Intelligent Information Management, 2015, 7, 7-21
Published Online January 2015 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2015.71002

Scientific Research Publishing

A Heuristic Text Analytic Approach for Classifying Research Articles

Steven Walczak[1], Deborah L. Kellogg[2]

[1]Integrated Information Technology Department, University of South Carolina, Columbia, USA
[2]The Business School, University of Colorado Denver, Denver, USA
Email: swalczak@hrsm.sc.edu

Received 5 January 2015; accepted 23 January 2015; published 26 January 2015

Abstract

Classification of research articles is fundamental to analyze and understand research literature. Underlying concepts from both text analytics and concept mining form a foundation for the development of a quantitative heuristic methodology, the Scale of Theoretical and Applied Research (STAR), for classifying research. STAR demonstrates how concept mining may be used to classify research with respect to its theoretical and applied emphases. This research reports on evaluating the STAR heuristic classifier using the Business Analytics domain, by classifying 774 Business Analytics articles from 23 journals. The results indicate that STAR effectively evaluates overall article content of journals to be consistent with the expert opinion of journal editors with regard to the research type disposition of the respective journals.

Keywords

Bibliometric, Business Analytics, Concept Mining, Heuristic, Research Continuum, Text Analytics, Text Mining

1. Introduction

The ever increasing growth of information available over the Internet and also the growing quantity of open access journals makes access of reliable and usable research information problematic. Stringer *et al*. [1] have noted that "the size and growth of the research literature places a tremendous burden on researchers". Digital libraries need to evolve to facilitate proper document identification and utilization to satisfy varying needs and constituencies [2]. Classification and analysis of published research is a goal of bibliometric research, seeking to classify a domain's literature to identify commonalities or ontologies and also extensions to existing areas of research [3] [4].

How to cite this paper: Walczak, S. and Kellogg, D.L. (2015) A Heuristic Text Analytic Approach for Classifying Research Articles. *Intelligent Information Management*, **7**, 7-21. http://dx.doi.org/10.4236/iim.2015.71002

Currently, most bibliometric research and search engines utilize keywords to classify the domain of research articles and identify relevant documents [5]. Going beyond basic keyword searches, text analytics seeks to derive information from natural language text. It is often used for text categorization, text clustering or concept extraction. Akin to text analytics is concept mining, where documents are classified based on similar concepts. Concept mining can provide insights into the meaning and similarity of documents [6] and has been previously utilized to classify documents based on their research findings [7].

Categorizing research by its domain or sub-domain within a field research is the most commonly employed classification methodology for research literature access, such as the Dewey Decimal Classification system or the Library of Congress Classification system [8]. However, research may also be classified with respect to either its theoretical (also called basic research) or applied (also called practical research) research perspective. The primary reason for classifying research with respect to its theoretical or applied nature is to determine the relevancy of research articles for varying constituencies, which may require either more theoretical or more applied research results. Classifying research manuscripts is critical in reaching an appropriate audience that will be receptive to the published research and consequently will be more likely to utilize the published research findings [9] [10].

Theoretical or basic research seeks to create new theory, disprove existing theory, or expand existing theory. Theory provides the foundations of belief and understanding for a discipline, whereas applied or practical research focuses on solving domain problems.

The research presented in this article demonstrates a methodology for classifying research documents with regard to a theoretical or applied orientation. Most research contains at least some elements of both theoretical and applied research and as such we propose a theoretical-applied research continuum, which is a continuous scale of the interplay between theoretical and applied emphases for the research reported within a document.

Our research question is: "Is it possible to effectively evaluate the relative position of research manuscripts on the theoretical-applied continuum, based on quantitative aspects of the manuscript?" A heuristic algorithm, the Scale of Theoretical and Applied Research (STAR), is developed to accurately classify research on the theoretical-applied research continuum. The research question and subsequent STAR heuristic classification methodology specifically focus on quantitative properties of manuscripts, because these may be accurately measured and present a more reliable and repeatable objective evaluation of the theoretical and applied research emphasis within a given research document. The goal of this research is not to develop a better text mining algorithm, but to develop a new process for applying existing text mining, text analytics, and concept mining methodologies to more accurately classify domain specific bibliometric data.

Every new classification methodology needs to be verifiable. Since the theoretical-applied research continuum is a novel concept, existing measurements do not exist. A standard evaluation criterion is recommended that consists of the expert opinion of editors for journals within the field being analyzed. The field of Business Analytics, where complex mathematical and decision science techniques are applied to the solution of business problems [11], is selected as the field to demonstrate the proof of concept for the STAR heuristic research classification method. The field of Business Analytics has both theoretical and applied perspectives.

A survey of editors of the top journals in the Business Analytic domain is administered to determine the relative positioning of their respective journals on the theoretical-applied research continuum. The editors' perceptions of their own journal and of other Business Analytic journals serve as the expert opinion standard for evaluating the effectiveness of the newly developed STAR heuristic classification metric.

A two-stage text analytics based approach is used to classify domain dependent bibliometirc data. Text analytics is first used to identify relevant terms for the heuristic algorithm to utilize in evaluating the relative theoreticalness (or appliedness) of research manuscripts. Next, the STAR heuristic algorithm is developed to perform concept mining on research manuscripts to classify their relative position on the theoretical-applied research continuum, utilizing the terms discovered by the first stage text analytics. Finally, the heuristic STAR technique is evaluated by measuring the placement of multiple articles from the field of Business Analytics on the theoretical-applied continuum, with articles ranging across the continuum.

2. Evaluation Criteria via Editorial Opinion for the Theoretical/Applied Nature of Business Analytics Journals

A survey is developed to verify if the classification of articles based on their theoretical or applied nature was a

reasonable research assumption. A list of editors and their contact information who were serving in 2007 for 23 top Business Analytics domain journals (journals shown in first column of **Table 1**) is compiled. Journals are selected based on the criteria that the journal had to appear on multiple journal rankings lists and had to be ranked in the relevant "top tier" of the particular ranking at least twice. Several of the journals had multiple editors and all of the editors for each journal for which current contact information was available are added to the list (e.g., [12]). The survey was then sent to the Business Analytics journal editors for establishing the efficacy of the proposed research, and who serve as the domain experts for establishing the research evaluation criteria. Twenty editor responses were received, representing 13 of the 23 total journals surveyed, or approximately 57% of the journals.

