ISSN 2007-9737

A Study on Content-based Reviewer Assignment in the Semantic Web and Computer Science Domains

Farid Bagheri[1], Davide Buscaldi[2], Diego Reforgiato-Recupero[1,*]

[1] University of Cagliari, , Department of Mathematics and Computer Science, Cagliari, Italy

[2] Sorbonne Paris Nord University, Laboratoire d'Informatique de Paris Nord, Paris, France

f.bagheri@studenti.unica.it, davide.buscaldi@lipn.univ-paris13.fr, diego.reforgiato@unica.it

Abstract. This paper underscores the pivotal role of high-quality paper reviews and their assignment to reviewers, delving into the intricate process of reviewer selection. Employing a comprehensive, multidisciplinary approach spanning computational science, information retrieval, and academic evaluation, our objective is to elevate the efficacy of the peer-review process. Our study involves a dataset in the Semantic Web and Computer Science domain, featuring 663 papers from 85 conferences and profiles of 524 reviewers. To assess the relevance of potential reviewers to scientific papers, we employ various similarity measures and representation strategies, including Jaccard similarity, dot product, and cosine similarity. Exploring different forms of representation, such as title-only, abstract-only, and a summary of the abstract generated with a Large Language Model-based tool, we utilize evaluation metrics like Mean Reciprocal Rank, Precision at k, and Mean Average Precision to validate the accuracy of reviewer recommendations. The culmination of our research offers valuable insights into effective reviewer selection strategies and optimal representation measures within the context of scientific paper evaluation. These findings contribute to the ongoing refinement of the peer-review process, enhancing its overall effectiveness.

Keywords. Reviewer assignment, semantic web, reviewer recommendation, large language models.

1 Introduction

In today's scientific research landscape, conferences have become a crucial pillar for the dissemination of new scientific knowledge and discoveries aimed at benefiting the community [18, 15]. These events provide an important opportunity for researchers to interact with each other and share their research findings, garner invaluable feedback from knowledgeable colleagues, and make connections within the scientific community.

However, the success of a scientific conference largely relies on the quality of reviews of submitted papers and their proper assignment to suitable reviewers. Assigning appropriate reviewers to papers is a critical process, and given its complexity it requires careful consideration.

A proper distribution of reviewers can improve the quality of reviews and ensure that each scientific contribution is properly evaluated. However, this task has proven to be an increasing challenge given the growing quantity of papers submitted to conferences and the diversity of expertise required for their review. The problem of suggesting reviewers for academic papers is commonly referred to as the reviewer assignment

problem (RAP) [1]. The initial attempt to tackle this problem was made by Dumais and Nielsen in [6], who approached the RAP as an information retrieval challenge. They employed the latent semantic indexing model to establish connections between reviewers and papers.

As the field advanced, others utilized more sophisticated models such as latent Dirichlet allocation and author-topic (AT) models [11]. They introduced the author-persona-topic model to enhance the representation of a reviewer's covered topics. These approaches primarily rely on semantic information.

In contrast, some researchers have explored word-based information to extract features from reviewers and papers. Peng and colleagues in [13] applied the term frequency-inverse document frequency (TF-IDF) to capture the statistical characteristics of reviewers and papers.

They integrated this method with the topic model, resulting in the time-aware and topic-based model. However, these approaches overlook the constraints of the RAP, including i) incomplete reviewer data and ii) potential interference from non manuscript-related papers in the reviewer data.

The first mentioned challenge is the incompleteness of reviewer data. Acquiring precise and up-to-date full-text papers from all reviewers is impractical due to the challenges associated with data collection and processing, compounded by the presence of multilingual data. Typically, only the titles and abstracts of reviewers' papers are used as reviewer data.

However, relying solely on incomplete reviewer data poses challenges in accurately and quantitatively representing the field or topic of the reviewer's expertise. The other challenge is related to the interference from non manuscript-related papers. When assigning reviewers, the focus has usually been on authors (reviewers) of papers highly similar to the underlying manuscript, without considering whether the author has a significant number of unrelated publications.

Conversely, in computing full-text similarity, documents with paragraphs closely matching the underlying manuscript information are examined, including those containing numerous dissimilar paragraphs.

Consequently, when assessing the overall similarity between reviewers' papers and the manuscript, a multitude of irrelevant papers may unduly diminish the perceived similarity.

To address the aforementioned challenges, in this paper, we aim to enhance the peer-review process of scientific paper evaluation, by exploring a multidisciplinary intersection of computational science, information retrieval, and academic evaluation, with the overarching goal of advancing the peer-review process in the context of scientific paper evaluation.

More in detail, we apply Natural Language Processing (NLP), Deep Learning, and statistical approaches to investigate emerging challenges and propose how to optimize the reviewer assignment process. The reader will note that our objective is not to furnish an exhaustive reviewer recommendation system.

