Inductive Learning of Concept Representations from Library-Scale Corpora with Graph Convolution

Inductive Learning of Concept Representations from Library-Scale Corpora with Graph Convolution

Lukas Galke[1] Tetyana Melnychuk[2] Eva Seidlmayer[3] Steffen Trog[4] Konrad U. Förstner[5] Carsten Schultz[6] Klaus Tochtermann[7]

Abstract: Automated research analyses are becoming more and more important as the volume of research items grows at an increasing pace. We pursue a new direction for the analysis of research dynamics with graph neural networks. So far, graph neural networks have only been applied to small-scale datasets and primarily supervised tasks such as node classification. We propose to use an unsupervised training objective for concept representation learning that is tailored towards bibliographic data with millions of research papers and thousands of concepts from a controlled vocabulary. We have evaluated the learned representations in clustering and classification downstream tasks. Furthermore, we have conducted nearest concept queries in the representation space. Our results show that the representations learned by graph convolution with our training objective are comparable to the ones learned by the DeepWalk algorithm. Our findings suggest that concept embeddings can be solely derived from the text of associated documents without using a lookup-table embedding. Thus, graph neural networks can operate on arbitrary document collections without re-training. This property makes graph neural networks useful for the analysis of research dynamics, which is often conducted on time-based snapshots of bibliographic data.

Keywords: machine learning; representation learning; neural networks; graph mining

1 Introduction

The investigation of bibliographic data enables rich insights into research dynamics including knowledge generation and diffusion, convergence of distinct scientific areas and substitution of some scientific fields with converged domains. New valuable knowledge is produced within scientific communities through collaboration of multiple actors [PHW12, WJU07]. A collaboration between researchers from different scientific fields fosters the diffusion of knowledge of one domain into other fields. The intensification of such collaborations leads to blurring the boundaries between separate scientific fields and to emerging scientific disciplines [CBL10].

[1] ZBW – Leibniz Information Centre for Economics, Kiel and Hamburg, Germany l.galke@zbw.eu
[2] Kiel University, Germany melnychuk@bwl.uni-kiel.de
[3] ZB MED – Information Centre for Life Sciences, Cologne, Germany seidlmayer@zbmed.de
[4] ZBW – Leibniz Information Centre for Economics, Kiel and Hamburg, Germany shtrog@gmail.com
[5] ZB MED – Information Centre for Life Sciences, Germany foerstner@zbmed.de
[6] Kiel University, Germany schultz@bwl.uni-kiel.de
[7] ZBW – Leibniz Information Centre for Economics, Kiel and Hamburg, Germany k.tochtermann@zbw.eu

doi:10.18420/inf2019_26

Library-scale corpora of scientific publications hold a large potential for automated analyses of research dynamics. Machine learning techniques that benefit from large amounts of data are essential for studying research dynamics. A major challenge in analysis of research dynamics is to derive a meaningful similarity measure. So far, most existing approaches rely on text-based similarity, co-citation analysis [NMF17, URU10], or scientometric methods [Je16, JLC18]. In contrast, we exploit concept annotations, as present in corpora of (scientific) digital libraries to derive a similarity measure between concepts. We make use of machine learning techniques to learn a low-dimensional continuous vector, i. e., a *representation* for each concept, from which a similarity measure can be derived.

Problem Statement In a paper-concept graph, we study the novel problem of learning representations for featureless concept nodes from paper nodes that have textual features. We evaluate whether the resulting concept representations are *meaningful*, i. e. correspond to human judgements, and *useful* in terms of their performance in downstream tasks.

Formally, we operate on a graph $\mathcal{G} = (\mathbb{P} \cup \mathbb{C}, \boldsymbol{X}, \boldsymbol{A})$, whose N vertices are either paper nodes $\mathbb{P}$ or concept nodes $\mathbb{C}$. Textual features of paper nodes are encoded in $\boldsymbol{X} \in \mathbb{R}^{|\mathbb{P}| \times L}$, where L is the textual feature dimension. Concept nodes have no features. Edges are encoded in the adjacency matrix $\boldsymbol{A}$ such that $A_{ij} > 0$ when either two papers $i, j < N$ have at least one common author or a paper i is annotated with concept j. The task is to learn a parametrized function f_θ that maps paper $\boldsymbol{X}, \boldsymbol{A}$ to concept representations $C \in \mathbb{R}^{|\mathbb{C}| \times d}$ of size d. To enable a fair comparison between methods, we keep d fixed because larger representation sizes tend to lead to increased performance in downstream tasks [ERG19].