Four open-ended questions of the survey asked each editor to define theoretical and applied research and to identify article specifics that would help them identify an article as theoretical or applied. The interesting result here is that 19 out of 20 of the responses were able to provide definitions and article characteristics to distinguish between theoretical and applied research, with only one editor indicating they did not or were unwilling to distinguish research based on its theoretical versus applied nature.

Each editor was also asked to classify each of the 23 Business Analytics journals using a 7-point Likert-like

Table 1. Editors' theoretical/applied rating of business analytics journals.

Journal	Mean Classification	St. Error (Mean)
Mathematics of Operations Research*	1.17	0.11
Annals of Probability*	1.50	0.19
Mathematical Programming	1.55	0.21
Annals of Statistics	1.75	0.18
Operations Research*	2.00	0.28
J. of the Royal Statistical Society, Series B	2.38	0.33
Biometrika	2.64	0.34
Management Science	2.64	0.31
J. of the American Statistical Association	2.85	0.32
European J. of Operational Research	3.21	0.33
Decision Sciences	3.60	0.40
Naval Research Logistics	3.69	0.29
Computers & Operations Research	3.83	0.34
J. of the Royal Statistical Society, Series A	3.85	0.60
Transportation Science	3.90	0.53
J. of Operations Management	3.91	0.53
OMEGA—Int. J. of Management Science	4.09	0.28
IIE Transactions	4.17	0.39
J. of the Operational Research Society	4.17	0.47
Int. J. of Production Research	4.33	0.45
Transportation Research Part B: Methodological	4.33	0.65
Production and Operations Management*	4.70	0.50
Interfaces*	6.43	0.23

*Journals selected for text analytics.

scale [13] where 1 represented purely theoretical and 7 represented purely applied, with 4 representing an even balance between applied and theoretical emphasis for that journal's articles. The results for the journal classifications are shown in **Table 1**. The respondents were also given the opportunity to skip rating any journal with which they were unfamiliar (which happened for 37% of the total responses).

The values shown in **Table 1** indicate the average perception by the expert group of editors for each journal's articles with regard to placement on the theoretical-applied research continuum. Interestingly, on average no journal was perceived as being either purely theoretical or purely applied by the panel of editor experts, though *Mathematics of Operations Research* came close to being considered purely theoretical. Additionally, of the editors who responded, only one of these editors rated their own journal as being on one of the two extremes.

Most editors' view of their own journal was relatively consistent with the peer group's evaluation, with the exceptions being *Computers & Operations Research* and *Journal of the Royal Statistical Society*, *Series A* both of which thought their journal focused on moderately applied research while the peer group felt these journals were both marginally on the theoretical side of a balanced distribution of research emphasis. Inter-rater agreement is calculated using an intra-class correlation coefficient (ICC), since multiple raters are being compared to the group as a whole. The ICC value is 0.968 indicating with very high reliability that the individual editor's ratings are consistent with the group ($p < 0.001$) [14] [15].

3. Method Part 1: Text Analytics to Identify Theoretical and Applied Research Terms

Text analytics in addition to being a tool used in Business Analytics research [16] is an increasingly popular methodology for examining literature and performing bibliometric analysis [17]-[19]. Text analytics may be used to answer such questions as what are new research areas, who is publishing about a specific research topic, and where is this research being published [20].

Yang *et al*. [20] classify commercial text analytics systems into three types: able to work with unstructured format documents, requires highly structured format in documents, and domain specific (patents). Although academic articles tend to have a certain framework, the type of data analyzed bibliometricly is unstructured; therefore the most general type of text analytics paradigm must be utilized.

Bibliometric text analytics enables identification of hidden patterns in data for use in analysis of large data sets [17]. Usage of statistically derived keywords to characterize texts is becoming an increasingly important research methodology to explore differences between and classification of texts [21]. Glänzel [22] states that identifying the common vocabulary used within research disciplines is a necessary step in identifying research trends within literature. Other research [23] has demonstrated the efficacy of utilizing keyword based searches for performing text analytics article classification, including using keywords gathered from prior text analytics to classify current research [4]. The research reported in this article claims that theoretical research will utilize its own vocabulary distinct from applied research within any particular field (e.g., Business Analytics).

A text analytics approach is used to identify the distinct applied research and theoretical research terms commonly used in the Business Analytics literature. The text analytics application is written in C#, and does a character by character search of prepared article text files. The text analytics classification built for this research may be classified as a generalized text-miner capable of working with unstructured data. The text analytics application was verified against two test articles that had all words counted and documented manually.

The articles to be text analyzed are first transformed into a simple text file to eliminate specialized application control codes inserted by most modern word processing applications (e.g., Microsoft® Word, Latex®, and Adobe® Acrobat) that would cause noise in the keyword search process. Care must be taken in the current text mining approach to remove redundant information that does not contribute to the content of the research article or that are intentionally repeated to satisfy manuscript formatting so as not to skew the text analytics results. Thus the reference section of each article, keyword lists, author biographies, and any headers and footers are eliminated. Future research may investigate the use of the text analytics algorithm solely on the reference sections of documents to identify frequently cited authors for literature review citation and co-citation analysis research and also identify specific seminal articles that are shaping the domain based on citation frequency.

Strings of text representing words are collected from the document and stored in an array along with the number of occurrences within each document analyzed. Whitespace and punctuation marks other than apostrophe are used to denote word boundaries. Keeping track of the frequency of occurrence of specific keywords

may better characterize texts versus just identifying the simple occurrence of a keyword [21]. Words less than three characters in length, words containing numbers or that are solely numeric, pronouns, adjectives, adverbs, articles, conjunctions, and prepositions are eliminated from the word list automatically as these typically are too common and do not express meaning and thus cannot be used to distinguish between anything, including the type of research being performed.

The arrays of words (and their occurrence counts) that remain may then be sorted either alphabetically or in decreasing order of occurrence. This final sorted list is then written out to a file as tab-separated values for consumption by an electronic spreadsheet or statistical analysis program for further text analytics. The sorted individual words are then organized and collected into regular expressions [24], called "word patterns" for this research. Regular expressions enable rapid identification of all word forms of single words, such as pluralization of a word. For example, the regular expression *industr** represents the words: industry, industries, and industrial. A preceding or trailing space included in the word pattern is used to indicate that the characters of the pattern must occur at the beginning or end of a word respectively. The * indicates a wildcard that may be filled by any alphabetic character or characters and is only used in the article for explanation purposes; the actual keyword patterns omit the star as it is implied in the STAR algorithm for all regular expressions not bracketed with a space.

