

Building an Explainable Graph-based Biomedical Paper Recommendation System (Technical Report)

Hermann Kroll
krollh@acm.org
Institute for Information Systems,
TU Braunschweig
Germany

Christin K. Kreutz
ckreutz@acm.org
TH Mittelhessen & Herder Institute
Germany

Bill Matthias Thang
m.thang@tu-bs.de
Institute for Information Systems,
TU Braunschweig
Germany

Philipp Schaer
philipp.schaer@th-koeln.de
TH Köln (University of Applied
Sciences)
Germany

Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de
Institute for Information Systems,
TU Braunschweig
Germany

Abstract

Digital libraries provide different access paths, allowing users to explore their collections. For instance, paper recommendation suggests literature similar to some selected paper. Their implementation is often cost-intensive, especially if neural methods are applied. Additionally, it is hard for users to understand or guess why a recommendation should be relevant for them. That is why we tackled the problem from a different perspective. We propose XGPREc, a graph-based and thus explainable method which we integrate into our existing graph-based biomedical discovery system. Moreover, we show that XGPREc (1) can, in terms of computational costs, manage a real digital library collection with 37M documents from the biomedical domain, (2) performs well on established test collections and concept-centric information needs, and (3) generates explanations that proved to be beneficial in a preliminary user study. We share our code so that user libraries can build upon XGPREc.

CCS Concepts

• **Information systems** → **Information retrieval**; *Document representation*; Web interfaces.

Keywords

Explainable Paper Recommendation, Biomedical Document Retrieval, System Design, User Interface, Digital Libraries

1 Introduction

Digital libraries provide effective access paths for users to explore their underlying collections. However, the number of scientific publications in such collections is increasing rapidly. This vast amount of options to look at combined with possibly under-specified information needs of users can lead to challenges when trying to find related work for a topic of interest as keyword-based options are insufficient [25]. As a remedy, paper recommendation systems suggest related literature based on some user's selected article(s) [12, 49]. The advantage of paper recommendation is that users get an impression of what other works are highly related to some given article, allowing them to explore a library's collection in this way. We consider the following definition for the paper recommendation

task: *Given an initial article, what are other relevant articles with regard to the input article's topic and content?*

Current paper recommendation systems do not focus on limiting computational complexity or performing the task with fewer resources, even though the emissions produced by neural retrieval methods are several orders of magnitude higher than those of BM25 [44]. Moreover, even worse, current approaches often rely on deep learning [25], i.e., a cost-intensive training or embedding step is necessary. Collecting training data on a large scale while also considering a variety of different information needs and topics is an exceptionally challenging task. If training data is *just* collected for a special subgroup of users, learned system will likely have a bias in the end when answering unseen information needs. In practice, the computational costs of neural retrieval methods and the collection of training data usually hinder their implementation in a digital library. Another understudied issue but desirable goal in paper recommendation is explainability [25]. Explanations are helpful to (1) justify individual recommendations, (2) understand how a system works, and (3) distinguish good from bad recommendations via users' feedback [6]. In contrast to deep learning approaches that may struggle to explain their retrieved documents to users, we choose to head for an alternative: concept-centric graph-based document representations.

We tackled the two issues of high costs (in terms of retrieval and training data collection) and missing explainability by building upon graph-based document representations in cooperation with PubPharm¹, the specialized service for Pharmacy (some of this paper's authors are part of the PubPharm team). In addition to keyword-based retrieval, PubPharm offers a graph-based discovery system called the Narrative Service² as a more sophisticated access path for users. The Narrative Service assumes that complex, concept-centric information needs can be represented as narrative query graphs, i.e., as short stories of interest involving relevant biomedical concepts and their interactions. The service comes with two central advantages [28]: First, narrative query graphs allow users to search for literature precisely by stating concept interactions explicitly. Second, users can explore a digital library's collection using variables, e.g., structuring the literature by possible

¹www.pubpharm.de

²<https://narrative.pubpharm.de>

treatment options in adults with diabetes. While the costs for harvesting graphs from texts can be quite high, we proposed effective and cost-moderate extraction workflows for digital libraries that we applied to build our and PubPharm’s system [29, 30].

We expand our and PubPharm’s existing Narrative Service and provide more functionality to our users. We design and implement an explainable paper recommendation system (a demo video is available here³), called **XGPREc**. We will integrate XGPREc into to our graph-based biomedical discovery system [28]. By consciously refraining from using computationally intense neural retrieval methods in our recommendation and instead relying on graph-based similarities and text-scoring through BM25, we bypass vast computational costs [44]. Graph representations can visualize complex interconnections and enable users from the biomedical domain to immediately grasp if a paper fits what they mean to find [26]. In this work, we focus on justifying recommendations [6], which we achieve by displaying overlapping graph patterns between an initial and a candidate paper to explain recommendations for users.

At the showcase of a real digital library, we demonstrate how graph-based recommendation can be implemented for the extensive MEDLINE document collection, which contains about 37 million documents. Our research objective for this work is thus: *How can we design a fast and reliable paper recommendation system?* More precisely: (1) *How can we find relevant candidate documents with as low costs as possible (first stage)?* (2) *How can we compute relevant and explainable recommendations (second stage)?*

Our code is available at GitHub⁴ and Software Heritage⁵.

2 Related Work

2.1 Paper Recommendation

Several surveys describe approaches and problems of the general paper recommendation domain [4, 7, 25, 46, 55]. Our approach operates on the *biomedical domain*, incorporates a *knowledge graph* for every paper and is *explainable*.

Biomedical Domain. EILEEN [1] operates on a dataset from MEDLINE and lets users indicate binary relevancy of search results. Tf-idf, SVD, Elasticsearch and key phrase extraction with RAKE are combined. ChemVis [8] uses PubChem and depicts chemical compositions of materials while using HyPlag [39] for general document similarity and RecVis [9] for recommendation. Guo et al. [18] use a linear combination of bibliographic coupling and textual similarity either computed with PMRA [34], BM25 or cosine similarity of tf-idf vectors. They evaluate their approach on the TREC Genomics dataset. Emati [24] works on articles which appeared in PubMed or ArXiv since 2000. Users can (dis)like papers or upload a list of articles they have cited as positive examples to train classifiers on relevant literature. The system uses BERT or tf-idf vectors of articles and their metadata and multinomial Naive Bayes to classify relevant and irrelevant papers. KnowCOVID-19 [41] uses different topic models on CORD-19 to identify topics described, genes explored and drugs used in papers. They present an analysis of papers according to the Levels of Evidence Pyramid based on paper tagging by users.

Two approaches are available online: PubMed displays similar articles to papers on their abstract page⁶. Word-weighting on stemmed titles, abstracts and MeSH terms as well as document length are considered when computing the neighboring recommendation candidates for an initial paper. The most similar documents for each paper are pre-computed and displayed to users to ensure fast response times. LitSuggest [3] classifies unseen papers based on users’ ratings of relevancy of papers per topic on PubMed.

Therefore, from the biomedical domain, we could compare our approach against LitSuggest or PubMed. LitSuggest requires positive and negative training data provided by an user in order to recommend papers, therefore we disregard the system in our analysis to focus on easiness of use as given by PubMed.

Knowledge Graphs. KERS [2] is a feedback-based system recommending articles from categories, a user has interacted with before. The categories stem from an expert-built KB. CV-Lattes [14] uses users’ published papers to build terms and concepts profiles (stored in a KB), weights of similarity vectors between them stem from a KB. Li et al. [33] compose a KG-based embedding of paper metadata and learn a mapping between users’ browsed and clicked recommended articles. Manrique and Marino [37] compare different edge- and node- weighting schemes on directed graphs of single papers using an external KB. Neethukrishnan and Swaraj [40] build personal ontologies for users from concepts of associated articles using the ACM taxonomy, WordNet and SVMs. The overall goal of IBM PARSe [45] is identifying papers a user should read to update an existing KG. Named entities of papers are extracted and papers closest related to the existing KG are recommended. Tang et al. [47] use an external KB to identify concepts in papers which they connect in a KG including users and papers. Wang et al. [50] feed embedded paths from a user-paper-interaction KG into an LSTM.

While previous works use information from KGs to enrich documents, we transform each document into a small KG. Each of these document graphs then represents the document’s essential information (concept interactions).

