TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

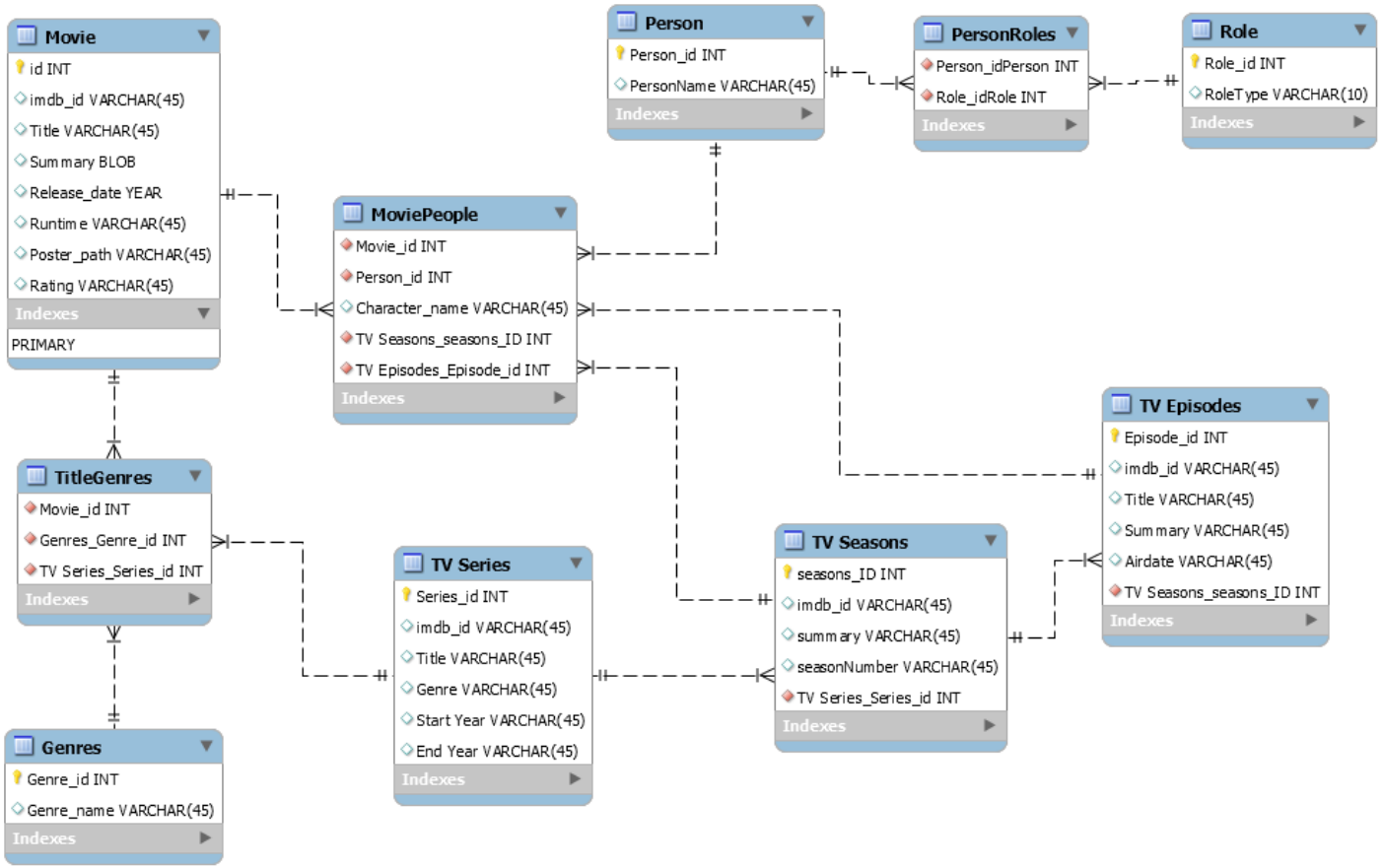
Evaluating Graph Embeddings and Graph Similarity Measurements for Database Systems

Josephine Rehak

Profilprojekt Anwendungsforschung in der Informatik

16.4.2019 13:00

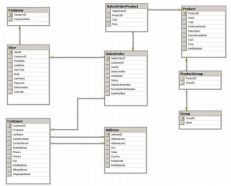
Motivation



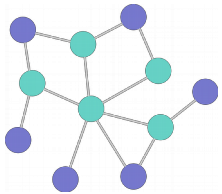
[1]

