Using Latent Semantic Indexing for Literature Based Discovery

Michael D. Gordon
Computer and Information Systems, School of Business, University of Michigan, Ann Arbor, MI 48109-1234.
E-mail: mdgordon@umich.edu

Susan Dumais
Microsoft Research, Redmond, WA 98052. E-mail: sdumais@microsoft.com

Latent semantic indexing (LSI) is a statistical technique for improving information retrieval effectiveness. Here, we use LSI to assist in literature-based discoveries. The idea behind literature-based discoveries is that different authors have already published certain underlying scientific ideas that, when taken together, can be connected to hypothesize a new discovery, and that these connections can be made by exploring the scientific literature. We explore latent semantic indexing's effectiveness on two discovery processes: uncovering "nearby" relationships that are necessary to initiate the literature based discovery process; and discovering more distant relationships that may genuinely generate new discovery hypotheses.

Introduction

Literature-based discovery uses the published, scientific literature as a source of new discovery. First discussed by Swanson (1986) in connection with Raynaud's disease, the problem can be characterized in this way: Beginning with the literature, R (for Raynaud's), on some subject, can you identify the literature on another subject that helps in better understanding R, even though no one has ever thought that these two subjects were related?[1] In a series of papers, Swanson (1986a, 1986b, 1987, 1988a, 1988b, 1989a, 1989b, 1989c, 1990a, 1990b, 1991, 1993) showed this could be done, both by intensive reading and study, and by semiautomatic methods involving text analysis. Subsequently, Gordon and Lindsay (1996) have replicated Swanson's results and used other statistical methods to help automate the literature discovery process.

As described by Swanson, there are two basic literature discovery processes. The first leads from the literature (R) associated with an initial topic to the literatures (I) of one or more related, intermediate topics. The second leads from one of these related topics to the literature (PD) associated with a potential discovery. Figure 1 illustrates these two steps (left to right).

We call these two processes *identifying intermediate literatures* and *identifying potential discovery literatures,* respectively (Fig. 1). Our interest is learning if latent semantic indexing (Deerwester et al., 1990), a statistical technique used with success in information retrieval, can help with either or both of these processes.

Identifying Intermediate Literatures

By definition, if we start with Raynaud's and discover a brand new concept (cure, cause, treatment, or physiological process) never before reported, there will be no document that discusses both Raynaud's and this new concept. But there may be a topic that is discussed along with Raynaud's and is also discussed along with the new concept, even though no single article on this topic discusses both. A literature that serves as such a bridge is an intermediate literature.

Finding intermediate literatures, then, is a central problem in literature-based discovery. Of course, one can read about Raynaud's and form impressions on that basis, but a systematic approach for identifying intermediate literatures would be more efficient and possibly more effective.

The following is an example of a MEDLINE record containing the term *Raynaud's* (with slight cosmetic modifications to illustrate more plainly the record's structure):

TITLE: Localized real-time blood flow measurements.
AUTHOR: van As H; Brouwers AA; Snaar JE
CITE: Arch Int Physiol Biochim 1985 Dec; 93 (5): 87–95
LANGUAGE: Eng; English

[1] Literature based discoveries generate scientific hypotheses; conventional scientific research must be conducted if the hypothesis is to be confirmed.

Received January 31, 1996; revised April 30, 1997; accepted April 30, 1997.

> ABSTRACT: A novel method for real time, localized, flow measurements is applied to blood flow in human fingers. Results for arterial and venous flow in normal subjects and patients with abnormal blood circulation are presented. Effects of blood flow regulation by the autonomic nervous system have been observed. Stricture of the digital arteries could be clearly demonstrated in a patient with Raynaud's phenomenon. Experimental signals due to pulsatile flow in a model system can be simulated in a quantitative way. The calibration, however, depends on the actual spin−spin relaxation time and the shape of the pulsatile flow vs. time curve. Due to these limitations, the volume flow rate can be measured with a relative error of approximately +/−25%. (AUTHOR)
> MAJOR TERMS: Blood Flow Velocity.
> MINOR TERMS: Fingers BS. Human. Nuclear Magnetic Resonance DU.
> Support, Non-U.S. Gov't.
> JOURNAL ARTICLE

Among other non-''noise'' words, this record contains *blood* and *flow* (from the title), *flow, blood, fingers,* etc. (from the abstract), plus other words from the remaining MEDLINE record fields. Similarly, the two-word adjacency phrases in this MEDLINE record include *localized real, real time, time blood, blood flow, flow measurements* (from the title), *novel method, real time, time localized, localized flow, flow measurements, blood flow,* and *blood circulation* (from the abstract). Standard information retrieval techniques can eliminate from consideration nonsubstantive words, such as *a, for,* and *is,* and can use sentence punctuation to prevent the inclusion of false phrases such as *fingers results* (from the abstract).

Gordon and Lindsay (1996) have investigated automated processes for supporting the identification of intermediate literatures from MEDLINE records such as these that are based on descriptive statistics similar to those used in information retrieval. Specifically, to identify intermediate literatures related to the topic Raynaud's, they downloaded the full MEDLINE records for all 1983–1985[2] documents that mention Raynaud's, parsed them as described for the sample record, and then computed the statistics shown in Table 1 for every term and two-word adjacency phrase.

For the MEDLINE record shown above, the word *time* had a token frequency of 4; *localized* had a token frequency of 2; and *Raynaud's* had a token frequency of 1. Similarly, the phrase *blood flow* had a token frequency of 4, whereas as *blood circulation* had a token frequency of 1. For this single MEDLINE record, the record frequency for each of these words and phrases is 1.

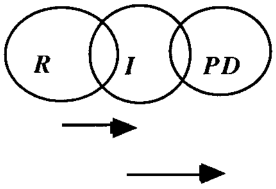

FIG. 1. The two steps in literature-based discovery.

TABLE 1. Four statistics used to identify intermediate literatures.

Statistic	Definition
token frequency[a]	number of tokens[b] of *X* within *R*
record frequency	number of records in *R* containing *X*
tf*igf = token frequency* log(inverse global record frequency)	token frequency* log(number of records in MEDLINE/number of records in MEDLINE containing *X*)
relative frequency	record frequency/number records in MEDLINE containing *X*

[a] Strictly, frequencies should be ratios, but the normalizing denominators in these statistics may be dropped since what is important is term (or phrase) rank *orderings,* which are identical with and without normalization.

