Exploiting Ontology Structure and Patterns of Annotation to Mine Significant Associations between Pairs of CV Terms

[RESEARCH PAPER SUBMISSION]

Woei-Jyh Lee¹, Louiqa Raschid¹, Hassan Sayyadi¹, and Padmini Srinivasan²

¹ University of Maryland, College Park, MD 20742, USA
{adamlee,louiqa}@umiacs.umd.edu, sayyadi@cs.umd.edu
² The University of Iowa, Iowa City, IA 52242, USA
padmini-srinivasan@uiowa.edu

Abstract. There is significant knowledge captured through annotations on the life sciences Web. In past research, we developed a methodology of support and confidence metrics from association rule mining, to mine the association bridge (of termlinks) between pairs of controlled vocabulary (CV) terms across two ontologies. Our (naive) approach did not exploit the following: implicit knowledge captured via the hierarchical is-a structure of ontologies, and patterns of annotation in datasets that may impact the distribution of parent/child or sibling CV terms. In this research, we consider this knowledge. We aggregate termlinks over the siblings of a parent CV term and use them as additional evidence to boost support and confidence scores in the associations of the parent CV term. A weight factor (α) reflects the contribution from the child CV terms; its value can be varied to reflect a variance of confidence values among the sibling CV terms of some parent CV term. We illustrate the benefits of exploiting this knowledge through experimental evaluation.

Key words: annotation, CV terms, generalized association rule mining, support and confidence, life sciences link (LSLink)

1 Introduction

The biomedical enterprise has generated an abundance of data that is captured using annotated and hyperlinked records in the life sciences Web. Records in each resource are typically annotated with terms from controlled vocabularies (CVs) or ontologies, forming a rich Web of knowledge. Consider a simplified Web of three publicly accessible resources Entrez Gene [1], OMIM [2] and PubMed [3], in Figure 1. Data records in each resource are annotated with terms from multiple CVs. The hyperlinks between data records in any two resources form a relationship between the two resources, represented by a (virtual) link. Thus, a record in OMIM, annotated with SNOMED terms [4] has multiple links to gene records in Entrez Gene, annotated with GO terms [5]; gene records further have hyperlinks to multiple records in PubMed annotated with MeSH terms [6].
A background LSLink (Life Sciences Link) dataset composed of termlinks is generated after executing a protocol to follow hyperlinks and to extract annotations; details are provided in [7]. Example hyperlinked records are shown in Figure 2. Each termlink (to be defined in the next section) associates a pair of CV terms, and contributes to an association bridge across two CVs or ontologies. A user dataset is a subset of the background dataset. In prior research, we mined the association bridge of termlinks of user datasets, to discover potentially new knowledge that is both meaningful and not well known a priori. Using support and confidence metrics, we can rank the pairs of associations of CV terms and identify potentially significant pairs. User validation confirmed that a majority of highly ranked pairs were meaningful. Several pairs were unknown and might lead to actionable knowledge [7].

There are two limitations of our prior research. First, while mining the association bridge of termlinks between pairs of CV terms, we treated each CV term (of the CV or ontology) independently. For example, is-a is a key relationship that exists amongst terms of a single vocabulary. Intuitively, termlink evidence existing for a child CV term could influence the confidence and support scores
of the parent CV term. By mining the termlinks of the child and parent CV
terms independently, we may be ignoring this potential contribution from the
structure of the ontologies.

The second limitation is that we did not consider any patterns of annotation
in a dataset of termlinks. Suppose we consider a user dataset of an OMIM record
conceptually linked to a set of Entrez Gene records. Such a set of gene records
have some biological affinity since they are all associated with the disease in the
OMIM record. Our analysis of such sets of gene records and the corresponding
datasets of termlinks indicates that patterns of annotation do exist. One such
pattern is an increase in the frequency of annotation using sibling CV terms.

This research will exploit both sources of knowledge, i.e., the is-a structure of
ontologies and the pattern of annotations. We aggregate the termlinks associated
with a parent CV term, so as to use this evidence to potentially boost the values
for confidence and support scores in associations of the parent CV term. A weight
factor ($\alpha$) determines the relative weight of evidence or the contribution from the
child CV terms. The value of $\alpha$ can also reflect a variance of confidence scores of
the sibling CV terms of some parent CV term, e.g., a high variance can reduce
the contribution from child terms.

Using a background dataset of OMIM, Entrez Gene and PubMed, and user
datasets that have observable patterns of annotation, we demonstrate the benefits of our research. One interesting result is that we find potentially significant associations involving parent GO terms, where this term does not occur among the termlinks, i.e., it was not used for annotation or it was not explicitly linked to the partner MeSH term. This suggests that our method can identify implicit annotations and there is scope to generate new knowledge as we identify such new associations.

