Literature-Based Discovery on the World Wide Web

MICHAEL GORDON
University of Michigan Business School, Ann Arbor
ROBERT K. LINDSAY
Mental Health Research Institute, University of Michigan, Ann Arbor
and
WEIGUO FAN
Virginia Pamplin College of Business, Virginia Polytechnic Institute, Blacksburg

Previous research has shown that researchers can generate medical hypotheses by using computers to analyze several, seemingly unrelated, medical literatures. In this work we suggest broader application for the ideas of literature-based discovery. Specifically, we suggest that literature-based discovery can be fruitful in areas other than medicine; that in addition to finding "cures" for "problems," literature-based discovery offers the possibility of finding new problems for existing technologies; that the analysis of a single literature may be sufficient for literature-based discovery; and that literature-based discovery can support individuals seeking to draw together ideas from various areas of inquiry, even if such connections have been previously made by others.

We describe literature-based discovery experiments conducted on the World Wide Web that support these ideas.

Categories and Subject Descriptors: H.3.3. [**Information Search and Retrieval**]: Subjects: retrieval models, search process; H.2.8. [**Database Applications**]: Subject: Data Mining

General Terms: Experimentation, Design

Additional Key Words and Phrases: Literature-based discovery

1. INTRODUCTION

D. Swanson began demonstrating over a decade ago that the MEDLINE medical database could be successfully analyzed by a computer in order to suggest

Authors' addresses: M. Gordon, Computer and Information Systems, University of Michigan Business School, Ann Arbor, MI 48109; email: mdgordon@umich.edu; R. K. Lindsay, Mental Health Research Institute, University of Michigan, Ann Arbor, MI 48109; email: lindsay@umich.edu; W. Fan, Department of Accounting and Information Systems, Pamplin College of Business, Blacksburg, VA 24061; email: wfan@vt.edu

new medical hypotheses [Swanson 1986, 1987, 1988, 1990]. The basic idea behind Swanson's method is to find a *bridge* that links two conceptually related topics that ought to have been studied together but never have been. For instance, in his initial work [Swanson 1986], the topic of *blood viscosity* served as a bridge between the topics of *Raynaud's disease* and dietary *fish oil*. The connections from *Raynaud's disease* to *blood viscosity*, and from *blood viscosity* to *fish oil* could be made on the strength of lexical statistics, heuristics, and a modest amount of human intuition and judgment; biological reasoning showed the medical relationship between Raynaud's disease and fish oil. Despite this, Raynaud's disease and fish oil had not been written about together at that time.

Since his first reports that literature-based discoveries were actually possible, others have taken up the task [Gordon et al. 1996; Gordon et al. 1998; Lindsay et al. 1999; Weeber et al. 2001] and variations on Swanson's methods have been attempted. What has been constant, though, are several features in undertaking the task of literature based discovery:

- a *scientific* literature as the source of discovery;
- discoveries have generally proceeded from *disease* to *cure* (problem to solution);
- the analysis of two or more literatures has been necessary to produce a discovery; and
- what constitutes a literature-based discovery has been a connection either *completely new* to the field or at least *overlooked by the vast majority* of its practitioners.

In this article we characterize the discovery process more broadly. We include nonscientific literature sources, connections leading from "cures" (or technologies) toward diseases (or suitable problems), discoveries supported by the analysis of a single literature, and ideas that provide personal enlightenment—even if they don't break completely new ground. At the end of this article we will give examples of how Web searching can, potentially, lead to new knowledge.

2. BACKGROUND

Information retrieval is a difficult problem. It involves translating an inquirer's need for information into a computer processable query and then (partially) matching that query against the representations created for documents in a document collection. Each of these activities involves a process that can be performed in many ways, with each variation having its virtues and defects.

Literature-based discovery relies on information retrieval-based techniques and insights, but is a much harder problem. Whereas information retrieval has, at the outset, the objective of finding documents relevant to a given need for information, the success of literature-based retrieval depends on finding topics (or documents) that are only *indirectly* relevant to the topic one uses to initiate the discovery process. In addition, what is found must be previously unknown in relation to the starting point.

