Artificial Intelligence and Law (2025) 33:43–78
https://doi.org/10.1007/s10506-023-09378-3

ORIGINAL RESEARCH

A large scale benchmark for session-based recommendations on the legal domain

Marcos Aurélio Domingues[1,2,4] · **Edleno Silva de Moura**[2,4] · **Leandro Balby Marinho**[3] · **Altigran da Silva**[2]

Accepted: 28 September 2023 / Published online: 25 October 2023

Abstract

The proliferation of legal documents in various formats and their dispersion across multiple courts present a significant challenge for users seeking precise matches to their information requirements. Despite notable advancements in legal information retrieval systems, research into legal recommender systems remains limited. A plausible factor contributing to this scarcity could be the absence of extensive publicly accessible datasets or benchmarks. While a few studies have emerged in this field, a comprehensive analysis of the distinct attributes of legal data that influence the design of effective legal recommenders is notably absent in the current literature. This paper addresses this gap by initially amassing a comprehensive session-based dataset from Jusbrasil, one of Brazil's largest online legal platforms. Subsequently, we scrutinize and discourse key facets of legal session-based recommendation data, including session duration, types of recommendable legal artifacts, coverage, and popularity. Furthermore, we introduce the first session-based recommendation benchmark tailored to the legal domain, shedding light on the performance and constraints of several renowned session-based recommendation approaches. These evaluations are based on real-world data sourced from Jusbrasil.

Keywords Legal documents recommendation · Session-based recommender systems · Recommender systems · Benchmark · Large scale dataset

1 Introduction

Searching for legal documents online is an integral part of the daily routine for legal students, professionals, and ordinary citizens. Legal search engines offer valuable assistance to lawyers by facilitating tasks like legal research, court decision tracking, and drafting pleadings. Similarly, ordinary citizens seek updates on their cases, while students find practicality in supplementing theoretical learning with real-world case laws. Given that legal documents are distributed across various courts

Extended author information available on the last page of the article

and employ specialized terminology, dedicated search engine platforms tailored to the legal domain have emerged (Sansone and Sperlí 2022).

Brazil's most prominent and widely used legal search engine platform is Jusbrasil.[1] Jusbrasil comprehensively crawls, indexes, and structures legal documents from numerous Brazilian courts, facilitating rapid and efficient legal information retrieval for millions of daily users across the nation. However, with the ever-growing number of legal documents of different types available, search engines are often not enough to mitigate the information overload problem. In such cases, recommender systems appear as an effective alternative since they anticipate users' information needs without requiring explicit user input, i.e., query formulation and submission. Moreover, recommender systems are well-known for improving the user experience when browsing large information spaces.

Despite notable advancements in legal information retrieval systems, the realm of legal recommender systems remains relatively unexplored. One possible reason is the lack of large and publicly available datasets and benchmarks. While some studies have delved into this field, an in-depth analysis of the diverse attributes inherent in legal data that influence the design of effective legal recommender systems, along with a benchmark comparing distinct recommendation models tailored to the legal domain, is notably absent from existing literature.

This paper seeks to address this gap by undertaking several contributions. Firstly, we curate a large-scale dataset of user sessions extracted from Jusbrasil, the largest legal search portal in Brazil. These sessions comprise user interactions with legal documents within predefined timeframes. Our investigation focuses on a session-based recommendation task, wherein we predict subsequent interactions based on the known portion of the ongoing session. We benchmark models suitable for a diverse audience within the Jusbrasil platform, catering to both legal professionals and non-experts alike. These session-based models operate by generating recommendations grounded in ongoing and past sessions, which can be particularly advantageous for users spanning various levels of expertise, leveraging content aligned with their comprehension and interests. Furthermore, given that a majority of these models are personalized, being influenced by user interactions, the resulting recommendations are tailored to the user's profile and specific needs. In our work, we analyze and discuss important features of legal recommendation data such as sparsity, coverage, and popularity. We also shed light on the performance and limitations of several well-known session-based recommendation models on our collected dataset, in contrast to other well-studied session-based recommendation domains such as music, e-commerce, news, and job postings.

The paper is organized as follows. Background concepts are presented in Sect. 2. Section 3 introduces the types of legal documents available at Jusbrasil. Section 4 introduces the experimental dataset collected from Jusbrasil. Section 5 describes our research methodology in more detail with respect to the compared algorithms, the evaluation protocol, the performance measures, and the results. In Sect. 6, we discuss relevant previous researches. Section 7 presents some limitations of our work. Finally, in Sect. 8 we conclude the paper and highlight future work directions.

[1] https://www.jusbrasil.com.br.

2 Preliminaries

A recommender system is an information filtering technology for recommending items that may be of interest to users (Ricci et al. 2011). Item is the general term employed to denote what the system recommends to users, like songs, products, and documents. In our case, an item is a legal document.

There are many different recommendation tasks according to the recommendation's purpose and the dataset's characteristics. In the following, we briefly describe the most popular ones.

Rating Prediction This is a well-known recommendation task (Ricci et al. 2011) that aims to predict user ratings for items that have not been rated yet by the user. The predicted ratings are computed from users' explicit feedback, i.e., ratings provided on some items in the past.
Item Recommendation This task, also known as top-*n* recommendations, consists of selecting the most relevant items for a user from an extensive catalog of items (Deshpande and Karypis 2004; Rendle 2022). Instead of leveraging explicit ratings, such as in the previous task, the recommendation models use implicit feedback, which consists of past user-item interactions, such as listening to a song, purchasing a product, and reading a document.
Sequential Recommendation In this task, the recommender models leverage the whole sequences of recorded user-item interactions to recommend the immediate next item or all subsequent items for the user (Wang et al. 2019).
Session-based Recommendation Here, the recommendation models are built from logs of recorded user-item interactions, where the interactions are grouped into anonymous sessions. A session is a sequence of interactions with a clear boundary (e.g., a shopping session on an e-commerce site). Then, given the interactions within the current session, the aim is to recommend the next item, the rest of the session, or even complete new sessions (Wang et al. 2021; Jannach et al. 2022).
Session-aware Recommendation This task is similar to the session-based one, given that the user-item interactions are also grouped into sessions. However, the users are not anonymous, and the models can leverage information about past user sessions when recommending the next item(s) or session(s) (Latifi et al. 2021).

Jusbrasil supports both logged and unlogged (i.e., anonymous) users. Thus, to consider both kinds of users, in this paper, we consider only the session-based recommendation task, which is currently a prevalent task in the recommender systems research community (Wang et al. 2021).

Session-based recommender systems involve both input and training data comprising user-item interactions grouped within specific time boundaries, referred to as sessions (Wang et al. 2021; Jannach et al. 2022). Each session may encompass activities such as a music listening session, an e-commerce shopping spree, or a legal document reading session within a search portal. Figure 1 illustrates the process of session-based recommendation. Each sequence of blocks constitutes an anonymous session, where green or blue blocks denote interactions (e.g., clicks, views, listens,

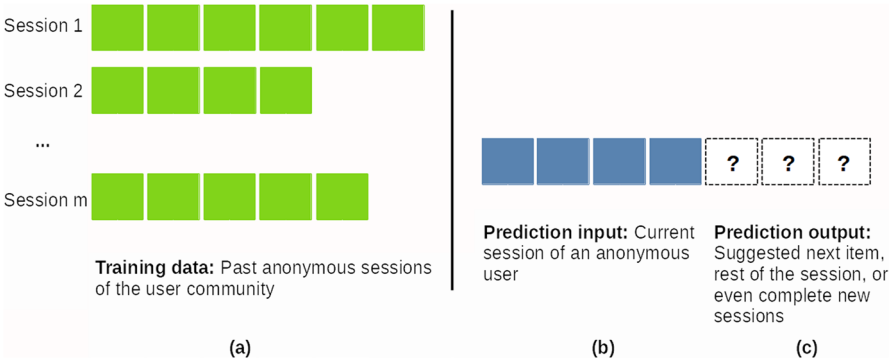


Fig. 1 Session-based recommendation process. Adapted from Latifi et al. (2021)

or purchases) between users and items within the session. White blocks symbolize recommendations. In this context, prior anonymous sessions (Fig. 1a) serve as training data for session-based recommendation models. Subsequently, based on the interactions within the ongoing session (Fig. 1b), the model suggests the next item, additional session content, or even entirely new sessions (Fig. 1c).

Similarly to previous work (Wang et al. 2021; Quadrana et al. 2018), we formally define a session-based recommender system as follows. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items, and $S = \{s_1, s_2, \ldots, s_m\}$ a set of historical anonymous sessions, where each session s is a sequence of interactions on items in I. Let $L = \{l_1, l_2, \ldots, l_p\}$ be a set of recommendable lists, where each list l is an element of the set of all permutations up to length k of the power set of I, i.e. $l \in \mathbb{S}_k(\mathbb{P}(I))$. Finally, let $f_S : S \times L \to \mathbb{R}$ be a function that takes a session $s \in S$ and a recommendation list $l \in L$ as input and returns a utility score denoting the relevance of l for s.

Now, a session-based recommender system selects the recommendation list $l'_s \in L$ that maximizes the utility score for the current session s. More formally,

$$l'_s = \underset{l \in L}{\operatorname{argmax}}\ f_S(s, l). \tag{1}$$

Session-based recommender systems is an active area of research with a vast number and variety of models available. In Sect. 5.1, we present and justify the choice of the models used in our experiments.

3 Jusbrasil's legal content

In this section, we provide more context about the types and purposes of legal information available at Jusbrasil.

- *Precedents* a set of decisions from Brazilian courts, including all the regional and federal general courts in Brazil and also specialized courts (the military courts, the labor issues courts, and the electoral courts). Users of Jusbrasil may

access precedent documents for several distinct purposes. The two major ones are understanding a legal issue by studying judicial decisions; and searching for precedents when writing their legal pleadings, seeking previous decisions which may support their requests.
- *Articles* specialized articles commenting on legal topics, which usually review legal cases, comment on court decisions, interpret laws, or explain in detail important legal concepts. Besides their informative and educational purposes, articles may also be used as a source of information to write pleadings.
- *News* news related to legal subjects, such as important legal cases, decisions from the Brazilian supreme court, and so on.
- *Models and Pleadings* sample documents to help lawyers in writing their own legal documents, including legal instruments such as contracts and wills, pleadings, and motions with examples of how to write persuasive requests to advocate in favor of a legal position. By reading pleadings, users can also study a legal issue and understand a legal rationale formulated by other lawyers.
- *Doctrines* documents containing doctrines followed in each field of law. In general, doctrines describe rules or principles largely followed in the law.

When searching for information, Jusbrasil's users may request copies of the content available on the platform or just view their content. Those interactions are logged and used to create the dataset we describe in the following section.

4 The JusBrasilRec collection

With millions of user interactions to billions of documents covering a wide range of different artifacts related to law in Brazil, Jusbrasil arises as a promising large-scale test bed for advancing research on the still scarce area of legal recommender systems. We first collect a large sample of session data.[2] Next, we characterize this dataset in terms of crucial session-based recommendation properties. Finally, we compare several session-based recommendation models on this dataset, and analyze their performances concerning a content-based model and also other session-based recommendation domains.

4.1 Dataset

We gathered a 30-day dataset from Jusbrasil spanning between 2021-02-23 and 2021-03-24. This specific duration was meticulously chosen to ensure the absence of recommendation bias, as it encompasses the final 30 days preceding Jusbrasil's inaugural recommender system launch. Consequently, the items users engaged with-some of which will serve testing purposes-were not influenced by any recommender system. This meticulous approach guarantees the fidelity and impartiality of our (large-scale) sample. Guided by the expertise of Jusbrasil's

[2] Available for download here: https://zenodo.org/record/8401278.

```
session_id;user_id;user_type;doc_id;doc_type;timestamp
5246659;931737;logged;21075;articles;2021-03-11T20:41:26.156-03:00
5246659;931737;logged;2708666;precedents;2021-03-11T20:51:41.120-03:00
5246659;931737;logged;2711326;precedents;2021-03-11T20:55:08.363-03:00
5246836;931781;unlogged;4821696;news;2021-03-11T11:31:55.895-03:00
5246836;931781;unlogged;4798653;news;2021-03-11T11:40:04.105-03:00
5246836;931781;unlogged;4880210;news;2021-03-11T11:40:20.909-03:00
5246840;931781;unlogged;60661;articles;2021-03-23T11:26:15.195-03:00
5246840;931781;unlogged;4813493;news;2021-03-23T11:55:00.665-03:00
5246799;931772;logged;4344438;precedents;2021-03-01T13:10:23.705-03:00
5246799;931772;logged;2228117;precedents;2021-03-01T13:10:38.374-03:00
```

Fig. 2 Example of the JusBrasilRec dataset

Table 1 Data statistics

Datasets	#Interactions	#Users	#Items	#Sessions
jusbrasilrec_all_users	22,442,232	2,310,247	4,225,874	5,415,623
jusbrasilrec_logged_users	13,308,981	673,580	3,083,495	2,787,995
jusbrasilrec_unlogged_users	9,133,251	1,636,667	1,862,639	2,627,628

developers and legal professionals, we methodically retrieved the data from the company's database, subjecting it to thorough validation and anonymization.

