Graph Embeddings for Enrichment of Historical Data

J. Baas M. M. Dastani A. J. Feelders
Information and Computing Sciences, Utrecht University

Utrecht University

Highlights

Objective

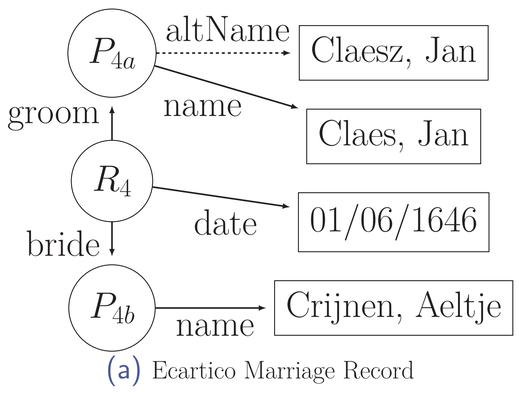
- Disambiguate references to people in historical records represented by knowledge graphs

Methodology

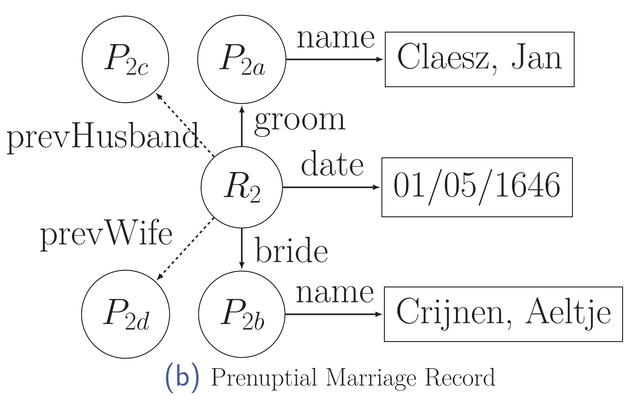
- Build a co-occurrence matrix using graph walks
- Use the co-occurrence matrix to create an embedding
- Find likely duplicate references by searching for nearest neighbors

Data

- We use data from the **Amsterdam City Archives** and **Ecartico**
- The archives contain baptism-, prenuptial marriage-, marriage- and burial records



(a) Ecartico Marriage Record



(b) Prenuptial Marriage Record

Figure 1: Input graph structures

Co-occurrence Matrix Creation

- Neighborhoods are created using the **Bookmark Coloring Algorithm** (BCA) [1]
- The combined neighborhoods together form a **sparse** co-occurrence matrix filled with **probability** values
- These probabilities are an approximation of the Personalized PageRank
- BCA can be biased with **weights** set by an expert user to emphasize certain relationships

pGloVe - Creating Embeddings

We have adapted the GloVe [2] loss function to work better with the **probability** values produced by the Bookmark Coloring Algorithm:

$$J = \sum_{i}^{|V|} \sum_{j}^{N(i)} X_{ij} \left(b_i + \bar{b}_j + w_i^\top \bar{w}_j - \log \frac{X_{ij}}{1 - X_{ij}} \right)^2 \quad (1)$$

- The weighing function $f(X_{ij})$ is dropped in favor of X_{ij}
- We transform the probability value into a real number in the range $(-\infty, \infty)$ using the **logit** function to facilitate the calculation of the vectors w_i and $\bar{w}_j$

We also experimented with gradient descent variations other than Adagrad. We obtained a **faster** convergence with **lower cost** when using **Amsgrad** [3] over Adagrad:

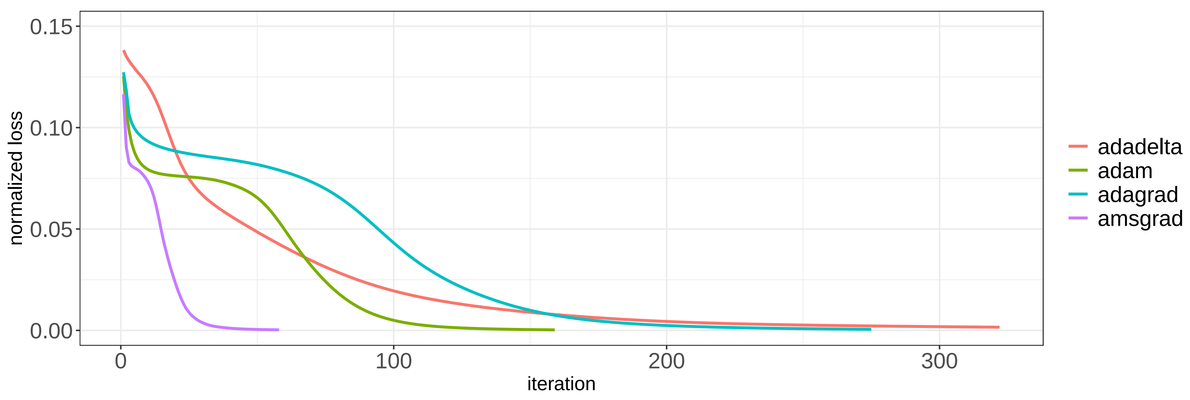


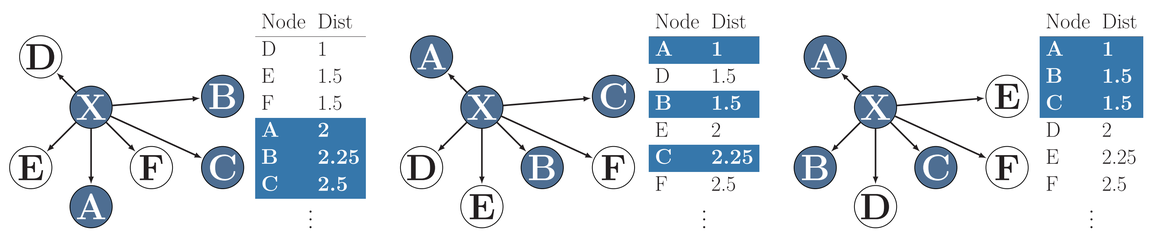
Figure 2: How different gradient descent algorithms impact pGloVe performance

Evaluation

We evaluate our embedding by comparison with a test-set created by a domain expert. The objective is to cluster together all references to persons which the expert annotated as being the same real-life person.

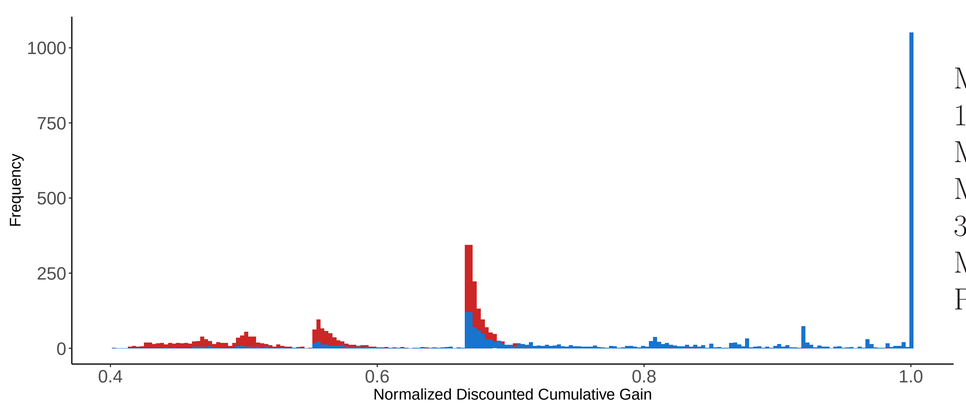
For each of the **2815** nodes annotated by the domain expert in the embedding:

- Rank all other annotated nodes based on **Canberra** distance
- Nodes which are identified by a domain expert as referencing the same person should appear on top
- Evaluate each ranking with **Normalized Discounted Cumulative Gain**



(a) Bad (b) Moderate (c) Perfect

Figure 3: Evaluations of node X based on the distance to relevant nodes



(a) NDCG score histogram of

Min.	0.40	0.42
1st Qu.	0.52	0.68
Median	0.67	0.88
Mean	0.60	0.84
3rd Qu.	0.67	1.0
Max.	1.0	1.0
Perfect	**0.2 %**	**37 %**

(b) Score summaries

Figure 4: Comparison between **GloVe** and **pGloVe**

Contact Information

- Web: www.goldenagents.org
- Email: j.baas@uu.nl
- Code: github.com/Jurian/graph-embeddings

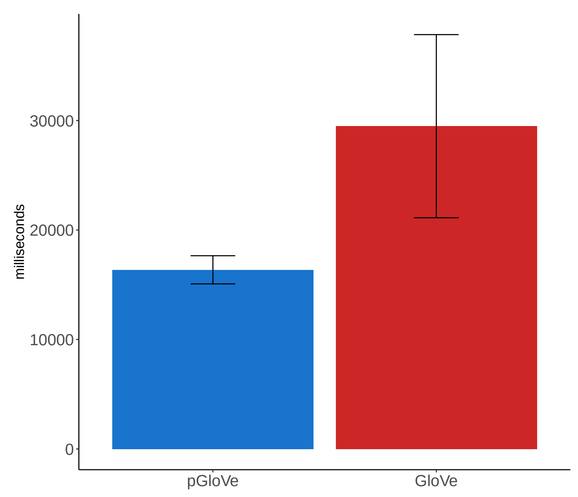


Figure 5: GloVe vs pGloVe mean convergence time using Amsgrad over 100 runs

Future Work

Historical data has its own peculiarities and difficulties. Such as the fact that often information is approximate: we do not know when someone was born exactly and there exist many variations of how to write down a certain name.

Conclusion

We have shown that

- Our adapted loss function performs better than the standard GloVe in combination with BCA using the Amsterdam City Archives data for entity resolution
- There is much room for improvement over the standard Adagrad in terms of convergence time and quality of local optimum

References

[1] Pavel Berkhin.
Bookmark-coloring algorithm for personalized pagerank computing.
Internet Mathematics, 3(1):41–62, 2006.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning.
Glove: Global vectors for word representation.
In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[3] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar.
On the convergence of adam and beyond.
arXiv preprint arXiv:1904.09237, 2019.

Acknowledgements

Supported by the Netherlands Organisation for Scientific Research. (NWO)