Exploiting LLMs to Improve Recommendation Diversity

Basar Yilmaz[1], Ali Eren Cankaya[1], Ismail Sengor Altingovde[1], Pinar Karagoz[1] and Ismail Hakki Toroslu[1]

[1]*Middle East Technical University (METU), Computer Eng. Dept., Ankara, Turkey*

Abstract

To enhance the diversity of recommendation lists, we propose leveraging semantic knowledge and representations from pre-trained LLMs for both *applying* and *evaluating* diversification approaches. Specifically, we incorporate pairwise item similarities obtained from LLMs into various implicit diversification algorithms based on the greedy best-first and local search meta-heuristics. Furthermore, we also evaluate list diversity by utilizing these semantic similarities in a widely-used metric that computes the average dissimilarity among items in the list. As a key contribution, we propose a knowledge distillation approach to fine-tune smaller text-embedding models, thereby capturing the LLM's semantic understanding for pairwise similarity scores without invoking the large model at run-time. Our experiments in the domain of movie recommendation reveal that semantic similarities from both the original LLMs and the more efficient, fine-tuned models with distilled knowledge effectively enable diversification, yielding gains up to 16.78% in semantic diversity in return for a negligible (≤1%) drop in NDCG scores. We also show that semantic diversity evaluation correlates well with traditional genre-based evaluation, suggesting that semantic representations capture meaningful diversity aspects and can reduce attribute dependency in evaluating recommendation list diversity.

Keywords

Recommendation, semantic representations, implicit diversification

1. Introduction

Traditional recommender systems often prioritize relevance, potentially leading to monotonous suggestions and "filter bubbles" that limit user exposure and satisfaction [1, 2]. Therefore, enhancing recommendation list diversity is a critical research direction.

Diversity-enhancing strategies are typically categorized as pre-processing, in-processing, or post-processing [3, 4, 5]. This work focuses on post-processing techniques, which re-ranks recommendations to enhance diversity without altering the underlying model.

Recent post-processing studies often adapt strategies from search result diversification [6, 7, 5]. Maximal Marginal Relevance (MMR) [8] is frequently used [9, 10], while other effective search diversification algorithms (e.g., [2, 11]) are less explored in recommendations. Evaluating diversification is also challenging, typically relying on dissimilarity from predefined attributes (e.g., movie genres, book categories, etc.), which can be biased or incomplete [12].

Our paper makes three novel contributions towards enhancing diversity in recommendation lists:

- **We investigate using rich semantic knowledge and representations from general purpose LLMs for both applying diversification algorithms and evaluating their outcomes.** Specifically, we propose two methods for computing pairwise item similarities that are required by typical diversification approaches. First, we directly prompt general-purpose Large Language Models (LLMs) to leverage their internal *knowledge* of the items to be recommended (as our work essentially focuses on recommending cultural artifacts such as movies, books, songs, on which LLMs are likely to capture deep knowledge covering several diverse aspects of them) and to generate a score representing the semantic similarity of an item pair.

Proceedings of the First Workshop on Small and Efficient Large Language Models for Knowledge Extraction (SmaLLEXT) @ CIKM 2025

basar.yilmaz@ceng.metu.edu.tr (B. Yilmaz); cankaya.ali@metu.edu.tr (A. E. Cankaya); altingovde@ceng.metu.edu.tr (I. S. Altingovde); karagoz@ceng.metu.edu.tr (P. Karagoz); toroslu@ceng.metu.edu.tr (I. H. Toroslu)

- **We then propose a knowledge distillation approach to make this semantic understanding practical.** To avoid the time, compute, and financial cost of computing such similarities using LLMs as in the former method, we fine-tune text-embedding models with the aforementioned LLM-based similarity scores (for only a relatively small subset of items) to obtain dense vector *representations* (and subsequently, similarity) of items. The item similarities obtained via our methods are not only leveraged within diversification algorithms, but also employed in a widely used metric, Intra-List Distance (ILD), yielding a variant called ILD-sem, to evaluate diversification effectiveness [13]. Unlike prior work employing item embeddings learned by the recommendation algorithm itself based on a limited set of content and interaction features (e.g., [14, 9, 5, 15]); our focus is on the systematic use of general-purpose LLMs and fine-tuned text embedding models with distilled knowledge to obtain *semantic item similarities* for both driving and evaluating diversity in recommendation lists.
- **We conduct a comparative analysis of several implicit diversification methods for recommendation lists.** We evaluate algorithms based on greedy best-first search heuristic (i.e., MMR [8], SY [16], MSD [17], GMC [2]) and greedy local search heuristic (i.e., Swap [18], BSwap [18], GNE [2]). While their performance is known in search result diversification [2], their effectiveness for re-ranking recommendation lists has not been explored before.

