

Similar Papers Recommendation for Research Comparisons

Vladyslav Nechakhin¹, Jennifer D'Souza²

¹L3S Research Center, Leibniz University Hannover, Hannover, Germany

²Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract

This paper outlines the core features and implementation details of the Similar Papers Recommendation service, which aims to expedite the creation and expansion of research paper comparisons within the Open Research Knowledge Graph (ORKG). By leveraging Semantic Scholar's Recommendations API and Elastic Search, the service provides a list of similar papers for a given comparison, considering both paper titles, abstracts, and their property values. The use of Semantic Scholar's Recommendations API allows the service to capitalize on machine learning techniques and semantic embeddings to generate relevant and tailored recommendations. The effectiveness of the service is demonstrated through the evaluation results, highlighting its potential as a valuable resource for the research community within the ORKG platform.

1. Introduction

The Open Research Knowledge Graph (ORKG) is an innovative research infrastructure based on the principles of open science, open data and open source methodologies. By facilitating the seamless interaction between human researchers and machine agents, the ORKG advances the landscape of research exploration and assistance. A core feature of the ORKG is its ability to create comparisons, condensed overviews that showcase the state-of-the-art for specific research questions, enabling efficient organization and analysis of contributions from different scientific papers. [2] Comparisons within the ORKG provide researchers with a comprehensive tabular view that allows them to compare and filter information based on various properties. By synthesizing insights from multiple papers within the same research field, comparisons offer an efficient and insightful means of understanding the latest developments. Whether it is calculating virological estimates, exploring material solubility parameters, or evaluating algorithm performances in computer science, comparisons facilitate the consolidation of literature from the same domain into a concise and informative overview. [1]

This paper presents the similar papers recommendation service implemented within the ORKG. The purpose of this service is to augment the capabilities of the ORKG by offering researchers a list of similar papers relevant to their comparisons. By streamlining the process of creating and expanding comparisons, this service enables researchers to gain deeper insights


SEMANTICS 2023 EU: 19th International Conference on Semantic Systems, September 20-22, 2023, Leipzig, Germany

✉ vladyslav.nechakhin@l3s.de (V. Nechakhin); jennifer.dsouza@tib.eu (J. D'Souza)

🆔 0000-0003-0146-1207 (V. Nechakhin); 0000-0002-6616-9509 (J. D'Souza)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

into the state-of-the-art, identify emerging trends, and establish connections across related scientific disciplines. The decision to display similar papers within the ORKG is driven by several significant advantages. Firstly, the service expedites the comparison creation process, optimizing the organization and presentation of research findings. Manual searches are replaced with a refined recommendation engine, significantly enhancing researchers' workflow. Secondly, the ability to discover newly published papers in their respective research fields provides researchers with a competitive edge, ensuring they remain up-to-date with the latest advancements. Additionally, the service simplifies the task of literature review by providing researchers with a curated list of relevant papers and promoting a more efficient exploration of existing knowledge.

2. Related Work

2.1. Recommendation Services' Content Types

Recommendation systems have become integral to various digital platforms, enhancing user experiences by providing personalized content suggestions. These systems cater to diverse content types, serving users with relevant recommendations based on their preferences and behavior. Popular examples of recommendation engines include YouTube [3], which suggests videos based on user viewing history, Amazon [4], which recommends products based on past purchases and browsing patterns, and Spotify [5], which offers personalized music recommendations based on users' listening habits and music preferences.

Among the various content types handled by recommendation systems, scholarly papers present a unique and distinct category. Unlike generic text data found on web pages or social media, research papers are meticulously crafted, formal documents containing scholarly contributions and insights. They adhere to specific formats and structures, reflecting the rigor of academic research. As a specialized content type, research papers demand a different approach to recommendation due to their emphasis on academic rigor and the critical nature of their content.