The paper is organized as follows: Section 2 defines a background LSLink
dataset, termlinks and support and confidence. Section 3 motivates the potential
benefits of aggregation over CV terms. Section 4 presents a methodology for
determining aggregate confidence and support scores. We discuss the limitations
of a simple solution (1-step link aggregation) and the features of a more robust
solution (2-step score-score aggregation). Section 5 presents the related work and
Section 6 illustrates evaluation results. Section 7 offers our conclusions.

## 2 Prior Results on LSLINK Association Mining

### 2.1 Support and Confidence for LSLINK Mining

A background LSLink dataset is associated with a specific experiment protocol
to gather a representative sample of data records, hyperlinks and annotations. Figure 2 illustrates three sample hyperlinks between two Entrez Gene and two
PubMed records. The hyperlinks are between records $e_1$ and $p_1$, $e_2$ and $p_1$, and
$e_2$ and $p_2$. The terms $g_a$, $g_b$, $g_c$ and $m_a$, $m_b$, $m_c$, $m_d$ annotate these records. Each
record is associated with two terms. If we consider the hyperlink between $e_1$ and
$p_1$, the two CV terms $g_a$ and $g_b$ annotating $e_1$, and the two CV terms $m_a$ and
An example termlink is the following: \((g_u, m_w, e_2, p_2) = (\text{DNA repair}, \text{Mitosis}, 675, 10749118)\). These three hyperlinks from Figure 2 generate 12 termlinks. Note that both hyperlinked data records must be annotated in order to generate a termlink.

The set of termlinks represents a bridge of associations between pairs of CV terms across two CVs or ontologies. We apply support and confidence metrics \(^3\) from association rule mining \(^8\), \(^9\) to identify significant pairs of associations among CV terms. The metrics reflect the extent to which the association between a pair of CV terms deviates from one resulting from chance alone (a random association). Support and confidence is defined as follows:

\[-(G, M, E, P)\] is the background dataset of genes from Entrez Gene \((E)\) annotated with GO terms \((G)\) with links to PubMed records \((P)\) that are in turn annotated by MeSH terms \((M)\). Termlinks are derived from this dataset.

\(#(G, M, E, P)\) is the cardinality of the termlinks in \((G, M, E, P)\).

\((G, M, E', P')\) and \(#(G, M, E', P')\) correspond to the user dataset, a subset of the background dataset.

\(#(g_u \land m_w, E, P)\) is the cardinality of termlinks containing the pair of terms \(g_u\) and \(m_w\) in the background dataset. \(#(g_u \land m_w, E', P')\) is the corresponding value in a user dataset.

\(#(g_u \lor m_w, E, P)\) is the cardinality of termlinks containing either term \(g_u\) or term \(m_w\) in the background dataset. \(#(g_u \lor m_w, E', P')\) is the corresponding value in a user dataset.

Term probability\(^4\) reflects how commonly a CV term is used to annotate a data record, as follows:

\[
Pr_{\text{term}}(g_u, E) = \frac{\text{number of annotations that are } g_u \text{ in } E}{\text{total number of annotations in } E} \quad (1a)
\]

\[
Pr_{\text{term}}(m_w, P) = \frac{\text{number of annotations that are } m_w \text{ in } P}{\text{total number of annotations in } P} \quad (1b)
\]

The link annotation probability \((Pr_{\text{link}})\), and the conditional probability \((Pr_{\text{cond}})\), for the pair \((g_u, m_w)\) estimated from the user query dataset is as follows:

\[
Pr_{\text{link}}(g_u, m_w, E', P') = \frac{\#(g_u \land m_w, E', P')}{\#(G, M, E', P')} \quad (2a)
\]

\[
Pr_{\text{cond}}(g_u, m_w, E', P') = \frac{\#(g_u \land m_w, E', P')}{\#(g_u \lor m_w, E', P')} \quad (2b)
\]

\(^3\) We note that given the universe of CV terms and annotations, data records and hyperlinks between data records, and termlinks, there were many possible approaches to obtain expressions for support and confidence. We used our judgment to pick some reasonable choices.

