Available online at www.sciencedirect.com

ScienceDirect

Technological Forecasting & Social Change 75 (2008) 186–202

Technological Forecasting and Social Change

Literature-related discovery (LRD): Methodology ☆

Ronald N. Kostoff [a,*], Michael B. Briggs [b], Jeffrey L. Solka [c], Robert L. Rushenberg [d]

[a] *Office of Naval Research, 875 N. Randolph St., Arlington, VA 22217, USA*
[b] *Arlington, VA 22204, USA*
[c] *Naval Surface Weapons Center Dahlgren Division, Dahlgren, VA 22448-5100, USA*
[d] *DDL-OMNI Engineering, LLC, 8260 Greensboro Drive, McLean, VA 22201, USA*

Received 30 July 2007; received in revised form 26 October 2007; accepted 31 October 2007

Abstract

Literature-related discovery (LRD) is linking two or more literature concepts that have heretofore not been linked (i.e., disjoint), in order to produce novel, interesting, plausible, and intelligible knowledge. LRD has two components: Literature-based discovery (LBD) generates potential discovery through literature analysis alone, whereas literature-assisted discovery (LAD) generates potential discovery through a combination of literature analysis and interactions among selected literature authors. In turn, there are two types of LBD and LAD: open discovery systems (ODS), where one starts with a problem and arrives at a solution, and closed discovery systems (CDS), where one starts with a problem and a solution, then determines the mechanism(s) that links them.

The generic methodology for identifying potential discovery candidates through ODS LRD, focusing mainly on its ODS LBD component, is described in this paper. A comprehensive flow chart showing the details of our systematic potential discovery generation process, including the evolution of the flow chart steps through each of the studies performed, is presented. Also shown is a vetting procedure that insures potential discoveries claimed are potential discoveries realized. The semantic filters that replace the numerical filters of other ODS LBD approaches are overviewed. The rationale for addressing the five topics studied (Raynaud's Phenomenon (RP), Cataracts, Parkinson's Disease (PD), Multiple Sclerosis (MS), and Water Purification (WP)) is summarized.

Keywords: Literature-Based Discovery; Text Mining; Information Retrieval; Clustering; Semantic Filters

☆ DISCLAIMER: (The views presented in this paper are those of the authors, and do not represent the views of the Department of the Navy or any of its components, or DDL-OMNI, LLC).

* Corresponding author.
E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).

0040-1625/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.techfore.2007.11.010

1. A systematic approach for accelerating discovery and innovation

1.1. Process requirements for discovery

The purpose of this paper is to identify characteristics of potential discovery, and to present a generic approach for targeting potential discovery. Equally important, as demonstrated by some of the problems in the Background section of the introductory paper [1], a vetting approach is described that insures verification of claimed potential discovery.

All LBD/LAD types share the common feature that more than one literature is required to address the problem of interest. If only one literature were required, then the solution would have been discovered by the producers and readers of that literature. What properties should these multiple literatures have for credible LBD/LAD?

- All the literatures contributing toward the solution of the problem should be complementary. Each literature should contain unique information that contributes to the total problem solution, and without each literature's unique contribution the overall problem cannot be solved.
- All the literatures should be disjoint. Otherwise, an individual's or group's knowledge of all literatures would eliminate literature-based discovery, since the information contributing to the solution could be pieced together by one individual.
- All these literatures should be as comprehensive as possible; otherwise, the disjoint-ness assumption may be a consequence of the limited literature selected, and may not be valid.
- All these literatures with unique information must be linked to form a whole that is greater than the sum of its parts.

The first author's text mining efforts over the past decade have been focused on developing methods to systematically access external sources of information that could contribute to problem solving for specific technical disciplines, technologies, systems, operations, or technical problems in general. Our group has applied text mining to assessing the technical structure and infrastructure of 1) single technologies [e.g., nanotechnology, anthrax] [2,3] and 2) country portfolios of myriad technologies [China–India] [4–7]. These methods have been integrated with the discovery literature characteristics above to form a systematic approach for accelerating discovery.

In particular, we have developed a generic approach to systematic acceleration of ODS LRD, and have applied six variants of this approach (mainly ODS LBD variants) to five problems: four medical (RP, cataracts, PD, MS) and one physical science (WP). After summarizing the generic approach, we will proceed in succeeding papers to the details of the approach and the potential discoveries made on the five problems.

1.2. Summary of generic approach to ODS LRD

1. Retrieve core literature of target problem
   - Generate query for core literature
   - Enter query into database search engine and retrieve core literature
2. Characterize core literature
   - Obtain technical infrastructure of core literature (key researchers, Centers of Excellence) through bibliometrics

- Obtain technical structure of core literature (pervasive thrusts, relationships among thrusts) through computational linguistics. Specifically, cluster core literature records to identify key technical thrust areas that characterize the core literature

3. Expand core literature
   - Generalize query (e.g., for water purification problem, generalize "water purification" as query term to "purification" as query term) for each of the core literature thrust areas obtained in the computational linguistics (clustering) step above
   - Identify and retrieve literatures related directly and indirectly to each core literature thrust area, to insure that potential discovery will impact all the major thrust areas that characterize the core literature
4. Generate potential discovery
   - Restrict classes of solutions. For example, in the RP problem, restrict solutions to non-drugs.
   - Examine all records in expanded literature that fall within the restricted classes.
   - For records that appear to contain potential discovery, perform vetting procedure as described later.

At this point, two general paths can be followed. In ODS LBD, the expanded literature is analyzed by different means for potential discovery candidates. The examples in this Special Issue will show six different, yet related, analytical approaches used successfully. In ODS LAD, the authors of the expanded literature are used to generate potential discovery candidates. The examples in this Special Issue will include one approach for ODS LAD.

1.3. Outline of generic approach to ODS LRD

We now proceed to examine the generic approach details further. Fig. 1 contains a schematic of our generic text mining approach to ODS LRD. The inner circle represents the core literature of the problem to be solved. In the example for Fig. 1, the problem to be solved is identifying 'improved' alternatives to existing water purification technologies, where 'improved' could encompass any combination of lower cost, lower energy use, lower maintenance, higher reliability, lighter weight, and improved modularity for field assembly. Thus, the core literature is the existing (more or less commonly accepted) water purification literature. The annular region between the inner and outer circles represents literatures related directly and indirectly to the core literature.

1.3.1. Front-end

1.3.1.1. Step 1. The front-end component (summarized to the left of the figure) contains two major steps: characterization of the core literature (Step 1) and characterization of the expanded literature, including identification of technical experts associated with this literature (Step 2). In Step 1, a query to retrieve the core literature is developed iteratively [8]. Once the core literature has been retrieved with this query, it is subject to text mining [9]. Bibliometrics provides the technical infrastructure (key authors/institutions/countries/journals, etc) of the core literature, e.g., [2], and computational linguistics (typically, some document and/or phrase clustering mechanism) provides the technical structure (technical thrusts, hierarchical taxonomies) of the core literature, e.g., [9]. Step 1 reflects the scope of many of our mono-technology text mining studies to date.

The criticality of Step 1 cannot be overemphasized. The core literature represents the starting point for the expansion processes. The derived expanded literature determines the pool of discovery candidates.

