Automated Knowledge Hierarchy Assessment

G. Nayak
Univ. of Minnesota, USA
nayak013@umn.edu

S. Dutta D. Ajwani P. Nicholson A. Sala
Nokia Bell Labs, Ireland
{firstname.lastname}@nokia-bell-labs.com

Abstract

Automated construction of knowledge hierarchies is gaining increasing attention to tackle the infeasibility of manually extracting and semantically linking millions of concepts. With the evolution of knowledge hierarchies, there is a need for measures to assess its temporal evolution, quantifying the similarities between different versions and identifying the relative growth of different subgraphs in the knowledge hierarchy. This work proposes a principled and scalable similarity measure, based on Katz similarity between concept nodes, for comparing knowledge hierarchies, modeled as generic Directed Acyclic Graphs (DAGs).

1 Contribution

Large knowledge repositories like DBpedia [6] represent concepts and relations as hierarchies expressing semantic connections via parent-child or *hypernym-hyponym* edges. These hierarchies (typically represented as DAGs) form the backbone of semantic search, personalization, recommendation and textual entailment, enabling easy navigation across concepts for information linking [1, 7, 4, 3] As the knowledge hierarchies evolve, there is a need for taxonomy evaluation, i.e., comparing hierarchies and quantifying their similarity. Intuitively, a principled similarity measure should demonstrate the following properties: (1) *Sensitivity to Concept Hierarchy*, (2) *Proximity of Least Common Ancestor*, and (3) *Importance of Relationship*. Additionally, the practicality of assessing large hierarchies imposes the factors of *scalability*, *tunability*, and *interpretability* for diverse applications.

Proposed Measure: To model similarities between DAGs, we adapt the *Katz similarity measure* [5], which captures *multiple short directed paths* between vertices, providing a good indicator of semantic subsumption. We define the notion of *Katz Similarity Vector* representing each concept node to obtain the **Katz Graph Similarity** measure. We theoretically show that our proposed measure conforms to the above salient properties.

Further, for scenarios where computing the Katz Graph Similarity might be expensive, we propose a faster approximate variant, the **Grouped Katz Similarity** measure. We also show that the *Katz Index* [5], a centrality measure capturing the influence of a vertex in a graph, is in fact a special case of the Grouped Katz Similarity (with 1 group) for measuring acyclic graph similarity. This provides the **Katz Index Graph Similarity** measure, a *linear time* variant for computing the similarity between hierarchies; albeit with a slight loss in structural information due to aggressive grouping. Interestingly, the above similarity measures demonstrates monotonic behavior with respect to the degree of structural difference between the DAGs.

We performed large-scale experiments on different sub-hierarchies of the DBpedia knowledge graph with varying the number of concept nodes. We demonstrated that our proposed measures were not only scalable (with run-time improvement of ~10000× compared to state-of-the-art Fowlkes-Mallows measure [2]), but also captures the structure and logical subsumption of concept relations in hierarchies. Additionally, we show that our measures are able to assess the *temporal evolution of concepts* in correspondence with category disruptions.

In: Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop on Data Search (DATA:SEARCH18). Co-located with SIGIR 2018, Ann Arbor, Michigan, USA – 12 July 2018, published at http://ceur-ws.org

References

[1] Fensel, D.: Spinning the Semantic Web: Bringing the World Wide Web to its full potential. MIT Press (2005)

[2] Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. American Statistical Association **78** (1983) 553–569

[3] Geffet, M., Dagan, I.: The distributional inclusion hypotheses and lexical entailment. In: ACL. (2005) 107–114

[4] Harabagiu, S.M., Maiorano, S.J., Paşca, A., M.: Open-domain textual question answering techniques. Natural Language Engineering **9**(3) (2003) 231–267

[5] Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1) (1953) 39–43

[6] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2) (2015) 167–195

[7] Maedche, A., Staab, S.: Measuring similarity between ontologies. In: EKAW. (2002) 251–263