\(^4\) We calculate term probability at the annotation level as follows and note that it can also be calculated at the data record level.
Finally, we define support and confidence. We incorporate a term-freq correction factor from Equations (1a) and (1b) for the background dataset as follows:

\[
Supp(g_u, m_w, E', P') = \log \left( \frac{Pr_{\text{link}}(g_u, m_w, E', P')}{Pr_{\text{term}}(g_u, E)Pr_{\text{term}}(m_w, P)} \right) \quad (3a)
\]

\[
Conf(g_u, m_w, E', P') = \log \left( \frac{Pr_{\text{cond}}(g_u, m_w, E', P')}{Pr_{\text{term}}(g_u, E)Pr_{\text{term}}(m_w, P)} \right) \quad (3b)
\]

We note that when presenting our methodology to aggregate termlinks and to calculate the boosted confidence and support in the associations of the parent CV terms, we use the simpler expressions for confidence and support from Equations (2a) and (2b). We do not apply the term-freq correction factor of the expressions of Equations (3a) and (3b).

### 2.2 Results from Mining

There can be a potentially large number of associations of pairs of CV terms, pairs even for a single gene. For example, for a user dataset defined for the human gene TP53 [7], there were 986,612 termlinks and they represented 83,116 distinct associations between pairs of GO and MeSH terms! The support and confidence metrics were used to rank these pairs of associations and identify the Top 25 potentially significant pairs for each gene. Experts rated the associations of pairs of CV terms along the following independent dimensions: (Meaningful, Maybe Meaningful, Not Meaningful), and (Widely Known, Somewhat Known, Unknown/Surprising). A majority of the Top 25 pairs of associations for user datasets such as BREAST CANCER, CFTR, TP53, etc., were identified as a true positive. Several of the pairs were unknown and might lead to new knowledge. For example, for BREAST CANCER, the previously unknown association of the GO term negative regulation of centriole replication with the MeSH term Fallopian Tube Neoplasms might be interesting, because it indicates that the tumor and the negative regulation might have a causal relationship [10]. The background dataset of termlinks from this study and the associations among pairs of GO and MeSH terms are available at the following site: http://www.cbcb.umd.edu/research/lslink/lodgui/

### 3 Motivation for Aggregation

We first illustrate the potential benefit of exploiting structural knowledge of is-a hierarchies and then discuss patterns of annotation.

The first set of examples are from termlinks generated from a user dataset of the human gene TP53 in Entrez Gene, PubMed records that are hyperlinked to it, and the corresponding annotations. Consider the GO and MeSH is-a hierarchies of Figure 3. In Figure 3(a), a termlink (negative regulation of progression through cell cycle, Cyclin-Dependent Kinases, 7157,
Fig. 3. Example parent-and-child hierarchies in GO and MeSH (each dotted line show an actual association generated in the human gene TP53 user dataset)

17612495) occurs between the parent GO term and the parent MeSH term. In addition, two termlinks (cell cycle arrest, CDC2-CDC28 Kinases, 7157, 14640983) and (cell cycle arrest, Cyclin-Dependent Kinase 2, 7157, 17371838) occur between the child terms. These latter two termlinks are evidence to boost the association between the pair of parent terms.

In Figure 3(b), the termlink (protein binding, Tosylphenylalanyl Chloromethyl Ketone, 7157, 12821135) occurs between the parent GO term protein binding and a child MeSH term Tosylphenylalanyl Chloromethyl Ketone. In addition, there are two termlinks from the parent MeSH term to two child GO terms. Note that there is no termlink between the two parent CV terms, protein binding and Amino Acid Chloromethyl Ketones in the termlink dataset; this is represented by a broken link between the pair of terms in the association bridge. However, the 3 termlinks in this Figure can be considered evidence to introduce a new association between the parent GO term protein binding and the parent MeSH term Amino Acid Chloromethyl Ketones.

To summarize, Figure 3 presented two examples of termlinks associated with combinations of parent/child CV terms. It seems intuitively apparent that the termlink evidence attached for example to the child GO terms should influence the evidence of the parent GO terms. By treating these termlinks as strictly independent, we may be ignoring potentially valuable information offered by the structure of the GO ontology. Note that this applies to each participating ontology involved in generating termlink, in this case GO and MeSH. Thus, analogously from the perspective of the MeSH hierarchy, parent MeSH terms may benefit from the termlink evidence of their child MeSH terms. Finally, new associations between pairs of parent CV terms may also be introduced, where the parent CV term was not used for annotation.

Note that in our experiments, we only exploit a limited amount of knowledge. For example, we limit aggregation of termlink evidence along the GO is-a hierarchy alone, and we only consider aggregation from the GO CV terms to the immediate parent CV term, in the next higher level of the GO hierarchy.

Next, we illustrate a pattern of annotation that results in a higher frequency of annotations that use sibling terms from the GO ontology. We note that there
is a similar pattern of higher frequency of annotation of parent and child terms, and that these patterns are also observed in individual Entrez Gene record annotations. For lack of space, we do not provide evidence on all such patterns.

