Context-Driven Automatic Subgraph Creation for Literature-Based Discovery

Delroy H. Cameron  
Wright State University - Main Campus, cameron.20@wright.edu

Ramakanth Kavuluru  
Wright State University - Main Campus

Thomas Rindflesch

Amit P. Sheth  
Wright State University - Main Campus, amit.sheth@wright.edu

Krishnaprasad Thirunarayan  
Wright State University - Main Campus, t.k.prasad@wright.edu

See next page for additional authors

Follow this and additional works at: http://corescholar.libraries.wright.edu/knoesis

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation
http://corescholar.libraries.wright.edu/knoesis/851

This Article is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu.
Context-Driven Automatic Subgraph Creation for Literature-Based Discovery

Delroy Cameron\(^a\)\*, Ramakanth Kavuluru\(^b\), Thomas C. Rindflesch\(^c\), Amit P. Sheth\(^a\), Krishnaprasad Thirunarayan\(^a\), Olivier Bodenreider\(^c\)

\(^a\)Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis)
Wright State University, Dayton OH 45435, USA
\(^b\)Division of Biomedical Informatics, University of Kentucky, Lexington, KY 40506, USA
\(^c\)National Library of Medicine, 8600 Rockville Pike, Bethesda MD 20894, USA

Abstract

**Background:** Literature-based discovery (LBD) is characterized by uncovering hidden associations in non-interacting scientific literature. Prior approaches to LBD include use of: 1) domain expertise and structured background knowledge to manually filter and explore the literature, 2) distributional statistics and graph-theoretic measures to rank interesting connections and 3) heuristics to help eliminate spurious connections. However, manual approaches to LBD are not scalable and purely distributional approaches may not be sufficient to obtain insights into the meaning of poorly understood associations. While several graph-based approaches have the potential to elucidate associations, their effectiveness has not been fully demonstrated. A considerable degree of *a priori* knowledge, heuristics and manual filtering is still required.

**Objectives:** In this paper we implement and evaluate a context-driven, automatic subgraph creation method that captures multifaceted complex associations between biomedical concepts for LBD. Given a pair of concepts, our method automatically generates a ranked list of subgraphs, which provide informative and potentially unknown associations between such concepts.

**Methods:** To generate subgraphs, the set of all MEDLINE articles that contain either of two specified concepts (A, C) are first collected. Binary relationships or assertions, which are automatically extracted from the MEDLINE articles, called *semantic predications*, are then used to create a labeled directed *predications graph*. In this graph, a *path* is represented as a sequence of semantic predications. The hierarchical agglomerative clustering (HAC) algorithm is then applied to cluster paths, which are bounded by the two concepts (A, C) based on the definition of the *context* of a path, as a set of Medical Subject Heading (MeSH) descriptors. Paths that exceed a threshold of semantic relatedness are clustered into subgraphs based on their shared context. The automatically generated clusters are then provided as a ranked list of subgraphs.

**Results:** The subgraphs generated using this approach facilitated the rediscovery of 8 out of 9 existing scientific discoveries. In particular, they directly (or indirectly) led to the recovery of several *intermediates* (or B-concepts) between A and C, while also providing insights into the meaning of each association. Such meaning is derived from predicates between the concepts, as well as the provenance of the semantic predications in MEDLINE. Additionally, by generating subgraphs on different thematic dimensions (such as *Cellular Activity*, *Pharmaceutical Treatment* and *Tissue Function*), the approach enables a broader understanding of the nature of complex associations between concepts in a domain. In a statistical evaluation to determine the interestingness of the subgraphs, it was observed that an arbitrary association is mentioned in only approximately 4 articles in MEDLINE on average.

**Conclusion:** These results suggest that leveraging the implicit and explicit context provided by manually assigned MeSH descriptors is an effective representation for capturing the underlying semantics of complex associations, along multiple thematic dimensions for LBD.

**Keywords:** Literature-based discovery (LBD), Graph mining, Path clustering, Hierarchical agglomerative clustering, Semantic Similarity, Semantic relatedness, Medical Subject Headings (MeSH)

1. Introduction

Leveraging rich representations of textual content from scientific literature could be effective for finding and elucidating complex associations. Rich representations exploit implicit, formal (or explicit) and powerful semantics [1] to capture context, which may be important in providing deeper insights into the nature of associations. Gordon and Dumais made this crucial observation in [2] after successfully applying the popular technique of Latent Semantic Indexing (LSI) for LBD. The authors reported that LSI was only slightly more effective than
traditional frequency-based metrics, such as token frequency, record frequency and term frequency-inverse global frequency (tf-igf) [3] for finding intermediates. While LSI was successful for knowledge rediscovery, the authors speculated that richer representations of textual content are needed to capture "evidence suggestive of 'causal' relationships in the literature (which may be revealed independently of their statistical prominence)." Moreover, they stressed the need for "semantics and category knowledge to improve the step of identifying intermediate and terminal concepts."

Many techniques for finding hidden connections (or associations) between biomedical concepts from scientific literature however, utilize frequency-based and graph-theoretic metrics. Few methods have been developed to seamlessly find and elucidate complex associations, by going beyond reliance on implicit semantics. The conventional wisdom has been that discoveries are likely to arise from logical connections between source (A) concepts, intermediates (B) and targets (C) that frequently or rarely (co) occur in the literature, or are highly or rarely connected in a knowledge base. The earliest frequency-based approaches utilized 'frequency of occurrence' mainly through measures of term (and concept) frequency [4, 5]. Other measures such as relative frequency, token frequency, term frequency-inverse global frequency (tf-igf) [3] and term frequency-inverse document frequency (tf-idf) [6, 7] were also used to rank intermediates. Subsequent approaches utilized 'frequency of co-occurrence' using techniques such as LSI [2], association rules [8, 9, 10, 11] and probability distributions [11, 12, 13, 14]. Torvik et. al. [12, 15] even used an ensemble approach to find intermediates that combined statistical and temporal features.

While distributional approaches have been used for LBD, a critical issue is the realization that the underlying frequencies only provide an indirect way of capturing the meaning of associations. For instance, consider the association in which Dietary Fish Oils (A) inhibit Platelet Aggregation (B) and the aggregation of blood platelets causes Raynaud Disease (C); a circulatory disorder that causes periods of severely restricted blood flow to the fingers and toes [38]. While Dietary Fish Oils, Platelet Aggregation, and Raynaud Disease may frequently co-occur in the literature, their precise association is not explicitly captured by their co-occurrence. A second issue is that the underlying frequency distribution may not be adequate for capturing related concepts, which may be important in elucidating causal relationships and mechanisms of interaction.

To address these problems, several relations-based techniques [16, 17, 18] have been developed, which use the explicit relationships (or predicates) between concepts. Such predicates are typically obtained from structured background knowledge or known a priori by domain experts. For example, Hristovski et. al. in [17], developed a relations-based approach that used ordered alternating sequences of predicates and classes (or semantic types) called discovery patterns. These patterns are also specified a priori using insights from background knowledge. Using discovery patterns, Hristovski argues that if a Disease causes a change in a Substance/Body Function and a Drug inhibits this change, then the Drug MAYBE TREATS the Disease. The CAUSES-INHIBITS sequence is used to uncover potentially new Drug treatments for the Disease.

While intuitive, the relations-based approach is mainly applicable in scenarios where both predicates and semantic types are known, or can be easily obtained. This is not always trivial. For example, consider the complex scenario depicted in Figure 1, in which Dietary Fish Oils produce several Prostaglandins, including Prostaglandin F3 (PGF3) and Epoprostenol (PGI2 also called Prostacyclin). The latter of these Prostaglandins (Epoprostenol) treats Raynaud Syndrome and also disrupts Platelet Aggregation. Since Platelet Aggregation causes Raynaud Syndrome one can reasonably conclude that a plausible mechanism by which Dietary Fish Oils treat Raynaud Syndrome is through the production of Prostaglandins, which actively disrupt Platelet Aggregation, deemed a cause of Raynaud Syndrome. Obviously such complex situations, involving several predicates and disparate concepts may be unknown initially by domain experts. The relations-based approach is therefore considered unsuitable for complex associations.

