Path Knowledge Discovery: Association Mining Based on Multi-Category Lexicons

Chen Liu*[1], Wesley W Chu*, Fred Sabb†[2], D. Stott Parker* and Joseph Korpela*
*Computer Science Department, University of California, Los Angeles, USA
†Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, USA

***Abstract*—Transdisciplinary research is a rapidly expanding part of science and engineering, demanding new methods for connecting results across fields. In biomedicine for example, modeling complex biological systems requires linking knowledge across multiple levels of science, from genes to disease. The move to multilevel research requires new strategies; in this paper we present *path knowledge discovery*, a novel methodology for linking published research findings. Path knowledge discovery consists of two integral tasks: 1) association path mining among concepts in a multipart lexicon that crosses disciplines, and 2) fine-granularity knowledge-based content retrieval along the path(s) to permit deeper analysis. Implementing this methodology has required development of innovative measures of association strength for pairwise associations, as well as the strength for sequences of associations, in addition to powerful lexicon-based association expansion to increase the scope of matching. In our discussions, we describe the validation of the methodology using a published heritability study from cognition research, and we obtain comparable results. We show how path knowledge discovery can greatly reduce a domain expert's time (by several orders of magnitude) when searching and gathering knowledge from the published literature, and can facilitate derivation of interpretable results.**

***Keywords*-path data mining; text mining; path knowledge discovery; content-based retrieval**

I. Introduction

Increasingly, scientific discovery requires the connection of concepts across disciplines, as well as systematizing their interrelationships. Doing this can require linking vast amounts of knowledge from very different domains. Experts in different fields still publish their discoveries in specialized journals, and even with the increasing availability of scientific literature in electronic media, it remains difficult to connect these discoveries. For example, an expert in neuropsychiatric syndromes such as schizophrenia and attention deficit hyperactivity disorder (ADHD) may know little about genetics, while an expert in genetics may lack knowledge of the cognitive phenotypes of syndromes such as ADHD. Although informatics tools such as search engines are very successful when it comes to helping people search for and retrieve information, these systems unfortunately lack the capability to connect that knowledge. Furthermore, using these systems to manually search a large corpus is not only time consuming, it can be infeasible and can lead researchers to be overly reductionist in a biased or arbitrary manner. To overcome this basic problem, new methodologies are needed for scalable and effective knowledge discovery and integration.

This work was motivated specifically by research on complex neuropsychiatric syndromes such as ADHD. Even with a large corpus of relevant work, it can be difficult for researchers studying such syndromes to find experimental results that examine direct relations between the syndromes and concepts from other scientific disciplines, such as genetics. Instead, they must investigate chains of associations that span multiple disciplines. For example, their research may be examining a hypothesized causal relation between mutations in the gene dopamine receptor D2 (DRD2) and the syndrome ADHD. However, the actual support for this is only formed by a chain of relations across multiple disciplines represented by a series of questions such as: *What symptoms are related to ADHD? Which parts of the brain would be affected? How is DRD2 related to the functioning of these parts of the brain?* While a search of the corpus for documents containing the terms "DRD2 AND ADHD" may fail to discover experimental results presenting this direct relation, a more elaborate query that connects the results of multiple experiments, such as "DRD2 $\rightarrow$ prefrontal cortex $\rightarrow$ working memory $\rightarrow$ attention deficit $\rightarrow$ ADHD," would allow the researchers to retrieve all the experimental results along this chain, so that they could rapidly examine the existing support for such a multilevel hypotheses, see Figure 1.

We can conceptualize and visualize this knowledge integration process as drawing a path from one concept to the other, connecting related concepts from one domain to concepts in other domains. When considering multiple domains, such connections form a path which represents the knowledge structure across the domains. We refer to the process of forming this path as path knowledge discovery.

Path knowledge discovery is challenging for the following reasons. First, a path describes a sequence of interrelated associations across multiple domains of knowledge. Although existing data mining methods such as Apriori [1] perform well when identifying high-confidence pairwise associations, mining interrelated associations still remains an open problem. Second, identifying possible associations alone does not provide sufficient information for path knowledge discovery. It is important to understand how the concepts

[1]Current affiliation: Google Inc., California, USA
[2]Current affiliation: Lewis Center for Neuroimaging, University of Oregon, USA

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

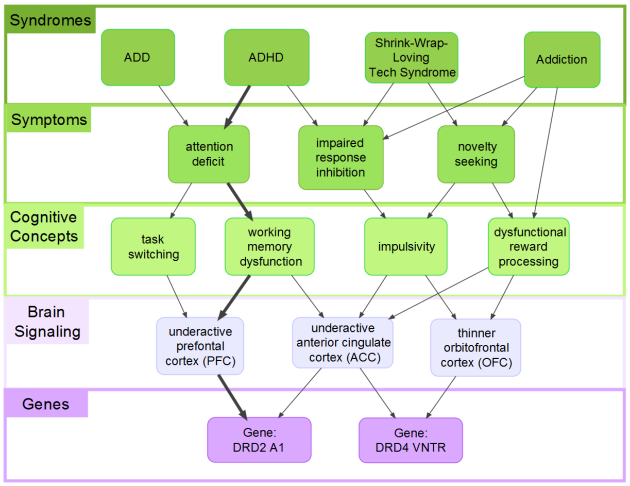

Figure 1. Knowledge paths. Each node represents a concept term found in a corpus made up of publications from the field of neuropsychiatry. Edges between terms represent the existence of published research results linking two concepts. In this example, the lack of an edge between ADHD and Gene: DRD2 A1 would indicate no publications in the corpus directly link these concepts; however, by tranversing several intermediate relations along the path "ADHD → attention deficit → working memory dysfunction → underactive prefrontal cortex → Gene: DRD2 A1" a multilevel hypothesis spanning the relations between concepts along the path could be examined.

are interrelated, and therefore it is necessary to retrieve information that can support the associations. Based on this description, the path knowledge discovery problem can be decomposed into two integral parts: 1) identifying paths describing relations among concepts at multiple concept levels, and 2) retrieving content corresponding to the paths from the corpus to explain the interrelations.

In this paper we propose a framework for path knowledge discovery. This framework uses a multilevel lexicon to index and query a corpus of scientific knowledge. Paths of interrelated concepts that span multiple domains are identified based on a user-supplied query, and these paths are evaluated to determine which paths show the strongest associations across the entire path. The strongest of these paths are then used to retrieve relevant content to allow the user to quickly evaluate this content and determine which paths best support their query, see Figure 2.