We consider a dataset of term links obtained from OMIM records conceptually hyperlinked to (one or a set of) gene records in Entrez Gene. We note that these gene records are biologically linked since they are associated with the same disease in the OMIM record. As of September 6, 2007, there were 14,851 OMIM records. The distribution of Entrez Gene records conceptually linked to an OMIM record is given in Figure 4. While 14,502 OMIM records are linked to a single gene, 193 records have links to two genes, and SCHIZOPHRENIA (MIM Number 181500) links to 22 genes.

To illustrate the annotation pattern, we compare two techniques to group pairs of gene records to create user datasets. For the first method (OM linked), we place a pair of genes in a user dataset only if both genes are conceptually hyperlinked to the same OMIM record. Next, we generate a similar number of pairs for Random; here we pick a pair of human genes at random from Entrez Gene. For each pair in OM linked and Random, we extract the GO annotations. Each dataset contains 1,000 pairs of genes. Figure 5 shows the distribution of the number of sibling GO terms that annotate the pairs of genes from OM linked and Random.

We observe that pairs of genes in OM linked have a much higher distribution of sibling GO terms. For example, there are 1,618 occurrences of (pairs of) term links involving a pair of sibling GO terms, and 148 occurrences of (a triple of) term links involving a triple of sibling GO terms, in OM linked. In contrast, the 1,000 pairs of genes in Random only have 559 occurrences of pairs and 34 occurrences of triples of sibling GO terms. To validate the pattern of annotation, we generated the 1,000 pairs of OM linked genes and the 1,000 pairs of Random genes 3 times. The 3 OM linked datasets had a mean of 1,499 pairs of sibling GO terms and a mean of 196 triples of GO terms. The 3 Random datasets had a mean of 487 pairs of sibling GO terms and a mean of 41 triples of GO terms. To summarize, user dataset such as OM linked with pairs of genes with biological affinity reflect a pattern of annotation with a higher frequency of annotation using sibling GO terms.
4 Methodology for Aggregation

We consider boosting the support and confidence scores of associations of the parent CV terms using the evidence of the term links of child CV terms. We use the unboosted score for support or confidence as a baseline, $Supp_B$ or $Conf_B$, respectively. While we provide expressions for boosted confidence and support scores, for space limitations we only report on the values of the confidence scores in the rest of this paper.

We propose two solutions for aggregation. The simple solution, 1-step Link aggregation (1L), will aggregate the term links from the child to the parent and use a counting approach. This approach has two limitations. One is that the percentage contribution from the term links of the child CV term cannot be controlled. The second is that a variance of confidence among the sibling terms of the parent CV term cannot be factored in by the 1L simple counting approach. We then present a comprehensive solution, 2-step Score-Score (2SS), that obtains a weighted score for the parent CV term. The weighted score allows the contribution from the child CV terms to be controlled. The value of the weight $\alpha$ can reflect the variance of confidence of the sibling CV terms. For example, a high variance can increase the contribution from the child terms.

4.1 Simple Solution for Aggregation (1L)

Consider the example in Figure 6(a) where $g_1$ and $g_2$ are two sibling child terms of parent GO term $g_u$. There are 2 term links, one from GO term $g_u$, and another one from $g_2$, to the MeSH term $m_w$. The confidence scores for the parent $g_u$, or for the child $g_2$, paired with $m_w$, are $\frac{1}{4}$ and $\frac{1}{3}$, respectively.

The simple 1L counting based approach to boost the confidence score of the parent CV term $g_u$ will accumulate all term links associated with $g_2$ and credit

$^5$ before the term-freq correction
it to the parent term. The 1L expression for the boosted support and confidence scores for the parent term is as follows:

\[
\text{Supp}_{1L}(g, m, E', P') = \frac{\#(g \land m, E', P') + \#(g_i \land m, E', P'|g_i \in \text{Child}(g))}{\#(G, M, E', P')}
\]

\[
\text{Conf}_{1L}(g, m, E', P') = \frac{\#(g \land m, E', P') + \#(g_i \land m, E', P'|g_i \in \text{Child}(g))}{\#((g \lor g_i) \land m, E', P'|g_i \in \text{Child}(g))}
\]

In this example, the original confidence for the association between parent \(g_u\) and \(m_w\) was \(\frac{1}{4}\), and the boosted confidence score is \(\frac{1+1}{5} = \frac{2}{5}\).

