Towards a probabilistic model for semantic similarity on concept set DAGs

R.W. Oldford

February 8, 2010

Consider a directed acyclic graph (DAG) of nodes where attached to each node is a set of "simple concepts" $\{a, b, c, \ldots, m\}$. A set of simple concepts having more than one element might itself be considered a "complex concept", though in reality this is only a distinction in the complexity of the representation not in the meaning of the concept.

Consider two simple concepts a and b appearing in the DAG of nodes. We say that a *leads to* b, denoted $a \rightsquigarrow b$, and use the probability calculus to express the strength of this relationship, $Pr(a \rightsquigarrow b)$, and make the following assertions.

If a and b are in constituent concepts of the same node, say N_i, in the DAG, then conditional on that node alone, we have either $a \rightsquigarrow b$ or $b \rightsquigarrow a$ and we assert that each has equiprobability of occurrence. That is,

$$Pr(a \rightsquigarrow b \mid a \in N_i, b \in N_i) = Pr(b \rightsquigarrow a \mid a \in N_i, b \in N_i) = 1/2 \tag{1}$$

Synonyms might then be defined in any given context C as follows: a and b are synonyms in context C, if and only if

$$Pr(a \rightsquigarrow b \mid C) = Pr(b \rightsquigarrow a \mid C) = 1 \tag{2}$$

Suppose we know that there exists in N_i a concept that leads to b, written $N_i \ni? \rightsquigarrow b$. Then we assert that under these conditions each constituent concept has a non-zero probability of leading to b and (again under these conditions) that only constituent concepts of N_i have non-zero probability of leading to b. In general, these non-zero probabilities could all be different provided they sum to 1. Without further means of specifying these, however, we will take them to be equiprobable and so assert that

$$Pr(a \rightsquigarrow b \mid a \in N_i, N_i \ni? \rightsquigarrow b) = n_i^{-1} \tag{3}$$

where n_i is the number of constituent concepts in node N_i.

Let $par(N)$, the "parents" of node N, denote the set of nodes in the graph that have directed edges connected to N. Let $anc(N)$ denote the set of nodes in the graph for which a

directed path exists from them to the node N. Both of these apply to sets of nodes so that, for example, for any set $\mathcal{A}$ of nodes,

$$par(\mathcal{A}) = \bigcup_{N \in \mathcal{A}} par(N)$$

and similarly for $anc(\mathcal{A})$. Then we can write

$$\begin{aligned} Pr(a \rightsquigarrow b, \;\; a \in anc(N_b) \mid b \in N_b) \\ = \; Pr(a \rightsquigarrow b, \;\; a \in par(N_b) \mid b \in N_b) \\ + Pr(a \rightsquigarrow b, \;\; a \in anc(par(N_b)) \mid b \in N_b) \\ - Pr(a \rightsquigarrow b, \;\; a \in par(N_b) \cap anc(par(N_b)) \mid b \in N_b) \end{aligned} \tag{4}$$

This first term can now be written as

$$\begin{aligned} Pr(a \rightsquigarrow b, \;\; a \in par(N_b) \mid b \in N_b) \\ = \; Pr(a \rightsquigarrow b \mid a \in par(N_b), \;\; b \in N_b) \\ \times Pr(a \in par(N_b) \mid b \in N_b) \end{aligned} \tag{5}$$

For simplicity, let's enumerate the nodes of $par(N_b) = \{N_1, N_2, \ldots N_{n_\mathcal{P}}\} = \mathcal{P}$, say, with $n_\mathcal{P} = |\mathcal{P}|$. Now denote by N_I, or equivalently (with some abuse) by $\cap_{i \in I} N_i$, the intersection of the concept sets of the nodes indexed by the index set $I \subseteq \{1, 2, \cdots, n_\mathcal{P}\}$. Finally, denote by A_I the event $a \in N_I$ for any such set I of indices. Now the application of the inclusion-exclusion principle to these events means the first term of the last equation's right side can be expressed compactly as the sum

$$Pr(a \rightsquigarrow b \mid a \in par(N_b), \;\; b \in N_b)$$

$$\begin{aligned} = \; & \sum_{|I|=1} Pr(a \rightsquigarrow b, \;\; A_I \mid a \in par(N_b), \;\; b \in N_b) \\ & - \sum_{|I|=2} Pr(a \rightsquigarrow b, \;\; A_I \mid a \in par(N_b), \;\; b \in N_b) \\ & + \cdots + (-1)^{n_\mathcal{P}-1} Pr(a \rightsquigarrow b, \;\; A_{\{1,2,\cdots,n_\mathcal{P}\}} \mid a \in par(N_b), \;\; b \in N_b) \end{aligned}$$

$$= \; \sum_{k=1}^{n_\mathcal{P}} (-1)^{k-1} \sum_{|I|=k} Pr(a \rightsquigarrow b, \;\; A_I \mid a \in par(N_b), \;\; b \in N_b) \tag{6}$$