II. Infrastructure for Path Knowledge Discovery

Path knowledge discovery is the exploration of the associations between concepts across different domains of knowledge, with the associations connected to form a path across the domains. These domains of knowledge are represented by a corpus of scientific research. Constructing paths across this corpus requires a knowledge of the hierarchy between the concepts in the domains. In this paper we build this hierarchy using a multilevel lexicon which gives a controlled vocabulary of the concepts belonging to different domains. Along with the corpus and lexicon, we build two indexes, an association index and a document index, which link the concepts in the lexicon to content in the corpus to facilitate search and retrieval.

A. Multilevel Lexicon

The multilevel lexicon is a controlled vocabulary of concepts at different levels that provide a knowledge of synonyms and a concept hierarchy. It is not fixed and can evolve with time and changing evidence base. The lexicon used in this research was constructed by domain experts according to a multilevel schema that groups concepts into levels corresponding to different domains [2]. Within a domain the concepts are further organized in a tree structure so that more specific concepts can be defined as sub-concepts of more general ones (see Figure 9 for an example lexicon). This hierarchical structure is useful for query preprocessing, such as query expansion (Section III-A). Beyond providing this hierarchical structure, the lexicon also includes lists of synonyms for common terms.

B. Corpus of Scientific Research

The corpus used for this research consists of a large number of full-text peer-reviewed publications retrieved from PubMed Central [3]. PubMed Central distributes these publications in XML format, which provides the text with tags marking the structural information of the document, allowing the content to be accessed at different granularities, i.e., paper, section, paragraph, or sentence. This is useful when conducting path mining as it allows one to search for associations at different granularities. Another added benefit of this structural information is the ability to identify document components such as captions and tables, allowing concepts to be associated with specific graphical elements within a document.

C. Indexes for Path Knowledge Discovery

Indexing is an essential infrastructure component for efficient retrieval of content and query answering. This research uses two types of indexes to facilitate path knowledge discovery. The first of these indexes is a document element index, which facilitates content retrieval. This index includes three fields: a document element id, a concept id, and the occurrence frequency for the concept appearing in the document element. Using this index, content at different granularities can be retrieved by either a document element id or a concept id. The corpus can then quickly be searched for the occurrence frequency of concepts, and relevant content can be retrieved at different granularities. The second of these indexes is an association index. If we envision a graph of concepts, where concepts are vertices and an edge exists between two vertices if and only if the corresponding concepts co-occur in some document element, then the association index is equivalent to the edge list of

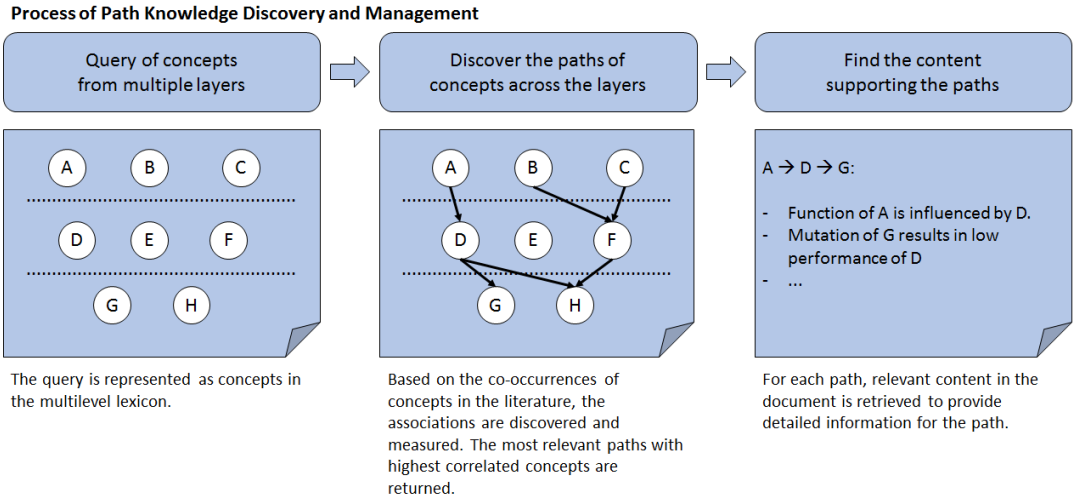

Figure 2. Process flow of path knowledge discovery

this graph. The association index describes relations between concepts, allowing us to combine such relations to answer path queries for interrelations across multiple concepts in different domains. This index also stores associations at multiple granularities, allowing for searches to be first conducted at a coarser granularity, e.g., paper or section, and then repeated recursively at finer granularities, e.g., paragraph or sentence, on the results obtained.

III. Path Mining

Path mining is the process of discovering path knowledge from a large corpus of text data. The objective is to search for the paths that satisfy a query pattern, and from those results identify the paths with the strongest associations.

A. Path Mining Queries

Path knowledge indicates a pattern of associations among concepts across different domains of knowledge, forming a path across those domains. These patterns can be represented by search queries in which one specifies k domains $D_1, D_2, \ldots, D_k$, with each domain having an associated concept $c_1, c_2, \ldots, c_k$, with these concepts connected across the domains using the set of all possible patterns $c_1 \rightarrow c_2 \rightarrow \ldots \rightarrow c_k$ such that concept $c_i \in D_i$ and $i = 1, 2, \ldots, k$. Unlike a simple keyword-based search of the corpus, these queries are not merely a list of concepts that should be searched for in all documents, but instead identify which pairings of concepts from which domains should be considered, increasing the likelihood that the paths discovered will be relevant to the research problem being queried.

Posing a path mining query requires some prior knowledge of each of the domains involved in the query, i.e., which concepts should be specified in which domains. More importantly, since the query defines the specific pairings of concepts (i.e., co-occurrences in a document element) to be found, the results are limited by which domains are included in the query. If a query does not include a specific domain, then any paths that are connected using that domain will not be discovered. To address this problem, we introduce the idea of "wildcard queries" in path mining, where queries can leave multiple intermediate domains as unspecified. If a wildcard is used for a domain, then all concept terms from that domain are considered, greatly increasing the number of paths considered between the explicitly queried domains. Viewing the associations among concepts as a graph, then we can think of this as a multipartite graph where each domain represents a separate partition and concepts represent the vertices within those partitions. Without wildcard queries, only paths with edges extending directly between the partitions specified in the query would be discovered. However, with wildcards, we can allow paths to traverse one or more unspecified intermediate partitions between those that were specified, see Figure 3.

