Supporting the Exploration of Semantic Features in Academic Literature using Graph-based Visualizations

Corinna Breitinger[1,3], Birkan Kolcu[2], Monique Meuschke[1], Norman Meuschke[1,3], Bela Gipp[1,3]

[1]University of Wuppertal, Germany (last@uni-wuppertal.de)
[2]University of Arizona, USA (birkankolcu@email.arizona.edu)
[3]University of Konstanz, Germany (first.last@uni-konstanz.de)

ABSTRACT

Literature search and recommendation systems have traditionally focused on improving recommendation accuracy through new algorithmic approaches. Less research has focused on the crucial task of visualizing the retrieved results to the user. Today, the most common visualization for literature search and recommendation systems remains the ranked list. However, this format exhibits several shortcomings, especially for academic literature. We present an alternative visual interface for exploring the results of an academic literature retrieval system using a force-directed graph layout. The interactive information visualization techniques we describe allow for a higher resolution search and discovery space tailored to the unique feature-based similarity present among academic literature. *RecVis* – the visual interface we propose – supports academics in exploring the scientific literature beyond textual similarity alone, since it enables the rapid identification of other forms of similarity, including the similarity of citations, figures, and mathematical expressions.

CCS CONCEPTS

• Information systems → Recommender systems; Specialized information retrieval; Content analysis and feature selection;
• Human-centered computing → Usability testing

KEYWORDS

Recommender Systems; Information Visualization

ACM Reference format:

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 *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'20). August 1-5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3383583.3398599

1 INTRODUCTION

The discovery of relevant scientific literature is a tedious and time-consuming task. Approximately 3 million new papers are published annually [1], and well over 50 million research papers are in circulation today [2]. Given this large volume and rapid growth of publications, researchers can easily miss the content most relevant to them. At the same time, identifying the specific features in publications that might be relevant to a researcher's information need is becoming more tedious and daunting.

Academic search and recommendation systems facilitate the information retrieval and discovery process. However, current literature recommendation systems are not using the full spectrum of available text-independent feature analysis methods to determine article relevance. Text-independent feature analysis methods for similarity assessment include citation-based measures [3], mathematical formulae analysis [4], and image-based retrieval methods [5, 6]. Thus far, no system has combined these methods to support the literature recommendation use case. Furthermore, the most commonly employed visualization method for literature recommendations remains a ranked list sorted in descending order of a predicted relevance score. However, this format entails two significant disadvantages:

1. Inefficient user exploration of recommended items. Since a list is, by definition, a space-saving format, it can only show the most condensed and superficial information, e.g., title, author names, and possibly the venue and publishing date. Lists do not support the discovery of deeper content, since metadata alone cannot inform readers about the type of content present in the recommended documents. To communicate the presence of *text-independent similarities* in the recommended literature, alternative visualization methods are required.

2. Relevance thresholds over which users have no control. If the predicted global relevance score is too low, an item will not be displayed in the list-based format, e.g., in a top-5 ranking. However, a paper at rank 16 might still be relevant to a reader if it contains certain features that the reader is keen to discover. For example, if a researcher seeks to find publications containing similar figures or citations as a given input file. Addressing this need requires a user-customizable interface.

We propose and evaluate a visualization approach that allows displaying the relatedness of documents and derives from the similarity of semantic features that commonly exist in scientific publications. The approach and visualization interface we describe is especially valuable for literature in the STEM fields (science, technology, engineering, and mathematics) since these disciplines contain high frequencies of text-independent features, such as mathematical formulae or charts. The presented interface supports users in the filtering and decision-making process to

JCDL '20, August 1–5, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7585-6/20/06...$15.00
https://doi.org/10.1145/3383583.3398599

discover and understand instances of both text-based and text-independent similarities within the recommended literature.

2 BACKGROUND

List-based results visualizations are the prevailing format for literature search and recommendation interfaces. Nonetheless, a range of *graph-based visualizations* has been proposed for recommending movies [7], TV shows [8], users in social networks [9], and talks to conference attendees [10]. However, research on *visualizations* designed to support the *academic literature recommendation use case* is sparse. Despite recommender systems research having emphasized the importance of transparency [11] and enhancing user control [12], these considerations are still lacking from today's systems for academic literature recommendation. We identify two related projects thus far.

Scienstein was a prototype of a hybrid research paper recommendation system employing a graph-based user interface (UI) [13]. The system combined four citation analysis algorithms, textual similarity assessments, and user ratings. The UI displayed papers as topically clustered icons in a graph (the higher the relevance score of a paper, the larger its icon). Users could filter the recommended papers, e.g., by impact factor, rating received, or publication dates. However, no other text-independent features beyond citations were considered.

