Journal of Biomedical Informatics 43 (2010) 240–256

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections

Trevor Cohen [a,*], Roger Schvaneveldt [b], Dominic Widdows [c]

[a] Center for Cognitive Informatics and Decision Making, School of Health Information Sciences, University of Texas, Houston, USA
[b] Applied Psychology Unit, Arizona State University, Arizona, USA
[c] Google Inc., USA

ARTICLE INFO

Article history:
Received 24 April 2009
Available online 15 September 2009

Keywords:
Distributional semantics
Literature-based discovery
Implicit associations
Indirect inference

ABSTRACT

The discovery of implicit connections between terms that do not occur together in any scientific document underlies the model of literature-based knowledge discovery first proposed by Swanson. Corpus-derived statistical models of semantic distance such as Latent Semantic Analysis (LSA) have been evaluated previously as methods for the discovery of such implicit connections. However, LSA in particular is dependent on a computationally demanding method of dimension reduction as a means to obtain meaningful indirect inference, limiting its ability to scale to large text corpora. In this paper, we evaluate the ability of Random Indexing (RI), a scalable distributional model of word associations, to draw meaningful implicit relationships between terms in general and biomedical language. Proponents of this method have achieved comparable performance to LSA on several cognitive tasks while using a simpler and less computationally demanding method of dimension reduction than LSA employs. In this paper, we demonstrate that the original implementation of RI is ineffective at inferring meaningful indirect connections, and evaluate Reflective Random Indexing (RRI), an iterative variant of the method that is better able to perform indirect inference. RRI is shown to lead to more clearly related indirect connections and to outperform existing RI implementations in the prediction of future direct co-occurrence in the MEDLINE corpus.

1. Introduction

This paper addresses the issue of indirect inference, finding meaningful connections between terms that are related but do not occur together in any document in a collection. Indirect inference is useful in many applications including information retrieval because documents that do not contain words in a query may be relevant to a user's information need. Thus, retrieval systems that reach beyond query terms can improve performance. Indirect inference is particularly important in the context of developing tools to aid discovery from literature because, by their very nature, discoveries are likely to involve bringing together ideas that have not occurred together previously.

In previous applications, implicit connections between two terms that do not co-occur have been discovered by finding a third bridging term that occurs directly with each of them, according to the discovery paradigm first proposed by Swanson [1]. Several automated methods of knowledge discovery based on this paradigm have been developed and evaluated in the literature [2]. However, given the number of possible combinations of bridging terms and potential discoveries, methods that are able to identify indirect connections without the explicit identification of bridging terms present an attractive alternative. The ability to directly identify implicit connections offers considerable computational advantages on account of the combinatorial explosion that occurs with the number of bridging terms permitted in the chain from cue concept to discovery. Even with only one linking term, the search for a novel discovery requires the following five stages, as described by Yetisgen-Yildiz and Pratt [3]:

1. Terms directly co-occurring with a given starting term are retrieved using a correlation mining approach.
2. A set of these ranked above some predetermined threshold are selected as linking terms.
3. Terms directly co-occurring with each of these linking terms are retrieved using a correlation mining approach, and are selected as target terms.
4. Those terms directly co-occurring with the starting term are excluded (consequently the end result of this process is indirect inference).
5. The remaining terms are ranked using a ranking approach.

This process carries considerable computational and disk I/O expense which limit the possibilities for highly interactive and responsive discovery support tools, unless significant constraints

* Corresponding author.
E-mail address: Trevor.Cohen@uth.tmc.edu (T. Cohen).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.003

are placed on the discovery search space. As the process does not utilize indirect inference, linking two terms that do not co-occur requires the construction of a path through the discovery space that traverses terms that co-occur directly. This process would be more accurately described as "direct inference" as an explicit pathway from source to target must be established before a discovery can occur. By comparison, corpus-derived statistical models of semantic relatedness such as Latent Semantic Analysis (LSA) [4] are able to identify directly meaningful associations between terms that do not co-occur. For example, in LSA, each term is represented as a vector, and meaningful connections between terms that do not co-occur can be retrieved and ranked using the following process:

1. Retrieve the vector for the starting term.
2. Compare this to the vector for all possible target terms.
3. Exclude those which directly co-occur with the starting term.

The performance of this approach is further enhanced by the condensed nature of the vector space representation—it is possible to maintain the vectors for all possible target terms in RAM, and disk lookup is required for the final step in this process only. LSA is one of several methods provided by the emerging field of distributional semantics that are able to learn meaningful associations between terms from the way in which they are distributed in natural language text. Of these methods, LSA [4] in particular has been shown to make meaningful estimates of the semantic relatedness between terms that do not co-occur directly. This suggests that such models may be useful as means to discover implicit connections in the biomedical literature without the need to explicitly identify a bridging term.

However, the generation of the condensed vector space representations utilized by such models often carries considerable computational cost. LSA, for example, is dependent upon the singular value decomposition (SVD), a computationally demanding method of dimensionality reduction to draw such associations. Consequently, LSA requires computational resources beyond the reach of most researchers to scale to large corpora such as the MEDLINE corpus of abstracts. In this paper we address this issue by evaluating the ability of Random Indexing (RI), which has recently emerged as a scalable alternative to LSA, to derive meaningful indirect inferences from general and biomedical text. We find the original implementation of this method is somewhat limited in its ability to indirectly inference, and propose and evaluate Reflective Random Indexing (RRI), a methodological variant that is customized for this purpose.

The primary motivation of this paper is to demonstrate that the indirect inferencing ability of RI is vastly improved when an iterative approach is utilized. As the scalability advantages of RI are retained, this improvement has significant implications for information retrieval and distributional semantics in general. In addition, we wish to evaluate these models as tools to support literature-based discovery.

The organization of the paper is as follows. Section 2 introduces indirect inference, provides illustrative examples of the ability of LSA to perform indirect inference, and discusses the significance of this ability for the discovery of implicit connections in biomedical text. Section 3 describes RI and its variants, sliding-window (or term–term) based RI, and Reflective Random Indexing, also with illustrative examples of the ability of these models to derive meaningful indirect inferences. In Section 4 we evaluate the ability of variants of RI to simulate Swanson's original discoveries of implicit connections between Raynaud's Disease and dietary fish oil, and migraine and magnesium. In addition, we present a large-scale evaluation of the abilities of these models to derive meaningful indirect inference from a time-delimited segment of the MEDLINE database. A discussion of these results and conclusion follow.

2. Background

2.1. Indirect inference

In the context of distributional models of semantic relatedness, an indirect inference is considered to be a measurable semantic relation between two terms that do not co-occur directly together in the corpus used to generate the model concerned. A simple network model of indirect inference can be generated by considering terms that co-occur in documents to be directly linked. In such a model, indirect inferences correspond to terms that are not directly connected but are connected to the same other terms. With such a model there are various ways to determine the "strength" of the indirect inferences including the number of shared intermediate connections and the strength of the direct connections involved. Paths with more than one intermediate node could also be considered. The important property to be preserved is to discriminate between terms that co-occur and those that do not co-occur but are connected by short paths. Indirect inference is particularly important in the domain of information retrieval, as an information request based on a search term would ideally retrieve related documents that do not contain this term. This issue was a primary motivation for the development of methods such as Latent Semantic Indexing (LSI) [5] that are able to retrieve with accuracy related documents that do not state a particular search term explicitly.

Indirect inference has much in common with the traditional use of "middle terms" in logic, as introduced by Aristotle (Prior Analytics Bk. 1 Ch. 4 and thereafter). For example, in the inference "Socrates is human, humans are mortal, therefore Socrates is mortal", "human" is the middle term, and empirically is the term we would like to find in text to answer the question "Is Socrates mortal?". Middle terms through which inferences are made often appear as bound variables in computational logic, and just as several paths may be chosen through a network, the same formal inference may be made by way of several different functional arrangements of bound variables. In practice, performing any such process by analyzing human language is fraught with difficulty, as Aristotle points out, for example:

> "*It is clear that the middle must not always be assumed to be an individual thing, but sometimes a phrase . . .. That the first term belongs to the middle, and the middle to the last, must not be understood in the sense that they can always be predicated of one another, or that the first term is predicated of the middle in the same way as the middle is predicated of the last.*"
>
> — Prior Analytics, Bk. 1. Ch. 35, 36.

In the context of literature-based knowledge discovery middle terms are generally referred to as "bridging terms" or "linking terms", terminology we will employ for the remainder of the paper. While accepting such complexities as term compounding and ambiguity of relationships, fields such as computational semantics and literature-based knowledge discovery have sought and to some extent found methods for traversing middle terms automatically in ways that can enable the more rapid discovery of potentially interesting connections in scientific information, as will be described in this paper.

2.2. Indirect inference and literature-based knowledge discovery

This capacity to indirectly inference is of particular interest to researchers in the field of literature-based discovery, a field which can trace its inception to the fortuitous discovery of a previously unpublished therapeutic relationship between fish oil and Raynaud's Disease, a circulatory disorder affecting the peripheral vasculature, by Don Swanson [1]. The premise underlying Swanson's

approach is that terms "A" and "C" from two disjoint literatures can be connected by a third term, term "B" that has some (direct) connection to both A and C. In the Raynaud's discovery, searching the literature for Raynaud's Disease led to the discovery of a few reports of raised blood viscosity and reduced red blood cell deformability in Raynaud's patients, suggesting literature on "blood factors" as a potential source of "B" terms. This literature was then searched for titles that did not include any reference to Raynaud's, revealing that blood viscosity was reduced, and red cell deformability increased, by dietary fish oil [1]. Swanson describes these literatures as "logically connected", in that they are linked by an implicit scientific argument (fish oil affects several patho-physiological factors which may be implicated in Raynaud's Disease, blood viscosity being one example). Note that this discovery, like other documented literature-based discoveries, constitutes an indirect inference: two terms that do not directly co-occur in the body of text used to generate this discovery have been associated with one another.

This scheme allows for two modes of discovery, open and closed. The open mode of discovery involves the generation of a new hypothesis, and consequently can be considered an example of abductive reasoning as proposed by Peirce [6]. This mode of discovery has two steps: first a "B" term (such as "viscosity"), or set of B-terms are identified. For example, if the "C" term under consideration concerns a disease, the choice of "B" terms may include patho-physiological mechanisms that present likely targets for therapeutic agents. Once these "B" terms have been identified the second step involves identifying potential "A" terms (such as the term "fish" or "oil"). In accordance with the argument of Bruza et al. [7], we propose that as the open mode of discovery simulates abductive reasoning, the constraint that only logically consistent connections between concepts should be considered is too strong for this process. This argument is consistent with cognitive models of abduction, and renders this process amenable to computationally tractable simulation. The closed mode of discovery involves the justification of the hypothesis once discovered, and is amenable to simulation using rule-based methods [8] although it is also possible to provide some explanation for indirect inferences using Pathfinder network scaling to reveal the most significant links between related concepts [9]. Numerous authors have explored the possibilities of automated knowledge-discovery systems based on Swanson's paradigm. These systems vary in their approach. Some are based on co-occurrence statistics of terms [10,11] while others draw upon knowledge resources such as the Unified Medical Language System (UMLS) [12], or Medical Subject Heading (MeSH) terms [13]. An exhaustive review of these systems is beyond the scope of this paper, but several such reviews exist in the literature and we refer the interested reader to the work of Weeber et al. [2], Ganiz et al. [14] and Kostoff et al. [15] for comprehensive reviews of developments in the field. For the purpose of this paper, we restrict our discussion to those systems that attempt to infer a quantitative measure of the indirect similarity between terms using corpus-based models of semantic relatedness that have been evaluated extensively in the cognitive science literature.

