TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

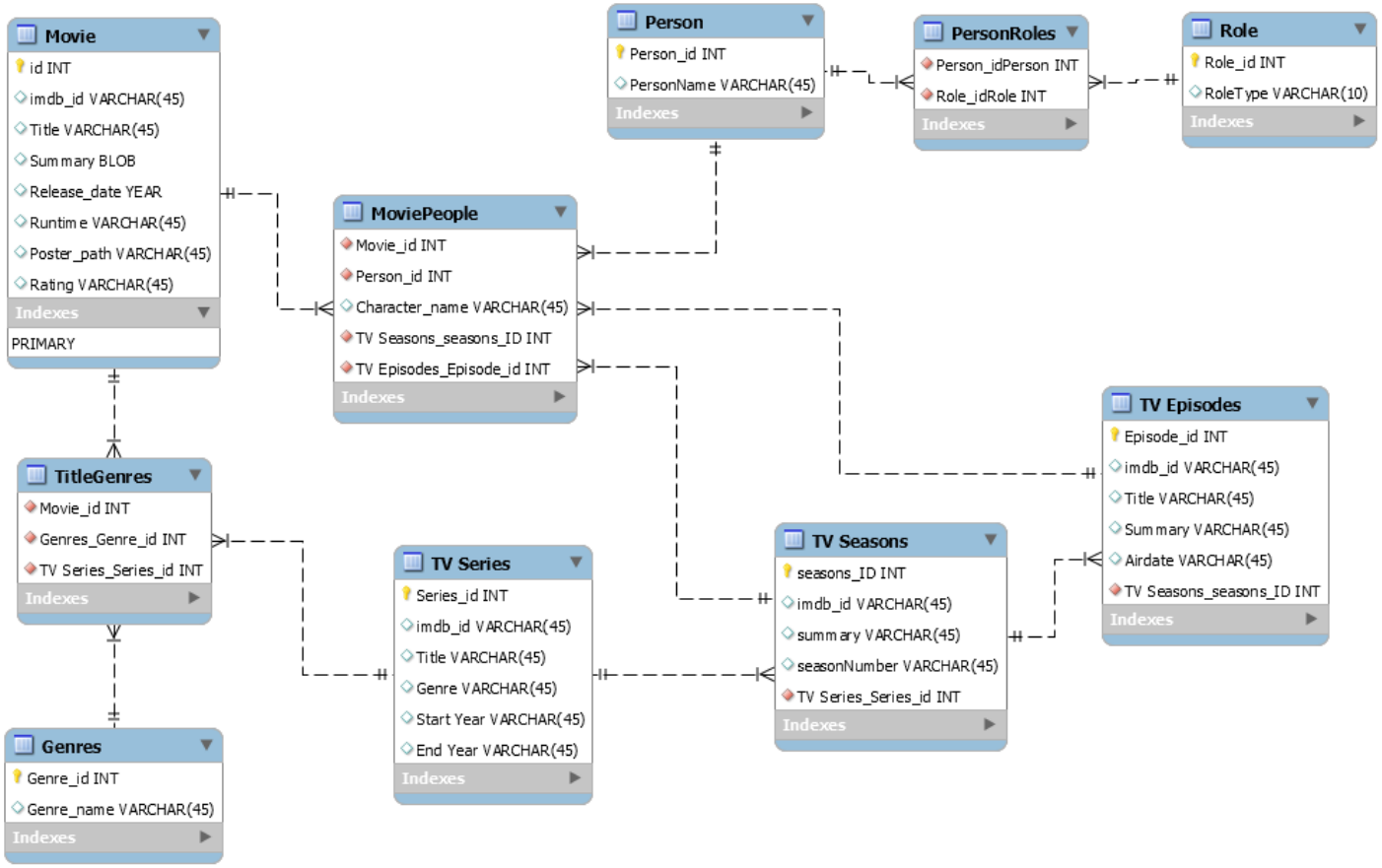
Evaluating Graph Embeddings and Graph Similarity Measurements for Database Systems

