Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents[1]

Olga Babko-Malaya, Andy Seidel, Daniel Hunter, Jason HandUber, Michelle Torrelli and Fotis Barlos

{olga.babko-malaya, andy.seidel, daniel.hunter, jason.handuber, michelle.torrelli, fotis.barlos}@baesystems.com
BAE Systems, Burlington, MA 01803

Abstract
This paper describes a multidisciplinary study and development effort to analyze full text and metadata of scientific articles and patents for indicators of new disruptive and game-changing technical breakthroughs. The system we are developing can scan millions of documents in two languages, English and Chinese, and extract meaningful trends and predictions. Whereas traditional approaches to innovation analytics rely on citation analysis to analyze impact or identify the most influential patents or researchers in the field, our system takes a step further and combines these methods with an analysis of text in order to identify and characterize emerging technologies. The paper describes the indicators and forecasting models, as well as presents the results of applying these indicators to forecast levels of interest in a particular technology based on the analysis of English and Chinese patents. It further shows how the indicators we developed can provide insights into the nature and the lifecycle of emerging technologies.

Conference Topic
Indicators

Introduction
This paper describes Abductive Reasoning Based on Indicators and Topics of EmeRgence, or ARBITER, an automated system whose purpose is to identify and characterize emerging technologies and emerging fields in science. It does so by processing very large collections of scientific publications and patents in multiple languages and identifies trends, associations, and predictions more rapidly than with current methods. Unlike previous approaches to detecting emergence, which are based on the citation analysis of papers and patents (e.g. Bettencourt et al., 2008; Shiebel et al., 2010; Roche et al., 2010), we are extracting information from the text of publications and patents, identifying authors, their affiliations, addresses, as well as classifying types of organizations and publications. Moreover, we apply natural language processing technologies to extract scientific terminology from the full text of the documents, to identify different types of relationships between citations, authors, terms, and organizations, including contrast, opinion, and related work, and to characterize maturity and other properties of terms based on their contextual patterns. This diverse set of features enables us to efficiently process multiple collections and various types of data without dependency on the presence of a specific feature in a collection. For example, our approach is not hampered by the lack of prior art references in Chinese patents, which is a problem for a standard, citation-based analysis of innovative technologies.

To define indicators of emergent technologies and scientific fields, we have developed a pragmatic theory of technoscientific emergence, described in Brock et al. (2012), which builds on Actant Network Theory (Latour, 2005). An Actant Network is a heterogeneous network of human and non-human elements, including people, institutions, funders, meetings, documents, and scientific terminology, interconnected by disparate relationships. The membership of elements within such a network, and the nature and extent of the relationships

[1] Approved for public release; unlimited distribution.

between these elements, is dynamic and constantly changing. To model emergence, we have developed indicators that measure the character and evolution of Actant Networks, including

- Extent of different types of elements in a network, including prolific and prominent entities
- Number of relationships and the volume of traffic in a network
- Growth of entities and relationships, including average growth rate and slope measures
- Novelty of elements and relationships
- Prevalence of the marketplace actant
- Extent of patenting activities
- Amount of disagreements and uncertainties.

In our previous work, we have shown how these indicators can be applied to characterize communities of practice (Babko-Malaya et al., 2013a), identify the presence of the debate in the community (Babko-Malaya et al., 2013b), as well as determine whether practical applications exist for research fields (Thomas et al., 2013). This paper presents the results of applying these indicators to forecast prominence of technology terms, as measured by a significant increase in term frequency. Whereas ARBITER processes both scientific articles and patents, the results presented in this paper are limited to the analysis of patents.

This paper contains three further sections. First, we give an overview of metadata and full text features, describe different categories of indicators designed to identify emerging technologies, as well as demonstrate how the indicators are combined via Bayesian networks into a forecasting model. The next section presents the results of the correlation analysis of indicators with future term prominence for English and Chinese patents, which measures the ability of our indicators to forecast a significant increase in term usage. The final section outlines how the system can be applied to characterize the nature and the lifecycle of the technology.

System Description

Feature Extraction

ARBITER extracts features from the metadata and full text of scientific papers and patents, including Lexis-Nexis Patent data, which includes granted patents and published patent applications from United States and Chinese national patent offices, and Thomson Reuters Web of ScienceTM (abstracts of journals and conference proceedings for the same time period, ~40M records). The features we extract from these sources include metadata features (such as title, author, author affiliation, patent assignees, etc.), as well as features that are based on the analysis of text. All feature extraction capabilities, including language features, are developed for two languages: English and Chinese. A summary of our features is shown in Figure 1. The entities we extract include people, organizations, documents, and scientific terminology, interconnected by different types of relationships.

To analyze persons, we extract authors from scientific articles and inventors from patents. In order to be able to count unique mentions of researchers, we developed a disambiguation component, which groups them into equivalence classes. Our analysis of researchers builds on features such as researcher impact, including Hirsch index and prolificness (measured by patent/paper productivity), as well as co-authorship and citation graphs.

