Session-based News Recommendation Using Cohesive Patterns

Mozhgan Karimi
University of Antwerp
Antwerp, Belgium
mozhgan.karimi2@uantwerpen.be

Len Feremans
University of Antwerp
Antwerp, Belgium
len.feremans@uantwerpen.be

Boris Cule
Tilburg University
Tilburg, The Netherlands
b.cule@tilburguniversity.edu

Bart Goethals
University of Antwerp
Antwerp, Belgium
bart.goethals@uantwerpen.be

***Abstract*—In the rapidly evolving field of news recommendation, where user preferences are highly dynamic and content quickly becomes obsolete, providing timely and relevant recommendations presents a significant challenge. Traditional recommender systems typically rely on complex collaborative filtering models that depend on extensive user histories. In the news domain, however, such histories are often scarce due to the high prevalence of anonymous users. To address these challenges, we introduce a novel session-based recommendation method that leverages *cohesive sequential pattern mining*. Rather than relying on traditional frequency-based pattern utility metrics, our approach prioritizes *pattern cohesiveness*, which captures the temporal proximity of item interactions within a pattern, resulting in recommendations that align more closely with the user's ongoing session.**

We conduct a comprehensive empirical evaluation of our approach using four large-scale real-world news datasets. The results demonstrate that our method, SEQCSP, significantly outperforms state-of-the-art session-based recommendation algorithms in terms of accuracy, ranking quality, as well as diversity. Furthermore, SEQCSP provides recommendations faster than most existing methods and is effective for both short and long user sessions, highlighting its robustness, adaptability, and efficiency.

***Index Terms*—large-scale recommendation systems, session-based news recommendation, pattern mining, pattern cohesion**

I. INTRODUCTION

In today's digital era, users are faced with an overwhelming amount of information, making it challenging to discover content that matches their interests. Recommender systems can help navigate these big data spaces by filtering the most relevant content, thereby enhancing user satisfaction. Traditional recommender systems, particularly in domains like movies, often leverage intricate collaborative filtering models, which predict users' interests based on other users' interactions, e.g., via stochastic methods like matrix factorization [1]. These approaches generally require extensive user profiles and are often trained overnight leading to recommendations that primarily focus on long-term user preferences.

However, in the news domain, where long-term user profiles are scarce and articles quickly lose relevance, such personalized approaches are often ineffective. Instead, the rapid emergence of new topics demands systems that adapt quickly to users' evolving interests and provide fresh suggestions [2]. Session-based recommender systems address these challenges by focusing on the sequence of actions within individual consumption sessions, thereby discerning users' immediate preferences [3]. State-of-the-art session-based approaches frequently employ relatively simple methods, such as k-Nearest Neighbors [4]–[6] or basic association/sequential rule mining [5], [7]. While effective, these may overlook intricate temporal patterns within user sessions, especially in fast-paced domains like news. Recently, advanced sequential pattern mining (SPM) approaches, like MARBLES [8], have demonstrated their ability to outperform state-of-the-art session-based recommendation approaches across various news datasets [9]. However, an unexplored question remains: can recommendation quality be further enhanced by focusing on *cohesive* patterns, i.e., patterns whose item interactions occur in close temporal proximity?

In this paper, we propose a novel application of *quantile-based cohesive sequential pattern mining* (QCSP) for session-based news recommendation. QCSP is a sequential pattern mining technique that introduces the concept of quantile-based cohesion, a measure of pattern cohesion that is robust to outliers. We posit that prioritizing cohesiveness over the traditional interestingness measure of pattern frequency could be particularly beneficial for news recommendation, where user interests constantly shift and breaking news needs to be identified quickly. In such scenarios, prioritizing patterns that appear less often, but proportionally closer together may serve as a more reliable indicator of their utility.

The remainder of this paper is structured as follows. In Section II, we review related work in session-based (news) recommendation with a focus on techniques that employ sequential pattern mining. Section III provides a formal definition of the session-based recommendation task. Section IV outlines our approach and details the adaptation of QCSP to the task of session-based recommendation. In Section V, we present the results of an empirical evaluation of our approach on four large real-world news datasets, comparing its effectiveness against state-of-the-art methods. Finally, in Section VI, we summarize our findings and discuss potential future work.

II. RELATED WORK

Already in the early days of recommender systems research, association rule mining and sequential pattern mining approaches have been used to predict, for example, which website a user is likely to visit next [10] or which e-learning resource a user should study next [11], [12]. However, since the 2006 *Netflix Prize* challenge [1], which aimed to improve the rating

prediction performance of Netflix's recommendation algorithm, academic research has shifted much of its focus to personalized recommendation by means of rating prediction [13]. Initially, research concentrated on traditional information retrieval methods, such as k-nearest neighbor (kNN) algorithms and stochastic approaches like matrix factorization [1]. However, nowadays, deep learning techniques are increasingly employed to improve prediction accuracy [14]. At the same time, many modern online services lack explicit feedback mechanisms, and users' short-term interests often outweigh long-term preferences, particularly in domains like news, making the aforementioned approaches less effective.

In these cases, session-based recommendation algorithms [3], which focus on users' recent interactions, have proven effective where traditional recommender systems fall short. A diverse range of session-based approaches has been explored, including methods based on association and sequential rules [7], [9], [15], nearest neighbors [4], [5], [16], and recurrent neural networks [17]–[19]. However, independent evaluations of session-based recommendations have shown that simple models generally outperform complex deep learning models [5], [6].