Literature-based discovery differs, too, from efforts in knowledge discovery and data mining in material ways. First, literature-based discovery uses as

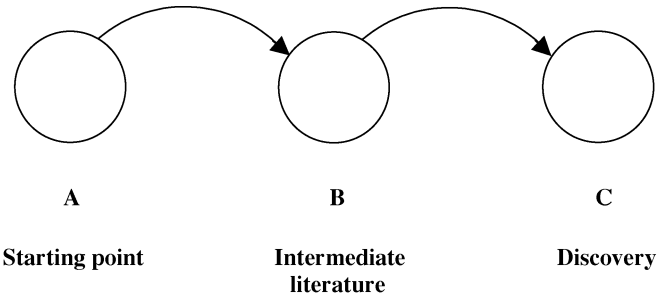
its input collections of ordinary documents—not transaction logs, database relations, or other structured information. Second (and more important), literature-based discovery seeks relationships that, by definition, are *not* contained within an existing textual corpus, unlike efforts that seek correlations, patterns, or rules within a defined set of texts.

Let us make the idea of literature-based discovery concrete. Raynaud's disease has served as a starting point for literature-based discovery. In 1986 when Swanson conducted his initial experiments, there was on MEDLINE a set of approximately 560 documents during the most recent five-year period whose bibliographic citations (including abstracts) mentioned the word *Raynaud*. The disease itself has no known cause or cure, and the goal of literature-based discovery was to uncover in MEDLINE novel suggestions for how Raynaud's disease might be caused, and how it might be treated. Of course, any idea that was already written about in conjunction with Raynaud's disease would not be novel. So the search for new connections had to exclude these ideas as new discoveries.

However, if one idea related to Raynaud's disease were itself related to a third idea, *that* idea might be related to Raynaud's disease, *and* the relationship might be brand new. In this regard, two key problems in literature-based discovery are first, to determine which already-established connection to Raynaud's disease should serve as a bridge; and second, to decide to which other novel-but-related ideas this bridge might link.

In Swanson's original work [Swanson, 1986], the literature on Raynaud's disease was downloaded and the terms occurring in the titles of those documents were analyzed statistically to find "clues" revealing topical themes contained in those documents. One or more of these indicated a new literature that was, itself, downloaded and statistically analyzed in a similar manner to search for an undiscovered idea in relation to Raynaud's disease. Swanson [1987, 1988, 1990]; Swanson et al. [1997]; Smalheiser et al. [1994, 1996a, 1996b, 1998]; Gordon et al. [1996, 1998]; Lindsay et al. [1999]; and Weeber et al. [2001] have explored variations of this approach by analyzing the full MEDLINE records of documents (rather than titles), considering various statistical methods to use for analysis, looking at phrases within documents (rather than individual terms), using concepts instead of textual tokens, and devising methods to display novel connections more readily.

The basic literature-based discovery process can be diagrammed like this:



3. GENERALIZING LITERATURE-BASED DISCOVERY

Common to all literature-based discovery efforts has been a process of successively analyzing multiple medical literatures in search of a novel treatment or explanation for a given medical problem. It is possible, however, to consider literature-based discovery in broader terms. We consider the use of *non*scientific literatures, literature-based discovery involving only a *single* literature, and discovery that proceeds from "cure" *to* "disease." We also discuss the importance of literature-based discovery approaches in situations where an individual's overriding aim is to become aware of connections between different bodies of information, even if he or she is not the first to make them.

3.1 New Sources of Discovery

Both experts and laymen have access to staggering amounts of information. The Internet and the World Wide Web have made it possible to access information on wines, organic produce, mergers and acquisitions, political terrorism, and seemingly everything else. Furthermore, publication costs are greatly reduced in comparison with traditional, nonelectronic means.

With this explosion of information, it is increasingly likely that there are connections between what is being written for different audiences. However, as Swanson so accurately pointed out, the explosion in written information is accompanied by increased specialization: the more there is to know, the more well-informed individuals will focus on fewer topics, lest they become overwhelmed. Indeed, in addition to the half-dozen or so well-known search engines, there are thousands of search engines specialized for specific content areas.