4.2 Limitations of the Simple Solution

We present two cases that illustrate the limitation of the simple 1L counting approach. Consider the termlinks of Figure 6(b). The original confidence scores for the associations of \(g_u\), and \(g_2\), with \(m_w\), are \(\frac{1}{5}\) and \(\frac{2}{5}\), respectively. We note that these values are equal to the scores in Figure 6(a). Suppose that we use the simple counting 1L approach to boost the confidence score. The boosted value for confidence for the association between \(g_u\) and \(m_w\) will be \(\frac{1+2}{5} = \frac{3}{5}\).

We note that the boosted confidence score of \(\frac{3}{5}\) in Figure 6(b) between \(g_u\) and \(m_w\) is different from the boosted value of \(\frac{2}{5}\) of Figure 6(a). However, in both cases, the original confidence scores between \(g_u\) and \(m_w\), and between \(g_2\) and \(m_w\), are identical. This is the first limitation. Intuitively, we would like to control
the contribution made by termlinks from the child CV terms, so that in a case such as Figures 6(a) and (b), when the confidence of the child CV term is the same, then there is an identical contribution to the parent CV term. With the 1L approach, the contribution to the parent CV term is not controlled by the confidence of the child CV term but instead it is controlled by the number of termlinks that refer to the child CV terms.

We next consider the situation where there is a variance in the confidence of the associations of the sibling CV terms. In Figure 7(a), the confidence scores for the associations of each of child terms, \( g_1 \) or \( g_2 \), with \( m_w \), is \( \frac{3}{8} \), i.e., they are of equal confidence. In Figure 7(b), there is a variance of the confidence scores of the child terms. The confidence score of the association of \( g_1 \) with \( m_w \) is \( \frac{1}{8} \), while the confidence score in the association of \( g_2 \) with \( m_w \) is 5 times higher and is \( \frac{5}{8} \).

In both Figures 7(a) and (b), the original confidence score of the association of the parent \( g_u \) with \( m_w \) is \( \frac{1}{8} \). Using the 1L approach, the boosted confidence score for the association between \( g_u \) and \( m_w \) is also \( \frac{1 + 3 + 3}{8} = \frac{7}{8} \), in both cases. Ideally, when there is equal confidence in the associations of the sibling terms (as in Figure 7(a)), this may be considered strong evidence that these siblings should boost the confidence in the associations of the parent term. On the other hand, when there is a significant variance in the confidence of the sibling terms (as in Figure 7(b)), it is unclear if these siblings are providing strong evidence to boost the confidence in the parent term. Thus, referring to Figures 7(a) and (b), when there is no variance in the confidence scores of the siblings as in Figure 7(a), the boost to the parent should be greater.

### 4.3 Comprehensive Solution for Aggregation (2SS)

We present the 2SS aggregation method; it will overcome both limitations of the 1L approach. It will use a weight factor \( \alpha \) to control the contribution to the parent CV term using the confidence of the child CV terms. The value of \( \alpha \) will be determined based on the variance of the confidence of the sibling CV terms. The support and confidence scores presented in Equations (5a) and (5b) are prior to the application of the term-freq correction from our prior research.

For space limitations we do not report on the expressions and the experimental evaluation of the 2SS boosted scores after applying the term-freq correction; they show similar behavior.

\[
Supp_{2SS}(g, m, E', P') = (1 - \alpha) * Supp_B(g, m, E', P') + \alpha * \text{Avg}(Supp_B(g_i, m, E', P') | g_i \in Child(g))
\]

**Equation 5a**

\[
Conf_{2SS}(g, m, E', P') = (1 - \alpha) * Conf_B(g, m, E', P') + \alpha * \text{Avg}(Conf_B(g_i, m, E', P') | g_i \in Child(g))
\]

**Equation 5b**
We summarize the features of the 2SS solution. First, we calculate the confidence score for each of the child terms, and then we average the confidence scores over all the child terms. We then use a weighting factor \( \alpha \) to determine the actual contribution from the child terms that should be used to boost the confidence score of the parent. We experiment with the following simple rule-of-thumb: The value for \( \alpha \) will depend on the variance for the confidence scores for the child terms. To explain, if there is high variance in the confidence score for each of the child terms of some parent \( g_u \), then we will be less confident that we should aggregate over these child terms and use the child terms to potentially boost the confidence score in \( g_u \).

We note that based on the above expression, the boost to the parent \( g_u \) is greatest when the confidence score of each of the child terms is independently high, and when there is low variance in the confidence score of the child terms. The boost to \( g_u \) is low when either the confidence score in each of the child terms is low, or when there is a high variance in the confidence scores of all child terms of \( g_u \). The boosted confidence score (with \( \alpha = \frac{1}{2} \)) in Figure 7(a) is \( \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times (\frac{3}{8} + \frac{2}{8}) = \frac{1}{2} \). This value is higher compared to the boosted confidence score in Figure 7(b) which is \( \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times (\frac{1}{8} + \frac{5}{8}) = \frac{3}{16} \). Although the difference between these two boosted confidence values is \( \frac{1}{16} \), this difference can have a major impact on the rank of the associations.