B. Measures for Path Strength

A path reveals knowledge about the relationships between concepts in different domains. If there are multiple paths that satisfy a query, then we seek those that are most relevant. In order to evaluate the potential relevance of paths, we evaluate the strength of pairwise associations in paths using their co-occurrence frequencies.

1) Measuring the Strengths of Pairwise Associations: An association between two concepts is the simplest path. Measuring the strength of pairwise associations is the first step towards measuring the strength of a more complex path. In the context of text mining, the co-occurrence of concepts is an indicator of association. If two concepts tend to appear in the same paper, the probability is higher

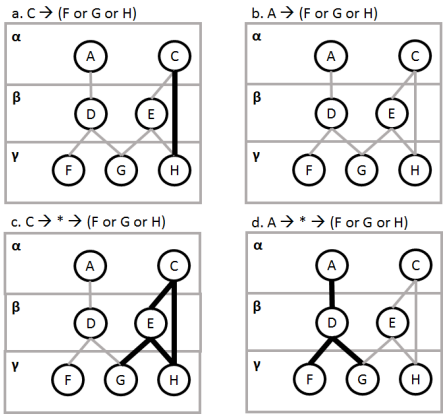

Figure 3. Use of wildcards in path queries. (a) and (b) show the results of two example queries without wildcards, with A-H representing concept terms and α, β, and γ representing domains. (c) and (d) show the results of those example queries with wildcards (represented by *) added. The discovered paths are represented in bold.

that these two concepts are related to each other. Such co-occurrence frequencies can be derived from the association index (Section II-C). The data mining community uses support and confidence to measure the strength of an association $A \rightarrow B$ between concepts A and B [1]:

$$support(A \rightarrow B) = \sigma(A \cap B) \tag{1}$$

$$confidence(A \rightarrow B) = \frac{\sigma(A \cap B)}{\sigma(A)} \tag{2}$$

where $\sigma(A)$ stands for the proportion of the documents in the corpus containing the concept A, and $\sigma(A \cap B)$ stands for the proportion of the documents in the corpus containing both concepts A and B.

Support measures the proportion of documents in which two concepts co-occur, and represents the probability of co-occurrence across the whole corpus. Confidence estimates the conditional probability of occurrence of B given A's occurrence. If we consider the occurrence of A and B as random events, we can also measure the strength of the association using the Pearson correlation $\rho_{A,B}$ between the two events

$$\rho_{A,B} = \frac{E(A,B) - E(A)E(B)}{\sqrt{E(A)(1-E(A))}\sqrt{E(B)(1-E(B))}} \tag{3}$$

where $E(A)$ is the expectation of the probability of occurrence for the concept A (i.e., $\sigma(A)$). Tan et al. [4] pointed out that $\rho(A,B)$ can be approximated by $IS(A,B)$

$$\rho_{A,B} \approx IS(A,B) = \sqrt{I(A,B) \times \sigma(A,B)} \tag{4}$$

where $I(A,B) = \frac{p(A,B)}{p(A)p(B)}$ is the interest factor [5]. The interest factor computes the ratio of the probability of co-occurrence and the expected probability of co-occurrence given that X and Y are independent of one another. The above approximation holds when $I(A,B)$ is high, and both $p(A)$ and $p(B)$ are very small, which in general fits the case of occurrences of concepts in a large text corpus. We can regard IS as an alternative interpretation of the association rule that does not indicate an inference from antecedents to consequents, but rather a measure of closeness between two concepts.

The conventional association rule mining problem is to find all associations whose strength indicators, such as support, confidence, and IS measure, are above given thresholds. Algorithms such as Apriori [1] solve the problem by generating the frequent item sets and then counting the support for the candidates in a bottom-up fashion. The FP-growth algorithm [6] solves the problem with the efficient data structure, the frequent pattern tree (FP-Tree). Path mining on the other hand goes beyond the individual associations, measuring the strength of a sequence of associations across all the concepts included in a path across multiple domains.

2) Measuring Strength of Associations in the Path Context: A path consists of a sequence of associations. In order to find paths with high association strength, we can impose a strength threshold on all the associations in the path. As with pairwise associations, each association in the path connects two concepts. However, since there are multiple associations in the path, measuring the strength of associations is more complicated. We use two approaches to measure the strength of associations: local strength and global strength.

The local strength measure considers the strength of individual associations as a "local" property. Each association in the path is independent of other associations and thus is only related to its direct antecedents and consequents. Therefore, the computation of association strength as a local strength measure is identical to the computation for pairwise relations (Equations (1), (2), and (4)).

The global strength measure considers the strength of individual associations as a "global" property of the entire path. In this setting, each association is related to the preceding associations. To compute association strength, we group all the concepts involved in previous associations in a path as the antecedent. For example, the second link of $A \rightarrow B \rightarrow C \rightarrow D$ would be regarded as $AB \rightarrow C$. In this case, the measurement of support and confidence differs from simple pairwise association mining. Specifically, support of the second link of $A \rightarrow B \rightarrow C \rightarrow D$ is $\sigma(A \cap B \cap C)$, and the confidence can be computed as

$$Confidence(AB \rightarrow C) = \frac{\sigma(A \cap B \cap C)}{\sigma(A \cap B)} \tag{5}$$

With this definition, the confidence is the conditional probability that C is part of the path given that $A \rightarrow B$ is part of the path. The correlation measure of the link can be derived by computing the correlation between two random events: the co-occurrence of all previous antecedents as one random event and the occurrence of the consequent as the other.

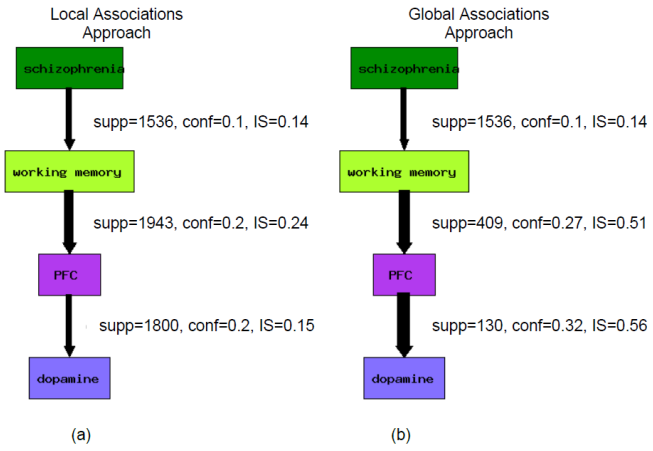

Figure 4. Two different approaches for measuring the strength of associations for the path "schizophrenia → working memory → prefrontal cortex (PFC) → dopamine": (a) the local strength measure, (b) the global strength measure. The thickness of the links in the path is proportional to the IS score of the corresponding associations. The support, confidence, and IS are derived from a sample of the PubMed Central corpus.