Figure 1: Complex association between Dietary Fish Oils and Raynaud Syndrome

It can argued that hierarchical relations from the schema of a domain specific knowledge base, such as the Unified Medical Language System (UMLS) can also be used to create such complex subgraphs, using measures like specificity. However, the semantic types for Prostaglandins and Platelet Aggregation are Eicosanoids and Cell Function, respectively. These semantic types share no common ancestors in their lineage in the UMLS Semantic Network. And while associative relations can be used instead, a proven and repeatable schema-driven approach that captures this level of complexity has not been forthcoming. Instead, contemporary approaches to LBD focus on creating subgraphs, which comprise of binary relations, called semantic predications, of the form (subject, predicate, object). These predications are extracted directly from assertions in scientific literature. Willkowski ei. al. [19] developed a graph-theoretic approach based on semantic predications that iteratively (and manually) uses a greedy strategy to create the best subgraph,

\[\text{There are three semantic types for Prostaglandins (i.e., Eicosanoid, Pharmacologic Substance, and Biologically Active Substance) and one semantic type for Platelet Aggregation (i.e., Cell Function) in the UMLS. These semantic types do not overlap in the UMLS Semantic Network hierarchy – https://uts.nlm.nih.gov/semanticnetwork.html}\]
by weighting edges using degree centrality. This approach was used to elucidate the association among Norepinephrine, Depression, and Sleep.

Wilkowski’s approach bears semblance to the approach by Ramakrishnan et. al. [20], in which a greedy strategy is applied, using an ensemble of features, to generate complex associations. While the approach is fully automatic, it uses class and property specificity, instance-level rarity and refraction, applied to statements in a synthetically generated dataset, instead of semantic predications. Ramakrishnan notes that this approach was used, in exploratory research, to recover the connections from the Raynaud Syndrome – Dietary Fish Oils discovery. However its broader applicability for LBD in general has not been fully demonstrated. Reliance on hierarchical relationships in the UMLS Semantic Network is subject to inconsistencies since the UMLS is a terminology and not a formal ontology. Also, by design, the trees in the UMLS Semantic Network are fairly disjoint, as for Prostaglandins and Platelet Aggregation.

Goodwin et. al. [21] developed a hybrid approach that uses spreading activation for LBD, deriving weights from relative frequencies (of concepts and semantic predications) and degree centrality. This approach was used to successfully recover the intermediate Cortisol in the Testosterone – Sleep discovery [22], and also to elucidate the Norepinephrine, Depression, and Sleep scenario from [19]. However, Goodwin generates a list of intermediates, instead of a graph. It is unclear how the spreading activation algorithm might be adapted to capture the context of complex associations. In [23] van der Eijk et. al. clustered only MeSH descriptors (not semantic predications) into subgraphs, based on frequency of co-occurrence and Hebbian Learning. This approach provided new insights into the association between Deafness and Macular Dystrophy, and between Insulin and Ferritin. In recent work, Spangler et. al. [24] also used distributional statistics (tf-idf) to weight edges in a kinase network, represented as a Laplacian Matrix. The approach creates an n-ary similarity tree in which 7 new p53 kinases were discovered, which could revolutionize Cancer treatments.

In spite of the successes of graph-based and frequency-based approaches for LBD, more effective methods for capturing the context of associations are desired. Gordon and Dumais suggested a possible independence between frequency and causal-ity for LBD in [2]. We believe that complex associations that elucidate the relationships among concepts depend both on implicit and explicit context. Further, we believe that capturing such context may be the important in segregating complex associations along multiple thematic dimensions. For example, Figure 2 shows that Dietary Fish Oils and Raynaud Syndrome are associated in at least the following three ways: 1) in terms of Cellular Activity involving Blood platelets/Prostaglandins, as shown in Figure 2a, 2) through Pharmaceuticals such as Nifedipine and Verapamil, as shown in Figure 2b and 3) through Lipids/Fatty Acids from Efamol and Evening primrose oil, as shown in Figure 2c. The approach for clustering of cliques developed by Zhang et. al. [25, 26] may be used to capture subgraphs on multiple thematic dimensions. However, the approach is based on degree centrality and is therefore more likely to create subgraphs that only consist of highly connected concepts from the literature. In this paper, we explore the idea that hidden connections, and their related concepts, which help elucidate underlying complex associations, are more dependent on context than frequency, connectivity or specificity.

This research builds on our previous approach in [27] in which we rediscovered and decomposed the Raynaud Syndrome – Dietary Fish Oils discovery by using domain expertise, as context to manually create the subgraphs. Here, we present a method to automatically create the subgraphs by gleaning implicit and explicit context using MeSH descriptors. To achieve this, we first specify the context of a semantic predication and then use it to infer the context of a path. Such paths are then clustered into coherent subgraphs on multiple thematic dimensions, based on their shared context.

The approach requires only three items from the user as input: 1) a list of concept labels for source (A) and target (C), 2) the maximum path length of paths to be generated (default k = 2, for ABC associations) and 3) a cut-off date dt for articles to be included from the scientific literature. If no cut-off date is provided all MEDLINE articles are used. The output of the approach is a ranked list of subgraphs $S$.

To facilitate understanding the meaning of associations present in the subgraphs, the predicates of the semantic predications and their provenance in MEDLINE are provided (see Section 3). Relationships that are not explicit in the subgraphs, but are inferred, can be explored by composing MEDLINE queries (as we will show). The collective use of predicates, provenance and MEDLINE queries for knowledge exploration constitute the notion of discovery browsing, introduced by Wilkowski [19] and extended by Cairelli [28].

\(^2\)Medical Subject Headings (MeSH) – https://www.nlm.nih.gov/mesh/

\(^3\)Discovery browsing is enabled when a system guides the user through their
To assess the efficacy of our approach, two forms of evaluation were conducted. An evidence-based evaluation showed that the generated subgraphs could facilitate the rediscovery of 8 out of 9 existing discoveries [29, 30, 31, 32, 33, 34, 22, 28]. A statistical evaluation to determine the interestingness of the subgraphs (discussed in Section 3.2) showed that an arbitrary association occurs only in approximately 4 articles in MEDLINE on average. This suggests that the subgraphs created using our approach provide an effective way of finding and elucidating poorly understood associations. In this paper we make the following specific contributions:

1. We develop a novel context-driven subgraph creation method for closed LBD, capable of finding complex associations. Our approach is distinct from previous approaches, which are mainly based on statistical frequency, graph metrics and specificity.

2. We implement an unsupervised clustering algorithm to automatically create complex subgraphs using implicit and explicit context, without the need for complex heuristics for filtering.

3. We illustrate the role of discovery browsing, through the use of predicates and provenance to supplement the subgraphs with insights from the scientific literature.

4. We show the effectiveness of this approach in facilitating the rediscovery of 8 out of 9 existing scientific discoveries.

The rest of this paper is organized as follows: The approach to automatic subgraph creation is discussed in Section 2. Experimental Results are presented in Section 3 and a thorough discussion on limitations and future work are presented in Section 4. Conclusions are presented in Section 5.

2. Automatic Subgraph Creation

To automatically create complex subgraphs our approach relies on three datasets. The first dataset is MEDLINE, which is a repository of more than 23 million bibliographic citations maintained by the National Library of Medicine (NLM). The second is SemMedDB [35], a database of more than 65 million semantic predications extracted from MEDLINE using SemRep\(^4\). The third is the Biomedical Knowledge Repository (BKR), a knowledge base consisting of statements from the UMLS Metathesaurus together with semantic predications extracted using SemRep. These components are used for automatic subgraph creation in five steps: 1) Query Specification, 2) Candidate Graph Generation, 3) Path Context Representation, 4) Path Clustering and 5) Subgraph Ranking. Each step is discussed in the following subsections:


2.1. Query Specification

The system (called Obvio\(^5\)) first requires a query, denoted \(q\), which can be specified initially by providing the labels of two concepts of interest (A, C). These terms are manually mapped to concept unique identifiers (or CUIs), using the UMLS Semantic Navigator\(^6\). For example, the A-term Dietary Fish Oil, maps to the UMLS concept C0016157, whose label is also Fish Oils. Initial A- and C-terms are also manually augmented with other closely related concepts. For example, the concepts Fish oil – dietary (C0016157) and Eicosapentaenoic Acid (C0000545) are closely related to Fish Oils (C0016157) and are therefore added to the query. Next, the cut-off date \(dt\) for the literature to be included may be optionally provided. If no cut-off date is given the system uses the entire MEDLINE database. The maximum path length \(k\) of paths to be generated between A and C may then also be optionally provided. If none is given, the system defaults to a maximum path length of \(k = 2\). An example query for Raynaud Syndrome – Dietary Fish Oils could be \(q = ((\text{Fish Oils}, \text{Fish oil – dietary, Eicosapentaenoic Acid}), \{\text{Raynaud Phenomenon, Raynaud Disease}\}, 11/01/1985, 3)\).