5 Related Work

We consider related work in generalized association rule mining, ranking and ontology matching. Generalized association rule mining \([11][12][13]\) creates an extended transaction set either by replacing an item with a new item representing a generalized concept, or by aggregating both the original item and the generalized item. We note that the generalized concept does not occur in their original transaction set. Their solution approach is similar to our counting based 1L approach and faces the limitations that were discussed, i.e., controlling the contribution of child CV terms and reflecting variance of confidence. \([14]\) proposed to assign a lower threshold of support for associations in the lower levels of ontology. Furthermore, in order to reduce the search space by filtering associations contain independent items, the metric usefulness or interest is suggested by \([15]\). They define R-interesting as a rule is interesting iff it has no predecessor or its adjacent interesting predecessor is R-interest\([16]\).

While there is extensive literature on ranking using link structure of a graph, the focus is on ranking nodes in a general graph \([17][18][19]\). There is no work on ranking an association bridge (edges) of a bi-partite graph and ranking typically does not consider metadata such as the is-a hierarchy.

There is also research on ontology matching or ontology alignment \([20][21]\). The objective is to determine matches or correspondence between concepts or between subgraphs. Their solutions are based on string similarity between the labels of concepts and structural similarity; the latter considers relationship pat-
terms in the ontology. While this research also exploits similar knowledge, since the objectives are different, we cannot apply any of their solutions.

6 Experimental Evaluation

6.1 Generating User Datasets

Disease related user datasets were generated using the corresponding OMIM record. The protocol follows links from OMIM to Entrez Gene and then to PubMed. Table 1 reports on the statistics of four disease related datasets. For e.g., for the BREAST CANCER user dataset, the OMIM record has hyperlinks to 13 Entrez Gene records that are annotated with 147 distinct GO terms. Following the hyperlinks from these 13 Entrez Gene records to PubMed, we obtain 3,237 distinct PubMed records that are annotated with 2,463 distinct MeSH descriptor terms (of selected UMLS semantic types [22]). We generate 1,232,086 termlink instances and collect 124,342 distinct associations pairs of a GO term and a MeSH term. The one-level aggregation using the GO structured is-a hierarchy introduces 24 new GO terms and 18,648 pairs of associations that did not occur among the original termlinks.

6.2 Examples of Identifying Significant Associations via Aggregation

We use several user datasets to illustrate a range of opportunities to boost the associations of the parent CV terms. We note that all these examples have been verified to be meaningful and some are previously unknown.

Table 1. Statistics in four disease-related user datasets

<table>
<thead>
<tr>
<th>MIM Number</th>
<th>Title</th>
<th>114480</th>
<th>114500</th>
<th>176807</th>
<th>191170</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>114480</td>
<td>114500</td>
<td>176807</td>
<td>191170</td>
<td></td>
</tr>
<tr>
<td>#($E'$)$^6$</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>1$^7$</td>
<td></td>
</tr>
<tr>
<td>#($G$)$^8$</td>
<td>147</td>
<td>135</td>
<td>117</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>#($P'$)$^9$</td>
<td>3,237</td>
<td>2,827</td>
<td>1,518</td>
<td>1,888</td>
<td></td>
</tr>
<tr>
<td>#($M$)$^{10}$</td>
<td>2,463</td>
<td>2,594</td>
<td>1,624</td>
<td>1,889</td>
<td></td>
</tr>
<tr>
<td>#($G,M,E',P'$)</td>
<td>1,232,086</td>
<td>1,189,379</td>
<td>339,491</td>
<td>986,612</td>
<td></td>
</tr>
<tr>
<td>#($G,M$)$^{11}$</td>
<td>124,342</td>
<td>123,343</td>
<td>57,735</td>
<td>83,116</td>
<td></td>
</tr>
<tr>
<td>#($G_{new}$)$^{12}$</td>
<td>24</td>
<td>23</td>
<td>20</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>#($G_{new},M$)$^{13}$</td>
<td>18,648</td>
<td>18,002</td>
<td>9,539</td>
<td>13,223</td>
<td></td>
</tr>
</tbody>
</table>