To identify organizations, we extract author affiliations and patent assignees from metadata, as well as funding organizations from the text of acknowledgements and footnotes of scientific papers. All organizations are classified into three classes: Commercial, Academic, and Government/Nonprofit. The organization classification component allows us to evaluate

the extent and changes in the Academic vs. Commercial involvement in a certain field, as well as the diversity of researchers and organizations.

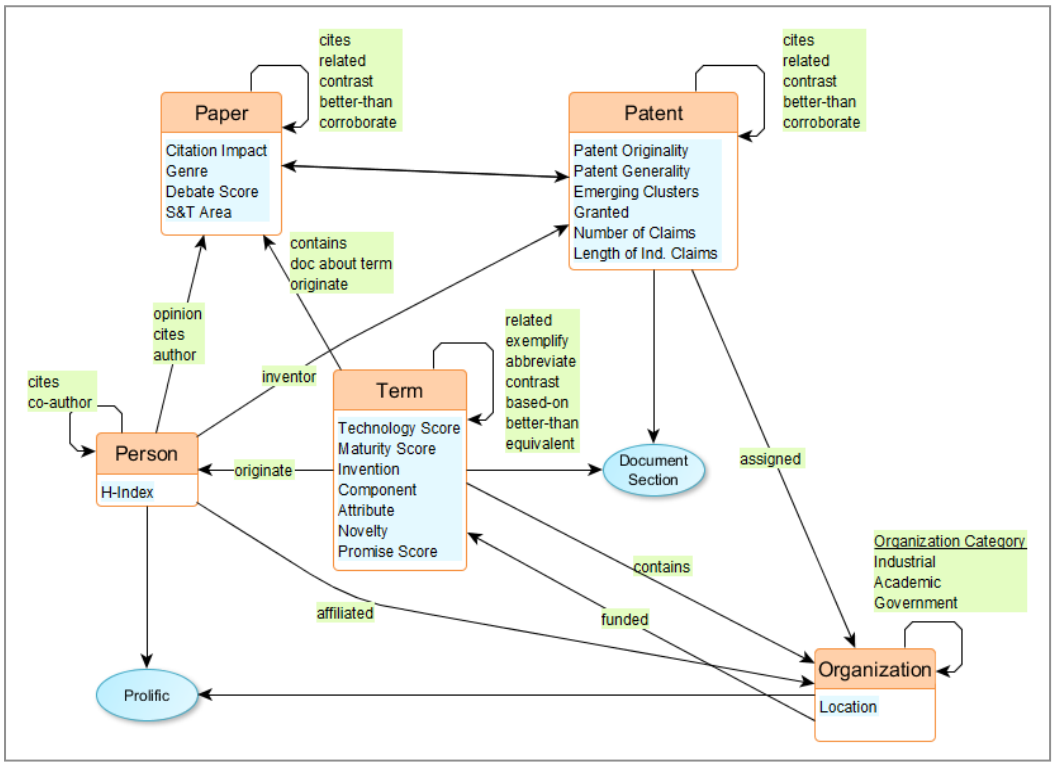

Figure 1. Actant Network extracted from metadata and text.

Our analysis of documents uses citation-based metrics developed by one of our team partners to measure generality, originality, and membership in "emerging clusters" (Breitzman & Thomas, 2015). We further measure mean citation impact of papers and patents, and analyze the structure and length of patent claims.

Our other partners have developed several modules for linguistic processing of text in English and Chinese. For example, to identify scientific terminology, we apply a technology described in Meyers et al. (2010) that extracts scientific noun phrases from the text of papers and patents. The extracted terms are noun phrases that tend to occur frequently in a set of articles from a specific field, but rarely occur in more general or popular articles.

In order to characterize these terms, we score terms based on the extent to which the term behaves like a technology (Anick et al., 2014), as well as assign a maturity score based on how often the term is mentioned in text as being used.

To analyse documents, we apply a genre classifier to evaluate the types of documents that are being published in a certain field, such as review articles or product reviews, as well as to classify documents based on the extent of the debate in the community (Babko-Malaya et al., 2013b). Using the document structure parser, we further identify different sections of documents and categorize claims in patents. To support Chinese extraction, we have adapted a tool to support word segmentation and part of speech tagging to scientific literature and patents (Li & Xue, 2014).

All entities we extract are linked by various types of relations. Whereas some relations are extracted from metadata (e.g. affiliated, invented, assigned, cites, co-author), many relations are extracted from text using information extraction techniques. These relations include opinion relations as well as relations like abbreviate, exemplify, and related work (based on,

better than, contrast, etc), which are described in more detail in Meyers (2013) and Meyers et al. (2014) and are illustrated below.