We call a method *transductive* if it relies on a static concept embedding given by a look-up table. A method is *inductive*, when the concept representation C can be derived solely from the input corpus $\boldsymbol{X}, \boldsymbol{A}$ without conducting further training.

In this paper, we propose to use graph convolution [KW16a, Hu18, CZS18] to tackle this problem. To enable unsupervised representation learning, we introduce a dedicated training objective. We compare the resulting concept representations to the ones of transductive DeepWalk and text-based latent semantic analysis [De90].

Our results show that the representations learned by graph convolutional networks are similarly useful and meaningful as the representations learned by DeepWalk [PAS14]. At the same time, our graph convolution approach has the advantage that it does not rely on a static concept embedding but rather learns a mapping from associated papers to concept representations. This turns graph convolution into a valuable approach for the analyses of research dynamics. A once-learned model can induce concept representations for any (sub-)set of annotated research papers such as annual snapshots. This is important for the analyses of research dynamics, which we consider as future work.

In summary, our contributions are: (1) We apply state-of-the-art graph neural networks on a dataset of 2.1M papers from the economics and business studies domain. (2) We introduce

a dedicated, reconstruction-based training objective that allows unsupervised representation learning of concept representation. (3) We show that the learned concept representations are similarly useful and meaningful as the ones of DeepWalk, while not depending on a static node embedding.

In the following Section 2, we give an overview of the related work. Then, we describe the employed methods in Section 3, before we outline the experimental setup in Section 4. We provide the results in Section 5, discuss them in Section 6, before we conclude.

2 Related Work

Salatino, Osborne, and Motta [SOM17] have shown that there is a strong correlation between the pace of collaboration and the emergence of new topics. The same authors have then developed an advanced clique percolation method [SOM18] to detect emerging topics at the early stages and evaluate against a synthetic ground truth. Wu et al. [WVC16] have studied the top 1% authors within the computer science domain and show that research topics are increasingly inter-related. Duvvuru, Kamarthi and Sultornsanee [DKS12] link keywords when they appear in the same scholarly article. The authors construct visual keyword maps that may aid identifying emerging research areas. He et al. [He09] propose to use citation data in conjunction with latent Dirichlet allocation to analyze topic evolution. Tseng et al. [Ts09] compare several methods to detect hot topics.

Several approaches have been proposed that are targeted specifically towards multi-relational graphs such as knowledge graphs or linked data [Bo13, So13, Ya14]. For homogeneous graphs, as faced in our context, the successful Word2vec algorithm [Mi13] has been transferred to graphs by sampling random walks, namely DeepWalk [PAS14]. Node2vec [GL16] generalizes DeepWalk and further analyzes how the window size affects capturing more structural or more semantic relationships. Yang, Cohen, and Salakhutdinov [YCS16] outline the difference between inductive and transductive learning settings and develop an approach that is suited for both cases. All previously described methods rely on look-up table embeddings and are, thus, not suited for inductive learning.

Numerous methods have recently emerged that generalize convolution to graphs [DBV16, KW16a]. In GraphSAGE [HYL17], the authors explore different aggregation functions and conduct experiments on representation learning in large-scale graphs by sampling adjacent nodes. Velickovic et al. [Ve18] suggest to incorporate an attention mechanism for neighbor aggregation. We refer to [Wu19] for a recent overview on graph neural networks.

3 Inductive Representation Learning with Graph Convolution

Graph convolution is an approach for graph-structured data that is capable of jointly exploiting textual and structural features. Approaches based on graph convolution yield

promising results on link prediction [KW16b], semi-supervised classification [KW16a], and representation learning [HYL17]. A benefit of graph convolution is the possibility to conduct inductive learning [YCS16, HYL17]. Inductive learning means that the textual features from paper nodes are aggregated to compose a representation of the featureless concept nodes. This property distinguishes this approach from other approaches that learn a static node embedding such as DeepWalk [PAS14] as well as TransE [Bo13] and their extensions. The inductive property allows computing concept representations on the basis of any subset of the data and also for entirely unseen data [GVS19].

