KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks

ICDM’15 Atlantic City, USA

Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, Jiawei Han
# Outline

<table>
<thead>
<tr>
<th>Motivation</th>
<th>The problem of current similarity measures.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KnowSim</td>
<td>Our way for computing similarity.</td>
</tr>
<tr>
<td>Experiments</td>
<td>The results on benchmark datasets.</td>
</tr>
</tbody>
</table>
Motivation


Bush portrayed himself as a compassionate conservative, implying he was more suitable than other Republicans to go to lead the United States.

Are the two documents similar?

Unstructured Data Similarity  Structured Network Node Similarity
Document-Based Heterogeneous Information Network Construction

• Machine learning with world knowledge framework [Wang et al. KDD’15]

World Knowledge Specification

Specified World Knowledge Representation

General purpose problem vs. Domain specific problem

Knowledge representation vs. Data representation

Documents

World knowledge bases

C. Wang et al. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15
World Knowledge Specification: Unsupervised Semantic Parsing for Documents


Composition rules: Join (between binary and unary); Intersection (between unary and unary).

Logic form construction: based on lexicon and composition rules recursively.

Entities are linked to Freebase.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].
World Knowledge Specification: Semantic Filtering

- Conceptualization based semantic filter (CBSF).
  
  Assumption: correct semantic meaning can best fit the context. Different entities can be used to disambiguate each other.

```
apple
software company, brand, fruit
```

```
adobe
brand, software company
```

```
software company, brand
```

P(type | related entities)

largest probability ones are selected

A cluster of entities of type features

Assumption:
Correct semantic meaning can best fit the context.

Different entities can be used to disambiguate each other.

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.
Specified World Knowledge Representation: Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.


Bush portrayed himself as a compassionate conservative, implying he was more suitable than other Republicans to go to lead the United States.

A good way to model real world data!

Network schema: High-level description of a network.
Meta-Path

Meta-path: A **path/link** in the network schema. [Sun et al., 2011]

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB’11.
KnowSim

KnowSim: An unstructured data similarity measure defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

\[
KS(d_i, d_j) = \frac{2 \times \sum_{m}^{M'} w_m | \{p_{i \rightarrow j} \in P_m \}|}{\sum_{m}^{M} w_m | \{p_{i \rightarrow i} \in P_m \}| + \sum_{m}^{M} w_m | \{p_{j \rightarrow j} \in P_m \}|}
\]

Semantic broadness: the number of total meta-paths between themselves.

• Intuition: The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.

• KnowSim is computed in nearly linear time.
Challenges

Number of meta-paths could be very large.

\[ KS(d_i, d_j) = \frac{2 \times \sum_{m=1}^{M'} w_m | p_{i \rightarrow j} \in P_m |}{\sum_{m=1}^{M'} w_m | p_{i \rightarrow i} \in P_m | + \sum_{m=1}^{M'} w_m | p_{j \rightarrow j} \in P_m |} \]

The weight/importance of each meta-path is different when the domain is different.

#1: How should we generate the large number of meta-paths at the same time?
Previous studies only focus on single meta-path, enumeration over the network is OK. In real world, what will happen when thousands of meta-paths are needed?

#2: How should we decide the weight of each meta-path?
Previous studies treat them equally. In real world, different meta-path should contribute differently in various domains.
Meta-Path Dependent Random Walk

Intuition: Discovering compact sub-graph based on seed document nodes.

• Compute Personalized PageRank around seed nodes.
• The random walk will get trapped inside the blue sub-graph.

Algorithm outline:
• Run PPR (approximate connectivity to seed nodes) with teleport set = \{S\}
• Sort the nodes by the decreasing PPR score
• **Sweep** over the nodes and find compact sub-graph.
• Use the sub-graph instead of the whole graph to compute # of meta-paths between nodes.
Meta-Path Selection

• Maximal Spanning Tree based Selection [Sahami, 1998]
  • Intuition: meta-paths that only weakly influence the remaining domain variables are candidates for elimination (Select meta-paths with the largest dependencies with others).

\[
\sum_{j \neq i}^{M} \cos(D_{:,j_1}, D_{:,j_2}) \times \frac{1}{M - 1}
\]

• Laplacian Score based Selection [He, 2006]
  • Intuition: Laplacian score represents the power of a meta-path in discriminating documents from different clusters.

\[
L_j = \frac{\bar{D}_{:,j}^T LD_{:,j}}{\bar{D}_{:,j}^T \wedge D_{:,j}}
\]
## Experiments

