Demonstrating Narrative Pattern Discovery from Biomedical Literature

Hermann Kroll[1][0000−0001−9887−9276], Pascal Sackhoff[1][0009−0005−3095−9794], Bill Matthias Thang[1][0009−0006−8321−8479], Christin Katharina Kreutz[2,3][0000−0002−5075−7699], and Wolf-Tilo Balke[1][0000−0002−5443−1215]

[1] TU Braunschweig, Braunschweig, Germany
krollh@acm.org, balke@ifis.cs.tu-bs.de
[2] TH Mittelhessen - University of Applied Sciences, Gießen, Germany
[3] Herder Institute, Marburg, Germany
ckreutz@acm.org

Abstract. Digital libraries maintain extensive collections of knowledge and need to provide effective access paths for their users. For instance, PubPharm, the specialized information service for Pharmacy in Germany, provides and develops access paths to their underlying biomedical document collection. In brief, PubPharm supports traditional keyword-based search, search for chemical structures, as well as novel graph-based discovery workflows, e.g., listing or searching for interactions between different pharmaceutical entities. This paper introduces a new search functionality, called narrative pattern mining, allowing users to explore context-relevant entities and entity interactions. We performed interviews with five domain experts to verify the usefulness of our prototype.

Keywords: Digital Libraries · Narrative Information Access · Graph-based Discovery · Pattern Mining · Entity Search

1 Introduction

Digital libraries maintain extensive collections of knowledge and need to provide effective access paths for their users. Ideally, two types of searches should be supported: precise search for relevant material and exploratory search [7]. In this paper, we focus on exploratory search to allow users to find new and interesting ideas for their own work. For instance, the connected papers service[4] allows users to explore the connection between different research articles, e.g., who cites whom or what is adjacent to a certain paper. This way users may find new and interesting articles for their own work.

We, as the specialized service for Pharmacy in Germany, build upon that idea. The biomedical/pharmaceutical domain is an entity-centric one, e.g., research focuses around certain drugs, diseases, targets, methods and more; see for instance a PubMed query log analysis [1] or entity-centric services like PubTator [12]. That is why we developed a new entity-centric search functionality

[4] https://connectedpapers.com

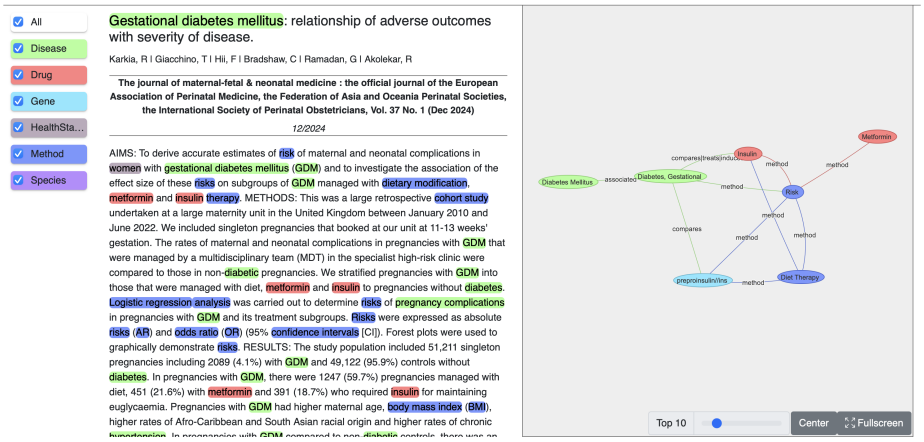

Fig. 1. Document Visualization: The left side shows the document text. Detected entities are highlighted in corresponding colors. A UI selection box on the left side allows to show or hide certain entity types. The right side depicts extracted interactions between entities as a labeled and colored graph.

for our platform, which we are describing in the following. The main idea is that users start their search with a set of relevant entities for their own work. Then, the service first retrieves documents that include these entities and second, mines patterns between the given and other context-relevant entities. All information is then visualized as a network so that users can explore context-relevant entities and entity interactions, so called *narrative patterns*. A click on a network's edge forwards users to corresponding literature supporting the selected entity-entity interaction. While our Narrative Discovery System has been published [7], this paper introduces our *narrative pattern*-driven discovery method and prototye.