[b] Token frequencies count each distinct occurrence of a word (or phrase). For instance, in the sentence ''Row, row, row your boat gently down the stream,'' the token frequency for *row* is 3; for *gently* it is 1. On the other hand, the record frequency of both of these items is incremented by 1 by this record.

For a term or phrase, *X*, these four statistics may be calculated in relation to the literature on Raynaud's literature, *R*.

Table 2 gives an example of the four statistics that would be computed for the term (or phrase) *X*, which occurs both within and outside the Raynaud's subset of MEDLINE.

Gordon and Lindsay (1996) used these statistics to try to identify intermediate literatures for further exploration. After calculating each of the four statistics for every term or two-word adjacency phrase in a downloaded literature (such as Raynaud's), they identified the twenty (or thirty) items with the highest values for each statistic. They then considered each of these items to be a query that that could be used to identify a different intermediate literature. Though the methods used were highly automated, the intended use of these methods was to provide support for a qualified medical researcher who could most effectively interpret and act upon the data provided.

In examining the four separate lists of highest-ranked items, Gordon and Lindsay concluded that three of the statistics—token frequency, record frequency, and token frequency * inverse global record frequency (*igf*)—were extremely predictive of each other. If a particular term or phrase, such as *blood,* was among the top 20 positions on one of the lists, very likely it was among the top 20 of another list as well. As a specific example, in analyzing the Raynaud's literature the four statistics were computed for each of the approximately 2,000 single-word terms that occurred at least four times in that literature. If the terms on the top 20 list for one statistic were statistically independent of those on another, a fractional number should appear on both lists. What was observed, instead, was that the token frequency and record frequency lists

[2] This date range was the same one that Swanson used and supported Gordon and Lindsay's replication of Swanson's results by new methods.

TABLE 2. Fictitious numerical example showing calculation of four statistics for term (or phrase) *X*.

Description	
Document collection characteristics	**Number of Documents**
All MEDLINE	*N*
R = documents mentioning Raynaud's	100
Subset of *R* mentioning term (phrase) *X*	40
Documents mentioning *X*	80
Token characteristics	**Number of Token occurrences**
X mentioned in *R*	70
Statistics	**Value of Statistic**
token frequency (*tf*)	70
record frequency	40
tf * inverse global record frequency (*tf* * *igf*)	70 * log (N/80)
relative frequency	40/80

had fifteen (of twenty) items in common; the token frequency and token frequency * *igf* lists had seventeen; and the record frequency and token frequency * *igf* had fifteen. In other words, an item's appearance on the top 20 list for one statistic was highly correlated with its appearance on the top 20 list of the other two. The same conclusion held when the number of items per list was increased; when two-word adjacency phrases were considered rather than single-word terms; and when literatures other than Raynaud's were analyzed.

There was not nearly the same degree of correlation between a term's occurrence on the top 20 list for relative frequency and its occurrence on the top 20 list of another statistic. Again considering Raynaud's as an example, the top 20 items sorted by relative frequency included one item in common with the top 20 token frequency items; one item in common with the top 20 record frequency items; and one in common with the top 20 token frequency * *igf* items. This pattern held for single words and two-word adjacency phrases, when the top *n* size was adjusted (to values other than 20), and when different literatures were evaluated.

Not only were the token frequency, record frequency, token frequency, and *tf* * *igf* lists quite similar, but they were effective in uncovering intermediate literatures on a discovery path from Raynaud's to fish oil. By looking at the very top items on any of the three lists, one was led from Raynaud's (the starting point) to the topic *blood.* Then, by downloading and analyzing the literature on the topic *blood* AND *Raynaud's,* one was led directly by any of the three statistics to the topic *blood viscosity* (see Fig. 2). Blood viscosity is indeed an intermediate, or ''bridge,'' literature: It is mentioned in the Raynaud's literature and is clearly accepted scientifically as being related to Raynaud's. It is also mentioned in the fish oil literature, and is scientifically related to that as well. Indeed, there are physiological connections implicating fish oil as a treatment for Raynaud's, including that fish oil reduces blood viscosity and that increased blood viscosity is one of the reasons Raynaud's patients suffer symptoms associated with peripheral blood deficiency. Despite this, the hypothesis that Raynaud's might be treated by fish oil lay dormant in the literature until Swanson (1986a, 1986b, 1987) uncovered it by methods of literature-based discovery.

To summarize, Gordon and Lindsay (1996) demonstrated three statistics that were useful for uncovering intermediate literatures to support literature-based discovery: token frequency, record frequency, and token frequency * inverse global record frequency. Each of them separately rank-ordered large lists of terms (and phrases) in quite similar ways. And from the starting point (Raynaud's in this example), a medical researcher using these statistics could be led first to *blood,* and then to *blood viscosity,* by any of these three statistics (Fig. 2). Gordon and Lindsay argued that an effective method for identifying an intermediate literature is finding one with strong conceptual similarity to the starting point and that each of the three correlated statistics can serve this purpose, since each has lexical prominence in the Raynaud's literature.

Latent semantic indexing (Deerwester et al., 1990) offers an entirely different way potentially to identify intermediate literatures and, thus, to support literature-based discovery. A standard term by document matrix, *D*, is mathematically equivalent to the product of three other matrices, as shown in Figure 3. *M* is a matrix of singular values computed by a ''factoring'' process—singular value decomposition (Forsythe et al., 1977)—

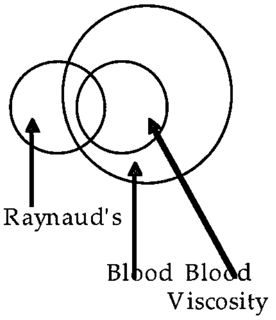

FIG. 2. Raynaud's and two intermediate literatures.

