Weighted Pseudo-distances for Categorization in Semantic Hierarchies

Cliff A. Joslyn[1] and William J. Bruno[2]

[1] Computer and Computational Sciences, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA, joslyn@lanl.gov
[2] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA, billb@lanl.gov

Abstract. Ontologies, taxonomies, and other semantic hierarchies are increasingly necessary for organizing large quantities of data. We continue our development of knowledge discovery techniques based on combinatorial algorithms rooted in order theory by aiming to supplement the pseudo-distances previously developed as structural measures of vertical height in poset-based ontologies with quantitative measures of vertical distance based on additional statistical information. In this way, we seek to accommodate weighting of different portions of the underlying ontology according to this external information source. We also wish to improve on the deficiencies of existing such measures, in particular Resnik's measure of semantic similarity in lexical databases such as Wordnet. We begin by recalling and developing some basic concepts for ordered data objects, including our pseudo-distances and the operation of probability distributions as weights on posets. We then discuss and critique Resnik's measure before introducing our own sense of links weights and weighted normalized pseudo-distances among comparable nodes.

1 Introduction

We are pursuing approaches to knowledge discovery based on combinatorial algorithms rooted in order theory [8], casting databases as ordered combinatorial data objects equipped both with inherent semantics and appropriate quantitative measures to support user-guided discovery tasks such as search, retrieval, discovery, anomaly detection, linkage, and alignment, with applications in intelligence analysis, homeland defense, computational biology, and law enforcement.

Ontologies, taxonomies, and other semantic hierarchies are increasingly necessary for organizing large quantities of data. Recent years have seen the emergence of a prominent example in the Gene Ontology (GO)[3] [5], a large, ($>$ 16K node) multi-taxonomy of biological functions and processes annotated to thousands of genes. Other cases include the UMLS Meta-Thesaurus [2], object-oriented typing hierarchies [12], and verb typing hierarchies in computational linguistics

[3] http://www.geneontology.org

F. Dau, M.-L. Mugnier, G. Stumme (Eds.): ICCS 2005, LNAI 3596, pp. 381–395, 2005.

[3]. Cast as Directed Acyclic Graphs (DAGs), these all entail partially ordered sets (posets) [16]. And other order-theoretical techniques such as formal concept analysis [4] and reconstructibility analysis, which uses lattices of reconstruction hypotheses of relational databases cast as irredundant covers of a space of variables [10], are available to provide order-theoretical representations of relational data objects. Research in fundamental properties and techniques for ordered data objects is sorely lacking; and their increasing size and the need for such tasks as linkage and federation of multiple ontologies constructed on similar domains by different organizations, creates unmet challenges in computer science and combinatorial algorithms.

In prior work we have explored some order-theoretical properties of semantic hierarchies such as measures of distance, level, and size in posets [6]. We have also developed the POSet Ontology Categorizer (POSOC)[4] for the task of "categorization" of gene lists in the GO, represented as a multi-labeled partially ordered set (poset), which we call a POSet Ontology (POSO) [7,9,17].

POSOC takes as a query a list of genes of interest, and then calculates a score for each node in the GO to represent how well that node best captures the overall distribution and location of the query in the poset. This score, whose ranks are as illustrated in an example in Fig. 1, depends on the vertical "separation" from the scored node to those nodes below it where query terms sit. We thus developed the concept of a pseudo-distance $\delta(a, b)$ between comparable nodes $a \leq b$ to measure this vertical distances as a property of the collection of lengths of the chains in the chain decomposition of the poset interval $[a, b]$.

It has been noted that our approach suffers from a deficiency, in that the structure of the GO is not uniform, but may be "denser" in certain regions and "sparser" in others, depending, for example, on the attention paid by the authors, or even the vagaries of funding of the research which supports construction of the GO. As an illustration, consider the taxonomy on the left of Fig. 2. While it appears that "grey wolf" should be as "close" to "animal" as is "ungulate", in fact we know that this isn't the case.

So our structural measure based on chains lengths should be supplemented by an approach to "stretch" or "shrink" links in the structure based on other available information. We propose to capture this additional information by a probability distribution p cast onto the POSO, and then use p to modify POSOC's current pseudo-distances $\delta(a, b)$ to become a weighted pseudo-distance $\delta^w(a, b)$ reflecting this degree of "stretch".

Here we do not presume any particular source for these probabilities. In practice, we're inspired by similar motivations as Lord *et al.* [13], who constructed p as the frequency with which GO node terms appeared in SWISS-PROT-Human protein database. They then used Resnik's measure of semantic similarity [15] developed for Wordnet[5] to measure the semantic distance between GO nodes.

[4] `http://www.c3.lanl.gov/~joslyn/posoc.html`
[5] `http://www.cogsci.princeton.edu/~wn`