2 Narrative Discovery System

PubPharm[5], the specialized information service for Pharmacy in Germany, aims to provide effective and innovative access paths to the pharmaceutical literature for our research community. In the past, we proposed and implemented a discovery system for narrative information access [7]. The system called the Narrative Service[6] allows users to formulate their information needs as graph patterns, i.e., interaction patterns between entities. This way, users may search for literature stating that *Metformin is used to treat diabetes mellitus in adult patients*. In addition, variables can be used to explore the literature, e.g., *any drugs* used to treat diabetes mellitus in adult patients. The service is capable of answering these queries through graph pattern matching. The Drug Overviews[7]

[5] https://www.pubpharm.de
[6] https://narrative.pubpharm.de
[7] https://narrative.pubpharm.de/drug_overview

service extends our discovery system by allowing users to generate overviews about specific drugs, i.e., known interactions like therapies, target interactions, administrations, and more. When users click on some information, they are directed to a corresponding search in the Narrative Service, e.g., searching for literature about a certain drug-disease therapy.

To enable graph pattern matching, the relevant biomedical literature is transformed into document graphs, i.e., relevant biomedical entities are detected, and their interactions are extracted from texts. Instead of building a single knowledge graph, we decided to transform each text into a small document graph and keep these graphs separated. Reasons for this decision were that 1) user queries are still answered by retrieving relevant literature instead of short answers and 2) the validity of statements is ensured [3], i.e., the system retrieves the corresponding context in which a statement is valid. A visualization of an enriched document in our system is shown in Figure 1 or online[8]. Entities were detected by performing a dictionary-based entity linking against existing vocabularies and by using existing biomedical annotation tools like GNormPlus [13], TaggerOne [10], and PubTator Central [12]. Statements were extracted by deploying PathIE, a method for extracting statements via the grammatical structure of sentences, and by extracting association statements if two entities co-occur within the same sentences. These methods are part of our extraction toolbox [6], which we described [2] and analyzed [8] comprehensively. In brief, most of our self-developed methods do not rely on supervision and thus bypass the need for training data, but they only come with a moderate extraction quality. The code for our discovery platform and the extension for this paper is available at GitHub[9] and SoftwareHeritage[10]. Our system (as of May 2025) includes 38M PubMed/Medline articles.

The Narrative Service itself provides precise literature searches due to graph-based queries and, thus, entity-interaction-aware searches [7]. A keyword-based search functionality assists users in formulating graph queries [4]. Exploratory searches are, as of now, supported by using variables in queries or using the Drug Overview functionality. This paper contributes a new access path for our service that allows users to discover entity interactions in contexts.

3 Mining Narrative Patterns

Entity interactions play a central role within the biomedical literature. The goal of our narrative pattern mining here is to allow an exploration of the literature, i.e., visualizing and thus summarizing what is known and often described between a set of searched entities. With that, the system can shed light on context-relevant entities and interactions between them, so that users can explore the entities' neighborhood. This kind of exploration should assists users

[8] https://narrative.pubpharm.de/document/?document_id=38844413&data_source=PubMed

[9] https://github.com/HermannKroll/NarrativeIntelligence/

[10] SoftwareHeritage ID: swh:1:dir:5b87566505d9f3ad0837cc91f105ee163515ec3d

in understanding the relationships between biomedical entities and possibly discover new relevant entities to the users' information needs.

System Architecture. In brief, our service expects a list of searched entities as its input. The output is a graph pattern that 1) puts the searched entities in relation and 2) adds more entities that play a central role in the searched entities' contexts. Therefore, we first identify relevant documents, retrieve the document graphs, score the graphs' edges, and sort these edges by their final score. Users can then select how many edges shall be displayed.