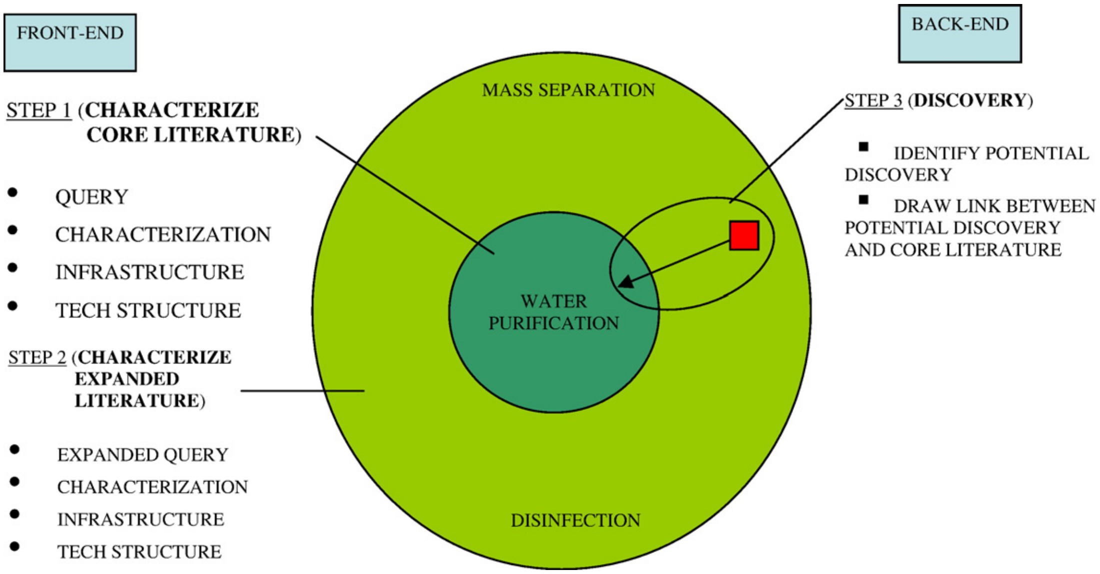


Fig. 1. The discovery process presented in the present paper is divided into two components, a front-end and a back-end.

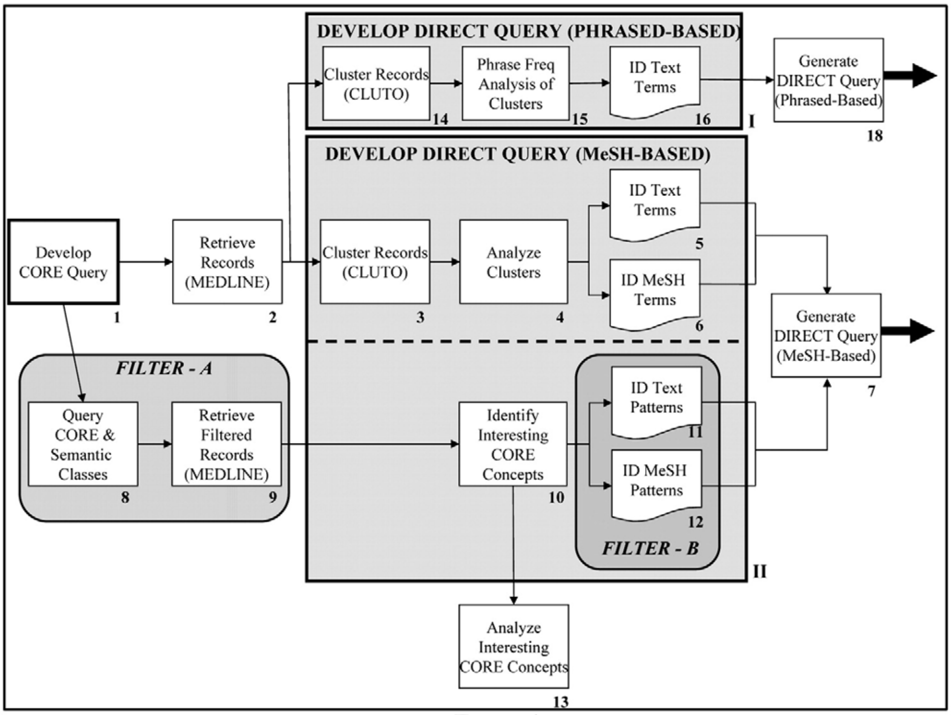
Any gaps in the core literature will be reflected as gaps in the pool of potential discovery candidates. Therefore, it is imperative that the core literature be as complete and comprehensive as possible for the discovery application.

In addition, for core literature retrieval, extensive exploitation of co-occurrence phenoma across many attributes should be made. For discovery purposes in particular, techniques that specifically exploit the underlying semantic structure of the core literature should also be used, in addition to strictly co-occurrence techniques. The first author has made extensive use of factor analysis in understanding the semantic/ conceptual structure of retrieved literatures, and has made less formal use of factor analysis for query refinement of the core literature. The factor matrix filtering technique [10] was developed to exploit the underlying semantic structure of a retrieved literature for the purpose of identifying high technical content phrases based on the strength of their contribution to semantic concepts. This is another approach for selecting new query terms.

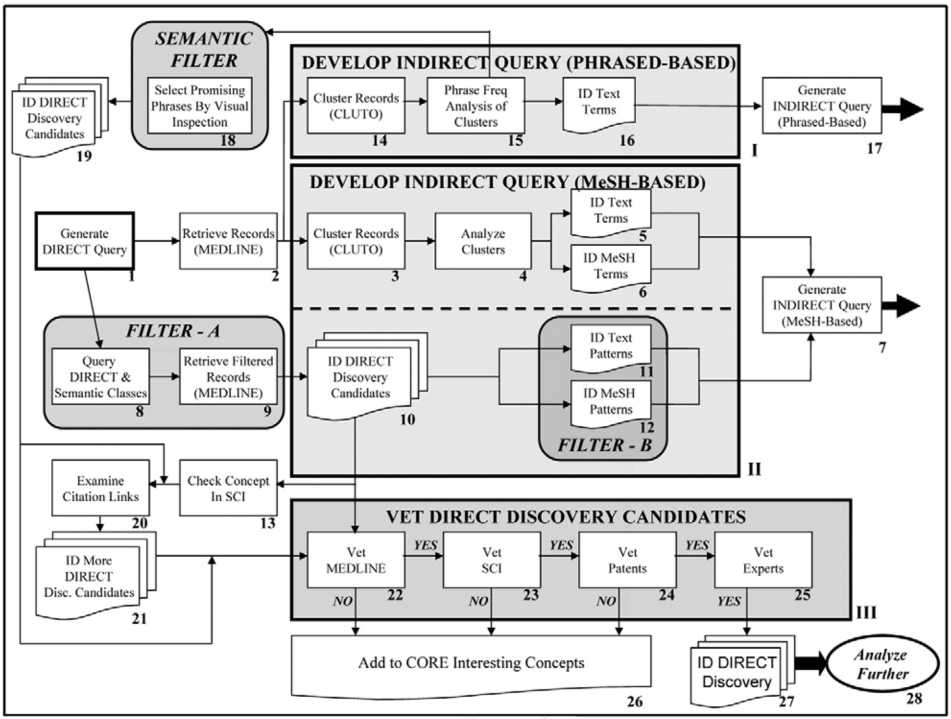
We have also made extensive use of citations for identifying additional relevant records. If a relevant record has been identified by, for example, the keyword approach, then the following may be explored to identify additional relevant records:

- A. Documents in the References section of the relevant record
- B. Documents that cite the relevant record
- C. Documents that share one of more References with the relevant record
- D. Documents in the References section of the new relevant records identified in A, B, and/or C.
- E. Documents that cite the new relevant records identified in A, B, and/or C.
- F. Documents that share one or more References with the new relevant records identified in A, B, and/or C.