Traditional recommendation engines, developed for generic content types, encounter challenges when applied to scholarly papers. The reliance on matching keywords as the primary criteria for similarity may lead to inadequate results in the context of scholarly research. The mere presence of shared words does not guarantee that two papers are related or belong to the same research field. Moreover, scholarly papers often employ domain-specific terminology and context, making conventional keyword matching less effective in capturing semantic relatedness.

Given the unique characteristics of scholarly papers and the limitations of default recommendation engines, there is a compelling need for a custom recommendation engine tailored specifically to the requirements of scholarly content within the Open Research Knowledge Graph (ORKG). Such a custom engine should be designed to address the challenges of recommending research papers effectively and accurately. By considering domain-specific knowledge and semantic understanding, a custom recommendation engine can provide researchers with precise and relevant recommendations, empowering them in their pursuit of knowledge exploration and state-of-the-art comparisons within the ORKG.

2.2. Recommendation Methodologies

Recommendation engines have emerged as powerful tools for enhancing user experiences by delivering personalized content suggestions. These systems play a crucial role in diverse digital platforms, ranging from search engines like Google, which offer personalized search results [10], to streaming services like Netflix, which recommend movies and TV shows based on users' viewing history and preferences [11]. The foundation of recommendation engines lies in various methodologies, including collaborative filtering, content-based filtering, and hybrid approaches [12]. Collaborative filtering leverages user-item interactions to identify similar users and recommend items based on their preferences. Content-based filtering, on the other hand, analyzes the attributes of items and user profiles to make recommendations. Hybrid approaches combine collaborative and content-based filtering to exploit their complementary strengths and provide more accurate and relevant recommendations.

Graph-based recommendation systems have gained popularity due to their ability to capture complex relationships between users and items. These systems represent users, items, and their interactions as nodes in a recommendation graph. Edges in the graph signify interactions or connections between users and items [13]. The graph structure allows for the utilization of graph-based algorithms such as Personalized PageRank [14] and random walks on the recommendation graph to infer user preferences and item relevance. By propagating preferences through the graph, personalized recommendations can be generated for individual users. Graph-based recommendation systems [15] excel at handling sparsity in the user-item interaction matrix and offer recommendations by exploiting indirect connections between users and items. However, they may face challenges in scalability when dealing with large-scale datasets and may struggle to capture the fine-grained semantics of complex content.

Semantic embedding-based recommendation systems have gained traction, particularly with advancements in deep learning. These approaches represent users and items as low-dimensional vectors in a semantic space, where the proximity of vectors reflects similarity. By leveraging neural network-based techniques such as Word2Vec and Doc2Vec, semantic embeddings can be learned from textual content, capturing the latent semantics of items and users [16]. Semantic embeddings are effective in capturing complex relationships between items, even when explicit user-item interactions are sparse. They excel at capturing fine-grained semantics and are inherently scalable to large datasets. Furthermore, semantic embedding-based methods can be adapted to capture domain-specific knowledge by training on domain-specific corpora, which is crucial in the context of scholarly papers and research content [17]. However, they may face challenges when dealing with cold-start scenarios for new items or users with limited historical data.

Comparing graph-based and semantic embedding-based recommendation methodologies requires a nuanced understanding of their respective strengths and limitations. Graph-based approaches are effective at capturing user-item interactions and delivering recommendations by exploring the connectivity of the recommendation graph. However, they may suffer from scalability issues and struggle to capture the fine-grained semantics of complex content, which is vital in the context of scholarly papers. On the other hand, semantic embedding-based methods excel at capturing fine-grained semantics, providing accurate and scalable recommendations even in sparse data scenarios. They are more suitable for the task of recommending similar papers

for creating comparisons, where semantic relationships play a significant role in identifying relevant research content.

Considering the unique characteristics of scholarly papers and the specific requirements of research paper comparisons within the ORKG, embedding-based recommendation methodologies appear more suitable. The nature of scholarly content demands precise and semantically meaningful recommendations. By employing semantic embedding techniques, the Similar Papers Recommendation service can effectively identify and suggest research papers that contribute to relevant research questions, facilitating the creation of comprehensive and condensed overviews of the state-of-the-art in specific domains.