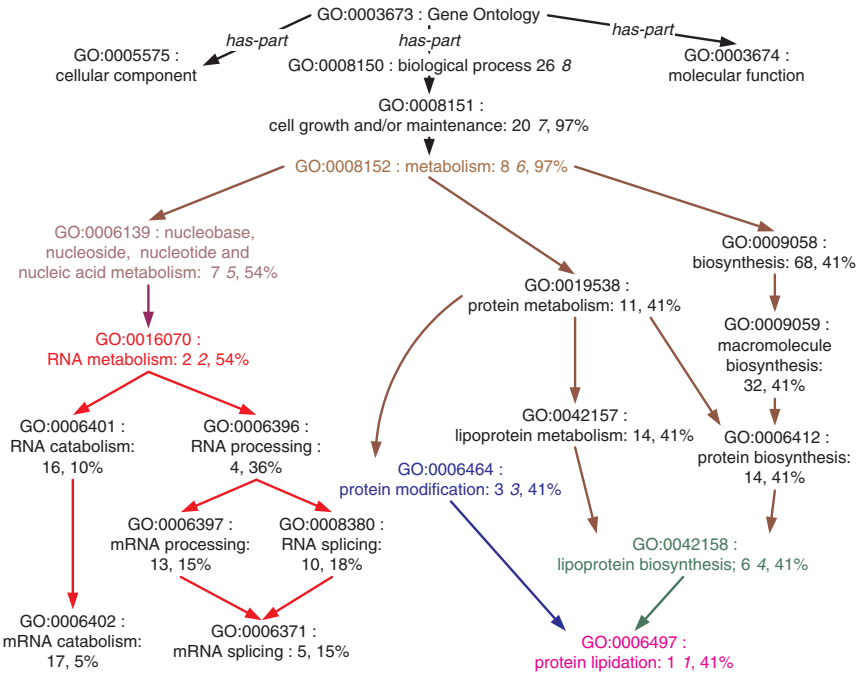

Fig. 1. Partial output from POSOC for a sample query [7]. Nodes in the GO are annotated by the rank of their score and the percentage of the query they cover.

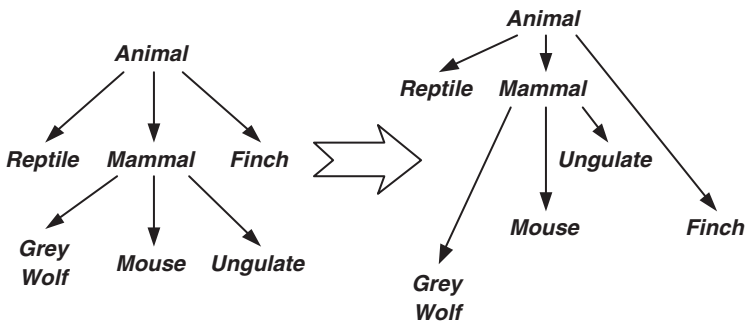

Fig. 2. A semantic hierarchy (left); with stretched links (right).

In this paper, we begin by recalling and developing some basic concepts for ordered data objects, including our current pseudo-distances δ and the operation of probability distributions as weights on posets. We then discuss and critique Resnik's measure of "semantic similarity" in weighted POSOS before introducing our own sense of link weights and weighted normalized pseudo-distances among comparable nodes.

2 Preliminaries on DAGs, Posets, and Covers

Assume a finite set of nodes P with $|P| \geq 2$. Then, assume a **directed acyclic graph (DAG)** $\Gamma \subseteq P^2$ where $\gamma = \langle a, b \rangle \in \Gamma$ is a directed edge from a node $b \in P$ to another $a \in P$. We also sometimes use $\gamma(a, b)$ to indicate a particular edge $\langle a, b \rangle \in \Gamma$. Then a **chain** (specifically, a **DAG chain**) of length $h \geq 2$ is a collection of edges

$$C = \langle \gamma_1, \gamma_2, \ldots, \gamma_i, \ldots, \gamma_{h-1} \rangle \subseteq \Gamma \tag{1}$$

where, letting $\gamma_i = \langle a_i, b_i \rangle$, for $h > 2$ and $1 \leq i \leq h - 1$, we have $b_i = a_{i+1}$.

We also represent Γ as a relational structure $\Gamma = \langle P, \Leftarrow \rangle$, where $\Leftarrow \subseteq P^2$ is a relation on P such that $\forall a, b \in P, a \Leftarrow b \leftrightarrow \langle a, b \rangle \in \Gamma$, so that there's a direct link in the DAG from b "down" to a.

The DAG Γ uniquely generates two mathematical structures:

- Let $\mathcal{V}(\Gamma) := \langle P, \prec \rangle$, be the **cover relation**, or just **cover**, where $\prec$ is the transitive reduction of $\Leftarrow$ [1]. So $a \prec b$ when $\gamma(a, b)$ is an edge in Γ which is non-transitive, that is, $\langle \gamma(a, b) \rangle$ is the only chain C with $a_1 = a, b_{h-1} = b$.
- Let $\mathcal{P}(\Gamma) := \langle P, \leq \rangle$ be the **partially ordered set (poset)** defined by Γ by transitive and reflexive closure of $\Leftarrow$. Below we sometimes use just $\mathcal{P}$ when clear from context. So $a \leq b$ when there's any chain C with $a_1 = a, b_{h-1} = b$.

Fig. 3 shows an example DAG Γ on $P = \{0, A, B, \ldots, K, 1\}$, with two transitive edges $D \Leftarrow B, I \Leftarrow 1$ shown as dashed lines, removal of which yields the cover relation $\mathcal{V}(\Gamma)$. $\mathcal{P}(\Gamma)$ would result from both retaining $D \Leftarrow B$ and $I \Leftarrow 1$ and adding all other transitive links, e.g. $H \Leftarrow B, E \Leftarrow C$. Note the inclusion of the special nodes $0, 1 \in P$ as global bounds, which we will always assume are present.