Josephine Rehak

Profilprojekt Anwendungsforschung in der Informatik

16.4.2019 13:00

Agenda

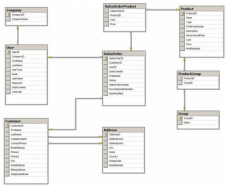

Movie
id INT
imdb_id VARCHAR(45)
Title VARCHAR(45)
Summary BLOB
Release_date YEAR
Runtime VARCHAR(45)
Poster_path VARCHAR(45)
Rating VARCHAR(45)
Indexes
PRIMARY
Person
Person_id INT
PersonName VARCHAR(45)
Indexes
PersonRoles
Person_idPerson INT
Role_idRole INT
Indexes
Role
Role_id INT
RoleType VARCHAR(10)
Indexes
MoviePeople
Movie_id INT
Person_id INT
Character_name VARCHAR(45)
TV Seasons_seasons_ID INT
TV Episodes_Episode_id INT
Indexes
TV Episodes
Episode_id INT
imdb_id VARCHAR(45)
Title VARCHAR(45)
Summary VARCHAR(45)
Airdate VARCHAR(45)
TV Seasons_seasons_ID INT
Indexes
TitleGenres
Movie_id INT
Genres_Genre_id INT
TV Series_Series_id INT
Indexes
TV Series
Series_id INT
imdb_id VARCHAR(45)
Title VARCHAR(45)
Genre VARCHAR(45)
Start Year VARCHAR(45)
End Year VARCHAR(45)
Indexes
TV Seasons
seasons_ID INT
imdb_id VARCHAR(45)
summary VARCHAR(45)
seasonNumber VARCHAR(45)
TV Series_Series_id INT
Indexes
Genres
Genre_id INT
Genre_name VARCHAR(45)
Indexes

[1]