Instead, our emphasis lies in refining the reviewer assignment process, notably by experimenting with diverse pairing metrics. In the endeavor to find a solution, several computational techniques were employed, ranging from numerical representation of articles to semantic similarity calculations. More in detail, the contributions we bring in this paper are:

– We have created a dataset on the Semantic Web and the Computer Science domain, composed of 663 papers from 85 conferences and the anonymized profiles of 524 reviewers in the form of the titles and summaries of their top 20 cited articles.
– We used various similarity measures (Jaccard similarity, dot product, and cosine similarity) to assess the relevance of potential reviewers to scientific papers.
– We considered various forms of representation of the papers and the reviewer profiles for matching: title-only, abstract, and a summary automatically extracted with a Large Language Model-based tool.
– To ensure the accuracy of the reviewer recommendations, we have used established evaluation metrics such as Mean Reciprocal

Rank (MRR), Precision at k (P@k), and Mean Average Precision (MAP).

– We have carried out experiments on the collected dataset and obtained some insights regarding the best representation strategies and measures and by considering a set of constraints that we have defined.

The remainder of this manuscript is organized as follows. Section 2 discusses related works about the task of reviewer recommendation. Section 3 describes the task we address in this paper. The dataset we have adopted is described in Section 4. Section 5 details the strategies we have adopted to assign a reviewer for a candidate paper.

Section 6 shows the performance evaluation we have carried out showing the different dimensions we have varied and the list of used metrics. The obtained results are illustrated in Section 7 together with a discussion and some remarks about them. Finally, Section 8 ends the paper with a summary, final considerations, and future directions where we are headed.

2 Related Work

The development of automated reviewer matching systems can be traced back to the pioneering work presented in [6] where a new automated assignment method sent reviewers more papers than they actually had to revise and then allowed them to choose the part of their review load. Following that, a multitude of studies has been done, concentrating on advanced review recommendation systems. To achieve an optimal assignment, three key points must be addressed:

Author Name Disambiguation, Expert Matching, and Expertise Representation. Disambiguating author names stands as a pivotal step in enhancing the accuracy of expert recommendations, ensuring precision and relevance by accurately identifying authors [16]. The challenge of ambiguity in author names arises from factors like incomplete or missing identifier metadata and interdisciplinary publications involving contributors from multiple institutions.

This task has been tackled using machine learning techniques [16, 7]. The primary objective of Expert Matching is to efficiently evaluate semantic correlations [1] between papers to review and reviewers' information. Semantic relatedness scores are computed by identifying keyword relations and leveraging collaborative intelligence.

The proximity between categories serves as an indicator of correlation, with shorter distances signifying stronger relationships. Numerous keywords are shared between submitted works and published articles, with similarity scores determined based on distance and depth.

The highest score within these pairs indicates the degree of semantic correlation, while the cumulative maximum scores assess the overall semantic association between a research submission and previously published scholarly works. Various studies employ the keyword pairing method for expert matching. For instance, Zhao et al. [22] utilized the Word Mover's Distance–Constructive Covering Algorithm to characterize reviewers by tags such as keywords and research interests.

The model utilizes Word Mover's Distance to measure the distance between submitted papers and reviewers. Framed as a classification problem, this task is tackled using the supervised learning method Constructive Covering Algorithm. The outcomes demonstrate an enhanced recommendation accuracy. Another work [5] addresses the same task by introducing the Sentence Pair Modeling-based Reviewer Assignment (SPM-RA) method.

The system automatically assigns reviewers to academic papers by leveraging supervisory information extracted from sentence pairs found in titles and abstracts. This process incorporates neural network models for training, and the resulting model is employed to predict the field distance between the reviewer and the manuscript.

Expertise Representation involves first representing the expertise of the reviewers and then comparing the subject field of the submitted paper with the current research field of the reviewer [14]. The study in [7] addresses this challenge by employing SVM-based multi-classifiers for automatic classification.

Expert pools are established based on research fields, utilizing SVM-based multiclassification and subject-specific keywords extracted from expert profiles. Liu and associates in [9] introduced an automatic reviewer recommendation system that integrates expertise, authority, and diversity. In their methodology, a graph is constructed based on potential reviewers and papers, combining information on expertise and authority.

The Random Walk with Restart model is applied to the graph, taking into account sparsity constraints and diversity. The results demonstrate that the model surpasses benchmark datasets in terms of expertise, authority, diversity, and similarity to human specialists' judgment.

Moreover, Maleszk et al. in [10] proposed a modular recommender system for reviewer recommendation, containing three modules: a keyword-based module, a social graph module, and a linguistic module. Instead of selecting a single best reviewer and then additional best-matched ones, the goal was to choose a reviewer and then form a diverse group of potential candidates which benefits larger groups of reviewers.

Furthermore, this system demonstrates flexibility in handling a wide range of topics. Moreover, for Expertise Representation, another study [20] introduces the modified Binary Butterfly Particle Swarm Optimization (MBBPSO) as a heuristic swarm intelligence optimization algorithm designed to address the reviewer assignment problem. Notably, MBBPSO represents an advancement over the original Particle Swarm Optimization (PSO) and Bare-Bones PSO (BBPSO) algorithms.

The primary objective of MBBPSO is to augment the population diversity of potential solutions by integrating a dynamic learning strategy. This strategic enhancement enables each particle to acquire knowledge from multiple channels, as opposed to a singular channel, thereby mitigating the risk of the algorithm converging to local optima. In [18], the Word and Semantic-based Iterative Model is introduced to address the limitations associated with incomplete reviewer data.