Identifying Relevant Documents. Our goal is to support users in exploring the literature by showing what is written about the set of searched entities. We, therefore, decided to extract these patterns from documents that include all of the searched entities. This way, only information appearing in a context (a document) that includes all searched entities is considered.

Users enter a list of strings. Each string represents a search for entities in our system. Each of these strings has to be translated into entities from our vocabulary. Therefore, we use the following translation paradigm: Suppose a user types the string *diabetes melli* in the search but does not complete their string insertion yet. All entities that include both the terms *diabetes* and *melli* in one of their synonyms are valid translations, e.g., *diabetes mellitus*, *diabetes mellitus type 1*, *diabetes mellitus type 2*, and many more. These valid translations are then suggested to users during the input as keywords. A user can pick one of the suggestions, e.g., *diabetes mellitus* as keyword[11] or finish their typing and lock in their inserted string as a keyword. With that, we can translate the user's input keyword into a set of entities. Next, we use an inverted index to retrieve document IDs in which a particular entity has been detected. This gives us the set of documents relevant for a specific keyword.

If a user enters multiple keywords, the translation is conducted for all keywords. Then we compute the intersection between those sets of documents relevant for *single* keywords to identify documents fitting *all* entered keywords.

Scoring Edges. Next, we retrieve the document graphs for the retrieved document IDs fitting all keywords and mine the actual narrative pattern. Combining all graphs and showing the resulting one to the user would likely to be overstraining them. For instance, when searching for entities like *diabetes mellitus* and *metformin*, thousands of documents are retrieved. The resulting graph would then also include hundreds of different statements. We tackled this problem in two ways: 1) We score each graph edge so that only the most important ones are shown to the users. Users can control how many edges should be visualized at once. 2) We only show edges that are incoming or outgoing from one of the user's searched entities so that these entities are put into the focus of the generated narrative pattern. We score graph edges as follows, which proved to be

[11] Note that selecting *diabetes mellitus* as keyword will still also translate the keyword to *diabetes mellitus type 1* etc.

Algorithm 1 Mining Narrative Patterns from Document Graphs

```
Input: list(set(entities translated from each keyword)) Output: a ranked list of documents
relDocPerKeyword = list()
for entitiesForKeyword ∈ listOfEntitiesPerKeyword do
    relDocPerKeyword.append(getRelevantDocs(entitiesForKeyword))
intersectRelDocs = set.intersection(*relDocPerKeyword)
edge2score = dict()
docs = retrieveDocumentData(intersectRelDocs)
for d ∈ docs do
    for e ∈ d.edges do
        if e.subject ∈ entities ∨ e.object ∈ entities then
            edge2score[s] += score(e, d)
edge2score = sortByScoreDescending(edge2score)
return edge2score
```

effective for ranking and recommending documents in our discovery system [5,9]: $score(e, d) = tf\text{-}idf(e, d) * coverage(e, d) * confidence(e, d)$

The scoring function takes a graph edge e and its corresponding document d as its input and returns a numeric score. In brief, tf-idf stands for term-frequency inverse-document-frequency and prefers edges that appear often within d but rarely within the whole collection. Coverage favors edges that include entities used across the document (and not just at the beginning or end of some text). Confidence boosts edges with high extraction confidence, i.e., the extraction method extracted the edge with high confidence. For more details we refer the reader to our prior works [5,9] or our actual implementation.

The score function allows us to score each document graph edge. For the narrative pattern, we compute the union of all retrieved graphs. Then, we score each of its edges as follows. Let D be the set of retrieved documents relevant for all keywords and e be an edge of the narrative pattern, we sum the scores for this edge in every document from D for our final score: $fscore(e) = \sum_{d \in D} score(e, d)$. $score$ returns 0 if the edge e is not present in document d.