Motivation I

Overview



dataset



graph

verse

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

embeddings

similarity measures

similarity rankings

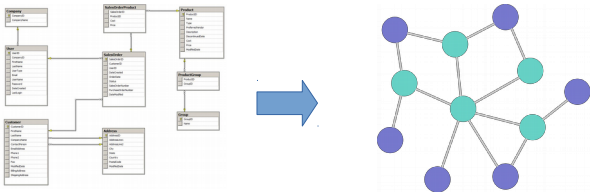
Fundamentals

Fundamentals I

Dresden Database Systems Group

Creating a directed graph from database

- Vertices: IDs + Table entries
  - Except: foreign keys, mn-tables



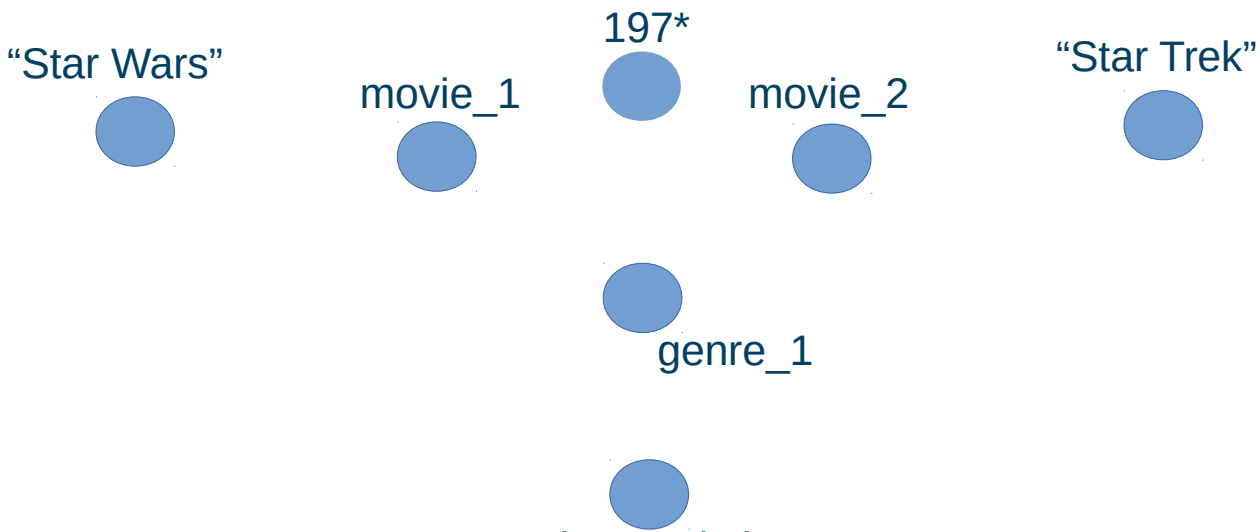
Fundamentals I

Dresden Database Systems Group

id	Name	Year
1	Star Wars	1977
2	Star Trek	1979

mid	gid
1	1
2	1

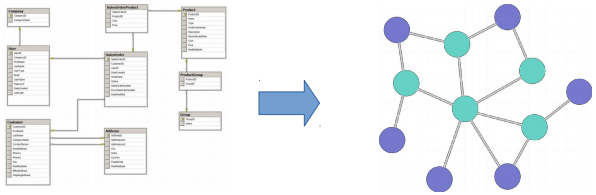
id	Genre
1	Science Fiction



Fundamentals I

Dresden Database Systems Group

Creating a directed graph from database



- Vertices: IDs + Table entries
  - Except: foreign keys, mn-tables
- Edges: Inner + Inter Table Relations
  - Directed edges: id to table values + Foreign keys
  - MN-Table-Relations to bidirected edges