Gordon and Dumais propose the use of Latent Semantic Indexing (LSI) for literature-based knowledge discovery[1] [16]. While the authors do demonstrate the ability of LSI to identify "B" terms, the first step of Swanson's two-step process, an attempt to simulate the Raynaud's discovery using indirect inference was not successful. The authors were not able to promote the terms 'fish' or 'oil' to near the top of a ranked list of indirect neighbors of the term "raynaud". Interestingly, eicosopentaenoic acid, the active ingredient of fish oil, was recovered as the 208th ranked indirect neighbor of "raynaud", but this is a lower rank than one would anticipate a human user of a knowledge-discovery system exploring. The study also revealed some other plausible therapeutic alternatives for Raynaud's. The authors acknowledge that computational limitations forced them to analyze a subset (18,499) of the desired MEDLINE records (780,000) from the period between 1980 and 1985.

Bruza, Cole and colleagues present a similar approach to literature-based knowledge discovery using the Hyperspace Analogue to Language (HAL) [17] approach [7,18]. Like LSA, HAL represents terms as high-dimensional vectors. However rather than treating each individual document as a context for co-occurrence, HAL employs a sliding-window that is moved through the text to generate a term–term co-occurrence matrix. Unlike most work in literature-based knowledge discovery, Bruza and Cole's takes a cognitive approach to this problem, highlighting the empirical and theoretical support for semantic spaces as models of meaning, and proximity within such spaces as a basis for abduction. Bruza et al. further motivate the use of semantic spaces from an operational perspective, arguing that (a) no automated system for the large-scale extraction of propositional logic from the literature exists, and (b) such systems are unlikely to yield a computationally tractable solution to this problem. This argument for economy in computational implementation is discussed in the context of an argument for cognitive economy presented by Gabbay and Woods [19], who make a similar point about the constraints placed on symbolic logic by the limitations of the human cognitive system. Bruza and Cole's research confirms Gordon and Dumais' finding that distributional statistics can support the discovery of B-terms. In addition, by weighting those dimensions of the vector for "Raynaud" that correspond to the vectors representing a manually curated set of these "B" terms, the terms "fish" and "oil" are promoted to among the top 10 ranked results when particular distance metrics and statistical weighting functions are used, effectively replicating Swanson's discovery using as a corpus the same set of MEDLINE titles from core clinical journals between 1980 and 1985 that Swanson himself employed. However, these terms are ranked far lower when other metrics and weighting functions are used, and it does not necessarily follow that the methods used to simulate this particular discovery would be the best choice for other potential discoveries.

2.3. LSA and indirect inference

Several computational models that derive estimates of semantic relatedness from unannotated natural language text have been developed and evaluated over the past decade (for a review, see [20]). Perhaps the best known of these within the cognitive science community is LSA. LSA seeks to identify "latent semantics", the meaning that underlies a particular term or passage of text regardless of the specific words used to convey this meaning. Consequently, the ability to generate indirect inferences is of fundamental importance to LSA. Synonyms tend not to co-occur with one another directly, so indirect inference is required to draw associations between different words used to express the same idea. LSA models the semantic relatedness between terms using a spatial metaphor: terms within a large corpus of text are projected into a high-dimensional semantic space by first generating a term-document matrix (usually applying local and global statistical weighting metrics rather than using raw term-frequency), and then reducing the dimensions of this matrix using SVD, an established technique of linear algebra. Distance between terms in this space is usually measured using the cosine metric, or normalized scalar product, providing a convenient measure of semantic relatedness. LSA has been shown to approximate human performance

[1] LSA was originally implemented to index documents for information retrieval, and is usually referred to as LSI when used for this purpose.

in a number of cognitive tasks including the Test of English as a Foreign Language (TOEFL) synonym test [4], the grading of content-based essays [21] and the categorization of groups of concepts [22].

LSA's performance has been attributed to its ability to make indirect inferences. According to Landauer and Dumais' analysis of the rate with which LSA's knowledge of TOEFL test terms improves as new text passages are introduced, "most of (LSA's) acquired knowledge was attributable to indirect inference rather than direct co-occurrence relations" [4]. We present a few examples to illustrate this inferencing capability of LSA. Table 1 shows the nearest-indirect neighbors, those terms in semantic space that are closest to a cue term but do not co-occur directly with it in any document, of a few cue terms. The semantic space used for this example is derived from the Touchstone Applied Sciences (TASA) Corpus, the same corpus used in the TOEFL test evaluations of LSA, using the General Text Parser software package [23].

LSA has identified a number of interesting relations between terms that do not co-occur directly. Take for example the nearest-indirect neighbors of the term "jazz". These include composers of classical music, musical instruments and other musical genre. The nearest-indirect neighbors of "nicotine" include another inhaled substance, terms related to the possible legal consequences of the use of this substance, and the "bronchioles", which are affected by lung disease caused by nicotine abuse. These terms were selected as they produce interesting indirect neighbors in a LSA-derived space, however finding them was not difficult as more often than not some meaningful indirect neighbors were obtainable.

3. Random Indexing and indirect inference

3.1. Random Indexing

Indirect inference has been shown to be an important component of LSA's knowledge acquisition process. In accordance with the position taken by previous authors [7], we posit that proximity within a semantic space can be employed as a basis for the computational modeling of abductive inference. The generation of such inferences within the biomedical literature has been used to replicate aspects of Don Swanson's seminal literature-based discovery. Given the results achieved with these methods with the smaller corpora described above, it is probable that a system able to generate meaningful indirect inferences from the entire MEDLINE corpus of abstracts (around 9,000,000 documents and 1.25 billion terms) could support the discovery of new links between disparate bodies of literature contained within the MEDLINE database. Extending this computational model of abduction to a corpus this size would support the generation of meaningful inferences from volumes of text that are far beyond human capacity to read, and include a broader range of literature in the search for novel connections. Dimension reduction using SVD underlies the generation of meaningful indirect inference in LSA. However, SVD requires the representation of a full term-document matrix (initially with 9,000,000-dimensional vectors in this case) in RAM, and the computational demands of the SVD itself preclude the computation of a reduced-dimensional representation of a matrix this size within the limits of the computational resources available to most researchers today.

Random Indexing (RI) [24,25] has recently emerged as a scalable alternative to LSA for the derivation of spatial models of semantic distance from large text corpora. For a thorough introduction to Random Indexing and hyper-dimensional computing in general, see [26]. In its simplest form, Random Indexing involves two phases, allocation of elemental vectors and training.

3.1.1. Elemental vector allocation

Let n be the dimension of the semantic space, and let $k \ll n$ be a small constant. The function `allocate_elemental_vector` proceeds as follows:

```
allocate_elemental_vector(n, k):
  let v be the zero vector of dimension n.
  for i up to k:
    Change one of the zero coordinates in v to +1, chosen arbitrarily.
    Change one of the zero coordinates in v to −1, chosen arbitrarily.
  return v.
```

Deterministic pseudo-random variants of `allocate_elemental_vector` can be implemented, by passing in the random seed as an argument, and by allocating a vector from values in a hash of the string identifier for the object in question (e.g., a term or a document path identifier), as used in Bloom filter algorithms. Elemental vectors generated in this fashion are sometimes referred to as "index vectors" or "basic vectors": we use the term elemental because it is not easily confused with other core concepts (such as learned vectors in an index, or basis vectors for the reduced space). One of the core properties of elemental vectors is that two elemental vectors are (on average) orthogonal to each other [26].

3.1.2. Vector allocation for a set

Let D be a set of p elements. We define the function `allocate_elemental_vectors` as follows:

```
allocate_elemental_vectors(p, n, k):
  initialize matrix D of size (p, n) to zero.
  for i less than p:
    D[i] = allocate_elemental_vector(n, k)
  return D.
```

Table 1
Nearest-indirect neighbors of terms in a semantic space (281 dimensions) derived from the TASA corpus using LSA. Each column contains the nearest-indirect neighbors of a cue term, as well as the strength of association between these neighbors and the cue term as measured using the cosine metric.

Nebula	Jazz	Semantic	Picasso	Nicotine
0.72: luminosity	0.79: beethoven	0.63: phonological	0.79: expressionism	0.81: casefinding
0.65: hubble	0.74: sonatas	0.61: phonics	0.79: impressionism	0.63: circumstantial
0.63: magellani	0.73: guitars	0.56: prefix	0.78: courbet	0.61: homicides
c 0.62: centauri	0.74: lullabies	0.55: morphemes	0.76: surrealists	0.58: spokesmen
0.58: algol	0.73: autoharps	0.54: confluence	0.76: pollock	0.57: bronchioles
0.58: pleiades	0.73: lyrics	0.51: morpheme	0.75: impressionist	0.54: burley
0.57: neutrinos	0.72: duple	0.51: suffixes	0.73: claes	0.52: peptic
0.56: supergiant	0.71: autoharp	0.50: correspondences	0.73: brushstrokes	0.52: golfing
0.56: proxima	0.71: motown	0.49: suffix	0.73: camille	0.52: cannabis
0.56: sirius	0.70: haydn	0.49: shetlands	0.73: expressionist	0.51: marijuana

Each row of the matrix D can be thought of as a vector in an n-dimensional space, so `allocate_elemental_vectors` creates a set of such vectors, expected to be almost orthogonal to one other, which are in one to one correspondence with the elements of the set D. In practice, the matrix D may be represented sparsely or regenerated in different locations if a deterministic `allocate_elemental_vector` function is used.

3.1.3. Term document corpus training

Now let each document d in D be also an ordered list of elements called tokens, each of which is a token of a particular type called a term. (The set D may now be called a corpus.) Let `T` be the set of terms and let the number of terms be q. Let M be the term-document matrix of the corpus, that is, M is of size (q, p) and $M[i][j]$ records the number of times the ith term occurs in the jth document, We define the function `train_model` as follows:

```
train_model(M, D):
  initialize matrix T of size (q, n) to zero.
  for i less than q:
    for j less than p:
      T[i] = T[i] + weight(M[i][j] * D[j]).


  (optional) foreach i less than q:
    normalize(T[i]).
  return T.
```

In this definition we have been less explicit in passing in dimension values, these are easily inferred from the other arguments, or accessed as global constants. Here `weight` is some weighting function such as tf-idf or log-entropy. The normalization phase is standard in most applications, to prevent search from giving excessive weight to either frequent or infrequent items. However, normalization is a lossy operation, which raises complications that must be taken into account when creating distributed or incremental implementations.

It should be noted that without the weight function or normalization phase, `train_model` is simply an implementation of matrix multiplication of the form $T = MD$, optimized to iterate over terms in a sparsely populated inverted index. This observation can be used to account for some of the observed properties of Random Indexing, including the convergence of the reflection algorithm over many iterations which is demonstrated in our experiments.

3.1.4. Computational observations

The Random Indexing process has an enormous computational advantage over methods requiring SVD for dimension reduction, in both space and time requirements. In space requirements, the matrix M can be computed in parallel and represented sparsely using standard indexing algorithms. Then in the training phase, M can be accessed sparsely one row at a time; only the matrices D and T representing the reduced vectors have to be kept in main memory throughout, so the main memory footprint of the training process scales with $n(p + q)$, or even $nq + kp$ if the elemental document vectors are represented sparsely. Compared with this, a standard SVD algorithm holds the full unreduced matrix in memory, and the matrix M, being of size pq, is much larger than $n(p + q)$, since the reduced dimension n is much smaller than either the number of terms q or the number of documents p. In time requirements, the process is essentially linear in the size of the document collection multiplied by the reduced dimension, whereas the time complexity of SVD is essentially cubic, often quoted as $O(qp^2)$ [27] In practice, our implementation in the open source Semantic Vectors package [28] has enabled models to scale to corpora several times the size of those processed using singular value decomposition, though a sparse SVD implementation may help to fill this gap if a suitable candidate were publicly available.

Note that our formulation of Random Indexing is independent of the number field over which the vector space is created, and the theory applies equally to any ground field including real and complex numbers. In practice, our implementation using the Semantic Vectors package is cast in terms of real numbers, represented using 4 byte floating point approximations.