The proposed approach enhances the similarity calculation method based on normalized discounted cumulative gain and integrates improved language and topic models. The results indicate a 2.5% increase in recommendation accuracy. In another work, researchers conducted a comparative analysis of various evaluation metrics, such as precision at k and mean average precision, while also assessing the error rates associated with each of these metrics [3].

With respect to the state of the art, in our proposed paper, we have employed diverse metrics to glean insights into the optimal strategies and criteria for evaluating the relevance of potential reviewers for scientific papers. This analysis contributes to the broader understanding of effective reviewer selection methodologies in the context of scholarly research.

3 The Targeted Task

In this paper, given a corpus of papers submitted to a conference and a pool of reviewers characterized by their known expertise, the task is to methodically assign a fixed number of reviewers to each paper. This assignment should be executed in a manner ensuring that the reviewers possess expertise pertinent to the field of the paper and share similar research interests with the authors of the candidate paper.

The goal is to enhance the quality of the peer review process by guaranteeing that each paper undergoes evaluation by experts well-qualified to assess its content and who are likely to furnish constructive feedback to the authors. In this study, we defined constraints to ensure the integrity and validity of the reviewer assignment process within the context of scientific conferences.

Firstly, we enforced a constraint where each paper is reviewed by a specified number of reviewers, namely, three reviewers per paper in our case. This constraint ensures a balanced and comprehensive evaluation of each paper while avoiding overburdening reviewers with excessive assignments. Additionally, we imposed a constraint to limit the maximum number of articles assigned to each reviewer within a conference to three.

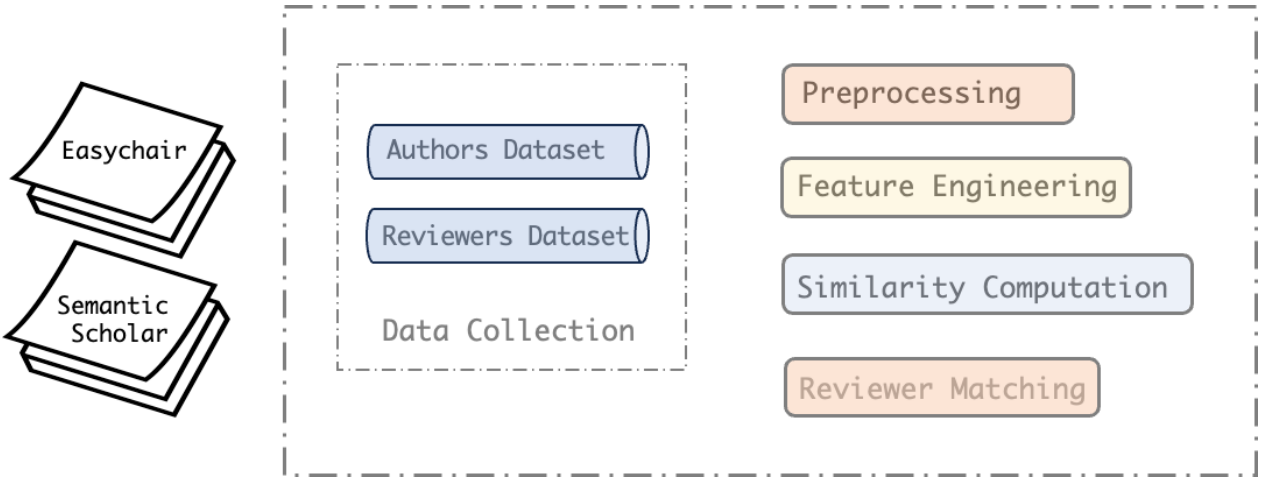

Fig. 1. Reviewer assignment pipeline

This constraint aims to distribute the workload evenly among reviewers and prevent any single reviewer from dominating the evaluation process. Finally, a reviewer cannot be assigned to a paper where one of the authors has previously collaborated on at least one publication with that reviewer. By adhering to these constraints, we strive to maintain fairness, objectivity, and efficiency in the peer review process, ultimately enhancing the quality and reliability of our evaluation outcomes. Three constraints that we have defined are the following:

– Firstly, each reviewer is limited in the number of papers they can review.
– Secondly, each paper must undergo review by a predetermined number of reviewers.
– Thirdly, a reviewer cannot be assigned to a paper where one of the authors has previously collaborated on at least one publication with that reviewer.

4 Dataset Creation

The dataset used for our study is composed of two subsets. The first one comprises information about authors' papers that have undergone a review process, sourced from the EasyChair website[1].

The second one contains the summaries of publications of potential reviewers and was extracted from the Semantic Scholar website[2]. The names of reviewers assigned to the papers and the list of each candidate reviewer have been anonymized. The anonymized dataset can be distributed by authors on request.

4.1 Authors' Dataset

As previously mentioned, this part of the dataset consists of abstracts and titles of papers submitted to conferences organized or co-organized by the authors of this paper or some of their colleagues. We collected 663 papers and their assigned reviewers from a set of 85 computer science conferences.

