MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-661

The Trajectory of Scientific Discovery: Concept Co-Occurrence and Converging Semantic Distance

Trevor Cohen[a] and Roger W. Schvaneveldt[b]

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

Abstract

The paradigm of literature-based knowledge discovery originated by Swanson involves finding meaningful associations between terms or concepts that have not occurred together in any previously published document. While several automated approaches have been applied to this problem, these generally evaluate the literature at a point in time, and do not evaluate the role of change over time in distributional statistics as an indicator of meaningful implicit associations. To address this issue, we develop and evaluate Symmetric Random Indexing (SRI), a novel variant of the Random Indexing (RI) approach that is able to measure implicit association over time. SRI is found to compare favorably to existing RI variants in the prediction of future direct co-occurrence. Summary statistics over several experiments suggest a trend of converging semantic distance prior to the co-occurrence of key terms for two seminal historical literature-based discoveries.

Keywords:

Literature-based discovery, Distributional semantics, Random indexing, Latent semantic analysis

Introduction

The field of literature-based knowledge discovery can be traced at its inception to the fortuitous discovery by Don Swanson of an implicit and therapeutically useful connection between fish-oil and Raynaud's Disease [1]. The idea underlying Swanson's approach to this problem is that two concepts may be meaningfully associated despite not yet having occurred together in the literature. The discovery of such implicit connections constitutes a literature-based knowledge discovery, and a number of computational approaches have been applied in an attempt to automate discoveries of this nature (for a review see [2-4]).

These approaches generally attempt to identify unknown connections by first identifying a "linking term" which co-occurs with the cue term. However, the explicit identification of linking terms imposes certain limitations on the discovery process. Firstly, the number of possible linking terms is generally large, and a combinatorial explosion in the size of the discovery search space occurs with the number of linking terms permitted on the path from cue to target. Consequently, automation of Swanson's discovery paradigm in this manner incurs I/O and computational costs that limit the possibilities for dynamic and responsive literature-based discovery tools. In this context, the development of methods to directly identify implicit connections, without the need for explicit identification of a linking term presents a desirable alternative. Such goals have led researchers in the field to explore methods of distributional semantics that directly identify implicit connections (without the need to explicitly identify a linking term), as an alternative [5, 6]. In our recent research, we have shown that Reflective Random Indexing (RRI) [7], a customized variant of the Random Indexing (RI) [8] approach to distributional semantics, is effective in identifying meaningful implicit connections.

However, a further limitation of existing approaches, including our own, is that they attempt to determine meaningful implicit connections by considering the distribution of terms or concepts in a corpus at a single point in time. From another perspective, the detection of meaningful implicit connections in a time-delimited set of documents can be viewed as the prediction of future explicit connections [9, 7]. One might hypothesize that changes in the strength of implicit associations over time would be important in predicting explicit connections in the future, as the associative strength between two concepts from disparate fields would be expected to grow as new connections are discovered between other concepts in these fields. In this paper we explore the extent to which changes in implicit associative strength over time are predictive of explicit connections in the future, in the context of two historical literature-based discoveries. In order to do so, we require a scalable model that is able to derive implicit connections from text, and is also easily incrementally updateable to accommodate adding new documents chronologically to the model without significant re-computation.

Both RI and RRI offer significant advantages in scalability over established methods such as Latent Semantic Indexing (LSI) [5] (for a computational details see [7]). In addition, both RI and RRI allow for incremental updates, as new documents are added to the corpus [8, 7]. As we have shown in our previous work [7], RI as originally implemented is ineffective in deriving meaningful associations between terms that do not occur together in any document. The reason for this is that RI produces a reasonably accurate reduced-dimensional approximation of the term-by-document matrix describing the

distributional statistics in the corpus concerned. Distance between terms is measured using the cosine (or normalized scalar product) between the vector representations of these terms in the reduced-dimensional space. Two terms that do not share any document between them will be represented as two vectors with no non-zero dimensions in common, and consequently their relatedness as measured by the cosine metric will be zero. RI does a good enough job of preserving the relative distances in the original matrix that there is a high probability of this being true in the reduced-dimensional matrix also [10]. In our previous work, we have addressed this limitation using RRI [7], however on account of its iterative nature the process for incremental updates using this method requires several steps.