Within this dataset, every interaction encompasses: **session_id** [session identifier established for each 30-min interval of user inactivity (Cooley et al. 1999)], **user_id** (user identifier), **user_type** (user classification, i.e., logged or unlogged), **doc_id** (legal document identifier), **doc_type** (document classification, i.e., Precedents, Articles, News, Models and Pleadings, and Doctrines), and **timestamp** (date and hour of the user-document interaction). An excerpt of the dataset is showcased in Fig. 2, where each line denotes a distinct interaction.

We analyzed this data in three different contexts:

- **jusbrasilrec_all_users:** data from all users;
- **jusbrasilrec_logged_users:** data from logged users only;
- **jusbrasilrec_unlogged_users:** data from unlogged users only.

Table 1 and Fig. 3 present some basic statistics of these datasets. In the following, we discuss some characteristics of this data with respect to session-based recommendations.

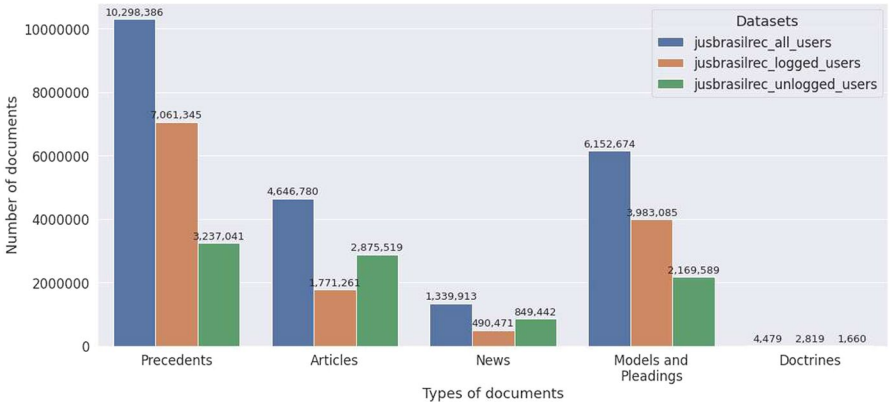


Fig. 3 Number of documents in JusBrasilRec

4.2 Session duration and size

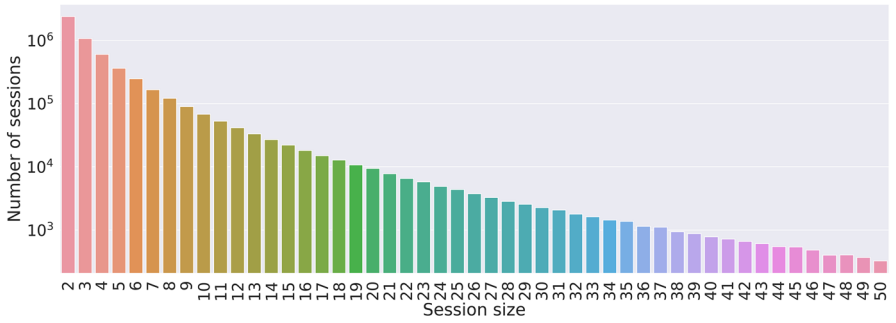
In Fig. 4, we show the session size distribution for the three datasets. Each dataset has a minimum of 2, and a maximum of 50 interactions per session, and between 80 and 90% of the sessions have no more than six interactions (cf. Fig. 5). Small session scenarios like this are challenging since the models will have limited information to train upon (Wang et al. 2021).

We consider the timestamp difference between the first and the last interactions in a session as the duration of the respective session. Figure 6 presents some summary statistics regarding session duration. Notice that session duration range from less than 1–345 min. Additionally, we have around 90% of the sessions with no more than 40 min (i.e., 92.85% for jusbrasilrec_all_users, 90.69% for jusbrasilrec_logged_users, and 95.17% for jusbrasilrec_unlogged_users).

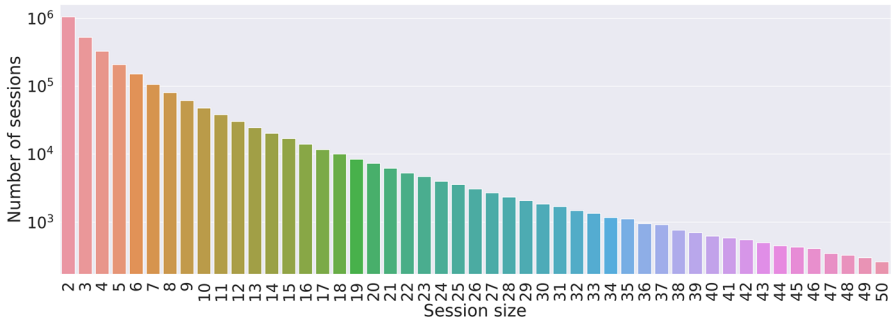
4.3 Number of sessions

Figure 7 depicts the distributions related to the number of sessions per user. For each dataset, we have at least one session per user. Additionally, we have 79.18% of the users in the jusbrasilrec_all_users dataset, 57.44% of the users in the jusbrasilrec_logged_users, and 88.13% of the users in the jusbrasilrec_unlogged_users dataset with at most two sessions. The maximum number of sessions per user is 172.

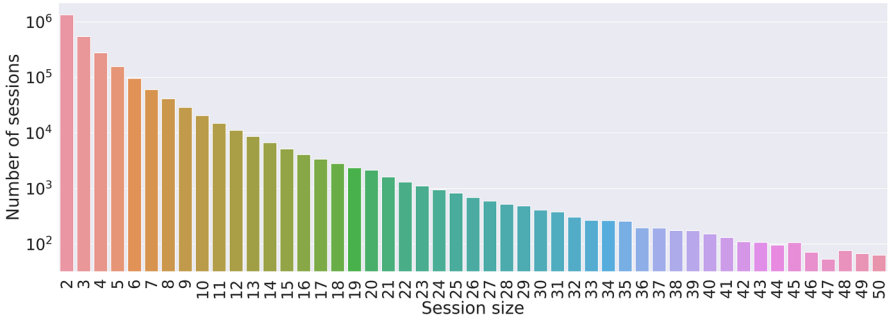
Although the collected data contains user information, we disregarded it and used the standard rule of creating a session after a defined user idle time of 30 min of inactivity (Ludewig and Jannach 2018). Thus, the sessions can be treated as anonymous, a fundamental assumption of session-based recommender systems.



(a) jusbrasilrec_all_users



(b) jusbrasilrec_logged_users



(c) jusbrasilrec_unlogged_users

Fig. 4 Sessions size distribution

5 Empirical evaluation

This section assesses the performance of 21 well-known session-based recommendation models applied to our compiled dataset. The models range from simple most-popular baselines to sophisticated deep learning-based models. While these models only leverage user-item interaction data, we also include a classic Bag of Words (BoW) model that uses the textual content of the documents. The outcomes of this assessment will facilitate the identification of the most promising strategies tailored

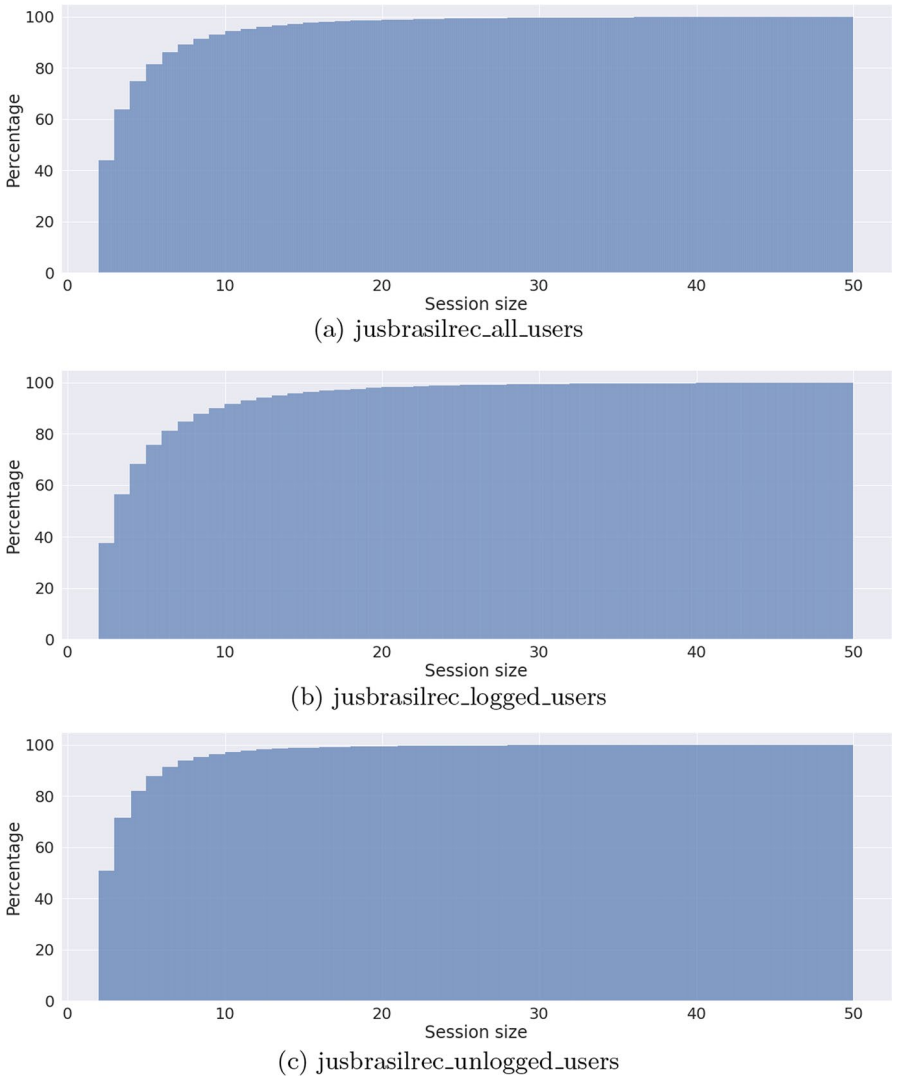


Fig. 5 Sessions size cumulative distribution

to our distinct dataset and domain, thereby providing valuable guidance to researchers and developers. This insight informs the optimal pathways for devising novel models or selecting apt existing ones as the foundation for practical implementation and deployment.

Additionally, we evaluate the performance of these recommendation models on other application domains (i.e., news, music, e-commerce, and job posting) to assess their commonalities and differences with our legal domain dataset. The following subsections introduce the recommendation models, evaluation protocol, datasets, and results.