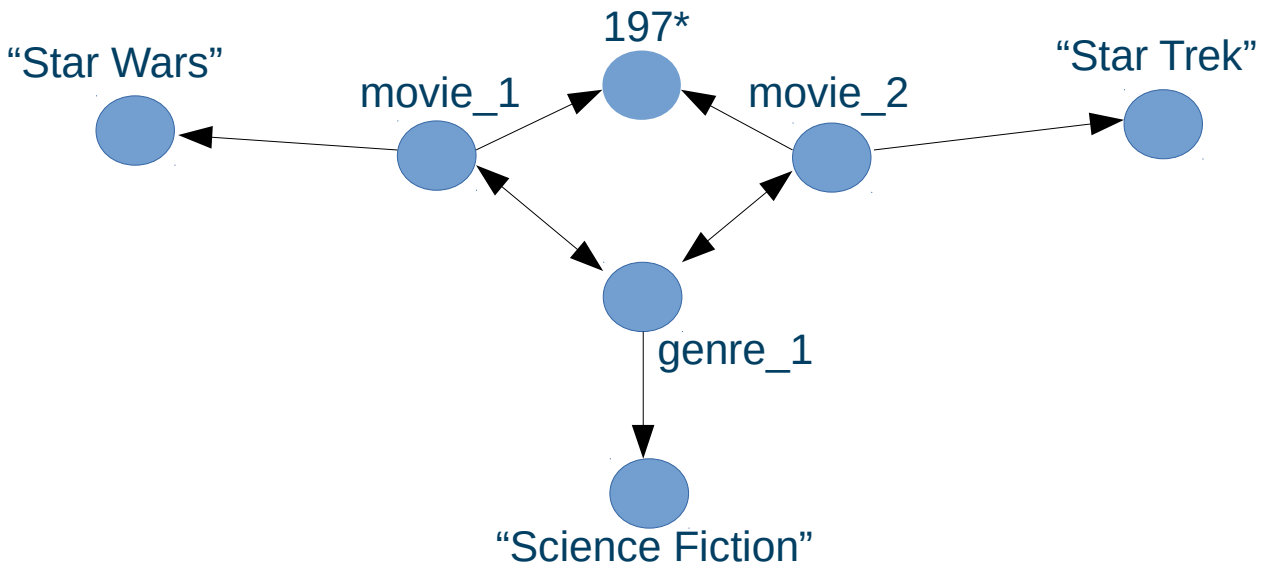
Fundamentals I

Dresden Database Systems Group

id	Name	Year
1	Star Wars	1977
2	Star Trek	1979

mid	gid
1	1
2	1

id	Genre
1	Science Fiction



Similarity Measures in Verse

- Adjacency Similarity:
  - Similarity to neighbours only
  - Low complexity

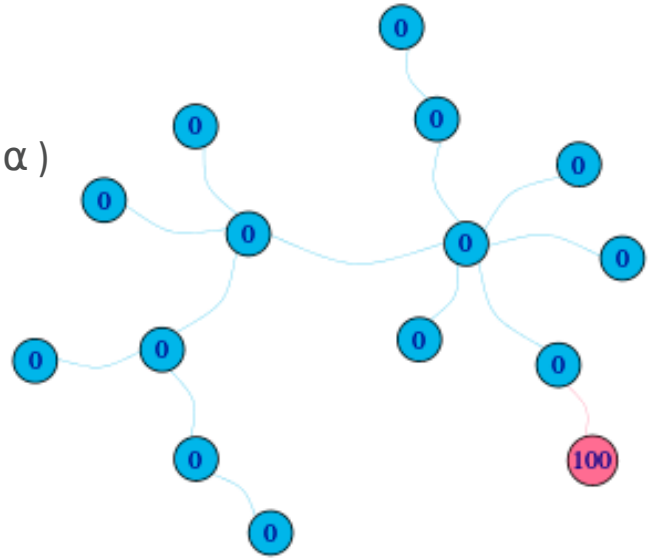
$$sim_{\mathrm{G}}^{\mathrm{ADJ}}(u,v) = \begin{cases} 1/Out(u) & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases}$$

Fundamentals III

Similarity Measures in Verse

- Personalized Page Rank:
  - Random walker with jump back probability (α)
  - Initial assignment, then recursive
  - Ranks converge to:

$$\pi_s = \alpha s + (1 - \alpha)\pi_s A$$



[3]

Fundamentals III

Similarity Measures in Verse

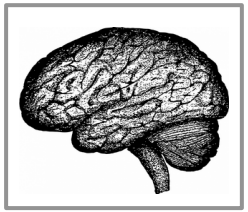
- SimRank:
  - Objects are similar when referenced by similar objects
  - Initial assignment, then recursive
  - importance of farther nodes : C
  - complexity of O(n4)

$$sim_G^{SR}(u,v) = \frac{C}{|I(u)|\,|I(v)|} \sum_{i=1}^{|I(u)|} \sum_{j=1}^{|I(v)|} sim_G^{SR}(I_i(u), I_j(v))$$

Fundamentals I

Dresden Database Systems Group

Verse (VERtex Similarity Embeddings)