Our final Algorithm is shown in 1. First, it retrieves relevant documents and then sums up the scores for each edge in every retrieved document. The edges are then sorted in descending order with regard to their scores.

4 Demonstration

Our new narrative pattern discovery component has been integrated into our main discovery system[12], and a tutorial video is available at[13]. In the following, we first describe the user interface and explain our design decisions. We then describe a preliminary user evaluation that we intend to extend in the future.

[12] https://beta.narrative.pubpharm.de, tab *Pattern Discovery (Beta)*
[13] https://pharmrxiv.de/receive/pharmrxiv_mods_00026752

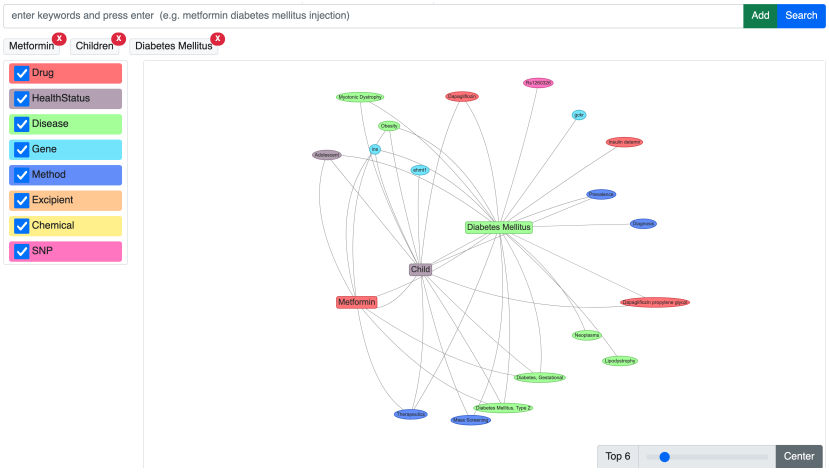

Fig. 2. Pattern Visualization: The extracted pattern is shown in the center of the screen. Nodes are colored depending on their entity type. The searched entities are depicted as rectangle nodes with a larger font. Users may hide certain entity types or select how many edges are visualized at once.

4.1 User Interface

Users can enter a list of searched entities by typing into a search bar. They are assisted with an autocompletion functionality that proposes known entity terms. They can then add entity terms to their search by pressing `enter` or clicking the `add` button on the right. Entities are then added to a list below the search bar. They can be removed by clicking on a red cancel sign. Next, users may start the search by pressing `enter` or clicking the `search` button.

The system replies with a color-coded graph representation. The searched entities are highlighted in the center of the visualization as rectangle nodes with a larger font. Every other node is a rounded oval. Nodes' colors depend on their entity types, e.g., red for drugs. A unified entity-type coloring is used across the whole discovery system. We decided to keep the representation simple, so we removed edge labels. Suppose users want more details on a certain interaction (edge) between two entities. In that case, they can click on an edge and are forwarded to a corresponding search in our discovery system, i.e., literature is shown that supports the clicked interaction between two entities. Entity types can be hidden or made visible via clicking the colored boxes on the left side of the screen. This way, users can narrow down the pattern to show only certain entity types, such as drugs, diseases, and targets. Our document graph representation has already established a similar feature; see Figure 1. At the bottom, users may select how many edges should be visualized simultaneously. By default, this is set to five edges per concept. If users scroll further down, they see a list of documents that contain the searched entities. This feature delivers Provenance, which allows users to screen the literature used to generate the pattern.

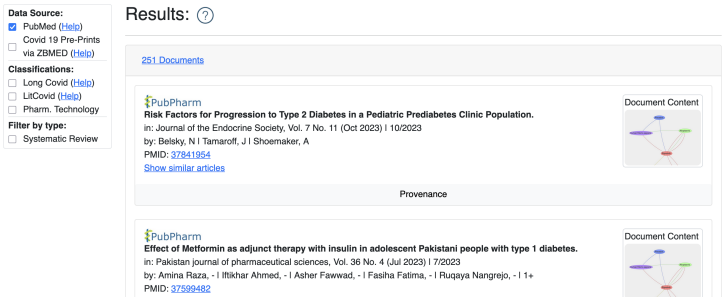

Fig. 3. Result Lists: The system shows the document lists used to generate the narrative pattern. Filter options are available to further narrow down searches.