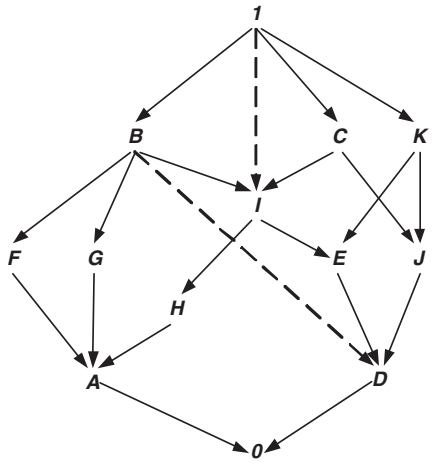

Fig. 3. A bounded DAG.

Given a DAG Γ, we generally focus on the corresponding poset and cover relations. In particular, below we will use $a \leq b$ to mean that $a \leq b$ in the poset $\mathcal{P}(\Gamma)$, and $a \Leftarrow b$, or $\gamma(a,b)$, to mean that $a \prec b$ in the cover $\mathcal{V}(\Gamma)$. $\forall a \in P$, let

$$\downarrow a := \{b \in P : b \leq a\}, \quad \uparrow a := \{b \in P : a \leq b\},$$

$$\dot{\downarrow} a := \{b \in P : b \prec a\}, \quad \dot{\uparrow} a := \{b \in P : a \prec b\}$$

be its **ideal, filter, children**, and **parents** respectively.

Two nodes $a, b \in P$ are called **comparable**, denoted $a \sim b$, when either $a \leq b$ or $b \leq a$. When $a, b \in P, a \neq b$ and $a \leq b$, we simply say $a \leq b \in P$. Two nodes $a, b \in P$ are **non-comparable** if $a \not\sim b$. A collection of nodes $A \subseteq P$ is also called a **chain** (specifically, a **poset chain**) if $\forall a, b \in A, a \sim b$, and an **antichain** if $\forall a, b \in A, a \not\sim b$. The **height** $\mathcal{H}(\mathcal{P})$ of a poset, or just $\mathcal{H}$, is the size of the largest chain, and the **width** $\mathcal{W}(\mathcal{P})$ is the size of the largest anti-chain.

Note that unlike in lattices, when $\mathcal{P}$ is a general poset then for a pair of nodes $a, b \in P$ the concepts of least upper bound $a \vee b$ and greatest lower bound $a \wedge b$, when they exist, are not singular. Rather, $a \vee b, a \wedge b \subseteq P$ are the sets of the (possibly multiple) least upper (greatest lower) bounds of a and b. For example, in Fig. 3 we have $E \vee J = \{C, K\} \subseteq P$.

Below assume two comparable nodes $a \leq b \in P$. Then let

$$\mathcal{C}(a,b) := \{C_1, C_2, \ldots, C_j, \ldots, C_M\} \subseteq 2^{2^P}$$

be the set of all DAG chains from a to b. $[a,b] := \{c \in P : a \leq c \leq b\}$ is the **interval** from a to b and is always a bounded sub-poset of $\mathcal{P}$ with $[a,b] = \bigcup_{j=1}^{M} C_j$. From Dilworth's theorem [16], we know that $M \geq \mathcal{W}([a,b])$. But otherwise, while bounded, the number M of chains is generally arbitrary with respect to a and b.

Let $h_j := |C_j| - 1$ be the length of a particular such chain $C_j \in \mathcal{C}(a,b)$. While in principle $0 \leq h_j \leq \mathcal{H} - 1$, note that $h_j = 0 \leftrightarrow a = b$, so in practice $h_j \geq 1$. Also define $\bar{h}_j := h_j/(\mathcal{H} - 1)$ as a chain length normalized to the height of $\mathcal{P}$. Let

$$\boldsymbol{h}(a,b) := \langle h_1, h_2, \ldots, h_j, \ldots, h_M \rangle, \quad \bar{\boldsymbol{h}}(a,b) := \boldsymbol{h}/(\mathcal{H} - 1)$$

be the **vectors of chain lengths** connecting a to b. We also denote

$$h_*(a,b) = \min_{h_j \in \boldsymbol{h}(a,b)} h_j, \quad h^*(a,b) = \max_{h_j \in \boldsymbol{h}(a,b)} h_j,$$

$$\bar{h}_*(a,b) = \min_{\bar{h}_j \in \bar{\boldsymbol{h}}(a,b)} \bar{h}_j, \quad \bar{h}^*(a,b) = \max_{\bar{h}_j \in \bar{\boldsymbol{h}}(a,b)} \bar{h}_j.$$

Finally, note that for $a \neq b$ and DAG chain C_j, we have $C_j = \{a, b\} \cup C$ for some, possibly empty, chain $C = \{c_2, c_3, \ldots, c_{h_j - 1}\} \subseteq P$, so that for $2 \leq i \leq h_j - 1$,

$$a \prec c_2 \prec c_3 \prec \ldots \prec c_{h_j - 2} \prec c_{h_j - 1} \prec b. \tag{2}$$

3 Pseudo-distances

Our approach begins with measures between comparable nodes $a \leq b$, which indicate how "high" b is above a. A **pseudo-distance** is a function $\delta\colon P^2 \to \mathbb{R}$ where $\forall a \leq b \in P, h_*(a,b) \leq \delta(a,b) \leq h^*(a,b)$, provideing some aggregate measure of the number of "hops" between two comparable nodes. There is also a normalized form $\bar{\delta} := \delta/(\mathcal{H}-1)$ which measures what proportion of the height of the whole poset $\mathcal{P}$ is taken up between a and b, so that $\bar{\delta}(a,b) \in [0,1]$.