- Trains unsupervised 1-layer-NNs
- **Scalable:** Processes 10^6 nodes in less than a day
- **Global**: similarity on any pair of graph nodes
- **Versatile:** supports any similarity measure between nodes

$$\sum_{v \in V} \mathrm{KL}\left(sim_{\mathrm{G}}(v, \cdot) \,||\, sim_{\mathrm{E}}(v, \cdot)\right)$$

Fundamentals II

Dresden Database Systems Group

Verse Vertex Embeddings

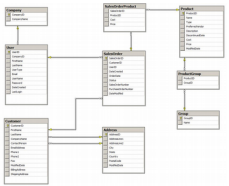
- Mapping of graph vertex to vector space
  - Maintains: neighborhood, connections, graph topology
  - Efficiency for large graphs
  - individual dimensions have no meaning
- Usecases:
  - ML
  - similarity measures
  - NearestNeighbour
  - ...

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

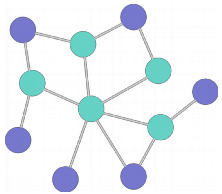
Motivation I

Dresden Database Systems Group

Overview



dataset



graph

verse

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

embeddings

similarity measures

similarity rankings

Fundamentals III

Spearman coeficcient

- Example, $n=3$

Movie id	r 1	r 2	d_i	d_i^2
1	2	1	1	1
2	1	3	-2	4
3	3	2	1	1

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

$$\rho = \text{-0.5}$$

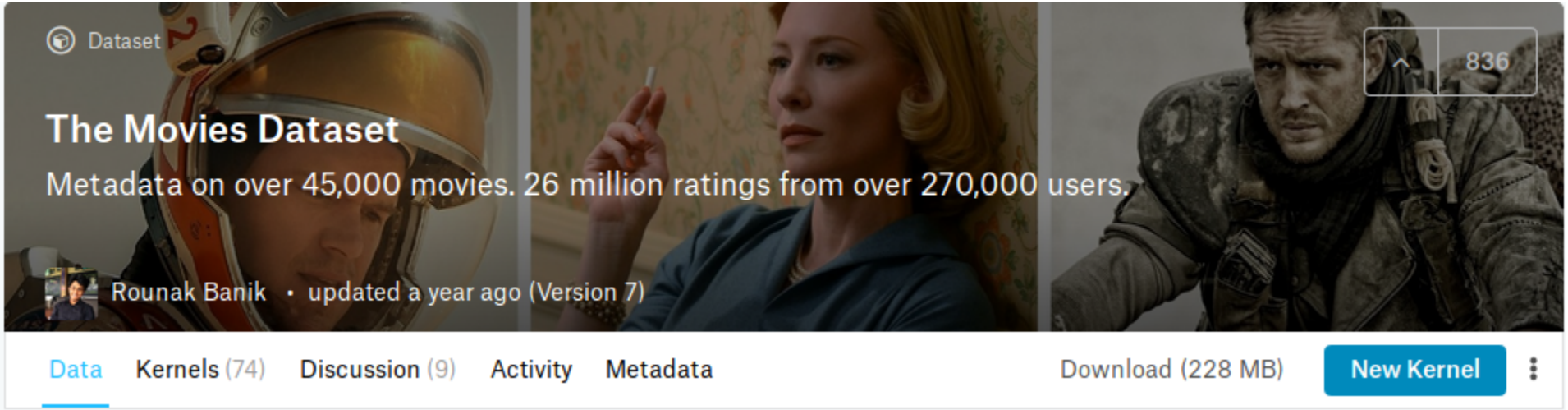
- Values between -1 and 1
  - -1 inverted correspondence
  - 0 no correspondence
  - 1 strong correspondence

Experimental Setup

Experimental Setup

Dresden Database Systems Group

Data Set



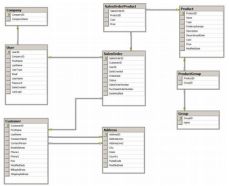
Used excerpt has: 45433 movies, 18001 directors, 20 genres