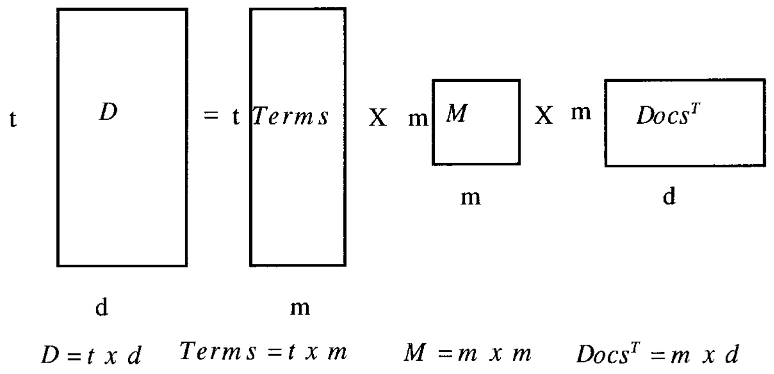

FIG. 3. Decomposition of term by document matrix.

that expresses each of the t original indexing terms and also each of the d original documents as a vector of m factors (where m is the number of linearly independent rows, and columns, in D). Technically and intuitively, each of the original indexing terms is now expressed as a vector of statistically independent factors (and represented by a row of the *Terms* matrix); each document is similarly represented by a column of the $Docs^T$ matrix. In other words, by means of singular value decomposition, terms and documents are represented in the same m-dimensional space.

The great benefit of representing D as a product of three matrices is that we can consider a representational space containing just the $k < m$ most important of these dimensions, for k of any size. We can then approximate D by the equation

$$D \approx D' = Terms' \times M' \times Docs^{T'}$$

where $Terms' = t \times k$; $M' = k \times k$; $Docs^{T'} = k \times d$.

The result is an optimal reduced dimensional approximation of D (by a criterion of least squares). Practically, this means that two documents that use strongly overlapping vocabulary may both be retrieved even if a particular query only uses the terms that index one of them. Similarly, terms will be considered ''close'' to each other if they occur in overlapping sets of documents.

Figure 4 suggests the way latent semantic indexing assists in information retrieval, using term co-occurrences to give support for document similarity. Pretend that the three documents shown are part of a larger collection where *term-a* and *term-b* tend to be used together in indexing documents, as do *term-b* and *term-c*. Then, the query *term-b* may still retrieve Doc-1, even though Doc-1 is not indexed by that term. Similarly, the query *term-c* may retrieve Doc-1 by virtue of ''transitive'' co-occurrence. In other words, *term-c* co-occurs often with *term-b*, which co-occurs with *term-a*. This gives support for retrieving Doc-1 for the query *term-c*. This is the ordinary spirit in which latent semantic indexing is used—to find similarity among documents based on their indexing, and thus retrieve documents that do not exactly match a query.

However, with equal applicability, latent semantic indexing can uncover relationships among terms. For instance, the terms *term-a* and *term-b* demonstrate semantic similarity by occurring together in Doc-2. Similarly, *term-a* and *term-c* will bear a transitive, but measurable, similarity to each other when a collection like the above is represented by means of latent semantic indexing.

This latter perspective suggests that, perhaps, latent semantic indexing provides an alternative approach to uncovering intermediate literatures. Specifically, if terms such as *Raynaud's* are thought to stand for underlying concepts (the concept *Raynaud's disease*), then we can see which terms lie near each other in LSI-space and, thus, make inferences about conceptual similarity.

To test the usefulness of this approach, we began with the 560 documents published during the years 1983–1985 containing mention of the term *Raynaud's*—the same documents used by Gordon and Lindsay and by Swanson. LSI scaling was then performed on this set of documents, and the top 100 factors were retained ($k = 100$). Each document, as well as each term used in any document, was thus represented as a vector in the same 100 dimensional space.

A central interest of ours was to determine if this method produced substantially different (possibly better) results than Gordon and Lindsay's method of selecting intermediate literatures on the basis of token counts, record counts, and *tf* * *igf* statistics.

A fairly crude measure of the similarity between the two methods of generating items associated with Raynaud's is to consider their overlap. To do this, a single list of items representing the ''best'' intermediate items

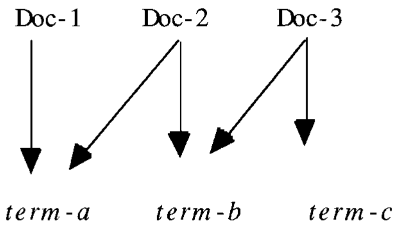

FIG. 4. Doc-x indexed by *term-y* is represented by $X \rightarrow Y$.

by Gordon and Lindsay's method was developed by taking the union of six top-40 lists:

- the top 40 terms, according to record counts
- the top 40 terms, according to token counts
- the top 40 terms, according to *tf * igf*
- the top 40 two-word phrases, according to record counts
- the top 40 two-word phrases, according to token counts
- the top 40 two-word phrases, according to *tf * igf*

This union contained 136 unique items (an item being either a term or a two-word phrase).

The 136 nearest neighboring terms to the term *Raynaud's* according to the LSI analysis were then identified. This was done by rank-ordering all terms by their cosine to the term *Raynaud's.* The size of the intersection of the two lists of 136 times was 57 items (approximately 42% of the elements in either list).

The best LSI-ranked items (i.e., those with lowest ranks) were most likely to be in the Gordon and Lindsay list. Table 3 shows all of the top 50 LSI-ranked terms that also appeared in the Gordon and Lindsay list. Practically every item very close to the term *Raynaud's* in LSI space was identified by the Gordon and Lindsay methods. In particular, of the top 10 items nearest *Raynaud's* according to LSI, Gordon and Lindsay's methods identified nine. Of the top 20 nearest items (by LSI methods), 15 were identified by Gordon and Lindsay's methods; of the top 30, 21; of the top 40, 27; and of the top 50, 31 (Fig. 5). Further, in just examining the very highest-ranked items (those that each method recommends most strongly as an intermediate literature), we find that each of the top 10 from Gordon and Lindsay is among the top 12 items by LSI. In other words, these two lists' top 10 items are nearly permutations of each other. In addition, the very highest items in one list tend to be right at the top of the other list, too.

