Similarity of hierarchical relationships in news portal datasets

Gergely Tibély[1], David Sousa-Rodrigues[2], Péter Pollner[3], Gergely Palla[3]

[1]Dept. of Biological Physics, Eötvös University, H-1117 Budapest, Hungary
[2]The Design Group, Faculty of Maths, Computing and Technology, The Open University, Walton Hall, Milton Keynes, MK7 6AA United Kingdom
[3]Statistical and Biological Physics Research Group of HAS, H-1117 Budapest, Hungary

Online datasets are becoming more and more prevalent. Besides presenting a huge amount of data available for analysis, they also bring the possibility of a new method for categorisation: tagging. Tags circumvent several limitations of the traditional expert-based hierarchical categorisation: they can be provided by several, non-expert users, more than one tag may be appointed to the same item, and tags do not require relations to be defined among them. Still, our notions are frequently organised in a hierarchical way, which is expected to appear in the way the tags are used. Indeed, in [1,2,3] we showed that tags various datasets like photo sharing websites, proteins, or even scientific publications can be given a meaningful hierarchy using network-based statistical methods.
Here, we introduce a new method to compare hierarchies. Usually, methods rely on comparing the ancestor-descendant relationships of two hierarchies [4,5,6], providing a vertical view on the similarity. We propose to compare the composition of branches, too, as a horizontal view, complementing the former method. The corresponding formula can be adjusted for random expectation values, thus providing a 0-1 measure of similarity, 0 occurring in independent, uncorrelated cases.
As an example, we apply the new measure to four news portal datasets (Spiegel Online, The Guardian, The New York Times, The Australian), where news items are tagged. The datasets show interesting differences, like the number of connected components in the constructed tag hierarchies, while maintaining a significant amount of similarity.

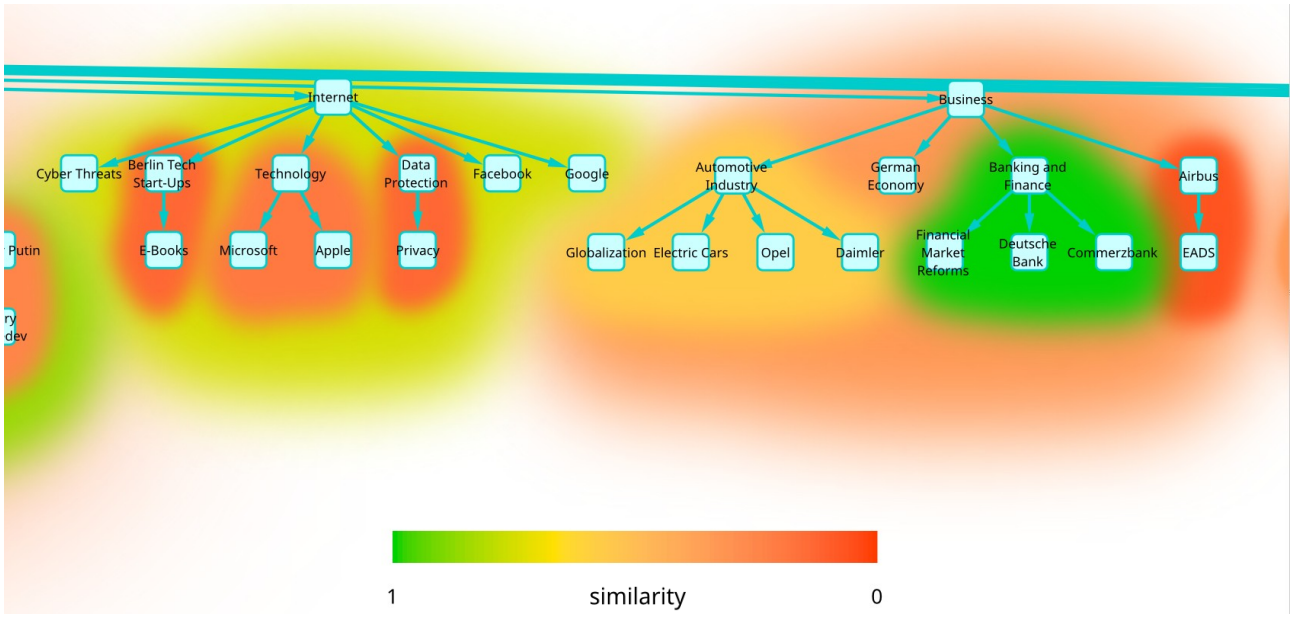

Figure 1: a sample from the constructed tag hierarchy of Spiegel Online. Colour-coded values show similarities to corresponding branches in The Guardian's tag hierarchy. 0 is set to the expected value of a random 0-model.

[1] G. Tibély, P. Pollner, T. Vicsek, G. Palla: Extracting Tag Hierarchies, *PloS ONE* **8**, e84133 (2013).
[2] G. Palla, G. Tibély, E. Mones, P. Pollner, T. Vicsek: Hierarchical networks of scientific journals, *Palgrave Communications* **1**, 15016 (2015).
[3] G. Tibély, P. Pollner, G. Palla: Comparing the hierarchy of author given tags and repository given tags in a large document archive, arXiv:1507.04930
[4] F. J. Brandenburg, A. Gleißner, A. Hofmeier. The nearest neighbor spearman footrule distance for bucket, interval, and partial orders. *J. Comb. Optim.*, **26**:310–332 (2013).
[5] R. Brüggemann, E. Halfon, G. Welzl, K. Voigt, C. E. W. Steinberg: Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests, *J. Chem. Inf. Comput. Sci*. **41**, 918 (2001).
[6] H. Era, K. Ogawa, M. Tsuchiya: On transformations of posets which have the same bound graph, *Discr. Math.* **235**, 215 (2001).