To make use of graph convolution, we first embed the textual features into a lower-dimensional space by averaging word vectors $\boldsymbol{h}^{(0)} = \frac{1}{|x|} \sum_{t \in \boldsymbol{x}} \boldsymbol{W}^{(0)}_{t,:}$, where x are the words of a document. Subsequently, we make use of graph convolution to aggregate neighbor representations after a nonlinear transform. The representation of node i in layer l is defined as:

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} \boldsymbol{W}^{(l)} \boldsymbol{h}_j^{(l)} + \boldsymbol{b}^{(l)}\right)$$

where $\mathcal{N}(\cdot)$ refers to the set of adjacent nodes and σ is a nonlinear activation function. We follow [HYL17, Hu18, CZS18] and use mean aggregation $c_{ij} = |\mathcal{N}(i)|$. The weights $\boldsymbol{W}^{(0)}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \ldots, \boldsymbol{W}^{(k)}, \boldsymbol{b}^{(k)}$, with k being the depth of the network, are then optimized with respect to the training objective, which we describe in the following section.

Training Objective Unsupervised deep learning techniques exploit auxiliary objectives such as auto-encoding [BCV13]. An auto-encoding objective refers to the task of reconstructing the input. It may happen that the input-output space is high-dimensional. For instance, consider the vocabulary of all words. In these cases, normalizing across all output probabilities via softmax can become computationally expensive. Negative sampling [Mi13] approximates the softmax by sampling few negative outputs. The task is then to distinguish the true output among the negative samples. Due to its higher efficiency, negative sampling often yields higher effective scores than the exact computation of the softmax [Mi13].

In the graph domain, link prediction is a common choice for learning node representations. The representation is trained for predicting whether a link between two nodes exists. This can be regarded as auto-encoding the adjacency matrix [KW16a]. Also here, negative sampling can be employed to approximate the full softmax [PAS14, HYL17].

In our case, the goal is to learn concept representations for a controlled vocabulary. While our models deal with millions of research papers, the dimension of the controlled vocabulary is rather small with 5,688 concepts. We chose to use only concepts from the controlled vocabulary as optimization objective. Thus, we can afford to compute the full softmax over the concepts. We employ a linear decoder $g : \mathbb{R}^d \rightarrow \mathbb{R}^{|C|}$ that uses the final representation

of the graph convolutional network to reconstruct the respective concept. The loss function is the softmax over the concepts:

$$\mathcal{L}_{\text{rec}}(\boldsymbol{X}, \boldsymbol{A}, y) = -\log \frac{\exp g(f(\boldsymbol{X}, \boldsymbol{A}))[y]}{\sum_j \exp g(f(\boldsymbol{X}, \boldsymbol{A}))[j]}$$

where f is a graph convolutional encoder, j iterates through all concepts and $[\cdot]$ denotes index access. We sample a set of documents $\boldsymbol{X}$ which are connected over at most two hops to the true concept y. The graph convolutional encoder f then constructs a low-dimensional concept representation $f(\boldsymbol{X}, \boldsymbol{A})$, which is then used by g to reconstruct the true concept. Since g is discarded after training, we deactivate its bias term such that all information for prediction of the concept is drawn from the representation.

Neighbor Sampling and Skip Connections The originally proposed graph autoencoders [KW16b] and graph convolutional networks [KW16a] operate on the whole graph in each optimization step. Storing the dense adjacency matrix is, however, not an option when the dataset is of large scale. The authors suggest to construct mini-batches with adjacent nodes. However, the receptive field still grows exponentially with the number of layers. Hamilton et al. [HYL17] instead propose to subsample adjacent nodes such that the growth factor is constant. Unfortunately, subsampling does not guarantee convergence to the full graph convolution [CZS18]. A control variate approach has been proposed that provably converges to the optimal, full graph convolution solution with only two sampled neighbors [CZS18]. The authors propose to keep track of past activations to incorporate the difference into the forward propagation path. Huang et al. [Hu18] propose a sampling approach that makes use of skip-connections to preserve second-order connections throughout the sampling process. We adopt these advances and employ a graph convolutional network with control variate sampling and skip-connections. We do not insert self-loops, such that the inductive property is retained. After training, we use all neighbors for creating the final representations.

4 Experimental Setup

In the following Section 4.1, we will describe the characteristics of the dataset and the processing of textual and structural features. We described the employed baselines in Section 4.2 and denote the selected hyperparameters in Section 4.3, before we describe the evaluation measures in Section 4.4.

4.1 Dataset

The EconBiz dataset comprises more than 11M records describing scientific publications from the economics and business studies domain. About 5.8M of these records are well

described by a controlled vocabulary and are used for our investigations. We filter these publications for English language and for annotations from the polyhierarchically-organized Standardthesaurus Wirtschaft[8]. These annotations are created by professional subject indexers. The resulting subset consists of 2.1M publications along with 5,688 subjects from the controlled vocabulary. As we focus on concept representations, we collate the authorship edges between authors and papers. We create an edge between two papers if the two papers have an author in common. This effectively enlarges the size of the receptive field of the models by one hop. This holds not only for graph convolution, but also for DeepWalk.