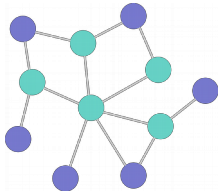
Agenda

Dresden Database Systems Group

Overview



dataset



graph

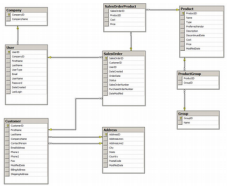
similarity measures

similarity rankings

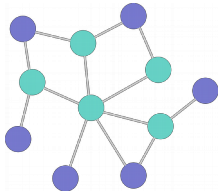
Agenda

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

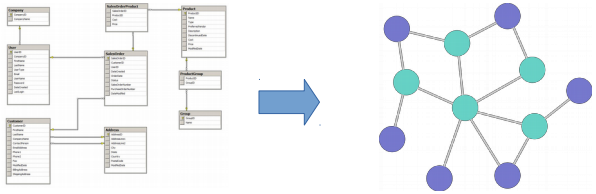
Fundamentals

Fundamentals I

Dresden Database Systems Group

Creating a directed graph from database

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



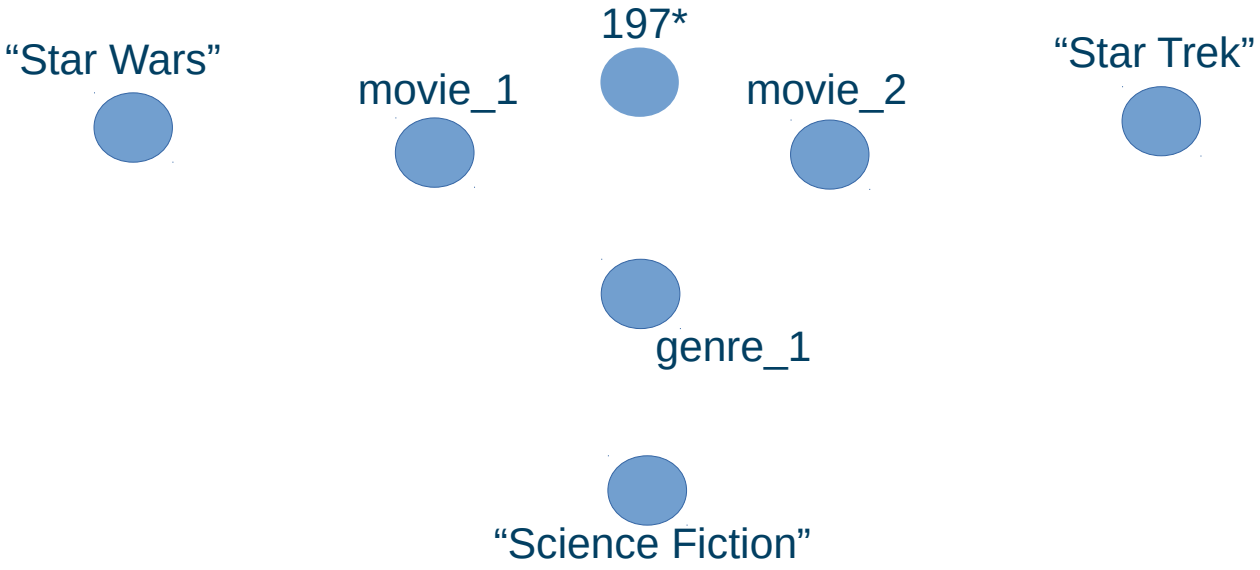
Fundamentals II

Dresden Database Systems Group

id	Name	Year
1	Star Wars	1977
2	Star Trek	1979

mid	gid
1	1
2	1

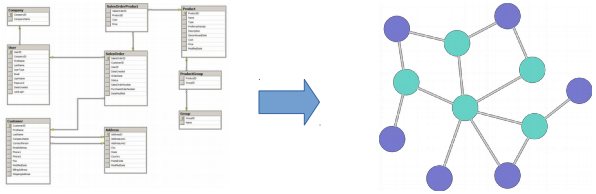
id	Genre
1	Science Fiction



Fundamentals III

Dresden Database Systems Group

Creating a directed graph from database



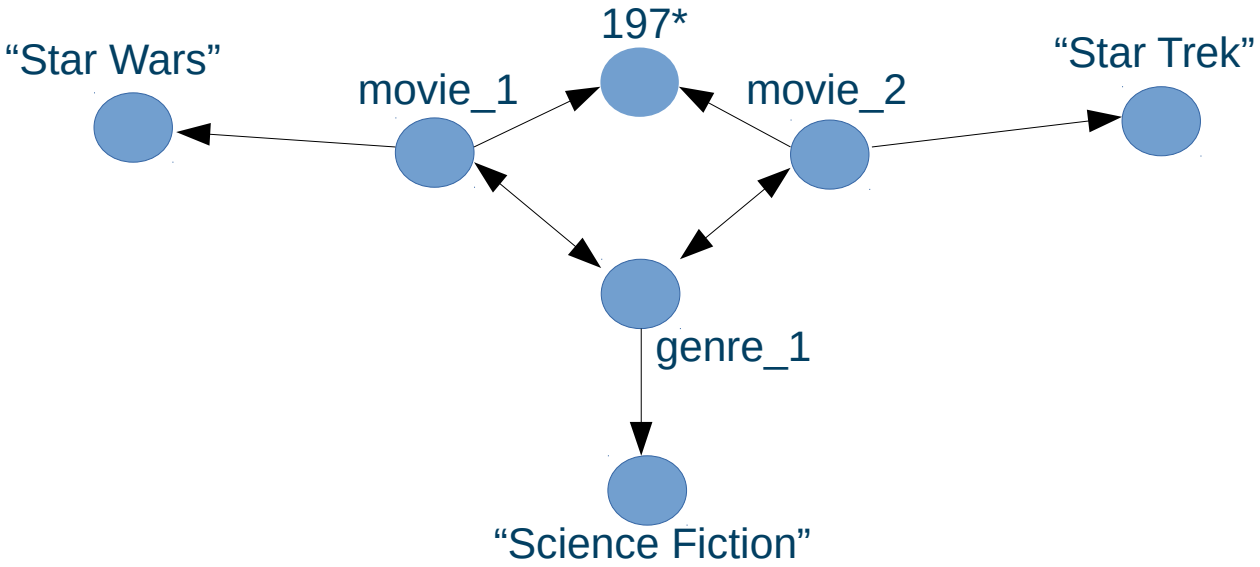
- Vertices: IDs + Table entries
  - Except: foreign keys, mn-tables
- Edges: Inner + Inter Table Relations
  - directed edges: id to table values
  - Bidirected edges: MN-Table-Relations + Foreign Keys