A pseudo-meta-analysis, as recommended for summarization or synthesis of research literature [25], is performed by utilizing the three of the five most theoretical journals and the two most applied journals identified by the editors' survey, as indicated in **Table 1**. Ten randomly selected articles from *Interfaces*, the most applied journal, and five randomly selected articles from each of the remaining journals are analyzed using the text analytics algorithm just described. The rationale for only using two journals versus three for the applied research text analytics is due to the rapid drop off in perception of applied-oriented research journals as indicated by the Business Analytics journal editors and the requirement to capture terms that are generalizable across a wide variety of Business Analytics journals.

The initial text analytics produced a collection of 2050 possible applied research word patterns and 2356 possible theoretical research word patterns. The collection of words captured across these 5 journals and 30 articles is then subjected to further heuristic analysis to determine if an identified word could serve to distinguish between theoretical and applied research. The heuristics perform a multi-criteria identification process similar to variable reduction processes found in other research [26]. The text analytics (TA) heuristic criteria for identifying research type specific words are:

TA heuristic criterion 1 The regular expression must occur at least 10 times across the collection of all articles for the specific research type;

TA heuristic criterion 2 The regular expression must appear in at least 3 separate articles of the specific research type;

TA heuristic criterion 3 The regular expression must appear in at least 2 distinct journals for the specific research type;

TA heuristic criterion 4 The regular expression may not appear in more than one article of the other research type.

Heuristic criteria 1 - 3 above follow from Conway's [21] observation that in text analytics frequent occurrence of words serves as a more reliable indicator than just the observation that a word occurred. Heuristic 4 assists in eliminating words that are commonly used by both types of research methodology and as such are insufficient to classify any individual research article. An alternative heuristic 4 could eliminate words that appear too frequently in other research type articles, thus enabling a fuzzy classifier depending on the range of values used to define frequency. These heuristics greatly narrowed the identifiable applied and theoretical research terms from the large collection of words occurring across all the articles. A total of 17 applied research word patterns and 19 theoretical word patterns are identified for the Business Analytics field and are shown in **Table 2**.

The average occurrence of the 17 applied and 19 theoretical word patterns as well as the average number of articles in which they appeared is shown in **Table 3**. An additional line in **Table 3** shows the occurrences of the word "theory". **Table 3** demonstrates that the frequency of the selected keywords is much greater in the corresponding article type. Additionally for the selected keywords, it appears that the theoretical keywords are used across a greater quantity of similar research type articles than the applied keywords indicating that these theoretical keywords are more commonly used across the corresponding literature.

Some examples of words that were rejected as keywords may help to illuminate the efficacy of the heuristics

Table 2. Applied and theoretical Business Analytics patterns.

Applied Research Word Patterns	Theoretical Research Word Patterns
business*	equalit*
financial	finite*
forecast*	arbitrar*
manufacture*	theorem*
_firm*	linear*
facilit*	vector*
industr*	corollar*
annual*	markov
capital*	induction
market*	verif*
commit*	asymptot*
consumer*	topolog*
attribut*	dimension*
_user*	partit*
_asset*	neighbor*
equipment	_law_
_npv	calculation
	invers*
	initializ*

The * is a wildcard character allowing for substitution of zero or more characters of any type and the _ indicates a required whitespace character in the regular expression pattern. The STAR text analytics approach allows for punctuation marks to substitute for whitespace characters.

Table 3. Average appearance of Business Analytics keywords across the 15 applied and 15 theoretical articles.

Type of Keywords	Occurrences in Applied Articles	Number of Applied Articles	Occurrences in Theoretic Articles	Number of Theoretic Articles
Applied	86	4.35	1.118	0.412
Theoretical	0.278	0.222	77.53	9.11
"Theory"	24	3	28	9

utilized. The pattern "cost*", which may seem to be more of an applied word, was used 417 times across 12 of the 15 applied articles, but conversely this pattern also occurred 66 times across 6 of the 15 theoretical articles, indicating that this particular word is used meaningfully across both types of research and is therefore not a good classifier. An example of a word pattern that might be interpreted as representing theoretical research in the Business Analytics field is "parameter*" which occurred 143 times across 10 of the 15 theoretical articles, but also occurred 110 times across 7 of the 15 applies articles. Interestingly, the word "theory" shown in the last line of **Table 3** appears to be used almost as frequently based on raw occurrences across articles from both types of research.

4. Method Part 2: Design of the Scale of Theoretical and Applied Research (STAR)

Once a prospective list of domain word patterns are identified from the extant literature for the Business Ana-

lytics domain, a method for utilizing these terms to classify the relative position of articles, manuscripts, and other research documents on the theoretical-applied research continuum is needed. As seen from **Table 3**, since the keywords occasionally occur in the opposite type of literature simply performing a keyword search on articles is insufficient to accurately classify each new article on the continuum. Also since regular expressions or patterns are being used to capture nuances in the various forms a word may take, a straight forward keyword lookup, which finds exact matching patterns in the text, is also problematic.

The solution is to perform a second round of specific text analytics, but this time focusing on the regular expression word patterns identified from the first round of generalized text analytics. The STAR methodology therefore requires two parameters to operate. The first is a simple text file with the list of applied research word patterns and theoretical research word patterns. Each set of word patterns is preceded by an integer value that specifies the quantity of word patterns for each research type, with the applied research type patterns always appearing first. The second parameter is a text file containing the article to be evaluated, which for best results should also have references, keywords, biographies, headers, and footers removed. Selection of the files for the two required parameters is accomplished by opening a file browser to allow the user to select the appropriate file.

Once the keyword pattern file and article text file are known to the STAR heuristic tool, STAR reads in the specified keyword patterns and maintains them in separate arrays. STAR behaves similar to grep [27], but utilizing multiple lists of multiple regular expression patterns. STAR scans the article text file one character at a time, similar to the initial general text analytics performed to identify the keyword patterns. As each new character is processed, STAR compares the currently acquired characters to each of the applied and theoretical patterns and records if a match is found. Since the STAR text analytics is looking for complete words that match regular expression patterns, white space characters as used to stop the current pattern held in STAR. The process is repeated for each new set of characters until the end of the file is reached. STAR then produces an output screen, with options to save the output in text or spreadsheet formats. **Figure 1** shows the output for a sample theoretical article, which is divided into three screenshots for readability.

