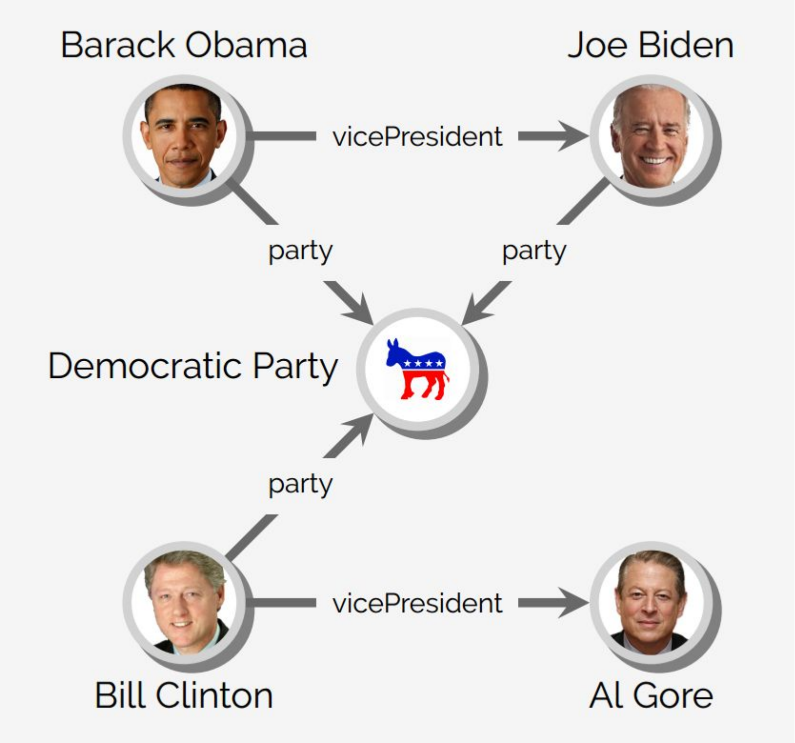
Seminar Knowledge Graph Analysis

Dr. Asja Fischer, Dr. Hajira Jabeen, Prof. Jens Lehmann

Knowledge Graphs

- Modelling entities and their relationships
- Analysis: finding underlying structure of graph e.g. to predict unknown relationships
- Examples: Google Knowledge Graph, DBpedia, Facebook, YAGO, Twitter, LinkedIn, MS Academic Graph, WikiData



Organisation

- Website: https://sewiki.iai.uni-bonn.de/teaching/lectures/kga/2016/seminar
- Presence dates:
    - October 25th Topic Selection
    - 3 slots for final presentations: 10./17./24. January
- Topic:
    - Each topic is a research paper
    - Summarise content and present it to the class (including your opinion on it)
- Final presentation: 15-20 minutes + questions
- Selection: listen to overview and pick one or two favorites

1. Knowledge Vault

Abstract: Recent years have witnessed a proliferation of large-scale knowledge bases, including Wikipedia, Freebase, YAGO, Microsoft’s Satori, and Google’s Knowledge Graph. To increase the scale even further, we need to explore automatic methods for constructing knowledge bases. Previous approaches have primarily focused on text-based extraction, which can be very noisy. Here we introduce Knowledge Vault, a **Web-scale probabilistic knowledge base that combines extractions from Web content (obtained via analysis of text, tabular data, page structure, and human annotations) with prior knowledge derived from existing knowledge repositories**. We employ supervised machine learning methods for fusing these distinct information sources. The Knowledge Vault is substantially **bigger than any previously published structured knowledge repository**, and features a probabilistic inference system that computes calibrated probabilities of fact correctness. We report the results of multiple studies that explore the relative utility of the different information sources and extraction methods.

Citation: Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 601-610). ACM.
PDF: https://www.cs.ubc.ca/~murphyk/Papers/kv-kdd14.pdf
Tutor: Prof. Jens Lehmann

2. Distant supervision for relation extraction

Abstract: Modern models of **relation extraction** for tasks like ACE are based on supervised learning of relations from small hand-labeled corpora. We investigate an alternative **paradigm that does not require labeled corpora**, avoiding the domain dependence of ACEstyle algorithms, and allowing the use of corpora of any size. Our experiments use Freebase, a large semantic database of several thousand relations, to provide distant supervision. For each pair of entities that appears in some Freebase relation, we find all sentences containing those entities in a large unlabeled corpus and extract textual features to train a relation classifier. Our algorithm combines the advantages of supervised IE (combining 400,000 noisy pattern features in a probabilistic classifier) and unsupervised IE (extracting large numbers of relations from large corpora of any domain). Our model is able to extract 10,000 instances of 102 relations at a precision of 67.6%. We also analyze feature performance, showing that syntactic parse features are particularly helpful for relations that are ambiguous or lexically distant in their expression.