2.2. Candidate Graph Generation

Given this query \(q = (A, C, dt, k)\), the Query Processor (Figure 3, top left) then retrieves the set of MEDLINE documents \(D\) that contain any of the terms (i.e., labels) in the A- and C- sets. These documents form the corpus from which semantic predications will be obtained. To obtain the predications, the set of PubMed identifiers (or PMIDs) for each article in \(D\) is passed to the Predications Graph Builder, which creates a labeled directed graph, called a predications graph denoted \(G\). To achieve this, the graph builder collects the semantic predications for each document in \(D\) that are also present in SemMedDB. The builder then creates a graph in which nodes are UMLS concepts

---


and edges are UMLS predicates. This graph is delivered as input to the Subgraph Generator, which first uses the Path Generator to extract all paths between (A, C) up to length k, using the Depth First Search (DFS) algorithm. DFS is selected because both A and C are known. However, the choice of Breadth First Search (BFS) may be equally effective for graph traversal, but has not been explored, since performance is not the primary focus. The path generator therefore uses the predications graph but has not been explored, since performance is not the primary focus. The path generator therefore uses the predications graph to produce paths (or ρ-path associations), where edges are oriented in either direction, consistent with our definition in [27]. This restricted set of paths is called the reachability relation R [36] (or candidate graph) between A and C at length k and date range dt. This candidate graph represents a more likely set from which discoveries may arise.

2.3. Path Context Representation

The candidate graph is then provided as input to the Path Clustering Module, which requires a definition for the context of a path p to cluster related paths into subgraphs. To specify path context, denoted C(p), we first specify the context of a semantic predication t, denoted c(t). The context of each predication in the path is then aggregated to obtain overall path context.

To define the context of a semantic predication, we make two assumptions, based on observations about MEDLINE articles. The first observation is that MeSH descriptors are manually assigned to MEDLINE articles (titles and abstract only) by MeSH indexers, based on human interpretation of the meaning of the entire article. These descriptors provide a concept-level semantic summary of the full text. Similarly, semantic predications also provide a semantic summary of the meaning of the content. However, semantic predications provide a relational semantic summary, by linking concepts using explicit predicates.

We therefore assume that the MeSH descriptors of an article capture its implicit context, which is shared by the semantic predications. A semantic predication can then be represented in terms of the MeSH descriptors of the article in which the predication occurs. This is the basis for our interchangeability assumption for subgraph creation, which states that the concept-level semantic summary and relational semantic summary of a MEDLINE article, are interchangeable. More specifically, given a semantic predication t and a MEDLINE article d such that t is extracted from d, the context of the semantic predication c(t) = M(d), where M(d) is the set of MeSH descriptors assigned to d.

If this assumption holds, then we can make a second assumption, which is that the overall context of a semantic predication t across the entire corpus can be represented as a vector of MeSH descriptors aggregated from each document containing t (based on distributional semantics). This is the basis for our context distribution assumption for subgraph creation, which states that the context of a semantic predication can be expressed as the distribution of all MeSH descriptors associated with all articles that contain the predication.

Since our fundamental premise for subgraph creation is that relatedness among concepts is independent of statistical frequency, graph metrics or specificity, our vector representation is downgraded to the Boolean-valued set representation, in which a MeSH descriptor is either present or absent in the distribution. The context of a path

\[ C(p) = \bigcup_{c(t) \in p} c(t) \] (1)

is therefore the aggregation of its predication context sets.

2.4. Path Clustering

The Path Clustering Module uses the context set C(p) for each path p in the candidate graph R to cluster related paths p₁ and p₂, based on their shared context. To compute this shared context between paths, the system initially computes the intersection s*(p₁, p₂) = C(p₁) ∩ C(p₂) of their shared MeSH descriptors. However, to account for inexact matches between MeSH descriptors across the two sets, this intersection is enhanced using the MeSH hierarchy, which provides explicit (or formal) semantics. Specifically, we use the Cartesian product of the two context sets C(p₁) × C(p₂) to determine which pairs of MeSH descriptors adequately indicate relatedness between the paths. Pairs of descriptors (mᵢ, mⱼ), whose similarity is above some threshold of MeSH semantic similarity are retained, while those below are discarded. The key idea is to maximize the weights of the in-context descriptors and minimize the weights of the out-of-context descriptors.

To compute semantic similarity between MeSH descriptors the measure of dice similarity is used. Dice similarity computes the proportion of common ancestors between descriptors in the MeSH hierarchy (MH). For two MeSH terms mᵢ and mⱼ, the dice similarity is computed as

\[ \text{dice}(mᵢ, mⱼ) = 2 \cdot \frac{|\text{ancestors}(mᵢ) \cap \text{ancestors}(mⱼ)|}{|\text{ancestors}(mᵢ)| + |\text{ancestors}(mⱼ)|} \] (2)

where \( \text{ancestors}(mᵢ) \) is the set of all ancestors of \( mᵢ \) in MeSH. The maximum similarity between two descriptors computed using dice similarity is 1. This occurs when the descriptors are equal. (i.e., \( mᵢ = mⱼ \)). The range of similarity values is therefore \([0, 1]\).

In this computation, pairs of descriptors, whose dice similarity exceed the threshold of semantic similarity (\( \tau_{\text{sim}} = 0.75 \)) are normalized to a value of 1. This normalized dice similarity

\[ \text{dice}_n(mᵢ, mⱼ) = \begin{cases} 1 & \text{if} \ \text{dice}(mᵢ, mⱼ) > \tau_{\text{sim}} \\ 0 & \text{otherwise} \end{cases} \] (3)

is therefore computed conditionally. The initial overall semantic relatedness

\[ s_r^*(pᵢ, pⱼ) = \sum_{(a, b) \in C(pᵢ) \times C(pⱼ)} \text{dice}_n(a, b) \] (4)

between \( pᵢ \) and \( pⱼ \) is the sum of the normalized pairwise dice similarity scores that exceed the threshold of semantic similarity, across the context sets \( C(pᵢ) \times C(pⱼ) \).

A consequence of this semantics-enhanced shared context, is that a broad range of relatedness scores may exist. Paths that are very similar, which have many exact (and inexact) MeSH
descriptors in common will have very scores, while others may have low scores. To dampen the major differences in similarity scores of different path pairs, we apply a log reduction on the normalized dice similarity scores. This is achieved by first computing the relatedness score between a given MeSH descriptor \( a \) in context set \( C(p_i) \) against the entire set of descriptors in the context set \( C(p_j) \). This computation yields the similarity score

\[
sim'(a, C(p_j)) = \sum_{b \in C(p_j)} \text{dice}_C(a, b).
\]

The log reduction

\[
\text{sr}_i^j(p_i, p_j) = \sum_{a \in C(p_i)} \log \left( 1 + \sim'(a, C(p_j)) \right)
\]

is then applied to \( \sim'(a, C(p_j)) \), and the overall semantic relatedness \( \text{sr}^j(p_i, p_j) \) between the two paths is the aggregate of the log-reduced scores for each descriptor in \( C(p_i) \) and the entire set in \( C(p_j) \). This metric is the basis for finding and elucidating complex associations among concepts, along multiple thematic dimensions, based on implicit and explicit context, alluded to by Gordon and Dumais in [2].

In the next step the hierarchical agglomerative clustering (HAC) algorithm is used to create subgraphs by clustering related paths. In the bucket population step, the algorithm initializes \( |R| \) buckets, one for each path in the candidate graph. For a given path, the relatedness score is computed for each of the remaining \( |R| - 1 \) paths. If two paths are sufficiently related, they must be placed in the same cluster. To achieve this, a method to automatically determine the threshold for path relatedness denoted \( \tau_{rel} \), is required.

To obtain the threshold for path relatedness the distribution of path relatedness scores between all pairs of paths in the candidate graph is pre-computed (i.e., \( (|R| \times |R| - 1)/2 \) scores). Figure 4 shows the distribution of relatedness scores for three experiments in the initial stages of our research. Each distribution approximates to a Gaussian (or normal) distribution.

### Table 1: Threshold Comparisons

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Path Relatedness Scores</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Std. Dev</td>
<td>Manual</td>
</tr>
<tr>
<td>Raynaud-Fish Oil</td>
<td>2.68</td>
<td>3.0</td>
</tr>
<tr>
<td>Testosterone-Sleep</td>
<td>3.35</td>
<td>3.5</td>
</tr>
<tr>
<td>DEHP-Sepsis</td>
<td>3.94</td>
<td>4.0</td>
</tr>
</tbody>
</table>

In statistics, the first standard deviation \((-\sigma, +\sigma)\) from the mean of a Gaussian distribution corresponds to the point of inflection. This point likely indicates a shift in a trend or phenomenon. When the manually determined thresholds for path relatedness for the same three experiments were compared to the \( \sigma, 2\sigma \) and \( 3\sigma \) of the Gaussian distribution, it was observed that the manual thresholds were consistently between the \( 2\sigma \) and \( 3\sigma \), as shown in Table 1. The second deviation from the mean of the Gaussian distribution \( (\tau_{rel} = 2\sigma) \) was therefore selected as the path relatedness threshold for clustering. During clustering, all pairs of paths with relatedness scores above this automatically determined threshold were added to the same cluster.