According to this definition, the correlation score of the second link of $A \rightarrow B \rightarrow C \rightarrow D$ can be computed as $IS(AB, C)$.

Figure 4 presents an example path measured by the two different approaches. The support, confidence and IS measure are computed using local strength measure and global strength measure in Figure 4(a) and Figure 4(b), respectively. For the global strength measure, since the association takes all preceding concepts as an antecedent, the support value of the association decreases when the path length increases, and confidence and IS scores change correspondingly. This property makes it more difficult to find a high-support path when more concepts are included in the path.

The major difference between the two approaches lies in the different requirements for co-occurrence of concepts in the paths. In the global approach, all the concepts are required to appear at least once in the same document element in order to ensure a non-zero confidence. On the other hand, the local approach only requires adjacent concepts in the path to appear in the same document elements. Therefore, the local approach can be applied to scenarios focusing on discovery of new paths and generating new hypotheses.

C. Path Mining Algorithms

Based on the choice of the association strength measurement, the path mining problem can be transformed into two different problems and solved by corresponding algorithms. When using the local strength measure, the path mining problem is equivalent to a graph search problem. For the global approach, the path mining problem can be viewed as an extension of traditional association rule mining.

1) Path Mining as a Graph Search Problem: For the local strength measure, the path discovery process is that of finding strongly connected pairwise associations across the domains specified in the path query. We can construct a graph of concepts whose edges are these associations. Then the path mining problem is equivalent to a graph search problem which finds paths in the graph that satisfy the path query, and for which the strengths of associations meet the desired threshold.

We can use graph traversal algorithms such as breadth-first search to examine the candidate associations. For example, assume concepts $a_1, a_2, \ldots, a_m$ are in level 1, concepts $b_1, b_2, \ldots, b_n$ are in level 2, and concepts $c_1, c_2, \ldots, c_l$ are in level 3, with each level representing a different domain of concepts from the lexicon. We can first draw an edge $a_1 \rightarrow b_1 \rightarrow c_1$, then draw an edge $a_1 \rightarrow b_1 \rightarrow c_2$, and so on. We can then pick the paths that meet the thresholds of association strength for measures such as support, confidence, and IS. In the case of answering wildcard queries, we will add wildcard levels between each level. The complexity of the computation can be $O(b^k)$ where b is the number of concepts in the level and k is the number of levels involved (including wildcard levels). This process can be computationally expensive when b or k is large. However, this high computational cost can be reduced significantly by introducing pruning steps when traversing the graph. According to our definition of association strength measures, computation of the strength of an association in a path is only affected by preceding associations (e.g., in $A \rightarrow B \rightarrow C \rightarrow D$, computation for strength of $B \rightarrow C$ would be affected by the strength of $A \rightarrow B$, but not by $C \rightarrow D$). Since all the associations are independent of each other, the measurements of path strengths are also independent. Therefore, if a link fails to meet the strength threshold, we can drop the link and all the possible paths containing the link. Although the worst-case time complexity is not reduced by this pruning process, in practice the computation time is largely reduced.

2) Path Mining as an Extension of Association Rule Mining: When using the global approach to measure the strength of association, the path mining algorithm can be viewed as an extension of traditional association rule mining. Path mining differs from traditional association rule mining in that a path has more than one association involved, and we need to check and maintain the strengths of all the associations in the path (such as confidence and IS). Although path mining provides more information, the computation cost is the same as traditional association rule mining. Based on how we define the association strength, the computation of confidence and correlation is only affected by preceding links in the paths. Therefore, as the path grows, we only need to compute the strength of newly added links, which makes the complexity equivalent to conventional association rule mining using the Apriori algorithm [1]. Moreover, since the strength of existing associations is fixed when a new association is added to the path, then similar to the pruning

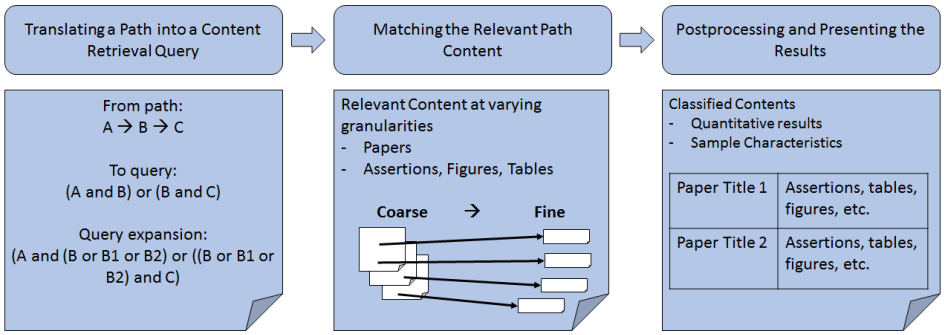

Figure 5. Process flow of path content retrieval

steps taken with the local strength measure, if a link fails to meet the strength threshold, we can drop the link and all the possible paths containing the link.

D. Ranking Path Relevance via Association Strength

When multiple paths exist, one goal is to determine which path is most relevant to the query. The association strength can be a good indicator of path relevance. So far, however, all the measurements focus on individual associations in the path. We still lack a uniform measure that we can use to evaluate the relevance of the entire path.

In our algorithm we take a heuristic approach to ranking the paths by comparing the "weakest link" of each path. For example, a path $A \rightarrow (0.4)B \rightarrow (0.4)C$ would be ranked higher than a path $A' \rightarrow (0.6)B' \rightarrow (0.3)C'$ (numbers in the parentheses indicate the strength of the link) because the weakest link in the first path (0.4) is stronger than the weakest link in the second one (0.3). This approach guarantees that higher ranked paths have reasonably high strength in all the links. In addition, this ranking approach can be exploited for pruning in the path discovery process. Consider the case for finding the top K paths satisfying a query. For any path containing a link whose strength is weaker than the weakest links of K existing paths, then no paths involving the link could be included in the result, and thus the path can be pruned. If we sort the associations by their strength before performing the search, then we can prune all the links with lower strength as well.

IV. Path Content Retrieval

In the previous sections we discussed our approach to identifying the paths and measuring path relevance via association strengths. After obtaining a list of paths, the next challenge is to study the paths to obtain more detailed knowledge about the interactions among the concepts. Since our paths are derived using text mining over a corpus of scientific literature, the relevant content from the literature is useful in studying these interactions. We refer to relevant document elements with knowledge about a path as "path content" and the process of searching for path content as "path content retrieval." Figure 5 presents the process of path content retrieval.