Citation: Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1003-1011). Association for Computational Linguistics.
PDF: http://web.stanford.edu/~jurafsky/mintz.pdf
Tutor: Dr. Hajira Jabeen

3. Wikidata

Abstract: UNNOTICED BY MOST of its readers, Wikipedia continues to undergo dramatic changes, as its sister project Wikidata introduces a new multilingual “**Wikipedia for data**” (http://www.wikidata.org) to manage the factual information of the popular online encyclopedia. With Wikipedia’s data becoming cleaned and integrated in a single location, opportunities arise for many new applications. Originally conceived in 2001 as a mainly text-based resource, **Wikipedia has collected increasing amounts of structured data, including numbers, dates, coordinates, and many types of relationships, from family trees to the taxonomy of species**. It has become a resource of enormous value, with potential applications across all areas of science, technology, and culture. This development is hardly surprising, given that Wikipedia is committed to “a world in which every single human being can freely share in the sum of all knowledge,”

Citation: Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.
PDF: http://korrekt.org/papers/Wikidata-CACM-2014.pdf
Tutor: Prof. Jens Lehmann

4. A Neural Knowledge Language Model

Abstract: Communicating knowledge is a primary purpose of language. However, current language models have significant limitations in their ability to encode or decode knowledge. This is mainly because they acquire knowledge based on statistical co-occurrences, even if most of the knowledge words are rarely observed named entities. In this paper, we propose a Neural Knowledge Language Model (NKLM) which **combines symbolic knowledge provided by knowledge graphs with RNN language models**. At each time step, the model predicts a fact on which the observed word is supposed to be based. Then, a word is either generated from the vocabulary or copied from the knowledge graph. We train and test the model on a new dataset, WikiFacts. In experiments, we show that the NKLM significantly improves the perplexity while generating a much smaller number of unknown words. In addition, we demonstrate that the sampled descriptions include named entities which were used to be the unknown words in RNN language models.

Citation: Ahn, S., Choi, H., Pärnamaa, T., & Bengio, Y. (2016). A Neural Knowledge Language Model. arXiv preprint arXiv:1608.00318.
PDF: https://arxiv.org/abs/1608.00318
Tutor: Dr. Asja Fischer

5. Rule Mining in Ontological Knowledge Bases

Abstract: Recent advances in information extraction have led to huge knowledge bases (KBs), which capture knowledge in a machine-readable format. **Inductive Logic Programming (ILP)** can be used to mine logical rules from these KBs, such as “If two persons are married, then they (usually) live in the same city”. While ILP is a mature field, mining logical rules from KBs is difficult, because KBs make an **open world assumption**. This means that absent information cannot be taken as counterexamples. Our approach AMIE [16] has shown how rules can be mined effectively from KBs even in the absence of counterexamples. In this paper, we show how this approach can be optimized to mine even larger KBs with more than 12M statements. Extensive experiments show how our new approach, **AMIE+**, extends to areas of mining that were previously beyond reach.

Citation: Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal, 24(6), 707-730.
PDF: https://people.mpi-inf.mpg.de/~chteflio/publications/amieplus.pdf
Tutor: Dr. Hajira Jabeen

6. YAGO2

Abstract: We present YAGO2, an extension of the YAGO knowledge base, in which entities, facts, and events are anchored in both **time and space**. YAGO2 is built automatically from **Wikipedia, GeoNames, and WordNet**. It contains 80 million facts about 9.8 million entities. Human evaluation confirmed an accuracy of 95% of the facts in YAGO2. In this paper, we present the extraction methodology, the integration of the **spatio-temporal dimension**, and our knowledge representation SPOTL, an extension of the original SPO-triple model to time and space.

Citation: Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 194, 28-61.
PDF: http://pubman.mpdl.mpg.de/pubman/item/escidoc:1323730/component/escidoc:1323729/MPI-I-2010-5-007.pdf
Tutor: Prof. Jens Lehmann

7. Reducing the Rank of Rel. Factorization Models

Abstract: Tensor factorization has become a popular method for learning from multi-relational data. In this context, the rank of the factorization is an important parameter that determines runtime as well as generalization ability. To identify conditions under which factorization is an efficient approach for learning from relational data, we derive **upper and lower bounds on the rank required to recover adjacency tensors.** Based on our findings, we propose a **novel additive tensor factorization** model to learn from latent and observable patterns on multi-relational data and present a **scalable algorithm** for computing the factorization. We show experimentally both that the proposed additive model does improve the predictive performance over pure latent variable methods and that it also reduces the required rank — and therefore runtime and memory complexity — significantly.