If the concept set DAG is such that, for some integer $k_0 > 0$, we have

$$Pr(a \rightsquigarrow b, \;\; A_I \mid a \in par(N_b), \;\; b \in N_b) \;\; \approx \;\; 0$$

for all $I \subseteq \{1, 2, \cdots, n_\mathcal{P}\}$ (with $|I| = k$) whenever $k > k_0$, then the previous sum is considerably simplified.

Suppose $k_0 = 1$, then

$$
\begin{aligned}
Pr(a \;\rightsquigarrow\; b \mid a \in par(N_b), \;\; b \in N_b) & \\
&= \sum_{|I|=1} Pr(a \rightsquigarrow b, \;\; A_I \mid a \in par(N_b),\; b \in N_b) \\
&= \sum_{i=1}^{n_{\mathcal{P}}} Pr(a \rightsquigarrow b, \;\; A_{\{i\}} \mid a \in par(N_b),\; b \in N_b) \qquad (7) \\
&= \sum_{i=1}^{n_{\mathcal{P}}} Pr(a \rightsquigarrow b \mid A_{\{i\}},\; a \in par(N_b),\; b \in N_b) \times Pr(A_{\{i\}} \mid a \in par(N_b),\; b \in N_b) \\
& \qquad (8)
\end{aligned}
$$

If, for example, $k_0 = 2$, then (7) is only an approximation (overestimating in this case) to the correct probability of (6).

The first term in the sum of equation (8) becomes

$$
\begin{aligned}
Pr(a \rightsquigarrow b \mid A_{\{i\}},\; a \in par(N_b),\; b \in N_b) &= Pr(a \rightsquigarrow b \mid a \in N_i,\; a \in par(N_b),\; b \in N_b) \\
&= Pr(a \rightsquigarrow b \mid a \in N_i,\; b \in N_b) \\
&= Pr(a \rightsquigarrow b \mid N_i \ni ? \rightsquigarrow b,\; a \in N_i) \\
&= \frac{1}{n_i} \qquad (9)
\end{aligned}
$$

where $n_i = |N_i|$. The second equality follows since $N_i \in par(N_b)$. The third is an assertion that in this case, the probability has the same meaning as that of equation (3), from which the last equality follows.

The second term in the sum of equation (8) becomes

$$
\begin{aligned}
Pr(A_{\{i\}} \mid a \in par(N_b),\; b \in N_b) &= Pr(a \in N_i \mid a \in par(N_b),\; b \in N_b) \\
&= \frac{1}{n_{\mathcal{P}}}. \qquad (10)
\end{aligned}
$$

The latter equality is simply asserted to be the case here for simplification of the model – there being $n_{\mathcal{P}}$ nodes in $par(N_b)$, all are given equal probability.

Together, the results of equations (9) and (10) can be inserted into the simple equation (7) to yield (at least as a coarse approximation):

$$Pr(a \;\rightsquigarrow\; b \mid a \in par(N_b), \;\; b \in N_b)$$

$$
\begin{aligned}
&\approx \sum_{|I|=1} Pr(a \rightsquigarrow b, \quad A_I \mid a \in par(N_b),\ b \in N_b) \\
&= \sum_{i=1}^{n_{\mathcal{P}}} Pr(a \rightsquigarrow b \mid A_{\{i\}},\ a \in par(N_b),\ b \in N_b) \times Pr(A_{\{i\}} \mid a \in par(N_b),\ b \in N_b) \\
&\approx \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right) \times \left(\frac{1}{n_{\mathcal{P}}}\right) \ = \ \frac{1}{n_{\mathcal{P}}} \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right) \qquad (11)
\end{aligned}
$$

the last line being an estimation where

$$
\mathcal{I}_{N_i}(a) = \begin{cases} 1 & \text{if } a \in N_i \\ 0 & \text{otherwise} \end{cases}
$$

is an indicator function to show whether a node $N_i \in par(N_b)$ actually contains a.

