Raytheon BBN Technologies

The Business of Innovation

Finding and Explaining Similarities in Linked Data

Jeff Sherman
Catherine Olsson
Plamen Petrov
Andrew Perez-Lopez

STIDS 2011
November 17, 2011

Raytheon
BBN Technologies

Outline

Raytheon
BBN Technologies

- Motivation
- Similarity overview, related work
- Extensions for similarity in linked data
- Preliminary results
- Future work
- Conclusions

Motivation

Raytheon
BBN Technologies

- Linked data is becoming increasingly pervasive in the intelligence/military communities
- Analysis of such data that goes beyond simple information retrieval (e.g. list all the transactions involving entity X) is crucial
- Many analysis algorithms that operate on linked data (or data stored in an RDBMS for that matter) are non-transparent

Similarity overview

Raytheon
BBN Technologies

- At a high level, “similarity” can be thought of as denoting any of the following:
  - Metric (e.g. Euclidian distance)
  - Divergence measure (e.g. KL divergence)
  - Kernel
- Similarity plays a crucial role in many of the most important data analysis tasks:
  - Query by example
  - Anomaly/normalcy detection
  - Prediction
  - Data summarization

Similarity for entities in linked data

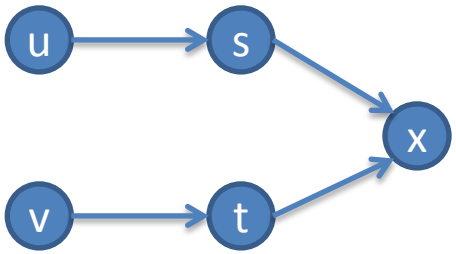
Raytheon
BBN Technologies

- There are roughly two facets to similarity in linked data
  - *Lexical* (e.g. edit distance between property sets of two entities)
  - *Structural,* which involves assessing the underlying link structure of the graph
- This work focuses on structural similarity

Related work for structural similarity

Raytheon
BBN Technologies

- We base our structural similarity formulation on the following intuition[1]:
  - Two entities are similar if they relate (are linked) to similar entities via directed edges
  - Entities are self-similar
- For convenience, we assume similarity scores are between 0 and 1



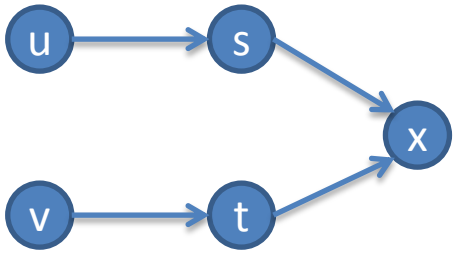
- *sim(x,x) = 1*
- *sim(s,t) > 0*
- *sim(u,v) > 0*
- *sim(u,v) < sim(s,t)*
- *sim(u,t) = 0*

1) Jeh & Widom 2002

Related work for structural similarity

Raytheon
BBN Technologies

- The expected meeting of two random walks from node u and v is (inversely) proportional to their similarity under the recursive definition[1]



- Walkers starting at u and v meet in two steps
- Walkers starting at s and t meet In one step

- This relationship can be exploited to develop scalable (web scale) approximation algorithm for recursive similarity systems[2]

1) Jeh & Widom 2002

2) Fogaras & Racz 2005

Our work

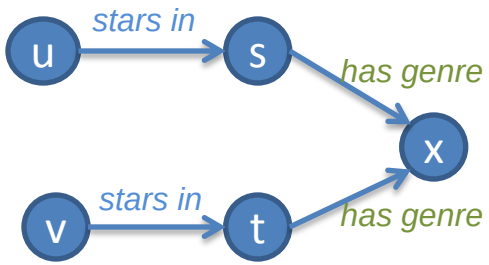
Raytheon
BBN Technologies

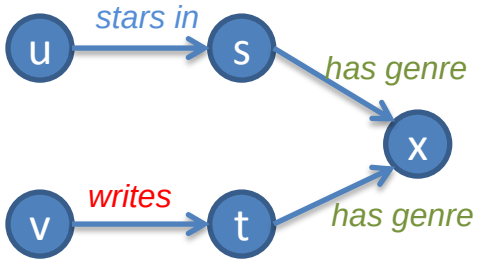
- We extend previous work in three important ways especially relevant for linked data
  - We account for different edge labels
  - We develop the concept of *salience*
  - We build an explanation framework
- We experiment with data from the Linked Movie Database which has information about
  - Films
  - Genres
  - Actors, Writers, Directors, etc.

Accounting for edge labels

Raytheon
BBN Technologies

- To motivate the importance of edge labels, consider the following graphs:





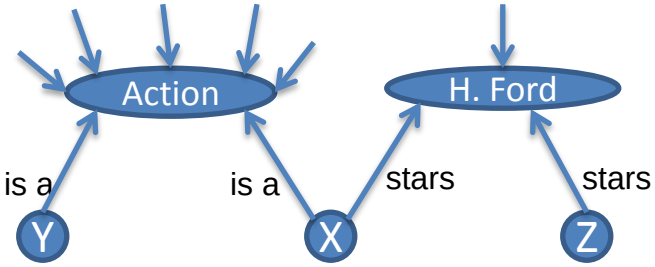
- Nodes u and v in the first graph should be more similar than u and v in the second graph because of the difference in v's relationship with t
- To account for this, we only allow walkers to meet when the sequence of edges they have traversed have the same labels