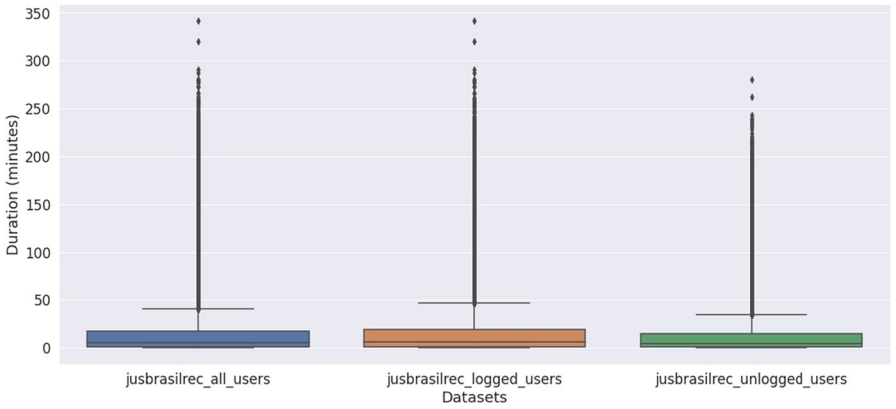


Fig. 6 Session duration

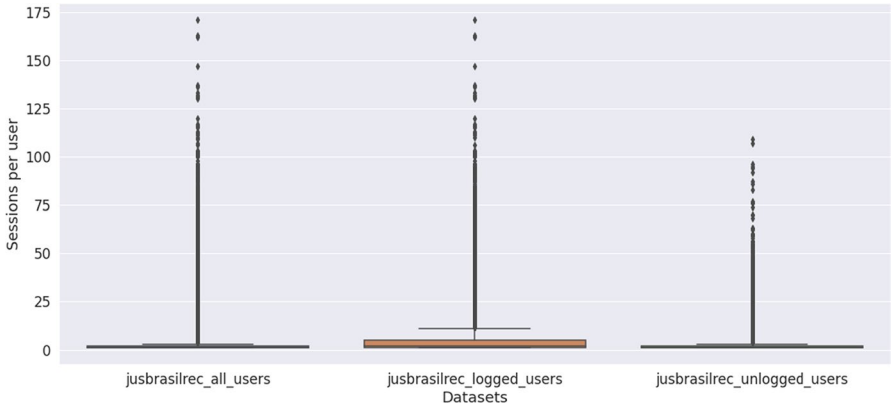


Fig. 7 Number of sessions per user

5.1 Session-based recommendation models

This subsection provides a succinct overview of the session-based recommendation models employed in our empirical evaluation. These models are publicly accessible via the *session-rec* recommendation framework,[3] where comprehensive implementation details can be found. The models underwent training with the default parameter configurations within the framework. Consequently, the outcomes presented here have the potential for enhancement through meticulous fine-tuning-a pursuit we defer to future work.

The models are categorized into the following five groups based on their underlying principles: non-personalized, pattern mining, nearest neighbors, matrix factorization, and neural networks.

[3] https://github.com/rn5l/session-rec.

5.1.1 Non-personalized models

A non-personalized recommendation model disregards users' consumption or interaction history, essentially omitting personal preferences. A good example of this category involves recommendations based on item popularity, wherein an identical recommendation list is offered regardless of individual tastes. In this work, we evaluate the efficacy of the subsequent non-personalized models:

- *random* A randomized list of items is generated each time recommendations are requested. This is the simplest baseline and is often used as a lower bound for model performance;
- *pop* Items are ranked in descending order of their frequency of occurrences (count) in the dataset;
- *rpop* Similar to the *pop* approach described above, but exclusively utilizing interactions from the last n days. For this study, $n = 1$ is employed to compute popularity solely for the current day;
- *spop* The Session Popularity (*spop*) model assigns higher scores to items with higher occurrence counts within the session. In cases of ties, the global popularity score of the item is factored in.

5.1.2 Models based on pattern mining

This category of recommendation model exploits item co-occurrence patterns. We have included the following models in our evaluation:

- *ar* This model is a simplified version of the Association Rules (*ar*) mining technique (Agrawal et al. 1993) with a maximal rule size of two items. In essence, for each rule $i \rightarrow j$ (indicating that interacting with item i suggests recommending item j), its score is calculated based on the frequency of co-occurrences of items i and j within the sessions dataset. Recommendations are derived by presenting items with the highest co-occurrence counts alongside the last item of the ongoing session (Ludewig and Jannach 2018). In this paper, we regard this model as a surrogate for Jusbrasil's recommender system. The current system is rooted in item co-occurrence counts, wherein the next item most frequently co-occurring with the presently visited item is suggested;
- *markov* A derivative of *ar*, the Markov Chain model extracts rules $i \rightarrow j$ from a first-order Markov Chain (Norris 1997). The model counts occurrences where users engage with an item j immediately after interacting with an item i within the training sessions;
- *sr* The Sequential Rules (*sr*) model, proposed in Kamehkhosh et al. (2017), is another variant of *ar* that quantifies pairwise item co-occurrences within the training sessions. However, to evaluate rules $i \rightarrow j$, *sr* accounts for the sequence of items within a session by employing a decay function to penalize inter-session distances between items.

5.1.3 Nearest neighbors models

Despite their simplicity, nearest neighbors-based recommendation models often perform surprisingly well (Kamehkhosh et al. 2017; Verstrepen and Goethals 2014; Jannach and Ludewig 2017). We have included the following five nearest neighbors-based variants in our benchmark:

- *iknn* The Item *k*-nearest Neighbors (*iknn*) model considers the last element in a given session as input and returns as recommendations the items most similar to it in terms of their co-occurrence in other sessions (Hidasi et al. 2016). Each item is encoded as a binary vector, where each element corresponds to a session and is set to 1 in case the item appeared in the session and 0 otherwise. The similarity of two items can be then determined, e.g., using the cosine similarity where the number of neighbors *k* is determined empirically;
- *sknn* Similar to *iknn*, but the Session-based KNN (*sknn*) model compares the entire ongoing session with the historical sessions in the training data to determine the items to be recommended (Jannach and Ludewig 2017). In this case, each session is encoded as a binary vector, where each element corresponds to an item and is set to 1 in case the item appeared in the session and 0 otherwise. Then, the model determines the *k* most similar past sessions (neighbors) by applying a suitable session similarity measure, e.g. the cosine similarity or the Jaccard coefficient. Finally, it recommends items from the top-*n* most similar sessions;
- *vsknn* The Vector Multiplication Session-based KNN (*vsknn*) model is a variant of *sknn* that puts more weight on the more recent interactions of the current session when computing the session similarities (Ludewig and Jannach 2018). The model encodes the current session as a real-valued vector instead of a binary vector as described previously. Thus, only the last item of the session obtains a value of 1; the weights of the other items are determined using a linear decay function that depends on the position of the item within the session, where items appearing earlier in the session get a lower weight. As a result, when using the dot product as a similarity function between the current weight-encoded session and a binary-encoded past session, more emphasis is given to items that appear later in the session;
- *stan* This model, called Sequence and Time-aware Neighborhood (*stan*), was proposed in Garg et al. (2019) and is based on the *sknn* model. It considers the following three factors for making recommendations: (1) the position of an item in the current session, (2) the recency of a past session concerning the current session, and (3) the position of a recommendable item in a neighboring session. Decay functions are used to put more emphasis on the more recent items;
- *vstan:* This model, which was proposed in Ludewig et al. (2021), combines the ideas from *stan* and *vsknn* in a single approach. It incorporates all the three previously described particularities of *stan*, which already share some similarities with the *vsknn* model. Furthermore, it adds a sequence-aware item scoring procedure and an IDF weighting scheme that gives lower weights to very frequent items in the dataset.

5.1.4 Factorization models

Factorization models have consistently shown to be highly effective in various recommendation tasks since the Netflix Prize was introduced (Koren et al. 2009). In these models, users and items are mapped onto a shared embedding space where their affinity for each other is modeled through dot products between their corresponding embedding vectors. Our experiments on session-based recommender systems have included several variants of factorization models from recent literature (Ludewig and Jannach 2018) such as:

- *bprmf* Originally, the Bayesian Personalized Ranking for Matrix Factorization (*bprmf*) model was designed for implicit feedback recommendation scenarios. It is usually cast as a matrix-completion problem based on long-term user preferences (Rendle et al. 2009). To use *bprmf* for session-based recommendations, where there are no long-term user profiles, the model is adapted to consider each session in the training set as a different user, i.e., each session corresponds to a user in the user-item interaction matrix. At the prediction time, the user vector is modeled as the average of the latent item vectors of the current session;
- *fpmc* The Factorized Personalized Markov Chains (*fpmc*) model was designed for the specific problem of next-basket recommendation (Rendle et al. 2010). It consists of predicting the entire next session (or basket), given the history of past sessions. To be used in session-based to recommend the next item, we need to limit the size of the recommended basket to one item and consider the current session as the history of baskets. Technically, *fpmc* combines factorized markov chain and traditional user-item matrix factorization in a three dimensional tensor factorization approach, where the third dimension captures the transition probabilities from one item to another. Similar to *bprmf*, the session latent vectors are estimated as the average of the latent factors of the individual items in the session (Rendle et al. 2010);
- *fism* The Factored Item Similarity (*fism*) model is based on an item-item factorization approach, which has the advantage of being directly applicable to the session-based recommendation task (Kabbur et al. 2013). The model does not incorporate sequential item-to-item transitions like *fpmc* does. Instead, it learns the item-item similarity matrix as a product of two low-dimensional latent factor matrices. This factored representation allows *fism* to capture and learn relations between items (Kabbur et al. 2013). It also has the advantage over *bprmf* that it does not build a representation for each user. This enables lightweight models and the ability to make recommendations to new users;
- *fossil* The Factorized Sequential Prediction with Item Similarity (*fossil*) model leverages both long-term (matrix factorization) and short-term (factorized markov chains) user preferences in a session-based recommendation model (He and McAuley 2016). To do that, it combines *fism* with *fpmc* to incorporate sequential information into the model. Unlike the previous *fpmc*, *fossil* can use higher-order markov chains;
- *smf* Similar to fossil, the Session-based Matrix Factorization (*smf*) model combines *fpmc* with standard matrix factorization (Ludewig and Jannach 2018). In

addition, the model also considers the cold-start situation of session-based recommendation scenarios as follows. In contrast to the traditional factorization-based prediction model, *smf* replaces the latent user vector with a session preference vector, which is computed as an embedding of the current session.

5.1.5 Neural network models

Models based on deep neural network techniques represent the most recent family of techniques for session-based recommender systems. In our experiments, we have included the following four different deep neural network-based models:

- *gru4rec* This model was the first neural approach for session-based recommendations. The model uses Gated Recurrent Units to deal with the vanishing gradient problem, in order to predict the probability of the subsequent interactions given the current session (Hidasi et al. 2016);
- *narm* The Neural Attentive Recommendation Machine (*narm*) is an extension of the previous *gru4rec* that improves the session modeling with the introduction of a hybrid encoder with an attention mechanism. The attention mechanism considers both the items that appeared earlier in the session and which are similar to the last accessed one. The score for each candidate recommendation is computed with a bilinear matching scheme based on the unified session representation (Li et al. 2017);
- *stamp* Different from *narm*, this model does not rely on a Recurrent Neural Networks. The model uses a short-term attention/memory priority approach, which is capable of capturing the users' general interests from the long-term memory of a session; and also takes the users' most recent interest from the short-term memory into account. Technically, the users' general interests are captured by an external memory built from all the historical interactions in a session (including the last interaction). The last interaction is also considered as short-term memory of the users' interests. The attention mechanism is built on top of the embedding of the last interaction that represents the user's current interest (Liu et al. 2018);
- *sgnn* In the session-based with graph neural network model, session sequences are modeled as directed graphs. Based on the session graph, the model is capable of capturing transitions of items and generating item embedding vectors. With the help of item embedding vectors, the model builds reliable session representations from which the next item can be predicted (Wu et al. 2019).

5.2 Evaluation protocol and performance measures

In general, the main task of a session-based recommender system is to generate a ranked list of items based on a given session. According to Ludewig and Jannach (2018), one can measure the efficacy of session-based recommenders offline by evaluating the model's ability to predict the withheld entries of a session. Various approaches in the literature exist for withholding certain session entries. One involves withholding all items and sequentially revealing each item, mirroring the

user's progression throughout a session (Ludewig and Jannach 2018). Another method involves withholding items and progressively unveiling subsequent items in a session from a specific juncture, predicting items until the session conclusion, and assisting the users in knowing when they have adequately researched a particular item or when all options of interest have been exhausted (Ludewig and Jannach 2018).

Building on the work of Ludewig and Jannach (2018), we incorporate both previously discussed evaluation scenarios into our benchmark. In the next item scenario, the goal is to predict the subsequent item the target user will engage with, using the first n items of the current session. For each session, we progressively increase n, evaluate the Hit Rate and Mean Reciprocal Rank (MRR), and then determine average values across sessions for varying recommendation list lengths. The Hit Rate represents the ratio of session iterations where the next item is part of the recommendation list. MRR calculates the average inverse rank-defined as one divided by the position of the next item in the recommendation list-across all session iterations. Consequently, the Hit Rate determines if the ground-truth item is on the recommendation list, while MRR examines its exact position.