Methods

To simplify the process of incremental updates for the purpose of exploring changes over time, we have developed a simple yet novel variant of RI which is also effective in deriving meaningful associations between terms that do not co-occur. The idea underlying this variant emerged from the observation that in the Singular Value Decomposition, which is used for dimension reduction in LSI, a reduced-dimensional approximation of the initial term-document matrix is obtained by finding the eigenvectors of the matrix AtA. This matrix is constructed by multiplying the term-document matrix by its transpose. Consequently a more consistent mapping between LSI and RI is obtained by adapting RI to reduce the dimensions of this symmetric matrix, in which each term is represented in terms of the extent to which its distribution across documents is similar to that of every other term. We will refer to this variant as Symmetric Random Indexing (SRI).

The procedure for dimension reduction of a term-document matrix using RI is as follows:

(1) Assign a *k*-dimensional (*k* in the order of 1000) zero vector to every **document**. We will term these elemental vectors.

(2) Set at random on the order of 10 of these zero values to +1, and the same number to -1.

(3) Assign a *k*-dimensional zero vector to every term. We will call these semantic vectors.

(4) Each time a term occurs in a document, add the elemental vector for this document to the semantic vector for this term (a weighting metric can optionally be applied here).

An advantage of RI is that the full term-document matrix is not ever represented in its entirety, which offers significant space advantages. As the elemental vectors are sparse, there is a high probability of their being orthogonal, or close-to-orthogonal to one another. Consequently, rather than assigning an orthogonal dimension to each document, as is the case in the full term-document matrix, a near-orthogonal elemental vector is assigned to each document, and semantic vectors for terms are projected directly into the reduced-dimensional space. The procedure for dimension reduction of the symmetric matrix AtA using SRI is as follows:

(1) Assign a high-dimensional (in the order of 1000) zero vector to every **term**. We will call these elemental vectors.

(2) Set at random on the order of 10 of these zero values to +1, and the same number to -1.

(3) Also assign a *k*-dimensional zero vector to every term. We will call these semantic vectors .

(4) Every time a term occurs in a document, add to its semantic vector the elemental vector for every other term occurring in the document, multiplied by the frequency with which this term occurs in the document (a weighting metric can be optionally applied here).

This procedure can be understood best by reviewing the process of matrix multiplication (Figure 1).

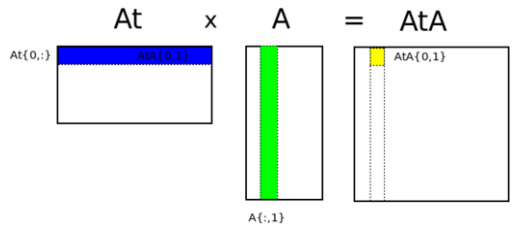

Figure 1- Matrix multiplication to generate AtA

A given cell, for example the cell {0,1} in the AtA matrix is obtained by taking the scalar product between the row At{0,:} and the column A{:,1} (Figure 1). To affect dimension reduction using random projection, we would like to assign a reduced-dimensional elemental vector to the column AtA{:,1}, thereby achieving dimension reduction by replacing each orthogonal dimension of the matrix AtA with a near-orthogonal approximation, in a similar manner to the process of RI. However, as the scalar product is a linear process, it is not necessary to calculate it simultaneously (Figure 2).

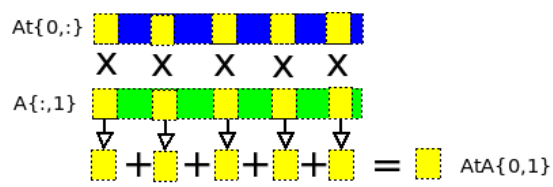

Figure 2-Components of the scalar product

Rather, we can calculate each linear component of the scalar product independently and multiply this by the elemental vector for AtA{:,1}. The sum of these linear components each multiplied by the elemental vector will equal the product of the elemental vector and the cell AtA{0,1}. In this way, we are able to create a reduced-dimensional approximation of AtA on a document-by-document basis, without constructing the full matrix. Consequently it is possible to derive implicit associations from large data-sets using this method. The space complexity of this model is $O(tk + ts)$ where t is the number of terms to be indexed, k is the dimensionality of the reduced-

dimensional space, and *s* is the number of non-zero values in each elemental vector. As with other variants of RI, time complexity is linear to the number of documents in the corpus.