Explainable Systems. Missing explainability is one of the widespread problems in paper recommendation [25]. TIGRIS [11] provides a responsive graph linking keyword nodes to nodes representing papers fitting a user’s interests. The graph substitutes a ranked result list. ArXivDigest [17] is a living lab for providing personalized recommendations from arXiv papers with textual explanations. JTIE [53] describes explainability as an important aspect of their approach but does not describe the format of explanations. Bakalov et al. [5] highlight terms matching user’ interests in abstracts of recommended papers. Kangasrääsiö et al. [23] do not explain recommendations but let users weight the input keywords in a visual interface, thus indirectly explaining the suggested papers. LIMEADE [32] is a general purpose posthoc explanation approach producing weighted interpretable features, e.g., influential terms of recommended candidate papers.

Existing work often relies on textual features such as highlighting terms in abstracts. As users of the underlying DL have already been encountered to prefer graph over textual explanations [26] we decided on focusing on an inherently explainable recommendation method producing graphs as explanation.

³<https://youtu.be/oLZFCtVuQWU>

⁴<https://github.com/HermannKroll/NarrativeRecommender/>

⁵Software Heritage ID:swlh:1dir:eaac5c6a9ccb00542431398e43dec34d910faf

⁶<https://pubmed.ncbi.nlm.nih.gov/help#computation-of-similar-articles>

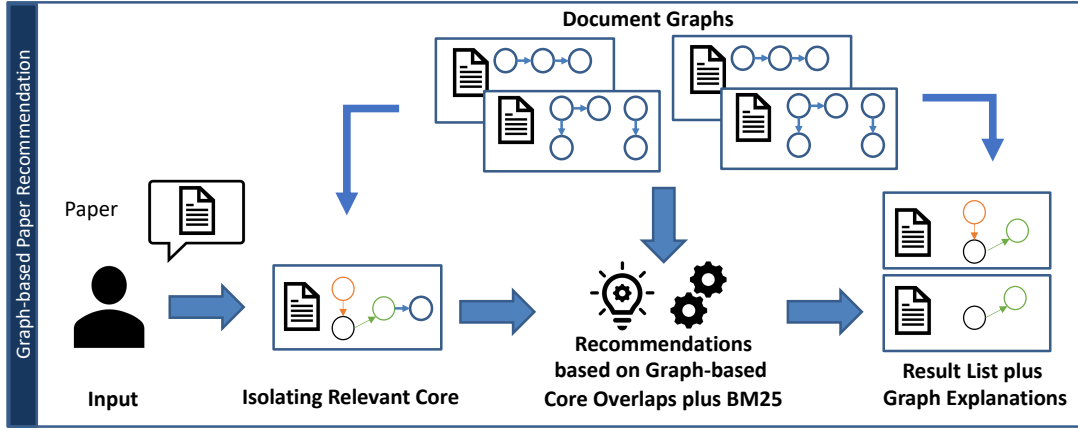


Figure 1: Systematic overview: Document graph representations are used to compute and explain relevant paper recommendations for users.

2.2 Narrative Service - Graph-based Discovery

PubPharm’s graph-based retrieval service, called the Narrative Service [28], allows users to express their information need as a narrative query graph, i.e., graph patterns with triple-like statements (concept, interaction, concept). A query is then answered by documents that contain the search graph pattern. For the query processing step, they transformed texts into graphs by detecting relevant concepts and extracting their interactions. Concepts were identified by deriving annotations from the PubTator service [51, 52] and performing a dictionary-based concept linking through vocabularies derived from ChEBML [38], Wikidata [48] and the Medical Subject Headings. Statements (concept interactions) were extracted by using PathIE and a sentence-based extraction method that extracts general *association* statements if two concepts were mentioned within the same sentence. Briefly, PathIE extracted a statement between two concepts if two detected concepts were connected on the dependency parse of a sentence (basically the grammatical structure) derived with the Stanford CoreNLP toolkit [36]. Details about the extraction methods have already been published in [27] and [28]. The methods’ and system’s code is freely available^{7, 8}. The discovery system currently features around 38 million publications from the MEDLINE collection and 70k COVID-19 pre-prints from ZB MED’s preVIEW service [31].

3 Graph-based Recommendation and Explanation

The discovery system contains a set of documents \mathcal{D} [28]. A document $d \in \mathcal{D}$ is represented by its text d_{text} (title and abstract), a list of concept annotations $d_{concepts}$ and a list of extracted statements $d_{statements}$. A concept annotation maps a specific text span of d to a pre-known concept of the concept vocabulary C (the set of all known concepts by the system). Concepts are identified by precise IDs and a type, e.g., *drug* or *disease*. An extracted statement is composed of a subject-predicate-object triple, e.g., (*Metformin*,

treats, *Diabetes*), the sentence it was extracted from, and a confidence value (how good the extraction was based on some NLP method). Sect. 2.2 contains details about applied methods and details. The **document graph** of d is the set of subject-predicate-object triples extracted from the document. A document graph is given by $graph(d) = (V, E)$ where V is the set of nodes (detected concepts) and E is a set of concept interactions (triples). This paper aims to reuse our existing document graph representation [28] to perform a reliable and explainable paper recommendation; see Figure 1.

Task Definition (Explainable Paper Recommendation):

Given an input document d , compute a ranked list of documents $\subseteq \mathcal{D}$. For each document d_i in that list provide an explanation e_i of why d_i is related to the input d .

3.1 Scoring Document Graph Components

First, we score the nodes and edges of a document graph to know which parts are the most important by using three distinct features.

Tf-idf. Some graph nodes or edges might carry more information (are more relevant) than very general ones. In information retrieval, term-frequency *tf* and inverse-document-frequency *idf* are two paradigms used to determine a term’s relevance concerning a document d . We follow that paradigm and design a *tf-idf* score for nodes and edges. We define *tf* for a concept c within a document d as $tf(c, d) = \frac{\#(c, d)}{\#C}$ with $\#(c, d)$ being the number of occurrences of concept c within d and $\#C$ being the number of all annotated concepts within d (for normalization). Next, we define *idf* for a concept c as $idf(c) = \log \frac{|\mathcal{D}|}{|\{d \in \mathcal{D} \mid c \in d\}|}$. $|\mathcal{D}|$ is the number of documents in our collection, and the denominator counts documents that include the concept c . With that, we can score nodes with *tf-idf* as follows:

$$n\text{-tf-idf}(n, d) = tf(n, d) \cdot idf(n). \quad (1)$$

For edges, we faced issues when maintaining an *idf* index: First, the index can get quite large (quadratic growth with regard to the size of the concept vocabulary). Second, our statement extraction methods are restricted to sentence levels and might be error-prone [28]. Many connections might be lost during that step,

⁷<https://github.com/HermannKroll/NarrativeIntelligence>

⁸Software Heritage ID:swb:1:dir:9e2435bb03d544039cc96fa1b17537050faec6e3

affecting the idf score. That is why we decided to approximate the tf-idf-score for an edge by combining the tf-idf scores of its subject and object plus multiplying it with a predicate specificity (basically a score determining how specific a predicate is: *treats* is more specific than a general *association*). We define the tf-idf score for an edge $e = (s, p, o)$ concerning a document d as:

$$e\text{-tf-idf}(e, d) = (n\text{-tf-idf}(s, d) + n\text{-tf-idf}(o, d)) \cdot \text{specificity}(p) \quad (2)$$

Coverage. The discovery system is designed for biomedical abstract retrieval. Each abstract typically starts with some background information in the corresponding field. Concept mentions within that background part might be less important than concepts mentioned across the whole abstract. We therefore define *coverage* of a node (concept) n within a document d as:

$$n\text{-coverage}(n, d) = \frac{\text{last_position}(n, d) - \text{first_position}(n, d)}{\text{text_length}(d)} \quad (3)$$

The method calculates the difference between the concept's last mention and the first mention within the document and normalizes it by its text length. Coverage approximates whether the concept is used from the beginning to the end or briefly mentioned somewhere as a side note. The higher the coverage is, the more relevant a concept c should be. We then define the coverage of a document graph's edge $e = (s, p, o)$ as:

$$e\text{-coverage}(e, d) = \min(\{n\text{-coverage}(s), n\text{-coverage}(o)\}). \quad (4)$$

Confidence. As mentioned, our extraction methods come with a confidence score, i.e., a score of how sure the tool is about the corresponding extraction. Please note that a document graph's edge could be extracted from different sentences within d . The **confidence** for an edge e is defined as the maximum confidence value of the statement extractions within $d_{\text{statements}}$ that support e . We do not have confidence values for concept annotations because the detection methods are dictionary-based linking methods that perform a binary decision.