Fundamentals IV

id	Name	Year
1	Star Wars	1977
2	Star Trek	1979

mid	gid
1	1
2	1

id	Genre
1	Science Fiction

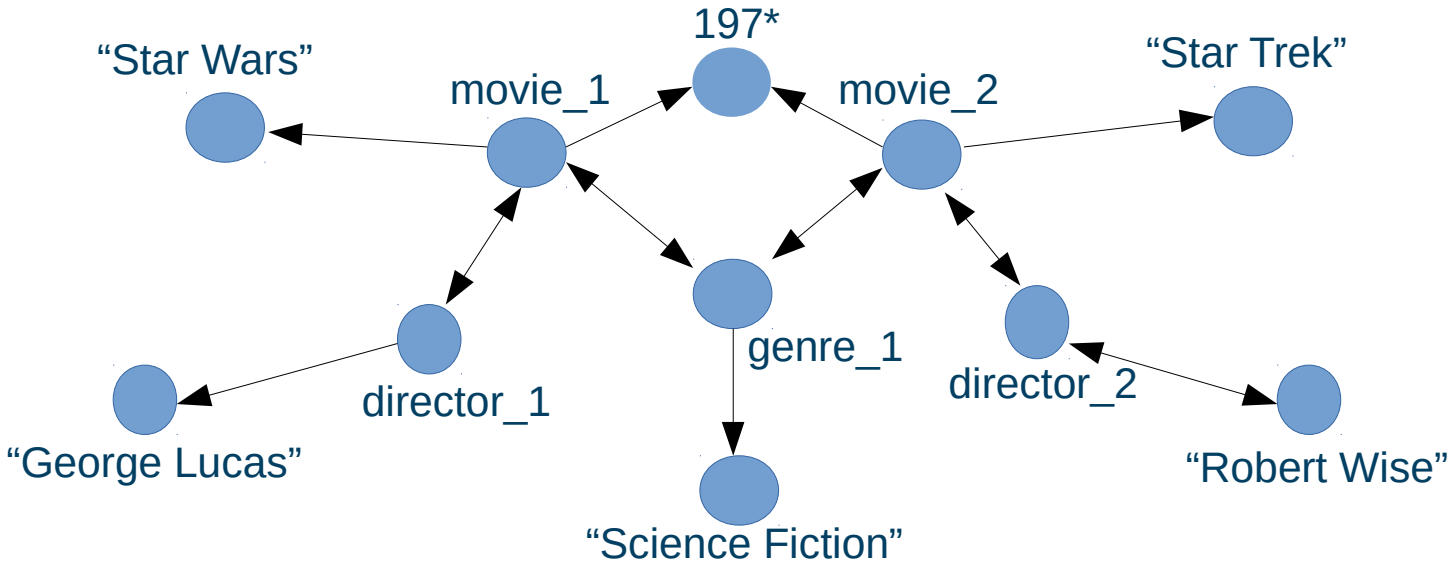


Fundamentals V

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



Fundamentals VI

Dresden Database Systems Group

Similarity Measures in Verse

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

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

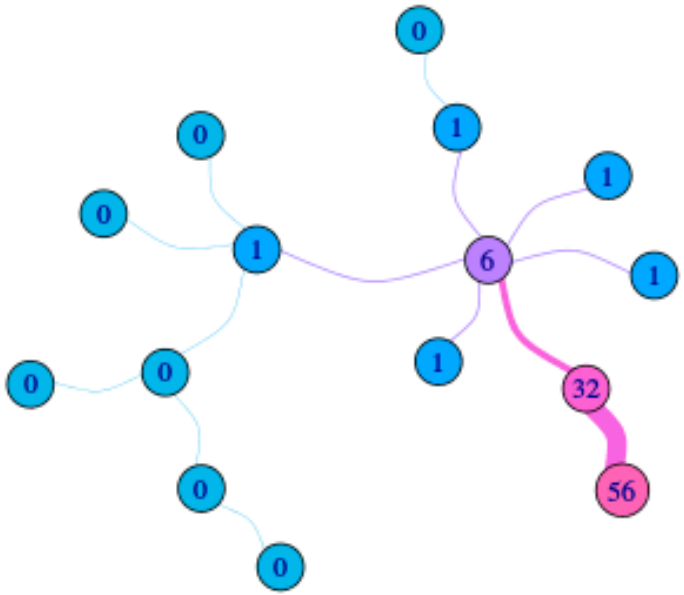
Fundamentals VII

Dresden Database Systems Group

Similarity Measures in Verse

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

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



[3]

Fundamentals IX

Dresden Database Systems Group

Similarity Measures in Verse

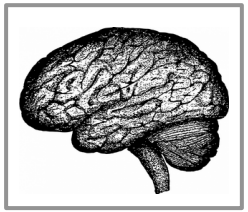
- SimRank:
  - Objects are similar when referenced by similar objects
  - Initial assignment, then recursive
  - importance of farther nodes : C
  - complexity of O(n4)

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

Fundamentals X

Dresden Database Systems Group

Verse (VERtex Similarity Embeddings)



- Trains unsupervised 1-layer-NNs
- Exceeds previous embedding methods
- **Scalable:** Processes 10^6 nodes in less than a day
- **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 XI

Dresden Database Systems Group

Verse Vertex Embeddings

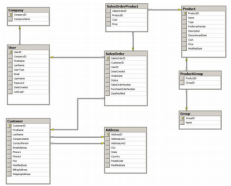
- Mapping of graph vertex to vector space
  - Maintains: neighborhood, connections, graph topology
  - Efficiency for large graphs
  - individual dimensions have no meaning
- Usecases:
  - ML
  - similarity measures
  - NearestNeighbour
  - …

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

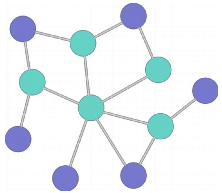
Fundamentals XII

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 XIII

Dresden Database Systems Group

Spearman coeficcient

- Example,

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

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

$$\rho = -0.5$$

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

Experimental Setup

Experimental Setup I

Dresden Database Systems Group

Data Set

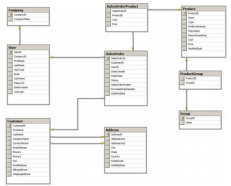
Used excerpt has: 45433 movies, 18001 directors, 20 genres

