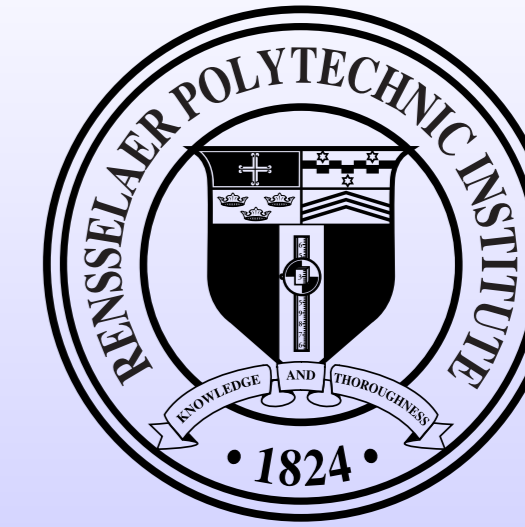


CENTRALITY-BASED RANKING FOR RDF NODES

Alvaro Graves, Sibel Adalı, James Hendler
{agraves, sibel, hendler}@cs.rpi.edu
Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY 12180



Introduction

This work focuses on finding ways to sort nodes of an RDF graph in a “relevant” order. By relevant we mean some reasonable order, generally accepted by people with expertise in the information represented in the graph. We base our method on what is called the **closeness centrality**, which can be defined as

$$C_i = \frac{r_i}{\sum_{j \in A_k \setminus \{i\}} d(i,j)}$$

where

- d is a distance function between nodes,
- A_k is the connected component that i belongs to, and
- r_i is the *reachability* of the node i , the number of nodes in A_k .

Finding the centrality for every node in a graph is equivalent to solving the *All-Pairs Shortest Path* problem. Our current solution is based on Dijkstra’s algorithm and has complexity $O(nm \log(n))$ where n is the number of nodes and m is the number of edges. This approach allows us to solve the problem in a parallelized platform improving the speed.

Moreover, we would like to rank nodes in the bigger connected component higher, hence we scale c_i with $\log(r_i)$.