Despite the success of these classic approaches in session-based recommendation, research on more *sophisticated* versions of these methods remains somewhat limited. Many approaches simply employ rules of size two [5], [7], effectively mining item-item co-occurrences within sessions. Notable among these is the SR method [7], because similar to our approach, it also features an interestingness measure based on cohesiveness. However, as this approach uses a simple aggregation mechanism to quantify cohesiveness, it is likely more susceptible to outliers than our proposed method. To test this assumption, we include SR as a baseline in our performance evaluations. On the opposite end of the spectrum, some related methods from the news domain interpret each session in the training data as one contiguous pattern [4], [20]. Thus, they prioritize longer patterns, potentially at the expense of applicability, as news consumption sessions are short on average.

Lastly, some approaches leverage additional contextual information, such as the user's location or time of day, to enhance recommendation accuracy [21], [22]. Among these, another method incorporates the concept of cohesiveness. Rather than weighting pattern occurrences by their cohesiveness, this approach segments consumption sessions into *cohesive units* that share similar contexts, such as movies from the same genre, and then extracts association rules from these units [22]. However, due to its dependence on contextual information, it lacks flexibility compared to our proposed method.

III. Problem Definition

As previously mentioned, session-based recommender systems solely utilize the user's most recent activity. The session-based recommendation task has been defined in various ways in the literature [3]. For the purpose of this paper, we define it as follows: Let the training set S_{train} consist of a number of historical sessions $\{s_1, \ldots, s_n\}$ from different users. Each session comprises an ordered list of item interactions, denoted as $s_x = (i_1, \ldots, i_m)$, where $i_x \in I$ (the set of all items).

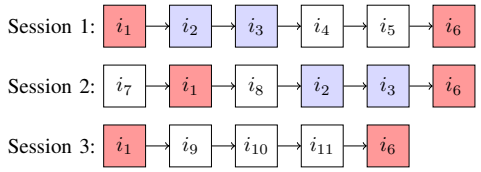

Fig. 1. Example illustrating the importance of *cohesiveness* when mining patterns from sessions. Which pattern is more valuable: (i_1, i_6) or (i_2, i_3)?

Given the training set S_{train}, algorithms are tasked with providing recommendations for a further set of sessions, S_{test}. Crucially, these sessions are not entirely revealed to the recommendation algorithms. Instead, a portion of items from the latter part of each session is withheld and used as ground truth for evaluating the algorithm's performance. For example, given a session $s_x = (i_1, i_2, i_3, i_4, i_5)$ from S_{test}, depending on the evaluation methodology, an algorithm might be tasked with predicting item i_4 and/or i_5, based on items i_1, i_2, and i_3.

IV. Proposed Approach

In this section, we first motivate the usage of cohesive patterns for session-based recommendation. Then, we provide an introduction to the pattern mining approach that underpins our recommendation scheme. Lastly, we detail how the mined patterns can be applied to session-based recommendation.

A. Motivation

Sequential pattern mining approaches have been used in various session-based recommendation scenarios [4], [7], [9], [23]. In these cases, sequential patterns are mined from the training sessions S_{train}, to be leveraged during the testing phase for generating recommendations for user sessions in S_{test}. However, when estimating the interestingness of patterns, most works rely on the frequency with which patterns appear in the training data, a measure commonly referred to as *support*. While this measure of interestingness is straightforward to compute, it presents several limitations in practical applications. One key drawback is that most algorithms require a support *threshold* to facilitate pruning, which must be determined empirically, as it is highly dependent on the characteristics of each dataset. Additionally, weighting patterns based on their support prioritizes popular items, which leads to low diversity.

We aim to address these issues by focusing on *cohesiveness* as an alternative measure of a pattern's interestingness. For example, consider the scenario illustrated in Figure 1. Here, the pattern (i_2, i_3) appears only twice, while the pattern (i_1, i_6) appears three times. Consequently, most traditional sequential pattern mining approaches [24], [25] would assign a higher value to (i_1, i_6) than to (i_2, i_3). Nevertheless, it is evident that the pattern (i_2, i_3), despite occurring less frequently, could still be valuable because its items are interacted with in closer proximity, i.e., the pattern is more cohesive.

We hypothesize that cohesiveness can serve as a more effective predictor of pattern utility in recommender systems,

particularly in domains such as news recommendation, where item spaces are large and item relevance evolves rapidly. By scoring patterns primarily based on their cohesiveness instead of their support, we anticipate several advantages over traditional pattern mining approaches when using these patterns for recommendation:

- In environments where short-term user interests can change quickly, recommendations may be more relevant when prioritizing patterns that appear closer together in the training data, as these patterns are more likely to follow a common contextual thread (as illustrated by the example in Figure 1).
- Recommendations would likely be less skewed towards popular items, addressing the common "rich get richer" phenomenon observed in recommender systems [26].
- Cohesive patterns could also be valuable in item cold-start scenarios, particularly in the context of breaking news, where prior interactions are scarce [27], and thus, support is low.
- Long-tail items, which can be highly relevant for users with strongly focused sessions, are often overlooked due to suboptimal support thresholds. Focusing on cohesion can eliminate the need for support thresholds, thereby enhancing recommendation diversity.

B. Cohesive Pattern Mining via QCSP

We leverage the sequential pattern mining approach by Feremans et al. [28], which offers several advantages over traditional pattern mining techniques, making it particularly suited for this scenario. This method, referred to as *top-k quantile-based cohesive sequential pattern mining*, or QCSP, is designed to extract interesting patterns from input sequences. However, unlike traditional pattern mining approaches [24], [25], QCSP defines pattern interestingness based on the *proportional cohesiveness* of patterns within the input data.

