MetaExp: Interactive Explanation and Exploration of Large Knowledge Graphs

Freya Behrens[1], Sebastian Bischoff[1], Pius Ladenburger[1], Julius Rückin[1], Laurenz Seidel[1], Fabian Stolp[1], Michael Vaichenker[1], Adrian Ziegler[1], Davide Mottin[2], Fatemeh Aghaei[2], Emmanuel Müller[2], Martin Preusse[3,4], Nikola Müller[3,4], Michael Hunger[5]
[1,2]Hasso-Plattner-Institute [3]Helmholtz Zentrum München, Institute of Computational Biology
[4]Knowing Health [5]neo4j Inc.
[1]first.last@student.hpi.de [2]first.last@hpi.de [3]first.last@helmholtz-muenchen.de
[5]first.last@neo4j.com

ABSTRACT

We present MetaExp, a system that assists the user during the exploration of large knowledge graphs, given two sets of initial nodes. At its core, MetaExp presents a small set of meta-paths to the user, which are sequences of relationships among node types. Such meta-paths do not overwhelm the user with complex structures, yet they preserve semantically-rich relationships in a graph. MetaExp engages the user in an interactive procedure, which involves simple meta-paths evaluations to infer a user-specific similarity measure. This similarity measure incorporates the domain knowledge and the preferences of the user, overcoming the fundamental limitations of previous methods based on local node neighborhoods or statically determined similarity scores. Our system provides a user-friendly interface for searching initial nodes and guides the user towards progressive refinements of the meta-paths. The system is demonstrated on three datasets, Freebase, a movie database, and a biological network.

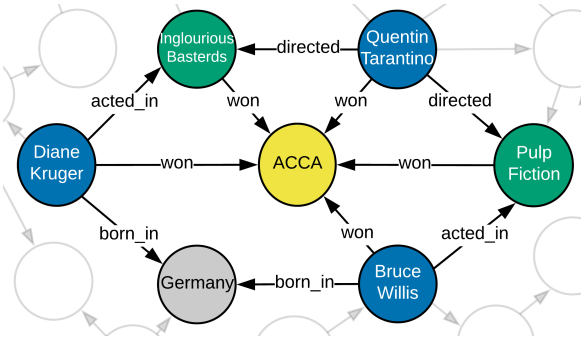
ACM Reference Format:
Freya Behrens, Sebastian Bischoff, Pius Ladenburger, Julius Rückin, Laurenz Seidel, Fabian Stolp, Michael Vaichenker, Adrian Ziegler, Davide Mottin, Fatemeh Aghaei, Emmanuel Müller, Martin Preusse, Nikola Müller, Michael Hunger. 2018. MetaExp: Interactive Explanation and Exploration of Large Knowledge Graphs. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3186978

1 INTRODUCTION

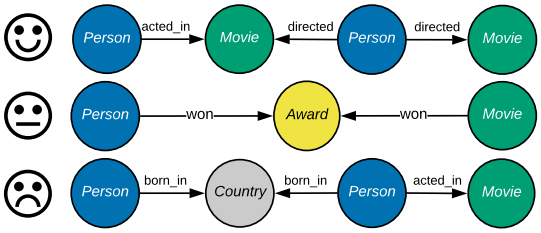
In the last few years, we have experienced a significant increment in the adoption of *knowledge graphs* to model complex phenomena in various domains. A knowledge graph is a heterogeneous network of entities, such as the actress *Diane Kruger* and the movie *Inglorious Basterds*, connected via named relationships, such as *acted_in*. These networks can conveniently represent various information like human knowledge, biological processes, and research work.

Nevertheless, expressiveness comes at the cost of complexity, since knowledge graphs usually have no predefined schema, and query languages, such as SPARQL, are hard to understand for novice users. Moreover, commonly available knowledge graphs like YAGO [5] and The Movie Database (https://www.themoviedb.org/), contain millions of nodes and edges. Hence, analyzing such graphs with common tools like Pegasus [3] and SNAP [1] is usually difficult for inexperienced users. As a remedy, exploration tools have been introduced to assist the user formulating queries with simple interfaces [7], or restricting the area of interest to some subset of the nodes [8]. The main shortcomings of such approaches are the inability to adapt to the *individual user preferences* [7] on the one hand, and the lack of a *global view* on the graph [8] on the other hand.