Scoring. Finally, we can define the scores of nodes and edges. Coverage and tf-idf are combined to compute the score for each node:

$$n\text{-score}(n, d) = n\text{-coverage}(n, d) \cdot n\text{-tf-idf}(n, d) \quad (5)$$

For edges, we combine confidence, coverage, and tf-idf:

$$e\text{-score}(e, d) = \text{confidence}(e, d) \cdot e\text{-coverage}(e, d) \cdot e\text{-tf-idf}(e, d) \quad (6)$$

Graph Cores. With our previous scoring functions, we can now determine the relevance of different components of each document graph. For instance, the statement extraction step might yield multiple edges (with different predicates) between two nodes, e.g., a general *association* and a specific *treats* predicate. For our recommendation step, we filter the graphs by only keeping the most relevant (best-scored) edge between two concepts and by only keeping edges between different concept types, e.g., drug-disease treatments. A **graph core** is a scored document graph (i.e., the scoring functions have been applied to each node and edge) that only keeps the best-scored edge between two nodes. The function **graph-core**(d) = d_{core} takes a document and returns its core.

3.2 Candidate Retrieval (First Stage)

The discovery system contains about 37 million documents (as of 04/2024). Given some input document d , comparing its core to that of every other document is obviously too expensive. That is why we headed for a two-step approach: A cheap first stage for initial candidate retrieval and a more expensive second stage that utilizes our graph cores and re-scores the candidate documents. Given the document collection \mathcal{D} and some input document d , its graph d_{graph} and its graph core d_{core} , we propose the following first stages:

Edge-driven (FSCore). The idea for this first stage is that the more edges some document $d_{\text{candidate}}$ contains with the same concepts in relation as in d_{core} , the better it is. We iterate over each edge $e = (s, p, o)$ of d_{core} . We then search for documents containing an edge between s and o (we do not force the edge's predicate to be p to be more flexible). We add those documents to a set of candidate documents $D_{\text{candidate}} \subseteq \mathcal{D}$. We add the score $e\text{-score}(e, d)$ to each candidate document that contains the searched edge. We continue with our iteration. Finally, we retrieve a scored list of documents: The higher the score is, the more overlap it has with edges of d_{core} .

Node-driven (FSNode). FSCore requires that the document d has a core (more than zero edges must have been extracted). However, due to error-prone NLP methods and extractions on sentence levels (but not across sentences), we might lack recall from the underlying discovery system. That is why we also introduce a concept-driven first stage. We apply a similar idea to that in FSCore, but this time, we do not search for edges; instead, we search for nodes. That means we iterate over the annotated concepts d_{concepts} , retrieve documents that contain the concept as one of its nodes, and add the concept's score to them.

Concept-driven (FSConcept). FSNode does not force that the input document d has a graph but searches for candidate documents with d 's concepts on their graphs. As an alternative, we propose FSConcept, which just requires that the searched concepts be detected in the candidate documents but do not necessarily need to be contained on their graphs. FSConcept allows a concept-driven retrieval without relying on graphs in the input or candidate documents.

Cutoff. We restrict our first stages by a fixed cutoff value, k , so we return only the best-scored k documents. If the score is equal, we sort documents by their IDs in descending order as our system maintains PubMed IDs and higher IDs usually mean newer publications. While BM25 computes nearly continuously distributed scores because it also considers the tf-idf scores of terms within candidate documents, our first stages come with a step function (either a component of the input is contained or not). For instance, the documents between rank 900 and 1200 could have the same score, as they contain the same overlap to the input. For that, we propose a flexible cutoff which considers the score at position k and then cuts the list at the next position the score drops again. However, a flexible cutoff may result in very large lists, that is why, we use a hard cutoff at $2 \cdot k$ in any case, to have a maximum boundary.

3.3 Recommendation (Second Stage)

The first stage returns a list of candidate documents $D_{\text{candidates}}$. For our second stage, we compute each candidate document's score by comparing their cores to our input document. Our proposed method is shown in Algorithm 1. In brief, if the input document

Algorithm 1 Recommendation Strategy based on Document Cores

```

1: Input:  $d_{input}, D_{candidates}$  Output: a ranked list of documents
2:  $core_{input} = graph-core(d_{input})$ 
3: if  $core_{input} = \emptyset$  then
4:   return candidate documents without sorting them
5: for  $d_{candidate} \in D_{candidates}$  do
6:    $score[d_{candidate}] = 0$ 
7:    $core_{candidate} = graph-core(d_{candidate})$ 
8:   for  $e \in (core_{input} \cap core_{candidate})$  do
9:      $score[d_{candidate}] = [d_{candidate}] + e-score(e, d_{input})$ 
10: return sorted candidate documents by their scores

```

does not have a core, we cannot compute scores and consider all input documents as equally relevant. If the input document has a core, we compare it to every core of the candidate documents. The score for the candidate document is defined as the sum of all edges shared between the input document core and the candidate document's core. An edge is considered as *shared* if it connects the same concepts. This strategy, however, relies on the existence of cores and the expression of relevant information in these cores. This might not always be the case: First, extraction methods are error-prone, i.e., relevant information might be lost. Second, our concept vocabulary might not contain *all* relevant concepts of that domain [28]. That is why we also integrate text-based scoring to consider not-as-graph-expressed information. Here, we use BM25 scoring by considering titles and abstracts. Let d_i be the input document and d_c be some candidate document. We compute the final score, as a weighted sum of the graph overlap and BM25 score:

$$XGPRC(d_i, d_c) = w_{graph} \cdot core-overlap(d_i, d_c) + w_{text} \cdot BM25(d_i, d_c) \quad (7)$$

The *core-overlap* returns the score of the core overlap based on our previous algorithm, and BM25 returns the BM25 score when comparing the text (title plus abstract) of d_i and d_c . Note that we normalize *core-overlap* and BM25 scores with regard to a candidate document list $D_{candidate}$ so documents receive scores between [0, 1] which makes both scores comparable and combineable.

3.4 Explanation Generation

Algorithm 2 shows our explanation approach. The algorithm takes an input document, some candidate document, and a parameter l as its input. The parameter l determines the length of the explanation to generate. Our idea is to take l edges shared between the input documents and the candidate and mark them as shared (later visualized in colors). In addition, we take up to $l \cdot 2$ edges of the candidate document, mark them as not shared (later visualized as dashed lines and not-colored nodes), and add them to our explanation. More precisely, we only consider the edges of the candidate core connected to one node shared between the input and candidate core. These additional edges should help the user to understand what the candidate documents add as new information to the shared pattern. This way users simultaneously see what is shared and what can be expected as new information in the candidate document.

Algorithm 2 Explanation Generation Algorithm

```

1: Input:  $d_{input}, d_{candidate}, l$  Output: A list of shared/not shared edges
2:  $core_{input} = graph-core(d_{input})$ 
3:  $core_{candidate} = graph-core(d_{candidate})$ 
4: for  $e \in sort\ by\ score(core_{input} \cap core_{candidate})$  do
5:    $explanation = explanation \cup (e, shared)$ 
6:   if  $|explanation| \geq l$  then
7:     break
8: for  $e = (s, p, o) \in sort\ by\ score(core_{candidate} \setminus core_{input})$  do
9:   if  $|explanation| \geq 2 \cdot l$  then
10:    break
11:   if  $s \in nodes(core_{input})$  then
12:      $explanation = explanation \cup (e, o\ not\ shared)$ 
13:   if  $o \in nodes(core_{input})$  then
14:      $explanation = explanation \cup (e, s\ not\ shared)$ 
15: return explanation

```

4 Implementation and Prototype

We implement our recommendation algorithm by building upon our [28] discovery system's code base. The discovery system already maintains indexes that can be used to estimate the idf scores for nodes and edges. More precisely, an index maps a concept to a list of document IDs in which the concept has been detected. For the idf computation, it is enough to store a dictionary that maps concepts to the number of documents it appears in. This index contains 1M concepts and takes 40MB of space. We keep it in memory. Suppose a user enters a document ID through manual input or via a link to our system. In that case, we retrieve the document's data from the database, perform first-stage retrieval, select the k best-scored documents, and then retrieve the actual candidate document data for the recommendation. Thus, we load complete document data only if required.