RI and other related dimension reduction algorithms such as Random Projection [29] rest on the Johnson–Lindenstrauss Lemma [30] which states that the distance between points in a high-dimensional space will be approximately preserved if they are projected into a lower-dimensional random subspace of sufficient dimensionality. RI has been shown to perform comparably well with SVD-based LSA on the TOEFL synonym test, which is often used to evaluate computational models of semantic relatedness, using the same training corpus [24,25]. In addition, RI has been shown to draw meaningful associations from the entire MEDLINE corpus, including a drug-disease association previously undetected by MEDLINE search between the terms ‘thrombophilia’ and ‘rppgf’ (an inhibitor of platelet aggregation) which do not co-occur directly in any MEDLINE abstract [31].

3.2. RI and indirect inference

3.2.1. RI as originally implemented

Despite this extraction of a meaningful indirect inference between ‘thrombophilia’ and ‘rppgf’, there is reason to believe that RI as originally implemented (and described above) is not optimal for the derivation of meaningful indirect connections. Each document is assigned an elemental vector that is almost orthogonal to that of every other document, and a term is represented as a linear sum of the index vectors for each document it occurs in. If term A does not occur in any document with term B, the representation of term A should be the linear sum of a set of vectors nearly orthogonal to all those that constitute the basis for the representation of term B. Consequently, one would expect the ability of this model to generate meaningful indirect inferences to be somewhat limited when compared to LSA. Table 2 shows the nearest-indirect neighbors of the same set of terms in Table 1, in a 2000-dimensional RI space derived from the TASA corpus using the Semantic Vectors software package [28]. The low cosine values in this example are not necessarily an indication of ‘weaker’ indirect connections, as while the mean cosine and standard deviation between sets of terms in RI spaces constructed in this manner tend to be much lower than those generated using LSA, the relative distance between terms as measured by this metric is still generally meaningful (at least as far as directly related terms are concerned). Also note that these cosine values have been rounded to three decimal places for the purpose of presentation, but are in fact not identical to one another even though they may appear the same once rounded.

Unlike the results obtained by LSA, the generation of a plausible connection between these terms and their neighbors (aside from a few exceptions such as “jazz” and “jazzing”) requires no small amount of inference on the part of the observer. While it may be possible to construct such associations (for example jazz-performer-chaplin), it is also possible that these associations were derived by chance overlap between elemental vectors. If the elemental vector for a document containing the term “jazz” happened to overlap with the elemental vector for a document containing the term “pda”, this may produce sufficient similarity to promote “pda” among the 10 nearest-indirect neighbors of jazz. The presence of meaningful conceptual clusters within some of

Table 2
Nearest-indirect neighbors of terms in a 2000-dimensional semantic space derived from the TASA corpus using RI. Each column contains the nearest-indirect neighbors of a cue term, as well as the strength of association between these neighbors and the cue term as measured using the cosine metric.

Nebula	Jazz	Semantic	Picasso	Nicotine
0.120: deux	0.107: menon	0.117: misting	0.114: cutlasses	0.106: foxhound
0.108: washlines	0.107: priates	0.116: scullary	0.109: kickingbird	0.106: froghopper
0.098: throught	0.102: whch	0.110: scullery	0.106: herter	0.100: outfought
0.098: tieds	0.107: pirsig	0.108: defers	0.102: lussier	0.100: moffitt
0.098: emilie	0.102: stoneface	0.108: lerner	0.102: defecation	0.094: producton
0.098: mutagen	0.099: chaplin	0.101: expalin	0.102: offshot	0.094: fernandezes
0.098: noncategorical	0.096: inflationalry	0.101: nuthatches	0.101: encompassed	0.094: genet
0.098: jeem	0.094: jazzing	0.099: stockinged	0.096: envelop	0.094: pegasos
0.098: multihandicapped	0.094: arteriosus	0.098: uncompleted	0.095: murstein	0.094: phorcys
0.098: strehler	0.094: pda	0.097: afterdinner	0.095: ratched	0.094: handcopyin

these neighbors supports this hypothesis. For example, even though 'pda' and 'arteriosus' are not obviously related to 'jazz', they are related to one another: pda stands for Patent Ductus Arteriosus, a congenital disorder of the cardiovascular system.

3.2.2. Sliding-window based implementations of RI

The RI space used in the previous example was constructed using term-document statistics: each document in the corpus is treated as a context, and each term is represented according to the contexts in which it occurs. LSA also derives an estimate of semantic distance from per-document statistics for each term. That is to say, before dimension reduction each term is represented by a vector with one dimension for every document in the corpus, a term-document matrix. An alternative approach to the derivation of semantic distance from unannotated electronic text is presented by the HAL model [32]. In contrast to LSA, HAL derives its estimates from the co-occurrence statistics between terms in a sliding-window that is moved across the entire corpus. Each term in the unreduced matrix is represented by a vector with a dimension for each of a series of terms, a term–term matrix. These two indexing procedures result in different models of the same corpus.

3.2.2.1. Sliding-window corpus training. With the sliding-window approach, we use the fact that each document d in the set *D* is an ordered set of tokens (*t_a, t_b, ... etc....*), where each token is an instance of one of the terms in the set T. Let *T0* be a matrix of elemental term vectors of size (q, n) created using the function `allocate_elemental_vectors(q, n, k)`, so that rows of *T0* correspond to elements of T just as rows of the matrix *D* corresponded to elements of the document set *D* earlier. Let *w* be the distance between terms within which they are considered related. The function `train_sliding_window_model` proceeds as follows:

```
train_sliding_window_model(D, T0, w):
  initialize Tl matrix of size (q, n) to zero.
  For each d in D:
    for each pair (t_a, t_b) in d:
      if distance(t_a, t_b) < w:
        Tl[a] = Tl[a] + weight(t_a, t_b) * T0[b].


  (optional) for each i less than q:
    normalize(Tl[i]).
```

The weight function in this case usually involves some function of the distance between the two terms, in addition to other weighting signals. In particular, setting a maximum on the distance between two terms reduces the complexity of the inner loop from being quadratic in the size of the document to being linear in the document size multiplied by the width of the sliding-window.

The implementation of `train_sliding_window_model` can be simplified and optimized using a standard postings list for the corpus, in much the same way that `train_model` can be optimized using a term-document matrix, though in this case we believe that the underlying algorithm can be understood more easily by directly using the notion of a document as an ordered list of terms.

3.2.2.2. Sliding-window associations. Recent work on spatial models of meaning [33,25] has attempted to distinguish between the different types of associations derived by sliding-window (term–term) and term-document based models. Term–term spaces are shown to generally perform better on synonym, antonym and part-of-speech tests, while term-document spaces perform better on word-association tests [33]. The performance of term–term spaces on synonym tests is best when narrow sliding-windows (for example two terms to the left and two to the right of a focus term) are used, a result that is consistent with that obtained by Rapp who uses a similar approach to obtain a score of 92.5% on the TOEFL synonym test [34]. Word-association tests measure correlation between semantic distance as estimated by the model and the frequency with which particular terms are retrieved from human memory in response to a cue term. These require the model to reproduce a more general sort of association that is not constrained by syntax. From the perspective of literature-based knowledge discovery, both of these sorts of association are valuable, but there is no reason to constrain the discovery process to terms with similar syntactic roles. What is clear, however, is that when RI is used, term–term spaces produce more meaningful indirect associations than term-document indexing, as illustrated by the examples in Table 3, which were also generated using the Semantic Vectors [28] package.

Several of the indirect neighbors derived using a term–term model are meaningfully related to the cue term. For example, the indirect neighbors of "nebula" include other celestial bodies (meteoroids and comets) and the luminescent "anglerfish". The indirect neighbors of "jazz" include a number of terms related to music such as "songwriters", "instrumental" "kapellmeister" and "lutenists".

3.3. Limitations of established methods

RI as it was originally implemented does not address the ability to make meaningful indirect inferences. While these do occur, the examples provided suggest they occur far less frequently than when RI with a term–term approach is used, and LSA produces more interesting indirect neighbors than either of these approaches on these examples. However, the computational demands of SVD limit the scalability of LSA. While this limitation does not prevent LSA building models based on corpora designed to approximate the lifetime reading of human subjects, for the pur-

Table 3
Nearest-indirect neighbors of terms in a 2000-dimensional semantic space derived from the TASA corpus using RI, a 2 + 2 sliding-window and no minimum term-frequency. Each column contains the nearest-indirect neighbors of a cue term, as well as the strength of association between these neighbors and the cue term as measured using the cosine metric.

Nebula	Jazz	Semantic	Picasso	Nicotine
0.47: prominences	0.52: julliard	0.47: flexibly	0.29: rona	0.55: spooks
0.35: subclustering	0.51: songwriters	0.44: generalizable	0.28: nouveau	0.42: staleness
0.36: crankarm	0.46: entrancing	0.42: sugg	0.27: azul	0.30: pipeful
0.33: anglerfish	0.44: plonk	0.41: protometabolism	0.26: consign	0.29: fluids
0.33: torr	0.40: gunk	0.41: effecitvely	0.25: philatelists	0.29: acrid
0.33: hurtle	0.36: instrumental	0.39: paralinguistic	0.25: munson	0.28: tissues
0.32: comets	0.36: fusiyama	0.37: peraltas	0.24: buchwald	0.28: katczinsky
0.32: stegosaur	0.36: kapellmeister	0.40: miniponds	0.22: dulwich	0.27: rigid
0.31: meteoroids	0.36: lutenists	0.35: foreshadow	0.20: lagar	0.26: hajji
0.31: monastary	0.35: frierly	0.30: diagnostician	0.20: ternura	0.26: greensboro

pose of literature-based discovery it is desirable to extend this mechanism of inference to much larger corpora. For this reason, we evaluate an iterative variant of RI which is customized to enhance its ability to draw indirect inference.

3.4. Reflective Random Indexing

3.4.1. Motivation

The intuition underlying our use of an iterative variant is that term vectors generated as a linear sum of near-orthogonal document vectors are unlikely to be optimal for the derivation of meaningful indirect inference. If truly orthogonal vectors were employed, a term could only accumulate vectors that are orthogonal to those accumulated by another term it does not co-occur with directly: the vectors representing these two terms would have a cosine value (or normalized scalar product) of zero. A similar result should occur in the reduced-dimensional space when RI is employed, as the Johsnon–Lindenstrauss Lemma predicts that if a cosine similarity between two vectors is zero in the initial space, it will be close to zero in the reduced-dimensional space with high probability. Those term similarities that arise from the basic `train_model` process in Section 3.1 are thus the result of explicit term co-occurrence. While the near-orthogonal nature of the index vectors used in RI does allow for some overlap, a more principled manner of representing documents containing similar terms is desirable. The idea of deriving term vectors from meaningful document vectors emerged from the observation that term vectors can be cyclically retrained in RI [35] and related models [36], as well as the observation that generating positional term vectors using pretrained term vectors increases the inferencing ability of a permutation-based variant of RI [37] (RI can be adapted to encode the relative position of terms using vector permutations [38]). We call this iterative, cyclical training process Reflective Random Indexing (RRI) as the system generates new inferences by considering what it has learned from a data set in a previous iteration.

3.4.2. Implementation

3.4.2.1. Retraining. RRI may be motivated by two observations about the `train_model` function of Section 3.1. Firstly, there is a structural symmetry between the term vectors T and the document vectors D, so these arguments can easily be interchanged. In practice, this means that training document vectors from term vectors is a natural counterpart to training term vectors from document vectors. Secondly, there is no stipulation that the document vectors D must be randomly allocated elemental vectors. The function can be used in just the same way using input vectors that were learned by a previous training phase, instead of basic random vectors.

To support the generation of meaningful indirect inferences while employing the smallest possible number of training cycles, we propose Term-based RRI (TRRI), a novel variant of RRI. In TRRI, rather than assigning elemental vectors to each document, we assign elemental vectors to terms in the corpus. For example, for the RI model evaluated in this section of the paper we have assigned an elemental vector to every term that does not contain any non-alphabet characters and does not appear on a list of frequently occurring stop words such as "if", "and" or "but" that do not carry semantic content. A vector representation for each document in the corpus is then constructed as the linear sum of all of the elemental vectors for the terms the document contains. This linear sum is frequency weighted such that the number of times a term occurs in a document affects the number of times the elemental vector for this term is added to the vector representation for the document. This weighting is discussed in greater detail in the following section. Document-based RRI (DRRI) requires an additional half-step of iteration: elemental vectors are assigned to every document in the corpus, and term vectors are generated from these using the original implementation of RI. In order to better support indirect inferencing, this process is repeated: the term vectors produced by the first iteration are used to support the generation of meaningful document vectors, which in turn are used to generate a new set of term vectors. As documents containing a similar distribution of terms will have similar vectors, two terms that do not co-occur directly should still be able to acquire similar vector representations when this method is used.