Together equations (5) and (11) suggest the following might be used:

$$
\begin{aligned}
&Pr(a \ \rightsquigarrow \ b, \ \ a \in par(N_b) \mid b \in N_b) \\
&\qquad \approx \left[\frac{1}{n_{\mathcal{P}}} \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right)\right] \times Pr(a \in par(N_b) \mid b \in N_b) \qquad (12)
\end{aligned}
$$

The probability $Pr(a \in par(N_b) \mid b \in N_b)$ might be estimated in a variety of ways. For example, let $m_{\mathcal{P}} = \sum_{i=1}^{n_{\mathcal{P}}} n_i$ be the total number of concepts (not necessarily unique) in $par(N_b)$ and $m_{\mathcal{A}}$ the total number in the ancestors $anc(N_b)$. The ratio $m_{\mathcal{P}}/m_{\mathcal{A}}$ might be used as an estimate. This would give

$$
\begin{aligned}
&Pr(a \ \rightsquigarrow \ b, \ \ a \in par(N_b) \mid b \in N_b) \\
&\qquad \approx \left[\frac{1}{n_{\mathcal{P}}} \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right)\right] \times \frac{m_{\mathcal{P}}}{m_{\mathcal{A}}} \\
&\qquad = \frac{\overline{n}}{m_{\mathcal{A}}} \ \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right) \qquad (13)
\end{aligned}
$$

where $\overline{n} = m_{\mathcal{P}}/n_{\mathcal{P}}$ is the average number of concepts per node in $par(N_b)$. Alternatively, if $\mathcal{G}$ is a local graph under consideration (or the whole DAG) and $m_{\mathcal{G}}$ the total number of (not necessarily unique) concepts, one might use the ratio $m_{\mathcal{P}}/m_{\mathcal{G}}$ instead to give

$$
Pr(a \ \rightsquigarrow \ b, \ \ a \in par(N_b) \mid b \in N_b) \approx \frac{\overline{n}}{m_{\mathcal{G}}} \ \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i}\right) \qquad (14)
$$

Another possible choice is to use only the ratio of the number of parent nodes, $n_{\mathcal{P}}$ to the total number of nodes in the graph $\mathcal{G}$, say $n_{\mathcal{G}}$. The estimate would then be:

$$Pr(a \rightsquigarrow b, \quad a \in par(N_b) \mid b \in N_b) \approx \frac{1}{n_{\mathcal{G}}} \sum_{i=1}^{n_{\mathcal{P}}} \left(\frac{\mathcal{I}_{N_i}(a)}{n_i} \right) \tag{15}$$

Equations (13), (14), and (15) each provide a rough estimate of $Pr(a \rightsquigarrow b, \quad a \in par(N_b) \mid b \in N_b)$. The principal distinctions are twofold: first between counting concepts as in (14) or nodes as in (15); the second depends on the choice of the graph $\mathcal{G}$ for comparison, the ancestor graph $\mathcal{A}$ as in (13) or any other "local graph". Any of these choices gives an estimate of the first term of equation (4). The quality of the estimate depends on both the quality of this choice and, perhaps more importantly, on the applicability of the simplifying equality (7).

The second term of (4) will now be treated in much the same way as the first – we begin by splitting the probability using inclusion exclusion of sets as in equation (6) and then apply conditional probability rules to each term as in (5). In this case, however, we now let A_I denote the event $a \in anc(N_I)$, where N_I is now $N_I = \bigcup_{i \in I} N_i$ and I denotes a set of indices $I \subseteq \{1, 2, \cdots, n_{\mathcal{P}}\}$. This yields

$$\begin{aligned} & Pr(a \rightsquigarrow b, \quad a \in anc(par(N_b)) \mid b \in N_b) \\ & = Pr(a \rightsquigarrow b, \quad a \in \bigcup_{i=1}^{n_{\mathcal{P}}} anc(N_i) \mid b \in N_b) \\ & = \sum_{k=1}^{n_{\mathcal{P}}} (-1)^{k-1} \sum_{|I|=k} Pr(a \rightsquigarrow b, \quad A_I \mid b \in N_b) \ . \end{aligned} \tag{16}$$