To clarify the concept of proportional cohesiveness, consider the following example: Given an input sequence $s = (i_1, i_2, i_3, i_4, i_5)$ and a potentially interesting pattern $p = (i_1, i_3, i_5)$, QCSP calculates the cohesion score $c(p,\, s)$ for the pattern's occurrence within the session by dividing the pattern's span within the input sequence (5) by the pattern's size (3), resulting in $c(p,\, s) = 5/3$. QCSP then applies a threshold (α) to each occurrence's cohesion score $c(p,\, s)$ to determine whether an occurrence is considered cohesive. For example, given $\alpha = 2$, the above-mentioned occurrence of p within s would be considered cohesive, while with a cohesiveness threshold $\alpha = 1.5$, it would not be considered cohesive. QCSP then divides the number of cohesive occurrences of a given pattern by the total number of occurrences in all sessions, yielding the pattern's proportional cohesiveness, also dubbed *quantile-based cohesion*, which can be defined as:

$$C_{\text{quan}}(p) = \frac{\sum_{s \in S_{\text{train}}} |p| \cdot 1_c(p,\, s)}{\sum_{i \in p} \sigma(i)} \tag{1}$$

with $\sigma(i)$ being the support of item i and

$$1_c(p,\, s) = \begin{cases} 1 & \text{if } p \text{ occurs in } s \text{ and } c(p,\, s) < \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that QCSP was originally designed to mine patterns in a single large sequence. We have consequently adapted the definition of C_{quan} for session-based data and omitted certain details for clarity, such as the calculation of minimal windows for pattern occurrences. The algorithm also imposes a limit on pattern lengths and incorporates several efficiency improvements, including prefix-projected pattern growth. For further details, please refer to the original publication [28]. Experimental results have demonstrated that QCSP is effective for standard pattern mining tasks on sequential data, outperforming other state-of-the-art SPM approaches. After training, QCSP provides us with the set of top-k patterns that exhibit the highest quantile-based cohesion, C_{quan}.

C. Utilizing Cohesive Patterns for Recommendation

After applying QCSP to the training sessions, we can harness the resulting top-k patterns to generate recommendations. We refer to this adaptation of QCSP for the task of session-based recommendation as SEQCSP. Specifically, given a set of top-k cohesive patterns P, we score each item i based on its relevance to the current user session s as follows.

$$\hat{r}(i\,, s) = \sum_{p \,\in\, P \,\wedge\, p_{|p|} = i} C_{\text{quan}}(p) \,\cdot\, \sigma(p) \,\cdot\, \phi\,(p \setminus i\,, s)^{\beta} \tag{3}$$

Here, p_x denotes the x-th item in the pattern p. Consequently, we include all patterns p in the score whose final item, $p_{|p|}$, matches the item we are trying to score. For instance, if we are calculating the score for item i_3, we sum up only the scores of patterns that conclude with i_3.

Each candidate pattern p then contributes to the score based on three major characteristics: C_{quan}, the previously described quantile-based cohesion of the pattern; σ, the pattern's support; and ϕ, a function that quantifies the match between the pattern and the current user session. Note that, when applying ϕ, the candidate item i is excluded from the pattern, as it is not relevant when determining the match between pattern and session. We further introduce a constant β, which can be used to adjust the influence of ϕ relative to the other two factors.

The session-pattern-match function ϕ allows us to refine the recommendation process, ensuring that the suggested items are highly relevant to the user's current context. It consists of four sub-components and is calculated as follows.

$$\phi(p\,, s) = \chi(p\,, s) \,\cdot\, \phi_{\text{O}}(p\,, s) \,\cdot\, \phi_{\text{L}}(p\,, s) \,\cdot\, \phi_{\text{C}}(p\,, s) \tag{4a}$$

$$\chi(p\,, s) = \begin{cases} 1 & \text{if } p \subseteq s \text{ and items of } p \text{ appear} \\ & \text{in the correct order in } s, \\ 0 & \text{otherwise.} \end{cases} \tag{4b}$$

$$\phi_O(p\,, s) = \frac{|p|}{|s|} \tag{4c}$$

$$\phi_L(p\,, s) = \frac{1}{|s| - \text{pos}\big(p_{|p|}\,, s\big)} \tag{4d}$$

$$\phi_C(p\,, s) = \frac{|p|}{\text{pos}\big(p_{|p|}\,, s\big) - \text{pos}(p_1\,, s) + 1} \tag{4e}$$

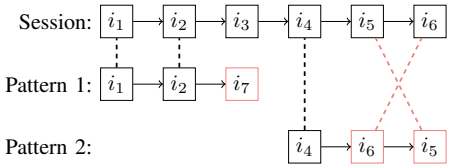

Fig. 2. Example illustrating $\chi(p, s)$. Patterns 1 and 2 cannot be applied to the session due to either an extra item or an incorrect item order, respectively.

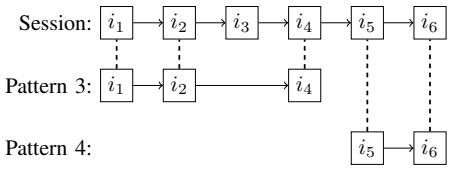

Fig. 3. Example illustrating $\phi_O(p, s)$ and $\phi_L(p, s)$. Pattern 3 overlaps more of the session, but pattern 4 aligns better with the user's recent interactions.

Here, $\text{pos}(i, s)$ denotes the 0-indexed position of item i within the current session s. The first component, χ, is an indicator function that evaluates whether the candidate pattern can be applied to the current user session based on two criteria. Figure 2 shows an example which illustrates these conditions. In the figure, pattern 1 cannot be applied to the current session, because the session does not contain item i_7. In contrast, all items of pattern 2 are present in the current session. However, the order in which the items of the pattern appear within the session does not match the order of the pattern. Thus, we consider neither pattern 1 nor pattern 2 to be applicable to the current session, resulting in $\chi = 0$ and, thus, eliminating their influence on the candidate item score $\hat{r}$.