In the next phase of HAC (bucket merging), buckets that contain multiple paths are merged if their inter-cluster similarity exceeds the threshold for path relatedness. That is, for each pair of paths \( (p_i, p_j) \) across a pair of buckets \( B_a \) and \( B_b \), the inter-cluster similarity

\[
\text{sim}_{inter}(B_a, B_b) = \frac{\sum_{(p_i, p_j) \in B_a \times B_b} \text{sr}_i^j(p_i, p_j)}{|B_a| \cdot |B_b|},
\]

was computed as the sum of the semantic relatedness scores, normalized by the sizes of the two buckets. The clustering algorithm terminated when the number of clusters between successive iterations remained unchanged.

#### 2.5. Subgraph Ranking

The generated subgraphs are then ranked. Subgraphs containing more than one path are ranked in descending order, based on their intra-cluster similarity, which measures the compactness of the cluster. To compute this measure

\[
\text{sim}_{intra}(B) = \frac{2 \cdot \sum_{(p_i, p_j) \in B \times B} \text{sr}_i^j(p_i, p_j)}{|B| \cdot (|B| - 1)},
\]

the aggregate of the relatedness score for each pair of paths \( (p_i \neq p_j) \) in a given cluster \( B \) is obtained and then normalized. Singleton clusters consisting of only one path were ranked in ascending order using the measure of association rarity. That is, given a path \( p_i \), an association \( A(p_i) \) derived of the path, is the set of unique concepts in the path. Association rarity is the number of MEDLINE articles \( f(A(p_i)) \) that contain only the concepts in the path. For singleton paths, bucket rarity

\[
r(B) = \frac{\sum_{p \in B} f(A(p))}{|B|}
\]

is the same as association rarity, since \( B = \{p\} \).

The ranked list of clusters is rendered to the user for inspection in the Discovery Browsing Interface. This interface is available for review online (live tool – http://knoesis-hpco.cs.wright.edu/obvio/, video demo – http://bit.ly/obviomdemo). Concepts are color-coded based on semantic groups obtained from the BKR, while predicates are color-coded based on a locally developed coding scheme, since none exists for predicates in the BKR.

Using this approach, 8 out of 9 existing scientific discoveries were recovered. These well-known discoveries are: 1) Raynaud - Fish Oil (1986) [29], 2) Magnesium - Migraine (1988) [30], 3) Somatomedin C - Arginine (1990) [31], 4) Indomethacin - Alzheimer’s Disease (1996) [32], 5) Estrogen - Alzheimer’s Disease (1996) [33] 6) Calcium-Independent Phospholipase A2 - Schizophrenia (1998) [34], 7) Chlorpromazine - Cardiac Hypertrophy (2004) [14], 8) Testosterone - Sleep (2012) [22] and 9) Diethylhexyl (DEHP) - Sepsis (2013) [28]. In the next section the application of this approach for the rediscoveries is discussed.
3. Experimental Results

Given the absence of a gold standard dataset in LBD research, knowledge rediscovery is considered a de facto standard for evaluating LBD systems. To assess the effectiveness of the context-driven subgraph method, both an evidence-based evaluation and a statistical evaluation were conducted, based on nine existing scientific discoveries. The evidence-based evaluation qualitatively determines the extent to which our approach is capable of rediscovering the known knowledge, while the statistical evaluation is intended to measure the likelihood that a domain expert might be motivated to explore an arbitrary subgraph generated by the system. Specifically, it measures the interestingness of a subgraph, by quantifying the rarity of associations in the subgraphs in MEDLINE. Associations that have never been mentioned in any MEDLINE article are considered most interesting and are called zero-rarity associations (ZR).

The obvious caveat is that rare associations are not necessarily all interesting. The next section discusses the evidence-based evaluation.

3.1. Evidence-Based Evaluation

The first aspect of the evidence-based evaluation reports on the number of intermediates from a discovery that could be retrieved by the system. The second aspect substantiates the meaning of each association using evidence from the literature. Such evidence can be derived first using the predicates of the semantic predications in the subgraph. However, when this is insufficient or contradictory, evidence is obtained using the provenance of the predications in MEDLINE. Additionally, queries can be composed and executed in PubMed to explore inferred associations, not explicitly stated in the subgraphs, as commonly practiced.

For each rediscovery scenario, no concept filters were specified, to exclude less informative UMLS predicates, such as ASSOCIATED_WITH, INTERACTS_WITH and AFFECTS. This limited degree of manual filtering is the extent of a priori knowledge required for subgraph generation in the system. Also, in the following tables, the letter Y (for yes) is used to indicate that the status S of an intermediate as ‘found directly in a subgraph at position P in the list of subgraphs. The symbol Y∗ indicates that an intermediate was found through discovery browsing.

Due to space limitations, only three experiments are discussed in detail: 1) Raynaud - Fish Oil, 2) Magnesium - Migraine and 3) Somatomedin C - Arginine. The six remaining experiments are discussed briefly in Section 3.1.4. Further details on each experiment are available in [43] and in the following online supplementary materials: 1) the Obvio wiki page - (http://wiki.knoesis.org/index.php/Obvio, section on Automatic Subgraph Creation), 2) a video demo - http://bit.ly/obviom and 3) a beta-version of the Obvio web application - http://knoesis-hpco.cs.wright.edu/obvio/. The next section discusses the application of our approach to rediscover the Raynaud - Fish Oil rediscovery.

3.1.1. Raynaud Syndrome – Dietary Fish Oils

In November 1985, American Information Scientist Don R. Swanson (1924 – 2012) explored the research question of the role of Dietary Fish Oils (from salmon, mackerel, albacore, etc.) in Raynaud Syndrome. Through the methods described in [29], Swanson discovered that “dietary fish oil might ameliorate or prevent Raynaud’s syndrome.” This is because Dietary Fish Oils: 1) inhibit Platelet Aggregation, 2) increase the flow of blood (by reducing Blood Viscosity) and 3) also have a regulatory effect on the smooth muscle (thereby preventing Vasoconstriction and stimulating Vasodilation). Each of these concepts is causally implicated in Raynaud.

We seeded our algorithm with three concepts as sources: 1) Fish Oils (C0016157), 2) Fish oil - dietary (C0556145) and 3) Eicosapentaenoic Acid (C0000545) and two concepts as targets: 1) Raynaud Disease (C0034734) and 2) Raynaud Phe-
nomenon (C0034735). The corpus consisted of the relevant 61 full text articles discussed by Swanson [29] in the pre-November 1985 period. There were only 4 articles from the Dietary Fish Oil set, which were in the Raynaud set. The path length was set to 3 and no predicate filter was specified. These choices are consistent with the choices in our earlier experiments in [27], in which we rediscovered and decomposed this hypothesis by manually constructing the subgraphs, using domain expertise as context.

Figure 5: Subgraph1 (k = 3, 3σ) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome

The algorithm terminated in less than 5 minutes, producing 4 subgraphs (and 134 singletons) at 2σ and one subgraph (and 164 singletons) at 3σ. There were 1035 unique concepts and 4143 unique predications in the predications graph and the candidate graph contained 171 paths of length 3. Figure 5 shows that at 3σ, subgraph1 directly contains the intermediate Platelet Aggregation, which many rediscovery approaches consider sufficient to constitute a rediscovery. However, to better substantiate the association, we utilize the predications in the subgraph, together with the provenance of the predications in MEDLINE, along with traditional PubMed search, to provide evidence.

The predications, which states that [Eicosapentaenoic Acid CONVERTS_TO Prostaglandins] was extracted from the following corroboration sentence, in the full text of the following article [PMID60629798] by Harris et. al. The authors state that the “recent discovery that the prostanoids derived from eicosapentaenoic acid have biological effects different than those derived from arachidonic acid (C20:4w6) has generated further interest in fish oils.” Two of the other 61 articles [PMID6321621, PMID6314583] contained this predications. Harris also refers to the 1979 article [PMID218223] by Needlemans et. al., which suggests further that [Eicosapentaenoic Acid CONVERTS_TO Prostaglandin (PGI3)] is in its metabolic pathway. And the full text of 1985 article [PMID5997286] by von Schacky et. al. confirms that Eicosapentaenoic Acid produces Prostaglandin (PGI3) and Epoprostenol (Prostacyclin (PGI2)). von Schacky notes that “dietary EPA is transformed in vivo in humans into prostaglandins I3, which is as active... as the vasodilatory and antiaggregatory prostaglandin I2.”