In fact, a more sensitive analysis was conducted to test for a correlation between the top LSI rank positions and the top Gordon and Lindsay ranks. Since the Gordon and Lindsay list was the union of six separate lists and an item could come from one or more of them, it would not have a unique rank across different lists. Arbitrarily, then, we selected the Gordon and Lindsay two-word phrase list ranked by record counts to provide ranks for use in our analysis. These 40 times were Spearman rank-correlated with the 40 highest-ranking two-word phrases identified by LSI scaling (retaining $k = 200$ factors). A two-word phrase that occurred in one list but not in the other was assigned a rank of 41 in the list in which it did not appear. The null-hypothesis tested was that the top 40 ranks of the Gordon and Lindsay and the LSI lists were uncorrelated. Data and results are shown in Tables 4 and 5.

By this analysis, we can conclude that the top 40 Gordon and Lindsay two-word phrases (by record counts) are rank-correlated with those found by LSI (even if assigning ranks of 41 to items not appearing on a list may suppress slightly the effects of outlying ranks). More simply, an item's approximate position on the Gordon and Lindsay list (whether near the top, in the middle, or near the bottom) will predict its approximate position on the LSI list.

TABLE 3. The 50 nearest neighbors to Raynaud's by LSI that were also identified by Gordon and Lindsay's statistical methods.

Term	Rank vs. Raynaud's	Cosine (term, Raynaud's)
Raynaud's	1	1.000
article	2	0.950
middle	3	0.847
phenomenon	4	0.832
bs	6	0.654
systemic	7	0.629
finger	8	0.624
blood	9	0.583
scleroderma	10	0.559
syndrome	11	0.539
digital	12	0.531
skin	15	0.504
vascular	16	0.502
normal	17	0.501
severe	18	0.500
ae	21	0.474
sclerosis	22	0.468
disorder	23	0.467
changes	26	0.461
di	27	0.458
primary	28	0.455
temperature	31	0.447
systemic sclerosis	33	0.443
flow	34	0.442
associated	35	0.436
trial	38	0.434
double blind	39	0.433
symptom	42	0.426
therapies	44	0.425
blood flow	46	0.424
test	49	0.417

Of course, other methodological approaches could be taken to compute rank correlations, including forming an ''average rank'' for each Gordon and Lindsay two-word phrase (based on its three separate ranks). However, because the three statistics Gordon and Lindsay used to determine intermediate literatures were so strongly correlated, this is unlikely to affect our finding in any appreciable way.

One surprising observation from Table 4 deserves a comment. The phrase *double blind* is the best-ranked phrase in LSI but is not among the top 40 items from the Gordon and Lindsay analysis (it had rank 45, occurring in 11 records). A possible explanation is that the term *Raynaud's* lies near the phrase *double blind* in MEDLINE. More likely, the prominence of *double blind* may be somewhat coincidental and actually result from the fact that the phrase occurred in just 11 of the 560 Raynaud's documents analyzed (14 times in total), but was near

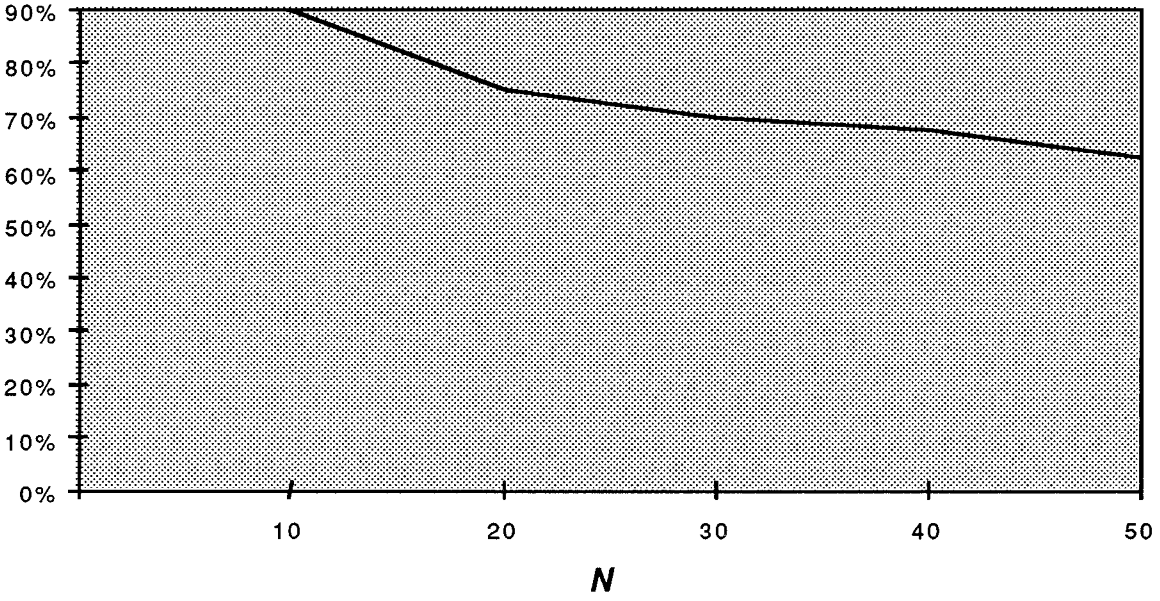

FIG. 5. Percentage of top *N* LSI items identified by Gordon and Lindsay.

Raynaud's by tending to co-occur with all of its chief factors.

What conclusions do these various analyses suggest? Principally, that there is a strong overlap among the terms uncovered by LSI scaling and by the Gordon and Lindsay techniques, and that this overlap is strongest among the very-top-ranked items by each method. Gordon and Lindsay have argued that the best terms for identifying intermediate literatures are those very close (semantically and statistically) to the starting point. By this argument, the two methods may provide similar, but complementary, approaches for identifying intermediate concepts.

In the next section, we change our focus and discuss the use of LSI for identifying potential discovery literatures.