To circumvent the aforementioned shortcomings, we introduce MetaExp[1], a system to *interactively* explore large knowledge graphs in a user-centric manner. Our system bypasses convoluted queries



(a) Knowledge graph including facts from the movie domain



(b) Discovered and rated meta-paths resulting from the connection of the actress *Diane Kruger* to the movie *Pulp Fiction*

Figure 1: Discovered meta-paths and user feedback.

[1]Video: https://youtu.be/7aBxVPUpUpM, Code: https://hpi.de/en/mueller/metaexp

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186978

and guides the users towards the part of the graph they are interested in by inferring their interests through an interactive process. Initially, MetaExp asks the user to provide two sets of nodes in a knowledge graph; such sets delineate the semantic boundaries of the interesting parts of the graph to explore. The interesting parts of the graphs are explained through meta-paths [6, 10], which are sequences of relationships (e.g., *acted_in*) between different entity types (e.g., *Movie*). Such meta-paths are simple yet expressive descriptions of large portions of the graph that embody information about the network structure. In this way, MetaExp allows the user to progressively refine their preferences, selecting the meta-paths of interest that represent the user's domain reflected in the graph structure. The discovered meta-paths also provide a convenient mechanism to explain the relationships among entities. As opposed to system-centric graph summarization techniques [4], MetaExp is user-centric in that it preserves the information that is relevant for the user, and at the same time, retains the global view on the graph through meta-paths that can be easily executed as queries.

Motivating Example: Imagine a film critic, who is working on research concerning the German film panorama. They would like to learn more about the reason behind the international inclination of certain actors. For this purpose, the film critic defines a set of representative German actors and another set of movies not produced in Germany. As shown in Figure 1a they have at their disposal a large knowledge graph of people, professions, countries, etc. Our MetaExp addresses the knowledge graph complexity and the need for personalization, asking simple questions to the film critic, who can decide, for instance, to explicitly ignore the language of the movie, represented as an edge in the graph. Afterwards, MetaExp suggests the meta-paths in Figure 1b. Those meta-paths describe the type of relationships between German actors and foreign movies. Given their interests, the critic marks as irrelevant the meta-path that represents the fact that actors are born in the same country; conversely, they deem as important the one that represents international movies directed by an international director (e.g., *Quentin Tarantino* for *Pulp Fiction* and *Inglorious Basterds*). These preferences delineate the characteristics of the inferred user similarity, which will help to find international working German actors like *Marlene Dietrich*.

The MetaExp Showcase: With MetaExp we showcase an innovative exploration system to explain the similarity between two sets of nodes using meta-paths while taking explicit domain knowledge into account. While previous approaches embed a static user preference into the system [8] or ask the user to explore the node relationships manually [7], our approach is a middle ground between those two. In particular, MetaExp engages the user by initially asking them to provide two sets of nodes to start the exploration; then, by interactively refining the user preferences showing few meta-paths to be rated. In the end, we provide an aggregated similarity score and simple statistics, such as the number of instances of a meta-path, to explain the relationships among the initial node sets.

2 THE METAEXP SYSTEM

MetaExp is a web application, which can work on any device. The sequence of operations and computations is shown in Figure 2. The system computes and stores the entire set of meta-paths for the

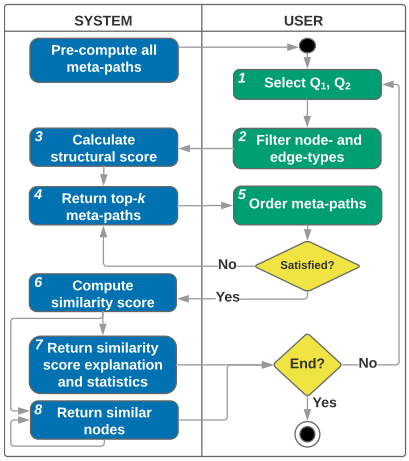


Figure 2: User interaction process with MetaExp.

graph in an offline manner. In the interactive online process, the user **(1)** first selects the initial sets of nodes $Q_1, Q_2 \subseteq V$ and then **(2)** filters node- and edge-types that are not relevant for the final result. After filtering, the system **(3)** computes a structural score and **(4)** returns the top-k meta-paths to be evaluated. Subsequently, the user **(5)** orders the meta-paths based on their preference and then decides whether to stop and let MetaExp **(6)** compute the final similarity score or ask the system to **(4)** show more meta-paths to be evaluated. Finally, we **(7)** return a set of exploratory statistics as well as **(8)** similar nodes based on the inferred similarity score.

Next, we describe the algorithms involved in the interactive exploration process of MetaExp. We recall that the system adaptively computes the similarity score and then returns the relevant meta-paths.