The specialization of which Swanson spoke is pervasive, and it is not simply a recent phenomenon. To his frustration, Darwin couldn't explain how variability within a species arose; and Mendel couldn't convince anyone of the significance of his experiments with cross-breeding plants. Ironically, the two were contemporaries, whose work strongly reinforced and complemented each other's—and Darwin even owned a book describing Mendel's experiments, though he never read the appropriate sections [Ridley 2000]. It seems plausible that similar phenomena occur far more often than we suspect (by definition they are unreported), and arise in situations outside the area of medicine as well as within it. Though there are certain advantages to performing literature-based discovery within the well-structured MEDLINE database, that does not rule out the possibility—or desirability—of coaxing discoveries from other sources.

3.2 Extension

A discoverer may be seeking new application areas for an existing technology or method. The practical applications of chaos theory, as an example, can presumably be extended to a great number of new problem areas. Literature-based methods may be appropriate in determining problems to which a technology applies.

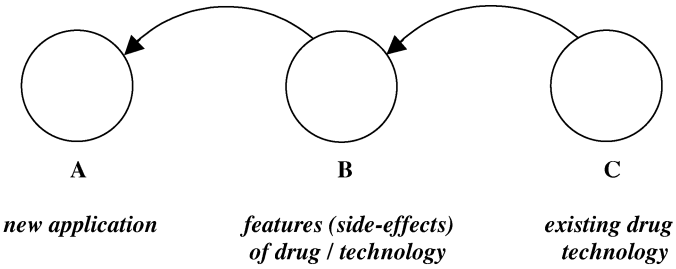
In fact, the history of commercialization of new ideas is full of examples of now indispensable technologies which at first appeared to have no practical

use. Edison failed to obtain the European patent for a device that evolved into the movie projector, saying, "It isn't worth it" [Boorstin 1974]. Western Union's President, William Orton, was approached by Alexander Graham Bell's backers about buying the telephone company and reportedly said, "What use could this company make of an electrical toy?" [MacKenzie 1928, p. 158]. Thomas Watson, Chairman of IBM, predicted in 1943 that the world market for computers was about five [Danzig 1999]. Personal computer sales failed to take off until the first "killer app" (spreadsheets) came to market. And even the Internet, a technology that is fundamentally changing our relationships with both machines and each other, was initially viewed by Microsoft as being an essentially frivolous application of computer technology [Edstrom et al. 1998].

Literature-based *extension* involves selecting some rich domain, using literature-based techniques to analyze it, and then attempting to see how it might be applied to problems in new areas. Starting with a given technology, one might use lexical statistics to determine its salient characteristics, which themselves serve as the source of a second round of lexical analysis whose objective is to find a new application area.

In medicine, for instance, one may analyze the literature on a known drug and determine some of its side-effects. The former side-effects of the drug minoxidil, originally developed to treat high blood pressure, have become the *main* effects of this drug, which now generates revenues of $150 million annually to treat baldness [Sheen 1999]. Conducting literature-based discovery to explore new uses for existing drugs is an approach being explored by Weeber et al. [2001].

Extension could be diagrammed this way:



In extension, a solution (or technology) with known characteristics is seeking new applications. The flow of associations is "backwards" compared to traditional literature-based discovery (where problems seek solutions).

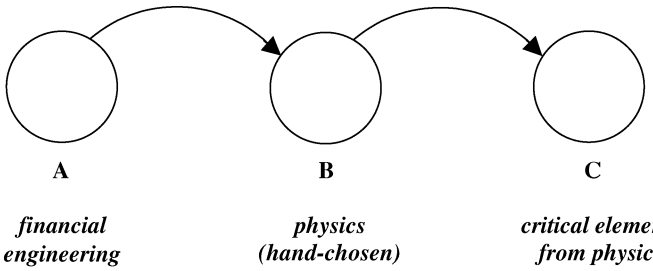
3.3 Inspiration

The idea behind lexically-driven literature-based discovery is to rely on statistics to guide the search from an initial literature, A, to an intermediate literature, B, and from there to a target (discovery) literature, C. However, the $A \rightarrow B$ step may be driven by the discoverer's intuition, instead. For instance, if one's goal were an *optimization algorithm*, one could immediately generalize that aim to the concept of *optimization* (independent of the idea of algorithm). Literature-based discovery techniques could then be used on this intermediate