3. ORKG Similar Papers Recommendation Service

The primary goal of the similar papers recommendation service in the ORKG is to provide researchers with a curated list of similar papers for a given comparison within the ORKG. This list aids researchers in condensing information from diverse scientific papers into comprehensive overviews, allowing them to gain insights into the state-of-the-art for specific research questions. The service recognizes that research papers are their own content type, necessitating a custom approach to recommendation to accommodate the unique characteristics of scholarly content. The process of generating recommendations involves a twofold approach, combining information from paper titles and their property values. Leveraging this information enables the service to identify research papers that contribute significantly to the relevant research field and align with the specific focus of the comparison. The workflow of the service is shown on the figure 1.

To initiate the recommendation process, the ORKG similar papers recommendation service queries Semantic Scholar's Recommendations API. Semantic Scholar [9] is widely recognized for its expertise in academic paper analysis and provides a rich source of scholarly knowledge. By utilizing the Recommendations API, the service retrieves a preliminary list of research papers that are likely to be relevant to the given comparison. However, to enhance the relevance and comprehensiveness of the recommendations, the service employs Elastic Search, a powerful search engine, to further refine the initial results. Elastic Search takes into account both contribution names and property values, offering a comprehensive analysis of research papers within the ORKG. This integration ensures that the service can effectively handle large-scale datasets of scholarly papers, providing users with precise and relevant recommendations.

The dataset used for re-ranking consists of papers from Semantic Scholar, including paper titles and abstracts, allowing Elastic Search to consider a wide range of textual information during the refinement process. By employing Elastic Search, the Similar Papers Recommendation service returns a final list of papers that are most relevant to the entire comparison. This final set of recommendations of relevant research papers is displayed on the ORKG platform, simplifying the process of creating or expanding the comparisons. The implementation of the Similar Papers Recommendation service within the ORKG enables researchers to navigate scholarly content in an efficient and insightful manner. By combining information from paper titles and property values, querying Semantic Scholar's Recommendations API, and refining the results using Elastic Search, the service provides access for researchers to an up-to-date list of