Compared to traditional information retrieval, path content retrieval poses many new challenges. First, we need to translate a path to a query so that it is digestible for an information retrieval system to find the relevant content describing relations between concepts. Second, the retrieved content should be in fine granularity so that specific information about the relations can be revealed. Third, according to the different demands of the research field, specific types of results are required in path content retrieval — such as quantitative experimental results or experiment sample characteristics.

A. Query Processing for Path Content Retrieval

The first step in path content retrieval is to translate a given path into a query that is digestible for an information system to find the relevant content. This can be accomplished by applying Boolean operators (e.g., AND, OR) among concepts in the path to form the basic query which reveals their relations. For example, a path "schizophrenia → working memory → prefrontal cortex (PFC) → dopamine" will be translated to a query "(schizophrenia AND working memory) OR (working memory AND PFC) OR (PFC AND dopamine)." Along with this translation, the lexicon can be exploited to perform query expansion. Query expansion can be done by matching a concept with its synonyms, as defined in the lexicon, or alternatively by using the hierarchical structuring of concepts in the lexicon to match concepts with more general parent concepts or more specific sub-concepts.

B. Finding Relevant Path Content

After translating a path into a query of concepts for content retrieval, we can utilize the document index which records the occurrences of concepts in documents (Section II-C) and standard information retrieval methods to retrieve the most relevant content. In path content retrieval, we want to retrieve document elements that include relations between concepts, such as sentences describing such a

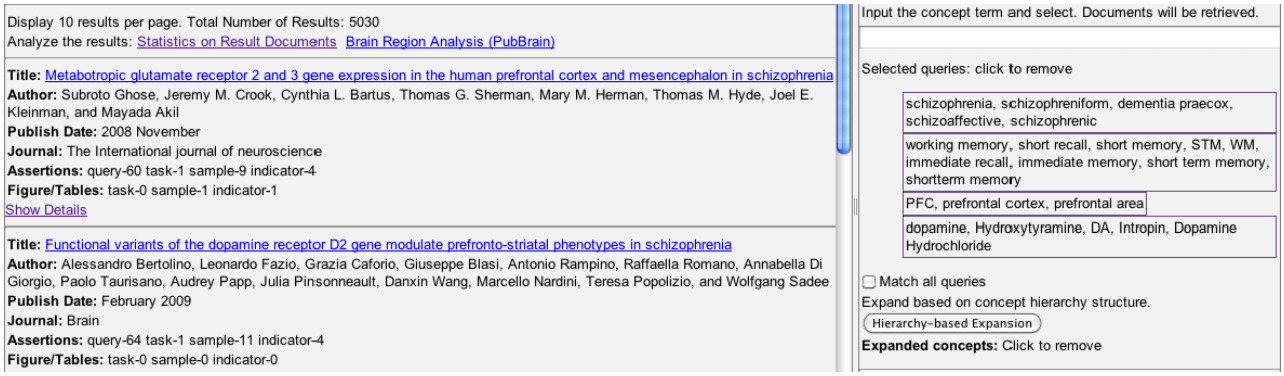

Figure 6. Example implementation of path content retrieval: user interface of the Document Content Explorer. User specifies the query in the "input panel" on the right and the relevant papers are displayed in the "results panel" on the left. The query shown includes four concepts: schizophrenia, working memory, prefrontal cortex (PFC) and dopamine, translated from the path "schizophrenia → working memory → PFC → dopamine."

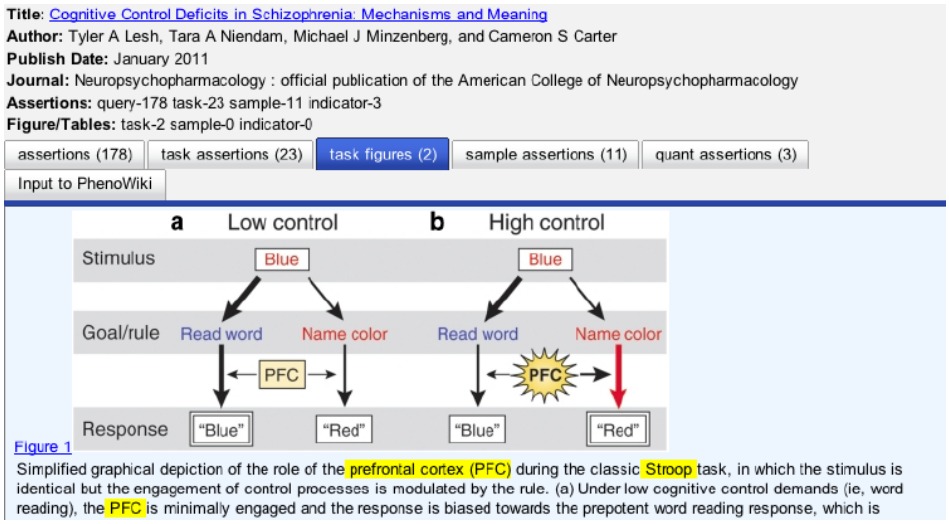

Figure 7. Example implementation of path content retrieval: detailed view of the Document Content Explorer. The paper "Cognitive Control Deficits in Schizophrenia: Mechanisms and Meaning" [7] has been retrieved with the path query "schizophrenia → working memory → PFC → dopamine." Fine-granularity content in the paper is classified into different categories such as task description, sample characteristics and quantitative results. Content of different categories is presented in different tabs. The selected tab (task figures) shows figures in the paper that include task descriptions. Both the task descriptor keywords ("Stroop") and the query keywords ("prefrontal cortex, PFC") are highlighted.

relationship, tables presenting experimental results explaining the correlation, or figures illustrating the interactions between concepts.

Fine-granularity path content, e.g., sentences and tables, better assists researchers in determining relations between concepts than coarse-granularity content, e.g., entire documents. However, due to the short length of fine-granularity content, the number of hits of concepts in fine-granularity document elements is usually very low, and these elements tend to have similar content frequencies. Therefore, it is difficult to retrieve enough fine-granularity content as well as to rank it. To remedy this problem, we take a two-step approach to finding relevant path content. We first retrieve and rank coarse-granularity content such as papers; then for each coarse-granularity element, we match fine granularity content such as sentences and tables, and return these as the path content. Since fine-granularity content is included within coarse-granularity content (e.g., a sentence is part of a paragraph, section, and so on), highly ranked coarse-granularity content most likely also contains relevant fine-granularity content. Such a two-step ranking scheme enables users to access the most relevant content describing the path.