literature ($B = optimization$). Similarly, the hand-selected B may have been a topic less directly related, such as *physics* or *chemistry,* in hopes that a new optimization algorithm could be invented by an examination of these topics. In fact, the optimization technique known as simulated annealing is based on an analogy to a physical process of heating and then slowly cooling a substance in a controlled way to obtain an optimal (minimal) energy state (a process known as annealing). In developing simulated annealing, existing concepts were viewed and put to use in a new context to inspire a new method of computation. Though not born of literature-based discovery, it is conceivable that literature-based discovery could have stimulated such an invention. Similarly, financial engineers are looking to chaos theory and nonlinear dynamics (from the fields of physics and complex adaptive systems) to find new ways to exploit the stock market [Bass 1999].

Naturally, a thorough reading (in chemistry or physics, say) can provide all the information one needs to understand the concepts from one domain and then apply them to another. But literature-based discovery techniques can assist in this process by helping decompose a subject area into various subtopics and giving an indication of their interrelationships, at least as revealed by lexical statistics. To date, such assistance has proven useful in guiding explorations of multiple literatures to suggest novel discovery hypotheses. It is possible that the same methods could be useful, as well, for a human discoverer who desires some computational support in developing the relationships between different fields of discourse.

Inspiration could be diagrammed this way:



Literature-based discovery techniques help the discoverer understand the new topic area, C—an area with which he or she may not be intimately familiar. In an exploratory, iterative fashion, the task then becomes to map these features onto the initial problem domain (A) to determine their applicability and usefulness to the problem at hand.

3.4 Personal Novelty

Literature-based discovery in the field of medicine has been based on an "intersection test." When a topic C (fish oil) is uncovered by a literature-based discovery process that began with the topic A (Raynaud's disease), one must then ask whether the topics are bibliographically disjoint. In other words, have fish oil and Raynaud's disease been written about together? Does one cite

the other? Or, in a way that indicates weaker coupling, do they cite common documents?

If topics *A* and *C* do intersect, then the claim of a new discovery must be ruled out. But this certainly does not mean that only brand new discoveries are of value. A system that brings attention to a connection that is new to *you* can be of great importance. It goes without saying that the field of knowledge is so broad that no person can be familiar with more than a tiny sliver. What is more, there are a far greater number of relationships among ideas than there are ideas in the first place, and keeping abreast of these is harder still.

Consequently, systems that acquaint an inquirer with relationships that are new to him or her have great value. Indeed, there are circumstances in which the concept of an intersection test becomes completely meaningless. As an example, the sales manager who wants to find examples of successful product launches from different (but related) industries is interested in information that is new to him (her), no matter who else might know it. Literature-based techniques may help surface such personally novel linkages.

4. EXPERIMENTS

We conducted a series of experiments in which we attempted to show that literature-based discovery could be performed on the Internet. Several of the ideas that we have described already are put in operation in these experiments, among them:

- allowing intermediate (*B*) literatures to be manually designated (as in inspiration);
- working from a "solution" to a problem (as in extension); and
- working on a corpus other than medical literature

These experiments embrace a continuum of methods. At one extreme is an experiment conducted almost entirely by automatic methods. At the other is a method that is purely intellectual. The remainder of this article discusses these experiments, not to demonstrate their absolute or even relative effectiveness, but to suggest what can be done and to stimulate further work in the area of literature-based discovery.

All of the experiments address a common question: Are there untried uses for genetic algorithms suggested by the literature? This question has practical import, as evidenced by the ever-increasing number of applications for which this method is well suited. It serves a theoretical purpose, by illustrating the notions of *extension* and *inspiration*. It also allowed us the special advantage of contrasting the results of our partially automated methods of literature-based discovery with the purely intellectual efforts of John Holland (the world's foremost expert on genetic algorithms; see Experiment 4).