As may be seen in **Figure 1**, the STAR methodology utilizes three heuristics for estimating the placement of an article on the theoretical-applied research continuum. The three heuristics are applied sequentially. Where R represents the occurrence of a specific regular expression within an article, with subscript A or T denoting applied or theoretical pattern type and i indicating the specific pattern, and C is the count of occurrences of a specific R within an article and S representing the total number of regular expressions for that type and H is the heuristic value (initialized at zero), the STAR heuristics are:

$$\begin{aligned} &\text{If } \sum R_{\mathrm{A},i} = 0 \quad \text{Then } H = -10\alpha \\ &\text{Else If } \sum R_{\mathrm{T},i} = 0 \quad \text{Then } H = 10\alpha \\ &\text{Else } H = \alpha\left(\sum R_{\mathrm{A},i} - \sum R_{\mathrm{T},i}\right) \end{aligned} \tag{1}$$

$$H = H + \beta\left(\left(100 * \left(\sum R_{\mathrm{A},i} / S_{\mathrm{A}}\right)\right) - \left(100 * \left(\sum R_{\mathrm{T},i} / S_{\mathrm{T}}\right)\right)\right) \tag{2}$$

$$H = H + \gamma\left(\left(\sum C_{\mathrm{A},i} / \sum R_{\mathrm{A},i}\right) - \left(\sum C_{\mathrm{T},i} / \sum R_{\mathrm{T},i}\right)\right) \tag{3}$$

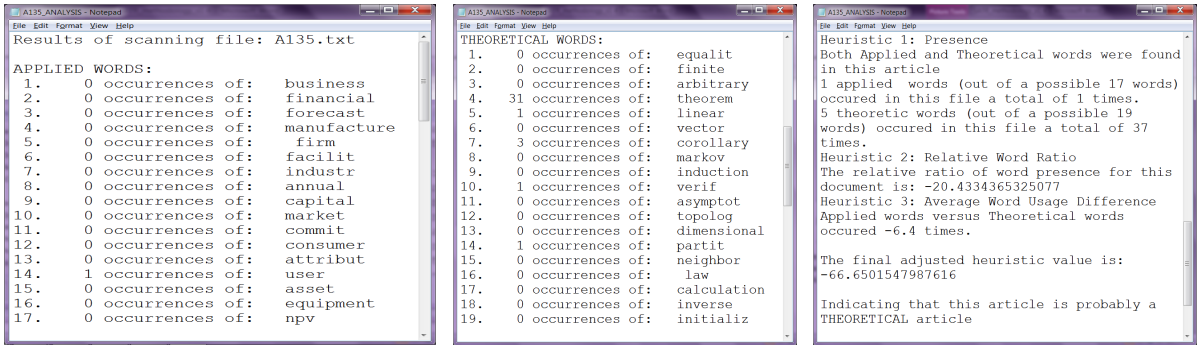

Figure 1. STAR heuristic algorithm output screens for a theoretical manuscript.

STAR heuristic 1 shows that the mere presence of applied or theoretical keyword patterns implies the type of research. If words of only one research type (applied or theoretical) are present then this lends strong evidence for that type, otherwise the difference of the applied pattern count and the theoretical pattern count becomes an indicator of the type of research, multiplied by the heuristic constant α. STAR heuristic 2 uses the relative usage ratio or coverage of the utilized applied and theoretical patterns to imply the type of research. Lastly, if applied or theoretical regular expressions are present from 1, then the corresponding part of STAR heuristic 3 is evaluated, such that the preponderance of word pattern usage supports the type of research.

The final STAR adjusted heuristic value H utilizes information from all three heuristics to determine the corresponding value. The first heuristic simply looks for applied or theoretical patterns to infer the research type for the manuscript. If both types of patterns exist then the heuristic calculates which patterns appear more frequently. The second heuristic makes adjustments to the first heuristic value to account for possible differences in the number of type of patterns (where a pattern type is a regular expression belonging to a specific research methodology) that exist for each research type. Finally, the third heuristic alters the heuristic value to account for the actual quantity of applied or theoretical patterns found. Thus if 100 percent of the theoretical word patterns occurred in a manuscript, but each pattern only occurred 1 or 2 times (1.5 average), and only 50 percent of the applied patterns occurred, but each pattern occurred on average 75 times, then this would indicate that even though more theoretical patterns existed, the overall pattern usage is more applied than theoretical.

The STAR final heuristic value H classifies the corresponding document with respect to its relative position on the theoretical-applied research continuum. Unlike traditional classifiers, STAR places each document on a continuous value scale, as opposed to a discrete value. These adjusted heuristic values are not probabilities, but rather represent a relative placement on the continuum. With the current heuristic constants, α, β, and γ, the STAR heuristic continuum position values range from −250 (absolutely theoretical) to 250 (absolutely applied). The largest negative (most theoretical) and largest positive (most applied) research type continuum placement scores recorded to date by STAR for Business Analytics articles are −180.59 and 201.74. These may be turned into pseudo-likelihood estimates by scaling the range appropriately.

STAR also provides statements to help guide the user in interpreting the STAR heuristic metric value, ranging from "cannot be differentiated as a <type>" to "is almost certainly (very high probability) a <type>". The example in **Figure 1** shows an article that is just below the near certainty value. The statements group the continuous classification values into 7 discreet sets, which enable better analysis against the expert opinion values. These statements are not meant to be construed as authoritative, but rather as applying a verbal interpretation to the STAR metric value. For example, the current cutoff used to indicate near balance, "cannot be differentiated", is an absolute value of 1 or less, but a more liberal interpretation of a balanced methodology might expand this to be any manuscript with an absolute value of 10 or less.

Currently, the heuristic constants α, β, and γ and the cutoff levels for each of the levels of advice are embedded in the STAR method. Future research and modifications to STAR will examine enabling the user to set the cutoff values to adjust the classifications, thus enabling more liberal or conversely more conservative interpretations of the corresponding theoretical-applied research continuum value.

