IC Based Set Similarity

- Extend the notion of minimum common ancestor (λ) to sets of terms as

$$\Lambda(S_i, S_j) = \bigsqcup_{c_k \in S_i, c_l \in S_j} \lambda(c_k, c_l)$$

- Information content of a set is defined as:

$$\rho_I(S_i, S_j) = I(\Lambda(S_i, S_j)) = -\log_2\left(\frac{G_{\Lambda(S_i,S_j)}}{G_r}\right)$$

Where $G_{\Lambda(S_i,S_j)} = \bigcap_{c_k \in \Lambda(S_i,S_j)} G_{c_k}$ is set associated with all terms in MCA of S_i, S_j

- This satisfies all 4 properties, can be extended

$$\rho_L = \frac{2\rho_I(S_i, S_j)}{\rho_I(S_i, S_i) + \rho_I(S_i, S_j)} \qquad \rho_{JC} = 1/(\rho_I(S_i, S_i) + \rho_I(S_j, S_j) - 2 * \rho_I(S_i, S_j) + 1)$$