---

$^6$ #($E'$): number of Entrez Gene records hyperlinked to the OMIM record.
$^7$ Corresponding to the human gene TP53 dataset in Sections 2 and 3.
$^8$ #($G$): number of distinct GO terms annotating $E'$.
$^9$ #($P'$): number of distinct PubMed records hyperlinked to $E'$.
$^{10}$ #($M$): number of distinct MeSH terms annotating $P'$.
$^{11}$ #($G,M$): number of distinct CV term associations.
$^{12}$ #($G_{new}$): number of new GO terms introduced by aggregation.
$^{13}$ #($G_{new},M$): number of distinct CV term associations generated by aggregation.
<table>
<thead>
<tr>
<th>GO Term</th>
<th>Parent GO Term</th>
<th>Conf</th>
<th>Rank</th>
<th>Conf2SS Rank2SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA binding</td>
<td></td>
<td>0.0180</td>
<td>156</td>
<td>0.0099 133</td>
</tr>
<tr>
<td>transcription factor activity</td>
<td>DNA binding</td>
<td>0.0045</td>
<td>2,572</td>
<td>3,522</td>
</tr>
<tr>
<td>damaged DNA binding</td>
<td>DNA binding</td>
<td>0.0005</td>
<td>31,030</td>
<td>38,349</td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>DNA binding</td>
<td>0.0005</td>
<td>31,030</td>
<td>38,349</td>
</tr>
</tbody>
</table>

Table 2. BREAST CANCER user dataset having MeSH descriptor term Cell Cycle Proteins

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Parent GO Term</th>
<th>Conf</th>
<th>Rank</th>
<th>Conf2SS Rank2SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>phosphoinositide 3-kinase activity</td>
<td></td>
<td>0.0125</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>phosphatidylinositol-4,5-bisphosphate 3-kinase activity</td>
<td>phosphoinositide 3-kinase activity</td>
<td>0.0325</td>
<td>29</td>
<td>32</td>
</tr>
<tr>
<td>1-phosphatidylinositol-3-kinase activity</td>
<td>phosphoinositide 3-kinase activity</td>
<td>0.0175</td>
<td>161</td>
<td>195</td>
</tr>
</tbody>
</table>

Table 3. BREAST CANCER user dataset having MeSH descriptor term 1-Phosphatidylinositol 3-Kinase

We calculate a baseline confidence score, \( Conf_B \), for associations of the parent CV term that does not reflect aggregation evidence, and a boosted confidence score \( Conf_{2SS} \). We also report on the original rank \( Rank_B \) and the new rank \( Rank_{2SS} \). Note that for each user dataset, \( Rank_{2SS} \) is determined over a combination (union) of both the original pairs of associations of CV terms and any new associations introduced via aggregation. For example, for the BREAST CANCER dataset, \( Rank_{2SS} \) will be determined over \((124,342+18,648)\) associations. We use values of \( \alpha = \frac{1}{2} \).

The first example in Table 2 involves a parent GO term DNA binding and its three child terms, transcription factor activity, damaged DNA binding and sequence-specific DNA binding. The associated MeSH term is Cell Cycle Proteins. We see that the parent term already has the highest confidence score (among these associations) and has a rank of 156. The confidence score of the child terms are low and they are farther back in rank. There is also high variance in the confidence score of the child terms. Nevertheless, there is a positive contribution from the child terms and the parent term’s boosted rank is 133. We note that the actual confidence score of the parent term has gone down after boosting and we note that in general the scores for confidence score tend to reduce after boosting. However, the rank is determined using the score relative to other associations. Thus, while the actual score may reduce, the rank may actually be improved.

In the second example in Table 3, the parent term phosphoinositide 3-kinase activity does not have a confidence score since there are no termlinks for this GO term to the MeSH term 1-Phosphatidylinositol 3-Kinase. The parent term has two child terms, phosphatidylinositol-4,5-bisphosphate 3-kinase activity and 1-phosphatidylinositol-3-kinase activity. Both child terms have high confidence scores and their ranks are also very good, at 29 and 161, respectively. The variance in the child terms is also low. This is
<table>
<thead>
<tr>
<th>GO Term</th>
<th>Parent GO Term</th>
<th>$Conf_{B}$</th>
<th>$Rank_{B}$</th>
<th>$Conf_{SS}$</th>
<th>$Rank_{SS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein binding</td>
<td></td>
<td>0.0165</td>
<td>147</td>
<td>0.0126</td>
<td>93</td>
</tr>
<tr>
<td>enzyme binding</td>
<td>protein binding</td>
<td>0.0174</td>
<td>101</td>
<td>0.0174</td>
<td>129</td>
</tr>
<tr>
<td>protein N-terminus binding</td>
<td>protein binding</td>
<td>0.0174</td>
<td>101</td>
<td>0.0126</td>
<td>93</td>
</tr>
<tr>
<td>protein C-terminus binding</td>
<td>protein binding</td>
<td>0.0004</td>
<td>40,481</td>
<td>0.0001</td>
<td>47,729</td>
</tr>
<tr>
<td>insulin receptor substrate binding</td>
<td>protein binding</td>
<td>0.0001</td>
<td>117,248</td>
<td>1.33,069</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. COLORECTAL CANCER user dataset having MeSH descriptor term Tumor Suppressor Protein p53