All entities and relations extracted from full text were evaluated against manually created gold standard corpora. Performance of extraction components is generally comparable across English and Chinese with the f-score above 70-75% in both languages.[2]

Indicators

Using this network, we have developed over 200 indicators that measure different characteristics and changes in the network associated with particular technologies and concepts. The indicators we developed are driven by our pragmatic theory, which defines emergence as the growth in the robustness of actant networks (Brock et al., 2012). The indicators we apply to identify potential disruptive technologies are therefore designed to analyze the relationships between the target entity and other elements in the actant network, including the extent and nature of these relationships, their novelty, dynamic changes, as well as impact, prominence and diversity. Other indicators we explore relate technology emergence to their practicality, as well as the presence of the debate in a community.[3]

Term Momentum Indicators. Our first set of indicators measures momentum in the usage of a particular term. These indicators are time series of annual counts, such as counts of term usage by inventors and organizations, with a further focus on prolific inventors and organizations. In addition, our 'section-based' indicators analyze term usage in independent claims, summary of invention, and abstract sections of patents. The rationale behind an analysis of term usage in specific sections is that these indicators can better measure the extent of the acceptance of the term by the community. For example, if a term occurs in independent claims of patents, it means that it has been legally accepted.

Term Characterization. Beyond indicators based on the momentum associated with individual terms, we also developed indicators that examine different characteristics of these terms. These characteristics include (1) the likelihood that the term describes a technology, (2) the maturity of the technology described by the term, (3) the degree to which the term functions as a description of an invention, and (4) the degree to which a term refers to a component of another technology.

Term characterization scores are calculated by collecting and aggregating evidence from the term's context. For example, to compute maturity scores, we define a set of 'usage' patterns, i.e. patterns that indicate that a term was used or applied: *We used [term] for …, [term] was used for …, employ [term], …* The maturity score is then derived from the number of times these 'usage' patterns are applied to the term. Likewise, the degree to which the term is used as a component is computed based on term usage in 'component'-specific contexts, as illustrated by the sentence *"A typical RFID tag consists of/contains an RFID antenna and RFID chip"*. The terms *RFID antenna* and *RFID chip* are tagged as components in this context, given that they occur as the objects of verbs *consist of* or *contains*. Our expectation is that a time series analysis of maturity of technologies, including their usage as an invention or a component, might be indicative of a change in the lifecycle of a technology, and therefore can be used to identify potentially disruptive technologies (Arthur, 2009).

Semantic Relations. Another class of language-based indicators is based on semantic relations we extract from text. These relations include Opinion, Abbreviate, Exemplify,

[2] Although performance is comparable, there is some variation in the frequency and the type of relations that we extract in the two languages. Some relations are very sparse in Chinese (such as Abbreviations, Contrast, Exemplify (Term1 is an example of Term2). Another difference is that text processing in Chinese is significantly slower than in English due to word segmentation.

[3] The indicators described in this section are focused on the analysis of patents. Similar indicators have also been developed for the analysis of scientific articles, but their analysis is beyond the scope of this paper.

Originate, and different types of Related Work, including Contrast, Based On, and Better Than (Meyers et, 2014). For example, Practical relations represent the author's view that the technology is either being used specially or is useful in some way. Therefore, the indicator that measures the number of Practical relations attached to a term may identify an increase in interest to using a given technology, or its new application. Meanwhile, the relation Abbreviate, which links scientific terms to their abbreviations, can be used to detect the timeline of the acceptance of the term by the community. Finally, relations like Contrast may help to identify the early stages of technology development, given that scientists developing innovative concepts tend to contrast their work with existing research, whereas as the technology becomes more accepted, the number of contrast relations declines.

Document and Inventor Characteristic indicators. This class of indicators measures characteristics of the papers or patents that are using the term. Some of these indicators measure citations to papers containing a given term, or the impact factor of the journals in which the term appears. Others compute dispersion of term usage across technologies or countries, or the number of prior art references in patents.

Inventor Characteristic indicators. In addition to characteristics of documents, we also analyse the inventors and patent assignees who use the term in patents. Examples include the Hirsch index of an inventor or the impact of prior patents granted to inventors or patent assignees.

Novelty. Term Novelty indicators measure the first appearance of a term anywhere in a patent document, as well as the first appearance of a term in specific sections of a patent, such as in the independent claims. Another Novelty indicator computes the first time a term appears with an abbreviation attached. These indicators are thus designed to analyse the timeline of the acceptance of the term by the community.

Most of the indicators described above are time series of annual counts or scores, such as a "number of prominent inventors per year using term in patents." To simplify the modelling process, we reduced each time series to a single value by applying three different methods:

(1) Find the slope of the regression line of indicator values against time (a measure of how fast the indicator is increasing over time);

(2) Calculate the average growth rate for the indicator value over the period selected for the time series;

(3) Compute the sum of indicator values for three years prior to the reference period.