3.4.2.2. Statistical weighting. In both cases when constructing document vectors, rather than using raw term-frequency, the vector for each term is weighted using the log-entropy weighting scheme [39], which is used to build document vectors in LSA. Log entropy weighting takes as local weighting log(1 + local term-frequency) and global weighting (1 + term entropy). The entropy of a term i over all j documents is the $\sum \frac{P_{ij}\log_2 P_{ij}}{\log_2 n}$ where $P_{ij} = \frac{tf_{ij}}{gf_i}$, tf_{ij} is the local frequency of term i in document j, gf_i is the global frequency of term i and n is the number of documents in the corpus. This weighting scheme has a number of desirable effects on the overall document representation. The impact of frequently occurring terms on a document is tempered by using a logarithmic function of the term-frequency. In addition, terms that occur focally in the corpus have a greater impact on account of the entropy function. Consequently, a term such as "platypus" which occurs in a few documents in the TASA corpus only, will have more impact on the representation of the documents in which it occurs than the term "egg" which is far more widely dispersed across the corpus.

3.4.2.3. Computational observations. Both forms of RRI maintain RI's desirable property of scalability. They also do not preclude the possibility of incremental updates, an attractive feature of RI. While this is possible with log-entropy weighting and DRRI, these would both require regenerating many document vectors on each iterative

update. A simpler approach would be to use TRRI and relinquish log-entropy weighting if it does not produce significant performance improvements for the task under consideration. Without log-entropy weighting, the procedure for incremental updates in TRRI is fairly straightforward. When a new document is added to the corpus, the elemental vectors for each term in the document are added together to generate a new document vector (any term that has not yet been encountered is simply assigned a new elemental vector). This document vector is then added to the stored semantic term vector for each term that occurs in the document.

3.4.2.4. Methodological variants. Pseudocode and a schematic representation for the RI variants evaluated in this paper are provided in Table 4 and Fig. 1. The formulations given here reuse the basic algorithms shown in Section 3.1, again assuming a document collection D of size p, a term set T of size q, and a term-document matrix M of size (p, q).

Fig. 1 illustrates the various document-based indexing approaches discussed in this paper. In RI as originally implemented, elemental vectors are assigned to each document to generate a set of elemental document vectors, D0. Term vectors are then constructed as the frequency-weighted vector sum of the elemental vectors for every document they occur in, to generate a set of term vectors, T1. In TRRI, elemental vectors are assigned to terms in the corpus, to generate a set of elemental term vectors, at position T0. Document vectors are constructed as the linear sum of the log-entropy weighted vectors for the terms they contain, to generate a set of document vectors, D1. Term vectors are then generated from these document vectors, to generate a set of semantic term vectors, T2. In DRRI, which adds an additional training cycle and log-entropy weighting to the RI model, elemental document vectors D0 are used to train term vectors T1, which in turn are used to train document vectors D2, and finally a set of term vectors T3 are generated.

3.4.2.5. Reflective Random Indexing associations. Table 5 shows the TRRI results from the TASA corpus with the set of test terms shown to produce interesting indirect near-neighbors when LSA is employed. TRRI identifies many more meaningful indirect associations on this set of terms than were identified using the original implementation of RI. DRRI produced comparable results, which are not presented.

There are many interesting and meaningful indirect associations in this table, including all of the 10 nearest-indirect neighbors of the terms "nebula", "picasso" and "semantic". While these neighbors are not necessarily of the same semantic class as the cue term, there are instances in which this is the case such as "picasso-pollock" and "nicotine-downers". However, many of the intuitively interpretable indirect associations are more general in nature. For example, the nearest-indirect neighbor of "nebula" is "astronomer", and "nicotine" is associated with several terms related to cancer ("sarcoma", "myeloma", "tumerous" and "sarcomatous").

In the remainder of this paper, we evaluate the ability of RI and RRI to make meaningful indirect inferences in the context of the literature-based discovery paradigm, by replicating two seminal historical literature-based discoveries, and comparing the ability of these methods to predict future direct co-occurrences in the MEDLINE corpus.

4. Evaluation

In the sections that follow, we describe two experiments to test the ability of variants of RI to draw meaningful indirect inference in the context of literature-based discovery. The first experiment (described in Section 4.1) investigates the ability of these models to replicate historical literature-based discoveries. While this is a common evaluation paradigm, the results may not generalize to other discoveries. Consequently in the second experiment (described in Section 4.2) we conduct a more extensive evaluation of the ability of each RI variant to predict terms that co-occur in the future with terms that are randomly selected from a time-delimited set of MEDLINE citations. As terms that are meaningfully related but do not co-occur at one point in time are likely to co-occur directly later, this evaluation can be considered as a measure of the extent to which the indirect inference derived by each model is meaningful.

4.1. Experiment I: simulating Swanson's discoveries

A common test for automated knowledge-discovery systems is to evaluate their ability to replicate historical literature-based discoveries [14], most commonly attempting to replicate Swanson's discovery of connections between Raynaud's Disease and fish oil, and migraine and magnesium. Bruza and his colleagues replicate aspects of both of these discoveries using a semantic space derived from the same corpus of titles of articles published in core clinical journals between 1980 and 1985 used by Swanson, by first finding near-neighbors of the "C" terms in semantic space, and then combining the vector for these "C" terms with those for a set of likely "B" (or linking) terms, chosen from among these neighbors [7,18].

Table 4
Pseudocode for evaluated algorithms.

Variant	Pseudocode	Description
Random Indexing (RI)	`D_0 = allocate_elemental_vectors(p, n, k)` `T_1 = train_model(M, D0)`	Assign elemental vectors to each document. For each term generate a semantic vector by adding the elemental vector for each document it occurs in

Document-based Reflective RI (DRRI)	`D_0 = allocate_elemental_vectors(p, n, k)`
	`T_1 = train_model(M, D_0)`
	`D_2 = train_model(M, T_1)`
	`T_3 = train_model(M, D_2)`
		Generate new document vectors using the semantic term vectors produced by RI, above. For each term, generate a new semantic vector by adding together the document vector of each document it occurs in

Term–term RI (TTRI)	`T_0 = allocate_elemental_vectors(q, n, k)` `T_1 = train_sliding_window_model(M, T_0)`	Assign elemental vector to each term. For each term, generate a semantic vector by adding the elemental vectors for each term it co-occurs with in a sliding-window moved through the text

Term-based Reflective RI (TRRI)	`T_0 = allocate_elemental_vectors(q, n, k)`
	`D_1 = train_model(M, T_0)`
	`T_2 = train_model(M, D_1)`
		Assign elemental vector to each term. For each document, generate a document vector by adding together the elemental vectors for each term it contains. For each term, generate a semantic vectorby adding the document vector of each document it occurs in

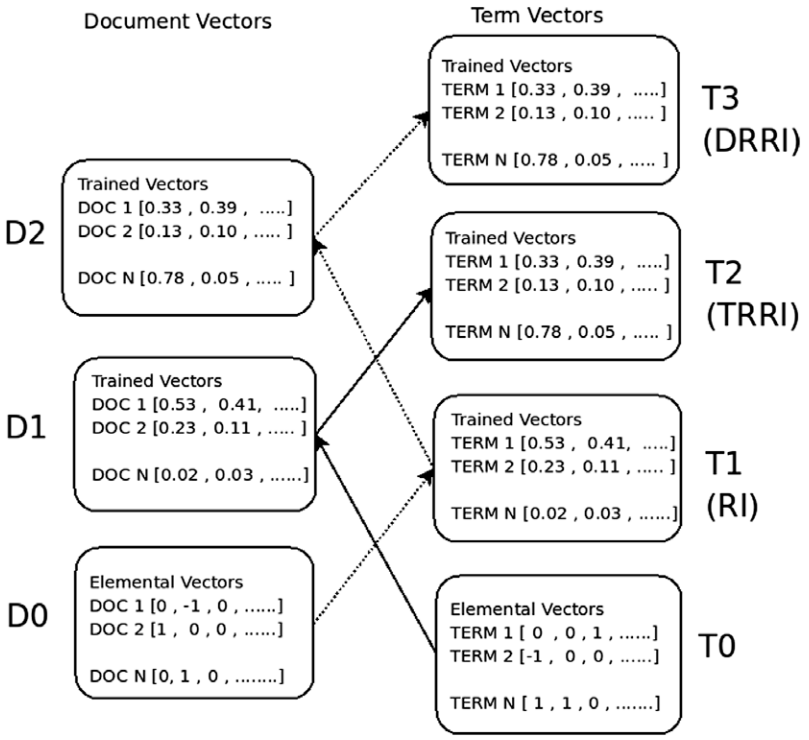

Fig. 1. A schematic representation of the generation of RI variants.

Table 5
Nearest-indirect neighbors of terms in a 2000-dimensional semantic space derived from the TASA corpus using a Term-based RRI (TRRI) approach. Each column contains the nearest-indirect neighbors of a cue term, as well as the strength of association between these neighbors and the cue term as measured using the cosine metric.

Nebula	Jazz	Semantic	Picasso	Nicotine
0.36: astronomer	0.25: performers	0.36: phonics	0.22: impressionism	0.32: casefinding
0.35: revolve	0.23: easier	0.33: generalizable	0.20: brushstrokes	0.31: doses
0.34: comets	0.22: contrast	0.33: preoccupy	0.20: architecture	0.29: depressants
0.33: solary	0.22: lyrics	0.33: nondirect	0.20: expressionism	0.29: comparatively
0.32: centauri	0.21: harmony	0.30: correspondences	0.20: pollock	0.29: sarcoma
0.32: supergiants	0.21: recently	0.30: phrases	0.20: craftsmanship	0.29: myeloma
0.31: globules	0.21: invented	0.29: phonological	0.19: sculptors	0.29: tumorous
0.30: gravitation	0.21: including	0.29: morpheme	0.19: surrealists	0.29: sarcomatous
0.30: neutrinos	0.21: performances	0.28: convey	0.19: beauty	0.29: nonepithelial
0.30: moons	0.21: popularity	0.28: writers	0.18: scenes	0.29: downers

Similarly, with all variants of RI it is possible to incorporate a linking term using vector addition at minimal additional computational expense. We find that while it is possible to reproduce these results using a similar approach with certain variants of RI, it is also possible to reproduce aspects of these discoveries directly, without the need for the explicit identification of a linking term.

In the work of Bruza et al., as in the work presented below, the choice of pertinent "B" terms is guided by hindsight as to which had proved useful in Swanson's research. As the derivation of "B" terms does not require indirect inference, we do not address this aspect of the knowledge discovery problem in this work. However, where B-terms are used, the B-terms selected would be intuitive choices for a domain expert, especially if promoted to near the top of a list of suggestions.

4.1.1. Methods

In order to simulate Swanson's discoveries of the migraine-magnesium and raynaud-eicosapentaenoic acid connections, we generate the following semantic spaces using variants of RI:

- RI: Random Indexing, as originally implemented.
- TTIDF: a (2 + 2) sliding-window based semantic space, weighted with the term-frequency/inverse-document frequency (TF-IDF) weighting scheme to limit the influence of frequently occurring terms.
- TRRI: Term-based Reflective Random Indexing.
- DRRI: Document-based Reflective Random Indexing.

