RDivF: Diversifying Keyword Search on RDF Graphs

Nikos Bikakis [1,2] • Giorgos Giannopoulos [1,2] • John Liagouris [1,2]
Dimitrios Skoutas [2] • Theodore Dalamagas [2] • Timos Sellis [3]

[1] National Technical University of Athens, Greece
[2] ATHENA Research Center, Greece
[3] RMIT University, Australia

Intro

RDivF: RDF + Diversity

RDivF *is a **diversification framework** for keyword search on RDF data. RDivF aims at exploiting several aspects of the **RDF data model** (e.g., **resource content**, **RDF graph structure**, **schema semantics**) to answer RDF keyword queries with a set of diverse results.*

Overview

From Keywords to Diverse RDF Subgraphs

In: Keyword query & RDF Data

Out: Ranked list of diverse RDF subgraphs

Retrieval Model

- Results are defined as RDF subgraphs
- RDF properties (relations) are treaded as first-class citizens
- Structural and semantic homogeneity among nodes, edges and paths of the same graph result and heterogeneity between different graph results

Ranking Model

- Textual similarity
- Classes and properties hierarchies
- RDF/S – OWL schema semantics

Retrieval Model

Query Result

Assume an ***RDF graph*** $\mathbf{G}(\mathbf{V},\mathbf{E})$, where $\mathbf{V}$ is the set of ***vertices*** and $\mathbf{E}$ the set of ***edges***.
Let $\mathbf{q} = \{\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_m\}, k, \rho\}$ be a ***keyword query*** comprising a set of m ***terms***, a parameter k specifying the ***maximum number of results*** to be returned, and a parameter ρ that is used to restrict the ***maximum path length*** between keyword nodes.

A subgraph $\mathbf{G}_q$ of $\mathbf{G}$ is a ***query result*** of $\mathbf{q}$ over $\mathbf{G}$, iff:

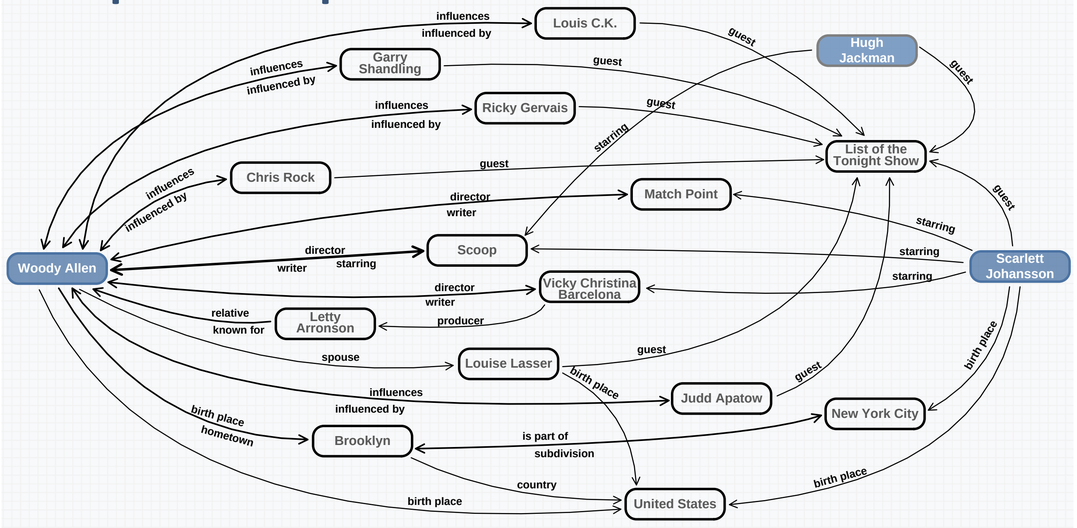
(a) for each keyword $\boldsymbol{t}$ in $\mathbf{q}$, there exists exactly one node $\boldsymbol{v}$ in $\mathbf{G}_q$ such that $\boldsymbol{v} \in \mathbf{V}_t$ (these are called ***keyword nodes***)
(b) for each pair of keyword nodes $\boldsymbol{u}, \boldsymbol{v}$ in $\mathbf{G}_q$, there exists a path between them with length at most ρ
(c) for each pair of keyword nodes $\boldsymbol{u}, \boldsymbol{v}$ in $\mathbf{G}_q$, there exists at most one direct keyword path between them
(d) each non-keyword node lies on a path connecting keyword nodes

Diversified Result Set

Let a function $\boldsymbol{r}$: $(\mathbf{G}_q, \mathbf{q}) \rightarrow [0, 1]$ that measures the ***relevance*** between the query $\mathbf{q}$ and a result $\mathbf{G}_q$.
Let function $\boldsymbol{d}$: $(\mathbf{G}_q, \mathbf{G'}_q) \rightarrow [0, 1]$ that measures the ***dissimilarity*** between two query results $\mathbf{G}_q$ and $\mathbf{G'}_q$.
The ***diversified result set*** $\mathbf{R}_k$ is a subset of the results $\mathbf{R}$ with size k that maximizes a combined measure of $\boldsymbol{r}$ and $\boldsymbol{d}$.

Motivating Example

DBpedia Snapshot



Example 1.

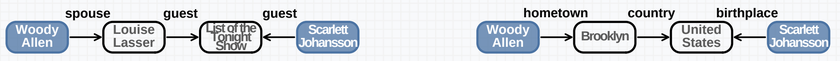
Keywords: *Woody Allen, Scarlett Johansson*

Intermediate Results

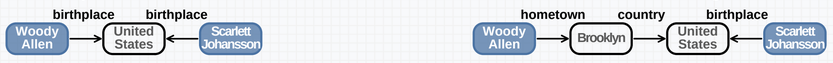
(a) Diverse paths w.r.t. properties of the searched entities

(b) Diverse paths w.r.t. the intermediate entities connecting searched entities

(c) Diverse paths w.r.t. the type (classes) of intermediate entities and the properties connecting searched entities



(d) Diverse paths w.r.t. the number of intermediate entities connecting searched entities



(e) Combination of c) and d)

(f) Non-diverse paths

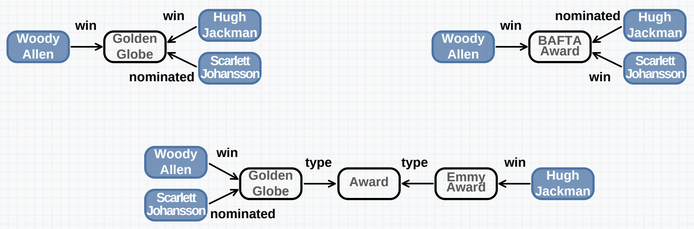
Example 2.

Keywords: *Woody Allen, Scarlett Johansson, Hugh Jackman*

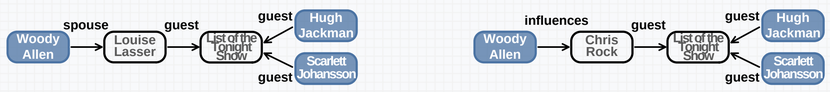
Diverse Graph Results

The specific diversification needs should determine the ranking model and the composition of the final results

(a) Return diverse results for the searched entities, w.r.t. awards they have won



(b) Return diverse results for the searched entities, w.r.t. the people they are related to



(c) Return diverse results for the searched entities