Accounting for salience

Raytheon
BBN Technologies

- As a motivating example, consider two statements about a movie X in a movie database
  - X is an Action movie
  - X stars Harrison Ford
- We refer to the predicate-object pair (in red) above as a fact about X

- The salience of a fact f, *sal(f)* is a measure of how *discriminative* f is



- Random walks can be biased to favor more salient facts (obscure), less salient facts (obvious), or something in between

Salience cont'd

Raytheon
BBN Technologies

- For a given fact *f*, let S_f denote the set of subjects to which *f* applies and let *S* denote the set of all entities

$$sal(f) = 1 - \frac{\log(|S_f|)}{\log(|S|)}$$

- The bias of f, w(f) can be defined in various ways, such as:

$w(f) = sal(f)$	*favors obscure facts*
$w(f) = 1 - sal(f)$	*favors obvious facts*
$w(f) = sal(f) * (1 - sal(f))$	*favors not-too-obvious/not-too-obscure facts*

- How we choose to bias the random walks depends on context

Explanation generation

Raytheon BBN Technologies

- Explanations are generated by a small number of random walkers starting at each of the similar nodes
- Walkers remember which facts they have traversed
- Explanations can be biased with salience (obscure to obvious)
- Example:

  X = "A New Hope"

  Y = "The Empire Strikes Back"

  sim(X,Y) = …… **Why?**

John Williams
  {A New Hope, hasMusicContributor, John Williams}
  {The Empire Strikes Back, hasMusicContributor, John Williams}

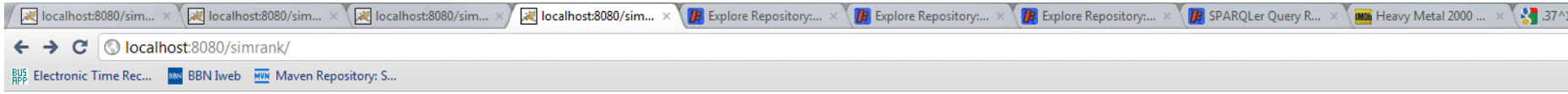
Harrison Ford
  {A New Hope, hasActor, Harrison Ford}
  {The Empire Strikes Back, hasActor, Harrison Ford}

Star Wars (Film Collection)
  {A New Hope, inCollection, Star Wars}
  {The Empire Strikes Back, inCollection, Star Wars}

  {Revenge of the Sith, hasSequel, A New Hope}
    {Revenge of the Sith, inCollection, Star Wars}
  {A New Hope, hasSequel, The Empire Strikes Back}
    {A New Hope, inCollection, Star Wars}
...

Preliminary results: query by example

Raytheon BBN Technologies



Similarity

Parliament URL: http://localhost:8081/parl
Query URL: http://data.linkedmdb.org
Other URL: http://data.linkedmdb.org
Similar | Similarity
Initialize

Getting similar items for http://data.linkedmdb.org/film/75: Star Wars
Time: 80 ms

http://data.linkedmdb.org/film/76, 0.1715689999999999	Empire Strikes Back
http://data.linkedmdb.org/film/68, 0.14936899999999986	Return of the Jedi
http://data.linkedmdb.org/film/36516, 0.10866899999999996	The Star Wars Holiday Special
http://data.linkedmdb.org/film/69, 0.090169	The Phantom Menace
http://data.linkedmdb.org/film/70, 0.086469	Attack of the Clones
http://data.linkedmdb.org/film/78, 0.0851	Revenge of the Sith
http://data.linkedmdb.org/film/32788, 0.075369	Return of the Ewok
http://data.linkedmdb.org/film/7676, 0.06290000000000001	The Emperor - a student film written and directed by George Lucas along with a frequent collaborator, Bob Hudson
http://data.linkedmdb.org/film/11532, 0.05550000000000001	Slipstream - a sci-fi movie starring Mark Hamill (Luke Skywalker)
http://data.linkedmdb.org/film/2900, 0.0407	A short film made by Lucas as a student, which he would adapt 4 years later into a full-length movie starring Robert Duvall
http://data.linkedmdb.org/film/3463, 0.03836900000000001	
http://data.linkedmdb.org/film/33762, 0.037000000000000005	
http://data.linkedmdb.org/film/1435, 0.034669000000000005	
http://data.linkedmdb.org/film/32615, 0.0333	
http://data.linkedmdb.org/film/25625, 0.029600000000000005	
http://data.linkedmdb.org/film/35257, 0.029600000000000005	
http://data.linkedmdb.org/film/33653, 0.027269000000000005	
http://data.linkedmdb.org/film/25531, 0.027269	
http://data.linkedmdb.org/film/2685, 0.027269	
http://data.linkedmdb.org/film/1729, 0.027269	
http://data.linkedmdb.org/film/7047, 0.027269	

Future work

Raytheon
BBN Technologies

- Test our similarity metric on other standard data mining tasks like clustering and classification
  - e.g. by using BBN network data
- Extend our algorithms to account for predicate similarity as well as node similarity
- Combine lexical similarity measures with structural similarity

Summary

Raytheon BBN Technologies

- We have introduced a number of novel extensions for computing and understanding similarities in linked data
  - Labeled edges
  - Salience
  - Explanation
- We have developed a scalable prototype implementation for our similarity metric and demonstrated its utility in the context of querying by example