The application of (TF-IDF) weighting to the sliding-window space was selected on the basis of its improvements in performance over an unweighted sliding-window approach in a preliminary experiment. All spaces are 2000-dimensional, and exclude terms occurring less than 10 times in the corpus as well as those terms on the stop list distributed with Swanson and Smalheiser's Arrowsmith [10] system, which has been customized for the purpose of knowledge discovery. In addition, any terms containing non-alphabet characters are excluded. Two versions of each of these spaces are created. The first is constructed from a corpus consisting of the titles of all MEDLINE articles (n = 190,129) published in core clinical journals between 1980 and 1985. These were the constraints employed by Swanson in the selection of titles for his seminal research, and they have also been employed

by many researchers attempting to simulate these discoveries using automated methods. With the additional constraints on term-frequency, non-alphabet characters and the stoplist, 6555 unique terms were extracted from this corpus. In order to assess the extent to which the various methods employed scale up to larger corpora, a second corpus was constructed, consisting of the abstracts of all articles in the MEDLINE database published between 1980 and 1985 (n = 844,289). 74,583 unique terms were extracted from this corpus after the application of the term-frequency and non-alphabet character related constraints and the Arrowsmith stoplist.

As these are considerably more terms than the number of unique terms (6555) in the corpus of titles, the relevant near-neighbors of the discovery-related terms are not ranked as high in the list of terms as when titles only are used. The term-document model used in RI suggests one solution to this problem: as abstracts in MEDLINE are indexed using Medical Subject Heading (MeSH) terms, it is possible to derive a distributional model of the relation between MeSH terms by altering the final step of either the TRRI or the DRRI process: rather than build term vectors from the document vectors, we build MeSH term vectors by normalizing the linear sum of the document vector for each document indexed by a particular MeSH term (Table 6). While not necessary for these experiments, this process has the additional useful property of deriving a distributional association from term to MeSH term. We will refer to these approaches to generating semantic similarity between MeSH terms as the TRRI_MeSH and DRRI_MeSH approaches. One issue we notice with the use of MeSH terms in this way, is that frequently occurring MeSH terms such as "Humans" and "Animals" tend to dominate the results of the "Raynaud_Disease" searches. As these frequently occurring terms have similar distributions, it is possible to exclude them using vector negation [40] (subtraction of the component of one vector that is parallel to another) of the exemplar high-frequency term "Humans". We use this strategy for the Raynaud's MeSH simulations only.

Within each space, the 2000 nearest-indirect (non-co-occurring) neighbors of the term 'raynaud' and the term 'magnesium' (or equivalent MeSH terms) are sought. In addition, the combined vectors for the terms 'raynaud + platelet' and 'migraine + calcium' (or equivalent MeSH terms) are generated using vector addition, to assess the extent to which vector combination with a linking term improves performance, and the 2000 nearest neighbors of these combined vectors that do not co-occur directly with the relevant cue term (for example "migraine" or "raynaud") are retrieved. We then evaluate the rank of the term "eicosapentaenoic" (eicosapentaenoic acid is the active ingredient of fish oil) when "raynaud" is used as a cue term, and "magnesium" when "migraine" is used as a cue term (or equivalent MeSH terms in the RRI-MeSH variants).

The random initiation of vectors in RI presents the possibility that the results obtained may not be repeatable, as the generation of another space with the same parameters but different Random Indexes may produce different results particularly in evaluations such as this in which the vector representation of a small number of individual terms is important. In order to confirm our results are reproducible, we repeat one hundred runs of all simulations on the smaller corpus, and 50 runs of all simulations on the larger corpus.

4.1.2. Results and discussion

4.1.2.1. Smaller corpus. The results of experiments on the smaller corpus (Table 7) confirm that Swanson's discoveries can be simulated using variants of RI. This table shows the ranking of the target terms "eicosapentaenoic" and "magnesium" among the nearest-indirect neighbors of the vector representations for individual or combined cue terms related to Raynaud's Disease and migraine over 100 repeated simulations. For each simulated discovery, the first row shows the number of runs (out of 100 possible runs) in which the target term obtained a rank less than 2000. The subsequent rows give the mean, minimum, maximum and standard deviation of the ranking for the target term across all 100 runs. For the purposes of simulated discovery it would be preferable for these target terms to be ranked among the top twenty or so nearest-indirect neighbors, as a user of a literature-based discovery system might reasonably be expected to explore this number of possibilities.

With this corpus, TTIDF and TRRI produce consistently better rankings on the Raynaud's and Migraine discoveries respectively. Of particular interest, TTIDF consistently produces a top 5 ranking for "eicosapentaenoic" among the nearest-indirect neighbors of "raynaud" over 100 simulations, without the need to explicitly identify a linking term.

TRRI is not effective in reproducing the "Raynaud" discovery, and despite a best ranking of 52 when searching for the term "raynaud" alone, repeated runs of this experiment do not recover the term "eicosapentaenoic" within the top 2000 ranks in 65 of 100 runs. One reason for this may be that 'documents' in this corpus consist of MEDLINE titles only, which are substantially shorter than the usual unit of analysis in most term-document based models. The performance of DRRI is generally comparable to that of TRRI in these experiments, although the term "eicosapentaenoic" tends to be ranked higher by DRRI in the Raynaud's experiments. Interestingly, however, on isolated runs DRRI produces top 10 rankings for both discoveries without the use of a linking term. One possible explanation for this is that as there are many more documents than terms in this corpus, a greater number of elemental vectors are introduced with the DRRI approach. This would lead to a higher probability of co-incidental overlap between ostensibly near-orthogonal index vectors, and irreproducible yet serendipitous results of this nature.

4.1.2.2. Larger corpus. The results of the simulation using the larger corpus are shown in Table 8. This table shows the ranking of the target terms "eicosapentaenoic" and "magnesium" or target MeSH terms "Eicosapentaenoic Acid" and "Magnesium" among the nearest-indirect neighbors of the vector representations for individual or combined cue terms related to Raynaud's Disease and migraine over 50 repeated simulations. For each simulated discovery, the first row shows the number of runs (out of 50 possible runs) in which the target term obtained a rank less than 2000. Where such runs occurred, the subsequent rows give the mean, minimum, maximum and standard deviation of the ranking for the target term across these runs.

Table 6
Pseudocode for the generation of the TRRI-MeSH space.

Variant	Pseudocode	Description

Reflective RI-MeSH (RRI-MeSH)	`T0 = allocate_elemental_vectors (T, n, k)`
	`D1 = train_model(M, T0)`
	`MeSH = train_model(MD,`[a] `D1)`
		Assign elemental vectors to each term. For each document, generate a document vector by adding together the elemental vectors for each term it contains. For each MeSH term, generate a semantic vector by adding together the document vector of each document it occurs in

[a] *MD* is the term-document matrix for MeSH terms.

Table 7
Results of 100 runs simulating each discovery in TTIDF, TRRI and DRRI spaces derived from the corpus of MEDLINE titles. The best overall results for each discovery are highlighted. TTIDF, inverse-document frequency weighted term–term based RI; TRRI, Term-based Reflective Random Indexing; DRRI, Document-based Reflective Random Indexing.

Target: "eicosapentaenoic"	TTIDF		TRRI		DRRI	
Cue: "raynaud"	Alone	+Platelets	Alone	+Platelets	Alone	+Platelets
N rank <2000	100	100	35	53	47	79
Mean	3.52	4.82	1055.09	1123.53	862.74	693.66
Min	3	1	52	60	6	50
Max	5	11	1978	1992	1926	1858
SD	0.52	1.99	602.04	588.4	579.56	527.05
Target "magnesium"	TTIDF		TRRI		DRRI	
Cue: "migraine"	Alone	+Calcium	Alone	+Calcium	Alone	+Calcium
N rank <2000	31	100	50	100	48	100
Mean	1256.32	21.86	1136.14	13.88	958.81	14.53
Min	131	11	11	7	5	11
Max	1943	50	1954	36	2000	21
SD	552.68	8.19	604.62	4.65	555.77	2.31

Table 8
Results of 50 runs simulating each discovery in TTIDF, TRRI and DRRI spaces derived from a corpus of MEDLINE abstracts dated between 1980 and 1985. The best overall results for each "discovery" are highlighted. TTIDF, inverse-document frequency weighted term–term based RI; TRRI, Term-based Reflective Random Indexing; DRRI, Document-based Reflective Random Indexing.

	TTIDF		TRRI		TRRI-MeSH[a]		DRRI		DRRI-MeSH[a]	
Target: eicosapentaenoic										
Cue: raynaud	Alone	+ Platelets	Alone	+ Platelets	Alone	+ Platelets	Alone	+ Platelets	Alone	+ Platelets
N rank <2000	0	0	2	50	15	50	1	48	0	41
Mean	—	—	1373.5	162.62	1413.87	59.04	1955	796.48	—	325.02
Min	—	—	925	89	516	44	1955	154	—	102
Max	—	—	1822	333	1913	100	1955	1859	—	534
SD	—	—	634.27	51.62	422.18	11.57	0	373.6	—	92.71
Target: magnesium										
Cue: calcium	Alone	+ Calcium	Alone	+ Calcium	Alone	+ Calcium	Alone	+ Calcium	Alone	+ Calcium
N rank <2000	0	0	50	50	50	50	0	0	1	41
Mean	—	—	252.18	1136.64	913.72	8.48	—	—	1843	217.41
Min	—	—	110	11	375	3	—	—	1843	161
Max	—	—	646	1954	1288	22	—	—	1843	266
SD	—	—	143.69	604.62	228.3	3.53	—	—	0	24.91

[a] With MeSH-based approaches, the terms used are "Migraine Disorders", "Calcium", "Raynaud Disease", "Platelet Aggregation" and "Humans".

We are not able to reproduce either discovery in the larger corpus using the TTIDF approach on any of the 50 runs (with or without "B" terms), suggesting that this approach does not scale well to larger corpora. TRRI generally produces a better ranking for the "raynaud-eicosapentaenoic" discovery when the "B" term platelets is employed, with an improvement in the mean ranking of "eicosapentaenoic" from more than 1000 (with the smaller corpus) to 162.62. Of note, this is a considerably higher ranking than it was possible to obtain consistently with the much smaller corpus of titles. The best-performing model on this corpus is the TRRI_MeSH model, suggesting the utility of this approach as a way to improve specificity. TRRI_MeSH consistently produces top 100 rankings for "Eicosapentaenoic Acid" and top 25 rankings for "Magnesium" when appropriate "B" terms are used. However, neither discovery is consistently reproducible in either the DRRI or DRRI-MeSH spaces.

These results suggest that TRRI in particular scales well to larger corpora, provided measures are taken to compensate for the increase in the number of distinct terms. The method we have employed to link terms to MeSH terms could also be adapted to derive distributional similarities between elements of other terminologies that have been used to index MEDLINE, such as UMLS concepts and Entrez gene ID's.

4.2. Experiment II: anticipating future connections

While the replication of historical discoveries provides an interesting demonstration of the potential usefulness of the variants of RI as tools to support knowledge discovery, evaluations of this nature do not evaluate the extent to which these approaches generalize. In the following section, we evaluate the ability of each of these methods to make meaningful indirect inferences from the MEDLINE corpus on a larger scale. As the domain knowledge required to interpret term associations in MEDLINE precludes qualitative evaluation at a large scale, we make the assumption that indirect relations in a time-delimited subset of MEDLINE documents that occur directly in future MEDLINE abstracts after this time can be considered as meaningful indirect connections, and consider the proportion of nearest-indirect neighbors of a set of random terms that co-occur in the future as a measure of the ability of each method evaluated to derive such connections.

4.2.1. Methods

Indirect connections in the biomedical literature that predict discoveries will become direct connections once the connection between the terms concerned becomes discovered public knowledge. In order to evaluate the ability of these different sorts of RI to derive meaningful indirect connections, we proceed to evaluate their ability to predict future connections in the MEDLINE corpus of abstracts. For this evaluation we use three different databases:

(a) **restrictedMEDLINE**: the term-document statistics derived from a corpus of MEDLINE abstracts and titles added to the index between the start of 1980 and the end of 1985 (n = 1,600,093),

(b) **futureMEDLINE**: a tabulation of the co-occurrence of terms occurring in abstracts added to MEDLINE after 1985, and
(c) **pastMEDLINE**: a tabulation of the co-occurrence of terms occurring in abstracts added to MEDLINE before 1980. This was used to eliminate indirect connections that had occurred in MEDLINE before 1980.