4.1 Experiment 1

This experiment demonstrates literature-based support for what we call *extension*—finding new applications for existing solutions. In this experiment, we began by locating and downloading the 50 most prominent documents on

Table I. Phrases Identified by Lexical Statistics as Being Related to Genetic Algorithms

fitness function
optimal solution
soft computing
search technique
image processing
stochastic search
bit string
objective function
connection machine
multi-modal
combinatorial optimization
Darwin theory

the World Wide Web related to the topic *genetic algorithms*. We did this by submitting the query "genetic algorithms" to the AltaVista search engine, and retrieving the contents of pages so identified with the commercial software package, MemoWeb (www.memoweb.com).

Using Unix-based software (described in more detail in Gordon et al. [1996] and Lindsay et al. [1999]) that was modified for Web pages, we proceeded as follows. First, we parsed the collection of retrieved Web pages (documents) into two-word phrases. We ensured that punctuation rules detected spurious phrases like *counts number,* an adjacency phrase occurring in the last sentence of this paragraph if punctuation is overlooked. Similarly, we used stop-word lists to eliminate noncontent bearing words. Next, we obtained two basic lexical statistics for the identified items: token counts (number of times a phrase occurs within a single Web page) and document counts (number of retrieved Web pages containing a given phrase), and several other statistics derived from these measures.

By examining lists of phrases ordered by these statistics, we selected twelve prominent phrases related to genetic algorithms, including, *fitness function*, *optimal solution*; *soft computing*; *search technique*; *bit string*; and *Darwin theory* (see Table I). In addition to ranking highly according to our statistical methods, each of these items seemed important (or interesting) based on our knowledge of genetic algorithms. We overlooked some other phrases with high statistical rankings and possibly meaningful connections to the concept of genetic algorithms; this is where human judgment can come into play in literature-based discovery. Altogether, over 3,000 phrases were evaluated at this stage.

For each of the twelve candidate phrases we identified, we went through a process similar to what we have already described: using AltaVista and MemoWeb, we obtained the top 100 Web pages identified by a query composed of just that phrase; we then parsed these pages and produced twelve separate lexical statistical analyses.

We then attempted to "pool" the results of these analyses. We did this in two different ways. First, we determined the number of lists in which any particular phrase occurred. Since we had twelve lists, this statistic produced a score

between 1 (one list) and 12 (all lists) for all the phrases we examined. Second, we determined an item's *normalized token frequency*. Within a given analysis, a phrase received a value equal to its token frequency for that analysis, divided by the maximum token frequency over all phrases within that analysis. Accordingly, the most frequently used phrase in that analysis received a score of 1.0, and all other phrases received a smaller, but nonnegative score. The normalized token frequency for the phrase was simply its score summed across the twelve analyses. This statistic attempted to determine the most frequently used phrase across analyses, accounting for differences in the size of the Web pages involved and in the number of different phrases each analysis contained.

From these analyses, we selected 42 prominent phrases that we thought might be important in relationship to genetic algorithms. Each of these might be a "discovery" (or what we deemed an extension) if we could first show that it had never been considered before along with genetic algorithms. Using the Web of Science (1987 to the present) and also UseNet, we determined the number of times each item had been written about in relationship to the phrase *genetic algorithm*. (This is the *intersection test* we performed for all experiments.) No items were nonintersecting with *genetic algorithm* when both sources were included. According to the Web of Science, the phrase *virtual reality* was nonintersecting with *genetic algorithm*. According to UseNet, *computer graphics* was nonintersecting. For a list of all terms considered, plus their intersections, see Table II. A few items, like *fluid dynamics* and *text retrieval*, are rather novel in the context of genetic algorithms, and may deserve further consideration in conjunction with this technique.

We also considered each of the twelve analyses separately, rather than pooling their results. By pooling results across twelve analyses, we were favoring any term appearing in a variety of contexts (i.e., appearing on many lists) and occurring frequently (high normalized token frequency). In the nonpooled analyses, we sought to identify a set of rarer occurring terms, on the assumption that this might improve the odds of identifying a discovery in relation to genetic algorithms. Thus, in considering each list separately, we identified terms in order of *increasing* document and token frequency, subject to the constraint that they appear at least twice within the particular literature being analyzed. The constraint was imposed to prevent consideration of terms that might be artifacts of the literature being analyzed.