The subgraph also contains the predications, which states that [Eicosapentaenoic Acid DISRUPTS Platelet Aggregation]. This predications was extracted from the full text of the article [PMID6320840] by Saynor et. al., who refers to the “Mechanisms underlying the inhibition of platelet aggregation by eicosapentaenoic acid and its metabolites.” The predications [Alprostadil DISRUPTS Platelet Aggregation] was extracted from the full text of the article [PMID3630714] by Dyerberg et. al., who pointed out that another author7 “was the first to show that [Prostaglandin E1] PGE1, inhibited platelet aggregation.” The previously article by von Schaky also alludes to this possibility.

Conversely, the predications [Epoprostenol TREATS Raynaud’s Phenomenon] was correctly extracted from two articles; by Dowd et. al. [PMID7037038], who discusses “Treatment of Raynaud’s phenomenon by intravenous infusion of prostacyclin (PGI2)” and by Belch et. al. [PMID3883365], who discusses “Increased prostacyclin metabolites and decreased red cell deformability in patients with systemic sclerosis and Raynauds syndrome.” Since both Alprostadil (PGE1) and Epoprostenol (PGI2) are synthetic forms of Prostaglandins, it is plausible that both Alprostadil and Epoprostenol actually treat Raynaud’s Syndrome by disrupting Platelet Aggregation. Indeed, the 1982 article [PMID6890719] by Pardy et. al., obtained through a date-restricted MEDLINE query8, confirms that Alprostadil (PGE1) treats Raynaud Phenomenon, instead of the weaker INTERACTS_WITH relationship, present in the subgraph. The role of Platelet Aggregation in causing Raynaud, which is inferred and not explicit in the subgraph, is easily confirmed using another MEDLINE query (Platelet Aggregation AND Raynaud AND 1865:1985/11[DP]), which yields the 1985 article [PMID39985417] by Soro et. al.

This subgraph together with discovery browsing suggest a richer relationship among Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome than would be provided by their co-occurrence. Rather, it appears that one mechanism by which [Eicosapentaenoic Acid TREATS Raynaud Syndrome] is by stimulating a series of Prostaglandins (namely, Prostaglandin I3 (PGI3), Prostaglandin E1 (PGE1) and Prostacyclin (PGI2)), which actually disrupt Platelet Aggregation.

An important observation is that the subgraph contains contradicting semantic predications. For example, the two predications [Eicosapentaenoic Acid CONVERTS_TO Prostaglandins] and [Eicosapentaenoic Acid INHIBITS Prostaglandins] are opposing. The full text of the article [PMID6827988] by Harris et. al., from which the predications [Eicosapentaenoic Acid CONVERTS_TO Prostaglandins] was extracted supports its claim. However, the full text of the lone article [PMID6301111] by Moncada from which the predications [Eicosapentaenoic Acid INHIBITS Prostaglandins] was extracted states that “It is clear, therefore, that both prostaglandin dependent and independent pathways of platelet aggregation are inhibited by EPA in vitro.” This is an incorrect extraction from SemRep. The author is noting that [Eicosapentaenoic Acid INHIBITS Platelet Aggregation], not Prostaglandins as the predications suggests. It is important to note that resolution of such discrepancies is part

8Query: Alprostadil AND Raynaud AND 1865:1985/11[DP]. Confirmed in search result #12
of the discovery browsing process, which requires adjudication by domain experts. We provide the infrastructure for achieving this through provenance.

The second intermediate Blood Viscosity, was found in the list of zero-rarity singletons (result #15 in Table 2). The actual singleton, which states that [Eicosapentaenoic Acid DISRUPTS Blood Viscosity], [Ketanserin DISRUPTS Blood Viscosity], [Ketanserin TREATS Raynaud Disease], suggests a causal relationship between Blood Viscosity and Raynaud Syndrome. This inferred relation that [Blood Viscosity CAUSES Raynaud Syndrome] is confirmed in the 1984 article [PMID6707529] by Larcan et. al through a MEDLINE search. The statement [Ketanserin DISRUPTS Blood Viscosity] is verified in the following articles [PMID401574], [PMID6303363] and [PMID2412054]. Likewise, the predication [Ketanserin TREATS Raynaud Disease] can be verified in the article [PMID6432198] by Roald et. al. and also [PMID6209510] by Bounamaeux et. al.

Table 2 shows the number of intermediates rediscovered for this experiment compared with 4 other approaches. The intermediate Vascular Reactivity (in reference to Vasoconstriction) was not found explicitly by our approach (although can be inferred from the article [PMID2997286] by von Schacky et. al.). This is not completely unexpected, since it is known from our reports in [27] that SemRep interprets “Vascular” and “Reactivity” as separate concepts. Hristovski in [17] was also subject to the same limitation.

Srinivasan [7] found all three intermediates in the top 2 of the top 30. However, note that Srinivasan’s approach relies on a priori knowledge of the semantic types of the intermediates for filtering and is manually intensive. Additionally, that approach does not create complex subgraphs, nor does it provide evidence for the meaning of associations using predicates. Hristovski et. al. [17] and Weeber et. al. [4] also require considerable domain expertise, particularly for specification of a priori relations (i.e., semantic types and discovery patterns). Gordon and Lindsay [6] find intermediates but make no attempt to elucidate the meaning of associations.

Figure 6: Subgraph1 \((k = 3, 2\sigma)\) on Dietary Fish Oils - Raynaud Syndrome (Blood Platelets/Prostaglandins)

To illustrate that our subgraphs capture different thematic dimensions of association between two concepts, consider the four subgraphs at 2\(\sigma\). Subgraph1 in Figure 6a is similar to subgraph1 (at 3\(\sigma\)) except that it includes the three additional intermediates, TIMP1, TIMP1 protein, human and Thromboembolism. This is naturally due to a lower threshold for path relatedness. By inspection, this subgraph elucidates the association between Dietary Fish Oils and Raynaud Syndrome through Blood Platelets/Prostaglandins, similar to the previous subgraph.

Figure 7: Subgraph2 \((k = 3, 2\sigma)\) on Dietary Fish Oils - Raynaud Syndrome (Pharmaceuticals)

Subgraph2 (shown in Figure 7) associates Dietary Fish Oils and Raynaud Syndrome from the perspective of Pharmaceuticals, including Nifedipine, Pentifylline, Thyrocalcitonin, and Trinitrin detailed especially in the article [PMID652267] by Kahan et. al., from which the predication [Nifedipine TREATS Raynaud Phenomenon] was extracted. Epoprostenol is the synthetic form of Prostacyclin (PGI2).

Figure 8: Subgraph3 \((k = 3, 2\sigma)\) on Dietary Fish Oils - Raynaud Syndrome (Lipids/Fatty Acids)

Subgraph3 in Figure 8 discusses the role of various Fatty Acids, which associate TIMP1, Epoprostenol, Efamol and Evening Primrose (see [PMID4082084, PMID6318123, PMID6321621]).

Figure 9: Subgraph4 \((k = 3, 2\sigma)\) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome (Blood Platelets)

Subgraph4 in Figure 9, which focuses more on Cellular Activity at the level of Blood Platelets involving Thromboembolism, is subsumed by subgraph1. Currently, subgraph subsumption has not been addressed in this work and remains a system limitation, discussed in Section 4. In the next section, the Migraine - Migraine experiment is discussed.
3.1.2. Magnesium – Migraine

In August 1987, Swanson explored the research question of the role of Magnesium in Migraine Disorder. Through the methods described in [30] he discovered 11 neglected connections between Magnesium and Migraine. He found that Magnesium deficiency might exacerbate Migraine due to complications involving Stress (Type A personality), Spreading Cortical Depression, Epilepsy, Platelet Aggregation, Serotonin, Substance P, Inflammation, Vasoconstriction, Prostaglandin formation and Hypoxia. Also, as a natural calcium channel blocker, Magnesium may further prevent Migraine attacks.

We seeded our algorithm with Magnesium (C0024467) as the source and Migraine Disorders (C0149931) as the target. The path length was 2 and no predicate filter was used, to be more consistent with the discovery. The corpus consisted of more than 47,000 articles from the pre-August 1987 period (i.e., 41,507 abstracts on magnesium and 6,171 on Migraine, 7 overlapping). There were 14697 unique concepts, 73,960 predications in the predications graph and 256 distinct paths of length 2 in the candidate graph. The algorithm terminated in less than one hour, producing 25 subgraphs (and 151 singletons) at 2σ and 6 subgraphs (and 231 singletons) at 3σ.