So far, our focus has been on items A, B, and C. For future studies, we plan to incorporate items D, E, and F.


DEVELOP DIRECT QUERY (PHRASED-BASED)
Cluster Records (CLUTO)
14
Phrase Freq Analysis of Clusters
15
ID Text Terms
16
I
Generate DIRECT Query (Phrased-Based)
18
DEVELOP DIRECT QUERY (MeSH-BASED)
Develop CORE Query
1
Retrieve Records (MEDLINE)
2
Cluster Records (CLUTO)
3
Analyze Clusters
4
ID Text Terms
5
ID MeSH Terms
6
Generate DIRECT Query (MeSH-Based)
7
FILTER - A
Query CORE & Semantic Classes
8
Retrieve Filtered Records (MEDLINE)
9
Identify Interesting CORE Concepts
10
ID Text Patterns
11
ID MeSH Patterns
12
FILTER - B
II
Analyze Interesting CORE Concepts
13

Frame 1


SEMANTIC FILTER
Select Promising Phrases By Visual Inspection
18
ID DIRECT Discovery Candidates
19
DEVELOP INDIRECT QUERY (PHRASED-BASED)
Cluster Records (CLUTO)
14
Phrase Freq Analysis of Clusters
15
ID Text Terms
16
I
Generate INDIRECT Query (Phrased-Based)
17
DEVELOP INDIRECT QUERY (MeSH-BASED)
Generate DIRECT Query
1
Retrieve Records (MEDLINE)
2
Cluster Records (CLUTO)
3
Analyze Clusters
4
ID Text Terms
5
ID MeSH Terms
6
Generate INDIRECT Query (MeSH-Based)
7
FILTER - A
Query DIRECT & Semantic Classes
8
Retrieve Filtered Records (MEDLINE)
9
ID DIRECT Discovery Candidates
10
ID Text Patterns
11
ID MeSH Patterns
12
FILTER - B
II
Examine Citation Links
20
Check Concept In SCI
13
ID More DIRECT Disc. Candidates
21
VET DIRECT DISCOVERY CANDIDATES
Vet MEDLINE
22
YES
Vet SCI
23
YES
Vet Patents
24
YES
Vet Experts
25
NO
NO
NO
YES
III
Add to CORE Interesting Concepts
26
ID DIRECT Discovery
27
Analyze Further
28

Frame 2

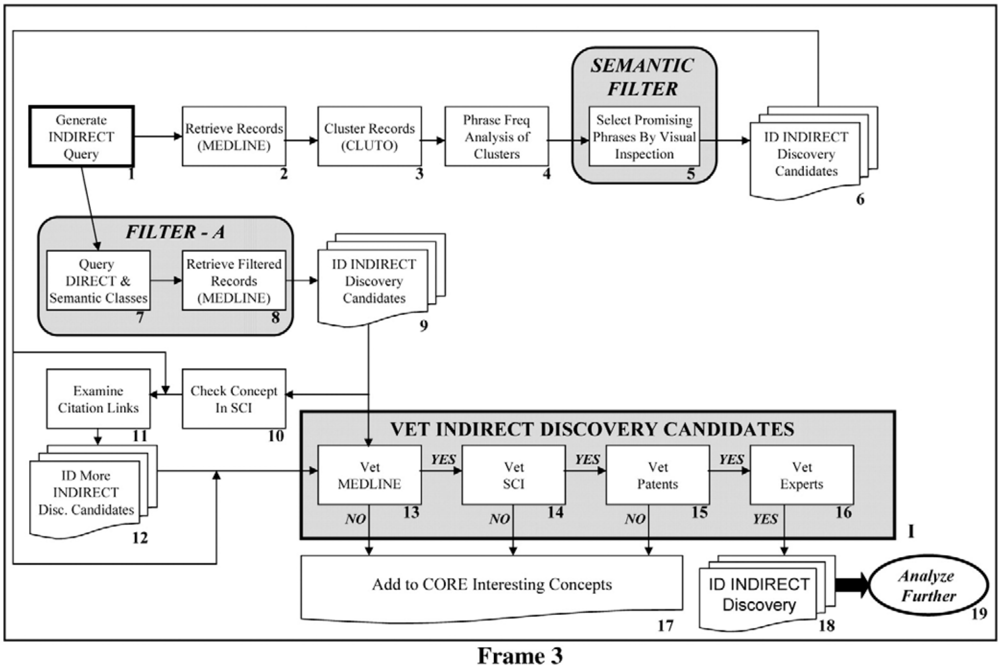

SEMANTIC FILTER
Generate INDIRECT Query 1
Retrieve Records (MEDLINE) 2
Cluster Records (CLUTO) 3
Phrase Freq Analysis of Clusters 4
Select Promising Phrases By Visual Inspection 5
ID INDIRECT Discovery Candidates 6
FILTER - A
Query DIRECT & Semantic Classes 7
Retrieve Filtered Records (MEDLINE) 8
ID INDIRECT Discovery Candidates 9
Check Concept In SCI 10
Examine Citation Links 11
ID More INDIRECT Disc. Candidates 12
VET INDIRECT DISCOVERY CANDIDATES
Vet MEDLINE 13
YES
Vet SCI 14
YES
Vet Patents 15
YES
Vet Experts 16
NO
NO
NO
YES
I
Add to CORE Interesting Concepts 17
ID INDIRECT Discovery 18
Analyze Further 19
Frame 3

Fig. 2. Medical discovery process flow chart.

One of the major differences between our techniques and those of other ODS LBD researchers is the size and scope of our core literature query. Some of our queries contain hundreds of terms, especially those for non-medical topics (because of the MeSH controlled vocabulary in the Medline database, medical core literature queries can be very short). We strongly believe that the most sophisticated analyses cannot compensate for incomplete data, and strive to make our retrievals as complete and comprehensive as possible.

Some of the more formal techniques (such as latent semantic indexing (LSI) [11,12]) that exploit the semantic structure should also be examined for core literature definition. It remains to be demonstrated in practice whether these techniques that exploit semantic structures offer more capabilities than properly conducted attribute co-occurrence techniques, e.g., [13].

The first author has also proposed latent feature indexing (LFI), of which LSI is a special case, for retrieving additional relevant records. In LFI, many of the bibliometric field features (e.g., authors, institutions, countries, citations) can be substituted for the words or phrases in LSI, and the mathematical operations remain the same. Probably the most promising for enhanced relevant record retrieval is latent citation indexing (LCI), where references are used as features and substituted for words or phrases in the LSI algorithms.

A multitude of information retrieval techniques have been examined for more than a decade at the TREC conferences [http://trec.nist.gov/pubs.html], and the interested reader is advised to examine the proceedings of these conferences for the state-of-the-art in information retrieval approaches.