If, on the other hand, a pattern fulfills the minimum criteria defined by χ, the three other session-pattern-matching sub-functions (ϕ_O, ϕ_L, and ϕ_C) come into effect. The first of these, ϕ_O, calculates the degree to which the pattern *overlaps* the current session. Based on the example in Figure 3, we can observe that pattern 3 would result in a score of $3/6$, whereas pattern 4 would only score $2/6$. By scoring patterns higher that cover more of the user session, ϕ_O prioritizes patterns that better align with the user's *overall* interests in the session.

The second scoring sub-component, ϕ_L, measures how well a pattern matches the user's *latest intent*. To this end, ϕ_L assigns higher scores to patterns whose last item is positioned more towards the end of the user's current session. Referring again to Figure 3, we see that pattern 3, which received a higher overlap score than pattern 4, only achieves a latest intent score of $1/3$. In contrast, ϕ_L assigns pattern 4 a perfect score of 1, because i_6 matches the last item in the user's session.

Lastly, ϕ_C scores patterns based on their *cohesiveness* within the user's current session, quantified as the inverse of how far patterns span across the session's interactions. By using the pattern's length as a denominator, we normalize the measure, resulting in a perfect score of 1 for patterns that occur without gaps in the session. For instance, in Figure 4, patterns 5 and 6 score the same with respect to overlap (ϕ_O) and latest user intent (ϕ_L). However, since pattern 5 spans six interactions

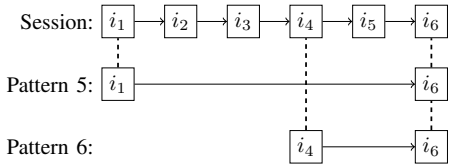

Fig. 4. Example illustrating $\phi_C(p, s)$. Pattern 6 occurs more cohesive within the user session than pattern 5.

and pattern 6 spans only three, pattern 6 receives a higher in-session cohesion score ($2/3$) compared to pattern 5 ($2/6$).

After calculating these three pattern-session-match scores and combining them with QCSP's quantile-based cohesion and support scores as outlined in Equation 3, we derive each item's final recommendation score $\hat{r}(i, s)$ as the sum over all patterns that include i as their last item. These scores are then sorted to generate a ranked list of item recommendations that best fit the user's current session.

V. Empirical Evaluation

To assess the effectiveness of our approach in comparison to state-of-the-art methods, we conduct a comprehensive experimental evaluation using four real-world news datasets.

A. Setup

First, we outline the setup of our experimental evaluation, including the software framework used to run the experiments, the datasets and their characteristics, and the baseline algorithms against which we benchmark our proposed approach.

1) Evaluation Framework: We use the open-source framework StreamingRec [4] for our performance evaluation, which provides reference implementations of leading-edge session-based recommendation algorithms. StreamingRec follows a streaming evaluation protocol wherein interactions are replayed in real time. During training, algorithms receive part of the data to build initial models. The remaining data is then replayed in the original order, and each algorithm must generate recommendations based on every new interaction, while also keeping its model up to date by learning incrementally.

Due to the inherent complexity of the top-k retrieval mechanism of SeqCSP, adapting it to incremental learning is difficult. However, because of SeqCSP's fast training times, it's model can simply be recalculated intermittently, i.e., every δ_t minutes of simulation time. By employing a queue of recent sessions δ_s, similar to other state-of-the-art session-based recommenders [4], [9], we keep the overhead of recreating the model manageable. Section V-B1b further examines SeqCSP's computational efficiency relative to baseline algorithms.

2) Datasets: The above-mentioned replay evaluation can only be conducted on datasets with timestamped interaction logs, as these allow us to accurately reconstruct the sequence of interactions. Consequently, a number of popular datasets, such as MIND [29]—which lacks an intra-session click order—are unsuitable. For a rigorous and realistic assessment, we select four datasets, all featuring timestamped interactions from real-world news platforms. The first dataset is based

TABLE I
DATASET CHARACTERISTICS

	Outbrain	Plista	EB-NeRD	Adressa
Interactions	1,067,641	2,052,213	3,334,389	17,928,078
Items	1,565	5,177	20,738	61,483
Sessions	421,620	667,339	701,341	6,237,110
Time frame	2 weeks	4 weeks	6 weeks	10 weeks
Avg. interact. per session	2.5	3.1	5.0	2.9
Avg. interact. per item	724	452	316	396

on the 2017 *Outbrain* click prediction challenge[1] and features anonymized data from several English news publishers, from which we use publisher "43". The *Plista* dataset originates from the CLEF NewsREEL competition in 2017 [30], in which researchers could benchmark their recommender algorithm implementations in a real-world multi-publisher setting. Here, we selected the large German language publisher "1677". The most recent dataset is the *EB-NeRD* dataset. It was published in the context of the 2024 RecSys challenge[2] and is exclusively based on data from the Danish tabloid news publisher Ekstra Bladet. Lastly, the *Adressa* dataset [31], which was collected in 2017 for an academic project, is the largest dataset in our comparison and contains data from the Norwegian local newspaper Adressavisen. Table I summarizes the key characteristics of each dataset, after preprocessing, e.g., by filtering duplicates.