Current pseudo-distances implemented in POSOC include:

- **Minimum chain length:** $\delta_m(a,b) := h_*(a,b), \bar{\delta}_m(a,b) := \bar{h}_*(a,b)$
- **Maximum chain length:** $\delta_x(a,b) := h^*(a,b), \bar{\delta}_x(a,b) := \bar{h}^*(a,b)$
- **Average of extreme chain lengths:**

$$\delta_{ax}(a,b) := \frac{h_*(a,b) + h^*(a,b)}{2}, \qquad \bar{\delta}_{ax}(a,b) := \frac{\bar{h}_*(a,b) + \bar{h}^*(a,b)}{2}.$$

- **Average of all chain lengths:**

$$\delta_{ap}(a,b) := \frac{\sum_{h_j \in \boldsymbol{h}(a,b)} h_j}{M}, \qquad \bar{\delta}_{ap}(a,b) := \frac{\sum_{\bar{h}_j \in \bar{\boldsymbol{h}}(a,b)} \bar{h}_j}{M}.$$

In our example in Fig. 3 (recalling that transitive edges are removed from the cover relation $\prec$), we have $\mathcal{H}(\mathcal{P}) = 6$. Considering $D \leq 1$, then $\mathcal{W}([D,1]) = 3$, while $M = 5$, and

$$\begin{aligned}
\mathcal{C}(D,1) &= \{\, D \prec E \prec I \prec B \prec 1, D \prec E \prec I \prec C \prec 1, \\
&\qquad D \prec E \prec K \prec 1, D \prec J \prec C \prec 1, D \prec J \prec K \prec 1\}, \\
\boldsymbol{h}(D,1) &= \langle 4,4,3,3,3\rangle\,, \quad \bar{\boldsymbol{h}}(D,1) = \langle 4/5, 4/5, 3/5, 3/5, 3/5\rangle\,, \\
\delta_m(D,1) &= 3, \delta_x(D,1) = 4, \delta_{ax} = 3.5, \delta_{ap}(D,1) = 3.4, \\
\bar{\delta}_m(D,1) &= 0.60, \bar{\delta}_x(D,1) = 0.80, \bar{\delta}_{ax} = 0.70, \bar{\delta}_{ap}(D,1) = 0.68.
\end{aligned}$$

4 Weighted Posets

Our overall motivation is to improve on the quantitative approach which Lord *et al.* [13] and Resnik [15] brought to measures of distance (in their language, "semantic similarity") in taxonomies equipped with probability distributions. To do so, we need to understand the basic operations of probabilities on posets.

Definition 1 (Weighted Poset). *Define a weighted poset as a structure* $\mathcal{O} := \langle \mathcal{P}(\Gamma), p\rangle$*, where* $p\colon P \to [0,1]$ *is a probability distribution on the nodes* P *of the poset* $\mathcal{P}(\Gamma)$*, so that* $\sum_{a \in P} p(a) = 1$.

For any node $b \in P$, what we will call a "beta" function $\beta\colon P \to [0,1]$ is a kind of probability measure over Γ, defined as

$$\beta(b) := \sum_{a \leq b} p(a) = \sum_{a \in \downarrow b} p(a). \tag{3}$$

Fig. 4 continues our example with $\beta(a)$ to the right of each node, and $p(a)$ below it. Weights are also shown on links, which will be discussed in Sec. 7 below.

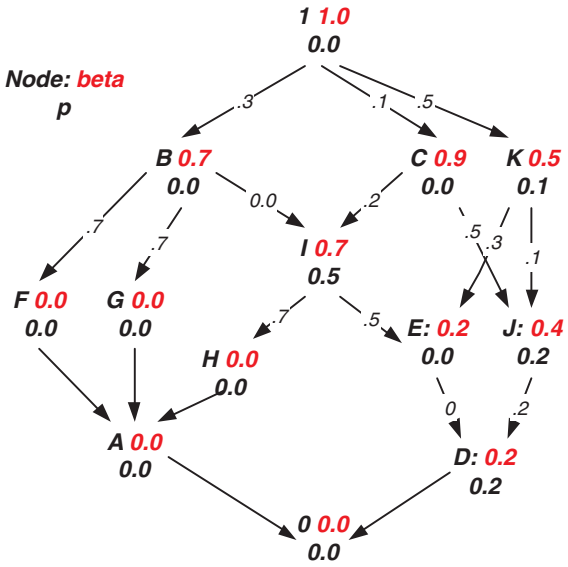

Fig. 4. An example of a weighted poset $\mathcal{O}$.

β is what's known as an isotone, or order-preserving, map on $\mathcal{P}$, which is crucial in Monjardet's general theory of metrics in posets [14]:

Proposition 1. $a \leq b \to \beta(a) \leq \beta(b)$.