We also experimented with (a) the x2 coefficient of the best-fitting, second-order polynomial for indicator value as a function of year (a measure of curvature, or rate of acceleration), and (b) the two-year prediction of this best-fitting polynomial. These indicators, while sometimes informative, were usually redundant with slope.

Forecasting Models

Our models are tree-augmented Naive Bayes networks (Friedman et al., 1997). Such networks have a structure like that of the network shown in Figure 2. For clarity, we display only a fragment of the model; a complete model may contain 30 to 50 indicator variables.

Bayesian networks provide a factorized representation of a joint probability distribution over a set of variables, and efficiently update the distribution, given evidence in the form of values for variables. In our models, there is a unique root node that represents the unobserved future prominence of an entity. In the above model, this is the node labeled "Prominence3." Prominence is normalized to be between 0 and 1, with a special value of -1 for cases in which the usage of the term decreases. As evidence is entered into the net, the probability distribution over the possible values of prominence is updated.

Bayesian Networks have shown good performance as classifiers (Friedman et al., 1997). We use a version of a Bayesian classifier in which links between indicator variables capture

synergistic effects among those variables – i.e. information about two or more variables tells us more about prominence than the sum of the information value of the individual variables. Capturing synergistic effects has been shown to improve classifier performance (Friedman et al., 1997).

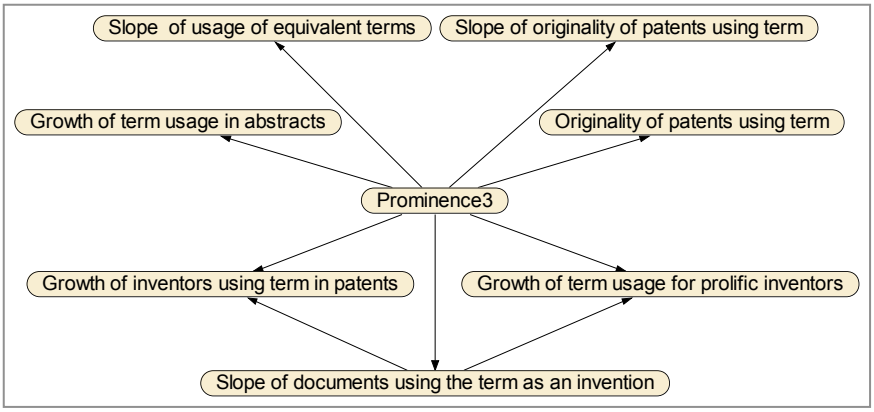

Figure 2. Fragment of model for predicting term prominence.

We chose to use Bayesian networks for several reasons. First, we executed a performance comparison between Bayesian networks (looking at common confusion matrix measurements such as the true and false positive rate, F1 score, etc.) and other classifiers such as JRip, J48, SVM, and meta-classifiers wrapping these, including Bagging and AdaBoostM1. Second, we chose Bayesian networks due to their flexibility and ease of interpretation. Finally, Bayesian networks provide insight into the contribution of indicator variables by supporting the computation of information-theoretic quantities such as mutual information and conditional mutual information.

We use a fine-grained discretization of prominence values instead of a binary prominent/not-prominent variable. This allows more precise computation of information-theoretic relations between indicator variables and prominence than does a binary variable. For example, some variables may be good at predicting very high prominence, while others merely discriminate prominent from non-prominent entities.

Although the prominence variable has a fine-grained discretization, it can be used as a binary classifier by choosing a threshold for prominence. The threshold is chosen through the multi-objective optimization process, described below.

Model Generation and Optimization

Automated model generation must answer the following questions in order to create the desired Bayes net:

- Which indicator variables should be included?
- Which indicator variables should be linked?
- How should continuous variables be discretized?
- How much weight should the training algorithm give to the training data relative to the untrained prior distribution so as to avoid over fitting?
- What threshold for predicting prominence provides the best trade off between recall, precision, and other performance goals?

All of these questions are answered by an optimization loop. This optimization loop uses a multi-objective elitist genetic algorithm (NSGA-II) to search the model parameter space (i.e. answers to the above questions) and rewards solutions that score well relative to specified recall and precision goals. The optimizer uses stratified 10-fold cross validation to compute metrics (e.g. recall and precision) for various combinations of system and ground truth

prominence thresholds. This process leverages the recall ↔ precision trade-off parameter. Finally, the optimizer promotes and further explores solutions that perform relatively well via: (1) uniform crossover, (2) Gaussian mutation for continuous variables, and (3) random flip mutation for discrete variables. The end result is an answer to the above questions that is optimized to the specified objectives.

Indicator Analysis

The analysis described below measures how well the indicators and models can forecast future term prominence, where a term is considered prominent if it has achieved a significant increase in usage.[4] To perform this analysis, we computed indicator values and generated models by processing all documents up to a given year (called the reference period), and then compared system outputs against a ground truth variable measuring an increase in term usage three years after the reference period. This analysis measures the ability of our models to forecast a significant increase in term frequency three years into the future.

