-0.2	-0.2	-0.4	-0.3	-0.4	-0.4	-0.5	-0.4	-0.3	-0.2	-0.2	-0.2	-0.1	0.3
pcr	0.0	0.3	0.1	-0.3	-0.6	-0.5	-0.6	-0.6	-0.6	-0.6	-0.6	-0.6	-0.6	-0.6	-0.6	-0.5	-0.5	-0.4	-0.4	-0.4	-0.3	0.2
polymerase chain reaction pcr	0.0	0.5	0.1	0.1	0.5	0.6	0.0	0.0	0.2	0.3	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.2
polymerase	0.5	0.5	0.5	0.5	0.0	0.4	-0.1	0.0	0.1	0.1	0.0	0.3	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1	0.0	0.0	0.4	0.2
reaction	0.2	0.6	0.5	0.5	0.1	0.3	0.0	-0.1	0.1	0.3	0.0	0.1	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	0.2	0.2
reverse transcription polymeras	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.4	0.7	0.5	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
reverse transcriptase	0.0	0.1	0.0	0.0	-0.1	0.0	0.3	0.2	0.1	0.0	0.1	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.7	0.2
dna amplification	0.7	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
cdna synthesis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.2	0.0	0.0	-0.1	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.1
primers	0.0	0.3	0.1	0.1	-0.2	-0.2	-0.2	-0.2	-0.1	-0.2	-0.3	-0.2	-0.3	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1	0.0	-0.1	0.1
products	0.0	0.2	0.1	0.0	-0.2	-0.2	-0.3	-0.3	-0.2	-0.2	-0.4	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1	0.1
rt	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.2	-0.2	0.0	-0.4	-0.3	-0.2	-0.1	-0.3	-0.1	-0.2	-0.1	-0.1	-0.1	0.2	0.1
cdna	-0.1	0.1	0.0	0.0	-0.1	-0.2	-0.3	0.1	-0.2	-0.2	-0.2	-0.3	-0.2	-0.4	-0.3	-0.3	-0.3	-0.2	-0.2	-0.1	-0.1	0.1
analysis	-0.2	0.1	-0.1	-0.1	-0.2	-0.3	-0.4	-0.3	-0.3	-0.3	-0.4	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.3	-0.2	-0.2	-0.3	0.1

Referring to **Table 9** in which the document context is set to those documents in which ***pcr*** occurs, we can see that after 1990, ***polymerase chain reaction*** occurs in fewer documents with ***pcr*** than would be expected given the feature associations of the two terms. And as we saw in **Table 5**, **Table 9** also clearly shows that the terms ***real time***, and ***real time pcr*** gradually increase in co-occurrence with ***pcr*** after 2000, indicating the rise of an automated and quantitative technology based on PCR. It can also be seen that the variance among term associations is a lot less in Table 9 (***pcr***) than in Table 8 (***polymerase chain reaction***). This is because Table 8 uses the term for the original disruptive technology that was quickly replaced to a large degree, whereas Table 9 uses a term that is still very much in use.

Table 9 Difference of similarities for document associations and feature associations by year to the term ***pcr***.

difference of czekanowski coefficients (doc - feature) in context of "pcr"																						
term	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	std. dev.
pcr	0.0	1.0	1.0	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.2
real time	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.4	0.3	0.3	0.3	0.1
polymerase chain reaction pcr	0.0	0.5	0.3	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
time	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.4	0.3	0.3	0.3	0.1
polymerase chain reaction	0.0	0.5	0.3	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
real time pcr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.2	0.1	0.3	0.1
rt	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.2	0.3	0.2	0.2	0.3	0.2	0.2	0.3	0.2	0.3	0.2	0.1	0.3	0.1	0.1
reaction	0.0	0.5	0.3	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.2	0.1
rt pcr	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.2	0.3	0.2	0.2	0.3	0.2	0.2	0.3	0.2	0.3	0.2	0.1	0.3	0.1	0.1
dna	-0.1	0.0	-0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.0	0.1	0.1	0.2	0.2	0.2	0.1	0.1
polymerase	0.0	0.4	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1
method	0.0	0.1	0.0	0.1	0.3	0.1	0.1	0.1	0.2	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.3	0.2	0.2	0.3	0.1
genes	0.0	0.1	0.3	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.3	0.1	0.2	0.2	0.0	0.1
amplification	0.0	0.3	0.4	0.4	0.3	0.3	0.2	0.2	0.3	0.2	0.3	0.3	0.3	0.2	0.2	0.3	0.1	0.2	0.3	0.2	0.2	0.1
pcr amplification	0.0	0.2	0.3	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.0	0.1	0.1	0.1	0.1

7.0 Conclusion

We have shown several methods that support the identification of morphing terminology and terminology transitions over time that aid users in identifying new technologies and their precursors, as well as new applications and combinations of science and technology. We expect that users will be interested in exploring terminology transitions in relation to a focused context, defined through a set of documents, search terms, or computed themes. Temporal analysis of terms' document occurrences, co-occurrences, and associations within the user-defined context will reveal terms that are transitioning into or out of that context. The surprise and emergence algorithms can be applied to term document occurrences and term document co-occurrences to identify terms whose statistically significant changes are temporally aligned, suggesting that one term is replacing or morphing into another term.