3) Baseline Algorithms: We compare our approach with a number of state-of-the-art session-based recommendation algorithms as well as a few simple baselines:

- Two simple, yet often moderately effective [4] baselines are RECENTLYCLICKED and RECENTLYPOPULAR. RECENTLYCLICKED recommends items that were most recently interacted with, regardless of user or session, and RECENTLYPOPULAR ranks items based on the number of interactions each item received in the last n minutes.
- The COOCCURRENCE algorithm counts how frequently each item co-occurred with other items in the training sessions. It then recommends items that co-occurred most frequently with the items in the user's current session [4].
- The SR (Sequential Rules) algorithm [7] mines sequential patterns of size two based on the training sessions. Similarly to SEQCSP, it calculates a cohesion score for each pattern based on the average distance of the two involved items within all occurrences of the pattern.
- The SEQ_r algorithm [15] mines sequential patterns incrementally in a pattern tree. It can thus react quickly to trends, but is not always as accurate due to its lack of pruning and sole reliance on pattern frequency.
- V-SkNN is one of the best-performing algorithms in the field of session-based recommendation, regardless of the application domain, and consistently outperforms even neural-network approaches [4]–[6]. We employ an implementation with a session-session similarity metric that prioritizes more recent user interactions [23] and which maintains a queue of recent sessions for efficiency and freshness [4].
- Finally, we include the session-based implementation of the MARBLES sequential pattern mining algorithm [8], which has been shown to perform well in news recommendation [9]. In contrast to SEQCSP, MARBLES mines *episodes*, which are partially ordered patterns, allowing it to capture more complex temporal session dynamics than other algorithms.

Several of the above algorithms, including SEQCSP, require hyper-parameter tuning for optimal performance. For example, SEQCSP's cohesiveness threshold α must be adjusted for each dataset to ensure effective operation. We experimentally determined these values using validation sets, comprising approximately 10% of each dataset.

4) Metrics: To quantify each algorithm's ability to produce relevant recommendations, we employ two well-known evaluation metrics from the field of information retrieval. We use the F1 score—a combination of the traditional Precision and Recall metrics—to measure retrieval accuracy. Furthermore, we utilize the Mean Reciprocal Rank (MRR) metric to quantify how high relevant items are placed in the recommendation list.

In addition to these standard accuracy metrics, we also evaluate the algorithms using beyond-accuracy criteria. To this end, we calculate the Gini coefficient, a measure of diversity which captures how well an algorithm distributes its recommendations among the available item space. Lastly, we measure each algorithm's efficiency in two dimensions: the time required to train the model and the latency experienced when generating a recommendation list.

B. Results

In this section, we detail the findings from our empirical evaluation, starting with a comprehensive performance and beyond-accuracy comparison of the different methods. We then examine how algorithm effectiveness varies with session length. Finally, we conduct an ablation study to identify key factors contributing to SEQCSP's performance.

1) Overall Performance: Table II summarizes the overall performance results of our empirical evaluation. Consistent with standard practice in recommender systems research, we focus on top-10 recommendation lists for our measurements.

a) Recommendation Effectiveness: Among the simple baselines, RECENTLYCLICKED performs significantly worse than RECENTLYPOPULAR, which occasionally matches or even surpasses more complex baselines. For most datasets, there is a noticeable improvement in performance, especially in terms of MRR, between the simple baselines and the more complex baselines, which utilize schemes based on item co-occurrences, sequential patterns, and nearest neighbors. From this group, the COOCCURRENCE approach is the least consistent and often underperforms relative to other methods.

Interestingly, the two simple sequential pattern mining approaches SEQ_r and SR perform comparably across datasets, in terms of F1 and MRR. This similarity could be due to the

[1] https://www.kaggle.com/c/outbrain-click-prediction
[2] https://recsys.eb.dk

TABLE II
OVERALL ALGORITHM PERFORMANCE

	Outbrain			Plista			EB-NeRD			Adressa		
	F1	MRR	Gini	F1	MRR	Gini	F1	MRR	Gini	F1	MRR	Gini
RECENTLYCLICKED	0.0760	0.1249	**0.8501**	0.0868	0.1359	**0.9202**	0.1242	0.1645	**0.8186**	0.0811	0.1311	**0.9520**
RECENTLYPOPULAR	0.1168	0.2090	0.9550	0.1224	0.1848	0.9692	0.1612	0.1687†	0.8675	0.1176	0.1830	0.9824
COOCCURRENCE	0.1404‡	0.2957	0.8917	0.1181†	0.1580	0.9619	0.1791	0.1688†	0.8964	0.1199	0.2104	0.9772
SEQ$_r$	0.1414†‡	0.3100†	0.8741	0.1186†	0.1676	0.9652	0.1912	0.2590	0.8788	0.1234	0.2472†	0.9729
SR	0.1425†	0.3106†	0.8786	0.1213	0.1710	0.9610	0.1985	0.2646	0.8865	0.1244	0.2455	0.9737
V-SkNN	0.1581	0.3444	0.9022	0.1462	0.2603	0.9652	0.2066	0.2498	0.8875	0.1384†	0.2571	0.9761
MARBLES$_f$	0.1895	0.3268	0.9042	0.1369‡	0.2071†	0.9671	0.1891	0.2422	0.9888	0.1406‡	0.2600	0.9769
MARBLES$_m$	0.2038	0.3155	0.9231	0.1358‡	0.1980	0.9609	0.1874	0.2395	0.9898	0.1395†‡	0.2481†	0.9772
MARBLES$_w$	**0.2147**	0.2984	0.9070	0.1398	0.2070†	0.9629	0.1816	0.2379	0.9890	0.1373	0.2013	0.9760
SEQCSP	0.1786	**0.3481**	0.8575	**0.1501**	**0.2816**	0.9608	**0.2252**	**0.3060**	0.8360	**0.1441**	**0.2867**	0.9652

Note: All pairwise differences in F1 and MRR were significant according to a Kolmogorov-Smirnov test with $p < 0.01$, except where indicated by † or ‡.

simple support-focused implementation of both approaches, whereas minor differences between these methods, e.g., on the Plista and EB-NeRD datasets, could be explained by SR's sole reliance on patterns of size two. When comparing SR and SEQCSP, which both use an interestingness measure based on cohesiveness, SR's performance lags notably behind, likely due to SEQCSP's more complex outlier-resistant *quantile-based cohesion* approach.