1.3.1.2. Step 2. In Step 2, the query developed in Step 1 is generalized and expanded, again iteratively. This expanded query will retrieve records from literatures related directly and indirectly to the core literature. Insights and principles from these disparate literatures/technical disciplines can be extrapolated to solve problems of the core literature. *To insure that all the major themes in the core literature are addressed in the expansion process, the core literature is clustered to identify these themes*. We have used a variety of approaches for the clustering, including document clustering, auto- and cross-correlation mapping of words/phrases, factor matrix analysis of words/phrases, and manual assignment of words/phrases/records to clusters. This is another major difference between our work and that of other ODS LBD researchers. *Clustering insures that all thrusts will be represented in the expanded literature.* We don't know anyone who has reported the use of clustering to identify the intermediate literature thrusts.

Thus, in the example on Fig. 1, the core water purification literature query is expanded to cover/retrieve all of mass separation and disinfection documents. Insights from very disparate mass separation and disinfection approaches can then be extrapolated to solve problems in water purification. Details on query expansion can be found in [13].

1.3.2. Back-end

The back-end component contains the discovery step, which itself contains two sub-components. The first sub-component is identification of potential discovery candidates from the expanded literature, and the second sub-component is drawing the linkages between the potential discovery candidates and the core literature. Ordinarily, ODS LBD researchers apply numerical filters at this point, to reduce the number of concepts that have to be examined for potential discovery linkages. We differ from the remainder of the ODS LBD research community in that *we do not use numerical filters to reduce the number of records that have to be examined*.

Recently, in our newly-developed streamlined medical studies approaches, we have been using two types of *semantic filters* for the medical studies, where the MeSH taxonomy in MEDLINE allows semantic classes to be defined. We restrict potential solutions to selected semantic classes (e.g., potential treatments are non-drug only), and use query term combinations (mainly of MeSH terms) characteristic of prior discovery in the semantic class (e.g., non-drug treatments in the core literature). *We believe these semantic filters are inherently superior to the numerical filters for potential discovery selection and identification!!!*

As will be shown in this section, there are many ways to identify potential discovery candidates, and to draw the subsequent linkages. These techniques differ mainly by the approach mechanics and the types of people used to identify the discovery and innovation candidates. The two main discovery approach types (literature-based, literature-assisted) are described now.

1.3.2.1. ODS literature-based discovery. ODS LBD is based strictly on analysis of literatures related directly and indirectly to the core literature. ODS LBD is useful in the planning and concept identification phases of the science and technology (S&T) development cycle. The literature-based approach can be viewed as a very sophisticated type of literature survey, and represents a somewhat different way of doing business for most S&T sponsoring agencies, researchers, and technical journals. *Done properly, ODS LBD has the potential of generating orders of magnitude more discovery than what has been reported so far in the ODS LBD literature (as this Special Issue will demonstrate).*

1.3.2.2. ODS literature-assisted discovery. We can identify technical experts associated with the 'external' directly and indirectly-related disciplines, and then have them focus their expertise on solving problems of interest from the 'internal' disciplines. Assembling of experts from multiple disciplines connected to a target discipline of interest could be done through workshops, panels, solicitation of proposals from various disciplines, etc. This literature-assisted people-based approach could easily be incorporated into most S&T sponsoring agencies' existing operational procedures.

However, in some applications, proper handling of the infusion of large numbers of concepts and insights from disparate disciplines will acquire the characteristics of 'disruptive technologies'. For example, in the use of ODS LAD to broaden the solicitation of proposals from disparate disciplines, a large number of reviewers may be required to handle the increased volume of proposals, and many of these reviewers will need to be from disparate disciplines.

Thus, the differences between paths ODS LBD and ODS LAD above are in the 'back-end', in 1) how the linkages between the 'external' and 'internal' disciplines are made, and 2) who makes the linkages.

The ultimate goal should be incorporation of both approaches in parallel, to exploit the strengths of each approach while eliminating the weaknesses. This synergy would provide the *comprehensiveness and objectivity* of the people-assisted literature-based approach coupled with the *interaction and feedback* of the literature-assisted people-based approach [14].

1.4. Details of generic approach to ODS LBD

Fig. 2 is a flow chart that describes a comprehensive process for generating medical ODS LBD. Because of resource limitations, none of our completed medical studies used the full process, although the MS study came closest. This did not present a problem in practice, since for all topics studied a large amount of potential discovery was generated. For the demonstration-type studies in this Special Issue, the potential discovery generated was more than adequate for all cases. We will describe the steps in the flow chart, and identify those that were used in each study. The steps on Fig. 2 have been numbered for ease of presentation.

1.4.1. Description of general medical studies approach

The implementation of the process for retrieving the core and expanded literatures portrayed schematically in Fig. 1 consisted of three steps. In the first step, a core literature was retrieved. In the second step, this core literature was clustered, and was expanded to generate a literature directly related to, but disjoint from, the core literature. In the third step, the directly related literature was expanded to generate a literature indirectly related to, and disjoint from, the core literature, as well as disjoint from the directly related literature.

Fig. 2 describes the total process in more detail, and consists of three frames. The first frame starts with the development of a core literature query, describes the development of a directly related literature query and the identification of interesting core literature concepts (potential treatments for the medical problem of interest already identified), and ends with the generation of the directly related literature query. The second frame describes the development of an indirectly related literature query and the identification of potential discovery from the directly related literature, and ends with the generation of the indirectly related literature query. The third frame describes the identification of potential discovery from the indirectly related literature. Each frame has its own numbering system.

1.4.1.1. Frame 1

1.4.1.1.1. Development of the directly related literature query. In Frame 1, Step 1 is development of the core query. It is based on a review of the background material for the medical topic being addressed, and discussions with experts in the medical topic. In all cases studied, the core queries were extremely simple, due to the existence of the MeSH terms. A single MeSH term tended to cover the core literature. Use of ‘cataract*, for example, covered both its use as a MeSH term and as a text term, and tended to retrieve the total core cataracts literature.

Step 2 is use of the query in the appropriate database search engine for a selected time frame to retrieve records for analysis. Medline was used as the main database for all studies, although the Science Citation Index (SCI) [15] was used at later stages for expanding discovery through the use of citation linkages.

Step 3 is grouping of the core literature concepts to identify the main medical thrusts, and insure these thrusts are represented in the expansion of the query for retrieving directly related literatures. Document clustering was used as the main core literature grouping technique, where documents are segregated into groups based on text similarity. The CLUTO software package [16] was used for document clustering. In some of the later studies, other core literature concept grouping techniques were used as supplements to the document clustering, including phrase autocorrelation mapping and factor matrix analyses.

Step 4 was analysis of the clusters, or groups obtained by other approaches. The main theme of each group was generated, and the key phrases within each group were identified. Additionally, only groups that focused on biomedical mechanisms and phenomena were selected for the final query expansion, since the potential treatments were focused on impacting these mechanisms and phenomena.

Step 5 was identification of the important text terms in the Abstracts, and Step 6 was identification of key MeSH terms. These first six steps constitute one of the two inputs in determining the directly related literature query. The next six steps constitute the second input in determining the directly related literature query (as well as identifying interesting core literature concepts).