Identifying Potential Discovery Literatures

A connotation underlying the phrase *latent semantic indexing* is that hidden relationships among concepts exist and, further, that they may be teased out statistically. Figure 4 has already illustrated how the concept identified by *term-c* may bear a latent relationship to the concept identified by *term-a* because both terms co-occur with *term-b.*

Is it possible, then, that LSI can form a bridge that connects two bilbiographically isolated literatures? From Swanson's work (1986a, 1986b, 1987), for example, we know that the concept *blood viscosity* is scientifically related to both *Raynaud's* and to *fish oil,* but that neither the Raynaud's nor the fish oil literature refers to each other, nor are they mentioned together by other documents.

Suppose, however, that one had conjectured that *blood viscosity* is an important intermediate literature linking Raynaud's and some unknown cause, cure, or treatment for this condition. This conjecture may have been stimulated by using a literature-based discovery tool, or it may have arisen simply by reading and thinking about Raynaud's. Figure 6 may help make this clearer. The suggestion is that a concept[3] that is related to blood viscosity but not directly to Raynaud's may be uncovered through latent semantic indexing.

Since blood viscosity is conjectured to be a bridge to some unknown discovery, it can be the focus for LSI scaling. Selecting the blood viscosity literature to perform LSI processing on would certainly appear to be an advantage in finding hidden connections to Raynaud's since blood viscosity is, in fact, a bridge to a hidden treatment (fish oil). To test the effectiveness of this use of latent semantic indexing, we proceeded as follows. The 809 MEDLINE records published between 1983–1985 and mentioning blood viscosity were downloaded and LSI-processed (retaining $k = 100$ most important factors). A list of closest neighbors to the term *Raynaud's* was then constructed according to their cosine to Raynaud's, but no element on the list could appear in any of the 560 Raynaud's documents from the same period. In other words, we constructed a list of terms that were ''near'' Raynaud's (from the perspective of blood viscosity) but were nonetheless bibliographically disjointed from it. The items on this list would certainly seem worthy of further investigation.

A specific interest was whether the phrase *fish oil* would appear prominently on this list. More generally, we wanted to see which *terminal* concepts contained in

[3] Implicitly, we are assuming that a term used in text represents the concept with the same name. Accordingly, the term *Raynaud's* would represent that medical concept.

TABLE 4. Top 40 phrases and ranks by LSI and Gordon and Lindsay analysis.

	Rank	
Phrase	LSI	Gordon and Lindsay
arch dermatol	37	41
adrenergic beta	41	25
antibody technique	41	34
anticentromere antibodies	41	26
antigen antibody	41	31
antinuclear antibodies	41	16
arterial occlusive	36	17
arthritis rheum	28	13
beta receptor	41	23
biofeedback psychology	41	27
bleomycins ae	34	28
blind trial	14	41
blood cell	19	41
blood flow	2	1
blood platelets	33	41
blood viscosity	40	14
calcium channel	23	18
channel blocker	26	24
connective tissue	5	3
controlled double	32	41
crest syndrome	30	12
digital arteries	29	41
digital blood	27	41
double blind	1	41
esophageal dysmotility	41	32
finger blood	38	41
finger systolic	22	41
finger temperature	15	41
flow velocity	25	29
fluorescent antibody	41	35
function tests	20	36
lupus erythematosus	8	4
mal vasc	17	22
mixed connective	41	11
nailfold capillaries	41	37
nifedipine ae	21	41
placebo controlled	16	41
platelet aggregation	18	15
progressive systemic	10	9
pulmonary hypertension	41	38
random allocation	7	19
receptor blockaders	41	20
regional blood	4	8
respiratory function	35	41
rheumatoid arthritis	39	21
rheum dis	24	39
skin temperature	6	6
systemic di	9	10
systemic lupus	13	7
systemic scleroderma	11	30
systemic sclerosis	3	2
thromboangiitis obliterans	41	40
thromboxane synthetase	31	41
vasodilator agents	41	33
vibration ae	12	5

the list might suggest a new discovery about Raynaud's. By way of an example, a substance such as aspirin was considered a *terminal* in the sense that it can be considered a possible cause, cure, or treatment for Raynaud's. A topic such as *tissue hypoxia* (hypoxia means a decrease in normal levels of oxygen) might be related to Raynaud's but cannot be considered a terminal concept by the same lines of reasoning, thus it was excluded from our list. The list of all Raynaud's neighbors that were bibliographically disjoint from Raynaud's and had a cosine (to Raynaud's) of 0.005 or more was examined by hand to remove nonterminals.[4] Table 6 shows the items that remained.

Ignoring the final column, each row in the table shows the value of cosine (Raynaud's, terminal term); a terminal term's ''Rank in LSI, non-Raynaud Terms,'' which only considers terms appearing in blood viscosity documents but not appearing in a Raynaud's document; and a terminal term's ''Rank in LSI space, all Terms,'' which tells how many terms had a larger cosine to Raynaud's, including all blood viscosity terms and two-word phrases (in any of the 809 blood viscosity documents). For instance, 149 terms had a larger cosine than the term *hydroxychloroquine,* but hydroxychloroquine's rank of six among non-Raynaud's items means that there were only five higher-ranked items, each judged a nonterminal, that appeared in blood viscosity documents but not in Raynaud's documents, including the items *viscosities* and *motor activity*. Notice that *hydroxychloroquine* is the only terminal term in Table 6 that has a cosine value of above 0.10.

By definition, none of the terms in Table 6 appeared in any of the 560 1983–1985 Raynaud's documents; the three-year time span was chosen to correspond as closely as possible to the documents Swanson (1986a, 1986b, 1987) examined in his Raynaud's studies. It is possible, of course, that some of the terms in Table 6 occurred along with the term *Raynaud's* before 1983. Because we are, in effect, investigating Swanson's literature-based discovery of the Raynaud's–fish oil connection, we can ignore co-occurrences after 1985. So we queried MEDLINE to determine the number of documents containing both the term *Raynaud's* and each one of the terms in Table 6 in any year before 1986. Results are shown in the last column of the table. This column indicates which terminal terms we can rule out as possible discoveries by

TABLE 5. Spearman rank correlation for LSI and Gordon and Lindsay top ranks.