The most accurate methods are V-SkNN, MARBLES, and SEQCSP. Among these, V-SkNN consistently performs well, often ranking second behind SEQCSP. In line with previous works [4], [15], V-SkNN performs especially well in terms of MRR, indicating that it is able to place relevant items high in recommendation lists. The MARBLES approach also performs well and even surpasses SEQCSP in terms of F1 on the Outbrain dataset. However, the different variants of MARBLES—fixed window (f), minimal window (m), and weighted minimal window (w)—vary strongly in performance across datasets. Consequently, in addition to requiring tweaking of numerous hyper-parameters and taking the longest to run, practitioners must also determine the variant of MARBLES that performs best on their dataset, which reduces its practical applicability compared to SEQCSP.

With the exception of the aforementioned F1 result on the Outbrain dataset, SEQCSP significantly outperforms all baseline approaches (with $p < 0.01$), often by a substantial margin. Notably, SEQCSP achieves markedly better results than state-of-the-art algorithms like V-SkNN and MARBLES on the two larger datasets, EB-NeRD and Adressa. Given that companies often share limited data with researchers, larger datasets like EB-NeRD and Adressa, with their extensive catalogs, are likely more reflective of real-world conditions, suggesting that our approach is well-suited for real-world big data applications.

b) Beyond-Accuracy Performance: Optimizing algorithms solely based on accuracy measures, such as F1 or MRR, can result in designs that favor a small subset of popular items, as these are more likely to produce "hits." However, users often *value* diversity in recommendations, seeking relevant yet novel content, and publishers also aim to prevent articles from receiving limited exposure. The Gini coefficient measures diversity by indicating how recommendations are distributed. A high Gini value suggests that (top-10) recommendations are concentrated around a few popular items, while a low Gini value reflects a broader variety of items included in recommendation lists.

The results show that SEQCSP is among the most diverse recommendation algorithms in our experiments. It is only outperformed in Gini value by the RECENTLYCLICKED baseline, which is expected due to the method's inherent randomness. Notably, other top-performing algorithms, like MARBLES and V-SkNN, consistently show relatively low diversity, often ranking behind algorithms like SR, SEQr, and COOCCURRENCE. Overall, SEQCSP's ability to generate consistently more diverse recommendations compared to other methods corroborates our initial hypothesis from Section IV-A that using pattern cohesion instead of support can mitigate popularity biases, leading to more long-tail recommendations.

Regarding efficiency (not shown in Table II), simple approaches, such as RECENTLYPOPULAR, COOCCURRENCE, and SR, require virtually no training time and can generate recommendations in under 1 ms. In contrast, the top-performing methods are significantly slower. For instance, on the largest dataset, V-SkNN takes an average of 11 ms to generate a recommendation list, because it computes session similarities in real time. Similarly, MARBLES generates recommendation lists in around 14 ms, but also requires intermittent retraining, which takes 90 s on average. As a result of this overhead, MARBLES cannot be retrained as frequently, making it prone to recommending outdated items.

In comparison, SEQCSP achieves more manageable intermittent training times (3 s on average) due to optimizations such as pruning. Moreover, SEQCSP generates recommendation lists in under 1 ms across all datasets, likely due to its top-k pattern retrieval mechanism. By focusing on a limited set of the k most

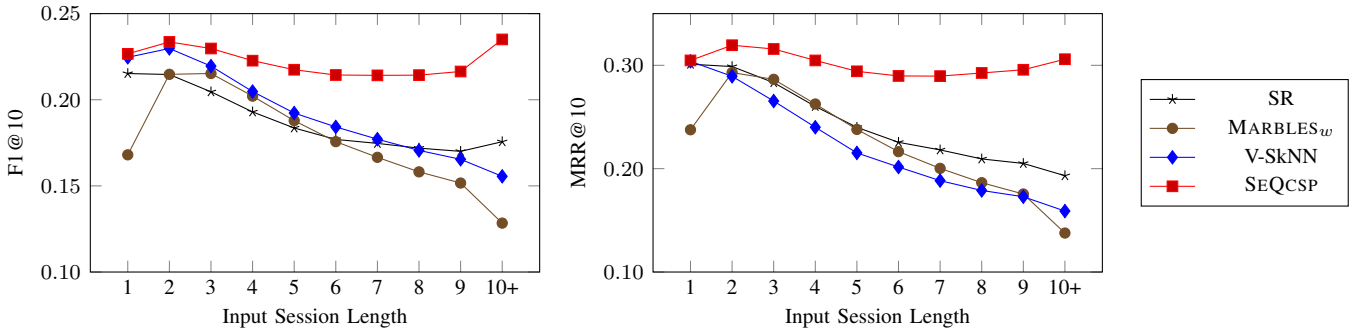

Fig. 5. Algorithm performance on the EB-NeRD dataset given different input session lengths

cohesive patterns, SEQCSP can very efficiently retrieve the most relevant patterns for a given user session. Overall, SEQCSP's ability to quickly generate recommendations combined with its superior recommendation accuracy, ranking performance, and diversity underscores its practical utility for large-scale, session-based recommendation in the news domain.