<table>
<thead>
<tr>
<th>Document datasets</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>#(Categories)</td>
<td>#(Leaf Categories)</td>
<td>#(Documents)</td>
</tr>
<tr>
<td>20Newsgroups (20NG)</td>
<td>6</td>
<td>20</td>
<td>20,000</td>
</tr>
<tr>
<td>GCAT (Government/Social)</td>
<td>1</td>
<td>16</td>
<td>60,608</td>
</tr>
</tbody>
</table>

GCAT is top category in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

<table>
<thead>
<tr>
<th>World knowledge bases</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>#(Entity Types)</td>
<td>#(Entity Instances)</td>
<td>#(Relation Types)</td>
<td>#(Relation Instances)</td>
</tr>
<tr>
<td>Freebase</td>
<td>1,500</td>
<td>40 millions</td>
<td>35,000</td>
<td>2 billions</td>
</tr>
</tbody>
</table>

publicly available knowledge base with entities and relations collaboratively collected by its community members.

The number is reported in [X. Dong et al. KDD’14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.
### Text Similarity Results

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Similarity Measures</th>
<th>BOW</th>
<th>BOW+TOPIC</th>
<th>BOW+ENTITY</th>
<th>BOW+TOPIC+ENTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>Cosine</td>
<td>0.2400</td>
<td>0.2713</td>
<td>0.2473</td>
<td>0.2768</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.2352</td>
<td>0.2632</td>
<td>0.2369</td>
<td>0.2650</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.2400</td>
<td>0.2712</td>
<td>0.2474</td>
<td>0.2767</td>
</tr>
<tr>
<td>KnowSim+UNI</td>
<td>Cosine</td>
<td>0.2860</td>
<td>0.2891</td>
<td></td>
<td>0.2913 (+5.2%)</td>
</tr>
<tr>
<td>GCAT</td>
<td>Cosine</td>
<td>0.3490</td>
<td>0.3639</td>
<td>0.2473</td>
<td>0.3128</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3313</td>
<td>0.3460</td>
<td>0.2319</td>
<td>0.2991</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3490</td>
<td>0.3638</td>
<td>0.2474</td>
<td>0.3156</td>
</tr>
<tr>
<td>KnowSim+UNI</td>
<td>Cosine</td>
<td>0.3815</td>
<td>0.3833</td>
<td></td>
<td>0.4086 (+12.3%)</td>
</tr>
</tbody>
</table>

Finding #1: Our method KnowSim is better than traditional measures. KnowSim can better leverage world knowledge (entity, meta-path) rather than just treating them as flat features (e.g., BOW+ENTITY).

Finding #2: More world knowledge will lead to better performance. Laplacian score based meta-path selection method (KnowSim+LAP) performs the best.
Spectral Clustering Using KnowSim Matrix

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Similarity Measures</th>
<th>BOW</th>
<th>BOW+TOPIC</th>
<th>BOW+ENTITY</th>
<th>BOW+TOPIC+ENTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>20NG</td>
<td>Cosine</td>
<td>0.3440</td>
<td>0.3461</td>
<td>0.3896</td>
<td>0.4247</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3547</td>
<td>0.3517</td>
<td>0.3850</td>
<td>0.4292</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3440</td>
<td>0.3457</td>
<td>0.3894</td>
<td>0.4248</td>
</tr>
<tr>
<td>GCAT</td>
<td>Cosine</td>
<td>0.3932</td>
<td>0.4352</td>
<td>0.2394</td>
<td>0.4106</td>
</tr>
<tr>
<td></td>
<td>Jaccard</td>
<td>0.3887</td>
<td>0.4292</td>
<td>0.2497</td>
<td>0.4159</td>
</tr>
<tr>
<td></td>
<td>Dice</td>
<td>0.3932</td>
<td>0.4355</td>
<td>0.2392</td>
<td>0.4112</td>
</tr>
</tbody>
</table>

- **KnowSim+UNI** 0.4304, **KnowSim+MST** 0.4412, **KnowSim+LAP** 0.4461 (+3.9%)

**Finding:** we can get the same results according to the clustering NMI. KnowSim is a better similarity measure. We can infer that KnowSim could have positive impact on other similarity-based applications, e.g., document classification and ranking.
Conclusion

Problem  Document similarity as network node similarity.

Approach  World knowledge specification; KnowSim: unstructured data similarity defined on network.

Results  Document similarity results and its application (clustering) show the power.

Thank You! 😊