By using automated model generation process described above, we generated domain-specific models for different technology areas in English and Chinese patents, including Computer Science, Communications, Biotechnology, and Semiconductors. The performance was higher for Chinese than for English, with the average recall of 0.49 and 0.52 for English patents and recall of 0.47 and precision of 0.61 for Chinese patents. The higher precision for Chinese patents is most likely due to Chinese patents containing a higher percentage of prominent terms than English patents.

To analyze individual indicators, we computed rank correlations between indicators and term prominence. Table 1 illustrates the performance of our indicators for English patents for the domain of Computer Science using Spearman's rank correlation coefficient (Rho) and three approaches to summarizing time series: slope, growth, and sum. For example, in Table 1, Rho slope for the indicator "Number of organizations per year using term in patents" shows the rank correlation for the indicator "the slope of the regression line fitted to the number of organizations using a selected term each year leading up to the reference period."

Table 1 reveals that indicators are significantly correlated with prominence for at least one computation (slope, growth, or sum), with the exception of one — the number of significant opinion relations. This is not unexpected, since opinion relations rarely occur in patents.[5] It also shows that term momentum indicators have the strongest rank correlations with prominence, i.e. measuring past momentum is particularly useful for predicting future prominence. Given that the other classes of indicators are conceptually very different from term momentum indicators, we expect that their effect on the forecasting model is additive to the momentum indicators, rather than duplicative. To test this hypothesis, we computed the partial correlations of non-momentum indicators with prominence, after the most basic term momentum has been accounted for (prior term usage in patents).

[4] One of the limitations of our system is that our analysis applies to individual terms, rather than sets of terms that are representative of technologies or research areas. This limitation is due to the problem of generation of ground truth data for training of our statistical models. In the future, we plan to extend this approach to analyse clusters of related terms, which are representative of technologies and scientific fields.

[5] Our analysis of scientific articles has shown that opinion-type relations (such as positive, standard, and negative opinion) are very infrequent in scientific literature as well, which suggests that opinion-based indicators are not particularly useful for the analysis of scientific literature and patents.

Table 1. Spearman rank correlations with future increase in term usage in English patents.

	Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Term Momentum Indicators	Number of unique organizations per year using term in patents	0.48	0.26	0.47
	Number of prolific organizations per year using term in patents	0.47	0.25	0.46
	Number of unique inventors per year using term in patents	0.50	0.13	0.47
	Number of prolific patenting inventors per year using term in patents	0.45	0.30	0.50
	Number of times per year term is used in patents	0.50	0.26	0.47
	Number of times per year equivalent terms are used in patents	0.48	0.25	0.45
	Number of times per year term is used in summary of invention section	0.52	0.26	0.51
	Number of times per year term is used in Independent claims	0.46	0.38	0.51
	Number of times per year term is used in Abstract section	0.47	0.33	0.52
	Number of industrial assignees using term per year	0.49	0.19	0.46
	Number of academic patent assignees using term per year	0.21	0.26	0.30
Term Characteristics	Annual technology score	N/S	N/S	0.19
	Annual maturity score	0.11	0.13	0.33
	Term usage as an invention	0.12	0.18	0.19
	Term usage as a component	0.23	0.25	0.27
Semantic relations	Annual counts of Exemplify relations	0.33	0.35	0.37
	Annual counts of Practical relations	0.33	0.33	0.37
	Annual counts of Opinion Significant relations	N/S	N/S	N/S
	Term usage with an abbreviation	0.19	0.23	0.24
	Annual counts of Contrast relations	0.20	0.26	0.26
	Annual counts of Based on relations	0.23	0.18	0.24
	Annual counts of Better than relations	0.17	0.13	0.18
Document Characteristic	Originality of patents using the term	N/S	N/S	0.19
	Average citation impact of documents about the term	N/S	N/S	0.31
	Term frequency in an emerging cluster	0.18	0.12	0.42
	Number of prior art references	0.02	-0.12	0.22
	Citations to high impact patents	N/S	N/S	0.31
	Dispersion of term usage across technologies	0.12	N/S	0.46
Inventor Char.	Number of patent inventors using the term as invention	0.12	0.17	0.19
	Hirsch index of the inventor	N/S	N/S	0.19
	Citation impact of prior patents granted to inventor(s)	N/S	N/S	0.29

Table 2 lists the indicators in the descending order of their partial correlations with prominence. An interesting finding is that the indicators that provide information over and above term momentum indicators include the ones that are based on language features, such as Practical and Exemplify relations, as well as term characterization. The indicators that have low or even negative correlations include document- and inventor-based indicators, such as the Hirsch index of the inventor, or the average citation index of document using the term. Having said that, it is important to note that document and inventor indicators are consistently selected by our forecasting models, which indicates that they are not really replaceable by other indicators.