Figure 6 presents an example interface for the first step in content matching, based on our Document Content Explorer developed for use in phenomics research (see Section V for more on the Document Content Explorer). Given a query, a list of relevant papers is returned. For each paper, basic information such as the title, authors, journal, and published date are displayed, along with basic occurrence statistics. By selecting each paper, its fine-granularity content is displayed

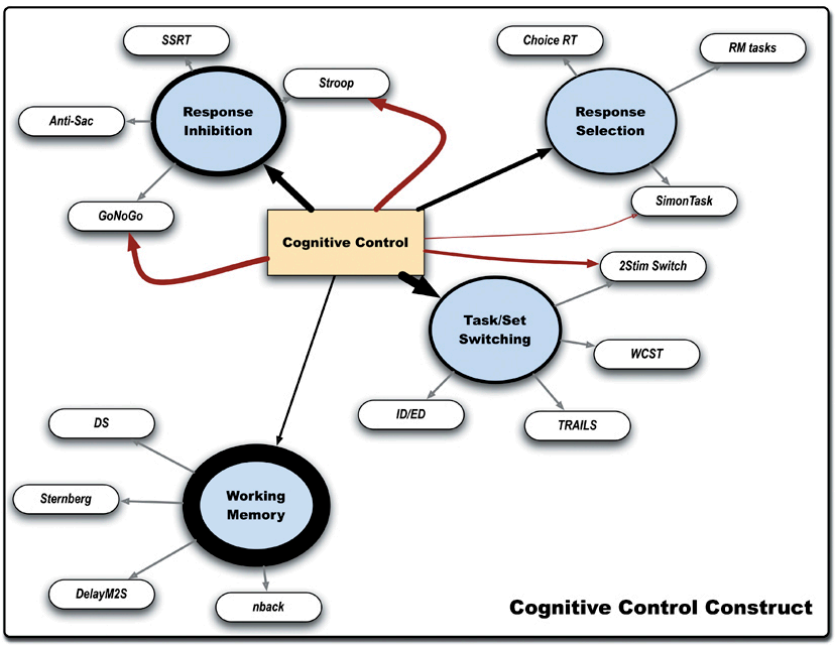

Figure 8. Components of the construct "cognitive control." This figure from [8] displays a graphical representation of the construct "cognitive control" as defined by the literature and expert review of behavioral tasks. The elongated ovals represent various cognitive tasks.

in the document details panel as shown in Figure 7.

C. Results Classification for Specific Research Goals

Based on different research goals, users may be interested in different types of content. For example, in phenomics, researchers are typically interested in quantitative experimental results (phenotype measurements) and experiment descriptions such as sample characteristics. To satisfy the demand of different users, we further classify our results and filter them according to different research goals. The results are classified using the category information from the concept lexicon. For each document element, we create a histogram vector by aggregating the count of concepts for different concept categories. The histogram of concept categories can be viewed as a feature that indicates the focus of the document element. In our lexicon there are several special categories introduced for content classification, such as sample characteristics, indicators, and sample species. We classify the content based on the majority category of its concepts. For example, in the sentence, "We tested WM [working memory] in infants at 6.5 and 9 months of age in a task that challenged them to remember the location of social and non-social targets." WM is a cognitive concept, and infants and months of age are concepts related to sample characteristics. In this case, the majority of the concepts are in the sample characteristics category, so the content is classified as sample characteristics.

In the document detailed view of Document Content Explorer (Figure 7), we can observe that the results are classified into different categories and are displayed by the corresponding tabs to permit users to choose content of interest. In each tab, results are broken down by sections and kept in the same order as in the original document, so that users can read them just as when reading the original paper. The detailed view of a paper provides a quick summary that permits users to quickly grasp the relevance of the results. In the list of papers returned for a query, the numbers of results classified into different categories in the paper are also presented; this helps users select the papers relevant to their interests.

V. PhenoMining: An Example Application of Path Mining

In this section we will present an example of using path knowledge discovery for knowledge discovery in phenomics. More specifically, we plan to answer the question of heritability for cognitive control phenotypes which was previously presented in [8]. We do this using tools developed to solve the path knowledge discovery problem in phenomics, called PhenoMining tools [9]. PhenoMining tools are able to identify associations among concepts in a multilevel phenotype lexicon in order to construct a path based on their co-occurrence in the corpus, and provide a quantitative way to measure the strength of associations. PhenoMining tools also provide a Document Content Explorer, which finds relevant published information for a specific path at fine granularity, so as to explain the interrelations (see Figures 6 and 7 for example displays from the Document Content Explorer).

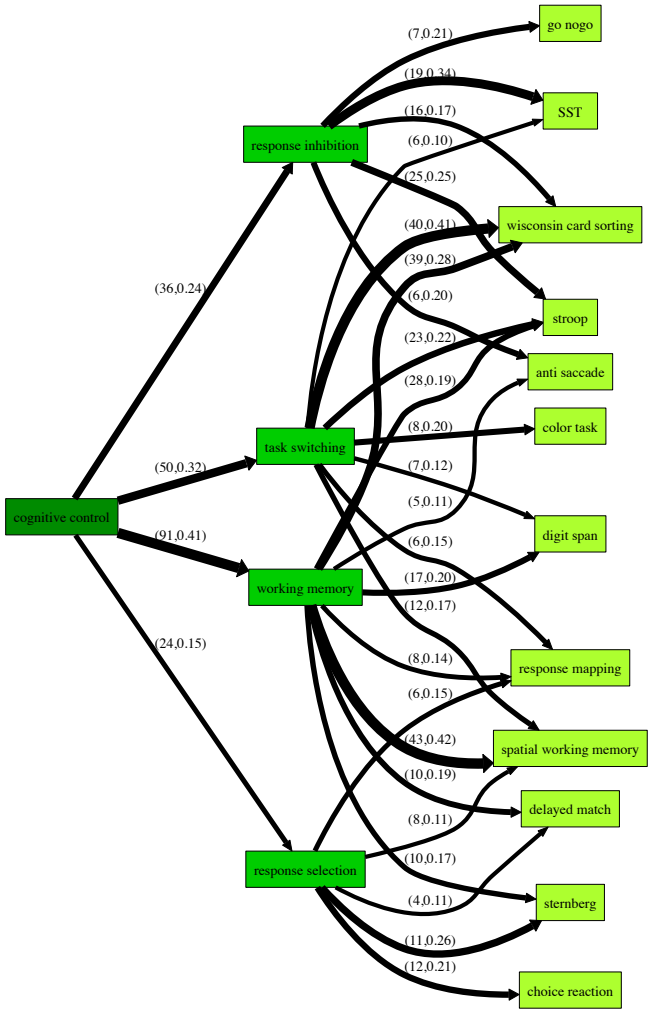

Figure 9. A PhenoGraph generated from path query "cognitive control → subprocesses → cognitive tasks". The strength of associations are computed based on the local strength measure. The numbers next to the links in the graph show the support (in absolute value of co-occurrence) and correlation scores of associations represented by the corresponding link. The thickness of links is proportional to their correlation score.