5. Results and Discussion

The STAR heuristic method for classifying manuscripts with respect to their placement on the theoretical-applied continuum is evaluated by applying it against published articles from each of the 23 top Business Analytics journals, with a total of 774 articles. The utilization of 774 articles from 23 journals provides a meta-analysis approach to analyzing the efficacy of the STAR methodology for classifying research manuscripts. None of the 774 articles were used in the initial text analytics to define the theoretical and applied ontologies for Business Analytics. Twenty randomly selected articles from the 774 articles were read and evaluated independently with respect to theoretical versus applied orientation of the research by a small focus group. The focus group found that the STAR heuristic methodology rated all articles judged to be applied by the focus group with a positive (applied) value and all articles judged to be theoretical by the focus group with a corresponding negative (theoretical) value, thus confirming the consistency of the STAR methodology to the focus group evaluations for these 20 articles.

The individual STAR metric values for all articles across a specific journal provides further supporting evidence for the efficacy of the STAR research type analysis methodology by examining how closely they align

with the editors' expert opinion of the various Business Analytics journals. The research hypothesis H_0: the STAR methodology consistently reflects expert editorial opinion of the overall research content type of Business Analytics journals, may be evaluated by such a comparison. Results are shown in **Table 4**, which includes the editors' ratings, a linearly based prediction in the same range as the editors' values (1 - 7) based on the STAR metric value, and the average, maximum, and minimum STAR heuristic values across all articles evaluated for each specific journal. The linearized STAR prediction value segments the complete range of STAR values into seven equal sized ranges of a width of approximately 54.62.

Since the third column that linearizes the average STAR heuristic metric is an integer value, if we round the editors' expert opinion of the theoretical-applied nature of the journals to also be an integer value, then the STAR metric has an average error of 0.696, or on average of less than 1 position away from the averaged editors' perceptions of journal research type. Thus the hypothesis that there is no difference between the editors' expert rating of the journals and the STAR predicted journal rating based on the sample of articles analyzed, is confirmed with a $p < 0.05$ (actual $p = 0.0359584$).

In fact, 8 (or almost 35%) of the STAR predictions are identical to the editors' expert opinion and only one prediction is more than one away, which was the prediction for the *Journal of Operations Management* (*JOM*). The editors' viewed *JOM* as being nearly balanced with just a slight edge toward the theoretical side, but STAR based on its current Business Analytics research type ontologies evaluated the content of the articles analyzed for this journal as being very applied (6 out of 7).

Table 4. Average STAR values for 23 Business Analytics journals.

Journal	Editor Rating	STAR Prediction	Avg. STAR*	Max. STAR*	Min. STAR*
Mathematics of Operations Research	1.2	1	−115.0	19.5	−173.6
Annals of Probability	1.5	1	−129.7	−77.8	−180.6
Mathematical Programming	1.6	1	−104.3	−35.8	−161.8
Annals of Statistics	1.8	1	−129.3	−79.7	−178.8
Operations Research	2.0	3	−21.7	126.8	−107.8
J. of the Royal Statistical Society, B	2.4	1	−91.8	−22.9	−132.2
Biometrika	2.6	2	−83.1	−38.2	−132.9
Management Science	2.6	4	28.6	150.2	−92.5
J. of the American Statistical Assoc.	2.8	2	−61.9	32.5	−117.3
European J. of Operational Research	3.2	3	−10.5	133.6	−135.2
Decision Sciences	3.6	5	77.0	118.8	14.7
Naval Research Logistics	3.7	3	−50.3	32.8	−131.3
Computers & Operations Research	3.8	3	−19.1	110.2	−167.3
J. of the Royal Statistical Society, A	3.8	3	−21.7	92.2	−86.0
Transportation Science	3.9	3	−41.7	−4.2	−121.1
J. of Operations Management	3.9	6	102.4	183.0	34.7
OMEGA	4.1	4	37.3	201.7	−45.4
IIE Transactions	4.2	3	−16.7	78.4	−114.5
J. of the Operational Research Society	4.2	4	8.6	92.0	−79.7
Int. J. of Production Research	4.3	4	34.4	161.1	−92.7
Transportation Research Part B: Methodological	4.3	4	−8.1	65.1	−105.3
Production & Operations Management	4.7	5	70.0	149.2	4.6
Interfaces	6.4	5	49.4	116.1	−42.6

*All star values rounded to nearest tenth.

Another observation from **Table 4** comes from the minimum and maximum STAR metric values' columns, which shows that most journals have a mixture of research type articles. Out of these 23 top Business Analytics journals, for the articles analyzed, only 3 had all positive STAR values or applied research type articles and only 6 had all negative STAR values or theoretical research type articles.

The STAR research type continuum values may also serve to demonstrate the concept of the theoretical-applied research continuum. **Figure 2** displays a scatter plot of the rounded STAR values for all 774 articles and demonstrates that the articles do in fact form a continuum of theoretical and applied aspects of research contained within articles. While not a goal of the STAR methodology, the scatter plot of **Figure 2** has a distribution that appears normal. The hypothesis H_1: that real-world articles distributed across the theoretical-applied research continuum will have a quasi-normal distribution; is evaluated using the Shapiro_Wilks W test (p = 0.0571). No other tested distribution fit the data better than the normal, thus supporting the hypothesis that the theoretical-applied research continuum is normally distributed with respect to real world research articles from a specific field. This side benefit enables the STAR to be used in other research where the assumption of normality is needed.

The research results should also show that words which are placed in the applied and theoretical research type ontologies are used across a large number of articles, especially articles in those journals of the same research type. **Table 5** shows the respective ontology words along with their utilization across the 774 articles from all 23 Business Analytics research journals. Evidence from **Table 5** implies that the majority of the words selected by the text analytics heuristics for the ontologies are valid choices with high respective article counts, recalling that more theoretical articles will not contain many applied ontology patterns and vice versa. STAR rated 337, or 43.6 percent as being articles on the theoretical side for type of research and 436, or 56.4 percent being on the applied side, with 1 article perfectly balanced.