Citation: Maximilian Nickel, Xueyan Jiang, and Volker Tresp. 2014. Reducing the rank of relational factorization models by including observable patterns. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). MIT Press, Cambridge, MA, USA, 1179-1187.
PDF: http://papers.nips.cc/paper/5448-reducing-the-rank-in-relational-factorization-models-by-including-observable-patterns.pdf
Tutor: Dr. Asja Fischer

8. Learning Entity and Relation Embeddings

Abstract: Knowledge graph completion aims to perform **link prediction** between entities. In this paper, we consider the approach of knowledge graph embeddings. Recently, models such as TransE and TransH build entity and relation embeddings by regarding a relation as translation from head entity to tail entity. We note that these models simply put both entities and relations within the same semantic space. In fact, an entity may have multiple aspects and various relations may focus on different aspects of entities, which makes a common space insufficient for modeling. In this paper, we propose TransR to build **entity and relation embeddings in separate entity space and relation spaces**. Afterwards, we learn embeddings by first projecting entities from entity space to corresponding relation space and then building translations between projected entities. In experiments, we evaluate our models on three tasks including link prediction, triple classification and relational fact extraction. Experimental results show significant and consistent improvements compared to state-of-the-art baselines including TransE and TransH. The source code of this paper can be obtained from https://github. com/mrlyk423/relation_extraction.

Citation: Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In In Proceedings of AAAI?15, 2015 Ebenfalls Latent Distance Model
PDF: http://nlp.csai.tsinghua.edu.cn/~lzy/publications/aaai2015_transr.pdf
Tutor: Dr. Asja Fischer

9. Knowledge Base Completion

Abstract: Knowledge bases (KBs) are often greatly incomplete, necessitating a demand for KB completion. A promising approach is to embed KBs into latent spaces and make inferences by learning and operating on latent representations. Such embedding models, however, do not make use of any rules during inference and hence have limited accuracy. This paper proposes a novel approach which **incorporates rules seamlessly into embedding models for KB completion**. It formulates inference as an **integer linear programming (ILP)** problem, with the objective function generated from embedding models and the constraints translated from rules. Solving the ILP problem results in a number of facts which 1) are the most preferred by the embedding models, and 2) comply with all the rules. By incorporating rules, our approach can greatly reduce the solution space and significantly improve the inference accuracy of embedding models. We further provide a slacking technique to handle noise in KBs, by explicitly modeling the noise with slack variables. Experimental results on two publicly available data sets show that our approach significantly and consistently outperforms state-of-the-art embedding models in KB completion. Moreover, the slacking technique is effective in identifying erroneous facts and ambig

Citation: Quan Wang, Bin Wang, and Li Guo. Knowledge base completion using embeddings and rules. In Proceedings of the 24th International Joint Conference on Artificial Intelligence, pages 1859–1865, 2015.
PDF: http://ijcai.org/Proceedings/15/Papers/264.pdf
Tutor: Prof. Jens Lehmann

10. Latent Relation Representations for Schemas

Abstract: Traditional relation extraction predicts relations within some fixed and finite target schema. Machine learning approaches to this task require either manual annotation or, in the case of distant supervision, existing structured sources of the same schema. The need for existing datasets can be avoided by using a universal schema: the union of all involved schemas (surface form predicates as in OpenIE, and relations in the schemas of pre-existing databases). This schema has an almost unlimited set of relations (due to surface forms), and supports integration with existing structured data (through the relation types of existing databases). To populate a database of such schema we present a family of matrix factorization models that predict affinity between database tuples and relations. We show that this achieves substantially higher accuracy than the traditional classification approach. More importantly, by operating simultaneously on relations observed in text and in pre-existing structured DBs such as Freebase, we are able to reason about unstructured and structured data in mutually-supporting ways. By doing so our approach outperforms state-of-the-art distant supervision systems.

Citation: Riedel, S., Yao, L., McCallum, A. (2013) Latent Relation Representations for Universal Schemas. arXiv preprintarXiv:1301.4293.
PDF: https://arxiv.org/abs/1301.4293
Tutor: Dr. Asja Fischer