We consider the titles of the documents as textual features. We have shown in prior work that using titles is competitive [Ge17] to full-text data for multi-label classification. When the amount of available title data exceeds the amount of full-text data, classifiers based on title data can even outperform classifiers based on full-text data [MGS18]. Thus, we employ the larger amount of available title data. For preprocessing, we remove punctuation and other non-alphanumeric characters, lowercase the text, and remove English stop-words. We compose a vocabulary of the 50,000 most-common words.

4.2 Baselines: LSA and DeepWalk

As baselines, we consider DeepWalk [PAS14] as a representative for a purely structural approach to graph representation learning along with latent semantic analysis [De90] as a well-known text-based approach for document-level similarity.

Latent semantic analysis [De90, MRS08] (LSA) is a technique to embed text documents into a lower dimensional space. The key idea of LSA is to factorize the term frequency–inverse document frequency [SB88] weighted term-document matrix. We apply LSA on the titles of the research papers [Ge17]. We employ truncated singular value decomposition to embed each document in a low dimensional vector space. Finally, we compute the centroid for each concept across those documents that are annotated with the respective concept.

DeepWalk [PAS14] is an approach for learning node embeddings in graph-structured data. The algorithm samples random walks through the graph structure. For each node in the path, its embedding is used to predict its predecessors and successors along the random walk. The embedding is initialized randomly and updated according to hierarchical softmax loss.

4.3 Hyperparameters

LSA uses 5 epochs for singular value decomposition of the term-document matrix. For DeepWalk, we generate 40,000 random walks for each concept node with a walk length of 3. We then run skip-gram optimization with window size 3 for 5 epochs over the generated

[8] http://zbw.eu/stw

random walks. The graph convolutional network uses two graph convolution layers. The text embedding size is 256 along with 128 hidden units and 128 output units corresponding to the representation size. We create mini-batches over concept nodes and sample 10 neighbors for each of the two hops. We run one sampling step per concept over 400 epochs. We use ReLU activation function and dropout [Ni14] with probability 0.5 within the GCN layers. We optimize the training objective via Adam [KB14] and an initial learning rate of 0.001. For a fair comparison, we fix the representation size to 128 for all models. Furthermore, the parameters are set such that both GCNs and DeepWalk are given the same number of sampled documents. We select a window size of 3 for DeepWalk such that the number of considered hops is the same as for GCNs.

4.4 Evaluation measures

To evaluate the resulting representations, we compare the performance on two downstream tasks: classification and clustering. For this purpose, we construct a dataset that maps each concept to its respective subthesaurus. The models have never seen the underlying concept hierarchy. As the thesaurus is organized in a polyhierarchic way, we use only those concepts, which belong to exactly one subthesaurus. We are left with 3,113 concepts and 7 classes.

Supervised Clustering We conduct a clustering on top of the learned concept representations with k-Means and k-Means++ as initialization strategy. We fix the number of clusters to 7 corresponding to the number of classes. We evaluate the supervised clustering metrics homogeneity, completeness, and V measure [RH07], as well as the adjusted rand index [HA85]. Homogeneity yields values between 0 and 1, which assess to which extent the clusters cover data points of the same class. Completeness is equivalent to homogeneity but switches the true and the predicted labels. V measure is the harmonic mean between homogeneity and completeness. The adjusted rand index is bounded between -1 and 1 and symmetrically assesses the similarity of a clustering result with the class labels. It is permutation-invariant and adjusted against chance. We report the mean scores of 100 k-Means runs for the raw concept vectors and L2-normalized concept vectors.

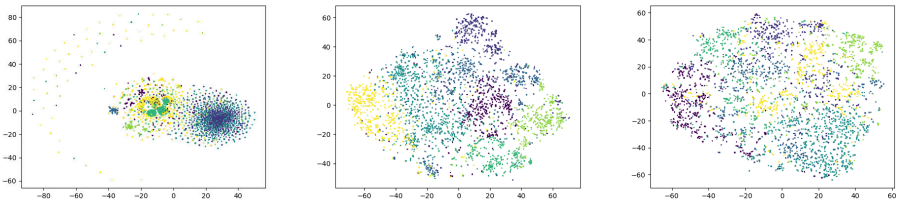
Unsupervised Clustering To gain more insights on the clustering tendency of the learned representations, we conduct a further unsupervised clustering experiment. Now we set the number of clusters to 101 corresponding to the number of top-level concepts across the 7 subthesauri. We evaluate the unsupervised clustering metrics silhouette coefficient [Ro87] and the Calinski-Harabasz criterion [CH74]. The silhouette coefficient is bounded between -1 and 1 and gives the ratio between intra-cluster distances and the pairwise distances to data points of the nearest cluster. The Calinski-Harabasz criterion compares the intra-cluster variance against the global between-cluster variance in distances. We also report these unsupervised clustering metrics for the supervised clustering experiments described above.