Proof. Follows from $a \leq b \to \downarrow a \subseteq \downarrow b$, Eq. (3), and $p(a) \geq 0$. □

5 A Mathematical Aside

We are regretfully ignorant of literature concerning discrete probability distributions like p on finite ordered sets, and yet we can note some intriguing connections to some powerful formalisms. In particular, let $\mathbf{B}(\mathcal{O}) := \{a \in P : p(a) > 0\} \subseteq P$

be the **base** of $\mathcal{O}$. Then consider the case when $\mathcal{P}$ is a Boolean lattice, in particular the power set 2^{Ω} on some underlying finite set Ω. β then is the **belief function** Bel on Ω from Dempster-Shafer evidence theory [11], and Eq. (3) becomes $\forall A \subseteq \Omega, \mathrm{Bel}(A) = \sum_{B \subseteq A} p(A)$. Bel is super-additive, with the modular property

$$\forall A, B \subseteq \Omega, \quad \mathrm{Bel}(A \cup B) \geq \mathrm{Bel}(A) + \mathrm{Bel}(B) - \mathrm{Bel}(A \cap B). \tag{4}$$

Expressed back in the lattice $\mathcal{P}$, Eq. (4) becomes

$$\forall a, b \in P, \quad \beta(a \vee b) + \beta(a \wedge b) \geq \beta(a) + \beta(b), \tag{5}$$

recalling that the single point $a \vee b \in P$ now always exists. There are some important special cases, for example:

- If $\mathbf{B}(\mathcal{O})$ is the antichain of the atoms of $\mathcal{P}$, then Eq. (5)

  $$\forall a, b \in P, \quad \beta(a \vee b) + \beta(a \wedge b) = \beta(a) + \beta(b),$$

  so that Bel becomes a classical probability measure Pr with p its discrete distribution, and

  $$\forall A, B \subseteq \Omega, \quad \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

- If $\mathbf{B}(\mathcal{O})$ is a maximal chain $C \subseteq \mathcal{P}$ with $|C| = \mathcal{H}(\mathcal{P})$, then

  $$\forall a, b \in P, \quad \beta(a \wedge b) = \min(\beta(a), \beta(b))$$

  so that Bel becomes a so-called "necessity measure" η with

  $$\forall A, B \subseteq \Omega, \quad \eta(A \cap B) = \min(\eta(A), \eta(B)).$$

Open questions remain about these properties when $\mathcal{P}$ is a general lattice, a complemented lattice, or a general poset (all finite). However, it's interesting to consider a potential form of Eq. (5) when $\mathcal{P}$ is a general poset:

$$\forall a, b \in P, \quad \sum_{c \in a \vee b} \beta(c) + \sum_{c \in a \wedge b} \beta(c) \geq \beta(a) + \beta(b),$$

and consider connections to general families of semimodular maps on posets [14].

6 Resnik's Semantic Similarity

A currently attractive way to approach semantic distance in semantic hierarchies is to use Resnik's measure of "semantic similarity" [15], originally developed for application to Wordnet, but then applied successfully by Lord *et al.* to the GO [13]. Not only do these concepts have a natural interpretation in our language, they also serve as a point of departure for our development.

Definition 2 (Resnik Semantic Similarity). *Given a weighted poset $\mathcal{O}$, then $\forall a, b \in P$, define*

$$\delta_{Resnik}(a, b) = \max_{c \in a \vee b} \left[-\log_2(\beta(c)) \right]. \tag{6}$$

Some issues are immediately evident when comparing the Resnik measure to our pseudo-distances δ. First, unlike δ_{Resnik}, our measure δ is definitely *not* a distance, most significantly because it is not defined for pairs of general nodes $a, b \in P$, but rather only for comparable nodes $a \leq b \in P$.

There is also some ambiguity as to the semantics of δ_{Resnik} as a measure of information content, since as discussed in Sec. 5, while the p's are definitely values of a discrete distribution on P, β is almost never actually a probability measure on P.

Also, if a portion of $\mathcal{P}$, and in particular the ideal $\downarrow b$ of a node $b \in P$, is a lattice, then the similarities between all the children of b are identical, no matter their β values.

Theorem 1. *Let $b \in P$ with $\downarrow b \subseteq P$ a lattice. Then $\forall a_1, a_2, a_3, a_4 \in \dot{\downarrow} b$, $\delta_{Resnik}(a_1, a_2) = \delta_{Resnik}(a_3, a_4)$.*

Proof. Since $\downarrow b$ is a lattice, therefore $\forall a, a' \in \dot{\downarrow} b, a \vee a'$ exists uniquely, and in particular $a \vee a' = \{b\}$. Thus $\forall a_1, a_2, a_3, a_4 \in \dot{\downarrow} b, \delta_{\text{Resnik}}(a_1, a_2) = -\log(\beta(b)) = \delta_{\text{Resnik}}(a_3, a_4)$. □

But most significantly, δ_{Resnik} cannot distinguish among links in a chain.

Theorem 2. *If $a \leq b \leq c$, then $\delta_{Resnik}(a, c) = \delta_{Resnik}(b, c) = \delta_{Resnik}(c, c)$.*

Proof. Since $a \leq b \rightarrow a \vee b = \{b\}$ uniquely in P, then $\delta_{\text{Resnik}}(a, c) = \delta_{\text{Resnik}}(b, c) = \delta_{\text{Resnik}}(c, c) = -\log(\beta(c))$. □

This is precisely contrasted with our pseudo-distances δ and $\bar{\delta}$, which satisfy

$$a \leq b \leq c \quad \rightarrow \quad \delta(b, c) \leq \delta(a, c), \quad \bar{\delta}(b, c) \leq \bar{\delta}(a, c).$$

7 Link Weights in Weighted Posets

Despite the weaknesses of δ_{Resnik}, it is valuable in pointing the way towards information theoretical distance measures in probability weighted posets, and in being defined on all $a, b \in P$. While we are also actively researching such measures for multiple purposes [6], categorization in general and POSOC in particular depends only on proper measures among comparable nodes $a \leq b \in P$.