We also analyze the rest of the session scenario. Here, with the objective of forecasting every subsequent interaction up to the session's conclusion, recognized information retrieval metrics such as Precision, Recall, Normalized Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP) are employed. Similar to the next item scenario, we gradually increase the number of starting items in the session and compute the average of these metrics over subsequent iterations. Precision and Recall check the presence of ground-truth items in recommendation lists, whereas MAP and NDCG focus on their ranks within those lists. It is a common practice in the recommender systems literature to consider 10 or fewer recommended items. This reflects commercial recommender systems that present only a few recommended items to users to alleviate the cognitive burden of processing the recommendation list. Hence, Metric@10 means that Metric is computed over recommendation lists of size 10.

Besides the accuracy measurements, we also consider the following metrics (Jannach and Ludewig 2017): Coverage and Popularity bias. Coverage is defined as the proportion of different items in the catalog that ever appear in the recommendation lists. Popularity bias can be used to measure if high accuracy values are correlated with the tendency of a model to recommend popular items (Jannach et al. 2015). We report the average popularity score for the items in the recommendation lists of each model, where the average score is the mean of the individual popularity scores of each recommended item. The individual scores are computed by counting how often each item appears in all training sessions, and then by applying min–max normalization to obtain a score between 0 and 1. By computing both coverage and popularity bias measures, it is possible to emphasize that different recommendation models can lead to quite different recommendations, even if they are similar in terms of prediction accuracy (Jannach et al. 2015).

In the experiments, the models were trained from the ground up by applying a sliding window protocol, i.e., sessions $s_1, s_2, \ldots, s_m$ are split into several slices of

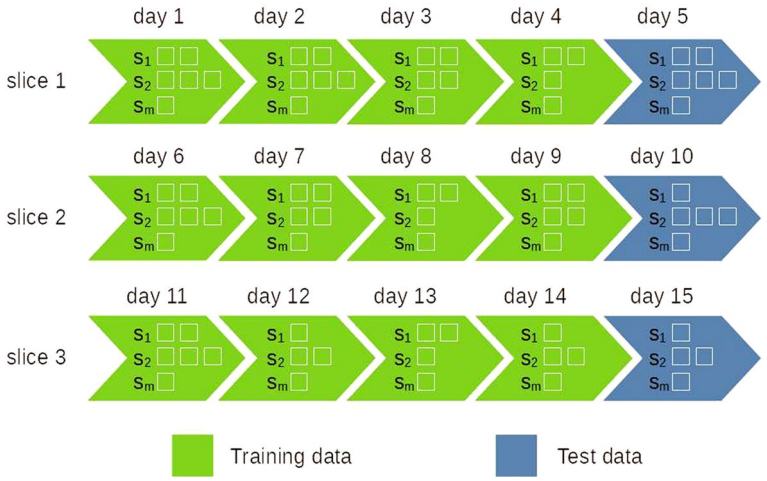


Fig. 8 Slice window protocol applied to 15 days of sessions

equal size in days to train and test the models (Ludewig and Jannach 2018). Figure 8 illustrates the protocol applied to 15 sessions to generate three slices of 5 days, four days for training, and one day for testing. This protocol allows us to evaluate the variability of models' performances by considering distinct data splits (i.e., slices) while preserving the chronological order of user-item interactions.

Since all three JusBrasilRec datasets contain 30 days of data, we decided to split each dataset into 5 slices of 6 days, with 5 days of data for training and one for testing. In addition, we adopt the common practice of terminating a session after 30 min of user inactivity (Cooley et al. 1999).

5.3 Other application domains datasets

We evaluated the same models on other datasets from four different application domains (i.e., news, music, e-commerce, and job posting).[4] By comparing the models' performances across these different datasets coming from other application domains, we can better identify which types of models are more particularly suitable for our legal dataset. In the following, we present a brief description of each dataset:

- *clef* This dataset was used in the 2017 CLEF NewsREEL Challenge.[5] It consists of a stream of news articles, which were collected by the company plista[6] from several publishers. We split it into 5 slices of 5 days, where we have 4 days of training data and 1 for testing;

[4] Available for download here: https://drive.google.com/drive/folders/1ritDnO_Zc6DFEU6UND9C8VCisT0ETVp5.

[5] http://www.clef-newsreel.org.

[6] https://www.plista.com.

Table 2 Datasets statistic

Statistics	Clef	Nowplaying	Retailrocket	Xing
Number of interactions	27,274,333	1,446,189	1,318,753	7,173,700
Number of items	1,597	249,183	132,923	969,349
Number of sessions	8,063,533	153,086	382,084	1,220,790
Interactions per session (min.)	2	2	2	2
Interactions per session (mean)	3.38	9.44	3.45	5.87
Interactions per session (max.)	50	50	50	50
Timespan in days	29	536	138	81

- *nowplaying* This dataset was created from music-related tweets, where users posted which music tracks they were currently listening to. The dataset was published by Zangerle et al. (2014). The dataset was split into 5 slices of 107 days, where we have 77 days of training and 30 days for testing;
- *retailrocket* This dataset was published by the e-commerce personalization company retailrocket for Kaggle competition[7] and contains user browsing activities. For this dataset, we split it into 5 slices of 27 days, where we have 20 days of training and 7 days for testing;
- *xing* This dataset was made available for the ACM RecSys Challenge 2016[8] and contains interactions of users with job postings. The dataset was split into 5 slices of 16 days, with 11 days of training and 5 for testing.

Notice that we preprocessed the datasets mentioned above in the same way we did for the JusBrasilRec dataset, i.e., sessions with only one interaction were removed, as well as items that appeared less than five times in the datasets and sessions with more than 50 interactions. Finally, we split the datasets into five slices with 70–80% of data for training and 30–20% for testing. Some basic statistics about the datasets are presented in Table 2.

5.4 Results

In this subsection, we present and discuss the benchmark results on the JusBrasilRec collection. We consider both the *next item* and *rest of the session* evaluation scenarios. We also detail a real recommendation example drawn from the best model. Since the session-based models yield similar performances across the three datasets, we present and discuss the results only for jusbrasilrec_all_users dataset. The results for jusbrasilrec_logged_users and jusbrasilrec_unlogged_users are presented in Appendix A. After discussing the performance of the session-based models, we compare them against a content-based recommendation model. Finally, we compare

[7] https://www.kaggle.com/retailrocket/ecommerce-dataset.

[8] https://github.com/recsyschallenge/2016.

Table 3 Hit rate, MRR, coverage, and popularity bias for top-10 recommendation lists considering the *next item* evaluation scenario in the JusBrasilRec dataset

Models	HitRate@10	MRR@10	Coverage@10	Popularity@10
random	0.000 ± 0.000	0.000 ± 0.000	**0.996 ± 0.006**	0.001 ± 0.000
pop	0.015 ± 0.001	0.008 ± 0.000	0.000 ± 0.000	**0.359 ± 0.038**
rpop	0.016 ± 0.001	0.008 ± 0.000	0.000 ± 0.000	0.335 ± 0.023
spop	0.599 ± 0.005	0.533 ± 0.006	0.199 ± 0.075	0.320 ± 0.034
ar	0.680 ± 0.004	0.540 ± 0.005	0.416 ± 0.110	0.022 ± 0.002
markov	0.664 ± 0.004	0.556 ± 0.005	0.323 ± 0.083	0.019 ± 0.002
sr	0.676 ± 0.004	0.553 ± 0.005	0.373 ± 0.096	0.020 ± 0.002
iknn	0.175 ± 0.005	0.108 ± 0.004	0.442 ± 0.097	0.009 ± 0.001
sknn	0.718 ± 0.002	0.548 ± 0.006	0.411 ± 0.121	0.019 ± 0.001
vsknn	0.748 ± 0.003	0.585 ± 0.006	0.419 ± 0.122	0.024 ± 0.002
stan	**0.751 ± 0.003**	0.601 ± 0.005	0.419 ± 0.121	0.023 ± 0.002
vstan	0.743 ± 0.003	**0.611 ± 0.005**	0.448 ± 0.121	0.018 ± 0.001
bprmf	0.555 ± 0.006	0.514 ± 0.006	0.685 ± 0.116	0.010 ± 0.001
fpmc	0.525 ± 0.011	0.500 ± 0.010	0.795 ± 0.112	0.005 ± 0.000
fism	0.119 ± 0.021	0.081 ± 0.015	0.681 ± 0.107	0.003 ± 0.000
fossil	0.008 ± 0.004	0.003 ± 0.002	0.648 ± 0.117	0.074 ± 0.068
smf	0.310 ± 0.010	0.220 ± 0.006	0.012 ± 0.001	0.135 ± 0.012
gru4rec	0.679 ± 0.004	0.536 ± 0.005	0.554 ± 0.114	0.010 ± 0.001
narm	0.732 ± 0.003	0.579 ± 0.005	0.526 ± 0.135	0.023 ± 0.002
stamp	0.698 ± 0.006	0.544 ± 0.008	0.415 ± 0.114	0.022 ± 0.002
sgnn	0.651 ± 0.005	0.519 ± 0.005	0.211 ± 0.047	0.029 ± 0.002

The best results are highlighted in boldface

the models' performances on the jusbrasilrec_all_users dataset to their performances on the other application domains datasets. From now on, we will refer to the jusbrasilrec_all_users dataset as JusBrasilRec.

5.4.1 Next-item recommendation

Table 3 shows the results for JusBrasilRec considering top-10 recommendation lists. In the sequel, we discuss these results according to the five categories of session-based recommender systems (see Sect. 5.1).

Non-personalized models
This category of recommendation approaches yields the worst results overall. The only exception was the *spop* model, yielding results on par with more sophisticated models, such as *bprmf*, for example, in terms of HitRate and MRR. A crucial difference between *spop* and the other models of the same type is the use of session data. The models in this category also provided the highest Popularity scores, which is

not surprising since the items are scored based on occurrence counts. Finally, the *random* model had the worst performance, although it has the highest Coverage as expected, since it has an equal probability of selecting any item from the available pool of candidate items.

Models based on pattern mining
This category of models proved to be very competitive yielding results on par with the more sophisticated models based on deep learning, which is the second-best-performing category of models. These results place this category of models as a good compromise between model complexity, quality of recommendations, and resilience to short sessions. Notice that the current recommender system adopted by Jusbrasil is included in this family of models.

Nearest neighbors models
In this category, except for the *iknn* model that only exploits the last item of the target session, we have the highest accuracy models for our legal dataset. In addition, the models that leverage recency information yield the highest accuracy-related values.

Factorization models
The performance of the models in this category was not consistent. While *bprmf* and *fpmc* performed reasonably well, *fism*, *smf*, and *fossil* performed poorly, sometimes in the same performance level as the worst non-personalized models. The poor performance of the models in this category is probably explained by the fact that factorization models do not perform well under severe sparsity scenarios, which, given the prevalence of short sessions, is the case here.

Neural network models
Finally, this is the second-best-performing category of models. It is worth mentioning that although these models have performed well, they are way more time-consuming to train than nearest neighbors-based models and demand expensive specialized acceleration hardware such as GPUs to make the training feasible.

We also measured the coverage and popularity bias of the models. Except for the popularity-based models, the models present good coverage (varying roughly between 40 and 80%), given the large size of the item catalog. In terms of popularity bias, the values are low for most models (i.e., less than 5%), which means that popular items do not seem to influence users' selection of items and consequently do not influence the recommendation models.

5.4.2 Recommending the rest of the session

The results of the *rest of the session* evaluation scenario are presented in Table 4.