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

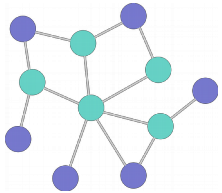
Experimental Setup II

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

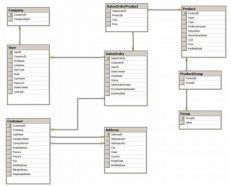
similarity measures

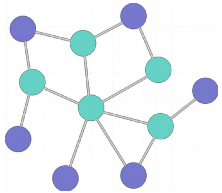
similarity rankings

Experimental Setup III

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
243410 edges

verse

3x 3h of calculation

embeddings

3 x 199753 x 180 matrix

similarity measures

3 measures

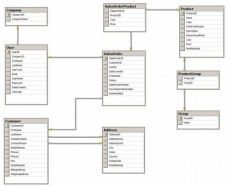
similarity rankings

6 x199753 rankings

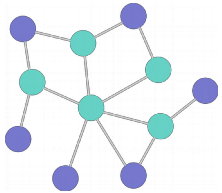
Experimental Setup IV

Dresden Database Systems Group

Investigation



dataset



graph

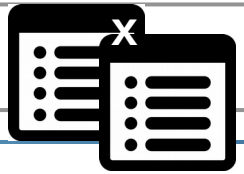
verse

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

embeddings

similarity measures

similarity rankings



Results

Results I

Dresden Database Systems Group

PPageRank Graph

	Movie Names
1	Star Wars
2	Star Wars Episode II: Attack of the Clones
3	American Grafitti
4	THX1138
5	Star Wars Episode III: Revenge of the Sith

AdjacencySimilarity Graph

	Movie Names
1	Star Wars
2	Hellboy
3	Get Carter
4	Der Totmacher
5	Begotten
6	Laws of Gravity

Results II

Dresden Database Systems Group

PPageRank Embeddings

	Movie Names
1	Star Wars
2	Star Wars: Episode III - Revenge of the Sith
3	THX 1138
4	Star Wars: Episode II Attack of the Clones
5	Star Wars: Episode I The Phantom Menace