The graph-based discovery system [28], which we used to implement XGPRC has only utilized concept and graph information of documents. The system maintains two indexes: A reverse index for concepts and a reverse index for edges, allowing it to retrieve documents that support a certain concept/edge (retrieve a set of document IDs). Our first stage can thus reuse both indexes. Both indexes are stored as database tables and details are available in the service's code repository. We ignored 20 very generic concepts like *therapy* or *humans* as they were detected in more than 1M documents and thus carry few information. For the BM25 computation, we created a new BM25 index by utilizing PyTerrier [35] (a Python Wrapper around the well-known Terrier toolkit). This index requires about 9.2GB of disk space and can optionally be retrieved from disk or directly from memory.

Parameters. For the final XGPRC score, we slightly prefer graph scores over text scores, i.e., we set $w_{graph} = 0.6$ and $w_{text} = 0.4$. We set predicate specificity scores (see tf-idf score for edges) based on each predicate's hierarchical level in the three-level predicate taxonomy (most-specific predicates received a score of 1.0, one level

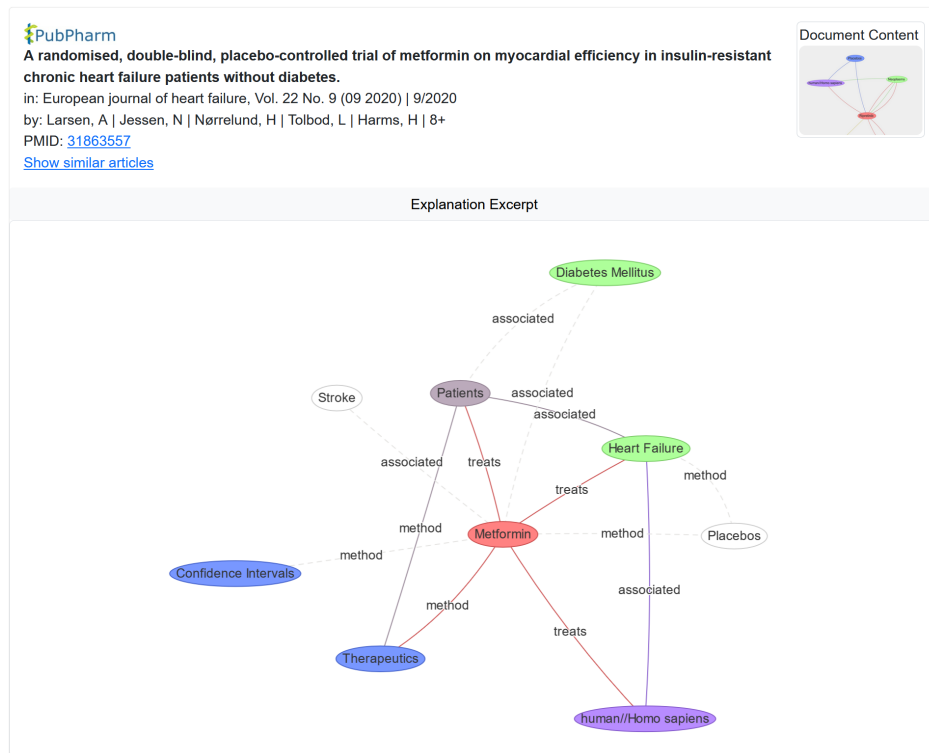


Figure 2: Screenshot of our prototypical system: The generated explanation why the candidate document should be relevant to the input document is shown. Shared information (nodes and edges) are visualized with colors whereas information that is added by the candidate document is visualized as dashed lines and not colored nodes.

higher 0.5, and the highest level (only associated) 0.25) defined by the discovery system⁹. We set $k = 1000$.

User Interface. Figure 2 shows a screenshot of our prototype. In our user interface the user can enter a document ID as an input. Then, a list of candidate documents is retrieved and ranked via our recommendation strategy. For each entry of that list, we generate an explanation and visualize it as shown in Figure 2. We used the same color encoding used in the Narrative Service to visualize document graphs (e.g., red for drugs, green for diseases [28]). Our graph patterns should help users quickly determine the information scent of the recommendation list – a feature not available in other systems. For the explanation visualization, we tested different l values (no. of shown edges when generating an explanation). We found that six (six edges of the core overlap and six additional edges from the candidate document) generates a graph pattern that fits into the user interface concerning the available space. We also improved our old document graph visualization [28]. The old system basically showed all extracted statements as one graph. Our improved visualization now shows document graphs that are reduced to the most essential parts by edge scoring, which should decrease the complexity for users. However, users can still control which concept types and thus, which interactions between concept types, are shown as well as the number of statements to show (default are ten statements). Our visualization is shown in Figure 3.

⁹See the taxonomy at <https://narrative.pubpharm.de/help/>.

5 Evaluation

We (1) describe the used test collections, (2) evaluate the recall and runtime of our first stage, (3) report the precision and recall of our recommendation strategy, and (4) close with a human perspective of our prototypical system.

5.1 Test Collections and Baselines

We use the following test collections:

TREC Genomics 2005 (Genomics) [19]: (50 topics/2525 input documents) contains natural language questions about genes, interactions, processes and methods for ad-hoc biomedical document retrieval. The information retrieval test collection was built with the MEDLINE 2005 (3.7M documents). It has already been used in previous paper recommendation evaluations [18, 34, 54]. We follow Zhang et al. [54] in selecting relevant articles (judged as 2 - relevant) per topic as input documents while considering all other articles belonging to the same topic as potential candidate documents with their corresponding judgments (2 - relevant/1 - partially relevant/0 - not relevant). We used the default configuration of trec_eval to compute the scores. Then, we iterate over all relevant (2) articles of a topic, perform the recommendation, compute scores, and then average the scores for all input documents per topic. In this way, we derive an average score per topic by iterating over all combinations (input vs. documents to recommend).

View Paper

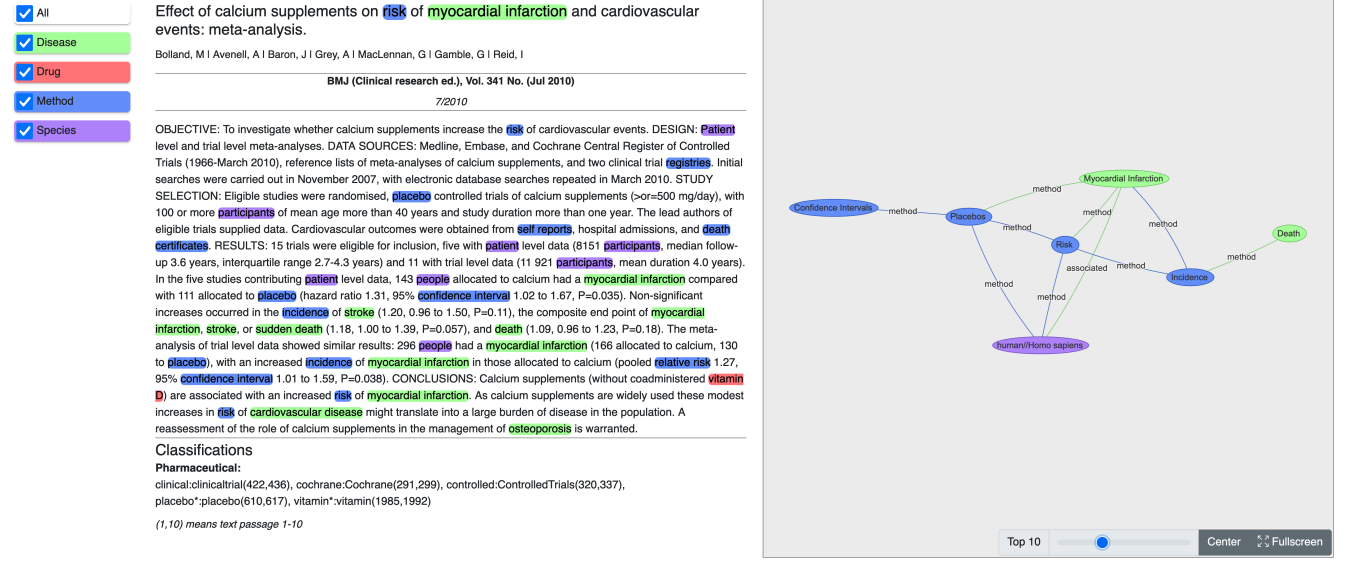


Figure 3: Screenshot of our improved document visualization: On the left side, detected concepts are highlighted in the text via a color encoding. On the right side, the essential document graph is shown, i.e., the most relevant extracted statements are shown to the users. Users can on the left side filter for certain concept types (diseases, drugs, etc.) and on the right side select the number of statements that should be shown.