Classification We evaluate the performance in a downstream classification task. We use the L2-normalized learned concept representations as input and the corresponding subthesaurus as class label. As a common classifier we employ a support vector machine with linear kernel. We conduct a ten-fold cross-validation and report the mean accuracy.

5 Results

Tab. 1: Silhouette score (S), Calinski-Harabasz score (CH), homogeneity (H), completeness (C), V measure (V), and adjusted rand index (ARI) of clustering results on the learned concept representations for LSA, DeepWalk, and GCNs. We provide the mean of 100 k-Means runs with 7 clusters on 3,113 concept representations. Higher is better.

Model	Norm	S	CH	H	C	V	ARI
Random	None	0.0062	13.83	0.0032	0.0030	0.0031	0.0000
Random	Unit L2	0.0062	13.92	0.0033	0.0031	0.0032	0.0001
LSA	None	-0.0207	53.45	0.0030	0.0071	0.0042	-0.0041
LSA	Unit L2	**0.1284**	96.44	0.0022	0.0025	0.0023	-0.0009
DeepWalk	None	0.0194	124.80	0.2165	0.2496	0.2318	0.1852
DeepWalk	Unit L2	0.0670	131.18	**0.2930**	**0.2810**	**0.2869**	**0.1981**
GCN	None	0.0667	171.13	0.1845	0.1761	0.1802	0.1178
GCN	Unit L2	0.0823	**193.64**	0.1992	0.1891	0.1940	0.1423



(a) (b) (c)

Fig. 1: t-SNE visualization (perplexity=30) of the L2-normalized learned representations by LSA (a), Deepwalk (b) and GCN (c). The colors correspond to one clustering result with 7 clusters.

Table 1 shows the results for the supervised clustering task. For the supervised clustering metrics, the Deepwalk representations achieve the highest scores of 0.2930 homogeneity, 0.2810 completeness, 0.2869 V measure and adjusted rand index 0.1981. The scores of GCN representations are behind with a margin of 0.1 V-measure and 0.08 adjusted rand index. LSA representations yield the highest silhouette coefficient while GCN representations yield the highest Calinksi-Harabasz score. We provide a visualization of one clustering run in Figure 1.

Tab. 2: Mean silhouette score and Calinski-Harabasz score across 100 k-Means runs for the unsupervised clustering experiments with 101 clusters on learend representations of 5,688 concepts.

Model	Norm	Silhouette	Calinski-Harabasz
LSA	None	0.0543 (SD: 0.01)	32.14 (SD: 0.26)
LSA	Unit L2	0.0909 (SD: 0.01)	25.05 (SD: 0.13)
DeepWalk	None	0.0383 (SD: 0.00)	52.50 (SD: 0.34)
DeepWalk	Unit L2	0.0688 (SD: 0.00)	53.31 (SD: 0.13)
GCN	None	0.0721 (SD: 0.00)	72.78 (SD: 0.24)
GCN	Unit L2	**0.1005 (SD: 0.00)**	**84.88 (SD: 0.20)**

The results for the unsupervised clustering tasks with 101 clusters are shown in Table 2. Here, the GCN representations lead to the highest silhouette and Calinski-Harabasz scores of 0.1005 and 84.88, respectively.

In Table 3, we show the nearest-concepts for manually-selected concepts. The LSA representations fail to yield consistently explainable responses. For example, *Tax* is closest to *Rehabilitation hospital* and *Abortion*. The responses by GCN's and DeepWalk's representations are similarly acceptable: the closest concepts to *Tax* are in both cases all related to taxes. In case of *Germany*, DeepWalk returns other European countries but also *Comparison*. GCN yields parts of Germany along with Western Europe and Austria. We note that also linear relationships are resembled by both GCN and DeepWalk. For instance, the sum of the *Tax* vector and the *Theory* vector has *Theory of Taxation* among the two nearest concepts in the representation space. Similarly, the addition of *Economic growth* and *Theory* leads to having *Growth Theory* among the top two nearest concepts.