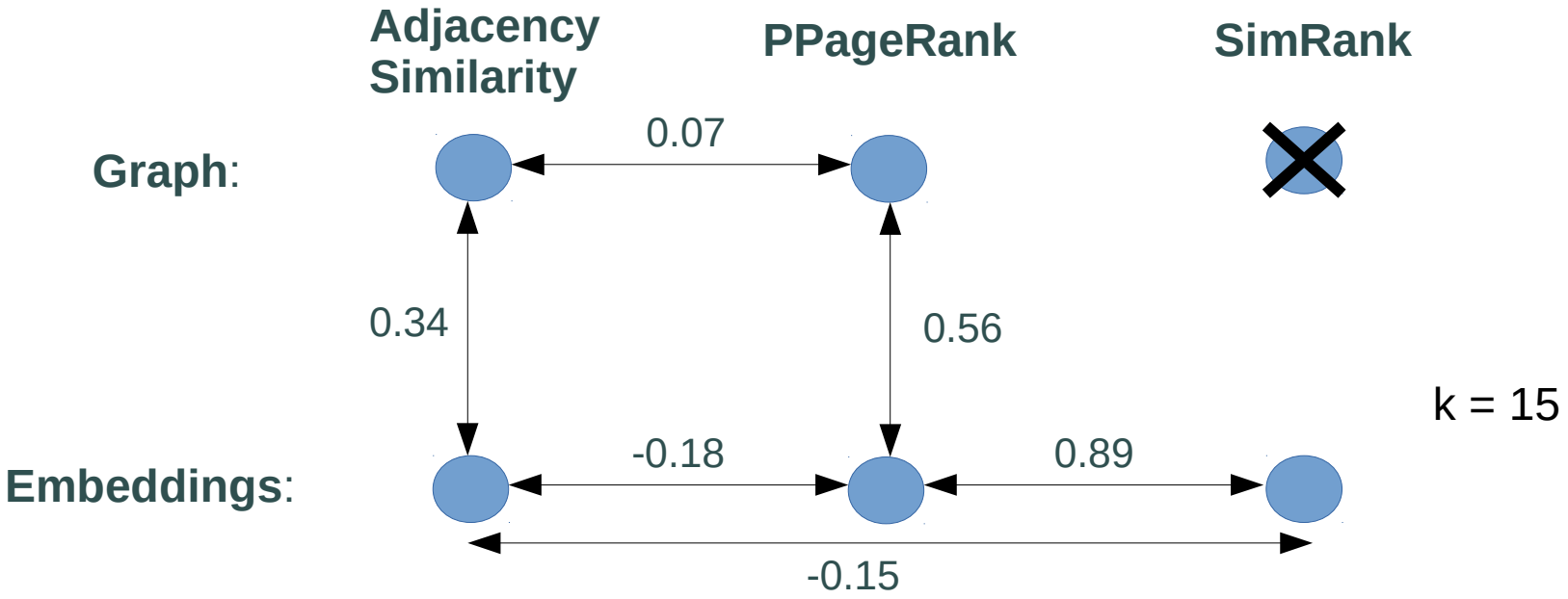
SimRank Embeddings

	Movie Names
1	Star Wars
2	Star Wars: Episode II Attack of the Clones
3	Electronic Labyrinth THX 1138 4EB
4	THX 1138
5	Star Wars: Episode I The Phantom Menace