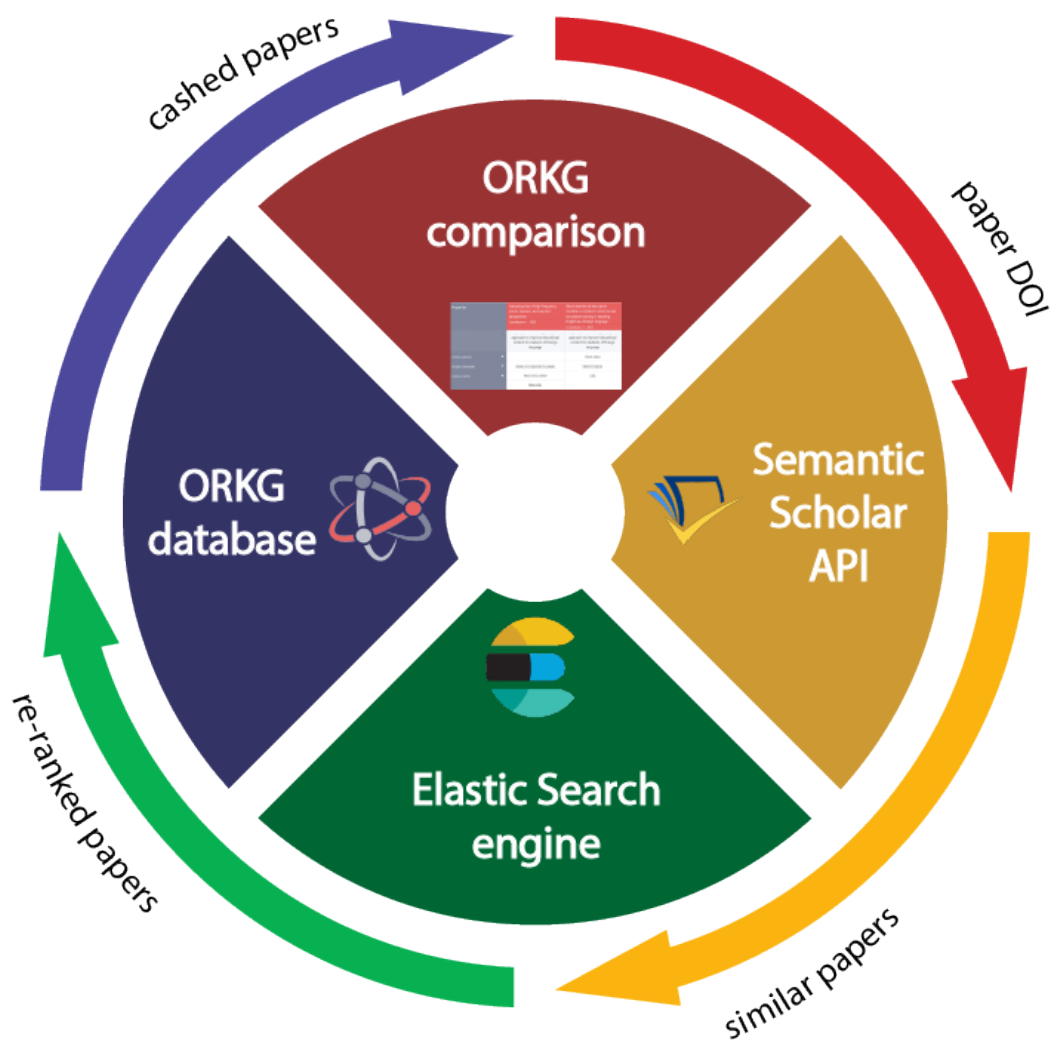


Figure 1: The workflow of similar paper recommendation service

relevant research papers that significantly improves their ability to create new contribution by comparing the results of the publications. In light of the overview description of the ORKG similar papers recommendation service above, next we elicit all the main components of the service in a detailed manner.

3.1. Workflow of the ORKG Similar Papers Recommendation Service

The initial approach envisioned creating an internal elastic search index encompassing the vast universe of scholarly publications from various research fields, spanning across Publishers, Meta-

repositories, and Pre-print servers. The process involved in-house fetching and update scripts to continuously populate and maintain the index. The service would then issue custom queries based on ORKG Comparison content and retrieve the most relevant papers with associated meta-information, enabling subsequent ranking based on date or relevance.

However, this approach presented several limitations that hindered its feasibility. The primary challenge was estimating the size of the internal paper database in advance, given the constantly expanding landscape of scholarly literature. This uncertainty rendered the workflow time-consuming and cumbersome, demanding substantial maintenance efforts for fetch and update scripts. Moreover, the presence of duplicated papers arriving from various sources necessitated the implementation of a dedicated deduplication system. Additionally, the approach relied heavily on establishing tie-ups with publishers and other entities, which might be impractical for comprehensive coverage.

To overcome the limitations of the first approach, the ORKG Similar Papers Recommendation service adopted an alternative strategy leveraging Semantic Scholar's API. Semantic Scholar already implements essential functionalities such as fetch and update scripts maintenance, deduplication of papers, and collaboration with publishers. This approach alleviates the need for the ORKG platform to create an extensive internal repository from scratch.

In the revised approach, the service utilizes an in-house script to query Semantic Scholar for each ORKG Comparison. It fetches up to 500 relevant papers, retrieving essential metadata such as paper titles and abstracts. This substantially reduces the size of the paper index, making it more manageable and focused on the most relevant publications. The paper index is then updated on a weekly or bi-weekly basis, ensuring the integration of the latest research into the recommendation service.