The remainder of this paper is organized as follows: Section 2 covers background, Section 3 details our semantic similarity framework, Section 4 presents experimental results, and Section 5 concludes.

2. Background

This section provides background on the implicit diversification strategies and evaluation metrics used in this study.

2.1. Implicit Methods for Diversification

Diversification techniques, often adapted from search literature, include explicit (relying on external knowledge) and implicit (using item properties and similarities) approaches [2, 7]. As explicit user diversity preferences are typically unavailable in recommendation scenarios, this paper focuses on implicit diversification. The aim is to re-rank an initial candidate list (e.g., top-100 items) into a smaller, diversified final list (e.g., top-10) based on relevance and inter-item similarity. We adopt the notation and framework from [2], which studies implicit approaches for search result diversification, and extend it to recommendations.

Problem definition. Let $S = \{s_1, ..., s_n\}$ be the candidate set of n items returned by a recommendation algorithm for the user u. Let k be the desired size of the result set $R \subseteq S$, with $|R| = k$ and $k < n$. The relevance of item s_i to user u is $\delta_{rel}(u, s_i)$, and the diversity (dissimilarity) between items s_i and s_j is $\delta_{div}(s_i, s_j)$.

The overall objective function $F(u, S')$ for a candidate results set ($S' \subseteq S$) of size k, given user u and trade-off parameter $\lambda \in [0, 1]$, is defined in Eq. (1).

$$F(u, S') = (k-1)(1-\lambda) \cdot \delta_{\text{rel}}(u, S') + 2\lambda \cdot \delta_{\text{div}}(S') \tag{1}$$

The goal is to find R that maximizes Eq. (1) (i.e., $R = argmax(F, S')$), an NP-complete problem. Thus, approximation algorithms are used, often based on two meta-heuristics, briefly reviewed as follows:

Algorithms based on Greedy Local Search These methods begin with an initial list and iteratively improve it by swapping selected items with candidates to enhance overall relevance and diversity, as expressed in the objective function shown in Eq. (1).

- **Swap [18]:** Swap initializes R with the top-k relevant items. It then iterates through remaining items $s_s \in S \setminus R$ (ordered by relevance). For each s_s, it considers swapping it with each $s_j \in R$. If a swap ($\{R \setminus s_j\} \cup \{s_s\}$) increases the objective function F (Eq. (1)), the one yielding the largest increase in F is permanently applied to R. The process continues with the next s_s.
- **BSwap [18]:** This method is similar to Swap, but it allows swapping an item $s_d \in R$ with an item $s_s \in S \setminus R$, if it increases diversity in R and the corresponding drop in relevance, $\delta_{\text{rel}}(u, s_d) - \delta_{\text{rel}}(u, s_s)$, does not exceed a predefined threshold θ.
- **GNE (Greedy Randomized w/ Neighborhood Expansion) [2]:** GNE employs the GRASP [19] meta-heuristic over $i_{\max}$ iterations. Each iteration involves: (1) *Randomized Construction:* Incrementally building a solution R by randomly selecting items from candidate list including items with highest MMC scores (Eq. (4)), (2) *Local Search:* Refining R by iteratively swapping items with diverse neighbors if it improves the overall objective F (Eq. (1)). The best R across all iterations is chosen.

Algorithms based on Greedy Best-First Search: These methods iteratively build a list by selecting the best candidate balancing relevance and diversity.