HyPlag is a prototype of a hybrid plagiarism detection system [14] that considers the similarity of in-text citations, figures, formulae, and text to retrieve instances of potential plagiarism in academic documents. Its similarity assessment methods are highly relevant to this work. The system's UI consists of two column-based views for presenting results. A first view shows an input document on the left and abstract representations for all documents exhibiting similarities to the input document on the right. A second view shows a side-by-side comparison of the input document with a potential source document. Similar features are highlighted and interactively linked in both documents.

3 RecVis SYSTEM

This section describes the similarity assessment methods of the newly conceived RecVis system, the system's architecture, and the novel interface for exploring literature recommendations.

3.1 Citation-based Similarity Assessment

To quantify the citation-based similarity of documents, RecVis employs the co-citation proximity analysis (CPA) measure [3]. Co-citation measures derive the similarity of two documents from the frequency with which the two documents are cited together in other documents. CPA improves upon the traditional co-citation by weighting the co-citation frequency with the smallest distance between the in-text citations that refer to the two documents in question. We use the more fine-grained CPA measure over co-citation, or bibliographic coupling, since CPA has outperformed these approaches in prior evaluations [15].

3.2 Text-based Similarity Assessment

RecVis uses the scoring function "More Like This" (MLT) implemented in Elasticsearch[1] to determine the textual similarity of documents. MLT combines a tf-idf weighted term vector space model with a Boolean retrieval approach by using the top k terms with the highest tf-idf values in a document to form a disjunctive query for finding related documents. Thus, MLT considers articles as being more similar the more specific terms they share.

3.3 Mathematics-based Similarity Assessment

To analyze the similarity of mathematical expressions, RecVis uses three measures we developed in our prior research [16]. All three measures consider mathematical identifiers, since these features achieved the best retrieval performance of all presentational math features in our prior experiments. The *Histo* measure reflects the global, order-agnostic overlap of identifiers in two documents by quantifying the difference of the identifiers' frequency histograms. The *Longest Common Subsequence of Identifiers* (LCIS) is the number of identifiers that match in both documents in the same order but not necessarily in a contiguous block. Like Histo, the LCIS measure quantifies the global similarity of mathematical content in documents. Finally, the *Greedy Identifier Tiles* (GIT) measure reflects the number of identifiers in the query document that are part of identifier tiles with a minimum length of five. Identifier tiles are the individually longest blocks of shared identifiers in identical order that cannot be extended to the left or right without encountering a non-matching identifier. Greedy tiles are well-suited to find confined regions in articles that feature high mathematical similarity.

3.4 Image-based Similarity Assessment

RecVis integrates four analysis methods to find similar or semantically related figures, charts or images. We previously developed this combined retrieval approach for the plagiarism detection use case [6]. However, with adjusted thresholds, we see this approach being equally applicable to the literature recommendation use case. The first of four image-based analysis algorithms we use is *perceptual hashing (pHash)*, a reliable and well-established method for retrieving highly similar images. The second and third analysis methods perform *order-agnostic* and *positional character trigram matching* for text extracted from figures using OCR. The order-agnostic matching compares all trigrams extracted from the images. The positional matching only compares trigrams occurring in similar relative positions of a figure, e.g., allowing a precise comparison of axis labels. The fourth method, *ratio hashing*, can identify bar charts that depict similar data even if the scales used differ.

3.5 System Architecture and Document Collection

RecVis is implemented as a 3-tier web application. The *frontend presentation tier* provides the recommendation interface we describe in this paper. The frontend uses HTML, CSS, and

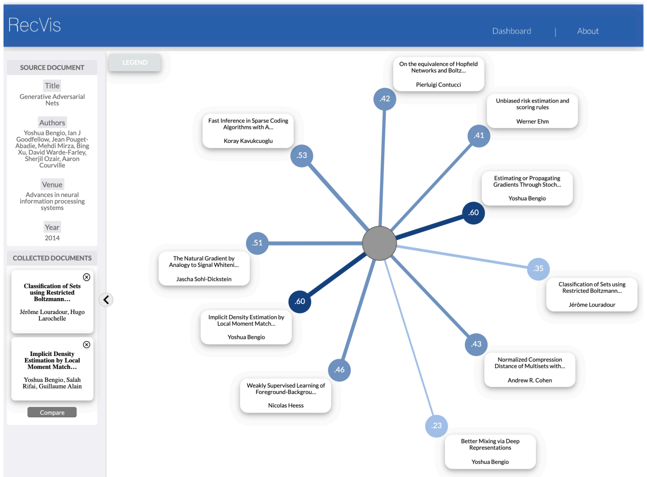
[1] https://elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html