These conferences cover a range of topics including Semantic Web and Linked Data, Artificial Intelligence and Machine Learning, Database and Information Systems, Computational Linguistics, and Natural Language Processing, as well as Ontology and Knowledge Graphs.

The common theme across these conferences highlights the substantial focus on various aspects of computer science.

[1] www.easychair.org/

[2] www.semanticscholar.org/

Table 1. Reviewer matching using full abstracts only and with the employment of constraints

	Mean of similarities across all reviewers' articles			Maximum of similarities across all reviewers' articles		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	0.6721	0.6653	0.3229	**0.7054**	0.4018	0.3196
Dot Product Similarity	0.6095	0.6166	0.3085	0.6452	0.3817	0.3118
Jaccard Similarity	**0.6993**	**0.6997**	**0.3617**	0.4929	**0.5301**	**0.3216**

4.2 Reviewers' Dataset

The second dataset consists of articles retrieved from SemanticScholar, specifically those linked to the reviewers assigned by EasyChair for the 663 papers in the previous dataset. Data retrieval was performed using the official SemanticScholar Application Programming Interfaces (APIs)[3].

Following data collection, the information was structured and stored in a JSON file. On average, each paper in the Authors' Dataset was assigned to three reviewers. For every reviewer linked to the 663 extracted papers, we identified their top 20 cited articles from SemanticScholar, arranging them in descending order based on citation count. In total, information for 524 reviewers was successfully extracted.

The disparity in the number of reviewers (expected to be 3×663) can be traced back to the absence of certain reviewers from the initial dataset in SemanticScholar.

This inconsistency is primarily due to challenges in disambiguating homonyms and reviewers with similar names. Additionally, limitations arise from the insufficient availability of articles with clear abstracts and the existence of multiple profiles for a single author.

5 Reviewer Assignment Pipeline

In this section, we will elaborate on the details of our reviewer assignment pipeline. Each paper, whether from the Authors' Dataset or the Reviewers' Dataset, can be effectively represented as vectors.

[3] api.semanticscholar.org/api-docs/

For instance, if we designate the profile of an article as X, it assumes the form $X = (x_1, x_2, \ldots, x_n)$, where x_i indicates a term within the textual content of the article. In the context of this project, the paper is represented by its title or abstract or its summary.

The pipeline we have designed consists of five phases, as indicated in Figure 1: Data Collection, Preprocessing, Feature Engineering, Similarity Computation, and Reviewer Matching.

The Data Collection phase deals with the creation of the datasets detailed in Section 4. Subsequently, the Preprocessing step entails a sequence of actions, including tokenization, stop-word removal, stemming, and lemmatization, aimed at readying the data for the subsequent steps.

Following the Preprocessing step, a Feature Engineering phase is executed. This step involves taking the preprocessed text and generating TF-IDF vector representations based on the Bag of Words model.

The TF-IDF vectors generated in the previous step will be employed in the subsequent phase, Similarity Computation. Various similarity methods, including Jaccard similarity, Dot Product Similarity, and Cosine similarity, will be computed to assess the degree of relatedness between different documents.

By utilizing these similarity measures, we can accurately evaluate the proximity of content in reviewers' profiles to that of the submitted articles. This facilitates the precise matching of reviewers to manuscripts. Jaccard similarity is a technique used to assess the similarity between two sets of words.

Table 2. Reviewer matching using full abstracts only and without the employment of constraints

	Mean of similarities across all reviewers' articles			**Maximum of similarities across all reviewers' articles**		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	**0.5662**	**0.5591**	0.3318	**0.6958**	0.3995	**0.3259**
Dot Product Similarity	0.4619	0.4616	0.3441	0.6263	0.3737	0.3071
Jaccard Similarity	0.4810	0.4728	**0.3485**	0.5737	**0.5881**	0.3125

It measures the commonalities in words shared between two sets, with a higher similarity ratio indicating a greater number of shared words. The computation entails dividing the size of intersection of two vectors in a vector space by the size of their union [19]. The dot product, also known as the inner product, is a mathematical operation that gauges the alignment and magnitude relationship between two vectors.

It produces a scalar value by multiplying the corresponding components of vectors. The dot product indicates the extent to which one vector projects onto another, with positive values indicating alignment and negative values indicating opposite directions.

A dot product of zero signifies perpendicular vectors. The formulation incorporates Euclidean magnitudes and the cosine of the angle between vectors. In the context of our system, the Dot Product Similarity employs the dot product to quantitatively measure the similarity between vectors, yielding a score between 0 and 1. Higher values in this score indicate greater similarity [12].

Cosine similarity, also referred to as cosine distance, refers to the angle between vectors rather than the distance between points [17]. Particularly effective when dealing with substantial distances between vectors, it calculates the cosine of the angle formed by the vectors.

The resulting cosine similarity score ranges from 0 to 1, where a higher value indicates greater relatedness between the two vectors in the vector space. In the project's context, each text is represented as a vector, with words serving as dimensions and their frequency determining values.

Cosine similarity is then computed by assessing the angles between these vectors, offering a quantitative measure of text similarity. The final phase of the proposed pipeline is the Reviewer Matching. In this phase, the system leverages the similarities computed by similarity methods to match reviewers with authors' manuscripts.

