SimCat: An Entity Similarity Measure for Heterogeneous Knowledge Graph with Categories

Arup Choudhury
Freelance Professional
xarup@hotmail.com

Shrey Sharma
Epic Systems
shreysharmaiit@gmail.com

Pabitra Mitra
IIT Kharagpur
pabitra@gmail.com

Cyril Sebastian
Yahoo Inc
write2cyril@gmail.com

Shrikant S. Naidu
Microsoft India

Muthusamy Chelliah
Flipkart
muthusamy.c@flipkart.com

ABSTRACT

Establishing similarity between heterogeneous entities in a complex knowledge graph is a challenging task due to the unrestricted nature of categories and relation types. In large graphs, the semantic roles of relation types and entity categories are strongly interdependent at a statistical scale and help in defining a key component of entity similarity measures. In this paper we define a measure that incorporates category information as well as aggregated relationship graph structures. Experimental evaluation in terms of entity retrieval performance on a large subset of the Freebase graph establishes the efficacy of the system.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION

A significant development in semantic web has been the evolution of large scale graph-shaped knowledge bases, referred as knowledge graphs or heterogeneous information networks. Such knowledge graphs typically consist of a large collection of entities, certain entity attributes, and a set of heterogeneous relations among the entities. The entities are also categorized in a flat or hierarchical taxonomy, with possibly multiple category tags associated with a single entity. Examples of such web scale knowledge graphs include, Freebase [1], Yago etc.

Structural context of an entity in a knowledge graphs is the natural way of characterizing semantic similarity of entities. SimRank [2] is a recursive measure based on the difference in neighborhood of entities. Various extensions to SimRank has been proposed in literature [4]. Recent research have focused on exploiting the heterogeneous link and entity type informations in defining object similarity [6, 5].

Approaches to knowledge graph mining discussed above characterize entities by their categories and the heterogeneous relations they participate in with other entities. However, the entity category and the relation it participates in are strongly interdependent. In this article we propose an entity similarity measure over a heterogeneous knowledge graph having both typed entities and links.

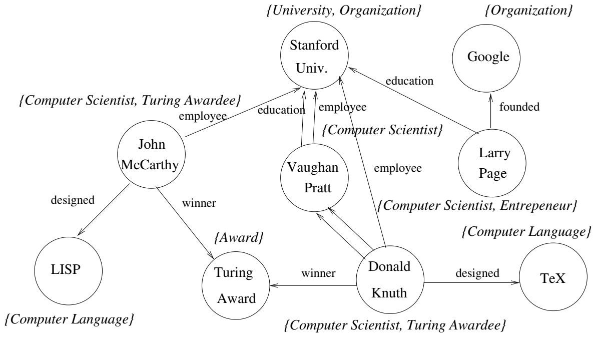

Figure 1: A heterogeneous knowledge graph G.

2. KNOWLEDGE GRAPHS

We define below some objects useful for definition of the entity similarity measure.

Definition 1. The *heterogeneous knowledge graph* is defined as a directed graph $G = (\mathcal{V}, \mathcal{E})$, with an entity category mapping $\tau : \mathcal{V} \to \mathcal{C}$ and relation type mapping $\rho : \mathcal{E} \to \mathcal{R}$. Each entity $v \in \mathcal{V}$ belongs to multiple categories $\tau(v) \subset \mathcal{C}$, similarly each edge $e \in \mathcal{E}$ may be of multiple types $\rho(e) \subset \mathcal{R}$.

An example knowledge graph is shown in Figure 1. The Freebase knowledge graph contains about 20 million entities, 100 million relations.

A directed edge $v_1 \to v_2$ in the heterogeneous knowledge graph may be represented by an *edge vector* $(e_{v_1 \to v_2})$ with dimension as the number of distinct relation types ($|\mathcal{R}|$) existing in the knowledge graph. The value of a component is 1 if that relation type relates entities $v_1 \to v_2$, and 0 otherwise. For example the edge vector $e_{DonaldKnuth \to VaughanPratt}$ is $\{advisee : 1, co - author : 1, \ldots\}$. A category vector for an entity may also be defined similarly.

Definition 2. The *participating relation vector* $PRV(v)$ for an entity v has the same dimension as the number of

CODS '15, March 18 - 21, 2015, Bangalore, India
Copyright 2015 ACM 978-1-4503-3436-5/15/03 $15.00
http://dx.doi.org/10.1145/2732587.2732604.

distinct relation types $|\mathcal{R}|$. The value of a component of the vector is the count of the number of other entities that v is related to via the corresponding relation. In other words, $PRV(v)$ is the sum of all the directed edge vectors of the knowledge graph emanating from v.

In the graph of Figure 1, $PRV(DonaldKnuth) = \{employee : 1, advisee : 1, award : 1, language_designed : 1, founded : 0, \ldots\}$. Next, we obtain the category graph G', by transforming the knowledge graph G.