"Cognitive control" is a complex process that involves different phenotype components. Deficits in cognitive control are apparent in many neuropsychiatric disorders with strong genetic components. Different behavioral tasks are used for measuring the performance of those components with specific indicators. Knowing whether these components are also under strong genetic control is important for neuropsychiatric research. As an example, "working memory" is a latent component of cognitive control associated with schizophrenia and bipolar disorder [8]. The "n-back task" is a behavioral task measuring a person's working memory

Table I
SUBPROCESSES AND THEIR CORRESPONDING COGNITIVE TASKS.

Latent Subprocess	Cognitive Tasks
Response inhibition	GoNoGo, SST (SSRT), Stroop, Anti Saccade (Anti-Sac)
Task switching	Wisconsin Card Sorting (WCST), Color Task, ID/ED
Working Memory	digit span (DS), spatial working memory, Delayed Match (delayed M2S)
Response Selection	Response Mapping (RM Tasks), Sternberg, Choice Reaction (Choice RT)

Note: The matching is based on the IS score of the associations between subprocesses and cognitive tasks. The association with the highest IS score for each task is selected. The names in the parentheses are the names of the tasks as they appeared in [8].

performance. One important indicator for the n-back task is accuracy. The heritability of cognitive control is associated with the heritability of the indicators of behavioral tasks (e.g., the heritability for accuracy in the n-back task). Formalizing the nature of cognitive control requires studying relations among cognitive control, its subprocesses, and phenotypes such as heritability scores and indicators of behavioral tasks. This can be viewed as a path knowledge discovery problem. With the pattern "cognitive control → subprocess → task → indicator" we can gather known results about cognitive control. The results of path knowledge discovery provide a basis for interdisciplinary analysis of the heritability of cognitive control.

Therefore, we can view the problem of path knowledge discovery as a three-step process. First, we complete a query schema to operationally define the construct of cognitive control by identifying candidate components, tasks, and indicators that exist in the literature (such as those in Figure 8). Then, we use path queries to obtain quantitative heritability results for the task indicators in the corpus. Finally, we explore this content and extract discoveries about the heritability of cognitive control.

A. Building the Infrastructure for PhenoMining

Prior to performing path knowledge discovery, we first created a lexicon from the concepts covered in PhenoWiki [8]. The lexicon consisted of concepts from four levels — latent complex constructs, latent processes, cognitive tasks and indicators. The corpus was then collected according to the domain of the lexicon. The 9000 papers used for the corpus were retrieved from PubMed Central using the search query *((Schizophrenia OR Bipolar Disorder OR Attention Deficit Disorder) OR (Working Memory OR Response Inhibition)) AND (Stop-Signal Task OR Go NoGo Task OR Spatial Capacity Task OR Digit Span Task OR Probabilistic Reversal Learning Task OR Spatial Manipulation Task OR Stroop Task)*. This query, designed by domain experts, includes important concepts at different levels so as to cover interactions among concepts and facilitate path knowledge discovery.

B. Validating the Path Mining Methodology

We shall use the path "cognitive control → subprocess → cognitive task" to determine if *cognitive control* is heritable from publications about the subprocesses *response inhibition*, *response selection*, *task switching*, and *working memory*. Figure 9 shows the resulting top paths returned by a path query of three layers that included cognitive control as the first layer, concepts in latent processes as the second layer, and cognitive tasks as the third layer. We refer to a graph representing the set of search paths returned from a path query as a PhenoGraph. This PhenoGraph demonstrates the matching between tasks and the subprocesses of cognitive control that the tasks measure. In the figure, each task has relations with multiple subprocesses. For clarity in the presentation, we choose a higher threshold to generate the PhenoGraph. Table I presents the results obtained by choosing the most highly correlated subprocesses for each task, which are derived using a lower threshold and thus match more tasks (e.g., ID/ED task). Compared to Figure 8, which was created by domain experts, the mining tool achieved very promising results; 12 out of 15 tasks are correctly associated with their corresponding subprocesses. Compared to the results from domain experts (as shown in Figure 8), Figure 9 includes extra links since some tasks match multiple subprocesses. False positives exist because the co-occurrence of tasks and subprocesses in a document element (using paragraph granularity in this example) does not necessarily indicate that the subprocesses are measured by the task. Also, it is entirely possible that two subprocesses are discussed in the same document element, and our system is unable to separate them. On the other hand, some tasks are not included in the top paths because the occurrences of those tasks in the corpus is so low that the correlation with subprocesses is too low to be included in the results. By setting the threshold lower, the missing tasks may appear, but this may also introduce more noise.

Even more important than this direct analysis of the PhenoGraph is the analysis of the path content. Using the Document Content Explorer, researchers have the ability to quickly scan the papers associated with each path, with relevant fine-granularity content such as sentences, tables, and figures extracted and classified into categories such as "task descriptions," "sample characteristics," and "quantitative results." Researchers can then assess the validity of these knowledge paths and begin the process of digesting the data from the path content into the format needed for their final analysis. Overall, using the PhenoMining tools, the time spent on collecting the relevant literature and deriving the knowledge structure is greatly reduced, and the results from the tools are comparable to human-derived results.

VI. Related Work

Traditional association rule mining studies [1], [6] have focused on finding recurring patterns. As described in [5], the association rules discovered are primarily intended to identify rules such as, "a customer purchasing item A is likely to also purchase item B." Extending this approach to the bioinformatics field, association rule mining has been used to profile gene expression [10] and study protein-protein interaction [11]. These studies focus on the discovery of individual associations.