Figure 10 shows subgraph1 (k = 2, 2σ) Magnesium - Migraine

It was known from the 1973 article [PMID4725298] by Vosgeru (one of the 7 overlapping) that Magnesium Glutamate was used to treat Migraine. Figure 10 shows that the intermediate Serotonin was found in subgraph1 at 2σ. The lone article [PMID3629724] by Pertseva et. al. from which the predication [Magnesium INTERACTS_WITH Serotonin] was extracted, is inconclusive. According to Swanson this association should be that [Magnesium INHIBITS Serotonin]. The article [PMID3512233] by Houston et. al. from which the predication [Serotonin CAUSES Migraine] was extracted (among three others), suggests that elevated levels of Serotonin can induce Vasoconstriction, which causes Migraine. Houston explicitly states that “much evidence has implicated serotonin (5-hydroxytryptamine) in the pathogenesis of migraine.” The article further notes that Serotonin is released from Platelet Aggregation and might reach sufficient levels to exacerbate Migraine, as noted by Swanson. The 1987 article [PMID2440758] by Briel et. al. (through a MEDLINE search) confirms that Magnesium inhibits Platelet Aggregation. It follows that elevated Magnesium levels may inhibit both Serotonin and Platelet Aggregation, and so treat Migraine.

Figure 11 shows subgraph4, which contains the intermediate Prostaglandins between Magnesium and Migraine. The lone article [PMID3871957] by Friedlander et. al. from which the predication [Prostaglandins INTERACTS_WITH Magnesium] was extracted, suggests a decrease in prostaglandin synthesis is accompanied by lower levels of magnesium (and calcium). This is based on the title: “Decreased calcium and magnesium urinary excretion during prostaglandin synthesis inhibition in the rat” as noted by Swanson. The 1986 article [PMID3016750] by Nigam et. al. confirms that [Magnesium STIMULATES Prostaglandins] as suggested by Swanson. The article [PMID89390] by Hakkarainen et. al. from which the predication [Prostaglandins ASSOCIATED_WITH Migraine Disorders] was extracted (among only three others) states that “Tolfenamic acid (a potent inhibitor of prostaglandin biosynthesis) was effective in treating acute migraine attacks.” The specific role of Prostaglandins in Migraine was unclear however, even after discovery browsing. Swanson suggested that [Prostaglandins INHIBITS Migraine].

Figure 12 shows that the intermediate Vascular Disease was found explicitly in subgraph9. The title of the article [PMID4260015] by Bristo et. al. from which the predication [Magnesium ASSOCIATED_WITH Vascular Disease] was extracted, suggests a role for magnesium in vascular reactivity. The title of the article reads in part, “...Findings in magnesium metabolism in vascular diseases.” Similar to the predi-
Table 3: Comparison of rediscoveries with other approaches for Magnesium - Migraine

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium Channel Blockers</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>Y+</td>
<td>9</td>
<td>–</td>
<td>Y</td>
<td>8</td>
<td>Y</td>
</tr>
<tr>
<td>Hypoxia</td>
<td>–</td>
<td>5</td>
<td>–</td>
<td>–</td>
<td>6</td>
<td>Y</td>
</tr>
<tr>
<td>Inflammation</td>
<td>Y+</td>
<td>ZR-3</td>
<td>Y</td>
<td>2</td>
<td>Y</td>
<td>170</td>
</tr>
<tr>
<td>Platelet Activity</td>
<td>Y+</td>
<td>1</td>
<td>Y</td>
<td>2</td>
<td>Y</td>
<td>2</td>
</tr>
<tr>
<td>Prostaglandins</td>
<td>Y</td>
<td>4</td>
<td>Y</td>
<td>1</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>Type A Personality</td>
<td>–</td>
<td>Y</td>
<td>Y</td>
<td>1</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>Serotonin</td>
<td>Y</td>
<td>1</td>
<td>Y</td>
<td>1</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td>Cortical Depression</td>
<td>–</td>
<td>Y</td>
<td>6</td>
<td>–</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>Substance P</td>
<td>–</td>
<td>Y</td>
<td>18</td>
<td>Y</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>Vascular mechanisms</td>
<td>Y</td>
<td>9</td>
<td>Y</td>
<td>1</td>
<td>Y</td>
<td>46</td>
</tr>
</tbody>
</table>

Figure 12: Subgraph9 ($k = 2, 2\sigma$) Magnesium - Migraine

Subgraph7 (shown in Figure 13) for example, links Theophylline and Caffeine, with Magnesium and Migraine, which have different semantic types, but belong to the general group of Stimulants. Subgraph6 (not shown) associates Epinephrine and Glucose from the perspective of Metabolism. Table 3 shows that ultimately, 7 out of the 11 associations found by Swanson could be found using our approach.

3.1.3. Somatomedin C – Arginine

In April 1989, Swanson explored the research question of the role of the dietary amino acid Arginine, in Growth along with the protein Somatomedin C (also called Insulin-Like Growth Factor 1 (IGF1)). Through the methods discussed in the brain), which leads to Brain Edema (referred to as or inflammation by Swanson), was found among the zero-rarity associations (see Table 3). The remaining intermediates Hypoxia, Spreading Cortical Depression, Stress (Type A Personality) and Substance P were not found among the subgraphs.

Interestingly, only subgraph22 on the calcium channel blockers was a complex subgraph in which existing knowledge was recovered. While several intermediates related to Vascular Reactivity, such as Vasospasm, Vascular Function, Vasoconstriction and Vascular Disease exists, their shared context did not meet our threshold for path relatedness and hence they were not grouped into the same cluster. The shortcomings of SemRep in extracting Vascular Reactivity may also have been a limiting factor. Still, altogether 10 out of the 25 subgraphs contained complex associations.
Swanson discovered 4 implicit connections between Somatomedin C and Arginine. He found that Arginine intake could: 1) stimulate Growth and protein synthesis, 2) promote Wound Healing and cell regeneration, 3) facilitate nutritional repletion and overcoming Malnutrition and 4) improve Body Mass (and Weight), especially in the elderly and debilitated.

We seeded our algorithm with Somatomedins (C0037657) and Insulin-Like Growth Factor I (C0021665) as the sources, and Arginine (C0003765) as the target. The corpus consisted of more than 11,000 articles (819 on Somatomedins and 10,698 on Arginine (with 53 overlapping), in the pre-April 1989 period. The path length was set to 2, this time with the STRICT predicate filter. There were 5195 concepts and 17,058 predications in the predicates graph and 239 distinct paths in the candidate graph. The algorithm terminated in less than one hour producing 10 subgraphs (and 153 singletons) at 2σ and 7 subgraphs (and 205 singletons) at 3σ.

Figure 14: Subgraph5 (k = 2, 3σ) Somatomedin C – Arginine

Figure 14 shows the intermediate Growth Hormone in subgraph 5 at 3σ. The sequence of predications [Arginine STIMULATES Growth Hormone] and [Growth Hormone STIMULATES Somatomedins] is entirely correct and requires no further proof (in terms of rediscovery). Still, for verification, we confirmed in the article [PMID6394628] by Chew et al. that dietary Arginine stimulates the release of Growth Hormones. These Growth Hormones then stimulate the production of Somatomedin C (IGF1), which leads to cell growth and increased body size and muscle (i.e., protein synthesis), as noted in article [PMID7194347] by Clemmons et al. The same association is captured in subgraph6 at 2σ (not shown).

In subgraph5, several articles from which the seemingly spurious predication [Arginine TREATS Child] was extracted, upon investigation, were shown to actually discuss Glucagon and Insulin. This includes the article [PMID7204541] by Blethen et al. whose title is “Plasma somatomedins in children with hyperinsulinism.” Likewise, the article [PMID6205015] by Bioux et al. from which the predication [Arginine TREATS Rat-tus norvegicus] was extracted, discusses observations regarding Insulin-like Growth Factor I in the serum of rats. The article [PMID7007553] by Ashby et al. from which the same predication was extracted, discusses the effects of Progesterone and Insulin in rats, resulting from Glucose and Arginine stimulation. Based on these observations, it is reasonable to conclude that this subgraph captures the shared context of role of Insulin in Somatomedin C and Arginine.