Definition 3. The *category graph* is defined as a directed graph $G' = (\mathcal{C}, \mathcal{E}')$, where the vertices are categories and the edges are vector valued quantities. The category edge vector $(e'_{c_1 \to c_2})$ between categories $c_1, c_2 \in \mathcal{C}$ is obtained by summing all the edge vectors $e_{v_i \to v_j}$ of the knowledge graph G such that $v_i \in c_1$ and $v_j \in c_2$.

The category and entity graphs are dual in the following sense. In knowledge graphs, the categories represent subgraph of entity nodes tagged with that particular category. For a category c, we represent this *induced knowledge subgraph* as $G(c)$. On the other hand, in category graphs, entities correspond to subgraph of category nodes to which that entity belongs to. For an entity v, we denote this *induced category subgraph* as $G'(v)$.

The category graph contains vector valued edges representing all existing types of relations between member entities. We next consider only the relation types that are relevant to a specific entity and project the category graph to a scalar edge weighted graph denoted as the projected induced category subgraph $G''(v)$.

Definition 4. The *projected induced category graph* $G''(v)$ for an entity v is derived from the induced category subgraph $G'(v)$ defined previously. $G''(v)$ has the same node and directed edge structure as $G'(v)$, but its edges have scalar weights. The scalar edge weights of $G''(v)$ are computed by evaluating the cosine similarity of the corresponding vector valued edge of $G'(v)$ with the participating relation vector $PRV(v)$.

3. ENTITY SIMILARITY MEASURE

Similarity between two entities is determined by the following four factors between entity pairs - overlap of the category tags, path similarity, aggregated nature of interrelation of categories, and, the participating relations profiles.

In order to integrate all the four factors, we define similarity $Simcat(v_1, v_2)$ between two entities v_1 and v_2 as the overlap between the projected subgraphs $G''(v)$ they induce in the category graph G' defined previously.

$$Simcat(v_1, v_2) = \mathcal{SG}(G''(v_1), G''(v_2)), \qquad (1)$$

where $G''(v)$ is the projected induced category subgraph of entity v in the category graph G'. We use Cauchy-Schwarz divergence as a subgraph overlap measure $\mathcal{SG}_{CS}$. Let $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ be to subgraphs of the graph $G(V, E)$. A graph interpretation of the Cauchy-Schwarz divergence [3] is defined below. Edges of either direction as well as self edges are used in computing the cut weight.

$$\mathcal{SG}_{CS}(G_1, G_2) = \frac{|\mathrm{Cut}(G_1, G_2)|}{\sqrt{|V_1||V_2|}}. \qquad (2)$$

Table 1: Top-20 retrieval accuracy (%) on Freebase entities using various similarity measures

Entity	*SimCat*	*PathSim*
Pixar	100.0	95.0
Roger Federer	100.0	90.0
American Idol	100.0	80.0
Golden Gate Bridge	100.0	60.0
Walmart	100.0	25.0
The Pentagon	75.0	90.0
Mean	95.8	73.3

4. EXPERIMENTAL RESULTS

We use an instance of the freebase graph consisting of about 1.9 million entities, 1800 categories and 100 million relations. We evaluate the efficacy of entity similarity measures by comparing its performance in entity search task. In Table 1 we present the similar entity retrieval performance of the proposed *SimCat* measure using Cauchy-Schwarz divergence score for six sample queries. We also present retrieval performance of the *PathSim* algorithm [6]. It is seen form the table that *SimCat* provides a higher accuracy than *PathSim* in most queries.

5. CONCLUSIONS AND DISCUSSION

We present a method for computing similarity among entities in heterogeneous knowledge graphs where entities as well as relations are marked with multiple types. This is achieved by a combination of bag of word representation of the relation vectors and transformations of the original knowledge graph to an aggregated category graph.

6. REFERENCES

[1] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (2008), pp. 1247–1250.

[2] Jeh, G., and Widom, J. SimRank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 538–543.

[3] Jenssen, R., Principe, J., Erdogmus, D., and Eltoft, T. The Cauchy–Schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute 343*, 6 (2006), 614–629.

[4] Liu, H., He, J., Zhu, D., Ling, C., and Du, X. Measuring similarity based on link information: A comparative study. *IEEE Trans. Knowl. Data Eng. 25*, 12 (2013), 2823–2840.

[5] Shi, C., Kong, X., Yu, P., Xie, S., and Wu, B. Relevance search in heterogeneous networks. In *Proceedings of the 15th International Conference on Extending Database Technology* (2012), pp. 180–191.

[6] Sun, Y., Han, J., Yan, X., Yu, P. S., and Wu, T. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *International Conference on Very Large Databases* (2011).