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Parent GO Term</th>
<th>$Conf_{B}$</th>
<th>$Rank_{B}$</th>
<th>$Conf_{SS}$</th>
<th>$Rank_{SS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>integral to membrane</td>
<td></td>
<td>0.0429</td>
<td>14</td>
<td>0.0394</td>
<td>4</td>
</tr>
<tr>
<td>integral to plasma membrane</td>
<td>integral to membrane</td>
<td>0.0366</td>
<td>26</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

Table 5. PROSTATE CANCER user dataset having MeSH descriptor term Kangai-1 Protein

A situation where the boost provided by the child terms should be the most significant, i.e., the confidence score in the child terms is high and variance in confidence score is low. Thus, after the parent term is boosted, it too has a very good rank of 71. We note that the rank of the child terms terms has worsened slightly. To explain, there are several parent GO term associations that did not occur in the original termlinks that have been introduced after aggregation. They tended to be ranked ahead of the child terms from the example.

In the third example in Table 4, we consider the parent GO term protein binding in the COLORECTAL CANCER user dataset. The parent GO term has four child terms, enzyme binding, protein N-terminus binding, protein C-terminus binding and insulin receptor substrate binding. The confidence scores of the associations of child terms enzyme binding and protein N-terminus binding is high and their rank is 101. The confidence scores of the other two child terms is very low. This is a case where the confidence scores in two child terms is high and there is also high variance among the child terms’ confidence scores. The boost should not be as significant as in the previous case. We see that the parent rank has improved from 147 to 93. Thus, the boost is not as significant as in Table 3.

In the final example in Table 5, we consider the PROSTATE CANCER user dataset. The parent term integral to membrane has only one child term integral to plasma membrane. The associated MeSH term is Kangai-1 Protein. Both parent and child have high confidence scores and their rank is within the Top 30. The boosted confidence score for the parent term pushes it to rank first among the $57,735+9,539$ associations for this user dataset! To summarize, we use a variety of GO is-a hierarchies, and range of confidence scores for the child terms, to illustrate the impact on the parent CV term.
6.3 Impact of $\alpha$ on Boosted Rank

We consider the BREAST CANCER dataset; it has 124,342 associations prior to aggregation and 18,642 associations are added after aggregation. We select the Top 300 associations (after 2SS boosting). Figure 8 reports on the rank $\text{Rank}_B$ before boosting (Y axis) and the rank $\text{Rank}_{2SS}$ after boosting (X axis), for the Top 300. If an association did not occur in the original termlink dataset, its rank is labeled no rank on the Y axis. We compare two $\alpha$ values, $\frac{1}{2}$ and $\frac{1}{4}$.

A 45 degree line in Figure 8 represents the case where there is no change in the rank from boosting. For $\alpha=\frac{1}{2}$ (labeled $+$), the contribution from the child terms is only 25%; hence we see many of these datum clustered around the no change in rank line. There are a few datum scattered above the line indicating cases where the ranks have improved after boosting.

For $\alpha=\frac{1}{4}$ (labeled $\bullet$), the situation is quite different since the contribution from the child terms is more significant at 50%. Many of the datum above the baseline indicate improvement of the rank. Among these improvements, there are six new associations (originally with no rank) and 21 associations whose original ranks were greater than 8,000 that now occur in the Top 300.

7 Conclusion

We have presented an approach and preliminary evaluation to exploit knowledge from ontologies and patterns of annotation to identify significant associations jointly offer a bridge between a pair of ontologies. In future work, we will consider further extensions, e.g., aggregating simultaneously using structure of both ontologies, aggregating up multiple levels, etc. We also plan an extensive evaluation on termlinks to identify interesting patterns of annotation, and study their impact on finding significant associations.

Acknowledgments This research has been partially supported by the National Science Foundation under grants IIS0222847, IIS0430915 and 0312356.
References

8. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Record 22(2) (June 1993) 207–216
12. Han, J., Fu, Y.: Discovery of multiple-level association rules from large databases. In: Proceeding of the 21th International Conference on Very Large Data Bases (VLDB 1995), Zürich, Switzerland (11-15 September 1995) 420–431