This method produced fourteen items that had never before been mentioned in connection with *genetic algorithms*, according to our intersection test using the Web of Science. These included *cancer detection*, *financial modeling*, and *secure communication* (see Table III). Terms with a document frequency of just 1 that had 0-intersection with genetic algorithm included *asset packaging, demand prediction, option price, and price search*. Once prompted with one of these phrases, an expert might devise a way that genetic algorithms could be applicable. For example, a genetic algorithm might be used in financial modeling in order to devise a portfolio that optimizes the trade-off between risk and return. Altogether, over 8,000 phrases were considered in selecting candidate discovery items.

Table II. Intersection Frequencies of Potential Discovery Phrases with *Genetic Algorithm* (These items were generated from nonpooled analyses. Each has a 0-intersection with *genetic algorithm* according to the Web of Science. The phrases *financial engineering* and *portfolio selection* have two or three intersections with *genetic algorithm* respectively, according to UseNet.)

	Intersection Frequency	
Phrase	Web of Science	UseNet
Virtual reality	0	34
Partial differential	1	6
Game playing	1	13
Fluid dynamic	2	3
Text retrieval	2	3
Crew scheduling	2	5
Civil engineering	2	14
Machine vision	2	18
Computer graphics	3	0
Vehicle scheduling	3	3
Project scheduling	5	3
Pattern matching	5	7
Character recognition	5	13
Image compression	5	48
Vehicle routing	6	17
Logic programming	6	52
Resource allocation	6	6
Edge detection	8	4
Speech recognition	8	28
Game theory	8	31
Image analysis	8	40
Computer vision	9	1
Feature extraction	9	20
Machine scheduling	10	2
Differential equation	10	3
Image segmentation	10	35
Autonomous agent	11	9
Travelling salesman	12	41
Cluster analysis	13	4
Signal processing	13	99
Constraint satisfaction	15	87
Linear programming	16	192
Mixed integer	17	32
Network design	19	43
Reinforcement learning	20	49
Engineering design	221	26
Mobile robot	29	48
Integer programming	31	42
Dynamic programming	33	19
Shop scheduling	33	27
Finite element	53	22
Pattern recognition	83	200

Table III. Potential Discovery Items that Do Not Intersect with *Genetic Algorithm*

Cancer detection
Encryption technology
Financial modeling
Modal transportation
Planning decision
Product management
Seasonal planning
Secure communication
Shadow price
Transaction processing
Travel forecasting
Financial engineering
Portfolio selection
Financial forecasting

4.2 Experiment 2

This experiment demonstrates literature-based support for *inspiration*, as we use this term. Recall that inspiration begins with an individual identifying a concept (intermediate or B literature) that might prove fertile in making a discovery and then using lexical statistics or other heuristics to follow up. We began by identifying, intellectually, four topics that we knew were central to the idea of genetic algorithms: *search algorithms* or *search techniques*; *optimization*; *machine learning*; and *adaptive algorithms* or *adaptive searches*.

We explored separately each of the literatures related to these topics, as follows:

(1) obtained from the World Wide Web documents found by issuing the query *phrase* AND NOT *genetic algorithm*, where *phrase* is one of the four topics shown above;
(2) parsed the documents obtained into their constituent phrases and computed lexical statistics; and
(3) determined the ten "best" items from each list so obtained.

The various literatures analyzed contained between 3,000 two-word phrases (when beginning with the topic of machine learning) and 10,000 two-word phrases (when beginning with the topic of optimization). We sorted the terms within a given analysis according to their document frequencies and then selected, by hand, the ten most interesting terms in order of increasing document frequency, for all terms with document frequency ≥ 2. The decision to consider documents with lowest document frequency first was made so as to strive for novelty: a term's occurrence in fewer documents suggested it was a less well-studied concept, and therefore might currently be unstudied in relationship to genetic algorithms. As before, the constraint of considering only terms occurring in at least two documents was to guard against considering terms that were more or less artifacts of the literature analyzed.