Table 1: Evaluation of our first stages: The set recall and average retrieval time per document is shown. We used a cutoff $k = 1000$. The flexible cutoff is reported as F .

Ranking	PM2020		Genomics		RELISH	
	Recall	T_{doc}	Recall	T_{doc}	Recall	T_{doc}
FScore	0.52 0.53 ^F	1.3s	0.22 0.22 ^F	0.1s	0.16 0.17 ^F	0.3s
FSNode	0.60 0.61^F	0.8s	0.27 0.29 ^F	0.1s	0.22 0.24 ^F	0.4s
FSConcept	0.60 0.61^F	0.9s	0.27 0.29 ^F	0.1s	0.23 0.25 ^F	0.5s
BM25 Title	0.50	1.1s	0.33	0.2s	0.61	0.6s
BM25 Title + Abstract	0.57	9.3s	0.41	1.2s	0.80	8.3s

TREC Precision Medicine 2020 (PM2020) [43]: (31 topics/1192 input documents) is a biomedical document retrieval test collection that asks for treatment options (drug), cancer forms (disease), and a gene variant (gene/target). The collection was built with the MEDLINE 2019 (29.1M documents). We performed our evaluation analogous to Genomics.

RELISH [10]: (3278 input document queries) is a biomedical paper recommendation test collection with 161184 distinct documents which have been rated by experts. It was built with the MEDLINE 2018 (26.7M documents).

Data Processing. Our discovery system [28] already included the latest version of MEDLINE, so the benchmark documents have already been processed and transformed into document graph representations. For our evaluation, we computed a list of document IDs belonging to the different MEDLINE versions (2005, 2018, and 2019). This list was used as a filter to retrieve only documents in our first stages that were initially considered in the benchmarks.

Baselines. We compare our first stages to BM25. Related work rarely publishes their code or describes their methodology in enough detail to enable reimplementations [25]. Therefore, we decided to compare against a method applicable to real-world data, that is a pre-existing system, namely PubMed [34] (see Sec. 2).

As another configuration for the second stage we compared XGPRC to the documents retrieved by BM25 on titles and abstracts (the best first stage) reranked with SPLADE [16]. SPLADE uses lexical matching and term expansion for the embedding of query and documents and achieved results on par with state-of-the-art dense retrieval models. Similarity between SPLADE-embedded¹⁰ candidate document vectors and query documents is computed with the Faiss [15, 22] L2 norm. We then identify the k documents to recommend by applying knn.

¹⁰By the naver/splade-cocondenser-ensembledistil model and max pooling.

Table 2: Detailed evaluation of our recommendation approach XGPRec.

Dataset	Strategy	T_{doc}	Recall	nDCG@10	nDCG@20	P@10	P@20	bpref
PM2020	XGPRec	$0.45 \pm 0.38s$	0.61	0.30	0.30	0.33	0.30	0.30
	- BM25	$0.44 \pm 0.38s$	0.61	0.25	0.25	0.28	0.26	0.29
	- CoreOverlap	$\leq 0.1 \pm 0.0s$	0.61	0.32	0.31	0.34	0.30	0.31
	BM25 Title	1.1s	0.50	0.23	0.23	0.25	0.23	0.25
	BM25 Title + Abstract	9.3s	0.57	0.29	0.28	0.31	0.28	0.28
	BM25 T+A + SPLADE	-	0.57	0.22	0.2	0.23	0.2	0.25
	PubMed Rec.	-	0.29	0.30	0.30	0.33	0.29	0.17
Genomics	XGPRec	$0.13 \pm 0.18s$	0.29	0.20	0.19	0.21	0.18	0.15
	- BM25	$0.13 \pm 0.18s$	0.29	0.14	0.13	0.16	0.13	0.14
	- CoreOverlap	$\leq 0.1 \pm 0.0s$	0.29	0.24	0.22	0.25	0.21	0.16
	BM25 Title	0.2s	0.33	0.19	0.18	0.21	0.18	0.17
	BM25 Title + Abstract	1.2s	0.41	0.26	0.25	0.28	0.24	0.21
	BM25 T+A + SPLADE	-	0.41	0.19	0.18	0.20	0.17	0.17
	PubMed Rec.	-	0.13	0.23	0.22	0.24	0.20	0.09
RELISH	XGPRec	$0.10 \pm 0.33s$	0.25	0.22	0.19	0.22	0.17	0.22
	- BM25	$0.10 \pm 0.33s$	0.25	0.09	0.08	0.09	0.07	0.21
	- CoreOverlap	$\leq 0.1 \pm 0.0s$	0.25	0.35	0.29	0.35	0.26	0.22
	BM25 Title	0.6s	0.61	0.33	0.29	0.33	0.27	0.47
	BM25 Title + Abstract	8.3s	0.80	0.54	0.49	0.57	0.47	0.59
	BM25 T+A + SPLADE	-	0.80	0.29	0.26	0.30	0.23	0.57
	PubMed Rec.	-	0.47	0.55	0.51	0.60	0.52	0.41

5.2 First Stage Evaluation

Table 1 shows the recall and average retrieval time per document for different first stages, i.e., FSCore, FSNode, FSConcept, BM25 with titles (BM25 T) and BM25 with titles and abstracts (BM25 T+A). We measured the runtime performance on our server, which has two Intel(R) Xeon(R) Gold 6336Y CPUs @ 2.40GHz (24 cores and 48 threads each), 2TB DDR4 main memory, and nine Nvidia A40 GPUs with 48GB memory. All time measurements were executed four times. The first run was a cold start to load all required data into main memory, and the reported results were averaged over the last three runs. All data for BM25 was also loaded into main memory, and no multi-threading/processing was used.

In concept-centric cases like PM2020, FSConcept, and FSNode achieved the highest recall of 0.6-0.61 and the lowest retrieval time of 0.8-0.9s per input document. In comparison, BM25 T+A achieved a recall of 0.57 by requiring about 9.3s per input document. On Genomics, our methods achieved a recall comparable to BM25 T (-0.04) with a retrieval time comparable to BM25 T (0.1s vs. 0.2s). BM25 T+A achieved a recall of 0.41 but required about 1.2s per document. Please note that these times are measured by using a BM25 index of only 3.7M documents, compared to 9.3s (BM25 T+A) on 29.1M documents in PM2020. On RELISH, BM25 T and BM25 T+A outperformed our first stages’ recall. In brief, the runtime for our first stages are always below BM25 retrieval and in concept-centric scenarios, our first stages are effective (PM2020) or comparable (Genomics). We discuss shortcomings on RELISH in Section 5.5.

5.3 Recommender Evaluation

Our first stages FSConcept and FSNode showed a comparable performance. We decided to use FSConcept with a flexible cutoff as the first stage for our recommendation approach because (1) we do not need an additional index (as the node graph index for FSNode) and (2) FSConcept is less restrictive than FSNode (it just requires concepts to be annotated in documents and not that these concepts need to appear on the document graph). Table 2 shows the results of our recommendation approach (XGPRec) compared to the PubMed Recommender. We also filtered the PubMed Recommender results so that only documents are used for the evaluation considered in the test collections. In addition, we report the results of using XGPRec without the core overlap score (-CoreOverlap) and without the BM25 score (-BM25). In general, our recommendation strategies took less than 1s per document for the computation (see T_{doc} in Table 2). On PM2020 and Genomics, XGPRec achieved comparable nDCG and precision to the PubMed Recommender but nearly doubled recall (0.61 vs. 0.29 and 0.29 vs. 0.13). However, PubMed’s lower recall could be explained by smaller result sets (maybe because of some internal precision-oriented cutting strategy). On RELISH, however, our method was outperformed by the PubMed Recommender. Another observation was that XGPRec without the CoreOverlap component achieved the highest scores, i.e., by just using the BM25 scores.