1.4.1.1.2. Identification of interesting core concepts. Steps 8 and 9 represent the application of the first semantic filter. The purpose is to extract the sub-set of core records that have pre-determined desired characteristics that we would like to see in potential discoveries. In the present group of medical studies, the classes of potential solutions were restricted to non-drug approaches, and these classes were defined (for the last three medical studies) by selection of appropriate MeSH terms that represented non-drug approaches (e.g., medicinal plants, phytotherapy, etc). In future studies, the classes could be expanded to include drugs, environmental effects, etc.

In Step 8, the core query is intersected with the semantic classes to generate a filtered core query. For example, if ‘cataract*’ is the unfiltered core query for retrieving the cataracts core literature, and if our semantic class of interest was ‘medicinal plants’, then the filtered core query would be (cataract* AND “medicinal plants). In Step 9, the filtered core query is inserted into the PubMed search engine, and the semantically filtered core records are retrieved.

In Step 10, the retrieved records (typically a few hundred) are read, and those that appear interesting (potential treatment for the medical problem being examined) are identified.

1.4.1.1.3. Development of the directly related literature query (cont'd). Steps 11 and 12 constitute implementation of the second semantic filter. Text and MeSH patterns (mainly combinations) characteristic of the ‘interesting’ core records are identified, for future use in helping to formulate the directly related literature query. In parallel, Step 13 is conducted to analyze the interesting core literature concepts. For example, are there classes of concepts that will allow generalization beyond individual interesting concepts and might offer further insights into treatment mechanisms?

In Step 7, the outputs from Steps 5, 6, 11, and 12 are combined to generate the directly related literature query. This query will reflect the thrusts defined by the different grouping procedures, and may include combinations of terms that reflect the patterns in the ‘interesting’ core literature records. Thus, while numerical filters are not employed as in other ODS LRD techniques, two semantic filters are used to narrow the scope of the retrieval, and sharpen the focus on promising concepts.

1.4.1.2. Frame 2

1.4.1.2.1. Development of the indirectly related literature query; identification of potential discovery candidates from the directly related literature. In Frame 2, Step 1 is generation of the directly related literature query, defined in Frame 1. Step 2 is insertion of this query into the PubMed search engine to retrieve the directly related literature records from Medline. Steps 3–6 are analogous to Steps 3–6 from Frame 1. Steps 8–9, application of the first semantic filter, are analogous to Steps 8–9 from Frame 1. Step 10 is identification of potential discovery candidates from the filtered directly related literature. The common feature Step 10 in Frame 2 shares with Step 10 in Frame 1 is that both search for interesting records/concepts. In Frame 1, these interesting concepts are not potential discoveries, since they are in the core literature, whereas in Frame 2 they are potential discovery candidates, since they are not in the core literature.

Steps 11 and 12 in Frame 2 are analogous to their counterparts in Frame 1. The combination of Steps 5, 6, 11, and 12 in Frame 2 to generate the indirectly related literature query is analogous to the similar process to generate the directly related literature query from Frame 1. The remaining steps in Frame 2 have no counterpart in Frame 1.

Steps 1–12 resulted in a) generation of a query for retrieving the indirectly related literature and b) identification of potential discovery candidates based on the Medline database and use of MeSH-based semantic filters. Sub-sets of these steps were used for the Cataracts, PD, and MS medical studies. Steps 14–19 use a different type of semantic filter to identify potential discovery candidates, and these steps formed the basis of the RP medical study. In Step 18, the analyst inspects all the phrases visually, and selects those from desired classes. In Step 19, the analyst then examines the records in which those phrases occur, and identifies potential discovery candidates.

The main difference between the two processes is that Step 18 involves visual inspection of all phrases generated by the phrase frequency analyzer, while Step 8 uses MeSH filtering for selecting the semantic classes desired. Step 18 is obviously much more labor intensive than Step 8. The semantic filtering is performed by the analyst selecting phrases of the class desired for the solution. Thus, if the analyst is interested in non-drug approaches to addressing the medical problem of interest, the analyst will select only those phrases that represent non-drug approaches for further analysis.

The benefit of the approach represented by Steps 14–19 is the independence of the process from third-party indexers and omissions of indexing. In theory, all records that contain phrases from the desired semantic classes will be accessed. The deficiencies of this approach are that applicable records that do not contain the desired phrases in their Abstract will not be accessed (whereas MeSH-based records would in theory access these records), and the labor intensity of the process. The combination of these two approaches, as depicted in Frame 2, would in theory eliminate the weaknesses of each approach and enhance the strengths. We did not combine the two approaches for any one study because of resource limitations.

1.4.1.2.2. Identifying potential discovery candidates through citation relations. Steps 1–12 and 14–19 represent two approaches for identifying potential discovery candidates that were used in part by different studies reported in this Special Issue. Steps 13, 20, and 21 represent another approach for identifying potential discovery candidates. After potential discovery candidates have been identified from Steps 10 and/or 19, their records are located in the SCI. Then, citation linkages are used to identify other potential discovery candidates.

Specifically, approaches A, B, and C below were explored to identify additional potential discovery candidates, and approaches D, E, and F will be explored in future studies to identify additional discovery candidates.

- A. Documents in the References section of the relevant record
- B. Documents that cite the relevant record
- C. Documents that share one of more References with the relevant record
- D. Documents in the References section of the new relevant records identified in A, B, and/or C.
- E. Documents that cite the new relevant records identified in A, B, and/or C.
- F. Documents that share one or more References with the new relevant records identified in A, B, and/or C.

We had only begun to scratch the surface of this relational citation approach; it was employed only at the very end of the RP study and at the end of some of the other studies as well. It appeared to offer enormous potential for uncovering additional potential discovery candidates.

1.4.1.2.3. Vetting potential discovery candidates. Irrespective of which of the above three processes were used to identify potential discovery candidates, the candidates had to be vetted before they could be considered as potential discoveries. Steps 22–25 constitute the vetting process that was used.

The purpose of our vetting procedures is to insure that what we report as potential discovery has not been found in the literature previously (i.e., no prior art), and obeys the criteria for discovery set forth at the beginning of the Introductory paper: *linking two or more literature concepts that have heretofore not been linked (i.e., disjoint), in order to produce novel, interesting, plausible, and intelligible knowledge*. If a concept has been found in the literature previously, but we believe its reporting would accelerate its development, we might report it as a potential innovation candidate. We have instituted a four step vetting process that balances thoroughness with pragmatism.

The first step (Step 22) is to check for appearance of the potential discovery concept in the core target problem research literature. How do we define this literature? There are two issues here: the database(s) selected as source material, and the technical scope of the problem. For database(s) selection, ideally, every research document published globally in the core problem area would constitute this core literature source(s). The practical compromise we have made is to define the core literature source(s) for the core target problem literature as the SCI and MEDLINE. While we believe this is a bare minimum core literature requirement to search for prior art/science, some examples overviewed in the Background section of the introductory paper and shown in more detail in [17,18] illustrate that even this threshold requirement was not met before potential discovery was claimed in the published literature. The technical scope is subjective, and flows from the original problem definition.