The adoption of the second approach offers notable advantages over the first, primarily stemming from the reduction of the universe of all possible papers to a collection of over 100,000 most relevant papers, aligned with the keywords present in ORKG Comparisons. By relying on Semantic Scholar's API, the service benefits from their established infrastructure and expertise in academic paper analysis, enabling the generation of targeted and high-quality recommendations.

While the Semantic Scholar API provides an initial index of relevant papers, the direct application of its keyword-based relevance ranking may not fully meet the specific needs of completing ORKG Comparisons. To address this, the Similar Papers Recommendation service re-ranks the results using Elastic Search. This step is essential to fine-tune the recommendations and ensure they align closely with the unique requirements of ORKG Comparisons. The use of Elastic Search, a powerful and versatile search engine, facilitates a more accurate and tailored ranking process, resulting in more usable and valuable results for researchers utilizing the ORKG platform. The integration of Elastic Search complements the semantic-based recommendations of Semantic Scholar, combining the strengths of both systems to optimize the Similar Papers Recommendation service within the ORKG.

3.2. Semantic Scholar Paper Recommendation API

The Semantic Scholar team at the Allen Institute for AI has released the Recommendations API [8], a publicly available service that generates recommendations for recently published

papers or preprints based on a learned model of researchers' topical interests. This API [7] enables third-party applications to suggest new and relevant papers to their users, delivering the latest scientific research. The Recommendations API operates in two modes: Single Paper Recommendations and List-Based Recommendations.

Single Paper Recommendations: In the Single Paper Recommendations mode, applications can request recommendations for a specific paper by providing its unique identifier. The API retrieves a list of top recommendations based on the given paper's content and the user's interests.

List-Based Recommendations: The List-Based Recommendations mode allows applications to pass a list of positive or negative paper examples to the API. This mode is useful for research servers and publishers to serve more relevant research to users. By leveraging the positive examples provided by users, the Recommendations API generates new recommendations that align with their interests.

Semantic Scholar's Recommendations API utilizes machine learning techniques to analyze the content of papers and understand the topical interests of researchers. It uses the contextual and semantic information within papers to identify related works and provide tailored recommendations. Semantic Scholar leverages semantic embeddings to represent research papers. These embeddings are trained using neural networks on large-scale academic datasets. The embeddings capture the semantic meaning of papers, enabling the identification of related works. By integrating Semantic Scholar's Recommendations API into the similar papers recommendation service in the ORKG, researchers can benefit from a curated list of papers that are similar to a given comparison. This recommendation service enhances the ability of researchers to discover relevant contributions and compare different works within their field of study.

3.3. Re-Ranking the Results with Elastic Search

In the implementation of the similar papers recommendation service in the ORKG, Elastic Search plays a vital role in re-ranking the initial recommendations obtained from Semantic Scholar. Elastic Search is a versatile and efficient search and analytics engine that allows for flexible querying and ranking of large datasets. In the context of the similar papers recommendation service, Elastic Search is applied to ORKG to re-rank the results and improve the relevance and quality of the recommended papers.

The re-ranking process in Elastic Search takes into account both the contribution names and property values provided in the ORKG. The service leverages a dataset of papers from Semantic Scholar, which includes paper titles and abstracts, to match and rank the recommended papers based on their relevance to the given comparison as a whole.

The application of Elastic Search in the similar papers recommendation service follows these key steps:

1. Indexing: The recommended papers, along with their associated contribution names and property values, are indexed in Elastic Search. This indexing process involves organizing and structuring the data in a way that allows for efficient retrieval and searching. The textual

content, such as paper titles and abstracts, is tokenized and analyzed to create an inverted index, enabling fast and accurate querying.

2. **Querying:** When a user submits a comparison, Elastic Search constructs a query based on the contribution names and property values provided. The query is designed to capture the essence of the comparison and retrieve papers that align with the overall context. Elastic Search utilizes its query language and search capabilities to match the query against the indexed papers.

3. **Scoring and Relevance:** Elastic Search assigns a relevance score to each recommended paper based on how well it matches the query. The scoring algorithm takes into account factors such as term frequency, inverse document frequency, and field length normalization. Papers that have higher relevance scores are considered more relevant to the given comparison.