The only ontology pattern that did not appear in at least 50 articles is the applied research pattern "_npv". This indicates that the pattern _npv merits further consideration for removal from the ontology. An experiment, which eliminated the _npv pattern showed only minor changes for 13 out of the 15 articles, with none of them changing the sign of their STAR metric value, indicating that all were classified correctly using the full ontologies with the pattern _npv present, but that the _npv pattern was not necessarily needed in the applied ontology for the Business Analytics field. Only 2 of the articles moved sufficiently to change the interpretation of the theoretical-applied research continuum placement using STAR's verbal description cutoff values, with one moving from near balanced but theoretical to likely theoretical research type and the other moving from near balanced but applied to likely applied research type.

Since STAR utilizes text analytics to gather data for utilization with its research type continuum measurement, one concern may be the amount of time that this application requires for performing this complex analysis.

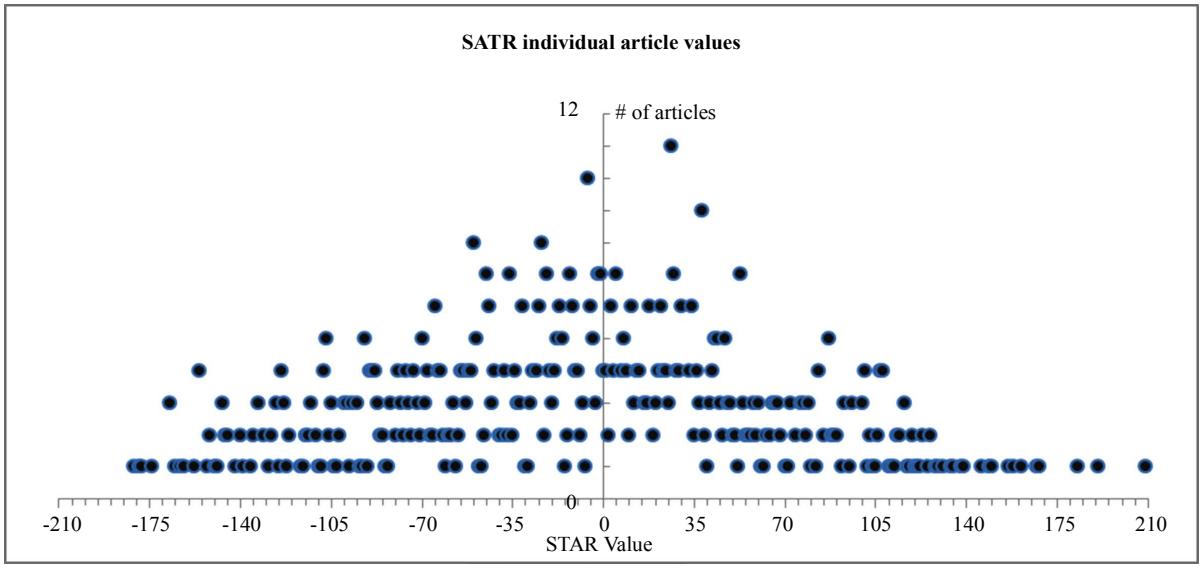

Figure 2. Scatter plot of STAR values for 774 Business Analytics articles published across 23 journals.

Table 5. Occurrence of ontology patterns across 774 articles.

Applied Pattern	Total Usage	Appeared in # Articles	Theoretical Pattern	Total Usage	Appeared in # Articles
business*	1862	217	equalit*	1976	286
financial	698	129	finite*	1937	257
forecast*	1646	133	arbitrar*	551	203
manufacture*	2059	173	theorem*	4360	318
_firm*	4260	149	linear*	4166	532
facilit*	1642	302	vector*	3656	368
industr*	2572	313	corollar*	651	138
annual*	551	137	markov	920	138
capital*	710	129	induction	149	63
market*	4284	301	verif*	539	202
commit*	510	119	asymptot*	1186	139
consumer*	1131	108	topolog*	275	54
attribut*	1454	218	dimension*	959	262
_user*	1167	212	partit*	808	159
_asset*	1046	114	neighbor*	688	120
equipment	787	139	_law_	318	92
_npv	137	15	calculation	753	262
			inverse*	524	167
			initializ*	210	94
AVERAGES	1559.8	171.1		1291.2	202.8

*Represents a wildcard that may be matched by zero or more of any characters; _ represents a required blank space, the _ is used instead of a space for visibility in the article only.

Recall, that typical research articles range from 5000 words to over 10,000 words. Although formal timing experiments have not been conducted, for the 774 articles analyzed using STAR in the current research, each was completed in less than 1 second of time on a Windows 7 notebook computer using an Intel™ i7 2.8 GHz chip, with 8 G of RAM.

5.1. Limitations

The empirical evidence shown indicates the efficacy of the STAR tool and methodology for classifying the relative position of Business Analytic articles on the theoretical-applied research continuum. Although the constituent parts of Business Analytics (e.g., statistics, decision sciences, management science, decision support systems and business intelligence [28]) have long histories, the combined field of Business Analytics is newer and thus the ontology of Business Analytics will likely grow and change over the coming years until the field achieves stability. This means that the current ontology discovered by the text analytics phase prior to the STAR tool deployment, which has been shown to accurately classify a large set of articles across 23 Business Analytics journals, will likely need to be modified in the future to capture newer and evolving elements of the ontology.

STAR is dependent on consistent usage of the domain's ontology by authors. Two examples of articles which caused problems for STAR are:

- An article was determined to be a moderately theoreticaltype by an independent panel of Business Analytics researchers. In this article, the author defined his own acronym NPV, meaning negative predictive value, but

this word pattern was already in the applied Business Analytics ontology as an acronym for "net present value". This article consequently received a STAR value of −5.3, still indicating a very slight theoretical leaning, but nearly balanced on the research type continuum. If the NPV pattern was removed, then the resulting STAR value for this article would have been −24.5, which shows a slightly stronger theoretical disposition.

- The second article that shows a limitation of STAR received a perfectly balanced score of zero. Further investigation revealed that this article, which subjectively would be classified as being of a more applied research type, did not contain any of the theoretical or applied regular expressions and thus could not be categorized by STAR.

The first article listed above indicates issues with overuse of terms and acronyms across closely related disciplines. Since NPV is already a recognized term in business research, then saying negative PV, would have prevented this confusion. The frequency of usage of the duplicate, but alternate research type pattern will dictate the net effect on the STAR metric.