We are therefore looking for a different mechanism to introduce link weights into our categorization algorithm. Our aim is to "lengthen" chains in the weighted

poset proportional to the amount of probability concentrated along them, while leaving them at "original length" where there is none. To do so, we also take an information theoretical approach, although somewhat distinct from Resnik's.

Definition 3 (Information Gain). *For all pairs of comparable nodes $a \leq b \in P$, let*

$$\iota(a,b) := \beta(b) - \beta(a) \tag{7}$$

be the amount of information gained when moving from b down to a. For each edge $\gamma(a,b)$, let $\iota(\gamma) := \iota(a,b)$ be the edge weight.

As an example, in Fig. 4 we have

$$\iota(D,K) = \beta(K) - \beta(D) = 0.5 - 0.2 = 0.3.$$

Also in Fig. 4, we have labeled each of the arcs γ with the information gain $\iota(\gamma)$.

It is obvious that $\iota(a,b) \in [0,1]$. It is also comforting that for all pairs of comparable nodes $a \leq b \in P$, no matter which chain is traversed from b down to a, the sum of the edge weights is always equal to the information gain from b to a.

Proposition 2.

$$\forall a \leq b \in P, \quad \forall C_j \in \mathcal{C}(a,b), \quad \sum_{\gamma_i \in C_j} \iota(\gamma_i) = \iota(a,b).$$

Proof. Fix $a \leq b \in P$, and a chain $C_j \in \mathcal{C}(a,b)$ of length h_j, and use the notation from Eq. (1) and Eq. (2). Then we have

$$\begin{aligned}\sum_{\gamma_i \in C_j} \iota(\gamma_i) &= (\beta(c_2) - \beta(a)) + (\beta(c_3) - \beta(c_2)) + \ldots + \\ &\qquad (\beta(c_{h_j-1}) - \beta(c_{h_j-2})) + (\beta(b) - \beta(c_{h_j-1})) \\ &= \beta(b) - \beta(a) = \iota(a,b).\end{aligned}$$

□

This development follows Monjardet almost precisely [14], with β an isotone map on $\mathcal{P}$, $\iota(\gamma)$ an edge weight, and $\iota(a,b)$ a path weight.

Continuing our example, for $D \leq 1$, where we still have the $M = 5$ chains, we also have

$$\begin{aligned}\iota(D,1) &= .8 \\ &= \iota(\langle D,E\rangle) + \iota(\langle E,I\rangle) + \iota(\langle I,B\rangle) + \iota(\langle B,1\rangle) = 0.0 + 0.5 + 0.0 + 0.3 \\ &= \iota(\langle D,J\rangle) + \iota(\langle J,K\rangle) + \iota(\langle K,1\rangle) = 0.2 + 0.1 + 0.5\end{aligned}$$

and similarly for the other chains. This is very convenient, because then we can deal with the information gain between comparable nodes and edge weights indiscriminately: information gain can always be calculated by edge weights, but we can also work with information gains whenever we want.

8 Weighted Normalized Pseudo-distances

For comparable nodes $a \leq b \in P$, we now want to take the chain lengths h_j (or the normalized chain lengths $\bar{h}_j$) and adjust them by the information gain $\iota(a,b)$. Thus for a particular chain $C_j \in \mathcal{C}(a,b)$ of length h_j $\left(\bar{h}_j\right)$, we wish to construct a weighted chain length $v_j(a,b)$ (or a normalized weighted chain length $\bar{v}_j(a,b)$) as a function of h_j $\left(\bar{h}_j\right)$ and $\iota(a,b)$, which will be equal to h_j $\left(\bar{h}_j\right)$ scaled up with $\iota(a,b)$. In this paper, we have only developed the normalized case for $\bar{v}_j$.

Considering a generic pair of comparable nodes $a \leq b \in P$ with information gain $\iota = \iota(a,b) = \beta(b) - \beta(a)$, and a particular chain $C_j \in \mathcal{C}(a,b)$ with length $\bar{h} = \bar{h}_j$, then how should the weighted, normalized chain length $\bar{v}$ be determined? Our motivations are as follows.

First, h and ι are in some sense independent quantities. As illustrated in cartoon form in Fig. 2, the size of ι indicates the amount of "stretch" in the chain C, no matter the number of "hops" along its length. So, "grey wolf" is farther from "animal" than "ungulate" is, despite the fact that they're both chains of length two. Similarly, "finch" and "mouse" are stretched to be a similar distance from "animal", despite the differences in those chain lengths. So despite the fact that by Thm. 1, as any particular chain C_j grows, ι increases with the chain length $\bar{h}_j$, nonetheless in any particular case $\bar{h}_j$ could be small while ι is large, or *vice versa*.

We thus wish to identify a function $f\colon [0,1]^2 \to [0,1]$ such that $\bar{v}_j := f\left(\bar{h}_j, \iota(a,b)\right)$, and list the properties desirable for $f(h,\iota)$ with $h, \iota \in [0,1]$:

1. In one limit, ι could be zero, indicating that no probability mass was encountered when traversing from b down to a. In this case, we wish $\bar{v}$ to be recovered simply as $\bar{h}$, requiring that $f(h,0) = h$.
2. Otherwise, we want ι to add in to increase the total length, requiring $\bar{v} \geq \bar{h}$, or in other words $f(h,\iota) \geq h$.
3. Then, considering some limit cases, if the entire structure is traversed by a maximal chain, then $\bar{v}$ should be maximum, so that $f(1,\iota) = 1$.
4. Similary, in the degenerate case of $a = b$ so that there are no chains and no ι gain, then $\bar{v}$ should be minimal, so that $f(0,0) = 0$.
5. Finally, if $\iota = 1$ so that *all* the probability mass is encountered when traversing from b down to a, then $\bar{v}$ is maximal, no matter than length h: $f(h,1) = 1$.