Subgraph7 at 3σ (not shown) contains the concept Growth as an intermediate instead of Growth Hormone (similar to subgraph2 at 2σ, also not shown). The sequence of predications [IGF1 CAUSES Growth] and [Growth PRODUCES Somatomedins], is interesting because the article [PMID3748655] by van Buul-Offers et al. from which the predication [IGF1 CAUSES Growth] was extracted states that IGF1 “increases body length and weight, as well as the growth of several organs of Snell dwarf mice,” which is consistent with Swansons report. The association between Malnutrition and Somatomedin production was found in the article [PMID7023246] by McCumbee et al., from which the predication [Growth PRODUCES Somatomedins], was extracted. No obvious association to Wound Healing was found using our methods. Table 4 shows that 3 out of 4 intermediates could be found using our approach.

Table 4: Comparison of rediscoveries with other approaches for Somatomedin C - Arginine

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron [S]</th>
<th>Srinivasan [7] [P]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somatomedin C - Arginine</td>
<td>Growth Hormone</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Body Weight</td>
<td>Y*</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Malnutrition</td>
<td>Y*</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Wound healing</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>–</td>
<td>4</td>
</tr>
</tbody>
</table>

3.1.4. Remaining Experiments

This section briefly presents the results for the remaining 6 rediscoveries attempted. Each intermediate marked as rediscovered was found using predicates, provenance and discovery browsing as previously done.

Scenario 4: For the Indomethacin - Alzheimer’s Disease discovery [32] by Smallheiser and Swanson in 1995, there were 15 subgraphs at 2σ. Srinivasan found all 8 intermediates, while we only recovered 6 out of 8 intermediates from subgraphs 2, 3, 4 and 14 (shown in Table 5). Recall that all our subgraphs are all automatically generated.

Table 5: Comparison of rediscoveries with other approaches in the literature for Indomethacin - Alzheimer’s Disease

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron [S]</th>
<th>Srinivasan [7] [P]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indomethacin - Alzheimer’s Disease</td>
<td>Acetylcholine</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Lipid peroxidation</td>
<td>Y*</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>M2-muscarinic</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Membrane Fluidity</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Lymphocytes</td>
<td>Y*</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Thyrotropin</td>
<td>Y</td>
<td>ZR-20</td>
</tr>
<tr>
<td></td>
<td>T-lymphocytes (T-Cells)</td>
<td>Y*</td>
<td>3</td>
</tr>
</tbody>
</table>

Scenario 5: For Estrogen - Alzheimer’s Disease [33] by Smallheiser and Swanson in 1995, we found 3 out of 8 interme-
Table 6: Comparison of rediscoveries with other approaches in the literature for Estrogen - Alzheimer’s Disease

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron S</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen - Alzheimer’s Disease</td>
<td>Antioxidant activity Y*</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alioprotein E (ApoE) Y*</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Calbindin D28k Y</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catepsin D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytochrome C oxidase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glutamate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receptor Polymorphism Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Superoxide Dismutase</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Comparison of rediscoveries with other approaches in the literature for Calcium-Independent PLA2 - Schizophrenia

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron S</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium-Independent PLA2 - Schizophrenia</td>
<td>Oxidative stress Y*</td>
<td>3</td>
<td>Y</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Selenium</td>
<td>Y*</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vitamin E</td>
<td>Y*</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Scenario 6: For Calcium-Independent PLA2 - Schizophrenia [34] by Smalheiser and Swanson in 1997, our algorithm produced 10 subgraphs at 2σ, all of which were singletons. Here, our results are comparable to Srinivasan’s, except that we are able to retrieve the article [PMID7782894] by Kuo et al. deemed crucial to the discovery, through discovery browsing from singleton2. The seemingly innocuous singleton in subgraph2, which states that [Phospholipase A2 INHIBITS Proteins] [Proteins CAUSES Schizophrenia] leads to the article [PMID7739414] by Berry, from which the predication [Proteins CAUSES Schizophrenia] was extracted. The article shows that the specific protein discussed was the selenium transport protein Selenoprotein P, as noted by Smalheiser. The article by Kuo is #4 in the search results of a MEDLINE search for Phospholipase A2 AND Selenium AND 1865:1997[DP].

Scenario 7: For Chlorpromazine - Cardiac Hypertrophy [14] by Wren et al. in 2002, there were 14 subgraphs at 2σ. The intermediate Isoproterenol was found in subgraph12 (as shown in Table 8). The article [PMID6165961] by Rossi et al. from which the predication [Chlorpromazine INHIBITS Isoproterenol] was extracted, together with the article [PMID203365] by Tsang et al. from which the predication [Isoproterenol CAUSES Cardiomegaly] was extracted, substantiated these predications. Subgraph5 contained the predications [Chlorpromazine INHIBITS Calcineurin] extracted from the article [PMID9001710] by Gong et al. and the predication [Calcineurin CAUSES Cardiac Hypertrophy] extracted from several articles, including [PMID9568714, PMID10679475, PMID11248077, PMID11773940, PMID10189350].

Scenario 8: For Testosterone - Sleep [22] by Miller and Rindflesch in 2011, which articulates that “testosterone enhances sleep quality by inhibiting cortisol,” we found 11 subgraphs at 2σ and 10 subgraphs at 3σ. Cortisol (or Hydrocortisone) was found in subgraph7 at 3σ and also in subgraph11 at 2σ. The article [PMID8548511] by Kern et al. confirmed that [Hydrocortisone DISRUPTS Sleep], while the crucial article [PMID15841103] by Rubinow et al., noted by Miller, confirms that [Testosterone INHIBITS Hydrocortisone].

Scenario 9: For Diethylhexyl Phthalate (DEHP) - Sepsis [28] by Cairelli and Rindflesch in 2013, which articulates one possible mechanism for the obesity paradox [38], we did not find the intermediate PParGamma altogether.

In summary, several approaches succeed in providing automation for finding intermediates. These approaches leverage keyword-based, concept-based relations-based, graph-based and hybrid techniques. Many also provide predicates between concepts, while more recent approaches are able to substantiate intermediates with provenance in MEDLINE. The main innovation of our approach is that we are able to retrieve and substantiate existing discoveries, on different thematic dimensions, using implicit and explicit context as suggested by Gordon and Dumais [2], not frequency, graph metrics or specificity. To the best of our knowledge, an approach that has rediscovered as many intermediates, with such degree of automation and substantiation of the knowledge, has never been developed. In the next section the statistical evaluation is presented.

3.2 Statistical Evaluation

In the previous section, we showed that our context-driven subgraph method facilitated the rediscovery of 8 existing discoveries with their substantiation in MEDLINE. While these are encouraging results, one might argue that our experiments were biased since we knew the intermediates to be found in the first place. Hence, it was easy to find them in the subgraphs. A more important question is how interesting are subgraphs in general, such that an arbitrary domain expert might be motivated to explore them altogether? To address this question, we conducted a statistical evaluation, which uses association rarity to compute interestingness. If the interestingness score of the subgraphs across an entire experiment is low, then the rediscoveries were fortuitous and the associations that led to the

Table 8: Comparison of rediscoveries with other approaches in the literature for Chlorpromazine - Cardiac Hypertrophy

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron S</th>
<th>S</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorpromazine - Cardiac Hypertrophy</td>
<td>Calcineurin</td>
<td>Y</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Isoproterenol</td>
<td>Y</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>
rediscoveries were serendipitous, rather than systematic. While this
not a complete loss, it is still less than ideal.

To perform this evaluation, for each path in each subgraph
across the 8 rediscoveries (excluding singletons), a PubMed
query was executed using the eUtils Web Service. This was
used to determine the number of documents that contain the as-
association in MEDLINE, with the date restriction enforced.
For example, for the path [Arginine STIMULATES Growth Hormone], [Growth Hormone STIMULATES Somatomedins], the
query “Arginine AND Growth Hormone AND Somatomedins
AND 1865:1989[DP]” was composed, where Arginine, Growth Hormone, and Somatomedins represent an association.

The interestingness of an experiment
was computed as the normalized reciprocal of rarity.

\[ r(E) = \frac{1}{r(A(p_i)) + 1} \quad (10) \]

of a set of associations across all subgraphs in an experiment
E, is computed as the average of the association rarity, where
\( f(A(p_i)) \) is the frequency of a unique association in MEDLINE.

The interestingness of an experiment
was computed as the normalized reciprocal of rarity.

\[ I(E) = \frac{1}{r(E) + 1} \quad (11) \]

was computed as the normalized reciprocal of rarity.