Specifically, this phase involves comparing the similarity scores between the vector representations of reviewers' profiles and the submitted articles (titles or abstracts or their summaries). We employ two distinct methods to calculate the similarity score. These methods evaluate different aspects of a reviewer's expertise about the manuscript's topic.

The first method is the "Maximum of Similarities" across all articles written by a reviewer. This approach identifies reviewers who have punctual experience on the topic, meaning they have written at least one paper that strongly matches the candidate's paper. The maximum similarity score reflects the highest degree of relevance between any single article by the reviewer and the submitted manuscript.

The second method is the "Mean of Similarities" across all articles written by a reviewer. This method assesses the reviewer's overall expertise on the topic by considering how well all of their papers match the candidate's paper.

A higher mean similarity score indicates that the reviewer's publications consistently align with the subject matter of the manuscript, suggesting that the reviewer has a more comprehensive knowledge of the topic. Reviewers whose profiles exhibit a high degree of similarity with the manuscript's content are considered suitable

Table 3. Reviewer matching using titles only and without the employment of constraints

	Mean of similarities across all reviewers' articles			Maximum of similarities across all reviewers' articles		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	**0.4475**	**0.3049**	**0.2866**	0.3669	0.2273	0.1409
Dot Product Similarity	0.4087	0.2643	0.2293	**0.6111**	**0.3611**	**0.1809**
Jaccard Similarity	0.0984	0.0469	0.0675	0.07692	0.0384	0.0192

candidates for reviewing the candidate article. The system ranks these potential reviewers based on their similarity scores, facilitating the selection of the most appropriate reviewers for each manuscript. The goal of this phase is to ensure that reviewers with expertise and interests aligned with the subject matter of the article are chosen, thereby enhancing the quality and relevance of the review process.

6 Evaluation

In this section, we will first outline the metrics considered and then we will present and elucidate the experiments we have carried out.

6.1 Evaluation Metrics

The following paragraphs introduce and elaborate on some of the most commonly used metrics for this purpose. In the following, the definitions are tailored for documents, which in our scenario, are mapped to reviewers. We consider a reviewer relevant if they were among the original reviewers of the underlying paper.

6.1.1 MRR

Mean Reciprocal Rank (MRR)[4] is characterized as a position-based metric, wielding utility in the qualification of recommendation and information retrieval systems. When the retrieval system operates on reviewers, it yields a ranked list of reviewers.

[4]www.evidentlyai.com/ranking-metrics/mean-reciprocal-rank-mrr

Each of these reviewers is accompanied by a score denoting its relevance to the paper described by the query. The MRR computation entails summing the reciprocal ranks of the first relevant reviewer retrieved for each paper and subsequently dividing the sum by the total number of papers. Mathematically expressed, MRR is defined as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}. \tag{1}$$

Here, $|Q|$ represents the total number of papers, and rank_i indicates the rank of the first relevant reviewer for the i-th paper. The MRR metric yields values within the range of 0 to 1, with proximity to 1 indicative of superior performance.

This implies that, on average, the first relevant reviewer tends to secure a higher rank across all papers. Conversely, a lower score signifies that related reviewers are relegated to lower positions in the list [2].

MRR focuses on putting the most relevant reviewers first, making it especially beneficial for systems where the top-ranked reviewers are of great importance.

In contrast, Mean Reciprocal, focusing exclusively on the initial item in the list, disregards subsequent items, rendering it an unreliable gauge of overall performance as it neglects non-relevant reviewers. In our scenario, a high MRR corresponds to the ability of the chosen methodology to rank a pertinent reviewer in the top positions.

Table 4. Reviewer matching using titles only and with the employment of constraints

	Mean of similarities across all reviewers' articles			**Maximum of similarities across all reviewers' articles**		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	0.5888	0.3789	**0.2888**	0.4444	0.2546	0.2000
Dot Product Similarity	**0.7380**	**0.4367**	0.2803	**0.6904**	**0.3928**	**0.2346**
Jaccard Similarity	0.2562	0.1799	0.1303	0.1944	0.1157	0.1083

6.1.2 P@k

Precision at k, denoted as P@k, holds significant importance in the realm of retrieval and recommendation systems. The variable k indicates a positive integer, representing, in our scenario, the number of reviewers considered. In essence, $P@k$ gauges the accuracy of our approach by assessing its ability to present pertinent reviewers within the top k recommendations.

The calculation of $P@k$ involves determining the precision of relevant reviewers within the top k suggested ones [8]. The mathematical formulation for $P@k$ is expressed as follows:

$$P@k = \frac{1}{k} \sum_{i=1}^{k} \begin{cases} 1, & \text{if } r_i \in \mathcal{T}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, k denotes the number of reviewers we want to extract, r_i represents the reviewer ranked at position i by the model, and $\mathcal{T}$ indicates the set of relevant reviewers. The resulting value of $P@k$ falls between 0 and 1, with proximity to 1 indicating superior performance.