JavaScript (Bootstrap for styling and D3.js for visualizations). It fetches result data from the backend and stores user credentials, user preferences, and the metadata for the recommendation exploration session. This tier uses a REST API realized using Node.js and Express.js, as well as a MongoDB database integrated via Mongoose.js. The existing HyPlag system [14] serves as the *backend tier* of the web application. The backend stores the document collection, performs the similarity computations and provides the recommendation results via the REST API to the frontend application tier. Since RecVis is connected to the HyPlag backend, it uses the PubMed Open Access Subset[2] and the NTCIR 11 Math Dataset [16] as the recommendation collections.

3.6 RecVis Recommendation Exploration Interface

Existing literature search and recommendation systems require researchers to examine the retrieved publications manually to discover the presence of any *feature-based similarity* they may find relevant. Our proposed supportive visualization process allows researchers to more quickly explore and compare academic literature with regard to the user-specified semantic features of interest. We define these features of interest to encompass similarities among the *figures* used, the *mathematical formulae* contained, or the literature *cited* – in addition to the presence of *textual similarity.*

For the recommendation exploration view, we conceived a visualization concept as shown in Figure 1. Users upload a seed document, for which they seek to receive recommendations, thus adhering to the query by example paradigm. The input document is displayed as the central node, around which the recommended literature is arranged in a force-directed graph layout.[3]



Title
Generative Adversarial Nets

Authors
Yoshua Bengio, Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville

Venue
Advances in neural information processing systems

Year
2014

Classification of Sets using Restricted Boltzmann...
Jérôme Louradour, Hugo Larochelle

Implicit Density Estimation by Local Moment Match...
Yoshua Bengio, Salah Rifai, Guillaume Alain

On the equivalence of Hopfield Networks and Boltz...
Pierluigi Contucci

Unbiased risk estimation and scoring rules
Werner Ehm

Fast Inference in Sparse Coding Algorithms with A...
Koray Kavukcuoglu

Estimating or Propagating Gradients Through Stoch...
Yoshua Bengio

The Natural Gradient by Analogy to Signal Whiteni...
Jascha Sohl-Dickstein

Classification of Sets using Restricted Boltzmann...
Jérôme Louradour

Implicit Density Estimation by Local Moment Match...
Yoshua Bengio

Normalized Compression Distance of Multisets with...
Andrew R. Cohen

Weakly Supervised Learning of Foreground-Backgrou...
Nicolas Heess

Better Mixing via Deep Representations
Yoshua Bengio

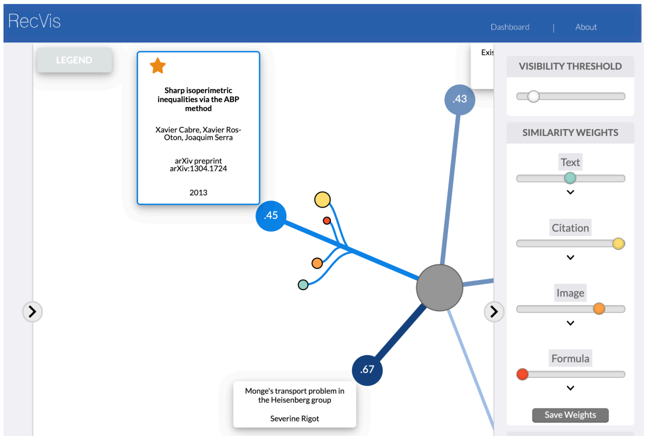
Figure 1: Interactive recommendation exploration view

Sliders in a right-hand panel allow users to determine their individual weighting preference for text, citation, figure, or formula-based similarities. A global relevance threshold slider additionally enables users to quickly filter for recommendations with the highest overall similarity score from zero to one. Nodes shaded in darker hues of blue and with thicker connecting lines indicate higher global relevance scores than lightly tinted nodes connected by thinner lines.

Changing the weights for any of the semantic features using the sliders (shown in Figure 2) result in an instantaneous update of the graph-based visualization. The responsiveness for the initial visualization of all recommendations for a seed document is 1-3 seconds, depending on the number of nodes. With the global sensitivity threshold set to the lowest point by the user, the interface displays up to 20 publications for exploration.