In this first vetting step, we check operationally for the intersection of the core target problem literature with the potential discovery literature. If the intersection is a null set, the first check is successful. Thus, if we check whether Fish Oil is a potential discovery for RP, we might use the query Fish Oil (or its many specific variants) and RP (or its variants), and see whether any records are retrieved. The real issue here is how broadly or narrowly we define the core target problem literature and the potential discovery concept literature. The breadth of definition could determine whether we have generated discovery, innovation, or nothing. For example, Fish Oil may or may not be a discovery for treating RP, depending on whether we define the core RP literature to include or exclude the Peripheral Vascular Disease literature.

The second vetting step (Step 23) could be viewed as a continuation of the first step. We go beyond simple intersection to see whether there are citation linkages between the potential discovery concept and the core target problem literature that would indicate researchers were aware of the linking between these literatures previously. Citation linkages were only used for the SCI database, since this database is structured to exploit citation relationships. There are many types of citation linkages (citing papers, cited papers, papers that share common references, papers that share common citing papers, etc). Depending on how far we plan to proceed with a potential discovery (e.g., do we want to patent the potential discovery), we check at least the citing papers for linkages between the concept literature and the problem literature.

The third vetting step (Step 24) is checking the patent literature. This is more difficult than the first step because of the breadth and scope of the claims in each patent. We read the claims thoroughly to check whether a linkage has been established, or whether the inventor has generated unsubstantiated claims. Most of the prior art exclusion of potential discovery candidates by the vetting process has occurred in this patent step.

Why does this exclusion occur mainly in the patent step, and how can it be overcome? In the ODS LBD medical studies reported in this Special Issue, and in essentially all the ODS LBD medical studies reported in the literature, the potential discovery algorithms are focused on Medline. This allows exploitation of the MeSH taxonomy capabilities. We exclude the medical problem core literature in Medline as part of the algorithm, which (except for some of the MeSH anomolies noted previously) essentially eliminates the vetting problem in Medline. Since there is much overlap between the laboratory research in Medline and SCI records, most (not all) prior art in the SCI will also be eliminated by the exclusion portion of the potential discovery algorithms.

The patent literature is very different from the SCI and Medline. Many of the authors are different; many people patent rather than publish, among other differences. Therefore, core literature concepts that were excluded from Medline (and effectively SCI) by the algorithms could (and do) occur in the patent literature. The obvious method for insuring that core literature concepts are excluded from the patent literature (and SCI) is to apply the potential discovery algorithms, with their core literature exclusion component, to the patent literature and SCI as well as to Medline. This approach would complicate the analytical procedure, since the SCI and patent literature do not have the MeSH capability, and the simplifications offered by the MeSH capability could not be exploited. For the proof-of-principle demonstrations reported in this Special Issue, we have chosen to exploit the MeSH capabilities and devote the extra effort required to vet the patent and SCI literatures.

All vetting steps are run serially. Once the first three vetting steps have been taken successfully, we then have the potential discovery candidate concepts examined by technical experts (Step 25). We access two types of technical experts: those expert in the core target problem literature (e.g., RP), and those expert in the potential discovery concept literature (e.g., Fish Oil). We ask the experts in the core target problem literature whether the potential discovery concept is indeed discovery (i.e., have they seen it before in the target problem context), and we ask the experts in the potential discovery concept literatures whether the concept could be extrapolated to the target problem. If we report potential discovery concepts that have been vetted partially, we state that fact.

How do our vetting procedures compare with those used by the remainder of the ODS LBD research community? We see little discussion of vetting in the open ODS LBD literature, and therefore it is difficult to compare our vetting approaches with others on the basis of published protocols. We applied the first and third vetting steps above to a number of potential ‘discoveries’ claimed in the mainstream literature, using only the terminology supplied by the authors and only the major databases, and have found that many of these potential ‘discoveries’ would have been excluded by our process [17–19].

Had we performed steps 1 and 2 only (Medline and the SCI) for our vetting procedure, we would have had substantially more potential ‘discoveries’. However, we believe that presenting such results as potential ‘discoveries’ to independent third parties would have impacted the credibility of our findings adversely, and would have cast doubt on the credibility of our whole approach. Therefore, we wanted to define discovery in the sense that it is understood by most of the technical community, and designed our vetting process to support that goal.

As shown on Frame 2, for Steps 22–24, there are two decision points. If a potential discovery candidate fails at any of these three steps, it is added to the ‘interesting’ concepts in the core literature (Step 26) defined in Frame 1 (shown in Step 13). If a potential discovery candidate passes all four decision points, it is then added to the pool of potential discovery (Step 27), and is subjected to further analysis (Step 28).

1.4.1.3. Frame 3

1.4.1.3.1. Identification of potential discovery candidates from the indirectly related literature. The steps in Frame 3 are analogous to those in Frame 2, with the exception that the steps necessary for defining a query (Steps 3, 4, 5, 6, 7, 11, and 12 in Frame 2) are eliminated in Frame 3, since we have terminated the process at the query for retrieving the indirectly related literature. Obviously, there are different levels of indirectly related literatures, and the frames could have been continued ad infinitum to identify literatures further and further removed from the core. Terminating the process at the first indirect literature represents a compromise between marginal recall and marginal effort.

1.4.2. Description of specific medical studies approaches

1.4.2.1. Raynaud's phenomenon. The RP study [20] was the first medical ODS LBD study we performed. For the RP study, we used Steps 1, 2, 14, 15, 16, 18 in sequence from Frame 1, Steps 1, 2, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28 in sequence from Frame 2, and Steps 1, 2, 3, 4, 5, 6, 11, 12, 13, 14, 15, 16, 17, 18, and 19 in sequence from Frame 3. Basically, we did not take advantage of the semantic class restrictions offered by use of the MeSH terms.

1.4.2.2. Cataracts. The Cataracts study [21] was the second medical ODS LBD study we performed. The main objective was to reduce the time required for the RP study, while still generating copious amounts of potential discovery. For this study, we used Steps 1–6, 8–13 and 7 in sequence from Frame 1, Steps 1–6, 8–12, 13, 20–28 from Frame 2, and Steps 1, 7–19 from Frame 3. We took advantage of MeSH terms as semantic filters (Filters A and B on the different Frames) to streamline the total process.

1.4.2.3. Parkinson's disease. The PD study [22] was the third medical ODS LBD study we performed. The objective was to build upon our experience with the Cataracts study, and make more extensive use of the semantic filter ‘B’ terms. For this study, we used Steps 1–6, 8–13 and 7 in sequence from Frame 1, Steps 1–6, 8–12, 13, 20–28 from Frame 2, and Steps 1, 7–19 from Frame 3.

1.4.2.4. Multiple sclerosis. The MS study [23] was the fourth medical ODS LBD study we performed. The objective was to build upon our experience with the Cataracts and PD studies, and use more of the features in the expanded flow chart depicted by Fig. 2. For this study, we used Steps 1–6, 8–13 and 7 in sequence from Frame 1, Steps 1–6, 8–12, 13, 20–28 from Frame 2, and Steps 1, 7–19 from Frame 3.

1.4.3. Description of non-medical studies approach

1.4.3.1. Overall. For the WP non-medical study [24], two separate techniques were used, and they each differed somewhat from the specific technique used for the medical studies. The two MeSH-based filters shown in Fig. 2 could not be used in the non-medical studies because of the lack of an available taxonomy for the non-medical studies.