$P@k$ demonstrates resilience in scenarios characterized by noisy datasets, yielding robust results. However, it is crucial to acknowledge that $P@k$'s sensitivity to the number of relevant reviewers retrieved poses a limitation. Additionally, it does not account for the relevance of reviewers retrieved beyond the top k elements.

In the context of this paper, $P@k$ serves as a measure for assessing the alignment between candidate papers and a list of suggested reviewers. In the experiments we have carried out, we have considered $k = 3$, to reflect the average assignment of 3 reviewers to each paper.

Therefore, a methodology obtaining high scores in P@3 indicates the ability to place pertinent reviewers in the top 3 choices.

6.1.3 MAP

Mean Average Precision (MAP)[5] denotes the average precision value of our approach across a specified set of n papers. It is formally defined as follows [21]:

$$\text{MAP@k} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}@k_i. \tag{3}$$

Here, n is the number of papers, k is the chosen cutoff point, AP denotes the average precision of the ranking list of reviewers at k for the paper i. This metric is employed to assess the relevance of predicted reviewers and to ensure that the most pertinent ones are positioned at the top.

Specifically, MAP is articulated as the average of the average precisions calculated for each given ranking list of reviewers. The average precision for an individual reviewer is derived by summing the precisions at all ranks, inclusive of and leading up to the rank of the respective reviewer, and subsequently dividing this sum by the total number of correct reviewers.

Therefore, in our scenario, MAP is usually higher than P@3 because a pertinent reviewer may be ranked just outside the top 3 choices, and P@3 would not take it into account. For instance, for one paper, if the pertinent reviewers are ranked 1, 3, and 4, P@3 would yield $2/3 = 0.66$ and MAP would yield $(1 + 0.6666 + 0.75)/3 = 0.80$.

[5] www.evidentlyai.com/ranking-metrics/mean-average-precision-map

Table 5. Reviewer matching using the summary of the abstracts and without the employment of constraints

	Mean of similarities across all reviewers' articles			Maximum of similarities across all reviewers' articles		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	**0.4937**	**0.4870**	0.2750	0.2380	0.1369	0.0845
Dot Product Similarity	0.4611	0.4435	**0.3299**	**0.5476**	**0.3452**	**0.3202**
Jaccard Similarity	0.3074	0.3120	0.2634	0.3095	0.1964	0.1673

6.2 Experiments

Utilizing the datasets outlined in Section 4, we conducted a series of experiments aimed at assigning reviewers to authors of submitted papers. The gold standard for this assignment was derived from the effective reviewer assignments extracted from Easychair, as reflected in the Authors' dataset.

Hence, a reviewer automatically assigned to a paper utilizing our method is deemed relevant if they were among the original reviewers for the underlying paper. As already introduced in Section 5, to calculate text similarities between an author's representative text and a reviewer's representative text, we employed two methodologies: the mean of similarities and the maximum of similarities.

The mean of similarities across all reviewer articles averages the similarity scores between the candidate author's paper and all papers authored by the reviewer. This approach assumes that a reviewer whose body of work consistently aligns with the topics of the candidate paper is likely to possess a broad and deep expertise in the subject matter. Thus, a high mean similarity score suggests that the reviewer is a well-rounded expert in the field covered by the candidate author's paper, capable of providing insightful and comprehensive feedback.

On the other hand, the maximum of similarities across all reviewer articles focuses on the highest similarity score between the candidate author's paper and any single paper authored by the reviewer. This method identifies reviewer who may have specialized or punctual experience relevant to the candidate author's paper, even if their overall body of work does not uniformly align with it. A high maximum similarity score indicates that the reviewer has at least one publication that closely matches the topic of the candidate author's paper, suggesting that they can offer valuable, targeted insights based on specific expertise.

Then we used three different similarity measures: Jaccard similarity, Dot Product Similarity, and Cosine similarity. Moreover, as a third dimension, we have varied the reviewer profile matching that has been performed with and without constraints. As already mentioned in Section 3, we define constraints in terms of the maximum number of assigned reviewers and the allocation of papers to each reviewer.

In the extraction of Authors' information, we uniformly set this number to 3 for all the conferences under consideration. Additionally, a key constraint prohibits assigning a reviewer to a paper authored by individuals who have previously collaborated with the reviewer on publications.

The inclusion of these constraints is essential to uphold fairness and reliability in the review process. Finally, as a fourth dimension, for the Similarity Computation phase, we have considered the titles of papers, their full abstracts, and the abstracts' LLM-generated summary. We used the CATTS summarizer [4] to produce TLDR summaries, of an average length of 21 words.

The rationale behind these options was to evaluate which of the three (titles only, full abstracts of papers, or a summarization of the abstracts) provides the most valuable information for the reviewer assignment phase. In general, full abstracts encapsulate the core objectives, methodologies, results, and conclusions of a

Table 6. Reviewer matching using the summary of abstracts and with the employment of constraints

	Mean of similarities across all reviewers' articles			**Maximum of similarities across all reviewers' articles**		
	MRR	**MAP**	**P@3**	**MRR**	**MAP**	**P@3**
Cosine Similarity	**0.6500**	**0.6689**	0.2883	0.3974	0.2307	0.1871
Dot Product Similarity	0.5666	0.5871	**0.295**	**0.5833**	**0.3541**	**0.3125**
Jaccard Similarity	0.4722	0.4674	0.2238	0.2923	0.1217	0.0840

paper, presenting a comprehensive and detailed overview crucial for matching papers with the most suitable reviewers. Titles, on the other hand, offer only a brief glimpse into the underlying paper, providing a partial suggestion at best. By creating a summary of the abstracts, we aimed to assess the capability of Large Language Models (LLMs) in enhancing the reviewer assignment phase.