Results III

Dresden Database Systems Group

Spearman on Similarity Rankings



Results IV

Dresden Database Systems Group

Computation Time

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

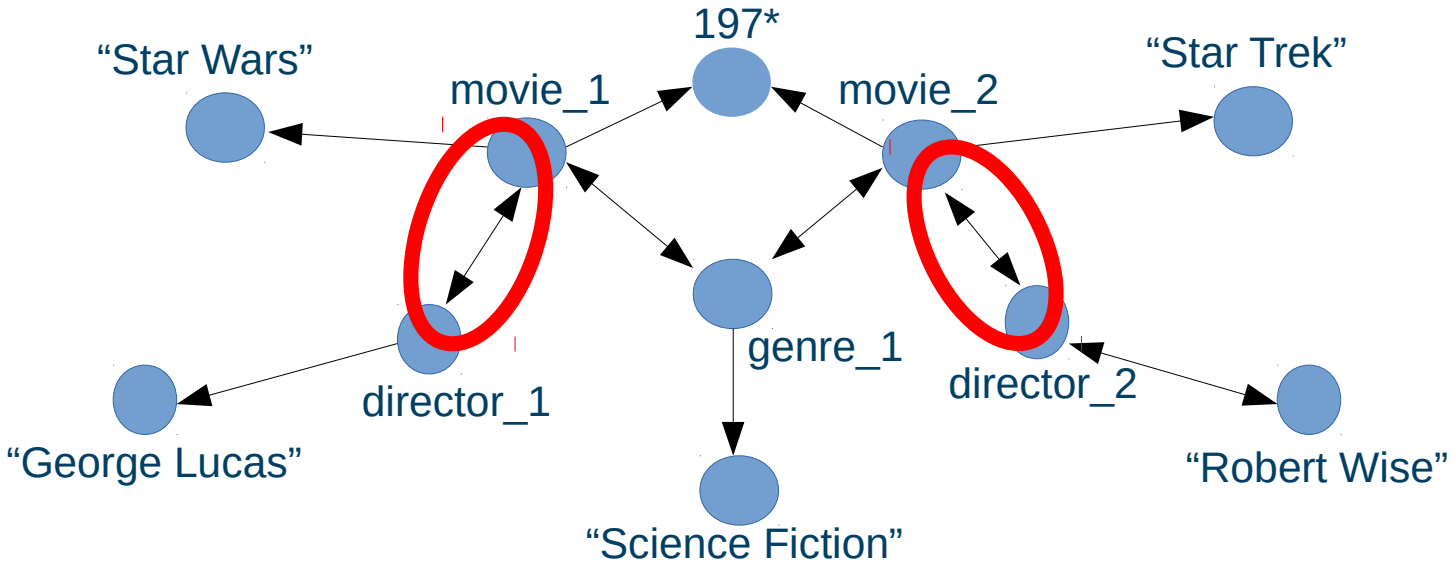
Excursus

Excursus I

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

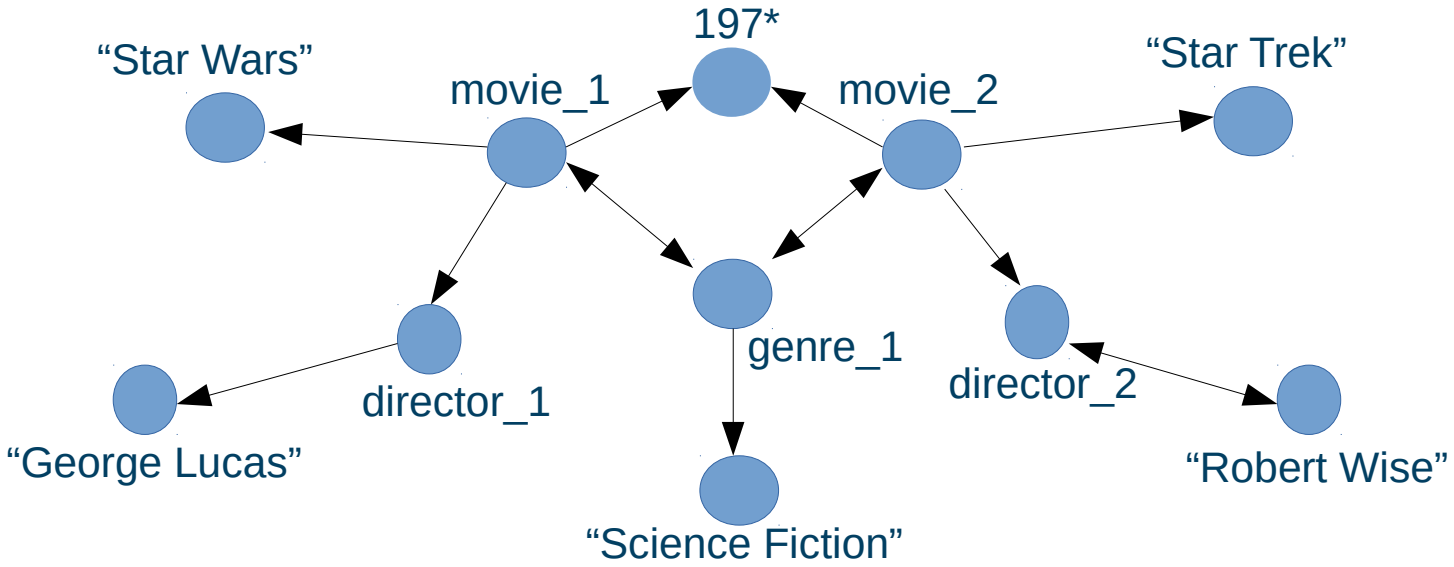


Excursus II

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



Excursus III

Dresden Database Systems Group

Directed: SimRank

	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

Bidirected: SimRank

	Movie Names
1	Star Wars
2	Star Wars: Episode II Attack of the Clones
3	Electronic Labyrinth THX 1138 4EB
4	THX 1138
5	Star Wars: Episode I The Phatnom Menace

Conclusion

Conclusion I

Dresden Database Systems Group

Computation Time

	Adjacency Similarity	PPageRank	SimRank
Graph:	*slow*	*slow*	∞
Embeddings:	*bad ranking*	*efficient* *good ranking*	*efficient* good ranking

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
- Similarity measure and graph construction coupled
- Needs parameter tuning

Fragen?

Extra Slides

Kullback-Leibler-Divergenz

- Also called „*Information Gain*“
- *Measure for divergence of probability distributions*

$$D(P\|Q) = KL(P,Q) = \sum_{x \in X} P(x) \cdot \log \frac{P(x)}{Q(x)}$$

Verse Steps

Dresden Database Systems Group

- Sampling: NCE-based
- *Dimensionality reduction: (linear)*
- *Normalisation: softmax*
- *Optimisation: KL + Stochastic gradient descent*
- *ML: 1-layer NN + input/output layer*

Vertice Similarity Measures

Dresden Database Systems Group

closeness

complexity

Structure-based

- Random Walks
- PpageRank
- SimRank

indirect neighborhood

- Cosine similarity
- Euclidean distance
- Jaccard coeficient
- PPageRank
- SimRank

Direct neighborhood

- Adjacency Similarity