A **knowledge graph** is a directed graph $G : \langle V, E, \phi, \psi \rangle$, where V is a set of nodes, $E \subseteq V \times V$ is a set of edges, $\phi : V \mapsto L_V$ and $\psi : E \mapsto L_E$ are node and edge labeling functions, respectively. We refer to the elements of L_V and L_E as node-types and edge-types.

A **meta-path** [10] for a path $\langle n_1, ..., n_t \rangle, n_i \in V, 1 \leq i \leq t$ is a sequence $\mathcal{P} : \langle \phi(n_1), \psi(n_1, n_2), ..., \psi(n_{t-1}, n_t), \phi(n_t) \rangle$ that alternates node- and edge-types along the path. They are commonly used to compute similarity measures on knowledge graphs [9].

Our approach is based on the computation of meta-paths between user-defined input node sets Q_1, Q_2, and a rating ρ of meta-paths by the user. This rating and the structural score of each meta-path, i.e., the number of instances of a meta-path, are combined to provide a similarity measure $\sigma(Q_1, Q_2)$.

Meta-paths computation. The system precomputes all meta-paths between every possible pair of node-types in the later defined node sets Q_1 and Q_2. After the definition of the input node sets Q_1 and Q_2 by the user, the system searches for instances of the meta-paths between the input sets. This precomputation empowers the system to present the initial meta-paths instantly to the user for weighting them.

Selection of node- and edge-types. The user has the possibility to influence which meta-paths will be shown to them by choosing

which node- and edge-types should be excluded from the presented meta-paths. This reduces the initial amount of meta-paths that a user needs to explore.

Ordering of meta-paths. Throughout several iterations, the user can rank the meta-paths relative to each other without the need of assigning absolute values. For this, they place the meta-paths on a line so that the distance between each pair of meta-paths represents the difference in their importance regarding the domain. In addition to the meta-paths from the current iteration, the system displays a representative selection of previously ranked meta-paths which helps the user to arrange the meta-paths from the current iteration in the total ranking. This relative arrangement to the previously ranked meta-paths allows interpolating a score based on the distance of the placed meta-paths. This weighting produces the *domain score*. As a result, the user can incorporate their expert knowledge in the similarity measure and can get a better intuition of what the similarity measure means. Since the user is only allowed to choose nodes of one node-type per input node set Q, the number of meta-paths they have to rate reduces heavily.

Diversification of shown meta-paths. To present k meta-paths with low redundancy, MetaExp includes a notion of diversity. Given the set of all meta-paths *mp*, we compute the set of k meta-paths that cover the largest amount of the remaining meta-paths and has a low redundancy [2]. A meta-path $\mathcal{P}$ covers another meta-path $\mathcal{P}_1$ if $\mathcal{P}$ is contained in $\mathcal{P}_1$. Coverage is the number of meta-paths that are covered by at least one of the chosen k meta-paths. Diversity quantifies the number of unique node- or edge-labels present in the k meta-paths.

Structural score. The structural score is a domain-dependent property of each meta-path and is inherent to the graph, e.g., the count of instances of meta-paths. For instance, with regards to the frequency of meta-path instances, a high frequency might correspond to very important meta-paths in one domain, while in another domain infrequent meta-paths could be more important to the similarity measure. Consequently, two different structural scores are necessary to reflect these properties. For this reason, MetaExp uses structural scores that can be adapted via parameters.

Similarity score. At the end of the interaction, the system exploits the collected information to calculate the similarity score for the two node sets. To this end, the system uses a similarity function $\sigma(Q_1, Q_2)$ as proposed by Zhang et al. [11]:

$$\sigma(Q_1, Q_2) = \sum_{\mathcal{P} \in mp(Q_1, Q_2)} s(\mathcal{P}) \cdot \rho(\mathcal{P}),$$

where *mp* is the function that returns the meta-paths between the two input node sets Q_1 and Q_2, and s and ρ map a meta-path to its structural and domain score respectively.

Discovery of similar nodes. The system calculates the similarity between a node in the graph and a user-selected node or node set to find the nodes with the highest similarity. By iteratively discovering similar nodes, the user can explore parts of the graph relevant to their input node sets.

3 SYSTEM DEMONSTRATION

During the demonstration the attendee will evaluate MetaExp with two real datasets. Additionally, we provide an exemplified scenario on biological data to establish the applicability of our approach to a real setting.

The Movie Database: A graph comprising of relationships between 63k movies, casts, and producers. It consists of either *Person* or *Movie* nodes and edges with six different types, such as *acted_in* or *directed*.