- **MMR (Maximal Marginal Relevance) [8]:** MMR greedily builds the diversified list R. In each step, it selects the item s_i satisfying Eq. (2) from the set $S \setminus R$. The selection maximizes a λ-weighted linear combination of s_i's relevance to user u ($\delta_{\text{rel}}(u, s_i)$) and its novelty, where novelty is achieved by penalizing similarity (i.e., $\delta_{\text{sim}}(s_i, s_j) = 1 - \delta_{\text{div}}(s_i, s_j)$) to items s_j in the set R.

$$\text{MMR}(s_i) = \underset{s_i \in S \setminus R}{\text{argmax}} \left[\lambda \cdot \delta_{\text{rel}}(u, s_i) - (1 - \lambda) \cdot \max_{s_j \in R} \delta_{\text{sim}}(s_i, s_j)\right] \tag{2}$$

- **MSD (Max Sum Dispersion) [17]:** In each iteration, MSD selects the item pair (s_i, s_j) from $S \setminus R$ that maximizes the sum of relevance and diversity, as shown in Eq. 3.

$$\text{MSD}(s_i, s_j) = (1 - \lambda) \cdot [\delta_{\text{rel}}(u, s_i) + \delta_{\text{rel}}(u, s_j)] + 2 \cdot \lambda[1 - \delta_{\text{sim}}(s_i, s_j)] \tag{3}$$

- **SY [16]:** SY incrementally builds R by first adding the most relevant item. Subsequent candidates from $S \setminus R$ are added only if their similarity to all items already in R does not exceed a threshold θ. This continues until R reaches size k.
- **GMC (Greedy with Marginal Contribution) [2]:** GMC builds the result set R by selecting, at each step p, the item $s^* \in S \setminus R_{p-1}$ that maximizes Eq. (4). The latter MMC score captures an item's relevance, its diversity w.r.t already selected items, and its potential diversity w.r.t. items yet to be selected.

$$\text{MMC}(s_i) = (1 - \lambda) \cdot \delta_{\text{rel}}(s_i, u) + \frac{\lambda}{k - 1} \left(\sum_{s_j \in R_{p-1}} \delta_{\text{div}}(s_i, s_j) + \sum_{l=1}^{k-p} \delta'_{\text{div}}(s_i, s_l) \right) \tag{4}$$

2.2. Evaluation Metrics

Evaluating the quality of diversified recommendations presents unique challenges, due to the absence of ground truth indicating user's desired level of diversity [12]. As a result, evaluation metrics typically asses diversity by measuring the dissimilarity among the items within the final recommendation list R.

A widely adopted metric for this purpose is Intra-List Distance (ILD) [9, 13], which measures the average pairwise dissimilarity between all items in the recommendation list R, as shown in Eq.(5):

$$\text{ILD}(R) = \frac{1}{k \cdot (k - 1)/2} \sum_{(s_i, s_j) \in R} d(s_i, s_j) \tag{5}$$

where $d(s_i, s_j)$ represents the distance between s_i and s_j, i.e., $d(s_i, s_j) = \delta_{\text{div}}(s_i, s_j) = 1 - \delta_{\text{sim}}(s_i, s_j)$.

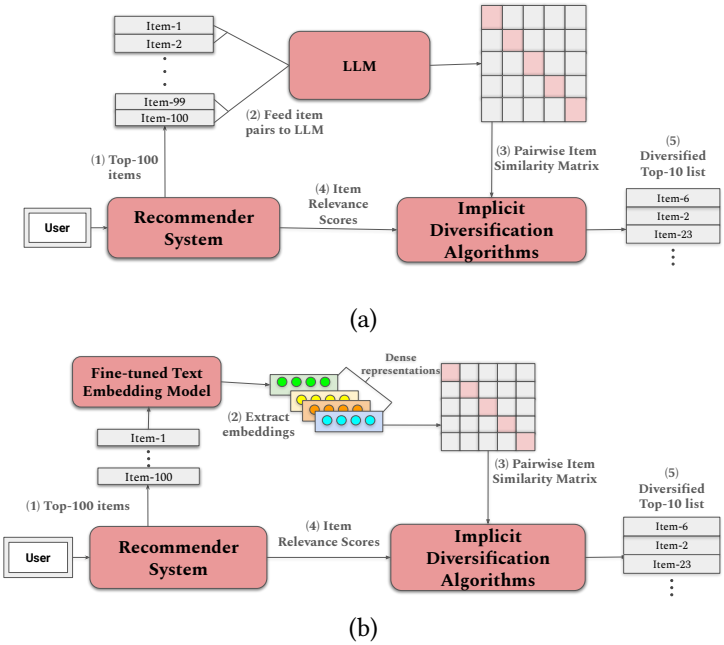

Figure 1: Pipeline for recommendation diversification using semantic pair similarities obtained via (a) LLMs, and (b) fine-tuned text embedding models.

