ScaDS.AI
DRESDEN LEIPZIG
CENTER FOR SCALABLE DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

Fast Hubness-Reduced Nearest Neighbor Search for Entity Alignment in Knowledge Graphs

Daniel Obraczka

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Users have complex information needs

”In what year did Richard David James win a Grammy?”

Users have complex information needs

”In what year did Richard David James win a Grammy?”

Users have complex information needs

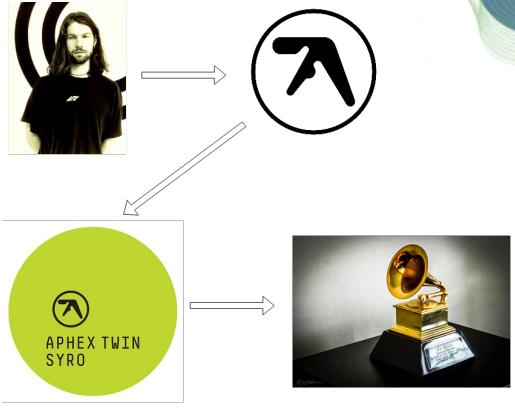
”In what year did Richard David James win a Grammy?”

Users have complex information needs

”In what year did Richard David James win a Grammy?”

Users have complex information needs

”In what year did Richard David James win a Grammy?”



Users have complex information needs

”In what year did Richard David James win a Grammy?” ⇒ 2015

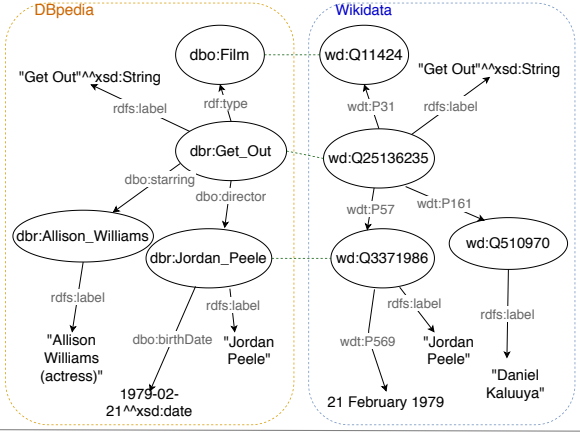
Entity resolution is a crucial first step

E.g. in

- E-commerce: Detect identical products across shops
- Medicine: Deduplicate patient records
- ...

Complex information needs require integrating various sources and find matching entities

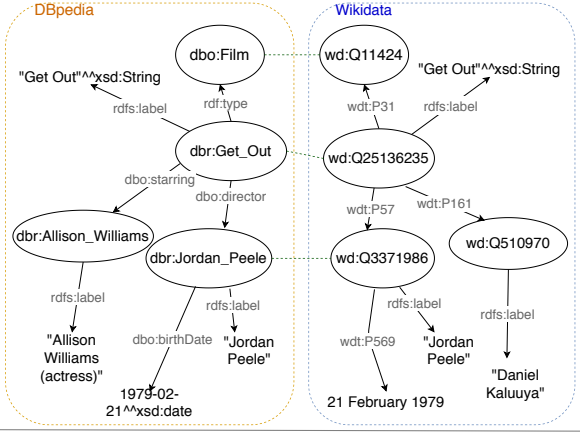
KGs pose specific problems for entity resolution



Flexible schema (usually) means:

- Many entity types
- different (number of) attributes
- various relationship types

KGs pose specific problems for entity resolution



Flexible schema (usually) means:

- Many entity types
- different (number of) attributes
- various relationship types

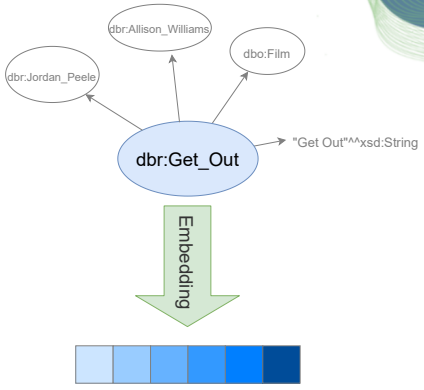
$\Rightarrow$ Challenging for classical entity resolution systems