Table IV. Potential Discovery Items Generated by Literature-Based Support for *Inspiration*

	Intersection Frequency	
Phrase	Web of Science	UseNet
Adaptive filter	1	4
Adaptive prediction	0	0
Adaptive wavelet	1	0
Asset allocation	0	0
Combinatorial optimization	93	68
Constrained optimization	23	41
Constraint satisfaction	15	87
Controller optimization	0	2
Differential equation	4	4
Dynamic programming	33	19
Feature selection	32	36
Finite element	53	22
Game playing	1	13
Image compression	5	48
Image enhancement	3	7
Investment analysis	0	1
Linear regression	25	30
Logic programming	6	53
Matching algorithm	5	2
Multi-agent	8	32
Multilevel optimization	1	0
Multiobjective function	21	32
Nonlinear optimization	31	92
Nonlinear programming	20	55
Parameter estimation	32	33
Pattern recognition	84	200
Portfolio optimization	0	1
Quadratic programming	11	44
Resource allocation	7	6
Resource management	1	4
Risk management	1	4
Sequential regression	0	0
Signal processing	13	96
Signal restoration	0	0
Spectral estimation	0	0
Text segmentation	0	0
Theorem proving	0	2

The four analyses produced 37 items (after removing 3 duplicates) to consider as possible discoveries. Though Step 1 above purposely excludes documents mentioning *genetic algorithms* in an attempt to identify the central ideas concerning the search phrase, the 37 items we selected might still have already been connected to genetic algorithms, so intersection tests were necessary. Table IV shows the terms we identified, along with their intersection frequencies with genetic algorithm according to the Web of Science and UseNet. Once again, a suitable expert might be able to find an appropriate and novel use for genetic algorithms in one of these problem areas.

Table V. Items that Have Low Intersection Frequencies with *Genetic Algorithm* Generated by a Highly Automated *Inspiration* Process

	Intersection Frequency	
Phrase	Web of Science	UseNet
Jazz programming	0	0
Tree search	1	8
Binary search	1	6
Search engine	1	0
Adaptive filter	1	4
Lm algorithm	0	0
Adaptive grid	1	0
Adaptive mesh	0	0
Filter algorithm	0	0
Portfolio optimization	0	1

4.3 Experiment 3

This experiment was a reanalysis of the data from the previous experiment. We were motivated by the following question: What would be the result of performing the previous experiment with little or no human judgment? In other words, starting with a discoverer who has in mind a topic to explore as a bridge to discovery, can literature-based statistics be used in an almost rote way to support discovery? While we do *not* believe that a task as intellectually challenging as literature-based discovery can be automated, we viewed this attempt as something of a "straw man" in a comparison to our previous approach, which relies on both judgment and statistics.

We discovered, however, that there was no way to automate the exact process we had followed. The problem was that there were hundreds of terms within each analyzed literature with a document frequency of two; so following an automated process to select *the* ten terms in order of increasing document frequency was impossible. (The problem, of course, would only be exacerbated by beginning with terms whose document frequency was one.)

As a result, we conducted a modified automated analysis. We selected terms in order of *decreasing* document frequency within a list, ensuring that a unique set of items would be identified due to fewer items having high document frequencies. We only used human judgment to eliminate items that we felt were irrelevant for discovery purposes, such as proper names and items containing stop words. The items so identified, along with the intersection test results from Web of Science and UseNet are shown in Table V. To our surprise (because we had selected high frequency terms), a few items, like *portfolio optimization*, *filter algorithm*, and *adaptive filter*, suggest new or overlooked ways that a genetic algorithm might be applied—and each of these phrases was generated with almost no human judgment.

4.4 Experiment 4

Here we asked our colleague, John Holland, to identify items he thought would be interesting to study in conjunction with genetic algorithms. Holland invented genetic algorithms several decades ago [Holland 1975], and has been extremely

active in continuing to study both their underlying theory and their application. In response to our request, he identified the following issues that, to his knowledge, were unpublished in relationship to genetic algorithms:

- optimizing taxonomic trees generated by cladistics;
- Interaction of "designer genes" in commercial "hybrid" corn, wheat, rice, and so on;
- the role of building blocks (such as the Krebs cycle) in the progressive evolution of embryogenetic systems; and
- adaptive guidance for space vehicles.