3. Semantic Knowledge & Representations for Diversifying Recommendation Lists

In this work, we investigate the utility of leveraging rich semantic knowledge and representations derived from LLMs to enhance recommendation diversification. We explore the application of these representations in two key aspects of the diversification pipeline: firstly, as the basis for computing the pairwise item similarities required by the implicit diversification algorithms (cf., Section 2.1), and secondly, as the foundation for evaluating the diversity of the resulting lists using the ILD metric (cf. Section 2.2). To compute the pairwise similarities, we explore two distinct semantic pipelines:

Similarity via LLM Prompting For each pair of items (s_i, s_j) from the candidate list S, we construct a prompt containing the textual description (e.g., titles) of both items. The LLM is instructed to assess their similarity and return a numerical score. The pipeline is shown in Figure 1a.

Similarity via Fine-tuned Text Embeddings Models To avoid the runtime costs of invoking an LLM for computing pairwise similarities, we fine-tune a pre-trained text embeddings model using item similarities obtained from an LLM. During diversification, we obtain embedding vectors for each item's textual description (e.g., title) from the fine-tuned model with distilled knowledge, and then compute Cosine similarity for each pair of items (see Figure 1b).

In addition to incorporating semantic similarities into diversification methods, we employ these similarities to enable a direct evaluation of the resulting list's diversity based on semantic content, in oppose to using some pre-defined attributes, as in the literature.

To this end, we instantiate the well-known ILD metric with these semantic similarities and refer to as **ILD-sem** metric. ILD-sem follows the general formulation presented in Equation (5), but specifically calculates pairwise distance $d(s_i, s_j)$ using the semantic similarities obtained from LLMs or fine-tuned text embedding models, as described in this section. In both cases, $d(s_i, s_j) = 1 - \delta_{\text{sim}}(s_i, s_j)$.

4. Experiments

We formulate three Research Questions (RQs) that guide our experiments:

- **RQ1:** Are item similarities obtained from a general-purpose LLM promising to be used for the recommendation diversification task?
- **RQ2:** Are implicit diversification methods based on semantic similarities effective for the recommendation diversification task?
- **RQ3:** Are ILD scores based on semantic similarities well correlated with the scores based on a specific attribute (namely, movie genre in our setup)?

4.1. Setup

Datasets Experiments utilized two movie recommendation datasets: **Movielens-100k** [20] (genres, titles) and **Amazon Movies & TV** [21] (reviews, ratings, titles). Both were processed with a 5-core strategy, retaining users with at least five interactions.

Baseline Recommendation Generation For each user in the test splits of the datasets, we first generate an initial list of top-100 candidate items. This list serves as the input to the diversification algorithm. These initial lists were generated using the Collective Matrix Factorization (CMF) [22], a well-known cross-domain recommendation approach, available in the RecBole-CDR library [23].

Semantic Similarity Models As described in Section 3, we compute pairwise item similarities in two ways:

- **LLM-based Similarity:** We prompt an LLM, specifically `Phi4` [24] to directly score the similarity between pairs of movies (i.e., titles). The prompt explicitly instructs the LLM to utilize its internal knowledge of these movies, including their plot, cast, genre, etc.
- **Fine-tuned Text Embeddings:** We fine-tuned a text embedding model, `multilingual-e5-large` [25] [1] , using 790K pairwise similarity scores obtained from `Phi4` LLM for all movies in Movielens dataset. Then, using the fine-tuned model, we obtain dense vector representations for movie titles & the similarity of a movie pair is computed as the Cosine similarity of these vectors.