Knowledge Graph Embeddings (KGEs)

Transform entities into a dense vector

If successful:

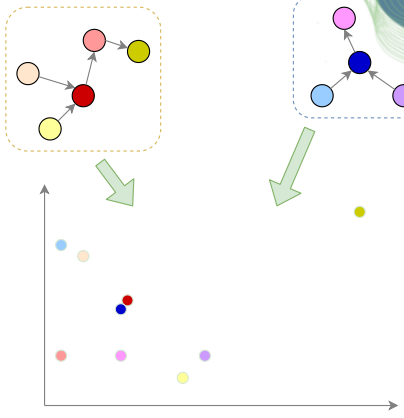
- similar entities close in the embedding space
- relational information retained



Entity Alignment with KGEs

Entity Alignment: find the same entities in different data sources

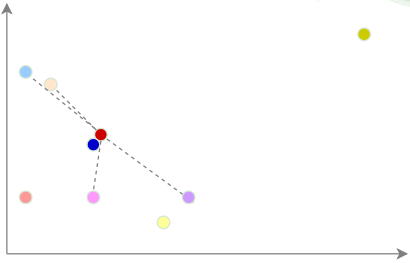
- similar entities close in the embedding space
- use nearest neighbors in embedding space to find matching entities



Entity Alignment with KGEs

Entity Alignment: find the same entities in different data sources

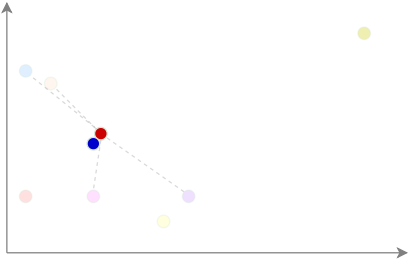
- similar entities close in the embedding space
- use nearest neighbors in embedding space to find matching entities



Entity Alignment with KGEs

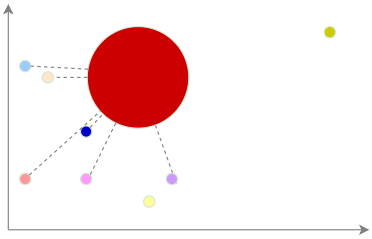
Entity Alignment: find the same entities in different data sources

- similar entities close in the embedding space
- use nearest neighbors in embedding space to find matching entities



Hubness Reduction for Entity Alignment

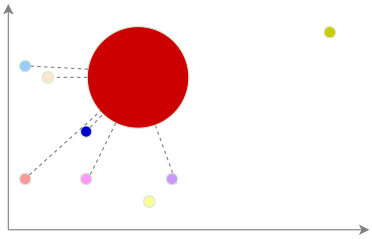
The hubness phenomenon



With increasing dimensionality:

- few points are nearest neighbors (NN) of many points
- many points are NN of no points

The hubness phenomenon



With increasing dimensionality:

- few points are nearest neighbors (NN) of many points
- many points are NN of no points

$\Rightarrow$ detrimental for alignment quality

Hubness reduction (HR)

Different ideas:

- Centering
- Repair asymmetric NN relationships

Hubness reduction (HR)

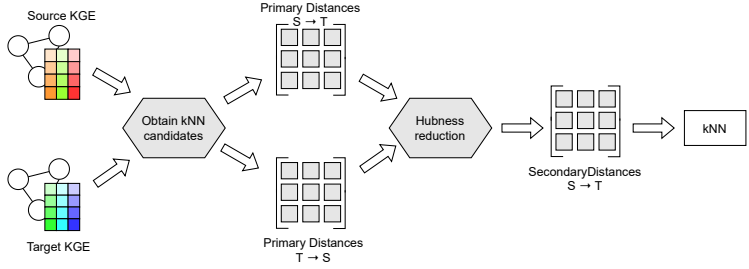
Different ideas:

- Centering
- Repair asymmetric NN relationships

Overview: Feldbauer and Flexer (2019): “A comprehensive empirical comparison of hubness reduction in high-dimensional spaces”

Our contribution

- Benchmark hubness reduction methods for entity alignment
- Provide an open-source library (kiez)



Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

kiez

Open-source python library for
hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

kiez

Open-source python library for
hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling Schnitzer et al. (2012)
- NICDM Schnitzer et al. (2012)
- CSLS Lample et al. (2018)
- Mutual Proximity Schnitzer et al. (2012)
- DisSimLocal Hara et al. (2016)

kiez

Open-source python library for
hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling Schnitzer et al. (2012)
- NICDM Schnitzer et al. (2012)
- CSLS Lample et al. (2018)
- Mutual Proximity Schnitzer et al. (2012)
- DisSimLocal Hara et al. (2016)

(Approximate) Nearest Neighbor Method:

- Sci-kit learn Pedregosa et al. (2011)
  - BallTree Omohundro (1989)
  - KDTree Bentley (1975)
  - Bruteforce
- NMSLIB: HNSW Malkov (2018)
- NGT Iwasaki (2016)
- Annoy *(github.com/spotify/annoy)*
- Faiss Johnson, Douze, and Jégou (2017)

kiez

Open-source python library for
hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling Schnitzer et al. (2012)
- **NICDM** Schnitzer et al. (2012)
- CSLS Lample et al. (2018)
- Mutual Proximity Schnitzer et al. (2012)
- DisSimLocal Hara et al. (2016)

(Approximate) Nearest Neighbor Method:

- Sci-kit learn Pedregosa et al. (2011)
  - BallTree Omohundro (1989)
  - KDTree Bentley (1975)
  - Bruteforce
- NMSLIB: HNSW Malkov (2018)
- NGT Iwasaki (2016)
- Annoy *(github.com/spotify/annoy)*
- Faiss Johnson, Douze, and Jégou (2017)

Non-iterative contextual dissimilarity measure

Schnitzer et al. (2012): “Local and global scaling reduce hubs in space”

$$NICDM(d_{x,y}) = \frac{d_{x,y}}{\sqrt{\mu_x \mu_y}}$$

Non-iterative contextual dissimilarity measure

Schnitzer et al. (2012): “Local and global scaling reduce hubs in space”

$$NICDM(d_{x,y}) = \frac{d_{x,y}}{\sqrt{\mu_x \mu_y}}$$

mean distance to the k-nearest neighbors

Evaluation

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

Sun et al. (2020): “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

Sun et al. (2020): “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”

- 15 KG embedding approaches

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

Sun et al. (2020): “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”

- 15 KG embedding approaches

$\Rightarrow$ 240 KGE pairs

Evaluation Metric

hits@k:

- suited for kNN-based tasks
- counts proportion of true matches in kNN

We use k=50, because we retrieve 50 nearest neighbors

Evaluation Metric

hits@k:

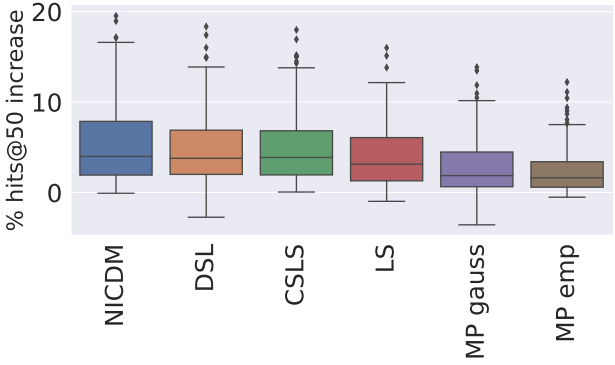
- suited for kNN-based tasks
- counts proportion of true matches in kNN

We use k=50, because we retrieve 50 nearest neighbors

Because absolute hits@k value is largely determined by KGE approach:

- look at improvement
- compare against no HR with same KGE

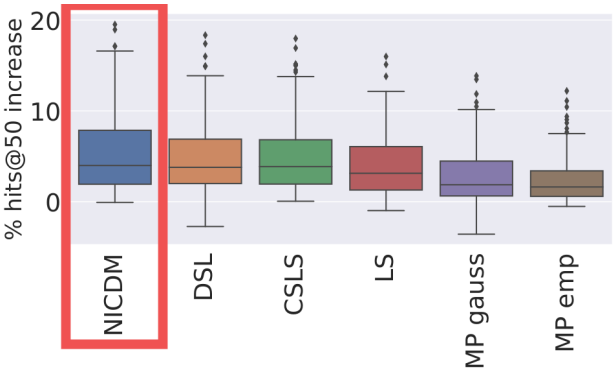
Does HR improve accuracy? (exact NN)



Improvement in hits@50 compared to no hubness reduction.
Aggregated over KGE approaches and datasets.

Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

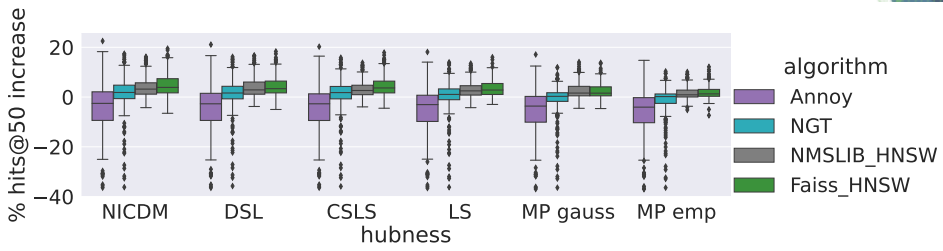
Does HR improve accuracy? (exact NN)



Improvement in hits@50 compared to no hubness reduction.
Aggregated over KGE approaches and datasets.

Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

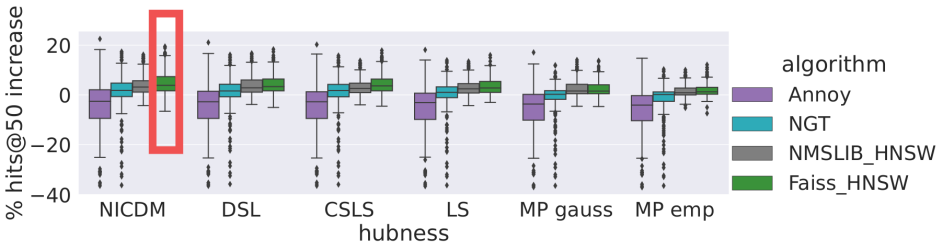
ANN + HR better than NN?



Improvement in hits@50 compared to baseline (no HR with exact NN).
Aggregated over KGE approaches and datasets.

Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

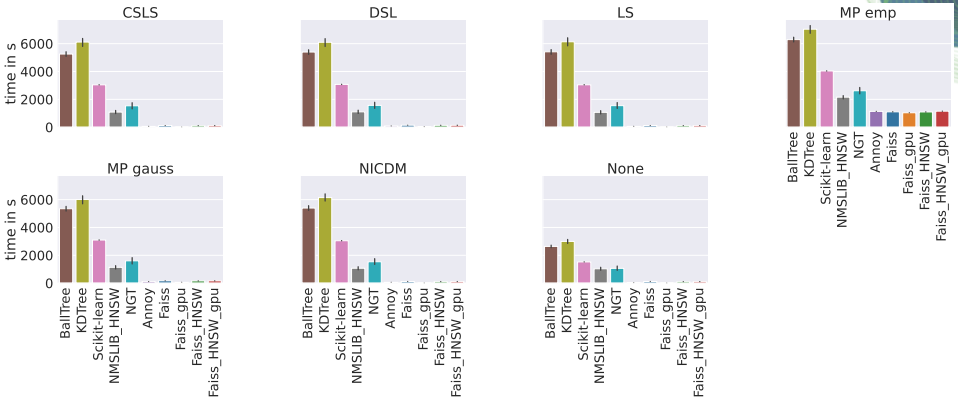
ANN + HR better than NN?



Improvement in hits@50 compared to baseline (no HR with exact NN).
Aggregated over KGE approaches and datasets.