https://www.kaggle.com/rounakbanik/the-movies-dataset/home

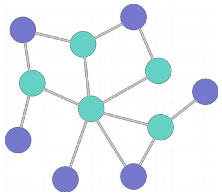
Motivation I

Dresden Database Systems Group

Overview



dataset



graph

verse

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

embeddings

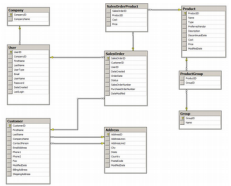
similarity measures

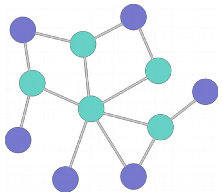
similarity rankings

Experimental Setup

Dresden Database Systems Group

Overview





$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

movie dataset

3 data tables
1 mn table

45433 movies
18001 directors
20 genres

networkx graph

199753 vertices
198865 edges

verse

3x 3h of calculation

embedding

3 x 199753 x 180 matrix

similarity measures

3 measures

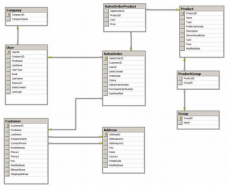
similarity rankings

6 x199753 rankings

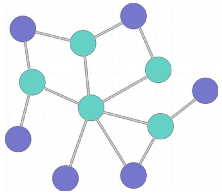
Motivation I

Dresden Database Systems Group

Investigation



dataset



graph

verse

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

embeddings

similarity measures

similarity rankings

Results

Results I

Dresden Database Systems Group

PPageRank Graph

	Movie Names
1	Star_Wars
2	Nana_Neul
3	Aki_Kaurismäki
4	The_Circle
5	Cross_Creek
6	Jean_Mach

AdjacencySimilarity Graph

	Movie Names
1	Star_Wars
2	Hellboy
3	Get_Carter
4	Der_Totmacher
5	Begotten
6	Laws_of_Gravity

Results I

Dresden Database Systems Group

PPageRank Embeddings

	Movie Names
1	Star_Wars
2	Soldier's Girl
3	Незнайка_на_Луне
4	Lignes_de_Front
5	Best_Night_Ever
6	London_Boulevard

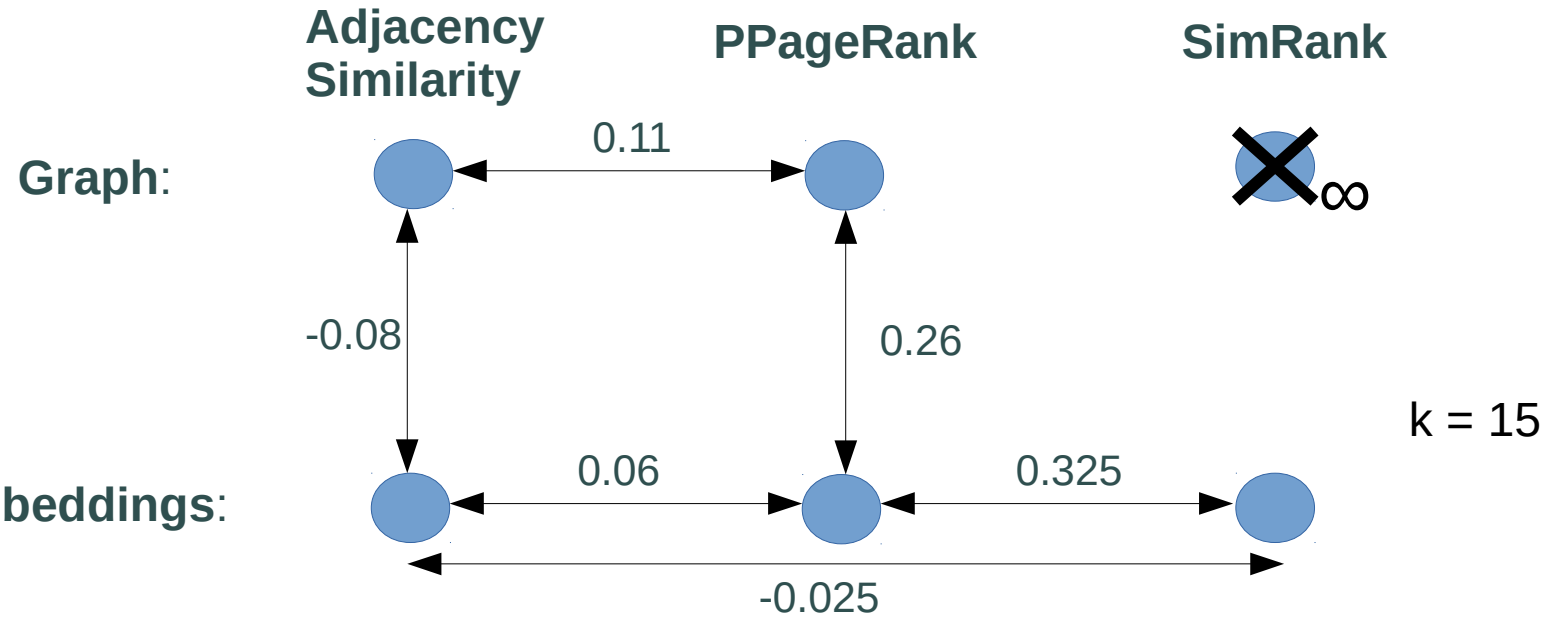
SimRank Embeddings

	Movie Names
1	Star_Wars
2	Roswell
3	THE_DARK_SIDE_OF_DIMENSIONS
4	Perfect
5	Love_and_Distrust
6	National_Lampoon's_Last_Resort