In comparison, using only BM25 on titles or titles and abstracts of input documents instead of employing a recommender system

Table 3: Differences between XGPRC, BM25 and the PubMed recommender. $J@k$ gives the average Jaccard coefficient over the sets of documents at top k recommendations for the three benchmarks PM2020, Genomics, and RELISH.

XGPRC vs.	Dataset	$J@5$	$J@10$	$J@20$	$J@100$
BM25 Title+Abstract	PM2020	0.24	0.25	0.26	0.30
	Genomics	0.23	0.23	0.23	0.26
	RELISH	0.10	0.10	0.10	0.09
PubMed	PM2020	0.13	0.14	0.16	0.18
	Genomics	0.13	0.14	0.15	0.13
	RELISH	0.08	0.07	0.07	0.07

achieves good results on the three benchmarks¹¹. The variant using abstracts unsurprisingly produces higher recall, nDCG, precision and bpref than the one using titles only. The comparably high execution time makes this strategy uneligible in a real system. Results produced by BM25 title are comparable to XGPRC except for the RELISH case.

For the combination of BM25 on titles and abstracts (BM25 T+A) as the first stage and using SPLADE and Faiss for reranking in the second stage we do not report recommendation times. SPLADE embeddings can theoretically be pre-computed for the whole document corpus¹² thus our measured times would not reflect the actual optimized usage options of Faiss (see [15, 22]). The BM25 T+A + SPLADE run produces results for nDCG, precision and bpref that are considerably worse than XGPRC on PM2020, similar on Genomics and considerably better on RELISH. As this variant is a reranking of the results of BM25 T+A, their recall does not differ.

5.4 Differences between XGPRC and PubMed

Table 3 contains average Jaccard coefficients of sets of documents from top k recommendation comparing XGPRC with BM25 using titles and abstracts as well as XGPRC with the PubMed recommender on our benchmarks. The overlap in documents is similar between different k per row, that is between benchmark and comparison. The overlaps between PM2020 and Genomics are comparable as well. The overlaps between recommended documents using XGPRC and the BM25 strategy as well as PubMed is considerably lower in the RELISH dataset. In general, XGPRC recommends documents more similar to the BM25 baseline than PubMed. Recommendations made by XGPRC are dissimilar from those made by the PubMed recommender. This dissimilarity can be regarded with the comparable performance of XGPRC and PubMed on the PM2020 and Genomics benchmarks to conclude that XGPRC presents documents to its users which they would not be recommended using PubMed but which are of similar relevance, as estimated by XGPRC. We therefore argue that XGPRC holds value in helping users to uncover relevant documents which the state-of-the-art in-use recommendation system does not suggest.

¹¹Note that both PM2020 and Genomics are retrieval benchmarks and that BM25 is a retrieval method. These benchmarks were built by annotation of documents which have been highly ranked by retrieval systems, see Sec. 5.5.

¹²Which we refrained from doing as we would have been faced with embedding 38M documents which was not feasible for our digital library use case.

5.5 Result Discussion

Usually, collections are crafted by using some initial pooling strategy of candidate documents for each topic and then showing them to judges who rate whether the document is relevant. RELISH was constructed by using BM25, tf-idf and the PubMed recommender for initial pooling [10]. That explains why the PubMed recommender shows good performance on RELISH. In general, when term-based retrieval is used for an initial pooling, term-based methods like BM25 are automatically favored. Advantages of graph-based retrieval, e.g., synonymous terms that refer to the same concept, are then rather useless than helpful. Especially, when documents that are not judged in the benchmarks are counted as wrong hits.

While some paper recommendation methods *just* divide relevant and irrelevant documents given by some test collection, as done in [10, 21, 42, 54], our system works on a comprehensive document collection. When evaluating the test collections, we observed two central issues: First, our graph-based approach retrieved many documents that have not been judged in the collection data. Those documents are then considered as incorrect results which hampers our nDCG and precision@k scores. However, the metric bpref [13], which ignores unjudged documents, shows that the difference between core overlap and BM25 is small (1-2% points). In fact, under the top-20-retrieved documents with XGPRC, we observed 6.2 ± 5.8 unjudged documents on average per topic for PM2020, 9.8 ± 6.3 for Genomics, and 15.2 ± 5.4 for RELISH. The number of unjudged documents decreased for XGPRC without CoreOverlap and only BM25 scores: 5.9 ± 5.6 for PM2020, 9.0 ± 6.2 for Genomics, and 12.9 ± 5.9 for RELISH. Second, not all test collection documents can be represented by a document graph and salient concepts; hence, our method cannot retrieve documents in those instances. We counted the ratio of input documents that have a core with five edges. This is the case for 96.0% documents for PM2020, 83.0% for Genomics, and 55.0% for RELISH. Thus, we clearly lacked suitable document graph representations on RELISH. The discovery system has been built for the pharmaceutical domain, i.e., mainly for gene, drug, and disease interactions (reflected in PM2020 and Genomics) and not for psychological or orthopedic topics as present in RELISH.

While both issues explain that BM25 achieved higher scores than our graph-based overlap method, our graph method has the advantage of preferring overlapping graphs. Moreover, these overlapping graphs can explain recommendations for users. We already know domain experts like using graphs in the context of literature search to quickly determine if results fit their information needs [26].

While PubMed and XGPRC produce comparable results on PM2020 and Genomics, the actually recommended documents are vastly different from each other. Therefore, we argue XGPRC being a suitable tool to uncover potentially relevant documents which PubMed does not suggest to users. We consider XGPRC not as a **better** recommendation system compared to PubMed for producing relevant recommendations but instead want to emphasize on XGPRC’s value to help discover **different** documents in an explainable manner.

In brief, XGPRC offers explanations by design, can handle a real-world digital library collection with about 37M documents, and offers a comparable performance to the PubMed recommendation (which is likely the most used recommender in the biomedical

domain). Beyond that, XGPRC retrieves result lists that differ from existing benchmarks (results are not judged there) and recommendation approaches. These results are worth of further investigation.

5.6 Human Perspective

To get initial user feedback on our recommendation system we separately interviewed two experts from the pharmaceutical domain for 30 minutes. The experts were also knowledgeable users of the underlying discovery system. An interviewer first briefly described the interface’s functionality, before the domain experts used the system for 20 minutes to satisfy information needs while thinking aloud. Afterwards the interviewer asked some clarifying questions regarding liked or disliked components, missed functionality and general feedback.

Results. Both users looked at five (individually chosen) initial papers and commented on their recommendations’ relevance. They both searched for 1) interactions or 2) keywords contained in an initial paper. They both observed broader and more focused information needs. Our users enjoyed using the system, were very positive about the quality of recommended papers and especially praised the explanation graphs. They were immediately able to use the visualization of the overlap between their initial paper and the candidate suggestions. They were able to interpret the graphs at a glance and distinguish relevant from less relevant suggestions instantly without having to read the title of the papers. One participant only observed the graphs of recommendations and did not click a single detail view while the other preferred to first read titles and if titles were hinting towards relevance, they checked out the graphs. When first screening titles, it would be desirable to have the graphs less prominent. It was mentioned that the graph was considered a better option for screening papers than a keyword list. One person used the option of observing similar articles of recommended articles, thus chaining recommendations to navigate the content of the underlying digital library.

Features which participants missed were a structuring of the result list, e.g., by types of overlapping edges/patterns and the possibility to emphasize/weight edges from input papers which a user considers highly relevant. Purely technical wished for functionality were the option to keep the initial paper displayed on the same page while checking the recommendations, a bookmarking or export option for relevant articles, filtering options on the result list, and a description how the recommendation system determines the suggestions, i.e., a brief description of our used method.