The first observation is that the models' performances drop drastically in this scenario, revealing their inability to predict interactions ahead of the next item. This is likely a consequence of the short sessions, a prevalent characteristic of our

Table 4 Precision, recall, NDCG, and MAP for a top-10 obtained for the *rest of the session* evaluation scenario in the JusBrasilRec dataset

Models	Precision@10	Recall@10	NDCG@10	MAP@10
random	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
pop	0.002 ± 0.000	0.011 ± 0.000	0.008 ± 0.000	0.001 ± 0.000
rpop	0.002 ± 0.000	0.011 ± 0.001	0.008 ± 0.000	0.001 ± 0.000
spop	0.072 ± 0.000	0.381 ± 0.002	0.394 ± 0.003	0.039 ± 0.000
ar	0.088 ± 0.000	0.451 ± 0.003	0.429 ± 0.003	0.049 ± 0.000
markov	0.084 ± 0.000	0.437 ± 0.003	0.429 ± 0.003	0.047 ± 0.000
sr	0.087 ± 0.000	0.447 ± 0.003	0.433 ± 0.003	0.048 ± 0.000
iknn	0.032 ± 0.001	0.137 ± 0.004	0.118 ± 0.003	0.015 ± 0.000
sknn	0.094 ± 0.000	0.479 ± 0.003	0.455 ± 0.002	0.052 ± 0.000
vsknn	**0.099 ± 0.000**	0.501 ± 0.004	0.475 ± 0.003	**0.055 ± 0.000**
stan	**0.099 ± 0.000**	**0.502 ± 0.004**	**0.481 ± 0.003**	**0.055 ± 0.000**
vstan	0.097 ± 0.000	0.494 ± 0.004	0.479 ± 0.003	0.054 ± 0.000
bprmf	0.064 ± 0.000	0.353 ± 0.003	0.368 ± 0.003	0.036 ± 0.000
fpmc	0.059 ± 0.001	0.331 ± 0.005	0.344 ± 0.006	0.033 ± 0.001
fism	0.013 ± 0.002	0.075 ± 0.013	0.065 ± 0.012	0.007 ± 0.001
fossil	0.001 ± 0.000	0.006 ± 0.003	0.004 ± 0.002	0.001 ± 0.000
smf	0.042 ± 0.001	0.208 ± 0.007	0.190 ± 0.005	0.022 ± 0.001
gru4rec	0.087 ± 0.000	0.448 ± 0.003	0.427 ± 0.002	0.048 ± 0.000
narm	0.096 ± 0.000	0.488 ± 0.004	0.466 ± 0.003	0.053 ± 0.000
stamp	0.090 ± 0.000	0.462 ± 0.005	0.439 ± 0.005	0.050 ± 0.001
sgnn	0.087 ± 0.001	0.436 ± 0.003	0.415 ± 0.002	0.047 ± 0.000

The best results are highlighted in boldface

datasets. Proportionally speaking, the overall performances follow the same trends as observed in the *next item* evaluation scenario.

5.4.3 Recommendations example

To highlight the efficacy and personalization capabilities of the best recommendation model, we detail a real example drawn from our legal dataset. In this example, we consider the profile of a law student keen on delving into the intricacies of the Maria da Penha Law and its varied applications.

The Maria da Penha Law was introduced in Brazil in 2006 to address the issue of domestic violence against women. This law has played a crucial role in providing greater legal protection for women who suffer from domestic violence and has become an important topic of study and discussion for legal professionals and students.

The law student's engagement and interest are evident from the spectrum of documents-**Precedents**, **News**, and **Articles**-she examined during her session:

Fig. 9 Example of the TF-IDF matrix

```
doc_id;word_id;tf_idf
24265;267634;0.0222561502549972
24265;25914;0.0218825896196252
24265;375761;0.0219427252053909
24265;164042;0.0373113681022015
24265;29865;0.0188419099004042
32601;397153;0.0186264218814076
32601;29369;0.0215799553677955
32601;99785;0.0393562906063653
4075;114554;0.0081329841086566
4075;474435;0.006497574770748
```

- ***(Precedents)*** Full Precedent Content RSE 15008748520198260052 SP 1500874−85.2019.8.26.0052.
- ***(News)*** The Maria da Penha Law can also put women as aggressors.
- ***(Precedents)*** Conflict of Jurisdiction CJ 70084034495 RS.
- ***(News)*** Unpacking the notion of double suicide.
- ***(Articles)*** Deliberations on crimes against life: distinguishing between necessary and sentimental abortions.

(Note: For clarity, we have correlated the document's unique ID to its respective title.)

Leveraging the aforementioned session data as input, our best-performing model, *stan*, computed the following top-10 recommendations:

1. ***(News)*** Can the Maria da Penha Law also classify women as aggressors?
2. ***(News)*** Men threatened by women: Can they invoke the Maria da Penha Law?
3. ***(Models and Pleadings)*** Template for Urgent Protective Action.
4. ***(Precedents)*** Full Precedent Content RSE 87123720188240023 Capital 0008712−37.2018.8.24.0023.
5. ***(News)*** The Maria da Penha Law's applicability in protecting men.
6. ***(News)*** Can the Maria da Penha Law address aggression from a mother to her daughter?
7. ***(Articles)*** Post domestic violence police report: Is retraction an option?
8. ***(Articles)*** Pre-requisites for invoking the Maria da Penha Law.
9. ***(Precedents)*** Conflict of Jurisdiction CJ 70077336691 RS.
10. ***(Articles)*** Countering a personal injury criminal case: Potential defense theses.

After carefully reviewing the list of suggested documents, it is clear that the recommendations are well-suited to the student's interests, specifically in regard to exploring the various aspects and applications of the Maria da Penha Law.

5.4.4 Comparing with a content-based recommender system

In this subsection, we compare the results of the best session-based model from each recommendation group (i.e., spop, ar, stan, bprmf and narm) against the results obtained with a content-based recommendation model. Although the legal domain is rich in textual content, we investigate whether leveraging user-item interaction data

Table 5 Comparing the best session-based models against a content-based model in terms of hit rate, MRR, coverage, and popularity bias for the *next item* evaluation scenario in the JusBrasilRec dataset

Models	HitRate@10	MRR@10	Coverage@10	Popularity@10
content-based	0.550 ± 0.006	0.503 ± 0.006	0.946 ± 0.239	0.010 ± 0.001
spop	**0.599 ± 0.005**	**0.533 ± 0.006**	**0.199 ± 0.075**	**0.320 ± 0.034**
ar	**0.680 ± 0.004**	**0.540 ± 0.005**	**0.416 ± 0.110**	**0.022 ± 0.002**
stan	**0.751 ± 0.003**	**0.601 ± 0.005**	**0.419 ± 0.121**	**0.023 ± 0.002**
bprmf	0.555 ± 0.006	**0.514 ± 0.006**	0.685 ± 0.116	0.010 ± 0.001
narm	**0.732 ± 0.003**	**0.579 ± 0.005**	**0.526 ± 0.135**	**0.023 ± 0.002**

Values with statistical significance to the content-based model are in bold face

Table 6 Comparing the best session-based models against a content-based model in terms of precision, recall, NDCG, and MAP considering the *rest of the session* evaluation scenario in the JusBrasilRec dataset

Models	Precision@10	Recall@10	NDCG@10	MAP@10
content-based	0.063 ± 0.000	0.347 ± 0.002	0.353 ± 0.002	0.035 ± 0.000
spop	**0.072 ± 0.000**	**0.381 ± 0.002**	**0.394 ± 0.003**	**0.039 ± 0.000**
ar	**0.088 ± 0.000**	**0.451 ± 0.003**	**0.429 ± 0.003**	**0.049 ± 0.000**
stan	**0.099 ± 0.000**	**0.502 ± 0.004**	**0.481 ± 0.003**	**0.055 ± 0.000**
bprmf	**0.064 ± 0.000**	**0.353 ± 0.003**	**0.368 ± 0.003**	**0.036 ± 0.000**
narm	**0.096 ± 0.000**	**0.488 ± 0.004**	**0.466 ± 0.003**	**0.053 ± 0.000**

Values with statistical significance to the content-based model are in boldface

alone is enough to enable high-quality recommendations in the context of our legal dataset.

We implement the traditional Bag of Words model, representing documents within a vector space where each word serves as a dimension. Despite its simplicity, this model is a hard-to-beat baseline in numerous domains, including the legal domain (Ostendorff et al. 2021). It first derives a TF-IDF matrix from the textual content of the documents, indicating the importance of a word to its respective document amidst a set of documents. Figure 9 provides an excerpt of our TF-IDF matrix, highlighting the **doc_id**, **word_id** (words from the document converted to numerical identifiers for anonymity), and the corresponding **tf_idf** values for each **doc_id/word_id** pair. To generate recommendations, the last item from the session works as the query document. We then determine its 10 most similar documents (i.e., the recommendations) using cosine similarity.

For the sake of reproducibility, we make available the computed TF-IDF matrix used by our content-based recommendation model.

Comparative evaluations were performed across two scenarios: *next item* and *rest of the session*. We employed a two-sided paired t-test at a 95% confidence level for this comparison (Mitchell 1997). The outcomes are detailed in Tables 5 and 6. Statistically significant differences to the content-based model are emphasized in bold.

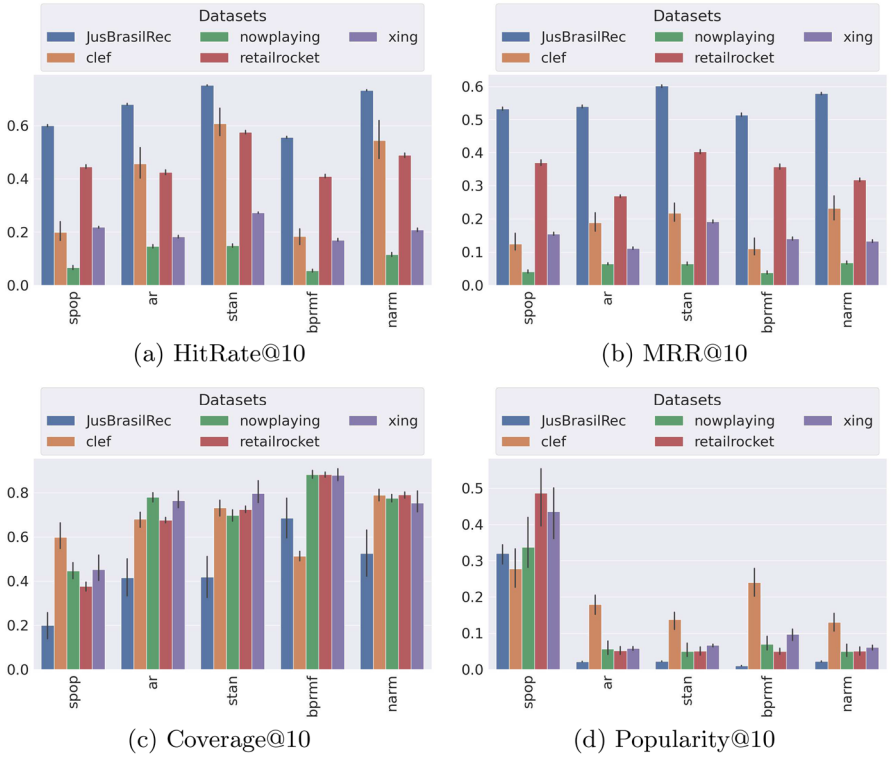


Fig. 10 Comparison of the results on JusBrasilRec to other datasets in the *next item* evaluation scenario

Observing both tables, it is clear that session-based models surpass the content-based approach in a majority of cases. The only exception is the *bprmf* model; its metrics for HitRate, MRR, and Popularity bias closely align with the content-based model. Regarding Coverage, the content-based model secured the top spot—an anticipated result given its dependency on document representation rather than user activity. In sum, this comparison underscores the potential of session-based models as viable alternatives to content-based models, especially within the legal domain.

5.4.5 Comparing with other application domains

In Figs. 10 and 11, we compare the results of the best model from each recommendation group (i.e., spop, ar, stan, bprmf and narm), obtained on the JusBrasilRec (i.e., jusbrasilrec_all_users dataset), with the results obtained on datasets coming from other application domains (cf. Sect. 5.3).

The aim of this comparison is to illustrate how challenging session-based recommendation is in the legal field when compared to other areas.