Results I

Dresden Database Systems Group

Spearman on Similarity Rankings



Results I

Dresden Database Systems Group

Computation Time

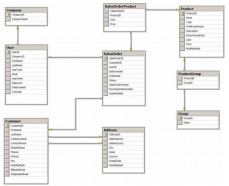
	Adjacency Similarity	PPageRank	SimRank
Graph:	*2.77s*	*12.82s*	∞
Embeddings:	*0.32s*	*0.36s*	0.21s

Discussion

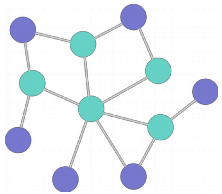
Discussion

Dresden Database Systems Group

What went wrong?



dataset



graph

verse

$$A_{ij} = \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

embeddings

similarity measures

similarity rankings

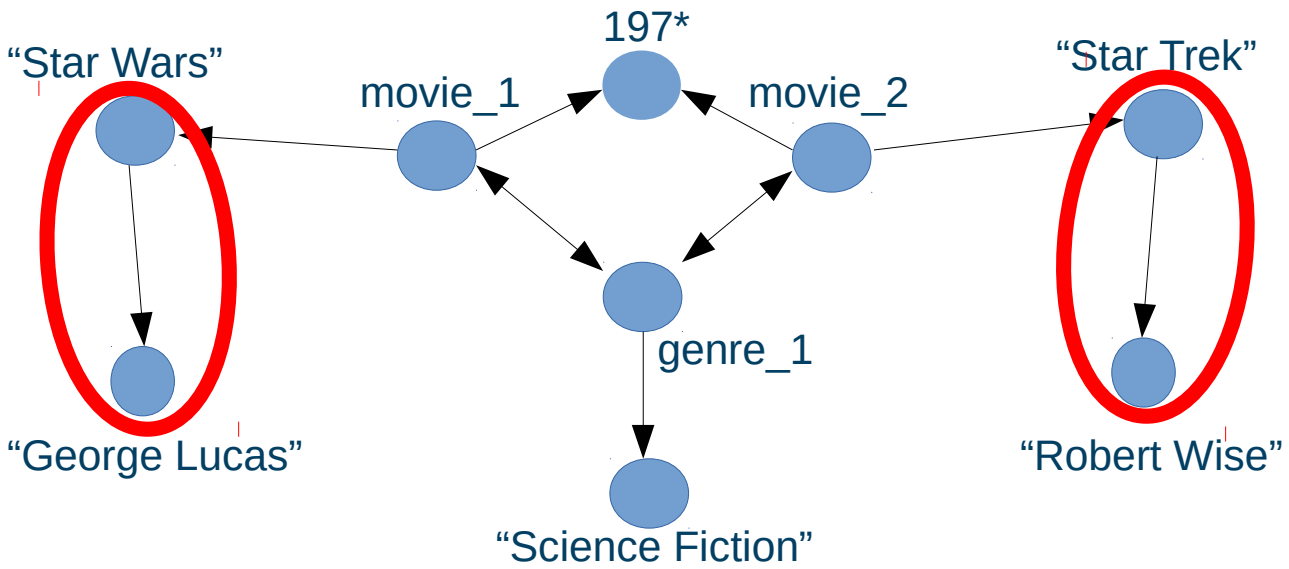
Discussion

Dresden Database Systems Group

id	Name	Year	d_id
1	Star Wars	1977	1
2	Star Trek	1979	2

M 1

id	Director
1	George Lucas
2	Robert Wise



Conclusion

Conclusion I

Dresden Database Systems Group

\+
- Translates across tasks/databases
- Speeding up similarity calculation
- Supports many similarity measures
- ‚Simple‘ solution for complex db problem

\-
- Verse needs preparation time
- Good parameters / similarity measure required
- Dependant of graph structure

Fragen?