The temporal variance of terms' proportionate counts of occurrences, co-occurrences or associations relative to a defined context provides a useful ranking metric to identify those terms that have transitioned the most in that context. The comparison of terms' feature associations and document associations provides a means to evaluate whether terms are being used as expected, or whether they are dropping out of (or into) a given context unexpectedly. This method also enables identification of terms that should appear in documents together based on their feature associations, regardless of whether they do in fact co-occur within documents.

8.0 References

Bookstein A, ST Klein, and T Raita. 1998. "Clumping Properties of Content-Bearing Words." *J. Am. Soc. Inf. Sci*. 49(2): 102-114. DOI= http://dx.doi.org/10.1002/(SICI)1097-4571(1998)49:2<102::AID-ASI2>3.0.CO;2-2

Engel, DW, PD Whitney, AJ Calapristi, and FJ Brockman. 2009. "Mining for Emerging Technologies within Text Streams and Documents." In *Proceedings for the Society of Industrial and Applied Mathematics (SIAM) International Conference on Data Mining (SDM09)*, Reno, NV, 2009.

Engel, DW, PD Whitney and NO Cramer. 2010. "Events and Trends in Text Stream." In *Text Mining: Application and Theory*, eds. M Berry and J Kogan. John Wiley & Sons, Chichester, United Kingdom.

Rose, SJ, DW Engel, NO Cramer, and WE Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." In *Text Mining: Application and Theory*, eds. M Berry and J Kogan, John Wiley & Sons, Chichester, United Kingdom.

Appendix A

200 Terms of Interest in BioTechniques dataset, as extracted by RAKE.

dna	cat activity	laser capture microdissection
site directed mutagenesis	sscp	hsv
polymerase chain reaction	strand cdna synthesis	reverse transcription rt pcr
green fluorescent protein	t7 dna polymerase	cell lines
real time pcr	dna methylation	reporter gene
gfp	automated dna sequencer	sequencing
dna sequence	denaturing gradient gel electrophoresis	sequences
pcr	methylation status	protein expression
single nucleotide polymorphism snp	dna sequence analysis	gel electrophoresis
agarose gel electrophoresis	gene expression analysis	sequencing reactions
gram negative bacteria	dna yields	real time quantitative pcr
taq dna polymerase	molecular weight dna	adherent cells
reverse transcription pcr rt pcr	pulsed field gel electrophoresis	purified dna
cell growth	quantitative real time pcr	apoptotic cells
dna fragments	endothelial cells	differential display dd
genomic dna	reverse transcription pcr	quantitative pcr
cdna library	ribonuclease protection assay	pcr technique
automated dna sequencing	quantitative rt pcr	peripheral blood cells
differentially expressed genes	virus	phage particles
performance liquid chromatography	peptide	extracted dna
dna polymerase	recombinant protein	polymerase chain reaction products
cells	gene	cloning strategy
polymerase chain reaction pcr	enzyme linked immunosorbent assay elisa	positively charged nylon membrane
dna sequencing	gel	flag peptide
protein	dna sequences	cloning vectors
rna degradation	dna polymerases	allele specific primers
rt	dna fragment	protein dna interaction
paraffin embedded tissues	oligonucleotide probes	fluorescent detection
fusion protein	cell	protein kinases
green fluorescent protein gfp	template dna	fluorescent proteins
rna	pcr amplification	qpcr
human genomic dna	dna binding proteins	quantitative reverse transcription pcr rt pcr
double stranded dna	beta gal	fret
plasmid dna	chromosomal dna	randomly amplified polymorphic dna rapd
dna templates	gene expression data	rapd
pcr products	aequorea victoria green fluorescent protein gfp	rapid cycle pcr
gene expression	scfv	contaminating genomic dna
rt pcr	single base mutations	recombinant antibody fragments
protein protein interactions	single stranded conformational polymorphism sscp	recombinant dna technology
beta galactosidase beta gal	direct dna sequencing	regulated expression
cloned dna	transgene expression	rna binding proteins
reverse transcription polymerase chain reaction rt pcr	dna damage	rna preparations
denaturing gradient gel electrophoresis dgge	vaccinia virus	dd
target protein	vh	silver staining
dna binding	live cell imaging	human dna
methylation	allele specific pcr	slab gel electrophoresis
microarray data	digoxigenin labeled probes	sscp analysis
gene expression profiling	methylation analysis	hybridoma cell lines
enhanced green fluorescent protein egfp	electrophoretic mobility shift assay	study protein protein interactions
single stranded dna	cell surface antigens	insect cells
single nucleotide polymorphisms snps	double stranded dna fragments	dna binding activity
expression	paraffin embedded tissue sections	transposon
cell line	pcr amplified dna	tta
recombinant virus	multiplex pcr	laser capture microdissection lcm
pcr product	protein dna interactions	cell death
system	basal expression	baculovirus expression system
phage	agarose gels	mab
total rna	probes	marker gene
target dna	dna probes	matrigel
cycle sequencing	dna microarray	detection system
assay	mass spectrometry ms	cell monolayers
pcr primers	real time reverse transcription pcr rt pcr	double stranded dna sequencing
program	single strand conformation polymorphism sscp	enhanced green fluorescent protein
expression level	transfected cells	microarray experiments
quality rna	method	filamentous fungi
gene expression patterns	dna template	ligation mediated pcr
gfp fluorescence	sodium dodecyl sulfate polyacrylamide gel electrophoresis sds page	

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

Proudly Operated by Battelle *Since 1965*

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnl.gov