Table 4 shows the results for the downstream classification task. The non-normalized GCN representation achieves the highest classification accuracy of 68%. The highest scores for LSA and DeepWalk are 23% and 67%, respectively.

6 Discussion

Our results show that the representations of graph convolution are comparable to the ones of DeepWalk. While DeepWalk has lead to higher scores in the clustering task, GCN's representations have lead to higher scores in the classification downstream task. By inspecting the representations with nearest neighbor queries, we could observe that both DeepWalk and GCN correspond to human intuition, while LSA falls behind.

We have further analyzed the usefulness of the learned representations in an unsupervised clustering task with 101 clusters, enforcing a more fine-grained setting. In this setting, the

Tab. 3: Most similar concepts according to learned representations of LSA, DeepWalk, and GCN. The responses are ordered by descending cosine similarity to the vector of the query concept. A plus in the query column indicates that we use the sum of two concept vectors as query.

Query	LSA	DeepWalk	GCN
Economic growth	Management information system	Economic adjustment	Stages of growth model
	Tobacco	Economic policy	Growth policy
	Internet Usage	Growth policy	Resource wealth
	Eurobond	Economic development	Kuznets curve
	Automobile engine	Economic reform	Export-led growth
Tax	Rehabilitation hospital	Fiscal administration	Tax policy
	Abortion	Tax system	Tax system
	Biodiversity	Tax policy	Tax reform
	Financial statement analysis	Sales tax	Taxation procedure
	Association agreement	Tax reform	Tax burden
Germany	Debt crisis	Italy	East Germany
	Mesoeconomics	France	Austria
	Population policy	Comparison	West Germany
	Complaint management	Netherlands	Lower Saxony
	Unemployment theory	Austria	Western Europe
Vehicle	Pigouvian tax	Transport research	Sustainable mobility
	Cargo shipping	Transport economics	Passenger transport
	Cyclical unemployment	Waste treatment	Freight transport
	Wage subsidy	Battery	Major electrical appliances
	Financial Statement analysis	Microsystems	Traffic
Tax + Theory	Tax	Tax	Theory of taxation
	Theory	Theory of taxation	Theory
	Financial statement analysis	Tax system	Second best
	Nursing profession	Capital income	Optimal taxation
	Rehabilitation hospital	Public economics	Welfare economics
Economic growth + Theory	Economic growth	Economic growth	Growth theory
	Banking services	Growth theory	Neoclassical growth model
	Producer cooperative	Economic model	Unbalanced growth
	Licence	Theory	Balanced growth
	Laboratory	Endogenous growth model	Functional income distribution

GCN's representations have yielded the highest silhouette coefficient and Calinski-Harabasz score.

The strong performance of the DeepWalk is to some extent surprising, as it does not use any textual features but only relies exclusively on the structure of the author-paper-concept graph. This, however, confirms the claim of the original work [PAS14] that meaningful node embeddings can be derived without using node attributes.

There is no ground truth for pairwise similarity between concepts. We could therefore evaluate only a small subset of nearest-concept queries manually. We, however, did create a dataset which maps each concept to the respective subthesaurus. The hierarchical

Tab. 4: Downstream classification performance with 3,113 concepts and 7 classes. We list mean and standard deviation from a 10-fold cross-validation using a linear SVM classifier.

Model	Norm	Accuracy
LSA	None	0.2345 (0.00)
LSA	Unit L2	0.2181 (0.02)
DeepWalk	None	0.6625 (0.04)
DeepWalk	Unit L2	0.6708 (0.03)
GCN	None	**0.6813 (0.03)**
GCN	Unit L2	0.6496 (0.03)

relationships were never presented to the models, but only used for evaluation. The assumption is that the learned representation should allow distinguishing the concepts on a very broad level such as "Economics", "Business economics", "Geographic Names". A limitation of our study is that this categorization could be too broad to fully assess similarity among concepts. Constructing a more fine-grained evaluation set is challenging because the underlying thesaurus is polyhierarchic, i.e., a concept can have multiple broader concepts. Our subthesauri-based evaluation set uses only concepts that belong to exactly one subthesaurus, despite following all paths upwards in the hierarchy, which renders it well-defined, even in the polyhierarchic case.

We have applied our concept representation learning method to a large-scale dataset with 2.1M publications from the economics and business studies domain. Our approach can be transferred to any other dataset that is annotated with concepts. These concepts may come from a controlled vocabulary as in our case but free-text author keywords can be used instead. When scaling the number of concepts up, it can become necessary to switch from softmax training to a negative sampling approximation. Our model is flexible in the sense that it allows incorporating further edges such as the broader and narrower connections between the concepts. For now, we held out these connections for evaluation purposes.