Thus we arrive at the following.

Definition 4 (Weighted Normalized Chain Lengths). *For $a \leq b \in P$, and for each chain $C_j \in \mathcal{C}(a,b), 1 \leq j \leq M$, define $\bar{v}_j := f\left(\bar{h}_j, \iota(a,b)\right)$, where*

$$f(h,\iota) := h^{1-\iota}, \quad h, \iota \in [0,1], \tag{8}$$

as the **weighted chain length**. *Construct $\bar{\boldsymbol{v}}(a,b) := \langle \bar{v}_1, \bar{v}_2, \ldots, \bar{v}_j, \ldots, \bar{v}_M \rangle$ as the* **vector of weighted chain lengths**, *with $\bar{v}_j \in \bar{\boldsymbol{v}}(a,b)$, and let*

$$\bar{v}_*(a,b) = \min_{\bar{v}_j \in \bar{\boldsymbol{v}}(a,b)} \bar{v}_j, \qquad \bar{v}^*(a,b) = \max_{\bar{v}_j \in \bar{\boldsymbol{v}}(a,b)} \bar{v}_j.$$

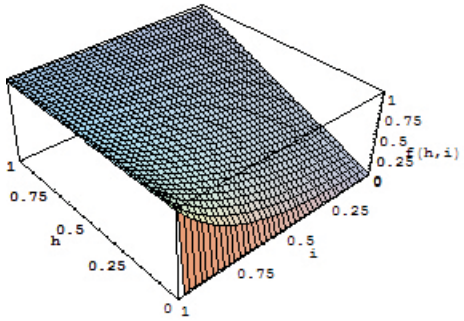

Fig. 5. $\bar{v}_j$ as a function of $\bar{h}_j$ and $\iota(a,b)$.

Theorem 3. *$\bar{v}_j$ as defined using f in Eq. (8) satisfies the conditions 1–5 above.*

Proof. Condition 2 follows from the fact that both $h, 1-\iota \in [0,1]$, so that $f(h,\iota) = h^{1-\iota} \geq h$. All the other conditions follow directly from the form of f.

Fig. 5 shows f as a surface in $[0,1]^2$, and Fig. 6 shows level curves of f for various values of h (left) and ι (right). Inspection of the figures reveals some interesting behaviors. For example, a large ι will bring up even a small h to a high value of f; for high h, f degrades gradually to be bounded below by h as ι drops, whereas for low h, this dropoff is more sudden.

We are now ready to introduce a modification to the prior pseudo-distances.

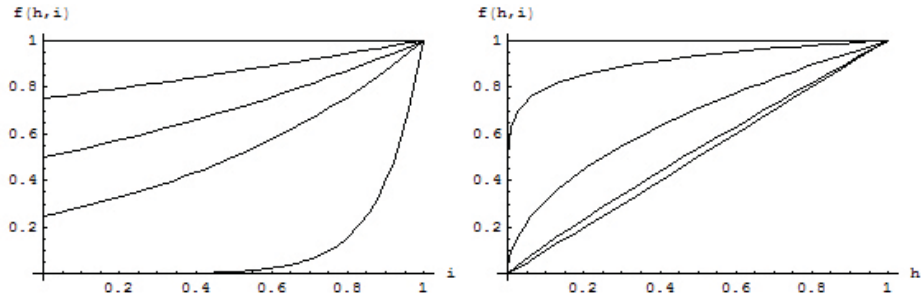

Fig. 6. (Left) Level curves of $f(h,\iota)$ for $h = 1 (f \equiv 1), .75, .5, .25, .0001$; (Right) Level curves for $\iota = 1 (f \equiv 1), .9, .5, .1, 0$.

Definition 5 (Weighted Normalized Pseudo-Distance). *Given a weighted poset $\mathcal{O}$, then for all $a \leq b \in P$, let a weighted normalized pseudo-distance $\bar{\delta}^w(a,b)$ be any function such that $\bar{v}_*(a,b) \leq \bar{\delta}^w(a,b) \leq \bar{v}^*(a,b)$. In particular, define the following weighted normalized pseudo-distances:*

- **Minimum Normalized Weighted Chain Length:** $\bar{\delta}^w_m(a,b) := \bar{v}_*(a,b)$.
- **Maximum Normalized Weighted Chain Length:** $\bar{\delta}^w_x(a,b) := \bar{v}^*(a,b)$.
- **Average of Extreme Normalized Weighted Chain Lengths:**

$$\bar{\delta}^w_{ax}(a,b) := \frac{\bar{v}_*(a,b) + \bar{v}^*(a,b)}{2}.$$

- **Average of All Normalized Weighted Chain Lengths:**

$$\bar{\delta}^w_{ap}(a,b) := \frac{\sum_{\bar{v}_j \in \bar{\boldsymbol{v}}(a,b)} \bar{v}_j}{M}.$$

Given a weighted pseudo-distance, then the POSOC methodology [7] is simply modified to substitute $\bar{\delta}^w$ for $\bar{\delta}$. The results of our example are shown in Tab. 1.

j	h_j	$\bar{h}_j$	$\bar{v}_j$
1	3.000	0.600	0.903
2	3.000	0.600	0.903
3	3.000	0.600	0.903
4	4.000	0.800	0.956
5	4.000	0.800	0.956