In the experiments presented in this paper, we utilize SRI to derive implicit associations from abstracts in the MEDLINE corpus, and monitor the strengths of these associations as new documents are added incrementally to the model. In addition to the four steps of SRI described above, we use the log-entropy weighting metric to minimize the effect of disparities in local term frequency and emphasize the effect of terms that are focally represented [11]. In addition, we perform an on-the-fly equivalent of normalization of the document vectors to minimize the effect of differences in document length. Both of these customizations are common practice in Latent Semantic Analysis (LSA) [12], which has been successful in the derivation of meaningful implicit associations from smaller corpora in a number of applications [12]. In all cases, we generate vectors for those unique terms occurring 10 times or more in the corpus concerned, that do not contain any non-alphabet characters. In addition, we exclude terms on the stopword list distributed with the Arrowsmith system [13], which has been customized for the purpose of literature-based discovery.

Experiment 1: SRI and Implicit Associations

For the first experiment, we use SRI to derive a model of a time-delimited segment of the MEDLINE database. As in our previous research [7], this segment consists of the titles and abstracts of all citations added to MEDLINE between 1980 and 1985. For comparison with existing models, two 2000-dimensional SRI spaces are constructed, one with and one without the use of log-entropy weighting. For each of a set of 2000 randomly selected cue terms, the same set as employed previously, the 50-nearest indirect neighbors (NINS) are retrieved. NIN's are the terms most associated with a cue term that do not co-occur with it in any document in the corpus. Finally the proportion of these NINS that co-occur directly with the cue term in citations added to the 2008 release of MEDLINE after 1985 are evaluated. The assumption underlying this evaluation and similar methodologies [9], [14] is that an implicit connection at one point in time will be stated explicitly once it becomes discovered public knowledge.

Experiment 2: Trajectory of Historical Discovery

Having evaluated the ability of our model to derive meaningful indirect connections, we proceed to explore the trajectory of scientific discovery in the context of two of Swanson's seminal discoveries. For this experiment, we derive a 1000-dimensional space from all of the abstracts in the 2009 baseline release of MEDLINE (n=9,573,614), generating vectors for every unique term meeting the constraints described previously (n=333,214). We build this model incrementally, processing those abstracts added to MEDLINE every year, and evaluating the strength of association (as measured with the cosine metric) between the term "raynaud" and the term "eicosapentaenoic" which represents eicosapentaenoic acid, the active ingredient of fish oil. In addition, we keep track of the points in time when these two terms occur together in an abstract. As a control, we also assess the strength of association between the term "raynaud" and the term "kuru" which was selected by the authors on the account of their inability to conceive of a reason for these two terms to be meaningfully associated. On account of the stochastic nature of SRI and the small sample size, this experiment is repeated 50 times establish an average and to assess the statistical significance of the observed results. We perform the same experiment using the terms "migraine" and "magnesium", another implicit association identified by Swanson in his early experiments.

Results and discussion

Experiment 1

Figure 3 shows the results of the first experiment, presenting the proportion of the 50 NIN's extracted prior to 1985 that co-occur directly with their cue term after 1985.

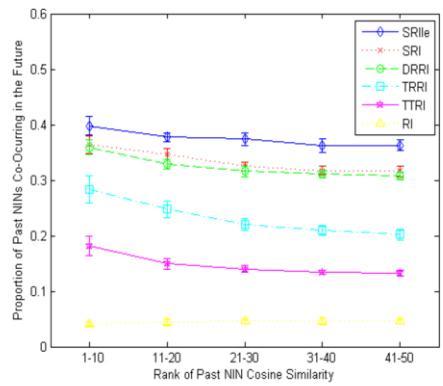

Figure 3-Comparison of RI variants. RI = Random Indexing as originally implemented [8], TTRI = term-term (or sliding window) RI [15], TRRI = term-based RRI [7], DRRI = document-based RRI [7], SRIIe = SRI with log-entropy weighting.

Each column in Figure 3 shows this metric for a range of rank, with the first column showing the proportion of the 10 nearest indirect neighbors that predict future co-occurrence. This column can be interpreted as precision at k=10, if future co-occurrence with the cue term is taken as a gold standard. SRI is compared to results obtained with other RI variants on this data set in a previous publication [7], which includes a detailed comparison of the methodological differences between these models. For the purposes of this paper we note that the SRI model is more effective in predicting future co-occurrence than any of the models evaluated in our previous work. As can be determined by the error bars on the graph (+-1 standard deviation) along with the sample size (n=2,000), most of the differences between SRI and other models are statistically significant, as are the changes in rank with decreasing cosine similarity. However, we note that SRI also favors NIN's with a higher global term frequency than other models, as illustrated in Figure 4. Terms with high global frequency are more likely to occur directly with cue terms by chance, and are generally less informative. Nonetheless, the results of this experiment show that log-entropy weighted SRI is able to predict future direct co-occurrence more effectively than any of the RI variants we have previously evaluated.