The second article helped uncover an issue with how to interpret STAR findings. STAR produces a metric value of zero, which indicates a balanced research type. However, no quantitative evidence from the article supports this valuation. A new decision rule has been incorporated into STAR so that if a zero value is produced by the heuristics, but no patterns were identified, then STAR will change its output to be a null value and indicate that it had insufficient evidence to make a classification of the article's theoretical-applied continuum position. Those articles that receive a STAR value of zero, but which have supporting evidence for the heuristics from the article contents, are still accurately classified as being perfectly balanced (midpoint) in the theoretical-applied research methodology continuum.

Future research is needed to evaluate the optimal number of domain articles to use in developing the theoretical and applied ontologies. The current research used 15 articles of each type. Is this a sufficient number? The early results indicate that 15 articles of each research type for the Business Analytics domain is a sufficient number to produce efficacious ontologies. Future research is needed to determine the threshold value for accurately acquiring theoretical and applied ontologies for any research domain. The number should be small for practical implementation reasons and this is a reasonable assumption since a domain's ontology will be used repeatedly across most domain articles. Future research should examine if adding additional articles would increase or decrease the identified research type regular expressions and if this resultant increase or decrease improves the performance of STAR's classifications.

5.2. Implications for Practice

A practical utilization of STAR is to enable researchers and practitioners to evaluate the potential research type of articles and to locate desired types of research. This will enable them to rapidly assess if the research type of any article in the domain matches the research type needed to further their current research or practice.

The perception of the research type orientation of a journal will influence its readership. Highly theoretical research should ideally be positioned to be read by individuals interested in theoretical aspects of research. Likewise highly applied research articles should be positioned to be readily available to practitioners and researchers interested in the application of research to solve real world problems [10]. STAR will enable authors to get an objective quantitative assessment of their article's research type, without introducing bias from desires of those too close to the research. This will enable authors to better select publication outlets that will most likely have a receptive readership that expects to encounter the specific type of research and is more likely to utilize the research findings [9].

The STAR results may also be employed by journal editors or editorial staff. From **Table 1**, the editors agreed with each other when evaluating the research type of the five most theoretical journals and the one most applied journal and two journals in the middle (where agreement is interpreted as a standard deviation of less than 1). The remaining 16 journals indicate that the expected article research type is less well known. Editors may utilize the STAR research type classification tool to better evaluate the possible perception of articles being published to make sure that these align with the desired research type orientation of the journal and the expectations of their readership.

As noted in **Table 1**, none of the 23 Business Analytics journals analyzed where evaluated by the editors as being purely theoretical or purely applied. This is supported by the range of STAR article values reported for

each journal in **Table 4**, with 14 of the 23 journals having at least one paper that was on the theoretical side and also at least one that was on the applied side of balanced research type. Thus, one must use caution interpreting the STAR values of individual articles when evaluating the overall research type of either a journal or other body of literature. From a journal editor's perspective, there are at least two ways to achieve a specific theoretical-applied research type continuum value for a journal. The first is to only publish articles that would have STAR values close to the desire research type orientation for the journal. Both the *Annals of Probability* and the *Annals of Statistics* appear to be following this paradigm. This will necessarily be the case for those journals whose editors desire to place them at either of the research type continuum extremes.

The other technique is to publish articles across a wide range of STAR values including very theoretical and very applied research type articles, but balancing the selection of articles such that the average STAR value for all articles published is close to the desired research type orientation. Based on the STAR values produced for the 23 Business Analytics journals, this appears to be the case for at least 14 of them. This approach enables greater flexibility in accepting articles outside of the desired research type classification, but still enables the journal to maintain the desired overall research type perception. This also means that researchers attempting to determine the research orientation of a specific journal (e.g., a new open access journal in the field) should evaluate several articles from that journal and then utilize an averaged STAR classification to determine the corresponding fit with their research needs. Future research may also examine how many articles and from how many issues of a journal are required to be evaluated by STAR so that the average STAR value for that journal is representative of the overall research methodology types for the journal as a whole.

The STAR classification tool is available publicly at http://win.itechcarolina.com/projects/satr/SATR.aspx. This is an ASP web application and is being made available for researchers interested in classifying research on the theoretical-applied research continuum. As new domain ontologies are uncovered, they too will be posted to the website. Every research field will need to undergo its own text analytics knowledge discovery of appropriate regular expressions since the same term will have different meanings across different domains [29].

6. Conclusions

The complete methodology for analyzing research and classifying its relative position on the theoretical-applied research continuum requires two distinct phases. The first is a broad and general text analytics solution to capture potential applied type research and theoretical type research regular expressions to create an ontology for each research type specific to a particular field. Analysis of the text analytics data to identify members of the ontologies is time consuming. Future research is needed to further automate this process and the application of the text analytics heuristics, so that collection of ontologies for additional domains may be accomplished expeditiously.

Collection of additional domain research type ontologies is another area for future research. Ongoing research is examining the development of research type ontologies for the domain of information systems research. Acquisition of additional domain ontologies will enable the application of STAR across a broader range of research fields, increasing the meta-analytic potential for STAR. Research may then evaluate the application of STAR in other domains and examine if the current STAR heuristics may accurately classify research type, given a specified research type ontology, or if additional domain-specific heuristics may be needed.

Once domain research type ontologies are acquired via text analytics, STAR utilizes its own concept mining and heuristics to classify the research type of manuscripts and articles on the theoretical-applied research continuum. STAR may be used to link research type to various bibliometric analysis variables. Through STAR's output, the propensity for each type of research may be linked not only to journals as shown, but also to other demographic article classifications, such as universities or business associated with the authors and geographic regions of the world.

Finally, the STAR methodology[1] is able to heuristically evaluate research articles and estimate their position on the research type continuum. Results from analyzing 774 articles from 23 Business Analytics journals have shown that using journal editors as expert evaluators, the STAR methodology accurately matches the average journal article research type for the majority of the 23 journals ($p < 0.05$). Knowing the research type of an ar-

[1]Just as a fun side note, the STAR value for this article using the Business Analytics ontology is 39.9, indicating it has an applied nature, but not significantly so. This stands to reason since this article develops a new algorithm, which may be seen as a theoretical development, but the algorithm is intended for application and demonstrated as a heuristic to be applied for solving the real-world problem of classifying research with respect to its placement on the theoretical-applied research continuum.

ticle and the research type continuum value of journals will aid researchers in finding appropriate sources of desired research types and also assist authors and editors in placing articles into publication forums that will lead to optimal utilization of the research results [9]. Being able to effectively determine an article's research type will hopefully lead to additional bibliometric-oriented research, applying the STAR methodology to new types of bibliometric analysis.