4. **Sorting and Filtering:** Elastic Search allows the results to be sorted based on the relevance scores assigned to each paper. The sorting can be performed in ascending or descending order, ensuring that the most relevant papers appear at the top of the list. Additionally, Elastic Search provides filtering options to refine the results based on specific criteria, such as publication year or author, enabling users to narrow down their search.

The integration of Elastic Search into the ORKG's similar papers recommendation service enables the system to consider both the contribution names and property values provided by users. By leveraging a dataset of papers from Semantic Scholar, Elastic Search enhances the relevance and accuracy of the recommended papers by considering their titles, abstracts, and other relevant metadata. Through the re-ranking process, Elastic Search provides a refined list of papers that are most relevant to the given comparison as a whole. By combining the contribution names, property values, and the dataset from Semantic Scholar, Elastic Search ensures that the recommended papers align closely with the user's research interests and requirements.

3.4. Results Evaluation

In order to assess the effectiveness of the similar papers recommendation service in the ORKG, we compare the re-ranked results obtained from Elastic Search with the un-ranked results from Semantic Scholar. The experimental dataset [6] was created by including 30 random ORKG comparisons, each of them is provided with 50 similar papers recommended by Semantic Scholar and 50 papers recommended by Elastic Search, including 10 most relevant papers that were manually labeled.

The evaluation is conducted based on precision and recall metrics at different values of k , where k represents the number of recommended papers. The precision ($P@k$) measures the proportion of relevant papers among the top k recommendations, while recall ($R@k$) quantifies the proportion of relevant papers retrieved out of all the relevant papers in the dataset.

The average precision and recall values for the Semantic Scholar results are shown on the figure 2. And the average precision and recall values for the Elastic Search results are shown on the figure 3.

From the evaluation results, it is evident that the re-ranked results obtained from Elastic Search outperform the un-ranked results from Semantic Scholar in terms of both precision and recall.

The precision values for Elastic Search consistently exceed those of Semantic Scholar across

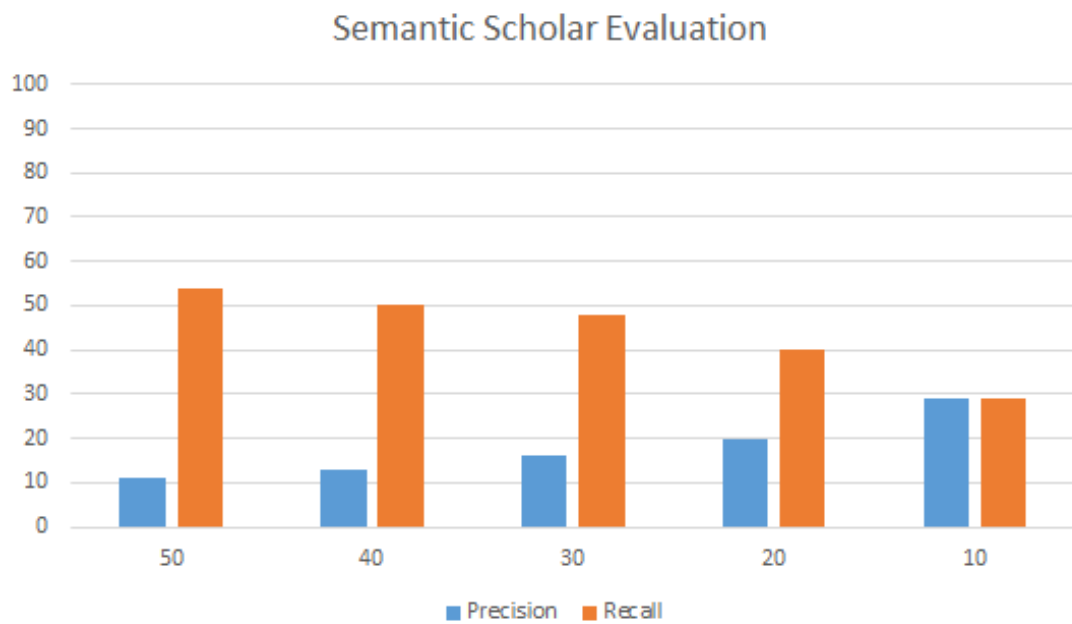


Figure 2: Evaluation of Semantic scholar results

Table 1

Semantic Scholar and Elastic Search Results.