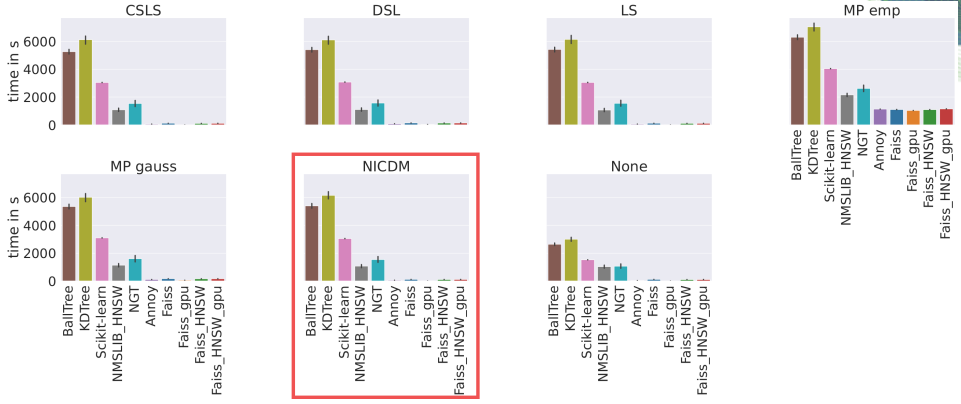
Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Speed comparison (large datasets)



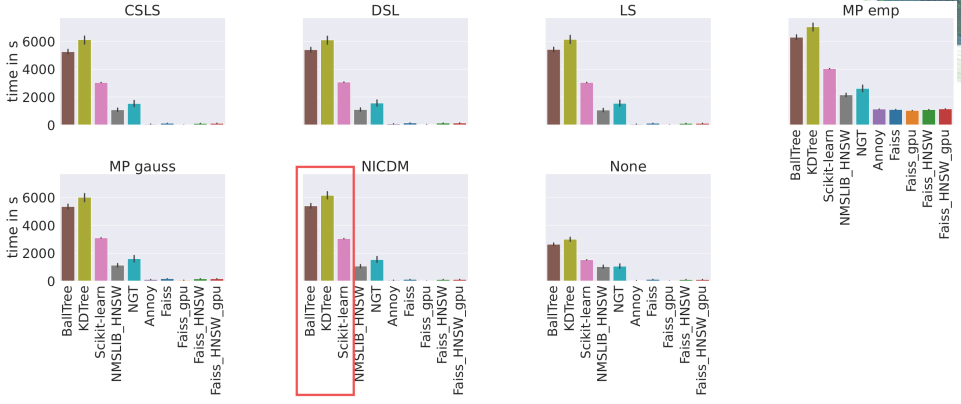
Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Speed comparison (large datasets)



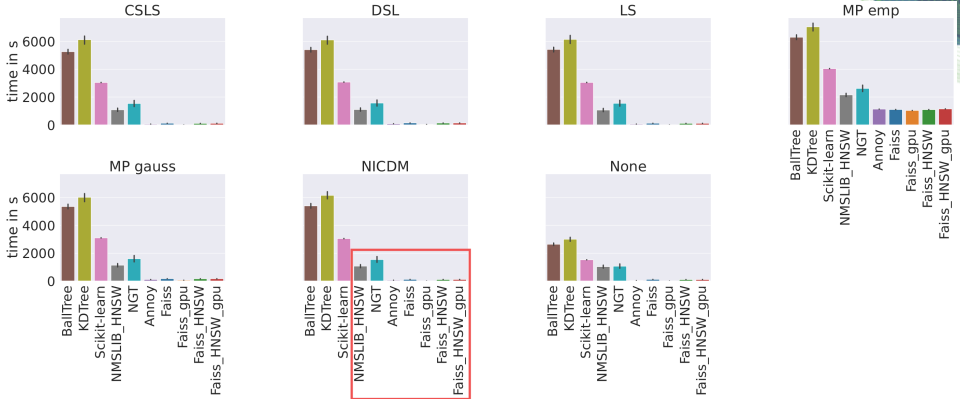
Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Speed comparison (large datasets)