The inductive property of the graph convolution approach enables us to map any set of annotated papers to concept representations without retraining. We can incrementally update the concept representations in a time-dynamic setting. This is important because fine-tuning pretrained, non-inductive, embeddings can be challenging: there are many options for weighting between the old embedding and the updates. We envision that this property will be crucial for analyses of the dynamics within and across research fields in our future work.

7 Conclusion

We conclude that the representations learned by graph neural networks are comparable to the ones learned by DeepWalk. Graph neural networks can induce representations for the

featureless concepts from the titles of associated research papers. To make graph neural networks applicable to library-scale bibliographic corpora, we have introduced a specific training objective for learning concept representations.

We have thoroughly analyzed the learned representations by conducting supervised and unsupervised downstream tasks. Furthermore, we have manually inspected the representations by conducting nearest neighbor queries. We have found that the nearest concepts are useful in downstream tasks and meaningful for humans, even in cases, where the vectors of two concepts are summed up.

Our findings suggest that concept embeddings can be solely derived from the text of associated documents without using a lookup-table embedding. In future work, we plan to make use of further structural features such as concept hierarchies. We further plan to use graph neural networks for dynamic research analyses based on annual snapshots of research papers. By analysing the trajectories, we can then make claims about the convergence and divergence of research areas.

Source Code: `github.com/lgalke/INFORMATIK2019-concept-representation-learning`

Acknowledgment: This work was supported by BMBF within the programme "Quantitative Wissenschaftsforschung" under grant numbers 01PU17013A, 01PU17013B, 01PU17013C.

Bibliography

[BCV13] Bengio, Yoshua; Courville, Aaron C.; Vincent, Pascal: Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell., 35(8), 2013.

[Bo13] Bordes, Antoine; Usunier, Nicolas; García-Durán, Alberto; Weston, Jason; Yakhnenko, Oksana: Translating Embeddings for Modeling Multi-relational Data. In: NIPS. 2013.

[CBL10] Curran, Clive-Steven; Bröring, Stefanie; Leker, Jens: Anticipating converging industries using publicly available data. Technological Forecasting and Social Change, 77(3), 2010.

[CH74] Caliński, T.; Harabasz, J: A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1974.

[CZS18] Chen, Jianfei; Zhu, Jun; Song, Le: Stochastic Training of Graph Convolutional Networks with Variance Reduction. In: ICML 2018. 2018.

[DBV16] Defferrard, Michaël; Bresson, Xavier; Vandergheynst, Pierre: Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In: NIPS. 2016.

[De90] Deerwester, Scott; Dumais, Susan T; Furnas, George W; Landauer, Thomas K; Harshman, Richard: Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 1990.

[DKS12] Duvvuru, Arjun; Kamarthi, Sagar; Sultornsanee, Sivarit: Undercovering research trends: Network analysis of keywords in scholarly articles. In: 2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2012.

[ERG19] Eger, S.; Rücklé, A.; Gurevych, I.: Pitfalls in the Evaluation of Sentence Embeddings. arXiv e-prints, June 2019.

[Ge17] Galke, Lukas; et al.: Using Titles vs. Full-text as Source for Automated Semantic Document Annotation. In: K-CAP. ACM, 2017.

[GL16] Grover, Aditya; Leskovec, Jure: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016.

[GVS19] Galke, Lukas; Vagliano, Iacopo; Scherp, Ansgar: Can Graph Neural Networks Go "Online"? An Analysis of Pretraining and Inference. In: Representation Learning on Graphs and Manifolds, ICLR Workshop. 2019.

[HA85] Hubert, Lawrence; Arabie, Phipps: Comparing partitions. Journal of Classification, 2(1), Dec 1985.

[He09] He, Qi; Chen, Bi; Pei, Jian; Qiu, Baojun; Mitra, Prasenjit; Giles, C. Lee: Detecting topic evolution in scientific literature: how can citations help? In: CIKM. ACM, 2009.

[Hu18] Huang, Wen-bing; Zhang, Tong; Rong, Yu; Huang, Junzhou: Adaptive Sampling Towards Fast Graph Representation Learning. In: NeurIPS. 2018.

[HYL17] Hamilton, William L.; Ying, Zhitao; Leskovec, Jure: Inductive Representation Learning on Large Graphs. In: NIPS. 2017.