Six approaches to generating semantic spaces from MEDLINE using RI are evaluated for their ability to derive meaningful indirect connections from MEDLINE abstracts in restrictedMEDLINE. They include:

- **RI**: RI as originally implemented [24].
- **TTRI**: RI using a narrow (2+2) sliding-window [25].
- **TTIDF**: as above, but the influence of other terms in the sliding-window on the focus term is weighted using inverse-document frequency.
- **DRRI**: Reflective RI starting with random document vectors.
- **TRRI**: Reflective RI starting with random term vectors.
- **TRRI2**: second iteration of TRRI. Term vectors produced by TRRI are used instead of elemental vectors as a basis for training of the model.

All spaces are 2000-dimensional, and as in the simulated discoveries terms were excluded if they occur on the Arrowsmith stoplist, contain non-alphabet characters or occur in the entire corpus less than 10 times. Two thousand (2000) cue terms are randomly selected from the **restrictedMEDLINE** index, and, for each index, the 50 nearest-indirect neighbors (NINs) of each of the targets are found in the **restrictedMEDLINE** database using each of the six indexing procedures. Then each of the indirect neighbors is checked to determine whether it co-occurs with its target after 1986. The ones that do co-occur are dubbed "future connections."

4.2.2. Results and discussion

4.2.2.1. Precision according to rank. All indexes are evaluated according to the proportion of nearest-indirect neighbors that occur directly in **futureMEDLINE**, which is equivalent to precision if direct co-occurrence in **futureMEDLINE** is considered as a gold standard. For example, the proportion of the 10 nearest-indirect neighbors that co-occur with the cue term in **futureMEDLINE** is equivalent to precision at $k = 10$. This proportion is evaluated for different rank strata (top 10, 11–20 and so forth) of nearest neighbors, in order to determine the extent to which the prediction of future co-occurrence is more accurate for neighbors of higher rank. The results are shown in Fig. 2 which plots the proportion of past NINs that co-occur in the future as a function of the rank of the cosine similarity for the past NINs and the index used to identify the past NINs. The *y* axis gives the mean proportion of NIN's in restrictedMEDLINE that co-occur directly in futureMEDLINE (but not in pastMEDLINE) with a cue term across all 2000 randomly selected cue terms. The *x* axis denotes the ranking of these NIN's, such that the first column gives the proportion of the top 10 ranked NINs that co-occur directly in the future, the second column gives the proportion of the NIN's ranked 11–20 and so forth.

Neither the RI index nor the TTIDF index produces many future connections, a maximum of 6% for the TTIDF index for the 10 nearest-indirect neighbors. The other indexes all show greater sensitivity to the cosine rank with percentage of future connections falling off across rank. The order of the indexes listed in the figure corresponds to the order of their precision, with precision at $k = 10$ of 0.4, 0.36 and 0.29 for TRRI2, DRRI and TRRI respectively. On account of the variability in the data (error bars in the figure show one standard deviation above and below the means) and the large sample sizes (2000 target items), most of the differences seen in the figure are statistically significant. TRRI shows the most pronounced decrease in future connections with decreasing cosine similarity (increasing rank). One indication of the meaningfulness of an index is the rate of decrease with decreasing similarity. Greater rates of decrease indicate greater sensitivity to the appropriate semantic characteristics.

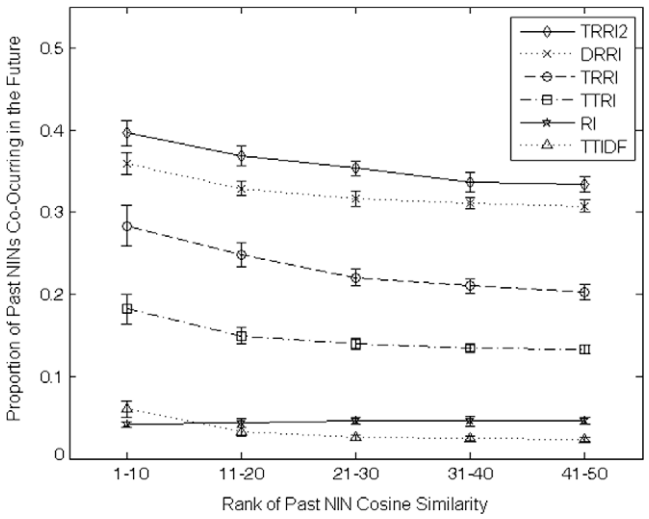

Fig. 2. Proportion of past indirect neighbors that co-occur in the future as a function of similarity rank for each index. RI, Random Indexing; DRRI, document-based RI; TRRI, term-based RI; TTRI, term–term based RI; TTIDF, inverse-document frequency weighted TTRI; TRRI2, a second iteration of TRRI (see Section 4.2.2.3). The rank "1–10" column gives precision at $k = 10$.

4.2.2.2. Term-frequency. The differences in precision may not necessarily indicate greater utility for the purpose of knowledge discovery, as it is possible that certain indexes favor terms that occur more frequently in futureMEDLINE. In order to determine whether there are differences in the models in terms of the influence of term-frequency, we examine the term-frequency for all of the future connections found. The results are shown in Table 9 which describes the term frequencies of each of the correctly predicted terms for each model. The first row gives the number of future connections, and the remaining rows give the summary statistics for the global term-frequency of each of these "discoveries".

The TRRI index tends to produce future connections of lower term-frequency[2] than the other productive models. There is substantial skew in the distributions, as illustrated in Fig. 3, which shows the distribution of the log of term-frequency for each of the indexes. This figure gives a global picture of the distribution of term-frequency across all future connections. The *y* axis gives the number of future connections that fall into the range of global term-frequency denoted by the *x* axis. It is possible that rather than selecting for meaningful indirect neighbors, some indexes are simply selecting for neighbors that occur more frequently in **fullMEDLINE**. Presumably connections with higher frequency terms are potentially less interesting than connections with terms occurring less frequently. However, the differences in term-frequency do not systematically follow the productivity of the indexes, as is illustrated by the fact that the distribution of global term frequencies for TRRI's future connections peaks earlier than in the distribution produced by other models.

[2] The picture is somewhat complicated by a particularly high maximum value for TRRI. However, it has the lowest median, and the distributions show that it produces future connections with generally lower frequency as shown in the distributions depicted in Fig. 3.

Table 9
Summary statistics on the term-frequency of future connections. RI, Random Indexing; DRRI, document-based RI; TRRI, term-based RI; TRRI2, second iteration of TRRI; TTRI, term–term based RI; TTIDF, inverse-document frequency weighted TTRI.

	Index					
	TRRI2	DRRI	TRRI	TTRI	RI	TTIDF
n	45,982	41,866	31,275	18,306	7025	5251
Median	3287	3622	1632	2433	1700	2129
Mean	10,793	11, 613	10,696	11,863	11,356	8570
SD	34,842	34,435	57,820	39,961	52,090	29,071
Minimum	13	13	13	13	10	13
Maximum	1,462,238	1,462,238	4,619,913	2,220,191	2,220,191	1,057,063

4.2.2.3. Cyclical retraining and statistical weighting. The DRRI model amounts to a single round of cyclical retraining of a RI space with the addition of log-entropy weighting when document vectors are generated. The TRRI model is also amenable to cyclical retraining. In order to investigate the effect of repeated rounds of retraining on each of these models, we generate 500-dimensional DRRI and TRRI spaces from a corpus consisting of all MEDLINE abstracts with publication dates between 1980 and 1985, and record how many of the 10 nearest-indirect neighbors of a set of 500 randomly selected terms in these spaces co-occur directly in an index of all the abstracts in MEDLINE. In early iterations, the proportion of predictions in the 10 nearest-indirect neighbors is somewhat higher than in the previous experiments, which used a larger corpus that included titles, a different set of randomly selected terms, and unlike these experiments excluded any direct connections that first occurred before 1980.

Fig. 4 shows the results of these experiments. The *y* axis gives the mean proportion of the 10 top-ranked NINs that co-occur directly with a cue term in futureMEDLINE, across all 500 randomly selected cue terms. The *x* axis gives the number of iterations of cyclical training, and the solid and dashed lines represent the results of starting with elemental term and elemental document vectors respectively. In this experiment, the second iteration (TRRI2) of TRRI predicts a greater number of direct co-occurrences than any iteration of DRRI. This result is consistent with the results on the larger corpus shown in Fig. 2: given a second iteration, TRRI produces more future direct co-occurrence than DRRI, and this result is statistically significant. However, this extra iteration also results in a shift in the term-frequency distribution such that more high-frequency terms are retrieved, much as they are by DRRI. As predicted, DRRI also produces far more predicted co-occurrences after the first round of training, which is equivalent to the original implementation of RI. There is a slight increase in the number of indirect neighbors in the next cycle, but the number of indirect neighbors produced by both models drops after this point, suggesting that while a single additional iteration improves the ability of both models to predict direct co-occurrence, cyclical retraining beyond these first two iterations is detrimental to the ability of these models to select for meaningful indirect inferences. This deterioration in performance can be explained on account of the similarity between RRI and iterative approaches to the problem of finding the principle eigenvector [41]. With repeated iterations of RRI, all vectors converge on a single vector, and the ability to discriminate between them is lost.

While these results are not featured in the tables or figures, we note that constructing document vectors without log-entropy weighting leads to a drop in performance with TRRI—the proportion of predicted future co-occurrence in the 10 nearest-indirect neighbors drops by about 0.5. There is also a drop in performance with DRRI when log-entropy weighting is not used. However, in both cases performance without log-entropy weighting still exceeds that of established models, and this drop in performance may be an acceptable trade-off for easy incremental updates in some applications.

The statistical analysis of the future connections provides some useful information about the relative productivity of the various indexing methods. All four of TRRI2, DRRI, TRRI, and TTRI are productive and they all show some sensitivity to the cosine similarity

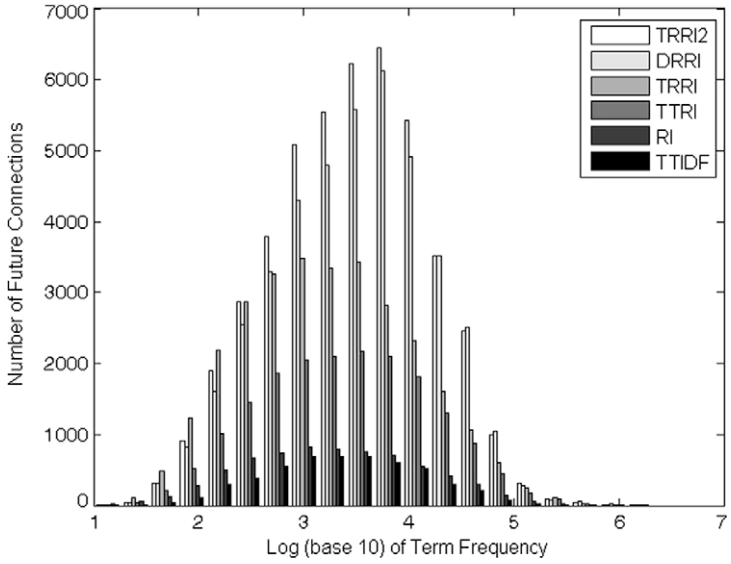

Fig. 3. The distribution of the log of future connection term-frequency for each index. RI, Random Indexing; DRRI, document-based RI; TRRI, term-based RI; TTRI, term–term based RI; TTIDF, inverse-document frequency weighted TTRI.