Table 2. Partial correlation of indicators with prominence, controlling for momentum indicator.

Indicator	Partial Correlations
Annual counts of Practical relations	0.199
Term usage as an invention	0.170
Annual counts of Exemplify relations	0.169
Term usage as a component	0.159
Citations to high-impact patents	0.149
Annual maturity score	0.134
Annual technology score	0.129
Annual counts of Based_on relations	0.120
Annual counts of Contrast relations	0.114
Originality of patents using the term	0.101
Term usage with an abbreviation	0.098
Annual counts of Better_than relations	0.080
Citation impact of prior patents granted to inventor(s)	0.019
Average citation impact of documents about the term	-0.023
Number of prior art references	-0.042
Term frequency in an emerging cluster	-0.057
Hirsch index of the inventor	-0.074

Comparing indicators with different rationale, such as practicality versus discursive interest, one interesting finding is that the indicators focusing on the practicality of a field have the strongest correlations with prominence. These indicators include maturity scoring, usage as a component, Practical relations, and term usage by industrial patent assignees. Indicators focused on discursive interest in the term, such as Contrast relations, Better Than relations, and term usage by academic researchers in the field, have weaker (although still significant) correlations with prominence (as shown in Table 1 above). This suggests that, while both practicality and discursive interest are useful characteristics for the analysis of patents, the former is of particular value in forecasting the future prominence of terms.

Our further analysis of indicators focused on trying to identify indicators with complementary strengths. For example, we discovered that many of our indicators are good at predicting whether term usage will increase or decline/remain stable, but there are only a few indicators that are good at predicting different degrees of positive changes in term usage. This is illustrated by Table 3, which shows rank correlations between indicators and future changes in term usage coded as positive versus non-positive (Rho+/), as well as rank correlations considering positive values only (Rho-Pos).

As Table 3 shows, the correlations for the classification problem (Rho+/-) are generally higher, which suggests that it is more straightforward for an indicator to forecast whether or not a term will have a positive prominence, versus forecasting different degrees of positive prominence. It also reveals that some indicators might have particular strengths. For example, while momentum indicators and some document characteristic indicators perform best for delineating between positive and non-positive cases, the best indicator for distinguishing between different levels of positive prominence is “the proportion of granted patents using term relative to published documents”.

Table 3. Spearman correlations for indicators based on different conditions.

	Time Series indicators	Rho+/-	Rho-Pos
Term Momentum Indicators	Number of unique organizations per year using term in patents - Slope	0.50	0.21
	Number of prolific patenting organizations per year using term in patents - Slope	0.49	0.19
	Number of unique inventors per year using term in patents - Slope	0.52	0.22
	Number of prolific patenting inventors per year using term in patents - Slope	0.52	0.22
	Number of times per year term is used in patents - Slope	0.53	0.22
	Number of times per year equivalent terms are used in patents - Slope	0.51	0.20
	Number of times per year term is used in summary of invention section - Sum	0.54	0.24
	Number of times per year term is used in Independent claims section - Sum	0.53	0.25
	Number of times per year term is used in Abstract section - Sum	0.55	0.26
	Number of industrial assignees using term per year - Slope	0.51	0.21
	Number of academic patent assignees using term per year - Sum	0.33	0.09
Term Character ization	Annual technology score - Sum	0.21	0.05
	Annual maturity score - Sum	0.33	0.14
	Term usage as an invention - Sum	0.17	0.12
	Term usage as a component - Sum	0.27	0.13
Semantic relations	Annual counts of Exemplify relations - Sum	0.36	0.19
	Annual counts of Practical relations - Sum	0.37	0.18
	Term usage with an abbreviation - Sum	0.22	0.15
	Annual counts of Contrast relations - Sum	0.24	0.15
	Annual counts of Based_on relations - Sum	0.21	0.15
	Annual counts of Better_than relations - Sum	0.14	0.14
Document Characteristic	Originality of patents using the term - Sum	0.21	0.07
	Average citation impact of documents about the term- Sum	0.30	0.03
	Term frequency in an emerging cluster - Sum	0.46	0.15
	Number of prior art references - Sum	0.27	0.05
	Citations to high-impact patents - Sum	0.33	0.16
	Dispersion of term usage across technologies - Sum	0.50	0.18
Inv- entor Char.	Number of patent inventors using term as invention-Sum	0.18	0.10
	Hirsch index of the inventor - Sum	0.30	-0.02
	Citation impact of prior patents granted to inventor(s) - Sum	0.36	0.07
Single value	Proportion of granted documents using term relative to published documents	0.39	0.29
	The year the term first appeared in a patent	-0.15	0.01
	The year the term first appeared with an abbreviation	0.25	0.17

We further evaluated performance of indicators across one-, two- and three-year gap periods and observed a significant difference. All indicators tend to perform better in predicting longer forecasts (such as three-year gap) than shorter periods (such as one- or two-year gap). This may be because a three-year forecast smoothed out some of the year-by-year volatility in term usage.