3.7 RecVis Feature Visualization

The detailed feature exploration view shown in Figure 2 displays 'feature nodes', which upon selecting a recommendation, extend from the edges. These nodes represent the presence of the individual similarity features (text, citations, figures, and formulae) for the inspected recommended publication. The aim of the interface is to support users in identifying the type of feature similarity and thus judging the relevance of the recommended literature more quickly.



Sharp isoperimetric inequalities via the ABP method
Xavier Cabre, Xavier Ros-Oton, Joaquim Serra
arXiv preprint arXiv:1304.1724
2013

Monge's transport problem in the Heisenberg group
Severine Rigot

Figure 2: Visualization depicting expanded feature nodes

The nodes are color-coded according to feature type and sized according to their relative presence in the publication to better assist the user in a quick discovery of which semantic features are most prevalent. Based on which features the user would like to explore in a subsequent detailed inspection view, the user can now save the most relevant recommendations in a 'collected documents' panel (shown on the left-hand side in Figure 1).

4 EVALUATION

To evaluate the perceived satisfaction and usability of RecVis' graph-based recommendation exploration interface, we conducted a study with STEM-academics at the Ph.D. and Postdoc level (N = 12). The average age of participants was 29, and 75% said they were 'very familiar' with using academic literature search or recommendation systems in their daily routines. Specifically, we examined (1) whether the graph-based visualization allows for efficient exploration of the recommended literature, and (2) whether the user-controllable similarity thresholds are intuitive

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3] Explore the RecVis interface for yourself at: https://purl.org/recvis (login ID: user@recvis.com, password: user)

for supporting the discovery of different feature-based instances of similarity.

We assigned participants with a series of tasks supported by the RecVis interface to assess successful task completion rates and identify areas of improvement. Participant's perceived cognitive workload for tasks was measured using NASA-TLX [17]. Finally, we assessed user's perceived satisfaction with individual design choices for recommendation interaction on a 5-point Likert scale.

Task completion & workload. We assigned participants with ten tasks, of which six tasks related to the recommendation exploration view (Figure 1) and four related to the detailed feature comparison view (Figure 2). Task completion success rate over all participants and tasks was 93%. The success rate was slightly higher (.96) for the six tasks relating to the graph-based exploration overview than for the detailed feature node examination view (.90). Participants 'struggled' most with Task 2.1: *Identify the recommended article that has the highest overall similarity with the source document and determine which feature(s) contribute the most to overall similarity.* Since this task required two steps, it was more complex to solve using the interface. However, even for this task, completion rate was .83 averaged over the 12 participants.

Given that this information need cannot be answered using existing recommendation exploration interfaces for academic literature, we were satisfied with the performance. Regarding participant's subjective workload for the tasks, mental demand on average was rated as 'low' (-1.9) (scaling NASA-TLX to a 7-point scale from -3 to 3). Performance, i.e. perceived success in accomplishing tasks, was rated at 2.8 (very high) and effort, i.e. difficulty to accomplish tasks was rated at -2.4 (very low). Frustration level was also rated as 'very low' (on average -2.4). These results are encouraging, since low cognitive complexity is crucial for the acceptance of new interfaces.

User-perceived satisfaction. Satisfaction with individual design choices was rated on a 5-point scale (-2 to 2) from 'strongly disagree' to 'strongly agree.' Responses to nine interface design decisions confirmed no critical interface design issues. Yet, we are currently making improvements to the feature node depiction as a result of the feedback collected. Due to space restrictions, we have made all survey questions, results, and charts available on GitHub[4].

In summary, the study showed that a graph-based visualization was perceived as an enjoyable and efficient format to explore and filter literature recommendations. Furthermore, the user-controllable thresholds for filtering recommendation sets were found to be intuitive and gave users a sense of control over the recommendation experience. This paper presents a first study of the RecVis concept. Subsequent evaluations will make use of standardized surveys, e.g. SUS or PSSUQ to enable comparability.

In the future, we will expand upon this recommendation visualization concept to also generate nested recommendation graphs for any of the currently displayed recommendations. We are now working on a detailed comparison view, which allows researchers to further inspect the instances of similarities in the literature in a subsequent detailed inspection view. We have made our system openly available on GitHub[5] and invite the scientific community to add semantic similarity measures or to customize the system to their own specialized information retrieval tasks.

5 CONCLUSION

RecVis demonstrates an alternative literature recommendation and visualization concept to help researchers identify specific features of interest within recommended sets of literature. An evaluation of our visualization concept showed good usability and demonstrated that the force-directed graph layout adds value for users when narrowing down recommendation results. Applying personalized discovery preferences and feature sensitivity values increased users' feelings of control. Our presented approach and interface allow a quick identification of academic publications that fulfill specific information needs, which existing recommendation interfaces cannot support. By conceiving more user-customizable recommendation interfaces, such as we introduce with the RecVis prototype, we hope that in the future, researchers will be better supported in their specialized academic literature search and discovery needs.