11. Graph-based Anomaly Detection

Abstract: Detecting **anomalies** in data is a vital task, with numerous high-impact applications in areas such as security, finance, health care, and law enforcement. While numerous techniques have been developed in past years for spotting outliers and anomalies in unstructured collections of multi-dimensional points, with graph data becoming ubiquitous, **techniques for structured graph data** have been of focus recently. As objects in graphs have long-range correlations, a suite of novel technology has been developed for anomaly detection in graph data.
This survey aims to provide a **general, comprehensive, and structured overview of the state-of-the-art methods for anomaly detection in data represented as graphs**. As a key contribution, we give a general framework for the algorithms categorized under various settings: unsupervised vs. (semi-)supervised approaches, for static vs. dynamic graphs, for attributed vs. plain graphs. We highlight the effectiveness, scalability, generality, and robustness aspects of the methods. What is more, we stress the importance of anomaly attribution and highlight the major techniques that facilitate digging out the **root cause**, or the 'why', of the detected anomalies for further analysis and sense-making. Finally, we present several real-world applications of graph-based anomaly detection in diverse domains, including financial, auction, computer traffic, and social networks. We conclude our survey with a discussion on open theoretical and practical challenges in the field.

Citation: Akoglu, L., Tong, H. & Koutra, D. Data Min Knowl Disc (2015) 29: 626. doi:10.1007/s10618-014-0365-y, arXiv preprint arXiv: 1404.4679.
PDF: https://arxiv.org/pdf/1404.4679.pdf
Tutor: Dr. Hajira Jabeen

12. The Anatomy of the Facebook Social Graph

Abstract: We study the structure of the social graph of active Facebook users, the largest social network ever analyzed. We **compute numerous features** of the graph including the **number of users and friendships, the degree distribution, path lengths, clustering, and mixing patterns**. Our results center around three main observations. First, we characterize the **global structure of the graph**, determining that the social network is nearly fully connected, with 99.91% of individuals belonging to a single large connected component, and we confirm the "six degrees of separation" phenomenon on a global scale. Second, by studying the average local clustering coefficient and degeneracy of graph neighborhoods, we show that while the Facebook graph as a whole is clearly sparse, the graph neighborhoods of users contain surprisingly dense structure. Third, we characterize the **assortativity patterns** present in the graph by studying the basic demographic and network properties of users. We observe clear degree assortativity and characterize the extent to which "your friends have more friends than you". Furthermore, we observe a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality, but we do not find any strong gender homophily. We compare our results with those from smaller social networks and find mostly, but not entirely, agreement on common structural network characteristics.

Citation: Ugander, J., Karrer, B., Backstrom, L., Marlow, C., (2011). The Anatomy of The Facebook Social Graph. arXiv preprint arXiv: 1111.4503.
PDF: https://arxiv.org/abs/1111.4503
Tutor: Dr. Hajira Jabeen

13. TransG: A Generative Embedding Model

Abstract: Recently, knowledge graph embedding, which projects symbolic entities and relations into continuous vector space, has become a new, hot topic in artificial intelligence. This paper proposes a novel **generative model** (TransG) to **address the issue of multiple relation semantics** that a relation may have multiple meanings revealed by the entity pairs associated with the corresponding triples. The new model can discover latent semantics for a relation and leverage a mixture of relation-specific component vectors to embed a fact triple. To the best of our knowledge, this is the first generative model for knowledge graph embedding, and at the first time, the issue of multiple relation semantics is formally discussed. Extensive experiments show that the proposed model achieves substantial improvements against the state-of-the-art baselines.

Citation: Huang, M., Hao, Y., Xiao, H., & Zhu, X.. (2015). TransG : A Generative Mixture Model for Knowledge Graph Embedding. CoRR, abs/1509.05488.
PDF: https://aclweb.org/anthology/P/P16/P16-1219.pdf
Tutor: Dr. Asja Fischer

14. Never-Ending Learning

Abstract: Whereas people learn many different types of knowledge from diverse experiences over many years, most current machine learning systems acquire just a single function or data model from just a single data set. We propose a never ending learning paradigm for machine learning, to better reflect the more ambitious and encompassing type of learning performed by humans. As a case study, we describe the Never-Ending Language Learner (NELL), which achieves some of the desired properties of a never-ending learner, and we discuss lessons learned. NELL has been learning to r**ead the web 24 hours/day since January 2010**, and so far has **acquired a knowledge base** with over 80 million confidence weighted beliefs (e.g., servedWith(tea, biscuits)). NELL has also learned millions of features and parameters that enable it to read these beliefs from the web. Additionally, it has learned to **reason over these beliefs to infer new beliefs**, and is able to extend its ontology by synthesizing new relational predicates. NELL can be tracked online at http://rtw.ml.cmu.edu, and followed on Twitter at @CMUNELL.

Citation: Never-Ending Learning
Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, J. Welling.
In Proceedings of the Conference on Artificial Intelligence (AAAI), 2015
PDF: http://www.cs.cmu.edu/~tom/pubs/NELL_aaai15.pdf
Tutor: Dr. Hajira Jabeen

OTE