Figure 2: Movie similarity matrix by Phi4 LLM.

[1]Our fine-tuned embedding model: basaryilmaz/e5-multilingual-large-finetuned-movielens

Evaluation Metrics To evaluate the relevance of a recommendation list, we employ two well-known metrics, i.e., **NDCG@10** and **MRR@10**. To evaluate diversity, we employ the ILD metric (cf. Eq. (5)) with two alternative instantiations to compute $d(s_i, s_j)$: **ILD-sem@10** and **ILD-genre@10**.

4.2. Results

To address **RQ1** and illustrate the usefulness of LLM-based similarity scores, we first present a toy scenario. We manually selected a small set of Movies & TV items from the Amazon Reviews (2018) dataset [26] to broadly cover different genres and subjects, and computed their pairwise similarities (based on their titles) using the `Phi4` LLM (cf. Fig. 1a). The resulting similarity matrix, visualized as a heatmap in Figure 2, allows a qualitative assessment of whether a general-purpose LLM captures intuitive semantic relationships.

Figure 2 reveals that `Phi4` LLM returns high similarity score (0.75) for the crime movies *Godfather* and *Goodfellas*, indicating their strong association. It also correctly identifies pairs of Sci-Fi movies, finding high similarity between *Men in Black* and *Alien* (0.65), and *Men in Black* and *Star Trek* (0.60). The romantic comedy *You've Got Mail* is again correctly identified as dissimilar to all the other crime and Sci-Fi films. While preliminary in the current state, these anecdotal observations show the LLMs potential for the task at hand.

To answer **RQ2**, we begin with the case where pairwise similarities obtained from the `Phi4` LLM. This experiment was conducted solely on the Movielens dataset, as the Amazon dataset contains an order of magnitude more movies, leading to a significantly larger number of pairs for LLM scoring, and longer experiment times.

In Table 1, we present the performance of implicit diversification algorithms for the Movielens dataset. Note that, we report the diversification performance observed at the trade-off parameter setting where the drop in NDCG@10 is closest to 1% but does not exceed it, relative to the original (non-diversified) top-10 recommendation list. In the table, next to the columns for diversification metrics ILD-sem and ILD-genre, we also show relative gains (%) wrt. the scores of non-diversified baseline.

Table 1
Performance of `Phi4` LLM on Movielens

Algorithms	NDCG@10	MRR@10	ILD-sem@10	ILD-sem Gain (%)	ILD-genre @10	ILD-genre Gain (%)
Non-diversified	0.4690	0.3886	0.7489	N/A	0.7472	N/A
Greedy Best-First Search Algorithms						
MMR	0.4644	0.3813	0.7762	3.65	0.7697	3.01
SY	0.4661	0.3874	0.7539	0.67	0.7488	0.21
MSD	0.4646	0.3856	0.7617	1.71	0.7586	1.53
GMC	0.4645	0.3852	0.7886	5.30	0.7811	4.54
Greedy Local Search Algorithms						
Swap	0.4644	0.3872	**0.7937**	**5.98**	**0.7852**	**5.09**
BSwap	0.4651	0.3888	0.7762	3.65	0.7697	3.01
GNE	0.4647	0.3862	0.7912	5.65	0.7825	4.72

Table 1 reveals that, across the greedy best-first search methods, GMC achieved the best trade-off with relative gains of 5.30% and 4.54% for ILD-sem and ILD-genre metrics, respectively. MMR and MSD also performed competitively, offering moderate diversity improvements with very small loss in relevance scores. Among the greedy local search algorithms, Swap emerged as the most effective method. It achieved the highest ILD-sem@10 score (0.7937) and genre diversity score (0.7852), resulting in the relative gains of 5.98% and 5.09%, respectively. GNE closely followed Swap in performance, showing strong diversity improvement across both metrics.