Again, dropping all higher order terms as having significantly smaller contributions, we have

$$\begin{aligned} & Pr(a \rightsquigarrow b, \quad a \in anc(par(N_b)) \mid b \in N_b) \\ & \approx \sum_{i=1}^{n_{\mathcal{P}}} Pr(a \rightsquigarrow b, \ A_{\{i\}} \mid b \in N_b) \qquad (17) \\ & = \sum_{i=1}^{n_{\mathcal{P}}} Pr(a \rightsquigarrow b, \ a \in anc(N_i) \mid b \in N_b) \qquad (18) \end{aligned}$$

The ith term in the sum has a structure that is similar to that of the left hand side of (4) except that in (18) we are not considering the ancestors of N_b (which appears in the conditioning event), but rather the ancestors of one of the parent nodes, N_i, of N_b. Following identical reasoning we have

$$Pr(a \rightsquigarrow b, \ a \in anc(N_i) \mid b \in N_b)$$

$$
\begin{aligned}
= \quad & Pr(a \rightsquigarrow b, \quad a \in par(N_i) \mid b \in N_b) \\
& \quad + Pr(a \rightsquigarrow b, \quad a \in anc(par(N_i)) \mid b \in N_b) \\
& \quad - Pr(a \rightsquigarrow b, \quad a \in par(N_i) \cap anc(par(N_i)) \mid b \in N_b) \qquad (19)
\end{aligned}
$$

the first term of which can be written as

$$
\begin{aligned}
Pr(a \ & \rightsquigarrow \ b, \quad a \in par(N_i) \mid b \in N_b) \\
& = \quad Pr(a \rightsquigarrow b \mid a \in par(N_i), \ b \in N_b) \\
& \qquad \times Pr(a \in par(N_i) \mid b \in N_b). \qquad (20)
\end{aligned}
$$

As before, the first term in (20) is separated and we follow the same reasoning used in equations (6), (8) and (9). Following this we have

$$
\begin{aligned}
& Pr(a \ \rightsquigarrow \ b \mid a \in par(N_i), \quad b \in N_b)) \\
& \approx \quad \sum_{j=1}^{|par(N_i)|} Pr(a \rightsquigarrow b \mid a \in N_{i,j} \ , \ b \in N_b) \times Pr(a \in N_{i,j} \mid b \in N_b) \\
& \approx \quad \sum_{j=1}^{|par(N_i)|} \left(\frac{\mathcal{I}_{N_{i,j}}(a)}{|N_{i,j}|} \right) \times \left(\frac{1}{n_{\mathcal{P}}} \times \frac{1}{|par(N_i)|} \right) \qquad (21)
\end{aligned}
$$

where $N_{i,j}$ is the j'th parent node of N_i (the i'th parent node of N_b), $|N_{i,j}|$ the number of concepts in $N_{i,j}$, $|par(N_i)|$ the number of parent nodes (or in-degree) of N_i, $n_{\mathcal{P}} = |par(N_b)|$ as before, and $\mathcal{I}_{N_{i,j}}(a)$ is an indicator function that is 1 when $a \in N_{i,j}$ and 0 otherwise.

The second term of (20) should be analogous to the choice made in (13), (14), or (15). In the present case, this would correspond, respectively, to using ratios .... Et cetera.

With a little more work a general formula can be written down.

Note that this approach takes into account both path distance and feature matching.

Put it all together

Continue in this way until the graph is exhausted (no more ancestors).

Could choose a local graph, instead of the whole graph (e.g. at most 4 levels of ancestors or path-lengths back).

Iterate over every node in the whole graph and over every concept in that node. Average to get final $Pr(a \rightsquigarrow b)$ for every pair of concepts (a, b). Note $\rightsquigarrow$ is not symmetric.

Possible measures of semantic similarity (a to b) include:

- $p_{a,b} = Pr(a \rightsquigarrow b)$
- information theoretic measure like $info(a, b) = -p_{a,b} \log(p_{a,b})$

- For the ith concept (in some order), construct the vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n})^T$ where n is the total number of concepts and $x_{i,j} = p_{i,j}$ (or $= info(i,j)$). Form Euclidean distance matrices based on these vectors and use the distances as a measure of *dis*similarity.

Concepts could be clustered using any one of these similarities.