2) Performance at Different Session Lengths: Examining the overall accuracy or ranking performance of an algorithm provides only a partial view of its effectiveness. In practice, algorithms may exhibit different strengths and weaknesses depending on user behavior. For example, some users consume news in short bursts, while others read multiple articles consecutively. Similarly, some users prefer to switch between topics, whereas others focus on a single topic before moving on. One way to assess how well an algorithm handles this diversity in user behavior is by analyzing its recommendation performance across different session lengths.

Figure 5 presents the results of such an analysis, showing the algorithms' performance in terms of F1 and MRR across various session lengths. Here, the leftmost data point represents the performance results when the user session for which the algorithms generated recommendations consisted of only one item. In contrast, the rightmost point reflects sessions with ten or more items. For this analysis, we discuss the performance of the top-performing methods, SR, MARBLES and V-SkNN on the EB-NeRD dataset. However, similar trends were observed across the other datasets.

The analysis reveals substantial variations in performance for SR, V-SkNN, and MARBLES across different session lengths. MARBLES in particular struggles with both very short and long sessions, achieving strong performance only for sessions with two or three items. While V-SkNN and SEQCSP also experience a slight drop in accuracy for single-item sessions, they do not decline as sharply as MARBLES. By contrast, SR performs better for these single-item sessions than at any other session length, likely due to its focus on short patterns. Interestingly, V-SkNN and SEQCSP show quite similar results for single-item sessions, suggesting that the benefits of considering cohesive patterns may be less pronounced in very short sessions.

However, SEQCSP stands out as the only algorithm whose performance consistently increases for five or more interactions. It exhibits a particularly sharp rise in F1 for sessions

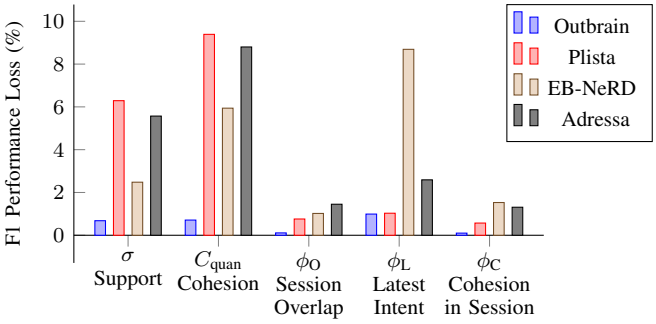

Fig. 6. Ablation study results. Each group of bars shows the performance loss when one of the scoring components from Equations 3 or 4 is left out.

exceeding ten items, whereas V-SkNN and MARBLES decline in performance for these long sessions. Most notably, SEQCSP consistently ranks as the best-performing algorithm across all session lengths in terms of both F1 and MRR, while other approaches frequently shift in ranking. These qualities make SEQCSP a promising choice for session-based news recommendation, where adapting to diverse user behaviors is crucial.

3) Ablation Study: To identify the key contributors to SEQCSP's performance, we additionally conduct an ablation study, in which we systematically remove each component of SEQCSP's scoring model, as defined in Equations 3 or 4. Figure 6 shows the results of this study.

The analysis confirms our hypothesis that pattern cohesion is a crucial factor for recommendation accuracy. In fact, quantile-based pattern cohesion is the most important scoring component across all datasets, with the exception of EB-NeRD. For this dataset, the latest-user-intent matching score has a higher impact on performance than the cohesion score. A possible explanation is that sessions in the EB-NeRD dataset are longer on average and, thus, users' short-term interests are more likely to fluctuate, making it crucial to prioritize their most recent interactions. Notably, across all datasets, support plays a significantly smaller role than cohesion, suggesting that the traditional emphasis on pattern frequency does not necessarily lead to optimal recommendations. Interestingly, the importance of pattern cohesion within the session—quantified by how far a pattern spans within the user's current session—varies across datasets. As expected, it becomes more impactful in

the EB-NeRD dataset, where sessions are longer. This finding suggests that pattern cohesion within the current session can be particularly valuable for application on news websites where users tend to have longer reading sessions.

VI. Conclusions and Future Work

In this paper, we propose SeqCSP, a session-based recommendation approach that leverages cohesive sequential patterns, i.e., patterns whose item interactions occur temporally close together in the training data. Our recommendation scoring model is based on pattern support and cohesion as well as the degree to which patterns overlap the user's current session. To further enhance recommendation accuracy, we also consider how cohesive patterns appear *within* the current user session and how well patterns align with the user's *most recent* interactions.

Our empirical evaluation, conducted on four large real-world news datasets, demonstrates that our method significantly outperforms state-of-the-art session-based approaches in terms of accuracy and diversity, sometimes by a substantial margin. A subgroup analysis further reveals that our approach performs well across short and long user sessions, underscoring its versatility. Lastly, the results of an ablation study indicate that pattern cohesion is an important factor for recommendation quality.

For future work, we plan to further investigate how session-based recommendations can benefit from cohesive sequential patterns in other domains where discovering temporal patterns has high potential, such as music and e-commerce. We also aim to hybridize our approach with conceptually different methods to potentially enhance performance through complementary recommendation schemes. Additionally, we intend to incorporate contextual factors, such as user location and time of day, to further refine our pattern mining process.

Acknowledgments

This research was supported by the Research Foundation Flanders (FWO) grant 12B0V24N to L. Feremans.

References

[1] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Matrix factorization and neighbor based algorithms for the netflix prize problem," in *Proceedings of the 2nd Conference on Recommender Systems (RecSys '08)*, 2008, p. 267–274.

[2] M. Karimi, D. Jannach, and M. Jugovac, "News recommender systems–survey and roads ahead," *Information Processing & Management*, vol. 54, no. 6, 2018.