Table 4. Spearman correlations for term prominence indicators in Chinese patents.

Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Number of unique inventors per year using term in patents	0.50	N/S	0.46
Number of prolific patenting inventors per year using term in patents	0.50	N/S	0.46
Number of times per year term is used in patents	0.50	0.06	0.46
Number of times per year term is used in Independent claims section	0.50	0.16	0.44
Number of unique organizations per year using term in patents	0.48	N/S	0.43
Number of prolific patenting organizations per year using term	0.48	N/S	0.44
Number of times term is used in summary of invention section	0.18	N/S	0.11
Annual maturity score	0.08	0.08	0.28

Finally, Table 4 shows correlation analysis for some of the indicators that were applied to Chinese Computer Science patents. It is important to note that citations rarely occur in Chinese patents, so indicators that are based on citation metrics cannot be used for the analysis of term prominence in Chinese. A comparison of correlations for English and Chinese (Tables 1 and 4) reveals that the general patterns across two collections are very similar, with Slope and Sum term momentum indicators performing particularly well, along with the Sum version of the Maturity Score.

Future Plans: Term Characterization

In addition to predicting future levels of interest to a technology, we expect that the indicators we developed can also provide some insights into the nature of the technology, its lifecycle, and other term characteristics. An example of this type of analysis is illustrated by 10 computer science terms, shown in Table 5.

Table 5. An analysis of 10 computer science terms.

Term	Pe	Term Characterization Analysis
RFID antenna	0.60	a device, becoming widely used in diff applications in 2007
Instant messaging	0.47	a technology or method, innovative, not a component
Robotics	0.31	a branch of technology, not a specific device, mature
XML	0.31	technology name, active area of research
Speech recognition	0.31	widely accepted technology, but best practice is being debated
Cellular telephone	0.31	a widely used standalone device, still of interest
RDF	0.31	technology name, becoming more widely used
Linux operating system	0.31	a widely accepted mature technology
GPS	0.30	a technology, widely used, mature, active area of research
Quantum computing	0	a principle or concept, innovative, no practical applications

The Pe column shows our predictions for the future changes in term usage, as described above, where zero value indicates that term usage will remain stable or decline in the future, whereas positive values predict that there will be an increased community interest in the term. The terms were analysed using 2007 as the reference period, forecasting term usage in 2010. The most interesting terms in this list include *RFID antenna* and *instant messaging*, the other terms, except for *quantum computing*, have slightly lower positive Pe values, indicating that there will be some growth in their usage between 2007 and 2010. The fact that quantum computing has zero value is not unexpected, considering that the data processed for this analysis included patent literature only, and this term has rarely been used in patents until 2007.

In addition to identifying terms with high prominence, we expect that the indicators described in the paper can also be used to characterize technologies, as illustrated in Table 5. For example, by using individual indicators or groups of indicators, we can potentially identify

widely accepted and mature technologies, terms that function as components of other technologies, active areas of research, as well as areas where best practice is being debated.

For example, Figure 3 reveals the values for the indicator that computes the average growth rate of term usage by academic institutions. This indicator can be used to identify innovative technologies that attract a growing attention from academia. Out of the 10 terms, technologies with the highest growth of academic assignees include *RFID antenna, instant messaging*, and *RDF*.

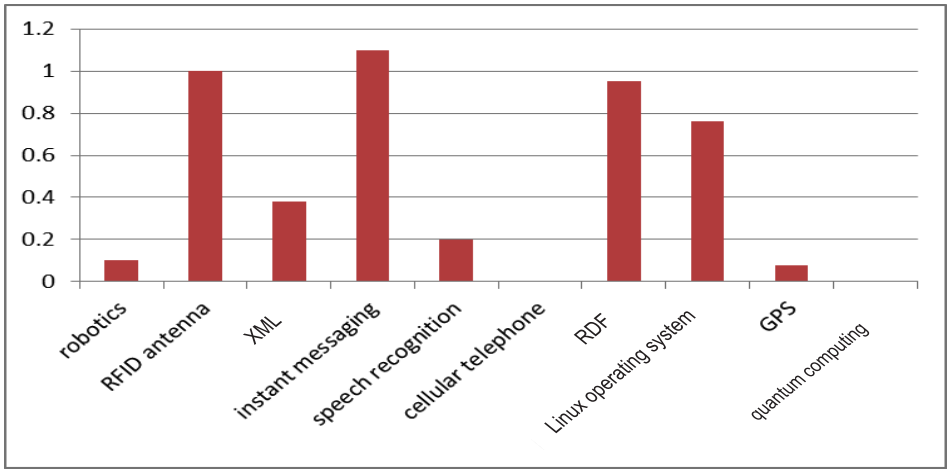

Figure 3. The average growth rate of academic assignees using term from 2002 to 2007.