Overall, a core literature query (consisting of hundreds of terms) was defined for WP using an iterative relevance feedback technique [8]. Then, an expanded (related) literature was defined through clustering of the core literature, and selecting/generalizing key phrases from each cluster. The expanded literature was searched by two different approaches for potential discovery candidates: Cluster Semantic Filtering (CSF) and Latent Semantic Indexing (LSI).

1.4.3.2. Cluster Semantic Filtering. Cluster Semantic Filtering was somewhat analogous to the semantic filtering approach used in the RP study. In the RP study, semantic filtering was performed on phrases from the expanded literature to identify topics of interest, whereas in the WP study, semantic filtering was performed on document clusters in the expanded literature. Specifically, the expanded WP literature was stratified using large numbers of document clusters. Each cluster was inspected visually, and those clusters that appeared to focus on desired semantic classes and novel topics were selected for more detailed analyses. All documents contained within these 'interesting' clusters were evaluated for potential discovery, and the promising candidates were subjected to the further analyses depicted on Fig. 2 (SCI linkages and vetting). It was found that the clustering served as a strong filter not only for

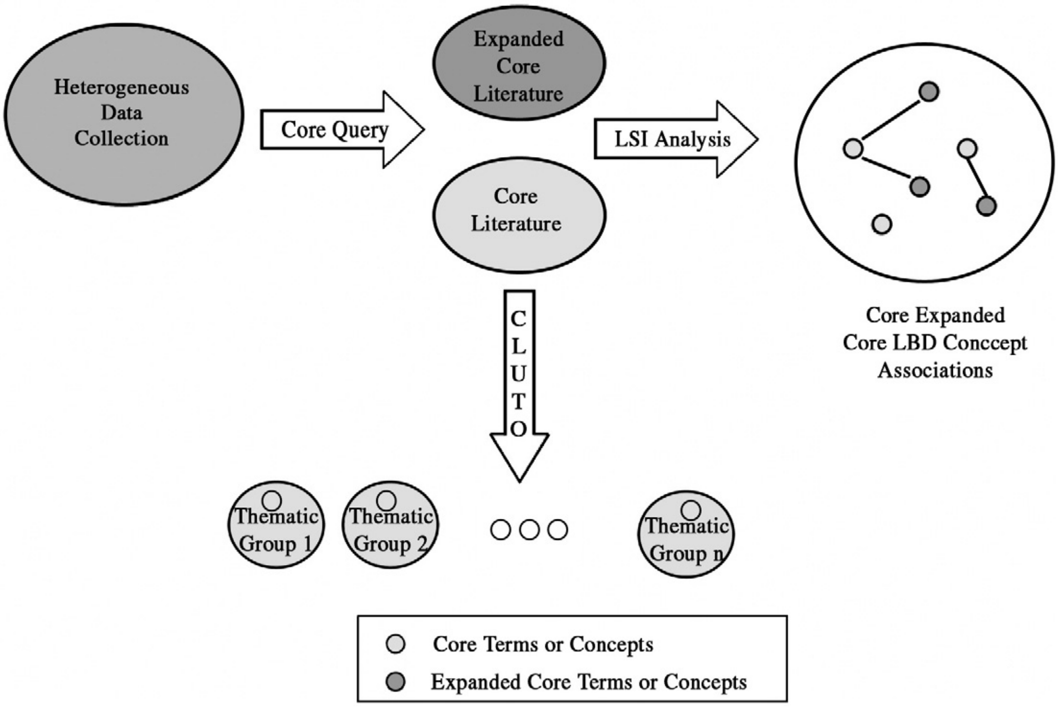


Fig. 3. LSI LBD methodology.

segregating the semantic classes, but for separating the promising discovery candidates from those much less promising. CSF had some of the manual intensity characteristic of the RP process, but was somewhat less labor intensive.

For example, one of the problems in WP is fouling of the separation membranes. This results in greater maintenance time and costs, and eventually in higher WP costs. It would be desirable to reduce membrane fouling.

We examined an expanded and clustered anti-fouling literature. One interesting cluster focused on anti-fouling properties of sponges. Since this theme seemed very disparate from the core literature, interesting, and promising, we examined it in more detail. We found there were sponges that had intrinsic anti-fouling defenses, using some combination of chemical and biological mechanisms. Potential discovery could involve extracting anti-fouling substances excreted by or intrinsic to the sponges and perhaps applying them to WP separation membranes, or studying the anti-fouling mechanisms used by the sponges and creating similar mechanisms artificially.

1.4.3.3. Latent semantic indexing. Operationally, we used a core article query to tag the core articles, and then tagged the remainder as expansion. We then clustered the core articles into thematic groups, and used terminology from these groups as 'seeds' for linking to related terms from the expanded literature. Following Gordon and Dumais [12], we then computed the ranking metrics (cosine similarity) score between the selected core terms and the expanded terms in the projected LSI space. After sorting the expanded core terms based on their cosine similarity to the selected core term, we obtained a number of interesting associations. High ranking terms were examined, and much potential discovery surfaced (including potential discovery from single frequency concepts). See Fig. 3 for a schematic of the LSI-based methodology.

1.5. Specific approaches and results

Now that the generic ODS LRD approach has been outlined, specific details of the approach for each variant used for discovery will be described, along with the results. Chronologically, the first problem addressed was the benchmark RP problem using ODS LBD, and it will be described in the next paper [20] in this Special Issue. A 2005 paper [10] outlined the clustering approach that was used to identify the main medical themes in the core literature.

The second medical problem addressed with ODS LBD was cataracts. A medical problem was selected as the second problem for two reasons: to show that the large number of potential discoveries from the first medical study (RP) were not a fluke, and to show that a 'streamlined' approach to radical discovery was possible with little loss in performance. The results of the cataracts study are reported in the second following paper [21].

A third medical problem, PD, was addressed through ODS LBD to gather further confirmatory data for the radical discovery approach. The semantic filters were applied more extensively in the PD study than in the cataracts study, and longer queries were used as well. The results of the PD study are reported in the third following paper [22].

A fourth medical problem, MS, was addressed with ODS LBD to further advance the technique, to improve the process steps depicted in the flow chart, and to 'push the envelope' on the bounds of the biomedical phenomena component of the query. The results of the MS study are reported in the fourth following paper [23].

Also, for the first time, a non-medical application (WP) was studied with the use of both ODS LBD and ODS LAD. The purpose was to show that ODS LBD and ODS LAD need not be limited to medical problems, and that much potential discovery was possible with non-medical topics as well. The results of the WP study are reported in the fifth following paper [24].

It should be emphasized that in all five topics studied, the results obtained (while substantial) are the tip of the iceberg of what is possible with adequately resourced studies. The reasons behind this statement will be included in the following papers.

References

[1] R.N. Kostoff, Literature-related discovery (LRD): introduction and background, Technol. Forecast. Soc. Change 75 (2) (2008) 165–185, doi:10.1016/j.techfore.2007.11.004.

[2] R.N. Kostoff, R.G. Koytcheff, C.G.Y. Lau, Global nanotechnology research metrics, Scientometrics 70 (3) (2007) 565–601.