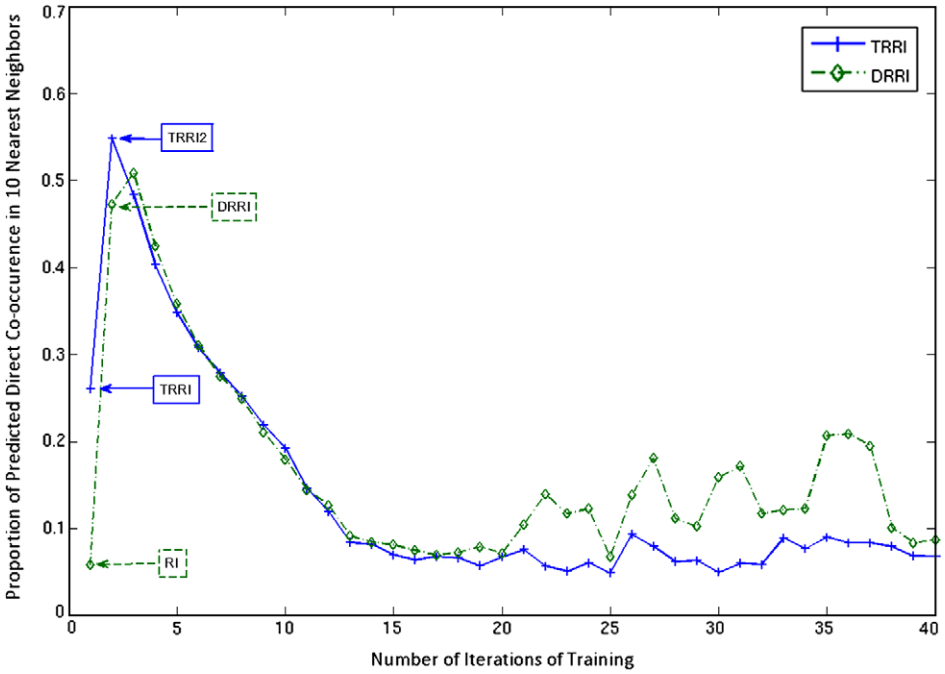

Fig. 4. Proportion of predicted direct occurrences in 10 nearest-indirect neighbors (precision at $k = 10$) with cyclical retraining of TRRI and DRRI on a corpus of MEDLINE abstracts. RI, Random Indexing; TRRI, Term-based Reflective Random Indexing; DRRI, Document-based Reflective Random Indexing; TRRI2, Second iteration of TRRI.

between indirect neighbors and the target terms. TRRI shows the greatest sensitivity to cosine similarity and lower term-frequency of future connections. By these criteria, TRRI may be preferred, however, it is another matter to determine the semantic value of future connections from all four indexes.

4.2.2.4. Qualitative evaluation of selected results. Most of the indirect neighbors retrieved during this study require a substantial amount of biomedical knowledge to interpret. It is unlikely that a human rater with sufficient expertise in the content domains of these 2000 terms to accurately annotate these results exists, and even if we were to assemble a team with such expertise the time required to annotate 45,000 'discoveries' would be prohibitive. In the interest of illustrating the sorts of associations that this approach has captured, we present an exhaustive analysis of the 'discoveries' generated by two terms related to snake venom, 'cobratoxin' (Table 10) and 'convulxin" (Table 11). These terms represent exemplars of the two broad categories of snake venom: Cobratoxin is a neurotoxin that causes paralysis by blocking the binding of acetylcholine to the nicotinic acetylcholine receptor

Table 10
Future connections (after 1985) predicted for term 'cobratoxin'. TRRI, Term-based Reflective RI; DRRI, Document-based Reflective RI; TTRI, Sliding-window RI.

'Discoveries' for term 'cobratoxin'		
Term	Space	Significance
butx	TRRI, TTRI	butx = 'bungarus toxin'.
elapidae	TRRI	Family of snakes that includes the cobra.
conotoxin	TRRI	Neurotoxin produced by the cone snail
dendrotoxin	TRRI, TTRI, DRRI	Neurotoxin produced by the "dendroaspis" (mamba) family of snakes
multicinctus	TRRI, DRRI,	Bungarus multicinctus = muti-banded krait
annulifera	TRRI, TTRI, DRRI	Naja annulifera = snouted cobra
acchr	TRRI, TTRI	The acetylcholine receptor
erythroidine	TRRI	A plant-derived neurotoxin with a similar action to cobratoxin
polylepsis	TRRI	Dendroaspis polyepsis = black mamba
haje	TRRI, TTRI, DRRI	Naje Haje = banded egyptian cobra
nachr	TRRI, TTRI	Nicotinic acetylcholinesterase receptor
bungarus	TRRI, DRRI	A genus of neurotoxic snakes, commonly referred to as kraits.
cholera	TRRI	Vibrio cholera has a toxin, although the action is different to cobratoxin
dendroaspis	TRRI, DRRI	Dendroaspis polyepsis = black mamba
scorpion	TRRI	Scorpion venom is neurotoxic
lsiii	TRRI	Laticauda semifasciata III, neurotoxin of sea snake Laticauda semifasciata, binds to *n*-AChr to induce paralysis
mamba	TRRI	Another neurotoxic snake
Btx,bgtx	TTRI	Bungarus toxin
machr	TTRI	The muscarinic acetylcholine receptor, which is not involved in the action of cobratoxin.
bonds	TTRI, DRRI	Various chemical bonds exist in the structure of cobratoxin, but this term is non-specific.
acchor	TTRI	Synonym for AChr
suberyldicholine	TTRI	Blocks the AChr
electroplax	DRRI	Electric fish muscle fibers rich in AChr and used in experiments with cobratoxin
maleimido	DRRI	Part of the chemical structure of a ligand that binds to the AChr
buried, displace	DRRI	Non-specific terms, relevance unclear

Table 11
Future connections (after 1985) predicted for term 'convulxin'. TRRI, Term-based Reflective RI; DRRI, Document-based Reflective RI; TTRI, Sliding-window RI.

'Discoveries' for term 'convulxin'		
Term	Space	Significance
citrated	TRRI, DRRI	Citrate is an anticoagulant used in the laboratory
hirudin	TRRI, DRRI	Anticoagulant occurring in leeches, which has since been synthesized as the active ingredient of the antithrombotic drug bivalirudin
apyrase	TRRI, DRRI	Anticoagulant secreted by the female mosquito
gpib	TRRI, TTRI, DRRI	Glycoprotein ib, the binding site for convulxin. This appears to have been first discovered in 2003
Echis	TRRI	Genus of venomous vipers with hemotoxic venom
antiaggregatory	TRRI, DRRI	Counteracting the effect of convulxin
prp	TRRI, DRRI	PRP: platelet-rich plasma used in experiments
atrox	TRRI	Croatulus atrox is a hemotoxic rattlesnake
aggregability	TRRI, TTRI, DRRI	Increased by convulxin
iiia	TRRI	Another glycoprotein involved in platelet aggregation
ristocetin	TRRI, TTRI, DRRI	Antibiotic agent, discontinued as causes platelet aggregation as side-effect
intraplatelet	TRRI, DRRI	Between platelets
aggregometer	TRRI, DRRI	Measures adhesiveness of platelets
elapidae	TTRI	Family of neurotoxic snakes that includes the cobra
thrombospondin	TTRI	Causes platelet aggregation
formosan, viper	TTRI	Formosan pit vipers have venom with similar action
trimeresurus	TTRI	Genus of venomous pit vipers
cobra	TTRI	Another venomous snake, but neurotoxic
bothrops	TTRI	Family of hemotoxic vipers
habu	TTRI	Japanese name for pit vipers
gpiib	DRRI	Glycoprotein iib, involved in platelet aggregation
Gpv	DRRI	Glycoprotein V, also involved in platelet aggregation

(nAChR), preventing the depolarization of the muscle cell that precedes muscle contraction. In contrast, Convulxin, the toxin of the rattlesnake Crotalus durissus terrificus, is hemotoxic. It acts in the bloodstream, causing platelet aggregation and the formation of thrombi (blood clots). Of note, one of the associations generated from the restrictedMEDLINE corpus appears to predict the discovery in 2003 of a binding site for Convulxin [42].

These tables illustrate several different types of inference generated by this method. For example, both toxins generate associations to other types of venomous snakes. In the case of TRRI, these are generally restricted to snakes producing the same class of toxin. Both spaces produce a number of neighbors related to the mechanism of the toxin under investigation, as well as other substances with a similar mechanism. Perhaps the most interesting of these is Glycoprotein ib, the binding site for Convulxin. This binding site appears to have been first discovered in 2003 [42], and as such represents a simulated discovery. Of note, this association was produced by the TTRI, DRRI and the TRRI approaches. A few future connections such as "bonds", "buried" and "displace" are non-specific and uninformative. While "bonds" was also recovered by TTRI, all three were produced by DRRI, which is not unexpected given the tendency of this model to recover higher frequency terms. We have not included TRRI2 in this table, as the additional future connections produced by this index consist of high-frequency terms such as "quaternary" and "recalcified" that are similarly uninformative.

5. Discussion

5.1. Summary of results

This study evaluates the ability of several scalable models of semantic distance to derive meaningful indirect connections between terms. We find that term–term based RI and TRRI, a novel variant of RI, are able to consistently simulate aspects of historical literature-based discoveries. In particular, we note that it is possible to reliably duplicate Swanson's seminal raynaud-to-eicosapentaenoic-acid discovery using indirect similarity alone, without the need for a linking term, using statistically weighted term–term based RI. However, our ability to replicate historical discoveries with this method deteriorates with larger corpora than the relatively small corpus of MEDLINE titles used in Swanson's original work. In contrast, TRRI seems better able to replicate historical discoveries as corpus size increases, particularly when vector representations of MeSH terms are used to enhance precision. Upon evaluation of the spaces derived using RI and its variants, we find that both RRI variants outperform established methodologies in their ability to predict future direct connections.

5.2. Implications for distributional semantics

This improvement in performance is interesting to consider in the light of a recent study by Sitbon and Bruza, which shows no improvement in the TOEFL synonym test evaluation with cyclical retraining [43]. While the TOEFL test is commonly used as a means of evaluating semantic space models, it is focused exclusively on the evaluation of synonymy, and in many questions indirect association in the TASA corpus is not required to obtain a correct answer. Our evaluation of the ability of semantic space models to predict future direct association differs from the traditional synonym test evaluation in several respects. Firstly, it accommodates more general semantic relations than synonymy alone. Arguably, it selectively evaluates other types of semantic relations than synonymy, as synonyms are unlikely to occur directly in the same context. In addition, it focuses specifically on indirect inference, and does not evaluate associations between terms that occur together in the same context. While we have used historical reference points to segment our corpora, this evaluation could be performed on general language corpora by using a smaller corpus (such as the TASA corpus) to try to predict direct associations in a larger corpus (such as the British National Corpus). This may be a more appropriate evaluation than the synonym test in situations where semantic

spaces are being constructed to support applications such as knowledge discovery or information retrieval where indirect associations of a more general nature than synonymy are likely to be useful.

5.3. Implications for literature-based discovery

While time-delimited corpora have been used previously to evaluate literature-based discovery tools [44,45] most prior evaluations have focused on the ability of a specific system to replicate historical discoveries concerning a small set of disease entities. An exception is the evaluation method proposed recently by Yetisgen-Yildiz and Pratt [3], which evaluates the ability of a system to predict future direct co-occurrence based on a set of randomly selected MeSH disease entities. While we arrived at our method of evaluation independently, it is similar in concept to this work, which provides a well-defined methodology for the evaluation of literature-based discovery systems based on their ability to predict future direct co-occurrence. However, this evaluation was somewhat smaller in scale, perhaps on account of the scalability limitations involved in directly identifying linking terms. 100 cue terms, all of which were MeSH terms categorized as diseases in the UMLS, were considered. In addition UMLS categories were used to limit potential target terms to only those MeSH terms occurring in two UMLS categories. Linking terms were similarly limited to a group of five categories which subsumed the other two. In addition, the training set was larger, which is justified by the authors as a means of ensuring that adequate linking terms could be identified. Consequently, this evaluation is not strictly comparable to ours. Nonetheless, we note that the proportion of future co-occurrences predicted by these methods for the top 10 ranked predictions ranges between 0.19 and 0.24, as compared to 0.29 (TRRI), 0.36 (DRRI) and 0.40 (TRRI2) in our full scale evaluation. While it is not possible to draw any strict comparisons due to the differences in the evaluation procedures and the additional constraints imposed to reduce the size of the search space, the fact that the precision of RRI and its variants is higher than any published estimate for related methods on a similar task suggests that RRI would make a useful addition to the methods currently utilized for literature-based discovery.