Regarding the *next item* evaluation scenario (Fig. 10), the results on JusBrasilRec are significantly better than on the other domain representatives, regardless

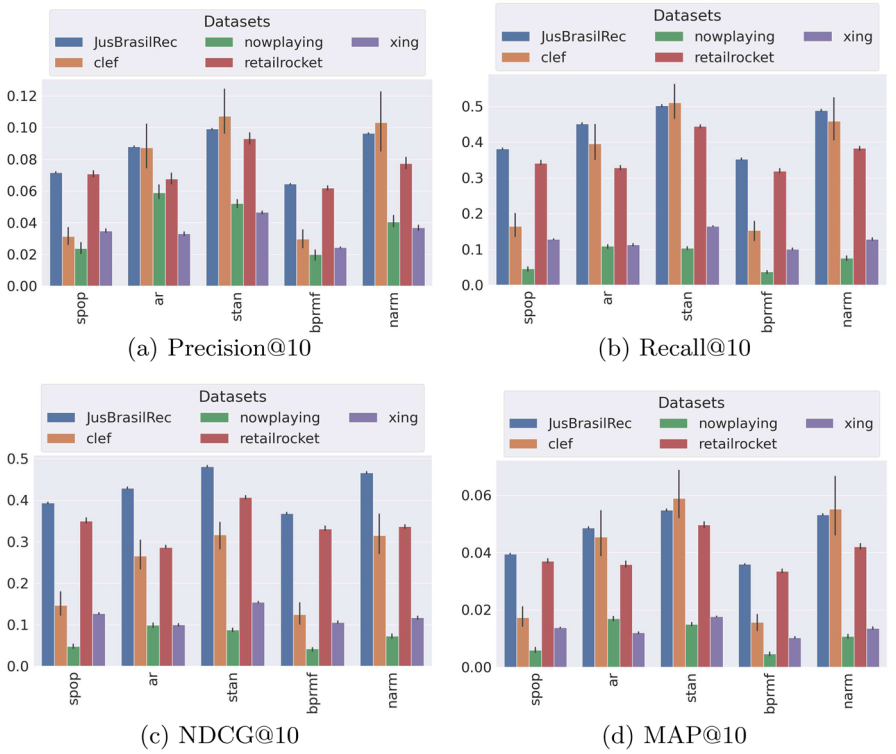


Fig. 11 Comparison of the results on JusBrasilRec to other datasets in the *rest of the session* evaluation scenario

of the recommendation model. The retailrocket dataset is probably the domain in which results more closely resemble the results observed on JusBrasilRec, proportionally speaking. Moreover, our dataset responds to all models better than the other datasets, reinforcing that the set of items within a session and sequential information carry strong preference signals in our target domain. Concerning Coverage, JusBrasilRec presents smaller values, which is probably a consequence of the much larger items catalog compared to the other datasets. As for popularity bias, the results across all datasets and models follow similar patterns. Still, JusBrasilRec appears among the ones with the smaller popularity scores.

Regarding the *rest of the session* evaluation scenario (Fig. 11), the recommendation models achieve the best results on clef and JusBrasilRec, where the worse results are observed on xing and now playing. However, unlike the *next item* evaluation scenario, the difference between the results here is less pronounced, especially in comparison to clef. In terms of NDCG, however, the differences are more pronounced in favor of JusBrasilRec.

In Table 7, we present a pairwise comparison between JusBrasilRec and each other domain dataset, summarizing the number of algorithms that perform better in one domain with respect to the other. The performance is measured in terms of

Table 7 Pairwise comparison between JusBrasilRec and other domain datasets with respect to the number of models with the best performance

Metrics	JBR/clef	JBR/nowp	JBR/retailr	JBR/xing
HitRate	14/7	18/3	18/3	20/1
MRR	15/6	20/1	19/2	21/0
Coverage	1/20	2/19	2/19	3/18
Popularity	4/17	2/19	0/21	0/21
Precision	3/18	17/4	15/6	19/2
Recall	10/11	18/3	17/4	20/1
NDCG	14/7	18/3	18/3	19/2
MAP	5/16	18/3	18/3	20/1

JBR JusBrasilRec, *nowp* nowplaying, *retailr* retailrocket datasets

the evaluation metrics and considers all algorithms from each recommendation group (cf. Sect. 5.1). Thus, the counting of algorithms with the best performances is made for the metrics and evaluation schemes, i.e., Hit Rate, MRR, Coverage, and Popularity bias for the *next item* evaluation scenario; and Precision, Recall, NDCG and MAP for the *rest of the session* scenario. In Table 7, we can see that the algorithms perform better in JusBrasilRec than in the other domains datasets in most of the cases.

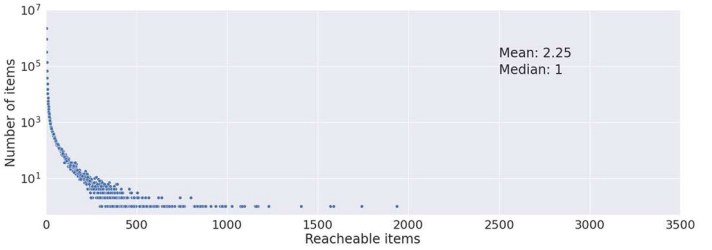
Overall, the results on our legal dataset tend to be better than the other evaluated domains. One possible reason is that legal users may conduct more focused browsing such that the number of possible next item(s) to choose from narrows down during browsing. This naturally reduces the number of candidate items to recommend, making the recommendation task "easier".

To test this hypothesis, we checked the out-of-degree distribution of items within the sessions, i.e., the number of different items a user clicks from a particular item. Figure 12 shows that in JusBrasilRec, the number of items that reach only one other item is bigger than in the other domain's datasets. This observation can also be seen in the mean and median of reachable items, where the values for JusBrasilRec are much lower than in the other application domain datasets.

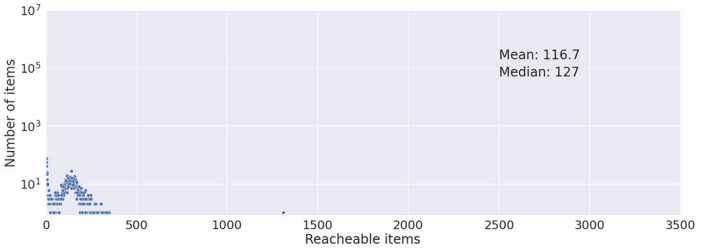
6 Related work

The first legal document recommender system we found, called ResultsPlus, date from 2007 (Al-Kofahi et al. 2007). This system recommends briefs and secondary law materials to attorneys engaged in primary law research. ResultsPlus generates a ranked list of recommendations using content-based similarity, and then, incorporates historical user and document usage data to enhance the ranking of its recommendations. The system was empirically evaluated and setup in production on Westlaw.[9] Infonorma was proposed in 2008 (Drumond and Girardi 2008). This is a multi-agent recommender system that provides its users with recommendations

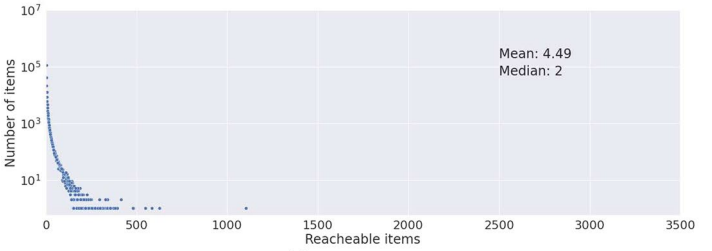
[9] https://legal.thomsonreuters.com/en/westlaw.



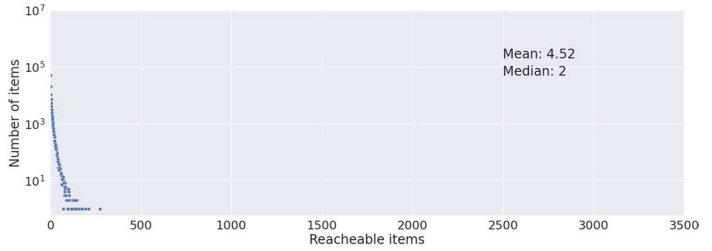
(a) JusBrasilRec



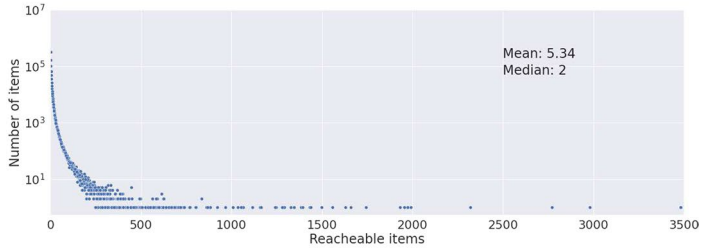
(b) clef



(c) nowplaying



(d) retailrocket



(e) xing

Fig. 12 Datasets scattering

of legal normative instruments. Infonorma classifies normative instruments, represented as semantic web documents, into legal branches and performs content-based similarity analysis.

A legal document recommender system that relies on a built-in topic segmentation algorithm was presented in Lu and Conrad (2012). Basically, the system delivers to its users the top-level legal issues underlying a case along with secondary and supplemental issues. So, the system provides additional documents that are closely related to a given current document. Additionally, the system also groups the recommendations based on issues that are discussed in the document. The system provided an encouraged performance after being evaluated by human legal experts with a collection of 7 million U.S. law case documents. Another recommender system was proposed in Winkels et al. (2014), where it is proposed a system that use a weighted graph of references between case laws and legislations to provide the recommendations. The system was tested in a prototype for Immigration Law in Netherlands.

From 2020 to present days, we have a few more recent researches. Quick Check is a system that extracts the legal arguments from a user's brief and recommends highly relevant case law opinions (Thomas et al. 2020). The system combines full-text search, citation network analysis, clickstream analysis, and a hierarchy of ranking models to generate the recommendations. The system was evaluated in a large corpus of graded issue-segment-to-case pairs; and provided 39% of highly relevant, 60.5% of relevant, and 0.5% of irrelevant recommendations.

In Yang et al. (2021) the authors proposed LegalGNN, which is a legal information enhanced graph neural network-based recommender system. The system first builds a heterogeneous network representation model (HLIN) to connect structural features (e.g., legal knowledge graph) and contextual features (e.g., the content of legal documents) for training. Then, to model user interests, the system incorporates queries submitted to legal systems into the HLIN and links them with both retrieved documents and the users who submitted those queries. Finally, a graph neural network with relational attention mechanism is applied to make use of high-order connections in HLIN for recommendations. The system was evaluated in a real-world Chinese legal dataset containing 64,642 interactions, 3327 users and 343,244 items. The results showed that LegalGNN significantly outperforms several state-of-the-art methods.

In legal domain, attorneys/practitioners analyze relevant previous judgments to prepare the arguments for an ongoing case. In this context, a recommender system needs to compute pairwise similarity scores between judgment pairs, which is an expensive time cost process. To address this issue, in Dhanani et al. (2021a) the authors proposed a graph clustering based recommender system which creates clusters of similar judgments and compute pairwise similarity only within the clusters. On the other hand, in Dhanani et al. (2021b) they proposed a recommender system that uses a customized domain-specific preprocessing and a legal dictionary to reduce the corpus and precisely recommend the relevant judgments. Both proposals were evaluated in a corpus of 48,000 Indian judgments, and the results demonstrated an encouraging performance in terms of Accuracy, F1-Score, MCC Score, and computational complexity.

Another proposal to reduce the time cost for computing pairwise similarity scores between judgment pairs was proposed in Dhanani et al. (2022). The authors proposed a legal document recommender system that uses a pre-learned word embedding to learn the Doc2Vec embedding of the documents, and the cosine measure to compute the similarity between pairs of judgments. The system was also evaluated with the previous Indian dataset, and compared with a similar system that compute the embedding of the documents using the traditional Doc2Vec method. Results showed that the proposed work performs significantly better than traditional Doc2Vec with an Accuracy of 0.88, F1-Score of 0.82 and MCC Score of 0.73.

Finally, we also found two benchmarks in the literature. The first is a short benchmark focused on legal document recommender systems (Huang et al. 2021). The authors compared four types of recommendation models, including a citation-list based model (i.e., a user-based collaborative filtering) and three context-based models (i.e., bag-of-words similarity, BiLSTM and RoBERTa). The benchmark was carried out with a subset of 324,309 cases of BVA, which is a corpus containing the full text of over 1 million appeal decisions. Results showed that BiLSTM and RoBERTa perform comparably, and outperform the collaborative filtering and bag-of-words models.

The second benchmark focused on document representation in content-based legal literature recommender systems (Ostendorff et al. 2021). Using two corpora (i.e., Open Case Book, which consists of 222 case books containing 3023 cases from 87 authors; and Wikisource, which is a collection of 2939 US Supreme Court decisions), the authors evaluated 27 representation methods grouped into four categories: baseline (i.e., tf-idf), word vector-based (e.g., paragraph vector, GloVe, fastText, etc), transformer-based (e.g., BERT, RoBERTa, Sentence-BERT, Sentence-RoBERTa, LongFormer, etc), citation-based (e.g., DeepWalk, Walklets, BoostNE, etc), and variation and hybrid methods (e.g., Sentence-Legal-AUEB-BERT-base, $\text{fastText}_{Legal}512$, etc). The evaluation using the document representation methods, and the cosine similarity to provide the recommendations, showed that document representations from averaged fastText word vectors yield the best results.

As we can see, this section brings evidences that the literature about legal document recommender systems is still scarce. Additionally, we could also verified the lack of large and publicly available datasets, and benchmarks for legal document recommender systems. This fact emphasizes the importance of the JusBrasilRec data collection and the session-based benchmark proposed in this work.