[3] R.N. Kostoff, S. Morse, S. Oncu, The seminal literature of anthrax research, Crit. Rev. Microbiol. 33 (3) (2007) 171–181.

[4] R.N. Kostoff, S. Bhattacharya, M. Pecht, Assessment of China's and India's science and technology literature – introduction, background, and approach". Technol. Forecast. Soc. Change 74 (9) (2007) 1519–1538.

[5] R.N. Kostoff, D. Johnson, C.A. Bowles, S. Bhattacharaya, A.S. Icenhour, K.F. Nikodym, R.B. Barth, S. Dodbele, Assessment of India's research literature. Technol. Forecast. Social. Change 74 (9) (2007) 1574–1608.

[6] R.N. Kostoff, M. Briggs, R. Rushenberg, C.A. Bowles, A.S. Icenhour, K.F. Nikodym, R.B. Barth, M. Pecht, Chinese science and technology - Structure and infrastructure. Technol. Forecast. Soc. Change 74 (9) (2007) 1539–1573.

[7] R.N. Kostoff, M. Briggs, R. Rushenberg, D. Johnson, C.A. Bowles, S. Bhattacharaya, A.S. Icenhour, K.F. Nikodym, R.B. Barth, S. Dodbele, M. Pecht, Comparisons of the structure and infrastructure of Chinese and Indian Science and Technology. Technol. Forecast. Soc. Change 74 (9) (2007) 1609–1630.

[8] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Database tomography for information retrieval, J. Inf. Sci. 23 (4) (1997) 301–311.

[9] R.N. Kostoff, J.A. Stump, D. Johnson, J. Murday, C. Lau, W. Tolles, The structure and infrastructure of the global nanotechnology literature, J. Nanopart. Res. 8 (3–4) (2006) 301–321.

[10] R.N. Kostoff, J.A. Block, Factor matrix text filtering and clustering, JASIST 56 (9) (2005) 946–968.

[11] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.

[12] M.D. Gordon, S. Dumais, Using latent semantic indexing for literature based discovery, J. Am. Soc. Inf. Sci. 49 (8) (1998) 674–685.

[13] R.N. Kostoff, Systematic acceleration of radical discovery and innovation in science and technology, Technol. Forecast. Soc. Change 73 (8) (2006) 923–936.

[14] R.N. Kostoff, Stimulating innovation, in: Larisa V. Shavinina (Ed.), International Handbook of Innovation, Elsevier Social and Behavioral Sciences, Oxford, UK, 2003, pp. 388–400.

[15] SCI, Certain data included herein are derived from the Science Citation Index/Social Science Citation Index prepared by the THOMSON SCIENTIFIC ®, Inc. (Thomson®), Philadelphia, Pennsylvania, USA: © Copyright THOMSON SCIENTIFIC ®, 2006 All rights reserved. (2006).

[16] CLUTO—A clustering toolkit, 2007 http://www.cs.umn.edu/~cluto.

[17] R.N. Kostoff, J.A. Block, J.L. Solka, M.B. Briggs, R.L. Rushenberg, J.A. Stump, D. Johnson, T.J. Lyons, J.R. Wyatt, Literature-related discovery, ARIST. (2008).

[18] R.N. Kostoff, J.A. Block, J.L. Solka, M.B. Briggs, R.L. Rushenberg, J.A. Stump, D. Johnson, T.J. Lyons, J.R. Wyatt, Literature-related discovery. DTIC Technical Report Number ADA473438 (http://www.dtic.mil/). Defense Technical Information Center, Fort Belvoir, VA, 2007.

[19] R.N. Kostoff, Where is the discovery in literature-based discovery? in: P. Bruza, M. Weeber (Eds.), Literature-Based Discovery, Information Science and Knowledge Management Series, Springer, 2008.

[20] R.N. Kostoff, J.A. Block, J.A. Stump, D. Johnson, Literature-related discovery (LRD): potential treatments for Raynaud's Phenomenon, Technol. Forecast. Soc. Change 75 (2) (2008) 203–214, doi:10.1016/j.techfore.2007.11.005.

[21] R.N. Kostoff, Literature-related discovery (LRD): potential treatments for cataracts, Technol. Forecast. Soc. Change 75 (2) (2008) 215–225, doi:10.1016/j.techfore.2007.11.006.

[22] R.N. Kostoff, M.B. Briggs, Literature-related discovery (LRD): potential treatments for Parkinson's Disease, Technol. Forecast. Soc. Change 75 (2) (2008) 226–238, doi:10.1016/j.techfore.2007.11.007.

[23] R.N. Kostoff, M.B. Briggs, T.J. Lyons, Literature-related discovery (LRD): potential treatments for Multiple Sclerosis, Technol. Forecast. Soc. Change 75 (2) (2008) 239–255, doi:10.1016/j.techfore.2007.11.002.

[24] R.N. Kostoff, J.L. Solka, R.L. Rushenberg, J.A. Wyatt, Literature-related discovery (LRD): water purification, Technol. Forecast. Soc. Change 75 (2) (2008) 256–275, doi:10.1016/j.techfore.2007.11.009.

Dr. Ronald Neil Kostoff received a Ph. D. in Aerospace and Mechanical Sciences from Princeton University in 1967. He has worked for Bell Laboratories, Department of Energy, and Office of Naval Research (ONR). He has authored well over 100 technical papers, served as Guest Editor of four journal Special Issues, obtained two text mining system patents, and presently manages a text mining program at ONR. He is listed in Who's Who in America, 60th Edition (2006), Who's Who in Science and Engineering, 9th Edition (2006), and 2000 Outstanding Intellectuals of the 21st Century, 4th Edition, (2006).

Michael Briggs is Reserve Lieutenant Colonel currently on Active Duty at the Marine Corps Warfighting Laboratory supporting Science & Technology Experimentation. He received a B.S. and M.S. in Nuclear Engineering, and a M.E. in Radiological Health Engineering from the University of Michigan. LtCol Briggs is a Naval Flight Officer with over 24 years of service in the USMC. His expertise includes electronic warfare, space operations, communications, intelligence, text mining, and technology road mapping. He is a Registered Engineer-In-Training, a member of several professional societies, and served as the USMC representative to numerous National Defense Forums, including the Defense Science Board.

Dr. Jeffrey Solka received his B.S. in Mathematics and Chemistry from James Madison University in 1978, his M.S. in Mathematics from James Madison University in 1981, his M.S., in Physics from Virginia Polytechnic Institute and State University in 1989, and his Ph.D. in Computational Sciences and Informatics (Computational Statistics) from George Mason University under the direction of Edward J. Wegman in 1995. Since 1984, Dr. Solka has been employed as a Principal Scientist at the Dahlgren Division of the Naval Surface Warfare Center and he currently works in the Advanced Science and Technology Division, Code Q20, of the Electromagnetic Sensors and Systems Department, Q Department. Dr. Solka has published over 120 journal papers, conference papers, technical reports, and book chapters and he holds 4 patents.

Mr. Robert Rushenberg has a B.S. in Mechanical Engineering and a minor in Astronomy from Iowa State University. He has worked in the field of Textual Data Mining for the past three years, focusing on technology assessments and country science and technology. Most recently, Mr. Rushenberg has worked on development and application of radical scientific discovery to improved water purification, China technology assessment, and promising new research directions for corrosion science and technology.