ACKNOWLEDGMENTS

This research was supported by the Carl-Zeiss Foundation and the German Research Foundation (DFG) Grant No.: GI 1259/1.

REFERENCES

[1] R. Johnson, A. Watkinson, and M. Mabe, "The STM report: An overview of scientific and scholarly journal publishing," 2018.

[2] A. E. Jinha, "Article 50 million: an estimate of the number of scholarly articles in existence," *Learned Publishing*, vol. 23, no. 3, pp. 258–263, 2010.

[3] B. Gipp and J. Beel, "Citation Proximity Analysis (CPA) - A New Approach for Identifying Related Work Based on Co-Citation Analysis," in *Proc. of the 12th Intl. Conf. on Scientometrics and Informetrics (ISSI'09)*, 2009, vol. 2.

[4] M. Schubotz *et al.*, "Semantification of Identifiers in Mathematics for Better Math Information Retrieval," in *Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2016.

[5] O. Chum, J. Philbin, and A. Zisserman, "Near Duplicate Image Detection: min-Hash and tf-idf Weighting," in *BMVC*, 2008.

[6] N. Meuschke, C. Gondek, D. Seebacher, C. Breitinger, D. Keim, and B. Gipp, "An Adaptive Image-based Plagiarism Detection Approach," in *Proc. of the ACM/IEEE Joint Conf. on Digital Libraries*, 2018.

[7] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer, "PeerChooser: visual interactive recommendation," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2008, pp. 1085–1088.

[8] E. Gansner, Y. Hu, S. Kobourov, and C. Volinsky, "Putting Recommendations on the Map: Visualizing Clusters and Relations," in *Proc. of the Third ACM Conf. on Recommender Systems*, 2009, pp. 345–348.

[9] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer, "Smallworlds: visualizing social recommendations," in *Computer Graphics Forum*, 2010, vol. 29, no. 3, pp. 833–842.

[10] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval, "Visualizing Recommendations to Support Exploration, Transparency and Controllability," in *Proc. of the 2013 Intl. Conf. on Intelligent User Interfaces*, 2013, pp. 351–362.

[11] R. R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *CHI*, 2002.

[12] D. Jannach, M. Jugovac, and I. Nunes, "Explanations and User Control in Recommender Systems," in *Proc. of the 23rd Intl. Workshop on Personalization and Recommendation on the Web and Beyond*, 2019, p. 31.

[4] https://github.com/ag-gipp/recvis-frontend/tree/master/study

[5] https://github.com/ag-gipp/hyplag-recvis-frontend

[13] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A research paper recommender system," *Proc. Intl. Conf. Emerg. Trends Comput.*, 2009.

[14] N. Meuschke, V. Stange, M. Schubotz, and B. Gipp, "HyPlag: A Hybrid Approach to Academic Plagiarism Detection," in *Proc. of the Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, 2018.

[15] M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp, "Evaluating link-based recommendations for Wikipedia," in *Proc. of the ACM/IEEE Joint Conf. on Digital Libraries (JCDL)*, 2016, pp. 191--200.

[16] N. Meuschke, V. Stange, M. Schubotz, M. Kramer, and B. Gipp, "Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations," in *Proc. of the ACM/IEEE Joint Conf. on Digital Libraries (JCDL)*, 2019.

[17] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183.

How to cite this paper:

C. Breitinger, B. Kolcu, M. Meuschke, N. Meuschke, B. Gipp "Supporting the Exploration of Semantic Features in Academic Literature using Graph-based Visualizations", in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2020. DOI: 10.1145/3383583.3398599

BibTex:

```
@InProceedings{Breitinger2020,
  title = {Supporting the {Exploration} of {Semantic} {Features} in {Academic}
{Literature} using {Graph}-based {Visualizations}},
  booktitle = {Proceedings of the {ACM}/{IEEE} {Joint} {Conference} on {Digital}
{Libraries} ({JCDL})},
  author = {Breitinger, Corinna and Kolcu, Birkan and Meuschke, Monique and
Meuschke, Norman and Gipp, Bela},
  year = {2020},
  month = {Aug.},
  location  = {Virtual Event, China},
  topic = {rec},
  doi = {10.1145/3383583.3398599},
  }
```

Related Publications: www.gipp.com/pub