Computer Science Department, University of Bonn

Knowledge Graph Analysis

Exercise Sheet 8

Dr. Asja Fischer, Prof. Jens Lehmann

January 18, 2018

1 In Class

1. **Latent Distant Models**

   a) What kind of distance measures (i.e. metrics) do you know? List at least two.

   b) Recall the Structured Embedding model

   $$f_{ijk}^{\text{SE}} = -||\boldsymbol{W}_k^s \boldsymbol{a}_i - \boldsymbol{W}_k^o \boldsymbol{a}_j||_1 \ .$$

   Explain the role of the parameters in the score-function. Do the same for the TransE model

   $$f_{ijk}^{\text{TransE}} = -d(\boldsymbol{a}_i + \boldsymbol{r}_k, \boldsymbol{a}_j) \ .$$

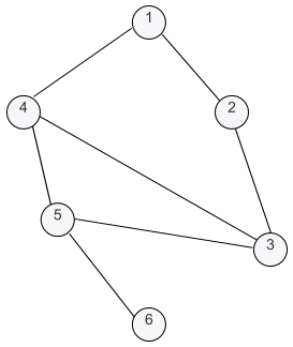
   c) Given $d(\cdot) = ||\cdot||_2$ and $||\boldsymbol{a}_i|| = ||\boldsymbol{a}_j|| = 1$, show that the score function of the TransE model can be written as:

   $$f_{ijk}^{\text{TransE}} = -\left(2\boldsymbol{r}_k^T(\boldsymbol{a}_i - \boldsymbol{a}_j) - 2\boldsymbol{a}_i^{\text{T}}\boldsymbol{a}_j + ||\boldsymbol{r}_k||_2^2\right) \ .$$

2. **Graph theory**
   Look at the undirected graph given in Figrure 2.

   a) What is the neighborhood of node 5?

   b) Give an example for a path between node 2 and node 6.



c) Consider a random walk starting in node 1 and moving in each step to a neighbor of the current note chosen uniformly at random. What is the probability of the path $\{(1,4),(4,5),(5,6)\}$?

3. **Graph Feature Models**

a) Explain the idea behind the **Common neighbors** index

$$f_{ij} = |N(e_i) \cap N(e_j)| \ ,$$

the **Adamic/Adar** index

$$f_{ij} = \sum_{e_h \in N(e_i) \cap N(e_j)} \frac{1}{\log(|N(e_h)|)} \ ,$$

and the **Katz** index

$$f_{ij} = \sum_{l=1}^{\infty} \beta^l \cdot |\Pi^l_{e_i,e_j}| \ ,$$

where $\Pi^l_{e_i,e_j} = \{\text{paths of length } l \text{ from } e_i \text{ to } e_j\}$.

b) Are these local or global similarity indices for uni-relational data?

c) Recall the score function of the **path ranking algorithm (PRA)**

$$f^{\text{PRA}}_{ijk} = \boldsymbol{r}_k^{\text{T}} \boldsymbol{x}^{\text{PRA}} \ ,$$

with

$$\boldsymbol{x}^{\text{PRA}}_{ijk} = [P(\Pi_{e_i,e_j}(t)) : \Pi_{e_i,e_j}(t) \subset \Pi_{ijk}] \ .$$

Consider the following relation sequences/paths and the wights learned by the PRA for the prediction of the college a person attended (i.e. the existence of triples of the form (p, college, c):

- t_1 =(*parents, education, institution*), $r_{k1} = 1.37$
- t_2 =(*placeOfBirth, peopleBornHere, education*), $r_{k2} = 6.4$
- t_3 =(*profession, peopleWithProfession, education, institution*), $r_{k3} = 2.19$

To which logic rules (i.e. Horn clauses) do the paths correspond and which is the most predictive for the existence of the relation?