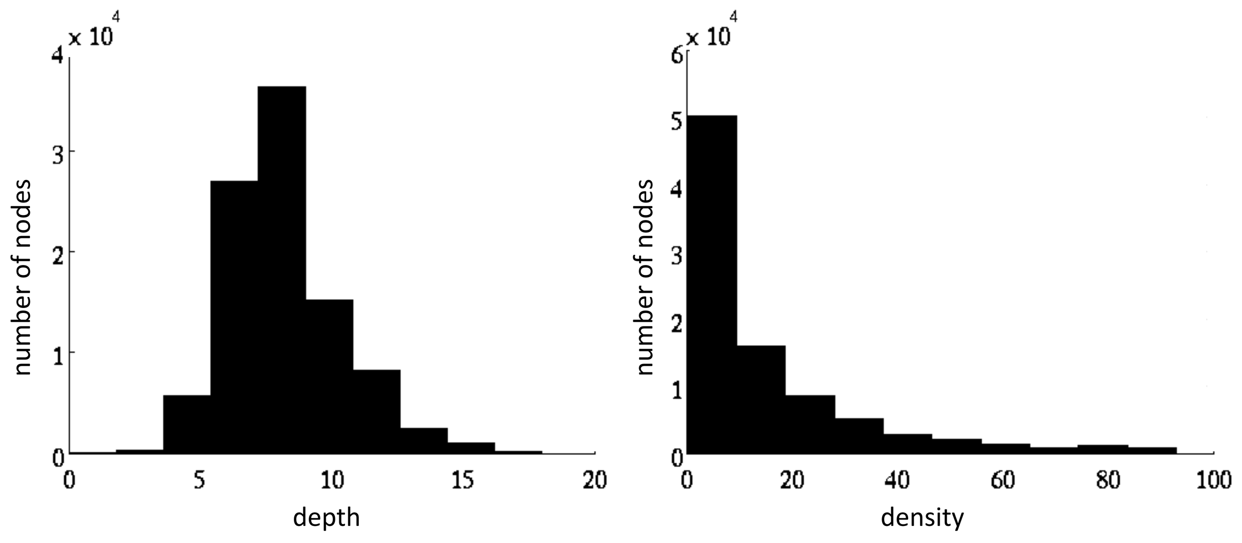
Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures

Tong Wang and Graeme Hirst
University of Toronto
{tong, gh}@cs.toronto.edu

1. Motivation

- Depth and density
  - Quantitative features of taxonomic structures
  - Used in *WordNet*-based semantic similarity measures (Sussna 1993; Wu and Palmer 1994; Jiang and Conrath 1997)
- Taxonomic depth and density as integers
  - Relative concepts used in absolute terms
  - Superimposing linearity on non-linear quantities