Discussion. Our interview partners used the graph overlap between the initial and a candidate paper intuitively to understand, why a paper was recommended. The explanation therefore satisfies the aims of justification [6]. One study participant mentioned the current interface lacking information how our systems generates the list of recommendations [6] which should be tackled by some info box. Chaining recommendations seems to be a good way of exploring a topic. Even though our evaluation with test collections did not produce the best results numerically, our preliminary user study showed participants’ immediate acceptance of the system and their satisfaction with the recommendations. This could hint at a mismatch between test collections and the human perspective in paper recommendation - collections are absolutely needed to

construct systems but they must be complemented with a user study to further shed light on advantages and limitations of systems.

As a next step in development and after confirming the user interface’s preliminary suitability, we intend to bring the system into a beta phase and run a large scale study in order to explore explanation options for recommendations.

6 Conclusion

This work extended our graph-based discovery system by an explainable paper recommendation component for a real-world digital library document collection. In contrast to many other works, our method (1) is unsupervised, i.e., we do not require training data and a library must thus not collect training data to implement a similar algorithm, and (2) it works on a real, large-scale collection with 37M documents. While our evaluation shows benefits and limitations, we demonstrated an overall comparable performance to the PubMed recommender while suggesting a vastly different set of documents and therefore potentially broadening a user’s exploration space. We argue that precise, concept-centric information needs are common in the biomedical domain, as seen in a PubMed query log analysis [20], PM2020 [43], our previous user study [28], or our discovery system’s query log analysis [26]. For these concept-centric use cases like in PM2020 or Genomics, our system shows its advantages: The first stage is fast and effective and the core overlap recommendation allows to derive graph pattern explanations which help our users. In brief, our recommendation strategy XGPRC is fast in handling an extensive collection, offers a comparable performance to PubMed’s real digital library recommendation system, and provides users with suitable graph explanations. This research demonstrates how graph-based document representations allow beneficial exploration in digital libraries. Beyond that, our code is freely available so that other digital libraries can use or adapt our effective graph-based recommendation implementation.

Explanations must be a focus for future work, i.e., to further improve the visualization strategy by adjusting it to possible user needs. Beyond that, combining graph-based explainable with traditional recommendation methods as well as a user-controlled weighting of graph components are worth of further investigation.

Acknowledgments

Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): PubPharm – the Specialized Information Service for Pharmacy (GepriS 267140244).

References

- [1] Daniel E. Acuna, Zijun Yi, Kartik Nagre, and Priya Matnani. 2021. EILEEN: A recommendation system for scientific publications and grants. *CoRR abs/2110.09663* (2021). arXiv:2110.09663 <https://arxiv.org/abs/2110.09663>
- [2] Mohammad Mehdi Afsar, Trafford Crump, and Behrouz Far. 2021. An Exploration On-demand Article Recommender System for Cancer Patients Information Provisioning. *The International FLAIRS Conference Proceedings* 34 (Apr. 2021). <https://doi.org/10.32473/flairs.v34i1.128339>
- [3] Alexis Allot, Kyubum Lee, Qingyu Chen, Ling Luo, and Zhiyong Lu. 2021. Lit-Suggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.* 49, Webserver-Issue (2021), 352–358. <https://doi.org/10.1093/nar/gkab326>
- [4] Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2019. Scientific Paper Recommendation: A Survey. *IEEE Access* 7 (2019), 9324–9339. <https://doi.org/10.1109/ACCESS.2018.2890388>

- [5] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An approach to controlling user models and personalization effects in recommender systems. In *18th International Conference on Intelligent User Interfaces, IUI 2013, Santa Monica, CA, USA, March 19-22, 2013*. ACM, 49–56. <https://doi.org/10.1145/2449396.2449405>
- [6] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 265–274. <https://doi.org/10.1145/3331184.3331211>
- [7] Jöran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* 17, 4 (2016), 305–338. <https://doi.org/10.1007/s00799-015-0156-0>
- [8] Corinna Breitinger, Kay Herklotz, Tim Fliegelskamp, and Norman Meuschke. 2022. Recommending research papers to chemists: a specialized interface for chemical entity exploration. In *JCDL '22*. ACM, 22. <https://doi.org/10.1145/3529372.3533281>
- [9] Corinna Breitinger, Birkan Kolcu, Monique Meuschke, Norman Meuschke, and Bela Gipp. 2020. Supporting the Exploration of Semantic Features in Academic Literature using Graph-based Visualizations. In *JCDL 2020*. ACM, 377–380. <https://doi.org/10.1145/3383583.3398599>
- [10] Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database J. Biol. Databases Curation* 2019 (2019), baz085. <https://doi.org/10.1093/DATABASE/BAZ085>
- [11] Simon Bruns, André Calero Valdez, Christoph Greven, Martina Ziefle, and Ulrik Schroeder. 2015. What Should I Read Next? A Personalized Visual Publication Recommender System. In *Human Interface and the Management of Information. Information and Knowledge in Context - 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 9173)*. Springer, 89–100. https://doi.org/10.1007/978-3-319-20618-9_9
- [12] Andrew Collins and Jöran Beel. 2019. Document Embeddings vs. Keyphrases vs. Terms for Recommender Systems: A Large-Scale Online Evaluation. In *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*. IEEE, 130–133. <https://doi.org/10.1109/JCDL.2019.00027>
- [13] Nick Craswell. 2009. *Bpref*. Springer US, Boston, MA, 266–267. https://doi.org/10.1007/978-0-387-39940-9_489
- [14] Jonathas José de Magalhães, Cleiton Caetano de Souza, Evandro Costa, and Joseana Macêdo Fachine. 2015. Recommending Scientific Papers: Investigating the User Curriculum. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida, USA, May 18-20, 2015*. AAAI Press, 489–494. <http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10417>
- [15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). [arXiv:2401.08281 \[cs.LG\]](https://arxiv.org/abs/2401.08281)
- [16] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM, 2288–2292. <https://doi.org/10.1145/3404835.3463098>
- [17] Kristian Gingstad, Øyvind Jekteberg, and Krisztian Balog. 2020. ArXivDigest: A Living Lab for Personalized Scientific Literature Recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 3393–3396. <https://doi.org/10.1145/3340531.3417417>
- [18] Hongmei Guo, Zhesi Shen, Jianxun Zeng, and Na Hong. 2021. Hybrid Methods of Bibliographic Coupling and Text Similarity Measurement for Biomedical Paper Recommendation. In *MEDINFO 2021 (Studies in Health Technology and Informatics, Vol. 290)*. IOS Press, 287–291. <https://doi.org/10.3233/SHTI220080>
- [19] William R. Hersh, Aaron M. Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe M. Roberts, and Marti A. Hearst. 2005. TREC 2005 Genomics Track Overview. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005 (NIST Special Publication, Vol. 500-266)*. National Institute of Standards and Technology (NIST). <http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf>
- [20] Jorge R. Hersh, Len Y. Tanaka, William Hersh, and Elmer V. Bernstam. 2007. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *Journal of the American Medical Informatics Association* 14, 2 (03 2007), 212–220. <https://doi.org/10.1197/jamia.M2191>
- [21] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* 39, 11 (11 2023), btad651. <https://doi.org/10.1093/bioinformatics/btad651>
- [22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [23] Antti Kangasräsä, Dorota Glowacka, and Samuel Kaski. 2015. Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, March 29 - April 01, 2015*. ACM, 247–251. <https://doi.org/10.1145/2678025.2701371>
- [24] Özge Kart, Alexandre Mestiasvili, Kurt Lachmann, Richard Kwasnicki, and Michael Schroeder. 2022. Emati: a recommender system for biomedical literature based on supervised learning. *Database J. Biol. Databases Curation* 2022, 2022 (2022). <https://doi.org/10.1093/database/baac104>
- [25] Christin Katharina Kreutz and Ralf Schenkel. 2022. Scientific paper recommendation systems: a literature review of recent publications. *Int. J. Digit. Libr.* 23, 4 (2022), 335–369. <https://doi.org/10.1007/s00799-022-00339-w>
- [26] Hermann Kroll, Christin Katharina Kreutz, Pascal Sackhoff, and Wolf-Tilo Balke. 2023. 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*. IEEE, 143–154. <https://doi.org/10.1109/JCDL57899.2023.00029>
- [27] Hermann Kroll, Jan Pirklbauer, and Wolf-Tilo Balke. 2021. A Toolbox for the Nearly-Unsupervised Construction of Digital Library Knowledge Graphs. In *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021*. IEEE, 21–30. <https://doi.org/10.1109/JCDL52503.2021.00014>
- [28] Hermann Kroll, Jan Pirklbauer, Jan-Christoph Kalo, Morris Kunz, Johannes Ruthmann, and Wolf-Tilo Balke. 2024. A discovery system for narrative query graphs: entity-interaction-aware document retrieval. *Int. J. Digit. Libr.* 25, 1 (2024), 3–24. <https://doi.org/10.1007/s00799-023-00356-3>
- [29] Hermann Kroll, Jan Pirklbauer, Florian Plötzky, and Wolf-Tilo Balke. 2022. A library perspective on nearly-unsupervised information extraction workflows in digital libraries. In *JCDL '22: The ACM/IEEE Joint Conference on Digital Libraries in 2022, Cologne, Germany, June 20 - 24, 2022*. ACM, 35. <https://doi.org/10.1145/3529372.3530924>
- [30] Hermann Kroll, Jan Pirklbauer, Florian Plötzky, and Wolf-Tilo Balke. 2023. A detailed library perspective on nearly unsupervised information extraction workflows in digital libraries. *International Journal on Digital Libraries* (13 06 2023). <https://doi.org/10.1007/s00799-023-00368-z>
- [31] Lisa Langnickel, Roman Baum, Johannes Darms, Sumit Madan, and Juliane Fluck. 2021. COVID-19 preVIEW: Semantic Search to Explore COVID-19 Research Preprints. In *Public Health and Informatics*. IOS Press, Amsterdam, the Netherlands, 78–82. <https://doi.org/10.3233/SHTI210124>
- [32] Benjamin Charles Germain Lee, Kyle Lu, Doug Downey, and Daniel S. Weld. 2020. Explanation-Based Tuning of Opaque Machine Learners with Application to Paper Recommendation. *CoRR abs/2003.04315* (2020). <https://arxiv.org/abs/2003.04315>
- [33] Xinyi Li, Yifan Chen, Benjamin Pettit, and Maarten de Rijke. 2019. Personalised Reranking of Paper Recommendations Using Paper Content and User Behavior. *ACM Trans. Inf. Syst.* 37, 3 (2019), 31:1–31:23. <https://doi.org/10.1145/3312528>
- [34] Jimmy Lin and W. John Wilbur. 2007. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinform.* 8 (2007). <https://doi.org/10.1186/1471-2105-8-423>
- [35] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, 2020*. ACM, 161–168. <https://doi.org/10.1145/3409256.3409829>
- [36] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, System Demonstrations*. The Association for Computer Linguistics, 55–60. <https://doi.org/10.3115/v1/p14-5010>
- [37] Ruben Manrique and Olga Marino. 2018. Knowledge graph-based weighting strategies for a scholarly paper recommendation scenario.
- [38] David Mendez, Anna Gaulton, A Patricia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michal Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodríguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. 2018. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* 47, D1 (11 2018), D930–D940. <https://doi.org/10.1093/nar/gky1075>
- [39] Norman Meuschke, Vincent Stange, Moritz Schubotz, and Bela Gipp. 2018. Hy-Plag: A Hybrid Approach to Academic Plagiarism Detection. In *SIGIR 2018*. ACM, 1321–1324. <https://doi.org/10.1145/3209978.3210177>
- [40] K. V. Neethukrishnan and K. P. Swaraj. 2017. Ontology based research paper recommendation using personal ontology similarity method. In *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 1–4. <https://doi.org/10.1109/ICECCT.2017.8117833>
- [41] Roland Oruche, Vidya Gundlapalli, Aditya P. Biswal, Prasad Callyam, Mauro Lemus Alarcon, Yuanxun Zhang, Naga Ramya Bhamidipati, Abhiram Malladi, and Hariharan Regunath. 2021. Evidence-Based Recommender System for a COVID-19 Publication Analytics Service. *IEEE Access* 9 (2021), 79400–79415. <https://doi.org/10.1109/ACCESS.2021.3083583>