7 Results

In this section, we will show the results we have obtained according to the campaign of experiments illustrated in Section 6. Table 2 includes the results when considering the full abstracts only of the reviewer' and authors' papers and without the presence of constraints. The results for the same settings but with the introduction of the constraints are illustrated in Table 1.

Table 3 shows the same information when only the titles of papers are considered for the similarity measures and no constraints are considered. For the same settings and the inclusion of the constraints we obtain the results illustrated in Table 4. Finally, Table 5 shows the results when we consider a summary of the abstracts without using the constraints.

Table 6 depicts the results for the same configuration and considering the constraints. The title-only approach exhibits the largest discrepancy between Mean Reciprocal Rank (MRR) and other measures. This suggests that while relying solely on the title may identify one pertinent reviewer, it generally fails to adequately cover the requirement for at least three reviewers.

Mean scores consistently surpass maximum scores, indicating that reviewer assignments in our gold standard were made with consideration of the reviewers' overall expertise within the domain.

This underscores a tendency towards specialist assignments. The automatic summarization method yields results comparable to the original abstracts, suggesting a high level of semantic coherence between the two.

This indicates the effectiveness of the Language Model-based summarization approach. Considering constraints leads to improved results, which aligns with expectations, given the inherent limitation of having a finite number of reviewers available for paper evaluations.

The evaluation of different similarity measures (cosine, dot product, Jaccard) reveals varying effectiveness. Dot product performs well when maximizing similarity, which is logical given its lack of normalization.

Jaccard, particularly with abstracts and constraints, emerges as the preferred choice for mean similarity. However, cosine similarity generally yields better results across other metrics and text representations.

In summary, while the title-only approach falls short in meeting reviewer coverage requirements, considering reviewer profiling and employing appropriate similarity measures significantly enhances the efficacy of the review process.

Further exploration and refinement of these strategies are warranted for optimal reviewer assignment outcomes.

8 Conclusions

In this work, we presented an overview of the reviewer assignment problem with a comparative evaluation of some matching strategies to address this task on a novel dataset that we created from scratch, on the Semantic Web and Computer Science domains.

The matching strategies were evaluated according to some well-known retrieval measures. The matching between the candidate paper and the reviewers was carried out considering either the mean or the max of similarities between the candidate and the reviewer's publications. We also considered various representations of the contents: title only, abstract, or an automatically generated summary of the abstract.

Finally, we considered also two scenarios: one in which we do not set any constraint on the choice of the reviewer (that is, a reviewer may theoretically review all the candidate papers), and one in which some constraints reflect the usual matching process for conference organisation.

The results show that the title-only matching strategy falls short in meeting reviewer coverage requirements, and LLM-generated summaries are good representations of the semantic content of the works. As indicated in the previous section, the best results are obtained for the mean of similarities. Among the evaluated metrics, the highest MRR achieved stands at 0.7380, attained when considering constraints and titles.

Then, the optimal MAP reaches 0.6997 when analyzing full abstracts under constraints. Furthermore, the top P@3 score of 0.3485 emerges from the evaluation of full abstracts without constraints.

These findings underscore a notable preference within our gold standard dataset for reviewers possessing a nuanced understanding of the topic, as indicated by their mean similarity scores, over those who have merely published at least one paper on the subject, as indicated by maximum similarity scores.

Further investigation will be required to understand some discrepancies in the strategies' results depending on the scenario. In particular, we will investigate the implications of utilizing embedding generated by advanced large language models, such as SciBERT[6]. Additionally, we will consider leveraging established datasets, such as RevASIDE, to rigorously evaluate the efficacy of our approach[7].

References

1. **Aksoy, M., Yanik, S., Amasyali, M. F. (2023).** Reviewer assignment problem: A systematic review of the literature. Journal of Artificial Intelligence Research, Vol. 76, pp. 761–827. DOI: 10.1613/jair.1.14318.
2. **Ali, Z., Ullah, I., Khan, A., Ullah-Jan, A., Muhammad, K. (2021).** An overview and evaluation of citation recommendation models. Scientometrics, Vol. 126, pp. 4083–4119. DOI: 10.1007/s11192-021-03909-y.
3. **Buckley, C., Voorhees, E. M. (2000).** Evaluating evaluation measure stability. Proceedings of the 23rd annual international Conference on Research and Development in Information Retrieval, pp. 33–40. DOI: 10.1145/345508.345543.
4. **Cachola, I., Lo, K., Cohan, A., Weld, D. (2020).** TLDR: Extreme summarization of scientific documents. Findings of the Association for Computational Linguistics, Empirical Methods in Natural Language Processing, pp. 4766–4777. DOI: 10.18653/v1/2020.findings-emnlp.428.
5. **Duan, Z., Tan, S., Zhao, S., Wang, Q., Chen, J., Zhang, Y. (2019).** Reviewer assignment based on sentence pair modeling. Neurocomputing, Vol. 366, pp. 97–108. DOI: 10.1016/j.neucom.2019.06.074.
6. **Dumais, S. T., Nielsen, J. (1992).** Automating the assignment of submitted manuscripts to reviewers. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 233–244. DOI: 10.1145/133160.133205.