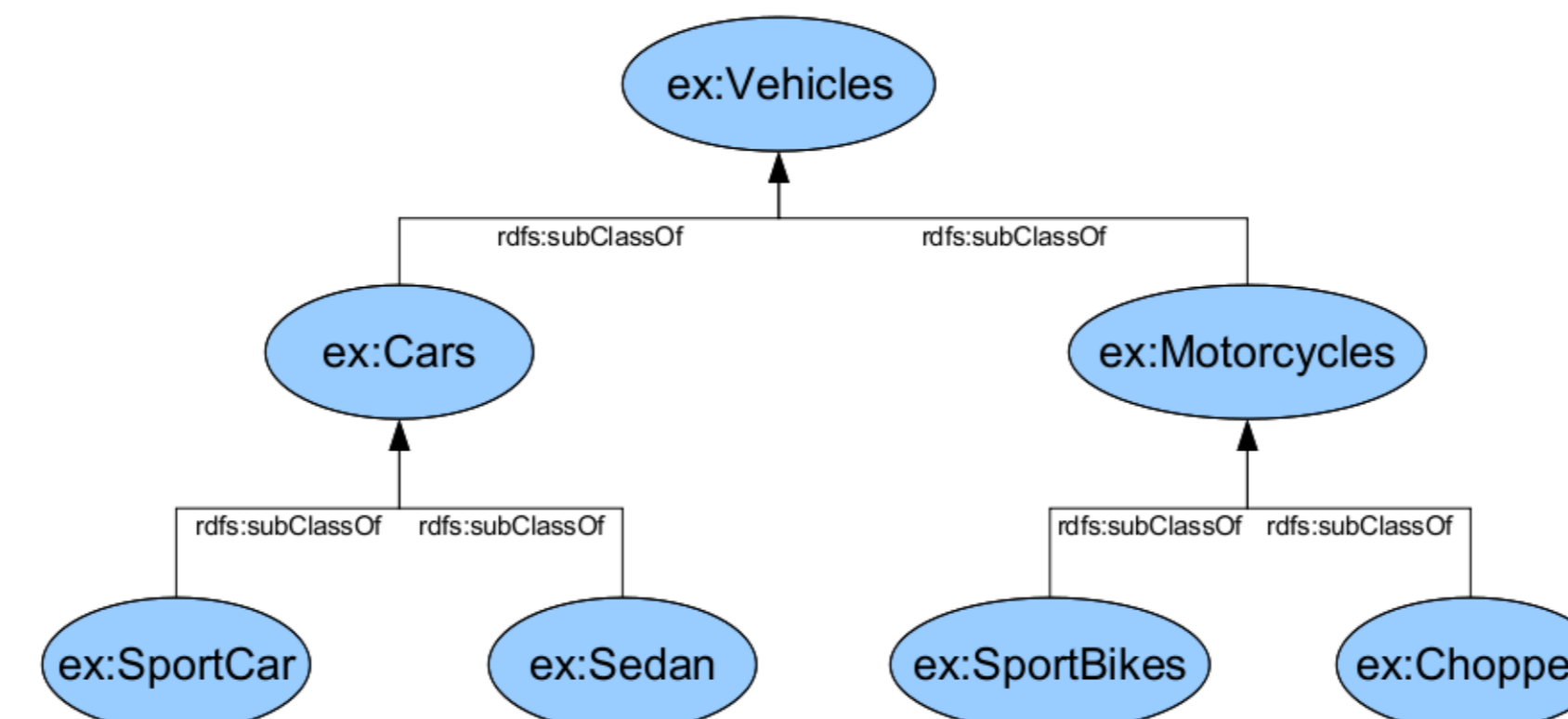
Weights and distance functions

The distance d of a path is given by the sum of the edge weights along this path. We consider two different types of weights:

- **Constant:** Each edge has weight 1. This basic distance function does not take into account the semantics of the graph.
- **Predicate-based weights:** Edges weights are determined as a function of the predicate they contain. An example function is given by: let $p(u,v)$ be the predicate associated with edge (u,v) , then $d(u,v) = \frac{|p(u,v)|}{|E|}$, where E is the set of all the edges.

A simple example

The simplest type of graph to rank is a **taxonomy**. Which are the most relevant nodes?



Future Work

- **Scalability:** Currently we are running our algorithm for graphs with over 1 million nodes. We are investigating different methods to make the method more scalable.
- **Customizability:** We would like to study the effect of different weight functions and different models of centrality based ranking on real data.
- **Stability:** The study of the stability of the rankings with respect to the properties of the different data sets and the relation with the degree distribution.

References

- [AMS05] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. Semrank: Ranking complex relationship search results on the semantic web. In *In Proceedings of the WWW Conference*, pages 117–127, 2005.
- [BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the ACM WWW Conference*, pages 107–117, 1998.
- [DFJ+04] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the 13th CIKM*, 2004.
- [LM05] R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. *Information Retrieval*, 8(2):245–264, 2005.

CIA Factbook

This dataset is an RDF version of the annual publication made by the CIA about the different countries in the world. It contains information about geography, demographics, history, etc. There are more than 30,000 nodes and almost 98,000 edges without considering the literals. The following table compares the ranking based on the in-degree against our centrality-based ranking for all the nodes in the graph. It is interesting that the top ranked nodes are international organizations and some globally important issues, like “Climate Change” and “Biodiversity”.

Position	In-degree top 10	Centrality top 10
1	rdf:Statement	org:IOC
2	cia:Estimate	org:WHO
3	cia:CommodityPercent	org:UN
4	cia:CountryPercent	org:ITU
5	cia:AirportBreakdown	org:UNCTAD
6	cia:Port	org:UPU
7	cia:SexRatioBreakdown	org:ICAO
8	cia:EthnicGroupPercent	org:UNESCO
9	cia:ReligionPercent	env:Biodiversity
10	cia:LanguagePercent	env:Climate_Change

Wine Ontology

The wine ontology contains 720 nodes and almost 2,000 edges without considering the literals. In this table, we show the results for weighted and non-weighted centrality ranking. We can see that the weights favor the more specific instances of the ontology over more general concepts.

Position	Weighted centrality top 10	Constant centrality top 10
1	wine:WhitehallLaneCabernetFranc	wine:Wine
2	wine:MariettaOldVinesRed	wine:Winery
3	wine:CorbansDryWhiteRiesling	wine:Region
4	wine:MountadamRiesling	wine:Dry
5	wine:StGenevieveTexasWhite	wine:WineGrape
6	wine:SelaksIceWine	wine:DessertWine
7	wine:KathrynKennedyLateral	wine:LateHarvest
8	wine:WhitehallLanePrimavera	wine:Moderate
9	wine:adjacentRegion	wine:WineColor
10	wine:ChiantiClassico	wine:EarlyHarvest

Terrorist Ontology

The terrorist ontology was developed by the Mindswap group from University of Maryland. It contains information about terrorists, events, organizations, victims and countries among others. It contains more than 11,000 nodes and 30,000 edges. This case is interesting mainly because it includes noisy data, ranging from mistyped instances to the inclusion of information not related with terrorism. Since this ontology is more focused on middle-eastern terrorists, they are better connected within the ontology and hence are highly ranked.

Position	Weighted centrality bottom 10	Weighted centrality top 10
1	Pir Farouq	Ayman al-Zawahiri
2	Ricardo Palmera	Midhat Mursi
3	Naim Kassem	Osama Bin Laden
4	Rodney Colorado	Mustafa Kamel
5	Jamal Abu Samhadena	Omar Abd al-Rahman
6	Qari Thair Yuldashev	Abu Musab al-Zarqawi
7	Shoko Asahar	Ramzi Yousef
8	Juan Jose Martinez Vega	Khalid Sheikh Mohammed
9	Erminso Cuevas Cabrera	Hafiz Mohammed Saeed
10	Tomas Molina Caracas	Huthaifa Azzam