7 Limitations

We will address the limitations of our work in the future. Although the session-based models yielded good results for the diverse audience on the Jusbrasil platform-encompassing legal professionals and non-experts-an underlying risk exists: models may absorb and disseminate misleading information from non-expert users to legal professionals (Abdollahpouri et al. 2020; Zheng 2019; Abdollahpouri et al. 2017). This stresses the necessity of adapting models based on user expertise. A promising approach is the adoption of multistakeholder techniques, which address

the various stakeholders involved in the recommendation process. Examples include group-based and fairness-aware recommendations, capable of accommodating user expertise levels (da Silva et al. 2021).

Moreover, the session-based models demonstrated superior performance in the legal domain compared to other application domains. However, deploying recommendation models in legal contexts involves significantly higher challenges and stakes. An erroneous legal research recommendation could lead to substantial financial losses due to wasted attorney time or, in the worst scenarios, result in the forfeiture of property, freedom, or even life due to weak arguments. Consequently, the legal domain mandates the incorporation of expert validation stages, the employment of sophisticated benchmarks (Guo 2023; Guo et al. 2023) that weigh the gravity of legal outcomes, and/or the integration of user expertise levels into the recommendation process (Abdollahpouri et al. 2020; Zheng 2019; Abdollahpouri et al. 2017; da Silva et al. 2021).

8 Conclusions and future work

This paper presented the first recommender systems benchmark on a legal domain dataset. We have considered the session-based recommender task, a prevalent task in recommender systems. First, we have collected a large-scale dataset from Jusbrasil, the largest legal search portal in Brazil, which will be made available to the research community. Next, we broadly characterized properties that may affect session-based recommenders. Then, we compared a wide range and variety of session-based recommender models in two evaluation scenarios, recommending the *next item* or the *rest of the session*. Additionally, we also detailed a real recommendation example drawn from the best model. Moreover, we also compared the session-based models against a content-based recommendation model. Finally, we compared the results attained in our collected dataset with those observed on datasets from four domains.

Among our findings, we highlight the following:

- The session size on our legal dataset is smaller than in other domains. This may suggest that users in this domain conduct focused browsing with little exploration;
- Models that leverage the whole session or the order of items in the session attain the best results, which means that both the set and order of items in the session carry strong preference signals;
- Popularity bias is minimal in virtually all cases. This indicates that legal users are not influenced by documents' popularity, such as in other domains like music;
- Session-based models are a good alternative to content-based ones. We showed that models that leverage user-item interaction data only outperform models that use documents' textual content;
- The overall results on our legal dataset were better than in other domains, with a few exceptions. One possible reason is that legal users conduct focused browsing such that the number of possible next item(s) to choose from narrows down dur-

ing browsing. This naturally reduces the number of candidate items to recommend, making the recommendation job "easier".

For future work, we plan to evaluate the impact of sequence and session-aware models on the logged users dataset. We also plan to assess other content-based and hybrid algorithms. This domain is rich in textual content, which can benefit models based on modern NLP and NLU techniques. We also plan to consider other evaluation metrics such as novelty and diversity, and other evaluation scenarios such as A/B testing. Finally, we plan to explore some multistakeholder techniques to improve the relevance of the recommendations, particularly for the legal professionals using the platform.

Appendix A: Results for logged and unlogged users in JusBrasilRec collection

In this appendix, we present the results for jusbrasilrec_logged_users (Tables 8, 10) and jusbrasilrec_unlogged_users (Tables 9, 11) considering top-10 recommendation lists for both the *next item* and *rest of the session* evaluation scenarios.

Table 8 Hit rate, MRR, coverage, and popularity bias for top-10 recommendation lists considering the *next item* evaluation scenario in the jusbrasilrec_logged_users dataset

Models	HitRate@10	MRR@10	Coverage@10	Popularity@10
random	0.000 ± 0.000	0.000 ± 0.000	**0.988 ± 0.016**	0.002 ± 0.000
pop	0.015 ± 0.001	0.006 ± 0.000	0.000 ± 0.000	**0.603 ± 0.055**
rpop	0.016 ± 0.001	0.006 ± 0.001	0.000 ± 0.000	0.590 ± 0.050
spop	0.597 ± 0.003	0.520 ± 0.003	0.188 ± 0.078	0.516 ± 0.047
ar	0.657 ± 0.003	0.511 ± 0.003	0.404 ± 0.115	0.048 ± 0.004
markov	0.640 ± 0.003	0.535 ± 0.002	0.315 ± 0.087	0.041 ± 0.004
sr	0.653 ± 0.003	0.529 ± 0.003	0.362 ± 0.100	0.044 ± 0.004
iknn	0.165 ± 0.004	0.100 ± 0.003	0.426 ± 0.101	0.021 ± 0.002
sknn	0.711 ± 0.003	0.527 ± 0.002	0.393 ± 0.125	0.046 ± 0.004
vsknn	0.737 ± 0.003	0.568 ± 0.002	0.403 ± 0.128	0.054 ± 0.004
stan	**0.739 ± 0.003**	0.590 ± 0.003	0.404 ± 0.126	0.051 ± 0.004
vstan	0.732 ± 0.003	**0.601 ± 0.003**	0.432 ± 0.125	0.040 ± 0.003
bprmf	0.548 ± 0.002	0.500 ± 0.003	0.661 ± 0.134	0.028 ± 0.003
fpmc	0.519 ± 0.002	0.490 ± 0.002	0.783 ± 0.129	0.011 ± 0.001
fism	0.174 ± 0.009	0.128 ± 0.009	0.677 ± 0.124	0.008 ± 0.001
fossil	0.006 ± 0.001	0.002 ± 0.000	0.714 ± 0.140	0.054 ± 0.024
smf	0.283 ± 0.007	0.196 ± 0.005	0.012 ± 0.001	0.316 ± 0.021
gru4rec	0.658 ± 0.002	0.515 ± 0.002	0.551 ± 0.124	0.021 ± 0.002
narm	0.716 ± 0.003	0.562 ± 0.003	0.531 ± 0.147	0.050 ± 0.004
stamp	0.677 ± 0.004	0.510 ± 0.008	0.441 ± 0.127	0.046 ± 0.006
sgnn	0.660 ± 0.004	0.512 ± 0.003	0.318 ± 0.085	0.061 ± 0.006

The best results are highlighted in boldface

Table 9 Hit rate, MRR, coverage, and popularity bias for top-10 recommendation lists considering the *next item* evaluation scenario in the jusbrasilrec_unlogged_users dataset

Models	HitRate@10	MRR@10	Coverage@10	Popularity@10
random	0.000 ± 0.000	0.000 ± 0.000	**0.999 ± 0.002**	0.001 ± 0.000
pop	0.020 ± 0.002	0.012 ± 0.001	0.000 ± 0.000	**0.289 ± 0.036**
rpop	0.021 ± 0.001	0.012 ± 0.000	0.000 ± 0.000	0.260 ± 0.015
spop	0.624 ± 0.007	0.575 ± 0.008	0.260 ± 0.081	0.269 ± 0.033
ar	0.712 ± 0.007	0.582 ± 0.007	0.481 ± 0.082	0.015 ± 0.001
markov	0.691 ± 0.007	0.591 ± 0.007	0.419 ± 0.071	0.013 ± 0.001
sr	0.702 ± 0.007	0.589 ± 0.007	0.446 ± 0.073	0.014 ± 0.001
iknn	0.170 ± 0.005	0.105 ± 0.004	0.500 ± 0.067	0.008 ± 0.001
sknn	0.758 ± 0.004	0.605 ± 0.007	0.495 ± 0.094	0.013 ± 0.001
vsknn	0.772 ± 0.005	0.629 ± 0.007	0.496 ± 0.093	0.016 ± 0.001
stan	**0.773 ± 0.005**	0.638 ± 0.007	0.500 ± 0.092	0.015 ± 0.001
vstan	0.768 ± 0.006	**0.644 ± 0.007**	0.527 ± 0.089	0.012 ± 0.001
bprmf	0.599 ± 0.007	0.563 ± 0.008	0.744 ± 0.096	0.012 ± 0.001
fpmc	0.590 ± 0.008	0.562 ± 0.008	0.859 ± 0.079	0.005 ± 0.000
fism	0.205 ± 0.003	0.136 ± 0.006	0.766 ± 0.082	0.004 ± 0.000
fossil	0.025 ± 0.015	0.012 ± 0.009	0.763 ± 0.073	0.031 ± 0.038
smf	0.283 ± 0.013	0.214 ± 0.008	0.022 ± 0.002	0.100 ± 0.010
gru4rec	0.709 ± 0.007	0.570 ± 0.006	0.648 ± 0.077	0.008 ± 0.001
narm	0.756 ± 0.006	0.611 ± 0.007	0.635 ± 0.103	0.015 ± 0.001
stamp	0.725 ± 0.008	0.564 ± 0.008	0.535 ± 0.101	0.015 ± 0.001
sgnn	0.724 ± 0.005	0.583 ± 0.006	0.416 ± 0.066	0.019 ± 0.002

The best results are highlighted in boldface

Table 10 Precision, recall, NDCG, and MAP for a top-10 obtained for the *rest of the session* evaluation scenario in the jusbrasilrec_logged_users dataset

Models	Precision@10	Recall@10	NDCG@10	MAP@10
random	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
pop	0.002 ± 0.000	0.011 ± 0.000	0.007 ± 0.000	0.001 ± 0.000
rpop	0.002 ± 0.000	0.011 ± 0.000	0.007 ± 0.000	0.001 ± 0.000
spop	0.073 ± 0.000	0.361 ± 0.005	0.375 ± 0.004	0.038 ± 0.000
ar	0.088 ± 0.001	0.414 ± 0.007	0.396 ± 0.005	0.045 ± 0.001
markov	0.083 ± 0.000	0.399 ± 0.007	0.397 ± 0.005	0.043 ± 0.001
sr	0.086 ± 0.001	0.410 ± 0.007	0.401 ± 0.005	0.045 ± 0.001
iknn	0.033 ± 0.000	0.127 ± 0.003	0.111 ± 0.003	0.014 ± 0.000
sknn	0.097 ± 0.000	0.454 ± 0.007	0.431 ± 0.005	0.050 ± 0.001
vsknn	**0.101 ± 0.001**	0.470 ± 0.007	0.449 ± 0.005	**0.052 ± 0.001**
stan	**0.101 ± 0.001**	**0.471 ± 0.007**	**0.457 ± 0.005**	**0.052 ± 0.001**
vstan	0.099 ± 0.001	0.464 ± 0.007	0.454 ± 0.006	0.051 ± 0.001
bprmf	0.065 ± 0.000	0.330 ± 0.006	0.348 ± 0.005	0.034 ± 0.001
fpmc	0.059 ± 0.000	0.309 ± 0.005	0.325 ± 0.004	0.031 ± 0.001
fism	0.020 ± 0.001	0.103 ± 0.007	0.095 ± 0.007	0.010 ± 0.001
fossil	0.001 ± 0.000	0.004 ± 0.001	0.002 ± 0.000	0.000 ± 0.000
smf	0.040 ± 0.001	0.182 ± 0.004	0.168 ± 0.004	0.020 ± 0.000
gru4rec	0.087 ± 0.000	0.411 ± 0.007	0.396 ± 0.005	0.045 ± 0.001
narm	0.097 ± 0.001	0.455 ± 0.007	0.438 ± 0.005	0.050 ± 0.001
stamp	0.090 ± 0.001	0.426 ± 0.007	0.406 ± 0.007	0.046 ± 0.001
sgnn	0.091 ± 0.001	0.421 ± 0.006	0.402 ± 0.005	0.046 ± 0.001

The best results are highlighted in boldface

Table 11 Precision, recall, NDCG, and MAP for a top-10 obtained for the *rest of the session* evaluation scenario in the jusbrasilrec_unlogged_users dataset