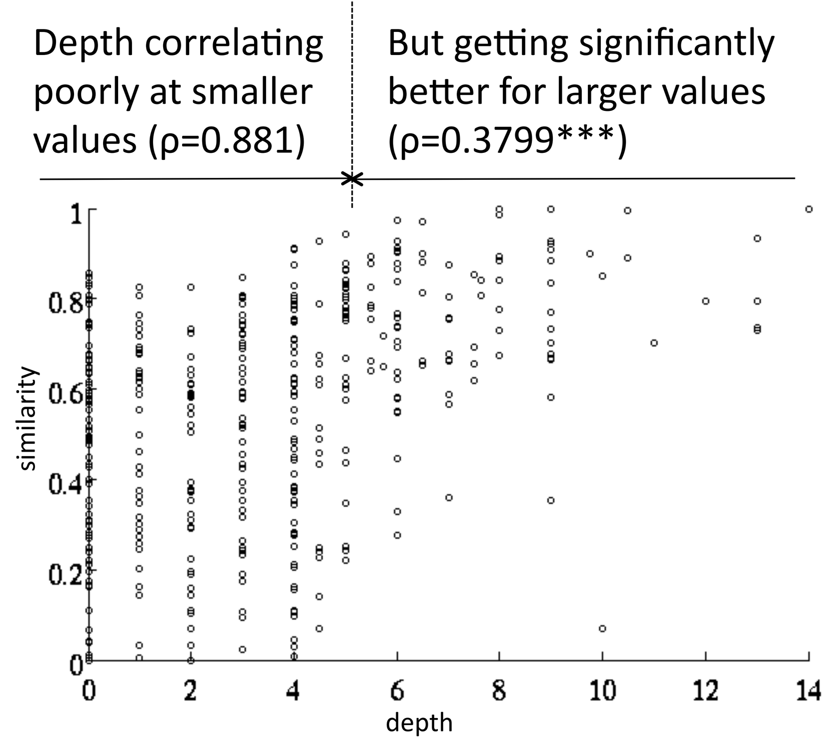


Distribution of depth and density of nodes in WordNet

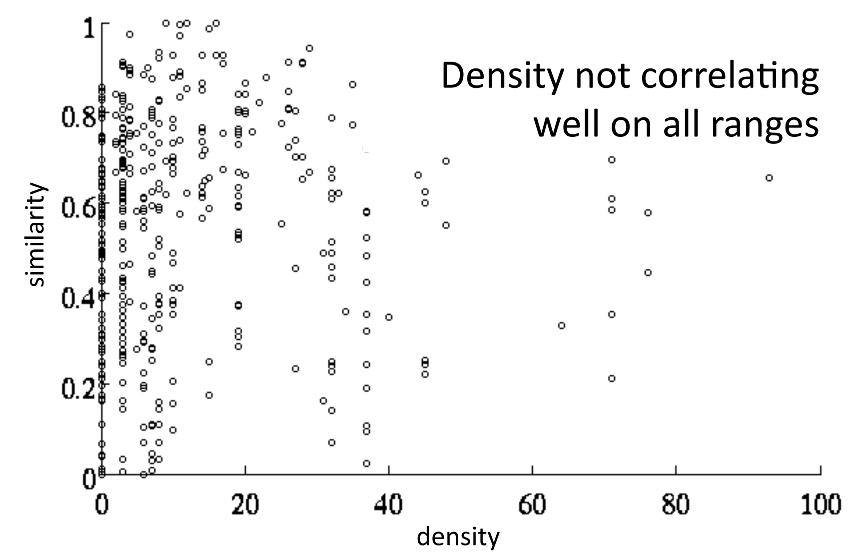
2. Limitations of current definitions

Are depth and density actually related to semantic similarity?

- Gold standard of semantic similarity: human judgment of lexical semantic similarity from psycholinguistic experiments (the *MC*, *RG*, and *FG* datasets)
- Correlation between depth/density and human judgment of similarity: measured by Spearman's Rho



	MC	*RG*	*FG*
dep	0.7056***	0.6909***	0.3701***
den	0.2268	0.2660*	0.1033



*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

3. Redefining depth and density

Depth

- "Re-curving" a depth value from integer to its cumulative distribution curve:

$$dep_u(c) = \frac{\Sigma_{c' \in \mathrm{WN}} |\{c' : dep(c') \le dep(c)\}|}{|\mathrm{WN}|}$$

	MC	*RG*	*FG*
original	0.7056***	0.6909***	0.3701***
re-curving	0.7201***	0.6798***	0.3751***

Density

- Re-curving does not work well due to Zipfian distribution
- Incorporating *inheritance* in the definition of density:

$$den_i(r) = 0$$

$$den_i(c) = \frac{\Sigma_{h \in hyper(c)} den_i(h)}{|hyper(c)|} + den(c)$$

	MC	*RG*	*FG*
before	0.2268	0.2660*	0.1033
re-curving	0.2268	0.2660*	0.1019
inheritance	0.7338***	0.6751***	0.3445***

4. Experiments

Do the new definitions contribute to better similarity measures?

- Same similarity gold standard as in section 2
- Re-implement some similarity measures that use depth and density
- Replace depth and density definition to see if there is performance improvement (in terms of correlation with gold standard)
  - Wu and Palmer (1994) – depth only

$$sim(c_1, c_2) = \frac{2 \cdot dep^{\alpha}(c)}{len(c_1, c) + len(c_2, c) + 2 \cdot dep^{\alpha}(c)}$$

  - Jiang and Conrath (1997) – depth and density

$$w(c,p) = \left(\frac{dep(p)+1}{dep(p)}\right)^{\alpha} \times \left[\beta + (1-\beta)\frac{\bar{E}}{den(p)}\right] \times [IC(c) - IC(p)]T(c,p)$$

- Parameterization of depth/density components enables investigation of numerical stability

5. Results and conclusions

- Refining depth in Wu and Palmer (1994)

	Best[1]			Average[2]		
	MC	*RG*	*FG*	*MC*	*RG*	*FG*
dep	0.7671	0.7824	0.3773	0.7612	0.7686	0.3660
dep_u	0.7824	0.7912	0.3946	0.7798	0.7810	0.3787

1. Highest correlation with human judgment among all parameters
2. Average correlation across parameters (indicating numerical stability of parameterization)

- Refining depth and density in Jiang and Conrath (1997)

	Best			Average		
	MC	*RG*	*FG*	*MC*	*RG*	*FG*
dep, den	0.7875	0.8111	0.3720	0.7689	0.7990	0.3583
dep_u, den	0.8009	0.8181	0.3804	0.7885	0.8032	0.3669
dep, den_i	0.7882	0.8199	0.3803	0.7863	0.8102	0.3689
dep_u, den_i	0.8065	0.8202	0.3818	0.8189	0.8194	0.3715

- **Both new definitions improve similarity measure performance and numerical stability**
- **Best result achieved using both new definitions together**

References and acknowledgement

This study was inspired by lectures given by Gerald Penn, and was financially supported by the Natural Sciences and Engineering Research Council of Canada. The poster is based on a design by Julian Brooke.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414. ACM, 2001.

Jay Jiang and David Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, 19–33, 1997.

George Miller and Walter Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

Herbert Rubenstein and John Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*, pages 67–74. ACM, 1993.

Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.