Sum of squared differences	11,980
r_s	0.57
t ($\rho_s = 0.0$)	5.02
p	<0.001

[4] Currently, automatic processing of text is incapable of determining terminal concepts. Thus, identification of terminals must be conducted by hand. This manual step does not diminish our approach, whose objective is to support hypothesis discovery, not automate it. Terminals can rapidly be selected from lists of terms and phrases, especially by domain experts.

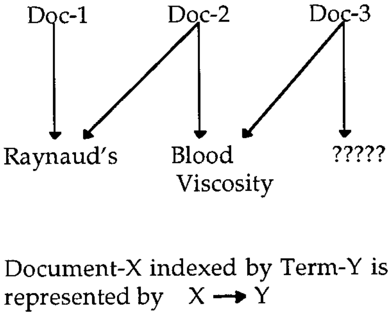

FIG. 6. Blood viscosity as intermediate literature.

virtue of having been already discussed with Raynaud's (those with a nonempty intersection).

In regard to our effort to discover directly the Raynaud's–fish oil connection, the results are disappointing. *Fish oil* is on the list of nonintersecting items, but is nowhere near the term *Raynaud's* (being its 1961st closest neighbor) and still behind almost 600 other terms that appear in the blood viscosity, but not Raynaud's, literature. However, eicosapentaenoic acid, the active agent in fish oil, fares much better, being the 208th-ranked newly uncovered item in relation to Raynaud's, and the fifth-ranked terminal when additional MEDLINE searching has ruled out terms and phrases that were mentioned along with Raynaud's in articles written before 1983. But its cosine, at 0.016, is a faint signal.

Two other items with null-intersection to Raynaud's may deserve further study if experts in the field should confirm their merit. Calcium dobesilate has been used for vascular diseases and diabetes retinopathy, among other conditions. It has been shown to reduce blood viscosity, improve venous insufficiency, and reduce platelet deposition. Niceritrol has been used to treat hyperlipidemia, and, in addition, it has beneficial effects on blood viscosity and platelet aggregation. The effects of these drugs are related to treating Raynaud's.

It is interesting to note, too, that among the items in Table 6 which have non-empty intersection with Raynaud's are substances such as isoxsuprine and dextran, which have been used to treat Raynaud's. In addition, some of the nonterminal, but nonintersecting, items produced by the analysis suggest possible avenues to examine in connection with Raynaud's. For instance, lysolecithin, an acid formed by an enzymatic process in the blood, is capable of breaking up red blood cells and thus may prove useful in treating Raynaud's. This conjecture, too, can be appropriately evaluated by medical experts.

A variation on this approach to finding new connec-

TABLE 6. Terminal concepts identified as Raynaud's neighbors plus their MEDLINE intersections with Raynaud's.

Terminal term	Cosine (term Raynaud's, terminal term)	Rank in LSI space, non-Raynaud terms	Rank in LSI space, all terms	Number of hits (Raynaud's ∩ terminal term before 1986)
hydroxychloroquine	0.178	6	150	1
fat	0.065	48	500	3
rutin	0.064	49	517	3
erythropoietin	0.060	61	547	0
uric acid	0.056	67	573	3
isoxsuprine	0.054	73	593	5
althesin	0.047	99	649	0
phospholipid	0.038	119	731	3
carbon dioxide	0.035	125	770	5
glutamyltransferase	0.034	129	788	0
hydrogen	0.030	141	831	21
lactate dehydrogenase	0.028	148	859	4
dextran	0.026	157	879	21
sulfonate	0.022	174	929	0
trental	0.021	180	942	1
melphalan	0.021	183	948	1
iron	0.020	187	957	4
eicosapentaenoic acid	0.016	208	1016	0
dobesilate	0.015	209	1019	0
calcium dobesilate	0.015	210	1020	0
suloctidil	0.015	213	1027	1
histidine	0.015	214	1029	1
glycerin	0.013	223	1063	0
eicosapentaenoic	0.013	229	1073	0
fenclofenac	0.009	250	1134	0
diclofenac	0.009	251	1135	0
niceritrol	0.005	279	1206	0
fish oil	−0.052	598	1961	0

tions to Raynaud's is to find the nearest neighbor to the centroid of all documents comprising the Raynaud's literature—instead of to the term *Raynaud's.* The centroid of a cluster of items is its central value, and is computable in a variety of ways. Experiments showed that considering Raynaud's to be a document centroid, rather than a term, barely made a difference.

Directly Identifying Potential Discovery Literatures

It is interesting to note that, since Raynaud's and fish oil truly are medically related, and since this relationship can be detected by other methods, LSI does not directly uncover this latent association, especially since LSI scaling was performed on the blood viscosity literature, which is known to connect them.

The problem may be one of scale. By analogy, a glance at a globe suggests that New York City and Boston are near each other. But they are anything but neighbors when considering only the northeast seaboard of the United States. The same may be true of the Raynaud's–fish oil association. In the broad context of medicine, these concepts clearly are linked by the bridge of blood viscosity. Nevertheless, blood viscosity (used as the focus for LSI processing) may be an improper vantage from which to detect the association. We may need to ''back up'' to gain some perspective, just as we can only see that Boston and New York City are near each other when our perspective is the globe.

An experiment was conducted to explore this possibility by attempting to identify a potential discovery literature without first selecting an intermediate literature. In principle, we desired to analyze all of MEDLINE from the period 1980–1985 (nearly 780,000 records). For practical and computational reasons, this was not possible. Instead, we tried to obtain an approximately random sample of MEDLINE from the given date rage. We did this by obtaining all MEDLINE records written in English, containing an abstract and at least one reference, and with a publication date between 1980 and 1985. By doing so, 18,499 records were identified and downloaded for processing (a sample of about 2.5% for the period). It is possible that including only English-language items in the sample may have introduced some bias, for instance in the areas of pharmacology, where different areas of the world have approved different drugs for the same illness. This concern is reduced by noting that research performed in Europe and elsewhere around the world has a significant representation in the sample, since much scientific publication is in English. It is also possible, though unlikely, that the constraint that all records contain an abstract and reference(s) distorted the sample in some unintended fashion. Of course, the size of this sample means that some very small topics were likely excluded from it. For instance (see Table 7), if there are only 50 documents about a given topic in MEDLINE during the period 1980–1985, there is an approximately 30.5% chance that it will not be represented in the of 18,499 documents drawn. On the other hand, by the time a topic is of size 200, there is a 95% chance that the sample will contain at least two documents on that topic. All told, the method of sampling used likely provided a fair approximation to a genuinely random sample.