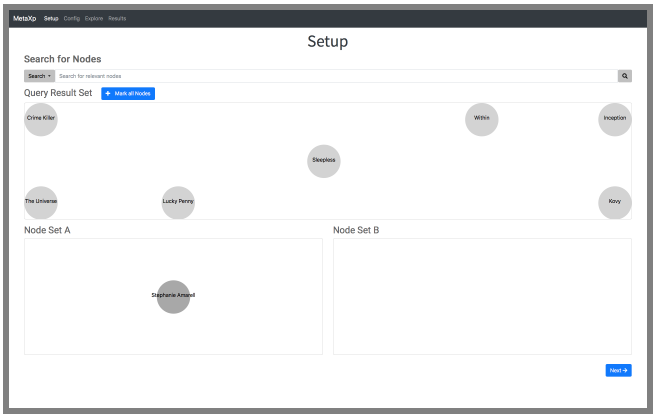
Freebase: A large knowledge graph of entities and relationships manually constructed with about 72M nodes and about 290M edges, 10k node types, and 4k edge types. Nodes represent entities (e.g., politicians), and edges relationships among such entities (e.g., partnership). For demonstration purpose, we will showcase MetaExp on specific Freebase domains, such as *sports*.

Biological dataset: This dataset models human biology and provides a real use-case of MetaExp. It contains relationships between phenotypes (e.g., diseases), genes and proteins in the human body. The datasets show interesting relationships between diseases and body processes and consists of 65M nodes and 91M edges. The dataset contains both public data the researchers collected from genome databases and proprietary research data.

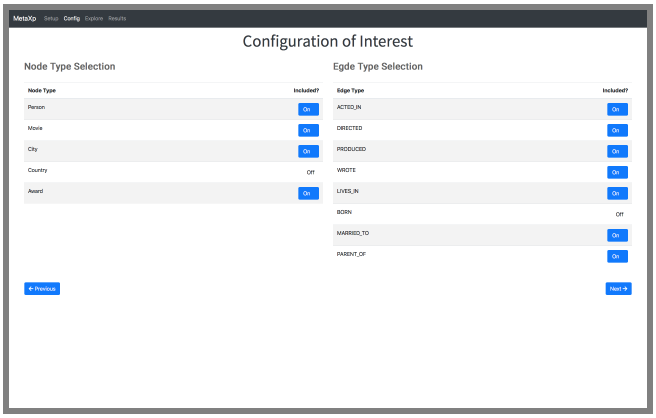
3.1 Use Case

Consider the scenario in which a marketing expert wants to start a campaign based on movies coming soon. Using our system the expert wants to retrieve other movies which can be advertised with the newcomers. The MetaExp demonstration scenario is the following.

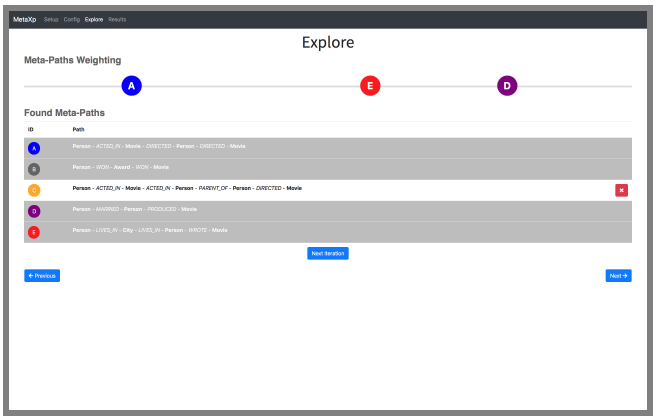
(1) The user starts with a set of recent movies (Figure 3a). Using the search bar, they can select movies by id to place in the first set and newcomers in the second set. Nodes can be searched by id, property or via a cypher graph query.
(2) Next (Figure 3b), the user chooses which are the relevant nodes and edges types from the list of all types.
(3) At this point (Figure 3c), the user provides their domain knowledge by means of scores on meta-paths. Meta-paths are shown in a batch of five, and the user can decide whether to rate such meta-paths or discard them. For example, the user might find that movies written by the same screenwriter are more similar than those with the same actors. Therefore, the user would place those meta-paths containing more *write* relationships further on the right of the line than those containing the *acted_in* relationship. On the other hand, long chains of *acted_in* relationships that not convey any relevant information about the similarity of the movies can be discarded. Note that the meta-paths shown to the user only contain node and edge types which have not been deactivated by the user in the previous step.
After the user has finished arranging the first five meta-paths, the user can click on *More Meta-Paths* to retrieve another batch of meta-paths to be rated. This process is repeated until the user feels that newly shown meta-paths do not convey further information. As a reference, the first, second (median) and third quantile of previously ordered meta-paths are always shown on the line.