- [42] Rohitha Ravinder, Tim Fellerhof, Vishnu Dadi, Lukas Geist, Guillermo Rocamora, Muhammad Talha, Dietrich Rebholz-Schuhmann, and Leyla Jael Castro. 2023. A Comparison of Vector-based Approaches for Document Similarity Using the RELISH Corpus (*Proceedings SeWebMeDa-2023: 6th International Workshop on Semantic Web solutions for large-scale biomedical data analytics, May 29, 2023, Hersonissos, Greece*). RWTH Aachen, Aachen, Germany. <https://nbn-resolving.org/urn:nbn:de:hbz:1044-opus-75198>
- [43] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. 2020. Overview of the TREC 2020 Precision Medicine Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*. National Institute of Standards and Technology (NIST). <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.PM.pdf>
- [44] Harris Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 2825–2837. <https://doi.org/10.1145/3477495.3531766>
- [45] Bhuvan Sharma, Van C Willis, Claudia S Huettner, Kirk Beaty, Jane L Snowdon, Shang Xue, Brett R South, Gretchen P Jackson, Dilhan Weeraratne, and Vanessa Michelini. 2020. Predictive article recommendation using natural language processing and machine learning to support evidence updates in domain-specific knowledge graphs. *JAMIA Open* 3, 3 (09 2020), 332–337. <https://doi.org/10.1093/jamiaopen/ooaa028>
- [46] Ritu Sharma, Dinesh Gopalani, and Yogesh Kumar Meena. 2023. An anatomization of research paper recommender system: Overview, approaches and challenges. *Eng. Appl. Artif. Intell.* 118 (2023), 105641. <https://doi.org/10.1016/J.ENGAPPAI.2022.105641>
- [47] Hao Tang, Baisong Liu, and Jiangbo Qian. 2021. Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation. *Concurr. Comput. Pract. Exp.* 33, 13 (2021). <https://doi.org/10.1002/CPE.6227>
- [48] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. <https://doi.org/10.1145/2629489>
- [49] Bangchao Wang, Ziyang Weng, and Yanping Wang. 2021. A Novel Paper Recommendation Method Empowered by Knowledge Graph: for Research Beginners. *CoRR abs/2103.08819* (2021). arXiv:2103.08819 <https://arxiv.org/abs/2103.08819>
- [50] Xiao Wang, Hanchuan Xu, Wenjie Tan, Zhongjie Wang, and Xiaofei Xu. 2020. Scholarly Paper Recommendation via Related Path Analysis in Knowledge Graph. In *2020 International Conference on Service Science, ICSS 2020, Xining, China, August 24-26, 2020*. IEEE, 36–43. <https://doi.org/10.1109/ICSS50103.2020.00014>
- [51] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research* 47, W1 (05 2019), W587–W593. <https://doi.org/10.1093/nar/gkz389>
- [52] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research* 41, W1 (05 2013), W518–W522. <https://doi.org/10.1093/nar/gkt441>
- [53] Yi Xie, Shaoqing Wang, Wei Pan, Huaibin Tang, and Yuqing Sun. 2020. Embedding Based Personalized New Paper Recommendation. In *Computer Supported Cooperative Work and Social Computing - 15th CCF Conference, ChineseCSCW 2020, Shenzhen, China, November 7-9, 2020, Revised Selected Papers (Communications in Computer and Information Science, Vol. 1330)*. Springer, 558–570. https://doi.org/10.1007/978-981-16-2540-4_40
- [54] Li Zhang, Wei Lu, Haihua Chen, Yong Huang, and Qikai Cheng. 2022. A comparative evaluation of biomedical similar article recommendation. *J. Biomed. Informatics* 131 (2022), 104106. <https://doi.org/10.1016/J.JBI.2022.104106>
- [55] Zitong Zhang, Braja Gopal Patra, Ashraf Yaseen, Jie Zhu, Rachit Sabharwal, Kirk Roberts, Tru Cao, and Hulin Wu. 2023. Scholarly recommendation systems: a literature survey. *Knowl. Inf. Syst.* 65, 11 (2023), 4433–4478. <https://doi.org/10.1007/S10115-023-01901-X>