Holland also commented on the surprising difficulty of identifying these items. We note that none of the items we identified was on his list, or vice versa. Indeed, even if every item we had identified were also to be identified by *some* expert, no single expert would even be close to capable of generating so many different discovery hypotheses.

5. CONCLUSION

We have argued that the idea of literature-based discovery may be applicable in ways previously not considered. We have suggested the merit of using the literature to search for novel applications for existing problem solutions, as well as using lexically-based analytic methods to support a traditional discoverer who already has developed a notion about where to search for new ideas to relate to a problem of interest.

We have performed several experiments using the World Wide Web that suggest the allure of this approach. These methods have varied in their use of lexical statistics as well as the amount of human judgment each has required.

A decade ago, the thought that information retrieval would be such a prominent technology occurred to very few individuals. Today, some of their ideas have found every day use in the search engines that are vital for successful negotiation of the Web. Perhaps a decade from now, discovery-support tools will be equally prevalent. Certainly, tools that help us better understand and take fuller advantage of what is known would be invaluable, even though they will never replace human judgment.

REFERENCES

BASS, T. A. 1999. *The Predictors*. Henry Holt and Company, New York.

BOORSTIN, D. J. 1974. *The Americans: The Democratic Experience*. Vantage Books, New York.

DANZIG, R. 1999. Edited remarks as delivered for the General Graves B. Erskine Lecture Series, Marine Corps University, Quantico, VA 28 April; in: http://www.chinfo.navy.mil/navpalib/people/secnav/danzig/speeches/sn_ersk.txt.

EDSTROM, J. AND ELLER, M. 1998. *Barbarians Led by Bill Gates: Microsoft from the Inside*. Henry Holt and Company, New York.

GORDON, M. AND DUMAIS, S. 1998. Using latent semantic indexing for literature based discovery. *J. Am. Soc. Inf. Sci. 49*, 8 (1998), 674–685.

GORDON, M. AND LINDSAY, R. K. 1996. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *J. Am. Soc. Inf. Sci. 47*, 2 (1996), 116–128.

HOLLAND, J. H. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

LINDSAY, R. K. AND GORDON, M. 1999. Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci. 50*, 7 (1999), 574–587.

MACKENZIE, C. D. 1928. *Alexander Graham Bell, the Man Who Contracted Space*. Houghton Mifflin, Boston.

MemoWeb. http://www.memoweb.com.

RIDLEY, M. 2000. *Genome*. Harper Collins, New York.

SHEEN, M. In book obtainable at http://www.health-library.com/library/health/.

SMALHEISER, N. R. AND SWANSON, D. R. 1994. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neurosci. Res. Commun. 15*, 1 (1999), 1–9.

SMALHEISER, N. R. AND SWANSON, D. R. 1996a. Indomethacin and alzheimer's disease. *Neurology 46*, 2 (1996), 583.

SMALHEISER, N. R. AND SWANSON, D. R. 1996b. Linking estrogen to alzheimer's disease: An informatics approach. *Neurology 47*, 3 (1996), 809–810.

SMALHEISER, N. R. AND SWANSON, D. R. 1998. Calcium-independent phospholipase A_2 and schizophrenia. *Archives of General Psychiatry 55*, 8 (1998), 752–753.

SWANSON, D. R. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine 30*, 1 (1986), 7–18.

SWANSON, D. R. 1987. Two medical literatures that are logically but not bibliographically connected. *J. Am. Soc. Inf. Sci. 38*, 4 (1987), 228–233.

SWANSON, D. R. 1988. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine 31*, 4 (1988), 526–557.

SWANSON, D. R. 1990. Somatomedin C and arginine: Implicit connections between mutually-isolated literatures. *Perspectives in Biology and Medicine 33* (1990), 157–186.

SWANSON, D. R. AND SMALHEISER, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell. 9*, 2 (1997), 183–203.

WEEBER, M., KLEIN, H., DE JONG-VAN DEN BERG, L. T. W., AND VOS, REIN. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech. 52*, 7(2001), 548–557.

Received February 2001; revised July 2002; accepted August 2002