Table 2
Performance of fine-tuned `e5-large` on Movielens

Algorithms	NDCG@10	MRR@10	ILD-sem@10	ILD-sem Gain (%)	ILD-genre @10	ILD-genre Gain (%)
Non-diversified	0.4690	0.3886	0.3767	N/A	0.7472	N/A
Greedy Best-First Search Algorithms						
MMR	0.4644	0.3823	0.4115	9.24	0.7711	3.20
SY	0.4647	0.3863	0.3846	2.10	0.7539	0.90
MSD	0.4645	0.3856	0.3983	5.73	0.7629	2.10
GMC	0.4645	0.3842	0.4170	10.70	0.7738	3.56
Greedy Local Search Algorithms						
Swap	0.4645	0.3869	**0.4399**	**16.78**	**0.7878**	**5.43**
BSwap	0.4649	0.3875	0.4096	8.73	0.7660	2.52
GNE	0.4644	0.3858	0.4296	14.04	0.7812	4.55

Overall, our findings in Table 1 demonstrate that approaches based on greedy local search are more effective in diversifying recommendation lists; and Swap outperforms all its competitors in terms of ILD-sem and ILD-genre metrics.

Next, we evaluate the performance of `e5-large` model, which is fine-tuned with similarity scores from `Phi4` obtained for the Movielens dataset. In Tables 2 and 3, we report the performance over the Movielens and Amazon datasets, respectively.

Table 2 reveals that, compared to the previously presented results, the relative performance trends of the algorithms are generally preserved. The Swap method remains dominant, delivering the highest relative gains in diversity, namely, 16.78% and 5.43%, in terms of ILD-sem and ILD-genre metrics, respectively. GNE is again the second-best performer (with a relative gain of 14.04% for ILD-sem and 4.55% for ILD-genre). GMC achieves substantial gains (namely, 10.70% for ILD-sem and 3.56% for ILD-genre) but as before, it is inferior to Swap and GNE.

Table 3
Performance of fine-tuned `e5-large` on Amazon

Algorithms	NDCG@10	MRR@10	ILD-sem@10	ILD-sem Gain (%)
Non-diversified	0.4274	0.3589	0.3781	N/A
Greedy Best-First Search Algorithms				
MMR	0.4231	0.3552	0.3941	4.23
SY	0.4232	0.3566	0.3831	1.32
MSD	0.4231	0.3354	0.3953	4.55
GMC	0.4231	0.3560	0.4008	6.00
Greedy Local Search Algorithms				
Swap	0.4231	0.3574	**0.4102**	**8.49**
BSwap	0.4231	0.3571	0.3939	4.18
GNE	0.4232	0.3568	0.4068	7.59

In Table 3, we report the results for the Amazon dataset in the same setup. The findings are similar to previous ones: GMC and Swap yield the highest diversity gains among the methods based on best-first search and local search heuristics, respectively; and Swap again outperforms GMC by far.

Our findings over Movielens dataset indicate that fine-tuning `e5-large` model for the movie similarity prediction task is effective, as the trends in Tables 1 and 2 overlap. Furthermore, proposed knowledge distillation approach from a large LLM (i.e., `Phi4` with 14B parameters) to a smaller text-embedding