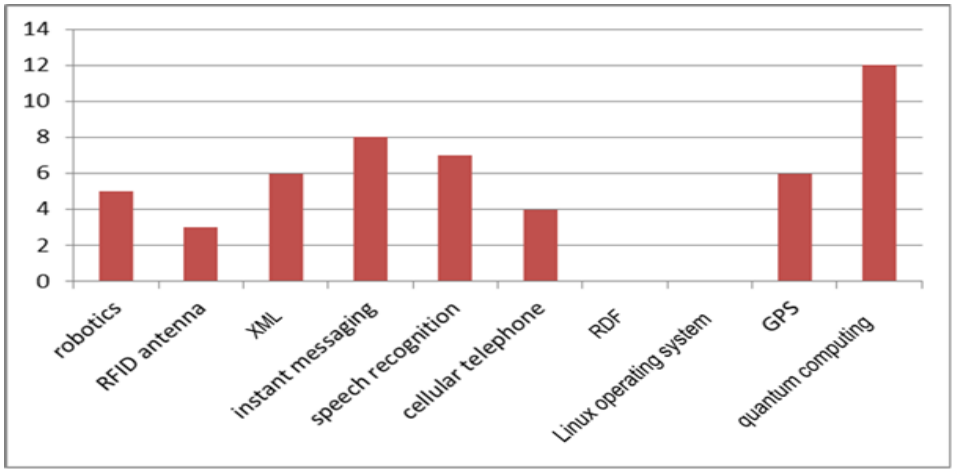

Figure 4. The number of inventors using term as an invention from 2005 to 2007.

Figure 4, on the other hand, illustrates the indicator values for "the number of inventors that were using the term as a description of an invention". Interestingly, the term that has the highest indicator value in this case is *quantum computing*. The terms with the higher values in Figure 3, *RDF and RFID antenna* have the lowest indicator values in Figure 4. This example suggests that individual indicators or groups of indicators may be used to detect different types of emerging technologies and that these differences might be related to their nature or lifecycle. It further illustrates that individual indicators can help to identify newer terms like *quantum computing*, and that high values of specific indicators may be indicative of the future potential of the term.

Conclusion

The system presented is capable of scanning millions of technical documents, extracting key indicators from both text and metadata, and forecasting meaningful trends and predictions from the extracted metrics. In particular, the extracted indicators are useful in predicting levels of interest in particular technologies. We also showed how the indicators provide insight into the nature and the lifecycle of emerging technologies, including their maturity, practicality, stages of development, and acceptance by the community.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. government.

References

Anick P, Verhagen M., & Pustejovsky J. (2014). Identification of technology terms in patents. In *Proceedings of LREC 2014*.

Arthur, B. (2009). The Nature of Technology: What It Is and How It Evolves. Free Press.

Babko-Malaya O., Thomas P., Hunter D., Meyers A., Pustejovsky P., Verhagen M., & Amis G. (2013a). Characterizing communities of practice in emerging science and technology fields, In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Babko-Malaya O., Meyers A., Pustejovsky J., & Verhagen M. (2013b). Modeling debate within a scientific community. In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495–518.

Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, 44(4), 195-205.

Brock, D.C, Babko-Malaya O., Pustejovsky, J., Thomas, P., Stromsten, S., & Barlos, F. (2012). Applied actant-network theory: Toward the automated detection of technoscientific emergence from full-text publications and patents. In *Proceedings of the AAAI Fall Symposium on Social Networks and Social Contagion 2013*.

Friedman N, Geiger, D., & Goldszmidt, M. (1997). Bayesian networks classifiers. *Machine Learning*, 29, 131-163.

Latour B. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press.

Li, S., & Xue, N. (2014). Effective document-level features for Chinese patent word segmentation, In *Proceedings of ACL 2014*.

Meyers, A., Zachary, G., Grieve-Smith, A., He, Y., Liao, S., & Grishman, R. (2014). Jargon-Term Extraction by Chunking. In *Proceedings of SADAATL 2014*.

Meyers, A. (2013). Contrasting and corroborating citations in journal articles, In *Proceedings of Recent Advances in Natural Language Processing 2013*.

Meyers, A., Lee G., Grieve-Smith A., He, Y., & Taber, H. (2014). Annotating relations in scientific articles. In *Proceedings of LREC 2014*.

Schiebel, E., Hörlesberger, M., Roche, I., François, C., & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. *Scientometrics*, *83*(3), 765-781.

Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and characterization of technological topics in the field of molecular biology. *Scientometrics*, *82*(3), 663-676.

Thomas P., Babko-Malaya O., Hunter D., Meyers A., & Verhagen M. (2013). Identifying emerging research fields with practical applications via analysis of scientific and technical documents. In *Proceedings of ISSI 2013*.