Paper number (k)	SS Precision (P@k)	SS Recall (R@k)	ES Precision (P@k)	ES Recall (R@k)
50	0.11	0.54	0.20	1.00
40	0.13	0.50	0.25	0.98
30	0.16	0.48	0.32	0.97
20	0.20	0.40	0.46	0.92
10	0.29	0.29	0.63	0.63

all values of k . This indicates that a higher proportion of the recommended papers from Elastic Search are relevant to the given comparison, ensuring greater accuracy in the recommendations.

Similarly, the recall values for Elastic Search are significantly higher than those of Semantic Scholar. This means that Elastic Search is able to retrieve a larger proportion of relevant papers from the dataset, ensuring that fewer relevant papers are missed.

The improvements in both precision and recall metrics demonstrate the effectiveness of the re-ranking process performed by Elastic Search. By considering the contribution names, property values, and utilizing its advanced search capabilities, Elastic Search is able to provide more relevant and comprehensive recommendations compared to the initial results obtained from Semantic Scholar.

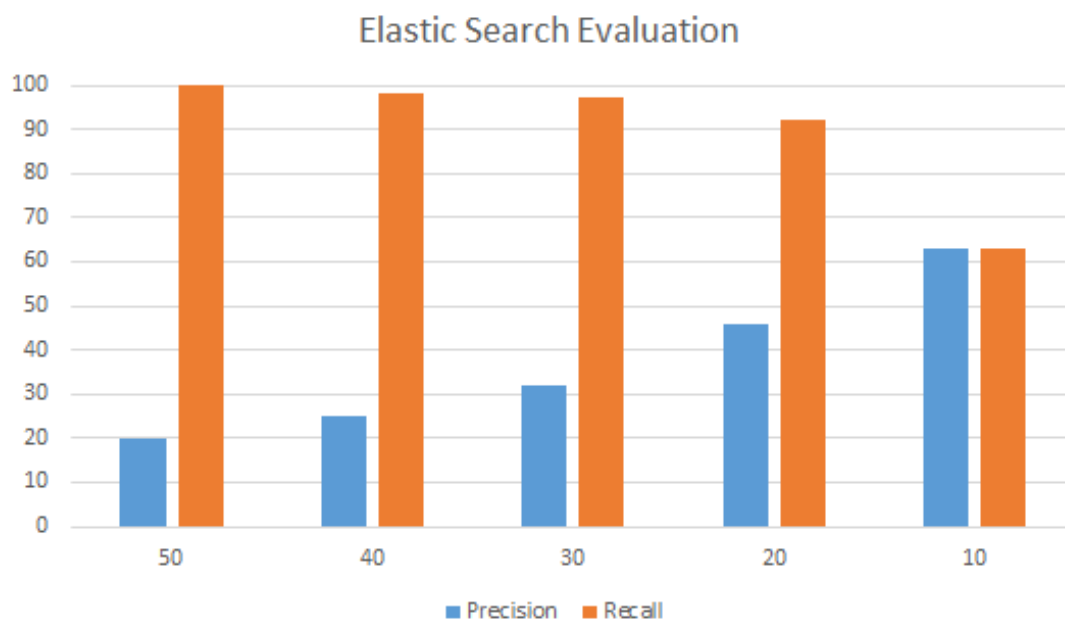


Figure 3: Evaluation of re-ranked Elastic Search results

4. Conclusion

In this paper, we explored the concept of a similar papers recommendation service for the ORKG. We compared the traditional graph-based recommendation engine methodology with the contemporary semantic embedding-based recommendation methodology. We examined the methodologies used by Semantic Scholar to generate recommendations and how Elastic Search re-ranks the results to enhance their relevance. The comparison revealed that semantic embedding-based methods offer advantages such as scalability, flexibility, and the ability to capture subtle relationships between documents. On the other hand, graph-based recommendation engines excel in leveraging the relationships between entities in a graph but may struggle with scalability. The evaluation of the results provided insights into the performance of the recommendation service. By considering metrics such as precision, recall, and F1-score, we assessed the quality of the recommendations and identified areas for improvement.

In conclusion, the implementation of a similar papers recommendation service in the ORKG can greatly enhance the usability and relevance of the platform. By leveraging the power of semantic embeddings and effective re-ranking algorithms, researchers can discover related works, explore new research avenues, and gain a comprehensive understanding of their research topics. However, further research and development are needed to optimize the recommendation methodologies and improve the overall performance of the service.

Future work could focus on exploring advanced techniques for semantic embeddings, incorporating user feedback to enhance recommendation relevance, and considering domain-specific