In [4] Tan et al. studied indirect associations, which are a special type of association rule describing associations $A \rightarrow B \rightarrow C$: "A customer purchasing item A is likely to also purchase item $B_i \in B$, and a customer purchasing item B_i is likely to also purchase item C," where $i = 1, 2, \ldots, n$. By introducing the intermediate item sets B, the rules reveal a "higher-order" (indirect) data dependency between A and C. This higher-order dependency is similar to the idea of path in our work. However, there are some major differences in path knowledge discovery. First, the goal of mining is different. The high-order dependency focuses on identifying pairs of indirectly related item sets connected by an intermediate item set. Our path mining not only identifies such indirect relations, but also requires that the intermediate relations satisfy a certain pattern specified in the path query. Second, our path mining is closely integrated with content retrieval. Instead of only identifying relations, our path knowledge discovery process also provides relevant content describing such relations.

Association analysis involving intermediate concepts has been applied in bioinformatics. Baker et al. [12] developed a method for mining connections between chemicals, proteins and diseases using the biomedical literature as a knowledge source. Voytek et al. [13] developed a semi-automatic way to extract the "cognome" — relationships between brain structure, function and disease. Both works essentially followed the model that "if A is related to B, and B is related to C, then A is likely to be related with C." These authors empirically evaluated their results by comparing them with human-generated ones. However they did not employ quantitative measurements in these relations, or extend their methods to an association with more than three concepts. Our work presents a methodology to evaluate sequences of associations and discover path associations with a multilevel lexicon from a large text corpus. Moreover, the introduction of wildcard concept levels greatly increases the path discovery scope and can lead to new hypotheses for further research.

There are also other literature-based discovery tools based on association rule mining. BITOLOA [14] is a tool mining the association pattern $X \rightarrow Y \rightarrow Z$ when two of the three concepts are specified by the user (e.g., the user may specify X and Y or X and Z). Arrowsmith [15] is a tool that finds the links between two separate sets of documents via common title words and phrases. Both of these tools are based on patterns of pairwise associations between three concept sets. By contrast, our tools provide not only the

ability to mine more complex path patterns, but also the ability to retrieve relevant path content.

VII. Conclusion

Path knowledge discovery consists of two integral parts—path discovery and path content retrieval—and focuses on the study of relations among concepts at multiple levels. Path discovery identifies and measures a path of knowledge, i.e., a sequence of associations among concepts across multiple domains, with these associations measured using an extension of the support, confidence, and correlation measures from traditional association rule mining. Path content retrieval takes those paths discovered by using path discovery and uses them to retrieve relevant content at various levels of granularity from the corpus. This path content reveals the semantics of the relations represented by those paths and provides a basis for deeper analysis by domain experts.

Manually conducting such interdisciplinary analysis requires effort, even from domain experts. This effort can be exceedingly time-prohibitive, especially as the literature base grows. Utilizing our path knowledge discovery process, initial search results can be compiled automatically, and path content can be retrieved for users. And while it may be that the actual derivation of final results cannot be automated, by employing our method, the whole process can be greatly accelerated. Whereas it takes our method seconds to retrieve the content and minutes for a user to browse and select what is relevant, the traditional manual approach may take several orders of magnitude longer to execute the same steps, and becomes infeasible when the number of papers to examine becomes too large. This typically results in severe reductionist approaches by domain experts when trying to identify a significant but manageable subset of the literature. Our method eliminates the need for drastic a priori approaches to reduce the scope of literature for review. Thus, with the aid of mining tools, the scope of research can be enlarged into a corpus of thousands of papers instead of the 150 papers used in [8]. And since our method scales well with an increasing corpus size, further research is desirable to extend and integrate different archives beyond PubMed Central to broaden this corpus. Human intelligence still plays an important role in this process, as selecting the best paths and content are quite subjective, but it is clearly beneficial to use text mining and information retrieval techniques to replace the mechanical aspects and speed up the process.

Acknowledgment

This work was supported by USPHS grants, including the NIH Roadmap Initiative Consortium for Neuropsychiatric Phenomics and including linked awards UL1DE019580, RL1LM009833. We thank Jianming He, Ying Wang, Jiacheng Yang, Jianwen Zhou, Xiuming Chen and Jiajun Lu of the CoBase research group in the UCLA Computer Science Department for their work on the PhenoMining tools and PhenoWiki+ implementation. We would also like to thank Professors Robert Bilder, Carrie Bearden, and Joseph Ventura from the Consortium for Neuropsychiatric Phenomics for testing of the tools and for their stimulating discussions during the development of this work.

References

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Intl. Conf. on Very Large Data Bases*, 1994, pp. 487–499.

[2] R. Bilder and et al., "Cognitive ontologies for neuropsychiatric phenomics research," *Cogn. Neuropsychiatry*, vol. 14, no. 4-5, pp. 419–450, 2009.

[3] "Pubmed central website," www.ncbi.nlm.nih.gov/pmc/.

[4] P.-N. Tan, V. Kumar, and J. Srivastava, "Indirect association: mining higher order dependencies in data," in *Princ. Data Min. Knowl. Discov.*, 2000, pp. 632–637.

[5] C. Silverstein, S. Brin, and R. Motwani, "Beyond market baskets: generalizing association rules to dependence rules," *Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 39–68, Jan. 1998.

[6] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proc. 2000 ACM SIGMOD Intl. Conf. on Mngmt. of Data*, pp. 1–12.

[7] T. Lesh, T. Niendam, M. Minzenberg, and C. Carter, "Cognitive control deficits in schizophrenia: mechanisms and meaning," *Neuropsychopharmacology*, vol. 36, no. 1, 2010.

[8] F. Sabb and et al., "A collaborative knowledge base for cognitive phenomics," *Mol. Psychiatry*, vol. 13, no. 4, 2008.

[9] "Phenomining tools," http://phenominingbeta.cs.ucla.edu/.

[10] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.

[11] T. Oyama, K. Kitano, K. Satou, and T. Ito, "Extraction of knowledge on protein–protein interaction by association rule discovery," *Bioinformatics*, vol. 18, no. 5, pp. 705–714, 2002.

[12] N. Baker and B. Hemminger, "Mining connections between chemicals, proteins, and diseases extracted from medline annotations," *J. Biomed. Informatics*, vol. 43, no. 4, 2010.

[13] J. Voytek and B. Voytek, "Automated cognome construction and semi-automated hypothesis generation," *J. Neurosci. Methods*, 2012.

[14] D. Hristovski, C. Friedman, T. Rindflesch, and B. Peterlin, "Exploiting semantic relations for literature-based discovery," in *AMIA Ann. Symp. Proc.*, vol. 2006, p. 349.

[15] N. Smalheiser, V. Torvik, and W. Zhou, "Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in medline," *Comput. Methods Programs Biomed.*, vol. 94, no. 2, 2009.