
Marco Ponza, Paolo Ferragina and Soumen Chakrabarti
University of Pisa  IIT Bombay

Motivations
Proliferation of the usage of Knowledge Graphs:
- Retrieval of Information (Bordino, WSDM '13; Corimbert, WWW '16)
- Entity Linking (Meij, WSDM '12; Piccinno, SIGIR '13; Ganea, WWW '16)
- Document Clustering, Classification and Similarity (Scalisci, WSDM '12; Fiaseu, ECIR '12; N. WSDM '13)

Need for computing relatedness between entities!

Contributions
1. A new entity-relatedness dataset WiRe, comprising judgments by human experts on 503 pairs of Wikipedia entities
2. Intrinsic evaluation of all recent relatedness measures:
   - Personalized PageRank (Haveliwala, WWW '02)
   - Link Prediction Liben-Nowell, JAIST '07
   - Word and Document Similarity (Gabrilovich, UCAI '07)
   - Word2Vec (Perozzi, KDD '14)
   - CoSimRank (Rothe, ACL '14)
   - ...and more in the paper!
3. A new efficient Two-Stage Framework for relatedness computation:
   - Configurable joint framework without any need of feature engineering
   - Improvements is more than 5% with peaks of 7% on WiRe
4. Extrinsic evaluation of the new framework on entity linking
5. Publicly available datasets and algorithms

Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>WiRe Pearson</th>
<th>Spearman</th>
<th>WiRe Spearman</th>
<th>Harmonic</th>
<th>WiRe Harmonic</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESA</td>
<td>0.61</td>
<td>0.60</td>
<td>0.67</td>
<td>0.63</td>
<td>0.62</td>
<td>0.645</td>
<td></td>
</tr>
<tr>
<td>Milne&amp;Witten</td>
<td>0.62</td>
<td>0.77</td>
<td>0.63</td>
<td>0.69</td>
<td>0.72</td>
<td>0.675</td>
<td></td>
</tr>
<tr>
<td>DeepWalk</td>
<td>0.71</td>
<td>0.74</td>
<td>0.71</td>
<td>0.68</td>
<td>0.71</td>
<td>0.710</td>
<td></td>
</tr>
<tr>
<td>Entity2Vec</td>
<td>0.68</td>
<td>0.70</td>
<td>0.69</td>
<td>0.74</td>
<td>0.70</td>
<td>0.705</td>
<td></td>
</tr>
<tr>
<td>Two-Stage Framework</td>
<td>0.74</td>
<td>0.83</td>
<td>0.75</td>
<td>0.79</td>
<td>0.765</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fair and comprehensive comparison of all relatedness methods present in the recent literature and properly adapted to our context (more experiments in the paper).

Efficiency

<table>
<thead>
<tr>
<th></th>
<th>Uncompressed</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space</td>
<td>5 GB</td>
<td>445 MB</td>
</tr>
<tr>
<td>Average Time</td>
<td>0.5 ms</td>
<td>3 ms</td>
</tr>
</tbody>
</table>

By carefully optimizing a few steps (details in the paper) the framework turns out to be space-efficient and computationally lightweight!

Our Two-Stage Framework

First Stage: Creation of a Wikipedia Subgraph

Choosing nodes of the subgraph

Creation of edges & their weights

Description of the First Stage. A Wikipedia subgraph is created by retrieving the top-k most related entities to the query entities (i.e. choosing nodes of the subgraph) and by subsequently linking them in a small and weighted sparse graph (i.e. creation of edges & their weights). Both the top-k retrieval and weighting schemes are fully configurable with a various set of algorithms, such as ESA (Gabrilovich, UCAI '07), Milne&Witten (Milne, AAAI '08), DeepWalk (Perozzi, KDD '14), Entity2Vec (Ni, WSDM '16).

Second Stage: Computing Relatedness

The relatedness between the two query entities is computed by running CoSimRank (Rothe, ACL '14) over this small and weighted graph. The overall computation (i.e. first + second stages) is fast and can be performed at query time.

Extrinsic Evaluation: Entity Linking

We replaced the relatedness method used by TagMe (i.e. Milne&Witten) with our Two-Stage Framework. Our relatedness measure not only improves TagMe, but also makes it more insensitive to choices of the $\epsilon$-parameter in TagMe.