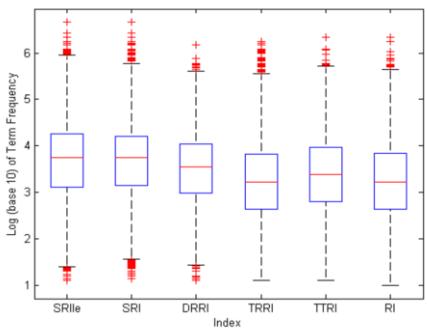

Figure 4-Boxplot of the global frequencies of NIN's.

This experiment proves SRI is able to predict future co-occurrence, which can be viewed as a measure of its ability to draw meaningful associations between terms that do not co-occur directly together in any document. When considering the ten NIN's, the precision of SRI with log-entropy weighting approaches 0.4. By comparison, in our previous research precision at k=10 on this data set ranged between 0.25 and 0.36 for RI variants, with one exception: a second iteration of the term-based RRI approach did produce a precision at k=10 of 0.4. However, repeated iterations make it more difficult to implement the incremental updates that are essential for the experiments presented here. All of these results exceed those obtained in a similarly-structured comparative study of discovery methods requiring explicit identification linking terms [9]. While these results are not strictly comparable due to the additional constraints and smaller size of the test set in the evaluation using linking terms, that the precision at k=10 of SRI in our experiment exceeds any published estimate for other methods in similar evaluations suggest it will make a useful addition to the set of tools currently employed for literature-based discovery. However, SRI tends to retrieve on average terms with higher global term frequency than other RI variants, suggesting that some of the predictions may be less informative than those produced by, for example, term-based RRI.

Experiment 2

Figure 5 illustrates the aggregated results of fifty runs of the second experiment. The figure show an increase in the association between "raynaud" and "eicosapentaenoic" that precedes the first time these terms co-occur together in a MEDLINE abstract, which is shown by a circular marker. We note that while the first observed co-occurrence occurs approximately five years after Swanson's discovery, the increase in associative strength begins before this. This increase in associative strength is also greater than that observed between "raynaud" and the control term, "kuru", and this pattern is consistent across runs. A similar pattern is observed in relation to the implicit association between "migraine" and "magnesium" (Figure 6).

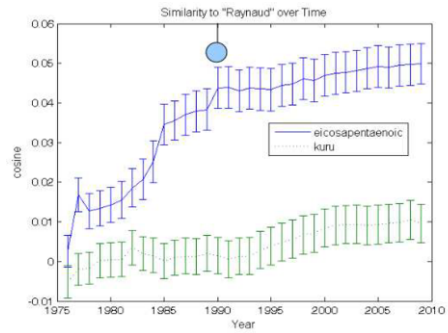

Figure 5-SRI associations between terms "raynaud" and terms "eicosapentaenoic" and terms "raynaud" and "kuru" between 1981 and 2009. Mean values over fifty runs depicted.

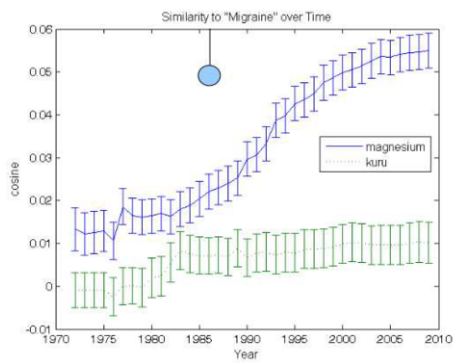

Figure 6-SRI associations between terms "migraine" and "magnesium" and terms "migraine" and "kuru" from 1981 to 2009. Mean values over fifty runs depicted.

In both cases, the increase in the strength of the implicit association precedes direct co-occurrence. This experiment provides some suggestive evidence that an approach incorporating changes over time may offer an advantage over approaches that view distributional statistics as a static snapshot in time. There is evidence from this study of a sharp rise in the strength of implicit association between the terms "raynaud" and "eicosapentaenoic", as well as between the terms "migraine" and "magnesium" that precedes both their first direct co-occurrence in an abstract and Swanson's seminal literature-based discoveries. However, as in all experiments simulating historical literature-based discoveries, there is the concern that findings related to a particular discovery may not generalize.

