Methods based on Information Content

Guangchuang Yu

February 20, 2009

The GOSemSim package implemented four methods which are based on information content were proposed by Resnik[Philip, 1999], Jiang[Jiang and Conrath, 1997], Lin[Lin, 1998] and Schlicker[Schlicker et al., 2006] respectively.

Information content is defined as frequency of each term occurs in the corpus. We used Bioconductor package GO.db to calculate the information content. The information content will update biannually as GO.db updated.

Given the information content, we applied the four measures to estimate the semantic similarity between terms.

As GO allows multiple parents for each concept, two terms can share parents by multiple paths. We take the mininmum p(t), where there is more than on shared parents. We defined p_{ms} as :

$p_{ms}(t1, t2) = \min_{t \in S(t1,t2)} \{p(t)\})$

where S(t1,t2) is the set of parent terms shared by t1 and t2.

The first method Resnik[Philip, 1999] is defined as:

$sim(t1, t2) = -\ln p_{ms}(t1, t2)$

The second method Lin[Lin, 1998] is defined as:

$$sim(t1, t2) = \frac{2 \times \ln(p_{ms}(t1, t2))}{\ln p(t1) + \ln p(c2)}$$

The third method Rel[Schlicker et al., 2006] combine Resnik's and Lin's method is defined as:

$$sim = \frac{2 \times \ln p_{ms}(t1, t2)}{\ln p(t1) + \ln p(p2)}$$

The last method Jiang[Jiang and Conrath, 1997] is defined as:

$sim(t1, t2) = 2 \times \ln p_{ms}(t1, t2) - \ln p(t1) - \ln p(p2)$

References

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of 10th International Conference on Research In Computational Linguistics*, 1997. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:cmp-lg/9709008`.

Dekang Lin. An Information-Theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296—304, 1998. doi: 10.1.1.55.1832. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.1832`.

Resnik Philip. Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999. URL `http://nzdl.sadl.uleth.ca/cgi-bin/library?e=d-00000-00--off-0jair-00-0-0-10-0--0--0prompt-10--4-----0-1l-11-en-50--20-about--00-0-1-00-0-0-11-1-0utfZz-8-00&cl=CL3.1.11&d=jair-514&x=1`.

Andreas Schlicker, Francisco S Domingues, JÃűrg RahnenfÃijhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302, 2006. ISSN 1471-2105. doi: 1471-2105-7-302. PMID: 16776819.