Our evaluation focuses on the ability of different variants of RI to predict direct co-occurrence over large sets of terms, without the requirement that these predictions either relate to a historical discovery or concern defined entity types. Consequently, we do not suggest that RRI alone constitutes a usable knowledge-discovery system. The generation of indirect inferences simulates but one aspect of the process of abductive reasoning: the generation of novel connections. Some constraints on this process are necessary to prevent the user of a system (or for that matter a creative thinker) being overwhelmed by irrelevant or unworkable hypotheses. In our future work we will attempt to model these constraints in order to select those indirect connections that are not only meaningful, but also represent useful hypotheses. We anticipate this aspect of the discovery process involving a high level of user involvement. Nonetheless, the development of a scalable and effective means of deriving indirect inferences from the scientific literature represents an important step toward a computational model of abduction as it pertains to scientific discovery. We note also that this study has certain limitations. Indirect inferences were derived from titles and abstracts in MEDLINE. However, as confirming co-occurrence in the larger fullMEDLINE index consumes time and computational resources, only abstracts were used. Consequently we may have missed some potential 'discoveries' that occurred directly together in future titles only, and by the same token it is also possible that some of the 'discoveries' proposed were already present in MEDLINE titles before 1980. We do not believe that either of these points detract from the strength of our findings, as the differences measured between methods were statistically significant, and indirect connection between terms in a pre-existing title may not be a discovery, but can nonetheless be considered as meaningful.

RRI is able to derive meaningful indirect connections from larger corpora such as the MEDLINE corpus of abstracts, and as the growth in complexity of the algorithm underlying RI is linear to the size of the data being processed, it should scale comfortably to accommodate the increasing size of the MEDLINE database. The derivation of meaningful indirect inferences from this rapidly expanding corpus is likely to be of use as a tool to support the discovery of implicit connections in the literature, and has certain advantages over existing approaches to this problem. Firstly, it is often possible to identify meaningful implicit connections without the need for the specification of a bridging term beforehand. This is something of a departure from Swanson's "open" discovery paradigm, which proceeds stepwise from cue concept through linking "B" concept to discovery. The ability of RRI to derive meaningful implicit connections directly, without the need for the specification of a "B" concept beforehand suggests an alternative approach in which interesting indirect connections are suggested by the system, and "B" concepts are sought for the purpose of explanation rather than as a prerequisite to the identification of possible discoveries. In previous work, we have illustrated the ability of Pathfinder networks constructed from the near-neighbors of a combined vector for the two indirectly connected terms to support the generation of explanatory hypotheses [9], but rule-based approaches could also be applied to this aspect of the problem. From a cognitive perspective, the identification of an indirect connection in this manner is appealing, as it provides a computational implementation of a two-stage model of abduction: first an initial possibility is arrived at based on similarity, and then this connection is subsequently explained using some more cognitively demanding mechanism. We note that the ability to derive meaningful indirect inference from large corpora presents opportunities for further research using RRI to generate document vectors for information retrieval. In addition to scalability advantages, these vectors would be amenable to incremental updates, which is particularly desirable with the large and rapidly growing document collections that contemporary information retrieval systems must manage.

5.4. Further implications

In addition, we have derived a scalable method of mapping between terms and controlled terminologies that have been used to index MEDLINE. While our primary use for this method was to increase precision, the derivation of semantic distance between terms and controlled terminologies in this manner has other possible applications such as automated indexing, and also provides a basis for a combination between distributional and ontology-based methods. While the use of term-document statistics to map between natural language text and controlled terminologies is not without precedent [46], the approach we have developed offers significant scalability advantages over established methods.

6. Conclusion

RI approaches to creating semantic spaces for large databases have several advantages. They can be scaled to handle very large corpora, and they can be incrementally updated as new documents are added without requiring completely new analyses. The reflective random vector method evaluated in this paper promises to provide superior recovery of indirect or latent relations compared to the original version which is critical in the arena of literature-

based discovery. The evaluations presented in this manuscript show that this scalable and direct approach to generating indirect inference is able to predict the future co-occurrence of terms from a time-delimited segment of the MEDLINE corpus without the need to explicitly identify linking terms. The estimates of the predictive ability of RRI exceed those obtained using existing literature-based discovery methods in similar, albeit smaller scale evaluations in the literature. While RRI was presented in the context of scientific discovery, the results should generalize to any sufficiently sized corpus of natural language text. In addition, the derivation of meaningful indirect connections has applications beyond this domain, such as the retrieval of documents related to but not containing a particular search keyword. RRI retains RI's desirable properties of scalability and the potential for incremental updates, and we anticipate the further application of this method to problems within and beyond the biomedical domain.

Acknowledgments

The authors would like to thank Pentti Kanerva for his insightful review of this paper. We would like to thank Tom Landauer for providing the TASA corpus, and to acknowledge Google Inc. for their support of author DW's ongoing research on the subject. Author T.C. would like to acknowledge Arizona State University for support of this research, which was conducted during his time as faculty there. We would also like to thank Jieping Ye, Pentti Kanerva and Ted Dunning for drawing our attention to the correspondence between RRI and iterative methods of finding the principal eigenvector.

References

[1] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30(1):7–18.
[2] Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Brief Bioinform 2005;6(3):277–86.
[3] Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inform 2009;42(4):633–43.
[4] Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104:211–40.
[5] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inform Sci 1990;41:391–407.
[6] Peirce CS. Abduction and induction. In: Buchler J, editor. Philosophical writings of Peirce. New York: Dover; 1955.
[7] Bruza P, Cole R, Song D, Bari Z. Towards operational abduction from a cognitive perspective. Oxford University Press; 2006.
[8] Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc 2006;3:49–53.
[9] Schvaneveldt Roger, Cohen Trevor. Abductive reasoning and similarity. In: Ifenthaler D, Seel NM, editors. Computer based diagnostics and systematic analysis of knowledge. New York: Springer; in press.
[10] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 1997;91:183–203.
[11] Gordon MD, Lindsay RK. 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 Inform Sci 1996;47(2):116–28.
[12] Weeber M, Vos R, Klein H, Berg LTWDJ-VD, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature. A case report of a search for new potential therapeutic uses for thalomide. J Am Med Inform Assoc 2003;10(3):252–9.
[13] Srinivasan P. Text mining: generating hypotheses from MEDLINE. J Am Soc Inform Sci Technol 2004;55(5):396–413.
[14] Ganiz M, Pottenger WM, Janneck CD. Recent advances in literature based discovery. Technical report. Lehih University; 2005. LU-CSE-05-027.
[15] Kostoff RN. Literature-related discovery (LRD): introduction and background. Technol Forecast Soc Change 2007;75(2):165–85.
[16] Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. J Am Soc Inform Sci 1998;49(8):674–85.
[17] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods Instrum Comput 1996;28:203–8.
[18] Cole R, Bruza P. A bare bones approach to literature-based discovery: an analysis of the Raynaud's/fish-oil and migraine-magnesium discoveries in semantic space. Lecture notes in computer science: discovery science, vol. 3735. Berlin/Heidelberg: Springer; 2005. p. 84–98.
[19] Bruza PD, Widdows D, Woods J. A quantum logic of down below. In: Engesser Kurt, Gabbay Dov, Lehmann Daniel, editors. Handbook of quantum logic and quantum structures: quantum logic. Elsevier; 2009. p. 625–60.
[20] Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform 2009;42(2):390–405.
[21] Landauer TK, Laham D, Rehder B, Schreiner ME. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In: Shafto MG, Langley P, editors. Proceedings of the 19th annual meeting of the cognitive science society. Mawhwah, NJ: Erlbaum; 1997. p. 412–7.
[22] Laham D. Latent semantic analysis approaches to categorization. In: Shafto MG, Langley P, editors. Proceedings of the 19th annual meeting of the cognitive science society. Mawhwah, NJ: Erlbauml; 1997. p. 979.
[23] Giles JT, Wo L, Berry MW. GTP (General Text Parser) software for text mining. In: Bozdogan Hamparsum, editor. Statistical data mining and knowledge discovery. Boca Raton, FL, USA: CRC Press Inc.; 2003.
[24] Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. In: Proceedings of the 22nd annual conference of the cognitive science society; 2000. p. 103–6.
[25] Karlgren J, Sahlgren M. From words to understanding. Found Real World Intell 2001:294–308.
[26] Kanerva P. Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. Cogn Comput 2009;1(2):139–59.
[27] Bau III David, Trefethen Lloyd. Numerical linear algebra. Philadelphia: Society for Industrial and Applied Mathematics; 1997.
[28] Widdows D, Ferraro K. Semantic vectors: a scalable open source package and online technology management application. In: 6th International conference on language resources and evaluation (LREC); 2008.
[29] Vempala SS. The random projection method. In: DIMACS series in discrete mathematics and theoretical computer science, vol. 65, Providence, RI: American Mathematical Society; 2004.
[30] Johnson W, Lindenstrauss J. Extension of lipshitz mapping to hilbert space. Contemp Math 1984;26:189–206.
[31] Cohen TA. Exploring MEDLINE space with random indexing and pathfinder networks. AMIA Annu Symp Proc 2008:126–30.
[32] Burgess C, Livesay K, Lund K. Explorations in context space. words, sentences, discourse. Discourse Process 1998;25(2–3):211–57.
[33] Sahlgren M. The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Doctoral thesis. Stockholm University, Faculty of Humanities, Department of Linguistics.
[34] Rapp R. Word sense discovery based on sense descriptor similarity. In: Proceedings of the 9th machine translation summit, New Orleans; 2003. p. 315–22.
[35] Widdows D. Retraining document and term vectors, and refactoring the interface to sparse vector stores [Internet]. Available from: http://groups.google.com/group/semanticvectors/browse_thread/thread/d5885d4822b09444/8ce877844c1cb0af?lnk=gst&q=retraining#8ce877844c1cb0af.
[36] Gallant SI. Context vectors: a step toward a "Grand Unified Representation". In: Wermter S, Sun R, editors. Hybrid neural systems (LNAI 1778). Berlin, Heidelberg: Springer-Verlag; 2000. p. 204–10.
[37] Widdows D, Cohen T. Semantic vector combinations and the synoptic gospels. In: Bruza P, Sofge D, Lawless W, Van Rijsbergen CJ, Klusch M, editors. Proceedings of the 3rd quantum interaction symposium (March 25–27, 2009—DFKI, Saarbruecken). Springer; 2009. p. 251–65.
[38] Sahlgren M, Holst A, Kanerva P. Permutations as a means to encode order in word space. In: Proceedings of the 30th annual meeting of the cognitive science society (CogSci'08), July 23–26, Washington, DC, USA; 2008.
[39] Martin DI, Berry MW. Mathematical foundations behind latent semantic analysis. In: Landauer T, McNamara D, Dennis S, Kintsch W, editors. Handbook of latent semantic analysis. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.
[40] Widdows D. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In: Proceedings of the 41st annual meeting of the association for computational linguistics (ACL); 2003.
[41] Ipsen I, Wills RM. Analysis and computation of Google's PageRank. In: 7th IMACS international symposium on iterative methods in scientific computing. Toronto, Canada: Fields Institute; 2005.
[42] Kanaji S, Kanaji T, Furihata K, Kato K, Ware JL, Kunicki TJ. Convulxin binds to native, human glycoprotein Ib alpha. J Biol Chem 2003;278(41):39452–60.
[43] Sitbon L, Bruza P. On the relevance of documents for semantic representation. In: Proceedings of the 13th Australasian document computing symposium (ADCS 2008). p. 19–22.
[44] Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Stud Health Technol Inform 2001:1344–8.
[45] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006;39(6):600–11.
[46] Yang Y, Chute CG. A linear least squares fit mapping method for information retrieval from natural language texts. Proceedings of the 14th conference on computational linguistics, vol. 2. Nantes, France: Association for Computational Linguistics; 1992. p. 447–53.