[6] github.com/allenai/scibert
[7] zenodo.org/records/4071874

7. **Im, Y., Song, G., Cho, M. (2023).** Perceiving conflict of interest experts recommendation system based on a machine learning approach. Applied Sciences, Vol. 13, No. 4, pp. 2214. DOI: 10.3390/app13042214.
8. **Khan, F., Rawajbeh, M. A., Ramasamy, L. K., Lim, S. (2023).** Context-aware and click session-based graph pattern mining with recommendations for smart EMS through AI. IEEE Access, Vol. 11, pp. 59854–59865. DOI: 10.1109/access.2023.3285552.
9. **Liu, X., Suel, T., Memon, N. (2014).** A robust model for paper reviewer assignment. Proceedings of the 8th ACM Conference on Recommender Systems, pp. 25–32. DOI: 10.1145/2645710.2645749.
10. **Maleszka, M., Maleszka, B., Król, D., Hernes, M., Martins, D. M. L., Homann, L., Vossen, G. (2020).** A modular diversity-based reviewer recommendation system. Intelligent Information and Database Systems: 12th Asian Conference, Asian Conference on Intelligent Information and Database Systems, pp. 550–561. DOI: 10.1007/978-981-15-3380-8_48.
11. **Mimno, D., McCallum, A. (2007).** Expertise modeling for matching papers with reviewers. Proceedings of the 13th ACM Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining, pp. 500–509. DOI: 10.1145/1281192.1281247.
12. **Nunes, I., Heddes, M., Vergés, P., Abraham, D., Veidenbaum, A., Nicolau, A., Givargis, T. (2023).** DotHash: Estimating set similarity metrics for link prediction and document deduplication. Proceedings of the 29th ACM Special Interest Group on Knowledge Discovery and Data Mining Conference on Knowledge Discovery and Data Mining, pp. 1758–1769. DOI: 10.1145/3580305.3599314.
13. **Peng, H., Hu, H., Wang, K., Wang, X. (2017).** Time-aware and topic-based reviewer assignment. Database Systems for Advanced Applications, pp. 145–157. DOI: 10.1007/978-3-319-55705-2_11.
14. **Peng, Q., Liu, H. (2022).** ExpertPLM: Pre-training expert representation for expert finding. Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing, pp. 1043–1052. DOI: 10.18653/v1/2022.findings-emnlp.74.
15. **Pradhan, T., Sahoo, S., Singh, U., Pal, S. (2021).** A proactive decision support system for reviewer recommendation in academia. Expert Systems With Applications, Vol. 169, pp. 114331. DOI: 10.1016/j.eswa.2020.114331.
16. **Shoaib, M., Daud, A., Amjad, T. (2020).** Author name disambiguation in bibliographic databases: A survey. DOI: 10.48550/ARXIV.2004.06391.
17. **Steck, H., Ekanadham, C., Kallus, N. (2024).** Is cosine-similarity of embeddings really about similarity? Companion Proceedings of the ACM Web Conference, pp. 887–890. DOI: 10.1145/3589335.3651526.
18. **Tan, S., Duan, Z., Zhao, S., Chen, J., Zhang, Y. (2021).** Improved reviewer assignment based on both word and semantic features. Information Retrieval Journal, Vol. 24, No. 3, pp. 175–204. DOI: 10.1007/s10791-021-09390-8.
19. **Vedavathi, N., Anil-Kumar, K. M. (2023).** E-learning course recommendation based on sentiment analysis using hybrid Elman similarity. Knowledge-Based Systems, Vol. 259, pp. 110086. DOI: 10.1016/j.knosys.2022.110086.
20. **Yang, C., Liu, T., Yi, W., Chen, X., Niu, B. (2020).** Identifying expertise through semantic modeling: A modified BBPSO algorithm for the reviewer assignment problem. Applied Soft Computing, Vol. 94, pp. 106483. DOI: 10.1016/j.asoc.2020.106483.
21. **Zhang, T., Zhang, Y., Xin, M., Liao, J., Xie, Q. (2023).** A light-weight network for

small insulator and defect detection using uav imaging based on improved YOLOv5. Sensors, Vol. 23, No. 11, pp. 5249. DOI: 10 .3390/s23115249.

22. **Zhao, S., Zhang, D., Duan, Z., Chen, J., Zhang, Y. P., Tang, J. (2018).** A novel classification method for paper-reviewer recommendation. Scientometrics, Vol. 115, No. 3, pp. 1293–1313. DOI: 10.1007/s11192 -018-2726-6.

Article received on 08/04/2024; accepted on 13/06/2024.
**Corresponding author is Diego Reforgiato-Recupero.*