In addition, a source selection filter is shown on the left side. Here, users can select which data sources are used for the pattern mining step. They can narrow down their searches to specific collections or certain document classes, e.g., articles relevant to Pharmaceutical Technology or published within a specified time span. Screenshots of our user interface are shown in Figure 2 and Figure 3.

4.2 User Evaluation

Setup. For our small-scale user evaluation, we conducted one on one interviews taking around thirty minutes with five participants in German. Participants were familiar with PubPharm and the Narrative Service as they already took part in earlier studies. Participants were able to access the tool via Zoom and remote desktop control. Two experienced interviewers lead the participants through the three-part study: **1. Introduction.** (5 minutes) Participants were informed on the purpose and terms of the study before they consented to take part in the study. Afterwards, the interviewer gave a brief overview of the tool. One of the interviewers took notes while the other interviewer lead the participant through the study. **2. Usage of the system with thinking aloud.** (15 minutes) Participants used the whole system freely while thinking aloud [11] to work on one of their research questions. **3. Semi-structured interview.** (10 minutes) We followed an earlier evaluation on parts of the system [4] for the semi-structured interview and asked participants the same questions on general thoughts regarding the tool, encountered problems, liked components, changes required for them to consider using the tool and if they had any other comments.

Results. Participants **encountered problems** related to the *search bar*: they did not immediately understand how keywords were supposed to be entered with one participant noting *you need to first understand what it wants then you are able to do it*. They had trouble with selecting keywords the system had in the vocabulary. The two buttons `search` and `add` were not descriptive enough. When *clicking edges* in the displayed graph, participants were surprised that the triggered search would lose the context of the graph and only look for the

entities associated with the edge (and not all entities that were searched in the first place). The click opened the results in a new tab leading to some difficulty to navigate back to the original graph. Participants identified some *interface problems* which are easy to correct from a technical point of view: cryptic error messages, small-sized symbols, similar colors used in the graph and uncertainty which fields can be clicked. We made some *additional observations*: in general it seemed to have been unclear what the additional values was, that the new tool brought. Participants would have required more time. They often did not scroll to see the documents resulting the search but only checked out the graph, sometimes they did not interact with the graph, they did not use the `top x` functionality or the possibility to restrict the depicted concept types.

As for **liked components** participants mentioned the fitting results for precise queries, the tool's power to provide a good overview of a topic, its intuitive and logical handling as well as the possibility to evaluate the results while searching which makes the tool stand out against using LLMs. The *interface* was praised: its visualization of relations and keywords, the possible interactions as well as opening clicked edges in a new tab. For participants, the tool and especially the interaction with the graph was behaving *according to their expectations*, clicking on an edge focuses on this specific part of the graph and shows literature backing the edge. We *observed* users clicking edges, deleting keywords and modifying or restricting their queries, using the `top x` functionality, restrict the depicted concept types in the graph, and recovering from involuntary clicks.

Participants mentioned some **required changes** or potential future features: Some participants were unsure how to express some technical terms in English which hints towards a language-independent query option. In terms of *search* there are many ideas: exclusion of edges, inclusion of terms which are not part of the tool's vocabulary, the option to query with variables, a `clear all` button for queries, an improved sorting for suggested keywords, typo resistance, and enabling the search between three interconnected components. Participants wished to *export or save* their searches, networks and interesting works. In terms of *visualization* they wished for seeing the pattern search and result list in one view. One participant asked for the inclusion of cross-references. *Other* wishes were video tutorials, better explanations and an extension of the data source to also incorporate news articles.