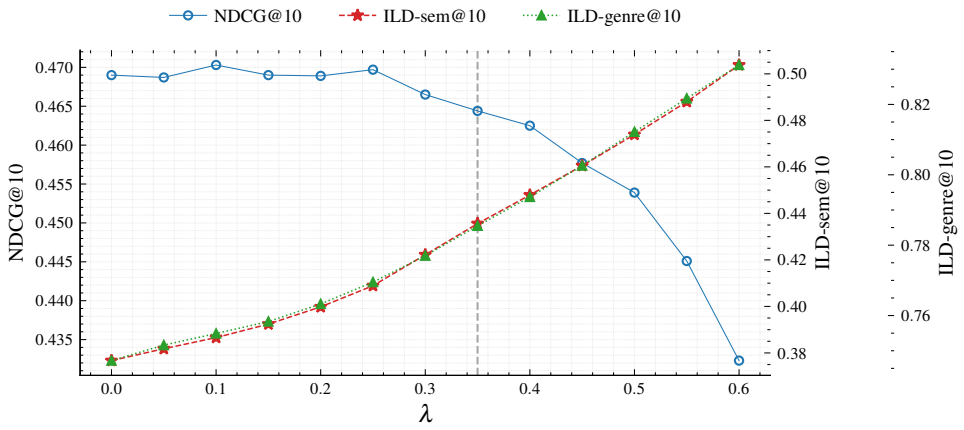

Figure 3: Performance of Swap wrt. λ

model (i.e., `e5-large` with 0.5B parameters) allows experiments over the Amazon dataset, which was otherwise infeasible due to the time cost of obtaining similarities from an LLM (as Amazon dataset includes 190M unique pairs), even for the in-lab experiments using `Phi4`. Although 98% of the movie pairs included in the test users's lists in the Amazon dataset have not been seen during the fine-tuning process (which involved only 714K pairs from the Movielens dataset); the distilled model yields trends (cf. Table 3) that align with those on Movielens.

To address **RQ3** and examine the correlation between embedding-based and genre-based ILD metrics, we compare their scores and relative gains in Tables1 and 2 reported over the Movielens dataset. We observe that while the absolute ILD values may differ, both metrics generally rank the diversification methods in a similar order based on the relative gains in diversity. Methods excelling in terms of ILD-sem (Swap, GNE, GMC) also tend to be the-best performing ones in terms of ILD-genre.

The parameter sweep visualization for the Swap algorithm in Figure 3 further illustrates the correlation between diversity metrics. In this plot, the ILD-sem curve and the ILD-genre curve almost overlap and they exhibit consistent upward trends as the trade-off parameter λ increases.

5. Conclusion

To obtain pairwise item similarities to be used in recommendation list diversification, we proposed leveraging LLMs (via prompting) and distilling their knowledge into fine-tuned text embedding models. Our experiments showed that using such semantic similarities, effective diversification can often be achieved with relatively simple methods, such as Swap, applied at the post-processing stage. We also incorporated these similarities into ILD metric, and validated a strong correlation between semantic ILD scores and traditional genre-based ILD scores, proposing semantic ILD as a generalizable alternative when attribute-based metrics are unavailable or insufficient.

Acknowledgments

We would like to express our gratitude to Erdeniz Aydogdu, Yagmur Tufekcioglu, Okan Dalan and Eren Colak for their valuable support and insightful feedback. This work is partially funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant no. 5230039 and METU under the grant no. ADEP-312-2024-11484.

Declaration on Generative AI

Generative AI tools were employed solely for polishing manuscript text, including grammar checks, word correction, and similar stylistic improvements, as permitted by the conference policies. No part of the scientific content, research methodology, or analytical results was generated by AI.

References

[1] Q. M. Areeb, M. Nadeem, S. S. Sohail, R. Imam, F. Doctor, Y. Himeur, A. Hussain, A. Amira, Filter bubbles in recommender systems: Fact or fallacy - A systematic review, WIREs Data. Mining. Knowl. Discov. 13 (2023). doi:10.1002/WIDM.1512.

[2] M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. T. Jr., V. J. Tsotras, On query result diversification, in: S. Abiteboul, K. Böhm, C. Koch, K. Tan (Eds.), Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany, IEEE Computer Society, 2011, pp. 1163–1174. doi:10.1109/ICDE.2011.5767846.

[3] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM Trans. Interact. Intell. Syst. 7 (2017) 2:1–2:42. doi:10.1145/2926720.

[4] Z. Fu, X. Niu, M. L. Maher, Deep learning models for serendipity recommendations: A survey and new perspectives, ACM Comput. Surv. 56 (2024) 19:1–19:26. doi:10.1145/3605145.

[5] H. Wu, Y. Zhang, C. Ma, F. Lyu, B. He, B. Mitra, X. Liu, Result diversification in search and recommendation: A survey, IEEE Trans. Knowl. Data Eng. 36 (2024) 5354–5373. doi:10.1109/TKDE.2024.3382262.

[6] R. L. T. Santos, P. Castells, I. S. Altingovde, F. Can, Diversity and novelty on the web: Search, recommendation, and data streaming aspects, in: Proceedings of the 24th International Conference on World Wide Web - Companion Volume, 2015, pp. 1529–1530. doi:10.1145/2740908.2741988.

[7] R. L. T. Santos, C. MacDonald, I. Ounis, Search result diversification, Found. Trends Inf. Retr. 9 (2015) 1–90. doi:10.1561/1500000040.