Table 10 shows the rarity and interestingness scores for each
of the eight successful rediscoveries. For the Raynaud Syn-
drome – Dietary Fish Oils experiment, there were 10 unique
intermediates/associations among the 4 subgraphs at 2 σ, all
of which had a zero-rarity in MEDLINE. This is not surpris-
ing, since Swanson noted in [29] that only four articles from
the Raynaud literature overlapped with the Fish Oil literature
by 1986. The rarity of these subgraphs is therefore 0.00, and
the interestingness is 1 (meaning absolutely interesting).

For Magnesium – Migraine there were 48 unique associ-
tions. The most commonly known intermediates were Hyper-
tensive Disease (3), Individual (3) and Vascular Diseases (4),
respectively among a total of 27 documents. The overall rarity
of the subgraphs in the experiment was therefore 27/48 = 0.56
and the interestingness is 0.64.

For Somatomedin C – Arginine there were 18 unique associ-
tions/intermediates among a total of 306 documents. The most
commonly known intermediates were Child (16), Somatropin
(63) and Growth Hormone (63). There were only two zero-
rarity associations, which were from the intermediates Mus (0)
and Falls (0). Clearly these are not interesting. Not surpris-
ingly, the overall interestingness score of these subgraphs was
306/18 = 17 and the interestingness is low (0.06). This suggests
that this field was better studied than others. It also partially
supports the observation by Gordon and Dumais [2] that while
frequency of intermediates may be sufficient for finding novel
intermediates in some cases, it may be insufficient to capture
the related concepts that elucidate complex associations.

There were 21 unique associations for Indomethacin –
Alzheimers. Hydrogen Peroxide (2), Interleukin-1 (2) and Free
Radicals (3) were the most commonly known intermediates
among a total of 9 documents. The overall rarity score was
9/21 = 0.43 and the interestingness is 0.70.

For Estrogen – Alzheimers there were 42 unique associations,
with 36 zero-rarity associations. Metabolism (6), Dementia
(10) and Senile dementia (10) were the most commonly known
intermediates among a total of 36 documents. The rarity score
is 36/42 = 0.86 and the interestingness is 0.54.

For Calcium-Independent PLA2 – Schizophrenia there were
10 unique intermediates/associations (singletons described in
Section 4.1.6), each of which was zero-rarity. Hence, the rarity
of this subgraph is 0.00 and the interestingness was high (1.0).

For Chlorpromazine – Cardiac Hypertrophy there were 21
unique associations, and 19 at zero-rarity. The most commonly
known were Catecholamines (1) and Hypertensive disease (1)
among a total of 2 documents. The rarity is therefore 2/21 =
0.10 and the interestingness is high (0.91).

For Testosterone – Sleep, there were 61 unique association-
s/intermediates and 20 at zero-rarity. The most commonly
known were Proteins (63), Symptoms (91) and Hormones (207)
among a total of 654 documents. The overall rarity score was
therefore 654/61 = 10.72 and the interestingness is low (0.09).
This is not surprising, since these two domains (Testosterone
and Sleep) are fairly well studied. This is another scenario that
suggests frequency may not be as effective for finding the 
related concepts that elucidate complex associations.

Across all 8 rediscoveries, the average rarity score is there-
fore 3.71 and the average interestingness was 0.62. This sug-
gests that an association chosen at random from the rediscover-
ies is likely to be known to only approximately 4 documents in
MEDLINE. Such a low rarity score suggests that the subgraphs
themselves might be quite interesting to a domain expert. This
is however not surprising, since most of the discoveries, at the
time when made would have been inherently interesting situa-
tions and possibly not well studied in the literature. Testos-
terone – Sleep (2011) and Somatomedin C – Arginine (1990)
are exceptional.

4. Discussion

This paper showed that the use of implicit and explicit con-
text to find and elucidate associations among concepts along
multiple thematic dimensions is effective for LBD. However,
our approach has several limitations. The first limitation is
the assumption that the context of a semantic predication, ex-
pressed in terms of the distribution of MeSH descriptors is rel-
liable for generating meaningful subgraphs. Not all MeSH de-
scriptors assigned to an article are relevant to all its semantic
predications, and hence the predication context vectors could
be noisy. Ideally, direct mappings between semantic predic-
actions and MeSH descriptors could help resolve this discrepancy.
Since, such mappings are unavailable our specification of con-
text is subject to limitations of distributional semantics.

The second limitation is the degree of domain expertise still
required for discovery browsing. Although impractical to elim-
nate the need for domain expertise, one improvement would be

---

providing additional background knowledge to supplement the subgraphs where appropriate. In this way, assertional knowledge from the literature would be complemented with definitional knowledge from structured knowledge sources (though deep integration). Metrics for determining interesting neighboring concepts in background knowledge need to be developed for concepts in the subgraph to overcome this limitation.

Another limitation is the inability to systematically detect contradicting semantic predications. While the provenance of predications in MEDLINE allows domain experts to adjudicate, a method for resolving conflicting predications could be beneficial. We believe that temporal analysis of semantic predications could enable conflict resolution. However, since many unresolved paradoxes inherent in science itself are reported in the literature, it is unclear whether one might reliably detect and resolve such contradictions automatically, using temporal, statistical and/or semantic approaches.

The reliability of the statistical evaluation is also another limitation of our approach. Rare associations are generally interesting but not always. While alternative methods for conducting statistical evaluation for LBD have been discussed [39], it is cumbersome to coordinate cut-off dates for each predication across the rediscoveries. The suggested techniques are therefore impractical to implement. We use association rarity to indicate interestingness, similar to existing research [20, 7].

A number of technical limitations exist in our approach. The first technical limitation is the manual selection of a threshold for MeSH semantic similarity based on dice similarity. While dice is advantageous because it is easy to implement, other similarity metrics and more principled ways of computing the threshold should be explored. Likewise, the threshold for path relatedness, which is based on the second (and third) standard deviation from the mean of the Gaussian distribution, could be unreliable. Our results show that the data distributions only approximate to Gaussian. The p-values from the $\chi^2$ test of the three Gaussian distribution in Figure 4 indeed more than the 0.05 value normally considered acceptable. To overcome this limitation, we anticipate that path relatedness could be recomputed relative to the minimum relatedness score. Torvik et. al. [12] and Smalheiser et. al. [15] implemented an approach based loosely on this idea, which normalized the distribution, using a mixture of Gaussian models.

Across some experiments, we utilized predicate filters to eliminate non-informative relationships (such as ASSOCIATED_WITH, INTERACTS_WITH, AFFECTS, etc). This is a compromise to achieve scalability. Ideally, the system should not require any predicate filters. In fact, the omission of some predications may be responsible for low recall in some of our experiments. Still, given that most experiments terminated in less than one hour, higher recall may not be too costly for performance. With the emergence of big data infrastructure, the performance limitations of our clustering may be resolved using additional computational resources.

The choice of HAC could be considered another limitation. HAC was selected because it is an unsupervised, deterministic clustering algorithm, for which the number of clusters does not have to be known or specified a priori. The time complexity of HAC is $O(N^2 \log N)$ in the best case. While approaches, such as those by Ramakrishnan et. al. [20] and van der Eijk et. al. [23] may be applicable for subgraph creation, it is unclear how they might be adapted to generate complex subgraphs along multiple thematic dimensions.

These and other limitations suggest the next steps in this research. In future, labels for subgraphs should be provided. This is a crucial task, since our approach is predicated on the idea that each subgraph captures a different thematic dimension of association between two concepts. Additionally, a comparative study using full text, compared to titles and abstracts, could be useful. However, full text is only available on a limited scale.

5. Conclusion

Leveraging rich representations of textual content from scientific literature based on implicit and explicit context can provide effective means for literature-based discovery, as illustrated in this paper. This paper also shows that these rich representations were important in facilitating the rediscovery of 8 out of 9 well-known associations and their substantiation. The approach advances LBD research specifically because it uses notions of context and shared context to facilitate making discoveries from scientific literature, which do not rely on statistical frequency, graph metrics or specificity. A beta-version of the Obvio web application is available online for optional viewing (http://knossos-hpc.cs.wright.edu/obvio/), along with various other resources (wiki page -

Conflict of interest: None.

6. Acknowledgement

This research was supported in part by the Intramural Research Program of the US National Institutes of Health, National Library of Medicine. We would especially like to thank Pavan Kapanipathi, Wenbo Wang and Shreyansh Bhatt for their insightful feedback on many aspects of this work. We also thank Swapnil Soni, Nishita Jaykumar, Vishnu Bompally, Gary A. Smith, Drashti Dave, Swapna Abhyankar, Mike Cairelli and Gaurish Anand, who contributed to other aspects of this work.

References

[34] Smalheiser NR, Swanson DR. Calcium-independent phospholipase a2 and schizophrenia. Arch Gen Psychiatry 1998;55(8):752–3.