References

[1] Stringer, M.J., Sales-Pardo, M. and Amaral, L.A.N. (2008) Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS ONE*, **3**, 1-8.

[2] García-Crespo, Á., Gómez-Berbís, J.M., Colomo-Palacios, R. and García-Sánchez, F. (2011) Digital Libraries and Web 3.0. The Callimachus DL Approach. *Computers in Human Behavior*, **27**, 1424-1430. http://dx.doi.org/10.1016/j.chb.2010.07.046

[3] Lee, J.Y., Kim, H. and Kim, P.J. (2010) Domain Analysis with Text Mining: Analysis of Digital Library Research Trends Using Profiling Methods. *Journal of Information Science*, **36**, 144-161. http://dx.doi.org/10.1177/0165551509353251

[4] Ma, J., Xu, W., Sun, Y., Turban, E., Wang, S. and Liu, O. (2012) An Ontology-Based Text Analytics Method to Cluster Proposals for Research Project Selection. *Transactions on Systems*, *Man*, *and Cybernetics—Part A*, **42**, 784-790. http://dx.doi.org/10.1109/TSMCA.2011.2172205

[5] Hjørland, B. (2012) Is Classification Necessary after Google? *Journal of Documentation*, **68**, 299-317. http://dx.doi.org/10.1108/00220411211225557

[6] Tseng, Y., Chang, C., Rundgren. S.C. and Rundgren, C. (2010) Mining Concept Maps from News Stories for Measuring Civic Scientific Literacy in Media. *Computers & Education*, **55**, 165-177. http://dx.doi.org/10.1016/j.compedu.2010.01.002

[7] Bichindaritz, I. and Akkineni, S. (2006) Concept Mining for Indexing Medical Literature. *Engineering Applications of Artificial Intelligence*, **19**, 411-417. http://dx.doi.org/10.1016/j.engappai.2006.01.009

[8] Ma, L. (2012) Principles of Classification. ALCTS Webinar. http://downloads.alcts.ala.org/ce/111212_principles_of_classification_slides.pdf

[9] Kellogg, D.L. and Walczak, S. (2007) Nurse Scheduling: From Academia to Implementation or Not? *Interfaces*, **37**, 355-369. http://dx.doi.org/10.1287/inte.1070.0291

[10] Susman, G.I, and Evered, R.D. (1978) An Assessment of the Scientific Merits of Action Research. *Administrative Science Quarterly*, **23**, 582-603. http://dx.doi.org/10.2307/2392581

[11] Kohavi, R., Rothleder, N.J. and Simoudis, E. (2002) Emerging Trends in Business Analytics. *Communications of the ACM*, **45**, 45-48.

[12] Harzing, A. (2007) Journal Quality List. http://www.harzing.com/jql.htm

[13] Likert, R. (1932) A Technique for the Measurement of Attitudes. *Archives of Psychology*, **22**, 1-54.

[14] McGraw, K.O. and Wong, S.P. (1996) Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, **1**, 30-46. http://dx.doi.org/10.1037/1082-989X.1.1.30

[15] Shrout, P.E. and Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, **86**, 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

[16] Balakrishnan, R., Qui, X.Y. and Srinivasan, P. (2010) On the Predictive Ability of Narrative Disclosures in Annual Reports. *European Journal of Operational Research*, **202**, 789-801. http://dx.doi.org/10.1016/j.ejor.2009.06.023

[17] Bragge, J., Thavikulwat, P. and Töyli, J. (2010) Profiling 40 Years of Research in Simulation & Gaming. *Simulation & Gaming*, **41**, 869-897. http://dx.doi.org/10.1177/1046878110387539

[18] Porter, A.L., Kongthon, A. and Lu, J.C. (2002) Research Profiling: Improving the Literature Review. *Scientometrics*, **53**, 351-370. http://dx.doi.org/10.1023/A:1014873029258

[19] Raghuram, S., Tuertscher, P. and Garud, R. (2010) Mapping the Field of Virtual Work: A Cocitation Analysis. *Information Systems Research*, **21**, 983-999. http://dx.doi.org/10.1287/isre.1080.0227

[20] Yang, Y., Akers, L., Klose, T. and Yang, C.B. (2008) Text Mining and Visualization Tools—Impressions of Emerging Capabilities. *World Patent Information*, **30**, 280-293. http://dx.doi.org/10.1016/j.wpi.2008.01.007

[21] Conway, M. (2010) Mining a Corpus of Biographical Texts Using Keywords. *Literary and Linguistic Computing*, **25**, 23-35. http://dx.doi.org/10.1093/llc/fqp035

[22] Glänzel, W. (2012) Bibliometric Methods for Detecting and Analysing Emerging Research Topics. *El profesional de la*

información, **21**, 194-201.

[23] Seol, H., Lee, S. and Kim, C. (2011) Identifying New Business Areas Using Patent Information: A DEA and Text Mining Approach. *Expert Systems with Applications*, **38**, 2933-2941. http://dx.doi.org/10.1016/j.eswa.2010.06.083

[24] Fitzgerald, M. (2012) Introducing Regular Expressions. O'Reilly Media, Sebastopol.

[25] Kepes, S., Banks, G.C., McDaniel, M. and Whetzel, D.L. (2012) Publication Bias in the Organizational Sciences. *Organizational Research Methods*, **15**, 624-662. http://dx.doi.org/10.1177/1094428112452760

[26] Anzanello, M.J., Albin, S.L. and Chaovalitwongse, W.A. (2012) Multicriteria Variable Selection for Classification of Production Batches. *European Journal of Operational Research*, **218**, 97-105. http://dx.doi.org/10.1016/j.ejor.2011.10.015

[27] Bambenek, J. and Klus, A. (2009) grep Pocket Reference. O'Reilly Media, Sebastopol.

[28] Elliot, T. (2011) Business Analytics vs Business Intelligence? http://timoelliott.com/blog/2011/03/business-analytics-vs-business-intelligence.html

[29] Storey, V.C. (2005) Comparing Relationships in Conceptual Modeling: Mapping to Semantic Classifications. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1478-1489. http://dx.doi.org/10.1109/TKDE.2005.175

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.