[8] J. G. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, J. Zobel (Eds.), SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, ACM, 1998, pp. 335–336. doi:10.1145/290941.291025.

[9] W. Chen, P. Ren, F. Cai, F. Sun, M. de Rijke, Multi-interest diversification for end-to-end sequential recommendation, ACM Trans. Inf. Syst. 40 (2022) 20:1–20:30. doi:10.1145/3475768.

[10] N. Bouarour, I. Benouaret, S. Amer-Yahia, Learning diversity attributes in multi-session recommendations, in: IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022, IEEE, 2022, pp. 465–474. doi:10.1109/BIGDATA55660.2022.10020476.

[11] K. D. Naini, I. S. Altingovde, W. Siberski, Scalable and efficient web search result diversification, ACM Trans. Web 10 (2016) 15:1–15:30. doi:10.1145/2907948.

[12] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: B. Mobasher, R. D. Burke, D. Jannach, G. Adomavicius (Eds.), Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011, ACM, 2011, pp. 109–116. URL: https://dl.acm.org/citation.cfm?id=2043955.

[13] M. Zhang, N. Hurley, Avoiding monotony: improving the diversity of recommendation lists, in: P. Pu, D. G. Bridge, B. Mobasher, F. Ricci (Eds.), Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, 2008, ACM, 2008, pp. 123–130. doi:10.1145/1454008.1454030.

[14] S. Wang, L. Hu, Y. Wang, Q. Z. Sheng, M. A. Orgun, L. Cao, Modeling multi-purpose sessions for next-item recommendations via mixture-channel purpose routing networks, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence,

IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 3771–3777. doi:10.24963/IJCAI.2019/523.

[15] C. Shi, P. Ren, D. Fu, X. Xin, S. Yang, F. Cai, Z. Ren, Z. Chen, Diversifying sequential recommendation with retrospective and prospective transformers, ACM Trans. Inf. Syst. 42 (2024) 132:1–132:37. doi:10.1145/3653016.

[16] K. Tao, F. Abel, C. Hauff, G. Houben, U. Gadiraju, Groundhog day: near-duplicate detection on twitter, in: D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates, S. B. Moon (Eds.), 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 1273–1284. doi:10.1145/2488388.2488499.

[17] S. Gollapudi, A. Sharma, An axiomatic approach for result diversification, in: J. Quemada, G. León, Y. S. Maarek, W. Nejdl (Eds.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, ACM, 2009, pp. 381–390. doi:10.1145/1526709.1526761.

[18] C. Yu, L. V. S. Lakshmanan, S. Amer-Yahia, It takes variety to make a world: diversification in recommender systems, in: M. L. Kersten, B. Novikov, J. Teubner, V. Polutin, S. Manegold (Eds.), EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24-26, 2009, Proceedings, volume 360 of *ACM International Conference Proceeding Series*, ACM, 2009, pp. 368–378. doi:10.1145/1516360.1516404.

[19] T. A. Feo, M. G. C. Resende, Greedy randomized adaptive search procedures, J. Glob. Optim. 6 (1995) 109–133. doi:10.1007/BF01096763.

[20] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 19:1–19:19. doi:10.1145/2827872.

[21] J. J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, 2015. doi:10.1145/2766462.2767755.

[22] A. P. Singh, G. J. Gordon, Relational learning via collective matrix factorization, in: Y. Li, B. Liu, S. Sarawagi (Eds.), Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, ACM, 2008, pp. 650–658. doi:10.1145/1401890.1401969.

[23] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian, Y. Min, Z. Feng, X. Fan, X. Chen, P. Wang, W. Ji, Y. Li, X. Wang, J. Wen, Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4653–4664. doi:10.1145/3459637.3482016.

[24] M. I. Abdin, J. Aneja, H. S. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024. doi:10.48550/ARXIV.2412.08905.

[25] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 text embeddings: A technical report, CoRR abs/2402.05672 (2024). doi:10.48550/ARXIV.2402.05672.

[26] J. Ni, J. Li, J. J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 188–197. URL: https://doi.org/10.18653/v1/D19-1018. doi:10.18653/V1/D19-1018.