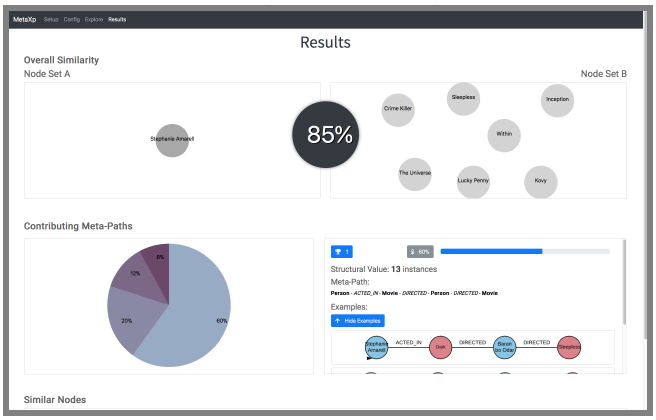
(a) Initial setup of node sets



(b) Filtering of node- and edge-types



(c) Exploration and rating of meta-paths



(d) Final similarity score, contributing meta-paths and similar nodes

Figure 3: The MetaExp user interface

(4) Finally, MetaExp presents results based on the structural and domain scores introduced in the previous sections (Figure 3d). A similarity score between the two input movie sets is shown and explained through a pie chart illustrating the contribution of different meta-paths. Furthermore, by clicking on a slice of the chart, it is possible to view detailed information about each of these meta-paths, such as instances of the meta-paths in the graph.
Additionally, the user can select one or multiple nodes in the input set and receive suggestions of similar nodes based on the similarity score. Moreover, the properties, the similarity to the selected nodes and the one-neighborhood for each of the suggested nodes are shown. The proposed nodes can be added to the input set, therefore refine the query, and start the process anew but maintaining the preferences computed so far.

4 ACKNOWLEDGEMENTS

The project was partially funded by the GeoForschungsZentrum (GFZ) in Potsdam and neo4j. We would also like to thank Knowing and the Helmholtz Association Munich for providing the biological dataset used in the demonstration.

REFERENCES

[1] David A Bader and Kamesh Madduri. 2008. Snap, small-world network analysis and partitioning: An open-source parallel graph framework for the exploration of large-scale networks. In *IPDPS*. 1–12.
[2] Allan Borodin, Aadhar Jain, Hyun Chul Lee, and Yuli Ye. 2017. Max-Sum Diversification, Monotone Submodular Functions, and Dynamic Updates. *TALG* 13, 3 (2017), 41.
[3] U Kang, Charalampos E Tsourakakis, and Christos Faloutsos. 2009. Pegasus: A peta-scale graph mining system implementation and observations. In *ICDM*. 229–238.
[4] Yike Liu, Abhilash Dighe, Tara Safavi, and Danai Koutra. 2016. A Graph Summarization: A Survey. *arXiv preprint arXiv:1612.04883* (2016).
[5] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. 2014. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*.
[6] Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, and Wangda Zhang. 2015. Discovering meta-paths in large heterogeneous information networks. In *WWW*. 754–764.
[7] Robert Pienta, Fred Hohman, Acar Tamersoy, Alex Endert, Shamkant Navathe, Hanghang Tong, and Duen Horng Chau. 2017. Visual Graph Query Construction and Refinement. In *SIGMOD*. 1587–1590.
[8] Stephan Seufert, Patrick Ernst, Srikanta J Bedathur, Sarath Kumar Kondreddi, Klaus Berberich, and Gerhard Weikum. 2016. Instant Espresso: Interactive Analysis of Relationships in Knowledge Graphs. In *WWW Companion*. 251–254.
[9] Chuan Shi, Xiangnan Kong, Yue Huang, S Yu Philip, and Bin Wu. 2014. Hetesim: A general framework for relevance measure in heterogeneous networks. *TKDE* 26, 10 (2014), 2479–2492.
[10] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB* 4, 11 (2011), 992–1003.
[11] Xiangling Zhang, Yueguo Chen, Jun Chen, Xiaoyong Du, Ke Wang, and Ji-Rong Wen. 2017. Entity Set Expansion via Knowledge Graphs. In *SIGIR*. 1101–1104.