[Je16] Jeong, Dae-hyun; Cho, Keuntae; Park, Sangyong; Hong, Soon-ki: Effects of knowledge diffusion on international joint research and science convergence: Multiple case studies in the fields of lithium-ion battery, fuel cell and wind power. Technological Forecasting and Social Change, 108, 2016.

[JLC18] Jeong, Daehyun; Lee, Kyuhong; Cho, Keuntae: Relationships among international joint research, knowledge diffusion, and science convergence: the case of secondary batteries and fuel cells. Asian Journal of Technology Innovation, 26(2), 2018.

[KB14] Kingma, Diederik P.; Ba, Jimmy: Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980, 2014.

[KW16a] Kipf, Thomas N.; Welling, Max: Semi-Supervised Classification with Graph Convolutional Networks. CoRR, abs/1609.02907, 2016. Published at ICLR 2017.

[KW16b] Kipf, Thomas N.; Welling, Max: Variational Graph Auto-Encoders. CoRR, abs/1611.07308, 2016.

[MGS18] Mai, Florian; Galke, Lukas; Scherp, Ansgar: Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text. In: JCDL. ACM, 2018.

[Mi13] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Gregory S.; Dean, Jeffrey: Distributed Representations of Words and Phrases and their Compositionality. In: NIPS. 2013.

[MRS08] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich: Introduction to information retrieval. Cambridge University Press, 2008.

[Ni14] Nitish, Srivastava; Hinton, Geoffrey E.; Krizhevsky, Alex; Sutskever, Ilya; Salakhutdinov, Ruslan: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 2014.

[NMF17] Niemann, Helen; Moehrle, Martin G.; Frischkorn, Jonas: Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. Technological Forecasting and Social Change, 115, 2017.

[PAS14] Perozzi, Bryan; Al-Rfou, Rami; Skiena, Steven: DeepWalk: online learning of social representations. In: KDD. ACM, 2014.

[PHW12] Phelps, Corey; Heidl, Ralph; Wadhwa, Anu: Knowledge, Networks, and Knowledge Networks: A Review and Research Agenda. Journal of Management, 38(4), 2012.

[RH07] Rosenberg, Andrew; Hirschberg, Julia: V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In: EMNLP-CoNLL. ACL, 2007.

[Ro87] Rousseeuw, Peter J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 1987.

[SB88] Salton, Gerard; Buckley, Christopher: Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 1988.

[So13] Socher, Richard; Chen, Danqi; Manning, Christopher D; Ng, Andrew: Reasoning With Neural Tensor Networks for Knowledge Base Completion. In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013.

[SOM17] Salatino, Angelo Antonio; Osborne, Francesco; Motta, Enrico: How are topics born? Understanding the research dynamics preceding the emergence of new areas. PeerJ Computer Science, 3, 2017.

[SOM18] Salatino, Angelo Antonio; Osborne, Francesco; Motta, Enrico: AUGUR: Forecasting the Emergence of New Research Topics. In: JCDL. ACM, 2018.

[Ts09] Tseng, Yuen-Hsien; Lin, Yu-I; Lee, Yi-Yang; Hung, Wen-Chi; Lee, Chun-Hsiang: A comparison of methods for detecting hot topics. Scientometrics, 81(1), 2009.

[URU10] Upham, S. Phineas; Rosenkopf, Lori; Ungar, Lyle H.: Innovating knowledge communities. Scientometrics, 83(2), 2010.

[Ve18] Veličković, Petar; Cucurull, Guillem; Casanova, Arantxa; Romero, Adriana; Liò, Pietro; Bengio, Yoshua: Graph Attention Networks. International Conference on Learning Representations, 2018.

[WJU07] Wuchty, Stefan; Jones, Benjamin F.; Uzzi, Brian: The Increasing Dominance of Teams in Production of Knowledge. Science, 316(5827), 2007.

[Wu19] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S.: A Comprehensive Survey on Graph Neural Networks. arXiv e-prints, January 2019.

[WVC16] Wu, Yan; Venkatramanan, Srinivasan; Chiu, Dah Ming: Research collaboration and topic trends in Computer Science based on top active authors. PeerJ Computer Science, 2, 2016.

[Ya14] Yang, Bishan; Yih, Wen-tau; He, Xiaodong; Gao, Jianfeng; Deng, Li: Embedding Entities and Relations for Learning and Inference in Knowledge Bases. CoRR, abs/1412.6575, 2014.

[YCS16] Yang, Zhilin; Cohen, William W.; Salakhutdinov, Ruslan: Revisiting Semi-Supervised Learning with Graph Embeddings. In: ICML. volume 48 of JMLR Workshop and Conference Proceedings. JMLR.org, 2016.