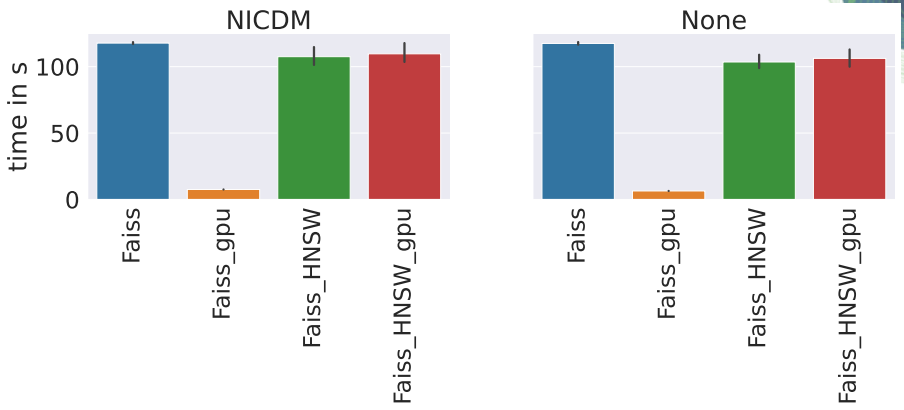
Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Speed comparison (large datasets)



Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Speed comparison (large datasets) only Faiss



Obraczka and Rahm (2021): "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings"

Conclusion

Conclusion

What we learned:

- Hubness reduction improves alignment results
- HR with ANN gives better AND faster results than exact NN
- Faiss + NICDM gave the fastest and most accurate results
- Faiss on a GPU goes a long way even with exact NN

Contact:
obraczka@informatik.uni-leipzig.de
github.com/dobraczka
dobraczka

Thank you for your attention!

References I

Bentley, Jon Louis (1975).
"Multidimensional Binary Search Trees Used for Associative Searching".
In: *Commun. ACM* 18.9, pp. 509–517. DOI: 10.1145/361002.361007.
URL: http://doi.acm.org/10.1145/361002.361007.

Feldbauer, Roman and Arthur Flexer (2019).
"A comprehensive empirical comparison of hubness reduction in high-dimensional spaces".
In: *Knowledge and Information Systems* 59.1, pp. 137–166. ISSN: 02193116.
DOI: 10.1007/s10115-018-1205-y. URL: https://doi.org/10.1007/s10115-018-1205-y.

Hara, Kazuo et al. (2016).
"Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness".
In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, pp. 1659–1665. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12055.

References II

Iwasaki, Masajiro (2016).
"Pruned Bi-directed K-nearest neighbor graph for proximity search".
In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9939 LNCS, pp. 20–33. ISBN: 9783319467580.
DOI: 10.1007/978-3-319-46759-7_2.
URL: https://link.springer.com/chapter/10.1007/978-3-319-46759-7_2.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2017).
"Billion-scale similarity search with GPUs". In: *ArXiv preprint* abs/1702.08734.
URL: https://arxiv.org/abs/1702.08734.

Lample, Guillaume et al. (2018). "Word translation without parallel data".
In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
URL: https://openreview.net/forum?id=H196sainb.

References III

- Malkov, Yu. A. (2018). "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs".
  In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 31–33.
  URL: https://arxiv.org/pdf/1603.09320.pdf.
- Obraczka, Daniel and Erhard Rahm (2021). "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings". In: *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2021, Volume 2: KEOD, Online Streaming, October 25-27, 2021*.
  Ed. by David Aveiro, Jan L. G. Dietz, and Joaquim Filipe. SCITEPRESS, pp. 28–39.
  DOI: 10.5220/0010646400003064.
  URL: https://dbs.uni-leipzig.de/file/KIEZ_KEOD_2021_Obraczka_Rahm.pdf.
- Omohundro, Stephen M. (1989). *Five Balltree Construction Algorithms*. Tech. rep.
  International Computer Science Institute. URL: https://omohundro.files.wordpress.com/2009/03/omohundro89_five_balltree_construction_algorithms.pdf.

References IV

Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”.
In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
URL: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Schnitzer, Dominik et al. (2012). “Local and global scaling reduce hubs in space”.
In: *Journal of Machine Learning Research* 13. ISSN: 15324435.
URL: https://jmlr.csail.mit.edu/papers/volume13/schnitzer12a/schnitzer12a.pdf.

Sun, Zequn et al. (2020).
“A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”.
In: *Proc. VLDB Endow.* 13.11, pp. 2326–2340.
URL: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.