Limitations and future work

In our current research, we are investigating the extent to which trends in time, as measured using SRI, can be of use in the prediction of future co-occurrence from a large sample set. Results up to this point have been inconclusive: re-ranking of SRI-based NIN's according to the positive change in their cosine to the cue term over time does not appear to improve predictive ability. Furthermore, while terms with the greatest increase in association to a given cue term (Table 1) are often meaningfully related, this measure of relatedness does not

appear to be as predictive of future co-occurrence as SRI alone. Further research is required to develop methods that utilize the additional information provided by trends over time.

Table 1 – terms with the greatest increase in (direct) association strength (1980 and 1985)

thrombolysis	Schizophrenia
0.127: intracoronary	0.065: dimond
0.109: stk	0.055: casenotes
0.100: reocclusion	0.054: multiconditional
0.083: recanalisations	0.053: constructivist
0.077: reoccluded	0.049: schizoform
0.075: recanalised	0.048: balogh
0.074: recanalization	0.047: nonsimultaneity
0.067: thrombolysin	0.047: palau
0.065: reperfusion	0.045: murky
0.063: myocardial	0.044: schizoaffectives

While this work does demonstrate the effectiveness of SRI in deriving meaningful implicit associations, and provides some suggestive evidence that additional useful information is obtained by following trends in these associations over time, the methods used to measure changes over time are rudimentary. In our future work we will explore linear regression models and other more sophisticated approaches to measuring change over time to determine the extent to which these improve the ability to predict future co-occurrence. In addition, we will evaluate other factors we have not attended to here, such as the point in time at which future co-occurrence occurs.

Conclusion

In this paper, we present and evaluate SRI, a method which enables highly efficient updating of term vectors as additional documents are added to a database. The ability to track changes in the strengths of implicit associations over time is one of the advantages of the method. We find suggestive evidence that the additional information provided by this method may be of use in the prediction of future co-occurrence, and consequently it is of interest for literature-based knowledge discovery. The finding that this method is productive in deriving implicit associations has implications for information retrieval also, as it is both scalable and more conveniently amenable to incremental updates than similarly productive models.

Acknowledgments

We would like to acknowledge Dominic Widdows, for originating the open source Semantic Vectors package, some of which was adapted to the ends of this research.

References

[1] Swanson DR. Two Medical Literatures that are Logically but not Bibliographically Connected. JASIS. 1987; 38; p. 228-33

[2] Ganiz M, Pottenger WM, Janneck CD. Recent advances in literature based discovery. http://www.dimacs.rutgers .edu/~billp/pubs/JASISTLBD.pdf. Accessed 03/08/10.

[3] Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. Briefings in bioinformatics. 2005; 6(3); p. 277-286.

[4] Kostoff RN. Literature-related discovery (LRD): introduction and background. Technological Forecasting & Social Change. 2007 ; p. 165-185

[5] Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. JASIS. 1998 ; 49; p. 674-685.

[6] Cole RJ, Bruza PD. A Bare Bones Approach to Literature-Based Discovery: An Analysis of the Raynaud's/Fish-Oil and Migraine-Magnesium Discoveries in Semantic Space. LNAI. 3735; p.84-98

[7] Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. JBI. 2009 Sep 15. [Epub ahead of print]

[8] Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. Proceedings of the 22nd Annual Conference of the Cognitive Science Society. 2000 ; p. 1036

[9] Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. JBI. 2009; 42(4): p. 633-643;

[10] Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics. 1984 ; 26; p.189-206.

[11] Martin DI, Berry MW. Mathematical Foundations Behind Latent Semantic Analysis. In: Landauer, TK. McNamara, DS. Dennis, S. Kintsch, W. Eds. Handbook of Latent Semantic Analysis. 2007. Lawrence Erlbaum, NJ.

[12] Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997 ; 104; p. 211-240.

[13] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 1997 ; 91; p. 183-203. Stopwords at: http://arrowsmith.psych.uic.edu/arrowsmith_uic/data/stopwords_swanson (Accessed 03/08/10)

[14] Hristovski, D. Stare, J. Peterlin, B. and Dzeroski, S. Supporting Discovery in Medicine by Association Rule Mining and UMLS. In: Proceedings of MedInfo Conf. 2001. London, England.

[15] Karlgren J, Sahlgren M. From words to understanding. Foundations of Real-World Intelligence. 2001 ; p. 294-308.