TABLE 7. Poisson approximations of size of topic in sample.

Size of topic = n = number of documents	E (number of documents ϵ topic in sample) = λ = $S*p$	Pr (sample contains k documents ϵ topic)		
		$k = 0$	$k = 1$	$k = \geq 2$
1	0.024	0.977	0.023	0.000
5	0.119	0.888	0.105	0.006
10	0.237	0.789	0.187	0.024
25	0.593	0.553	0.328	0.120
30	0.712	0.491	0.349	0.160
40	0.949	0.387	0.367	0.245
50	1.186	0.305	0.362	0.332
60	1.423	0.241	0.343	0.416
70	1.660	0.190	0.316	0.494
80	1.897	0.150	0.285	0.565
90	2.135	0.118	0.253	0.629
100	2.372	0.093	0.221	0.685
110	2.609	0.074	0.192	0.734
120	2.846	0.058	0.165	0.777
130	3.083	0.046	0.141	0.813
140	3.320	0.036	0.120	0.844
150	3.558	0.029	0.101	0.870
160	3.795	0.022	0.085	0.892
170	4.032	0.018	0.072	0.911
180	4.269	0.014	0.060	0.926
190	4.506	0.011	0.050	0.939
200	4.743	0.009	0.041	0.950

Size of MEDLINE (1980–1985) = 780,000;
S = sample size = 18,499;
p = topic base rate is MEDLINE = n/780,000.

LSI processing proceeded along the lines already described: The set of documents and terms was represented by $k = 300$ orthogonal factors (as opposed to 100 in the previous experiment) to adjust for the larger collection size. In this space, there were just over 36,000 terms or phrases that were not among those mentioned in the 1983–1985 Raynaud's document collection. From these new items, a list of the 1,000 closest neighbors to Raynaud's was generated. When we then hand-selected terminals from this list, we obtained a list of 37 items.

To ensure that an item did not occur in a Raynaud's document earlier than 1983, we consulted the entire MEDLINE document collection to find the number of documents published any time before 1986 that used both that item and the term *Raynaud's.* The cosine, rank, and intersection data for the hand-selected terminal items are shown in Table 8.

Among the list of items in Table 8 are those with already known connections to Raynaud's, including methysergide, hydralazine, and isoxsuprine. Although these cannot be considered discoveries, their inclusion rein-

TABLE 8. Terminal concepts identified as Raynaud's neighbors.

Term	Cosine (Raynaud's, term)	Rank in LSI	Rank among non-Raynaud's items	Number of hits* (Raynaud's ∩ term before 1986)
perhexiline	0.552	23	9	0
diltiazem hydrochloride	0.539	32	15	0 (13)
lidoflazine	0.425	116	69	0
dihydropyridine derivatives	0.409	135	84	0 (0)
nitrendipine	0.389	157	98	0
gallopamil	0.378	170	107	0
norverapamil	0.375	175	111	0
ergonovine maleate	0.347	227	145	0 (2)
nimodipine	0.337	254	162	0
dihydropyridine	0.295	349	228	0
methysergide maleate	0.294	351	229	0 (2)
methyldopa ad	0.288	372	242	0 (12)
oral nitrates	0.286	384	249	0 (1)
maleate	0.283	403	260	1
methysergide	0.260	494	313	2
indoramin ae	0.259	503	317	0 (2)
nisoldipine	0.258	504	318	0
nylidrin	0.238	626	408	0
ergonovine	0.229	692	454	2
antiplatelet therapy	0.225	711	467	1
hydralazine ae	0.203	899	598	3 (4)
chlorothiazide ae	0.194	1005	666	0 (0)
bepridil	0.192	1025	682	0
diazoxide	0.190	1071	710	0
nitrates	0.190	1073	712	1 (1)
aniline compounds	0.186	1138	759	1 (1)
ergotamine ae	0.183	1170	776	1 (7)
methoxamine	0.182	1185	788	0
captopril ad	0.182	1195	795	0 (4)
nicotinic acids	0.181	1203	802	24
isoxsuprine	0.179	1248	824	5
digoxin therapy	0.177	1266	837	0 (0)
hydralazine induced	0.173	1338	887	0 (4)
clonidine hydrochloride	0.169	1393	930	0 (2)
hydrazine	0.169	1404	936	0
lanthanum	0.169	1408	939	0
cyproterone acetate	0.164	1500	994	0 (3)

* Items with two values, like 0(13) for *diltiazem hydrochloride*, show (1) the size of the intersection with Raynaud's of the entire phrase (diltiazem hydrochloride ∩ Raynaud's = 0); and (2) the size of the intersection of its chief chemical constituent (diltiazem ∩ Raynaud's = 13).

forces the idea that LSI processing can help detect possible treatments for Raynaud's when a broad, unfocused literature (a random subset of MEDLINE) is processed without the benefit of a predefined connection, such as blood viscosity, to link them. Among the items in Table 8 are also substances never used before to treat Raynaud's that may deserve exploration as Raynaud's treatments if they were to pass the review of experts in medical therapeutics. These include vasodilating agents, such as perhexiline, diltiazen hydrochloride, nylidrin, and lidoflazine; drugs for treating ischemia, i.e., insufficient blood flow, such as dihydropyridine derivatives, including nitrendipine, gallopamil, nisoldipine, and bepridil; and antihypertensive drugs such as diazoxide, captopril, and clonidine hydrochloride.

We emphasize again that these analyses and all others we have shown support, but do not automate, discovery, and that their appropriate interpretation should come from medical researchers familiar with the topic. For instance, several of the drugs mentioned in Table 8 are calcium channels blockers; and the nonterminal phrase *calcium blocking* has a very high cosine (0.573). So, without additional evidence to the contrary, a possibility is that calcium channel blockers may be effective in the treatment of Raynaud's, and the nonterminal concept, calcium channel blocking, could itself be analyzed as an intermediate literature (its literature downloaded, parsed, and statistics computed) in the search for terminals disjoint from Raynaud's. On the other hand, research pharmacologists familiar with calcium channel blockers might know, for example, that those that affect peripheral blood flow (such as nifedipine) have already been tested as treatments for

Raynaud's, whereas the calcium channel blockers in Table 8 affect the heart, thus making them ineffective as Raynaud's treatments. So for those with the requisite background knowledge, the computed statistics should help stimulate useful conjectures that may lead to the discovery of scientific hypotheses.