	δ_*	$\bar{\delta}_*$	$\bar{\delta}^w_*$
m	3.000	0.600	0.903
x	4.000	0.800	0.956
ax	3.500	0.700	0.930
ap	3.400	0.680	0.924

Table 1. Results for $D \leq 1$. (Top) Chain lengths h_j, normalized chain lengths $\bar{h}_j$, and weighted normalized chain lengths $\bar{h}_j$ for chains $C_j \in \mathcal{C}(D,1), 1 \leq j \leq M = 5$; (Bottom) Pseudo-distances δ_*, normalized pseudo-distances $\bar{\delta}_*$, and weighted normalized pseudo-distances $\bar{\delta}^w_*$ for $* \in \{x, m, ax, ap\}$.

9 Conclusion and Discussion

We have demonstrated a method by which probabilities can be used to weighte pseudo-distances in ordered data objects. Space allows only an explication of this basic step, and not its application within the overall POSOC ontology categorization algorithm. While straightforward, this requires the determination of a real probability distribution p on the GO, perhaps in a manner similar to Lord *et al* [13]. The details and results of this application, along with development for non-normalized weighted chain lengths v_j and pseudo-distances δ^w, await further work.

We should also mention an additional step we wish to take in the future. As we've seen, $\bar{v}$ is not especially sensitive to even moderate information gains ι even when the chain lengths $\bar{h}_j$ are reasonable. But moreover, the example in Fig. 4 is perhaps somewhat unfortunately chosen, in that the probability mass is concentrated on the right-hand side of the structure. The pair $D \leq 1$ we considered has a relatively large gain $\iota(D, 1) = 0.8$. In practice, in a real ontology, mass can be expected to be widely distributed over a very wide and shallow structure (the width of the GO in some places exceeds 6000, while the height is only 16), meaning that it can be expected that for any typical pair of nodes $a \leq b, \iota(a, b)$ would actually be expected to be very low, and thus $\bar{v}$ quite close to $\bar{h}$. Thus we are also considering the viability of a non-linear function for information gain instead of Eq. (7) to heighten or tune the affect of small information gains. In particular, we are considering various logarithmic forms similar in spirit to Resnik's Eq. (6).

10 Acknowledgments

This work was sponsored by the Department of Energy under contract W-7405-ENG-36 to the University of California. We would like to thank the Los Alamos National Laboratory Protein Function Inference Group for their contributions to this work, and in particular Michael Wall (LANL Computer and Computational Science) for his consultation. We also appreciate the dialog currently ongoing with Alex Sanchez, Phillip Lord, and Robert Stevens of U. Manchester. Finally, we'd like to thank Petko Valtchev of the University of Montreal for the reference to Monjardet.

References

1. Aho, AV; Garey, MR; and Ullman, JD: (1972) "The Transitive Reduction of a Directed Graph", *SIAM Journal of Computing*, v. **1**:2, pp. 131-137
2. Bodenreider, Olivier; Mitchell, Joyce A; and McCray, Alexa T: (2002) "Evaluation of the UMLS As a Terminology and Knowledge Resource for Biomedical Informatics", in: *AMIA 2002 Annual Symposium*, pp. 61-65
3. Davis, Anthony R: (2000) *Types and Constraints for Lexical Semantics and Linking*, Cambridge UP
4. Ganter, Bernhard and Wille, Rudolf: (1999) *Formal Concept Analysis*, Springer-Verlag
5. Gene Ontology Consortium: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. **25**:1, pp. 25-29
6. Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work, Lecture Notes in Artificial Intelligence*, v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin
7. Joslyn, Cliff; Mniszewski, Susan; and Fulmer, Andy; and GG Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177
8. Joslyn, Cliff; Oliverira, Joseph; and Scherrer, Chad: (2004) "Order Theoretical Knowledge Discovery: A White Paper", *LAUR = 04-5812*, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/white.pdf`

9. Joslyn, Cliff; Cohn, JD; Verspoor, KM; and Mniszewski, SM: (2005) "Automating Ontological Function Annotation: Towards a Common Methodological Framework", submitted to *2005 Bio-Ontologies Meeting, ISMB 05*
10. Klir, George and Elias, Doug: (2003) *Architecture of Systems Problem Solving*, Plenum, New York, 2nd edition
11. Klir, George and Yuan, Bo: (1995) *Fuzzy Sets and Fuzzy Logic*, Prentice-Hall, New York
12. Knoblock, Todd B and Rehof, Jakob: (2000) "Type Elaboration and Subtype Completion for Java Bytecode", in: *Proc. 27th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*
13. Lord, PW; Stevens, Robert; Brass, A; and C Goble: (2003) "Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation", *Bioinformatics*, v. **10**, pp. 1275-1283
14. Monjardet, B: (1981) "Metrics on Partially Ordered Sets - A Survey", *Discrete Mathematics*, v. **35**, pp. 173-184
15. Resnik, Philip: (1995) "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", in: *Int. Joint Conf. on Artificial Intelligence*, pp. 448-452, Morgan Kaufmann
16. Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston
17. Verspoor, Karin; Cohn, J; Joslyn, C; SM Mniszewski, A Rechtsteiner, LM Rocha, and T Simas: (2004) "Protein Annotation as Term Categorization in the Gene Ontology Using Word Proximity Networks", *BMC Bioinformatics*, v. 6, suppl 1