factors to tailor recommendations to the needs of researchers in specific fields. With continued advancements in recommendation systems, the ORKG can become an invaluable resource for the academic community, facilitating collaboration, knowledge dissemination, and the advancement of scientific research.

References

- [1] Open Research Knowledge Graph Overview, <https://orkg.org/about/1/Overview>. Last accessed 31 Jul 2023
- [2] Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M. and Auer, S., 2019, September. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In Proceedings of the 10th International Conference on Knowledge Capture (pp. 243-246).
- [3] Covington, P., Adams, J. and Sargin, E., 2016, September. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems (pp. 191-198).
- [4] Smith, B. and Linden, G., 2017. Two decades of recommender systems at Amazon. com. *Ieee internet computing*, 21(3), pp.12-18.
- [5] Madathil, M., 2017. Music recommendation system spotify-collaborative filtering. Reports in Computer Music. Aachen University, Germany.
- [6] Vladyslav Nechakhin, Jennifer D'Souza (2023). Dataset: ORKG Similar Papers Recommendation Service Evaluation Dataset. <https://doi.org/10.25835/qftvbgo4>
- [7] Semantic Scholar – Recommendations API, <https://api.semanticscholar.org/api-docs/recommendations>. Last accessed 31 Jul 2023
- [8] Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A. and Crawford, M., 2023. The semantic scholar open data platform. arXiv preprint arXiv:2301.10140.
- [9] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V. and Kinney, R., 2018. Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262.
- [10] Das, A.S., Datar, M., Garg, A. and Rajaram, S., 2007, May. Google news personalization: scalable online collaborative filtering. In Proceedings of the 16th international conference on World Wide Web (pp. 271-280).
- [11] Amatriain, X., 2013, August. Big & personal: data and models behind netflix recommendations. In Proceedings of the 2nd international workshop on big data, streams and heterogeneous source Mining: Algorithms, systems, programming models and applications (pp. 1-6)
- [12] Geetha, G., Safa, M., Fancy, C. and Saranya, D., 2018, April. A hybrid approach using collaborative filtering and content based filtering for recommender system. In Journal of Physics: Conference Series (Vol. 1000, p. 012101). IOP Publishing.
- [13] Yang, K. and Toni, L., 2018, November. Graph-based recommendation system. In 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 798-802). IEEE.

- [14] Bahmani, B., Chowdhury, A. and Goel, A., 2010. Fast incremental and personalized pagerank. arXiv preprint arXiv:1006.2880.
- [15] Yang, K. and Toni, L., 2018, November. Graph-based recommendation system. In 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (pp. 798-802). IEEE.
- [16] Church, K.W., 2017. Word2Vec. *Natural Language Engineering*, 23(1), pp.155-162.
- [17] Huang, X., Zhang, J., Li, D. and Li, P., 2019, January. Knowledge graph embedding based question answering. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 105-113).