None of the terms in Table 8 with empty intersection with Raynaud's was among the list of nonintersecting terms in Table 6 (the equivalent table for the previous experiment, where the blood viscosity literature was LSI-processed). In fact, only one term, *isoxsuprine,* was common to both tables even when we consider both terms with empty and nonempty intersection with the term *Raynaud's.* As we suspected, a MEDLINE focus for LSI has certainly produced a different set of Raynaud's near neighbors than did a blood viscosity focus. In this sense, LSI processing of MEDLINE to search for potential discoveries directly is another tack to consider in attempting to uncover latent medical discoveries.

We considered a variation of this method in an attempt to adopt the MEDLINE focus for LSI processing while retaining some of the advantages of considering blood viscosity an intermediate literature: We restricted the list of items in Table 8 to those that were both bibliographically disjoint from Raynaud's and present in the blood viscosity literature. Only three items met these criteria: *methyldopa ad, methoxamine,* and *captopril ad.* However, in looking at articles on methyldopa and captopril, we learned that both had been studied as a treatment for Raynaud's. The reason for this apparent contradiction is that the phrases identified, *methyldopa ad* and *captopril ad,* where *ad* is a MEDLINE subheading meaning ''administration and dosage'' were not used in the Raynaud's literature, even though both of these drugs had been written about without the *ad* subheading. Methoxamine causes vasoconstriction and, as such, would be contraindicated for Raynaud's.

Summary and Discussion

Our investigation suggests that latent semantic indexing might be a useful tool in literature-based discovery. Because of the difficulty of the task, literature-based discovery may be totally unsuccessful for certain problems, or by certain methods. LSI provides another technique that can be considered in looking to uncover hidden discoveries.

We have shown that latent semantic indexing might be a useful technique in either of the two phases of literature-based discovery. During the search for intermediate literatures, it fairly closely reproduces (but extends) the same set of highly ranked terms and phrases that Gordon and Lindsay (1996) have shown are a useful starting point for literature-based discover. In helping identify potential discovery literatures, LSI can be used in either of two ways: by factoring a set of documents associated with a suspected intermediate literature, or by analyzing the larger literature (MEDLINE, in this study) that forms the universe of discourse. In studying new discoveries in connection to Raynaud's disease, the first method was able to identify fairly prominently a chief chemical constituent (eicosapentaenoic acid) in fish oil using the literature from a time when the healthful effects of fish oil on Raynaud's were unknown. The phrase *fish oil* was not nearly as prominent. The second method revealed a very different set of substances.

It is important to remember that tools and analyses like those we have described in this paper support, but do not in any way replace, scientists. The skilled scientist may see patterns in data like those we report that derive from his or her knowledge of the field. One scientist may see, for example, that a particular class of drugs is prominently represented in the data and begin to form hypotheses about this drug class's ability to treat Raynaud's. A pharmacologist with a more complete background in the area may know that certain of these drugs are primarily known for their effects on the heart, rather than the peripheral vascular system. This type of knowledge could help isolate the drugs that truly merit scientific investigation by suggesting a more focused analysis.

The premise behind literature-based discovery support is that medical specialization makes it virtually impossible for a scientist to stay abreast of developments in areas outside his or her area of direct interest. As a consequence, important connections crossing disciplinary boundaries may never be noticed. Literature-based discovery support tools can help organize the knowledge of scientific fields that lie outside a scientist's direct specialization, thus improving his or her ability to organize and make use of this information.

LSI is one tool that may help in this effort. Additional research is needed to provide a broader array of tools. Among other tools that we are investigating are those for: (*1*) reporting data at several levels of abstraction (e.g., counting as statistical evidence for calcium channel blockers any drug that is in this drug family); (*2*) looking for evidence suggestive of ''causal'' relationships in the literature (which may be revealed independently of their statistical prominence); and (*3*) using semantic and category knowledge to improve the step of identifying terminal concepts, which is now a completely intellectual process. Through these efforts, we hope to provide scientists methods that support their efforts to generate discovery hypotheses that lie latent in the published literature.

Acknowledgments

This research was conducted while Michael Gordon was on sabbatical at Bellcore. He thanks Tom Landauer and Michael Lesk for that opportunity. This work benefitted from discussions with Tom Landauer, George Furnas, Jeff Zacks, Robert Lindsay (University of Michigan), and Don Swanson (University of Chicago). The authors also thank the anonymous referees for their careful reviews and their help in strengthening the medical and methodological arguments contained in the paper.

References

Deerwester, S., et al. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1977). *Computer methods for mathematical computations* (chapt. 9). Englewood Cliffs, NJ: Prentice Hall.

Gordon, M. D., & 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. *Journal of the American Society for Information Science, 47,* 116–128.

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

Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly, 56,* 103–118.

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science, 38,* 228–233.

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

Swanson, D. R. (1988b). Unnoticed connections in the literature of medicine: Implications for knowledge representation and natural-language searching. 1988 ASIS Mid-Year Meeting, Ann Arbor, MI.

Swanson, D. R. (1989a). A second example of mutually isolated medical literatures related by implicit unnoticed connections. *Journal of the American Society for Information Science, 40,* 432–435.

Swanson, D. R. (1989b). Online search for logically related noninteractive medical literatures: A systematic trial and error strategy. *Journal of the American Society for Information Science, 40,* 356–358.

Swanson, D. R. (1989c). Medical literatures as a source of new knowledge. *USDE Final Report,* Dec. 1989.

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

Swanson, D. R. (1990b). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association, 78,* 29–37.

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. *Proceedings of the Fourteenth Annual International ACM SIGIR Conference,* (pp. 280–289).

Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends, 41,* 606–631.