5 Conclusion

In brief, this paper introduces a novel access path for PubPharm's Narrative Discovery System. Our narrative pattern mining approach assists users with entity-driven exploratory search. This way users can explore the neighborhood of searched entities. The conducted interviews verified the usefulness of our system and the entity-driven and network-based visualization.

Future work will tackle described problems and improve the search. We intend to provide users with the option to enter keywords similar to a Google search before employing, e.g., an LLM to suggested potential interesting patterns.

References

1. Herskovic, J.R., Tanaka, L.Y., Hersh, W., Bernstam, E.V.: A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. Journal of the American Medical Informatics Association **14**(2), 212–220 (03 2007). https://doi.org/10.1197/jamia.M2191
2. Kroll, H.: Narrative Information Access - A new Paradigm for Digital Libraries (Narrativer Informationszugriff - Ein neues Paradigma für Digitale Bibliotheken). Ph.D. thesis, TU Braunschweig, Germany (2023). https://doi.org/10.24355/DBBS.084-202401171145-1
3. Kroll, H., Kalo, J., Nagel, D., Mennicke, S., Balke, W.: Context-compatible information fusion for scientific knowledge graphs. In: Digital Libraries for Open Knowledge - 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12246, pp. 33–47. Springer (2020). https://doi.org/10.1007/978-3-030-54956-5_3, https://doi.org/10.1007/978-3-030-54956-5_3
4. Kroll, H., Kreutz, C.K., Sackhoff, P., Balke, W.: Enriching simple keyword queries for domain-aware narrative retrieval. In: ACM/IEEE Joint Conference on Digital Libraries, JCDL 2023, Santa Fe, NM, USA, June 26-30, 2023. pp. 143–154. IEEE (2023). https://doi.org/10.1109/JCDL57899.2023.00029
5. Kroll, H., Kreutz, C.K., Thang, B.M., Schaer, P., Balke, W.: Building an explainable graph-based biomedical paper recommendation system (technical report). CoRR **abs/2412.15229** (2024). https://doi.org/10.48550/ARXIV.2412.15229
6. Kroll, H., Pirklbauer, J., Balke, W.: A toolbox for the nearly-unsupervised construction of digital library knowledge graphs. In: ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, Champaign, IL, USA, September 27-30, 2021. pp. 21–30. IEEE (2021). https://doi.org/10.1109/JCDL52503.2021.00014
7. Kroll, H., Pirklbauer, J., Kalo, J., Kunz, M., Ruthmann, J., Balke, W.: A discovery system for narrative query graphs: entity-interaction-aware document retrieval. Int. J. Digit. Libr. **25**(1), 3–24 (2024). https://doi.org/10.1007/S00799-023-00356-3
8. Kroll, H., Pirklbauer, J., Plötzky, F., Balke, W.: A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries. Int. J. Digit. Libr. **25**(2), 401–425 (2024). https://doi.org/10.1007/S00799-023-00368-Z
9. Kroll, H., Sackhoff, P., Breuer, T., Schenkel, R., Balke, W.: Ranking narrative query graphs for biomedical document retrieval (technical report). CoRR **abs/2412.15232** (2024). https://doi.org/10.48550/ARXIV.2412.15232
10. Leaman, R., Lu, Z.: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics **32**(18), 2839–2846 (06 2016). https://doi.org/10.1093/bioinformatics/btw343
11. Lewis, C.: Using the "thinking Aloud" Method in Cognitive Interface Design. https://books.google.de/books?id=F5AKHQAACAAJ
12. Wei, C.H., Allot, A., Leaman, R., Lu, Z.: PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Research **47**(W1), W587–W593 (05 2019). https://doi.org/10.1093/nar/gkz389
13. Wei, C.H., Kao, H.Y., Lu, Z.: GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. BioMed Research International **2015**, 918710 (Aug 2015). https://doi.org/10.1155/2015/918710