[3] S. Wang, L. Cao, Y. Wang, Q. Z. Sheng, M. A. Orgun, and D. Lian, "A survey on session-based recommender systems," *ACM Computing Surveys*, vol. 54, no. 7, 2021.

[4] M. Jugovac, D. Jannach, and M. Karimi, "Streamingrec: a framework for benchmarking stream-based news recommenders," in *Proceedings of the 12th Conference on Recommender Systems (RecSys '18)*, 2018.

[5] M. Ludewig, N. Mauro, S. Latifi, and D. Jannach, "Empirical analysis of session-based recommendation algorithms," *User Modeling and User-Adapted Interaction*, vol. 31, no. 1, pp. 149–181, 2021.

[6] S. Moliński, "WSKNN - weighted session-based K-NN recommender system," *Journal of Open Source Software*, vol. 8, no. 90, 2023.

[7] I. Kamehkhosh, D. Jannach, and M. Ludewig, "A comparison of frequent pattern techniques and a deep learning method for session-based recommendation," in *Proceedings of the 1st Workshop on Temporal Reasoning in Recommender Systems (RecTemp@RecSys '17)*, 2017, pp. 50–56.

[8] B. Cule, N. Tatti, and B. Goethals, "MARBLES: Mining association rules buried in long event sequences," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 7, no. 2, pp. 93–110, 2014.

[9] M. Karimi, B. Cule, and B. Goethals, "Leveraging sequential episode mining for session-based news recommendation," in *Proceedings of the 24th International Conference on Web Information Systems Engineering (WISE '23)*, 2023, p. 594–608.

[10] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data," in *Proceedings of the 3rd International Workshop on Web Information and Data Management (WIDM '01)*, 2001, pp. 9–15.

[11] O. Zaiane, "Building a recommender agent for e-learning systems," in *Proceedings of the 2002 International Conference on Computers in Education (ICCE '02)*, 2002, pp. 55–59.

[12] C.-H. Lai and D.-R. Liu, "Integrating knowledge flow mining and collaborative filtering to support document recommendation," *Journal of Systems and Software*, vol. 82, no. 12, pp. 2023–2037, 2009.

[13] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, 2022.

[14] R. Mu, "A survey of recommender systems based on deep learning," *IEEE Access*, vol. 6, pp. 69 009–69 022, 2018.

[15] M. Karimi, B. Cule, and B. Goethals, "On-the-fly news recommendation using sequential patterns," in *Proceedings of the 7th International Workshop on News Recommendation and Analytics (INRA@RecSys '19)*, 2019.

[16] M. Karimi, "SessionPrint: Accelerating kNN via locality-sensitive hashing for session-based news recommendation," in *Proceedings of the 15th Conference and Labs of the Evaluation Forum (CLEF '24)*, 2024, pp. 159–165.

[17] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proceedings of the 4th Conference on Learning Representations (ICLR '16)*, 2016.

[18] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *Proceedings of the 2017 on Conference on Information and Knowledge Management (CIKM '17)*, 2017, pp. 1419–1428.

[19] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the 27th International Conference on Information and Knowledge Management (CIKM '18)*, 2018, pp. 843–852.

[20] P. Bons, N. Evans, P. Kampstra, and T. van Kessel, "A news recommender engine with a killer sequence," in *Working Notes of the 8th Conference and Labs of the Evaluation Forum (CLEF '17)*, 2017.

[21] A. Lommatzsch, "Real-time news recommendation using context-aware ensembles," in *Proceedings of the 36th European Conference on Information Retrieval (ECIR' 14)*, 2014, pp. 51–62.

[22] J. Kim and J.-H. Lee, "A novel recommendation approach based on chronological cohesive units in content consuming logs," *Information Sciences*, vol. 470, pp. 141–155, 2019.

[23] M. Ludewig and D. Jannach, "Evaluation of session-based recommendation algorithms," *User Modeling and User-Adapted Interaction*, vol. 28, p. 331–390, 2018.

[24] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, 1994, p. 487–499.

[25] H. Mannila, H. Toivonen, and A. Inkeri Verkamo, "Discovery of frequent episodes in event sequences," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, 1997.

[26] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, and E. Malthouse, "User-centered evaluation of popularity bias in recommender systems," in *Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization (UMAP '21)*, 2021, p. 119–129.

[27] Y. Wang, J. Wu, Z. Wu, H. Yuan, and X. Zhang, "Popular items or niche items: Flexible recommendation using cosine patterns," in *Workshop Proceedings of the 2014 International Conference on Data Mining (ICDM '14)*, 2014, pp. 205–212.

[28] L. Feremans, B. Cule, and B. Goethals, "Mining top-*k* quantile-based cohesive sequential patterns," in *Proceedings of the 2018 SIAM International Conference on Data Mining (SDM '18)*, 2018, pp. 90–98.

[29] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, and M. Zhou, "MIND: A large-scale dataset for news recommendation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*, 2020, pp. 3597–3606.

[30] A. Lommatzsch, B. Kille, F. Hopfgartner, M. Larson, T. Brodt, J. Seiler, and Ö. Özgöbek, "CLEF 2017 NewsREEL overview: A stream-based recommender task for evaluation and education," in *Proceedings of the 8th Conference and Labs of the Evaluation Forum (CLEF '17)*, 2017.

[31] J. A. Gulla, L. Zhang, P. Liu, O. Özgöbek, and X. Su, "The Adressa dataset for news recommendation," in *Proceedings of the 2017 International Conference on Web Intelligence (WI '17)*, 2017, p. 1042–1048.