Models	Precision@10	Recall@10	NDCG@10	MAP@10
random	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
pop	0.002 ± 0.000	0.014 ± 0.002	0.011 ± 0.001	0.001 ± 0.000
rpop	0.003 ± 0.000	0.015 ± 0.001	0.012 ± 0.001	0.001 ± 0.000
spop	0.071 ± 0.001	0.420 ± 0.005	0.432 ± 0.005	0.043 ± 0.001
ar	0.087 ± 0.001	0.499 ± 0.007	0.475 ± 0.006	0.053 ± 0.001
markov	0.083 ± 0.001	0.480 ± 0.007	0.469 ± 0.006	0.051 ± 0.001
sr	0.085 ± 0.001	0.490 ± 0.007	0.473 ± 0.006	0.052 ± 0.001
iknn	0.029 ± 0.001	0.141 ± 0.005	0.119 ± 0.004	0.015 ± 0.000
sknn	0.094 ± 0.001	0.534 ± 0.005	0.508 ± 0.005	0.057 ± 0.001
vsknn	**0.096 ± 0.001**	0.544 ± 0.006	0.519 ± 0.006	**0.058 ± 0.001**
stan	**0.096 ± 0.001**	**0.545 ± 0.006**	**0.522 ± 0.006**	**0.058 ± 0.001**
vstan	0.094 ± 0.001	0.539 ± 0.006	0.520 ± 0.006	**0.058 ± 0.001**
bprmf	0.067 ± 0.001	0.403 ± 0.005	0.416 ± 0.006	0.041 ± 0.001
fpmc	0.065 ± 0.001	0.395 ± 0.006	0.406 ± 0.006	0.040 ± 0.001
fism	0.022 ± 0.000	0.137 ± 0.003	0.116 ± 0.003	0.013 ± 0.000
fossil	0.003 ± 0.002	0.017 ± 0.010	0.012 ± 0.008	0.002 ± 0.001
smf	0.035 ± 0.001	0.195 ± 0.008	0.180 ± 0.007	0.020 ± 0.001
gru4rec	0.086 ± 0.001	0.495 ± 0.007	0.468 ± 0.006	0.052 ± 0.001
narm	0.093 ± 0.001	0.532 ± 0.007	0.506 ± 0.006	0.057 ± 0.001
stamp	0.088 ± 0.001	0.508 ± 0.008	0.477 ± 0.008	0.054 ± 0.001
sgnn	0.090 ± 0.001	0.511 ± 0.005	0.483 ± 0.005	0.055 ± 0.001

The best results are highlighted in boldface

Acknowledgements This research is partially supported by the Jusbrasil Postdoctoral Fellowship Program, the IPDEC Institute, the Brazilian funding agency FAPEAM-POSGRAD 2022, the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) financial code 001, and individual grant from CNPq to Altigran da Silva (307248/2019-4).

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, experiments and analysis were performed by MAD. The manuscript was written by all authors. All authors read and approved the manuscript.

Funding This research is partially supported by the Jusbrasil Postdoctoral Fellowship Program, the IPDEC Institute, the Brazilian funding agency FAPEAM-POSGRAD 2022, the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) financial code 001, and individual grant from CNPq to Altigran da Silva (307248/2019-4).

Data availability The data used in this work is available in the Zenodo repository (https://zenodo.org/record/8401278).

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no confict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Abdollahpouri H, Burke R, Mobasher B (2017) Recommender systems as multistakeholder environments. In: Proceedings of the 25th conference on user modeling, adaptation and personalization. UMAP '17. Association for Computing Machinery, New York, NY, USA, pp 347–348. https://doi.org/10.1145/3079628.3079657

Abdollahpouri H, Adomavicius G, Burke R, Guy I, Jannach D, Kamishima T, Krasnodebski J, Pizzato LA (2020) Multistakeholder recommendation: survey and research directions. User Model User-Adapt Interact 30(1):127–158. https://doi.org/10.1007/s11257-019-09256-1

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on management of data. SIGMOD '93. Association for Computing Machinery, New York, NY, USA, pp 207–216. https://doi.org/10.1145/170035.170072

Al-Kofahi K, Jackson P, Dahn M, Elberti C, Keenan W, Duprey J (2007) A document recommendation system blending retrieval and categorization technologies. In: 2007 AAAI workshop on intelligent techniques for web personalization and recommender systems in E-commerce, pp 9–16

Cooley R, Mobasher B, Srivastava J (1999) Data preparation for mining world wide web browsing patterns. Knowl Inf Syst 1(1):5–32. https://doi.org/10.1007/BF03325089

da Silva DC, Manzato MG, Durão FA (2021) Exploiting personalized calibration and metrics for fairness recommendation. Expert Syst Appl 181:115112. https://doi.org/10.1016/j.eswa.2021.115112

Deshpande M, Karypis G (2004) Item-based top-n recommendation algorithms. ACM Trans Inf Syst 22(1):143–177. https://doi.org/10.1145/963770.963776

Dhanani J, Mehta R, Rana D (2021a) Legal document recommendation system: a cluster based pairwise similarity computation. J Intell Fuzzy Syst 41:1–13. https://doi.org/10.3233/JIFS-189871

Dhanani J, Mehta R, Rana D (2021b) Legal document recommendation system: a dictionary based approach. Int J Web Inf Syst 17(3):187–203. https://doi.org/10.1108/IJWIS-02-2021-0015

Dhanani J, Mehta R, Rana D (2022) Effective and scalable legal judgment recommendation using prelearned word embedding. Complex Intell Syst. https://doi.org/10.1007/s40747-022-00673-1

Drumond L, Girardi R (2008) A multi-agent legal recommender system. Artif Intell Law 16(2):175–207. https://doi.org/10.1007/s10506-008-9062-8

Garg D, Gupta P, Malhotra P, Vig L, Shroff G (2019) Sequence and time aware neighborhood for session-based recommendations: Stan. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. SIGIR'19. Association for Computing Machinery, New York, NY, USA, pp 1069–1072. https://doi.org/10.1145/3331184.3331322

Guo H (2023) Fairness testing for recommender systems. In: Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis. ISSTA 2023. Association for Computing Machinery, New York, NY, USA, pp 1546–1548. https://doi.org/10.1145/3597926.3605235

Guo H, Li J, Wang J, Liu X, Wang D, Hu Z, Zhang R, Xue H (2023) Fairrec: fairness testing for deep recommender systems. In: Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis. ISSTA 2023. Association for Computing Machinery, New York, NY, USA, pp 310–321. https://doi.org/10.1145/3597926.3598058

He R, McAuley J (2016) Fusing similarity models with Markov chains for sparse sequential recommendation. In: 2016 IEEE 16th international conference on data mining (ICDM), pp 191–200. https://doi.org/10.1109/ICDM.2016.0030

Hidasi B, Karatzoglou A, Baltrunas L, Tikk D (2016) Session-based recommendations with recurrent neural networks. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings. arXiv:abs/1511.06939

Huang Z, Low C, Teng M, Zhang H, Ho DE, Krass MS, Grabmair M (2021) Context-aware legal citation recommendation using deep learning. In: Proceedings of the eighteenth international conference on artificial intelligence and law. Association for Computing Machinery, New York, NY, USA, pp 79–88. https://doi.org/10.1145/3462757.3466066

Jannach D, Ludewig M (2017) When recurrent neural networks meet the neighborhood for session-based recommendation. In: Proceedings of the eleventh ACM conference on recommender systems. RecSys '17. Association for Computing Machinery, New York, NY, USA, pp 306–310. https://doi.org/10.1145/3109859.3109872

Jannach D, Lerche L, Kamehkhosh I, Jugovac M (2015) What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Model User-Adapt Interact 25(5):427–491. https://doi.org/10.1007/s11257-015-9165-3

Jannach D, Quadrana M, Cremonesi P (2022) In: Ricci F, Rokach L, Shapira B (eds) Session-based recommender systems. Springer, New York, NY, pp 301–334. https://doi.org/10.1007/978-1-0716-2197-4_8

Kabbur S, Ning X, Karypis G (2013) Fism: factored item similarity models for top-n recommender systems. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '13. Association for Computing Machinery, New York, NY, USA, pp 659–667. https://doi.org/10.1145/2487575.2487589

Kamehkhosh I, Jannach D, Ludewig M (2017) A comparison of frequent pattern techniques and a deep learning method for session-based recommendation. In: RecTemp@RecSys

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37. https://doi.org/10.1109/MC.2009.263

Latifi S, Mauro N, Jannach D (2021) Session-aware recommendation: a surprising quest for the state-of-the-art. Inf Sci 573:291–315. https://doi.org/10.1016/j.ins.2021.05.048

Li J, Ren P, Chen Z, Ren Z, Lian T, Ma J (2017) Neural attentive session-based recommendation. In: Proceedings of the 2017 ACM on conference on information and knowledge management. CIKM '17. Association for Computing Machinery, New York, NY, USA, pp 1419–1428. https://doi.org/10.1145/3132847.3132926

Liu Q, Zeng Y, Mokhosi R, Zhang H (2018) Stamp: short-term attention/memory priority model for session-based recommendation. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '18. Association for Computing Machinery, New York, NY, USA, pp 1831–1839. https://doi.org/10.1145/3219819.3219950

Lu Q, Conrad JG (2012) Bringing order to legal documents—an issue-based recommendation system via cluster association. In: Filipe J, Dietz JLG (eds) KEOD 2012—proceedings of the international conference on knowledge engineering and ontology development, Barcelona, Spain, 4–7 October, 2012, pp 76–88

Ludewig M, Jannach D (2018) Evaluation of session-based recommendation algorithms. User Model User-Adapt Interact 28(4–5):331–390. https://doi.org/10.1007/s11257-018-9209-6

Ludewig M, Mauro N, Latifi S, Jannach D (2021) Empirical analysis of session-based recommendation algorithms. User Model User Adapt Interact 31(1):149–181. https://doi.org/10.1007/s11257-020-09277-1

Mitchell TM (1997) Machine learning, vol 1. McGraw-Hill, New York

Norris JR (1997) Markov chains. Cambridge series in statistical and probabilistic mathematics. https://doi.org/10.1017/CBO9780511810633

Ostendorff M, Ash E, Ruas T, Gipp B, Moreno-Schneider J, Rehm G (2021) Evaluating document representations for content-based legal literature recommendations. In: Proceedings of the eighteenth international conference on artificial intelligence and law. Association for Computing Machinery, New York, NY, USA, pp 109–118. https://doi.org/10.1145/3462757.3466073

Quadrana M, Cremonesi P, Jannach D (2018) Sequence-aware recommender systems. ACM Comput Surv. https://doi.org/10.1145/3190616

Rendle S (2022) In: Ricci F, Rokach L, Shapira B (eds) Item recommendation from implicit feedback. Springer, New York, NY, pp 143–171. https://doi.org/10.1007/978-1-0716-2197-4_4

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. UAI '09. AUAI Press, Arlington, Virginia, USA, pp 452–461

Rendle S, Freudenthaler C, Schmidt-Thieme L (2010) Factorizing personalized Markov chains for next-basket recommendation. In: Proceedings of the 19th international conference on world wide web. WWW '10. Association for Computing Machinery, New York, NY, USA, pp 811–820. https://doi.org/10.1145/1772690.1772773

Ricci F, Rokach L, Shapira B, Kantor PB (eds) (2011) Recommender systems handbook. Springer, New York. https://doi.org/10.1007/978-0-387-85820-3

Sansone C, Sperlí G (2022) Legal information retrieval systems: state-of-the-art and open issues. Inf Syst 106:101967. https://doi.org/10.1016/j.is.2021.101967

Thomas M, Vacek T, Shuai X, Liao W, Sanchez G, Sethia P, Teo D, Madan K, Custis T (2020) Quick check: a legal research recommendation system. In: NLLP@KDD

Verstrepen K, Goethals B (2014) Unifying nearest neighbors collaborative filtering. In: Proceedings of the 8th ACM conference on recommender systems. RecSys '14. Association for Computing Machinery, New York, NY, USA, pp 177–184. https://doi.org/10.1145/2645710.2645731

Wang S, Hu L, Wang Y, Cao L, Sheng QZ, Orgun M (2019) Sequential recommender systems: challenges, progress and prospects. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19, pp 6332–6338. https://doi.org/10.24963/ijcai.2019/883

Wang S, Cao L, Wang Y, Sheng QZ, Orgun MA, Lian D (2021) A survey on session-based recommender systems. ACM Comput Surv. https://doi.org/10.1145/3465401

Winkels R, Boer A, Vredebregt B, van Someren A (2014) Towards a legal recommender system. In: Frontiers in artificial intelligence and applications, pp 169–178

Wu S, Tang Y, Zhu Y, Wang L, Xie X, Tan T (2019) Session-based recommendation with graph neural networks. In: Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth aaai symposium on educational advances in artificial intelligence. AAAI'19/IAAI'19/EAAI'19. https://doi.org/10.1609/aaai.v33i01.3301346

Yang J, Ma W, Zhang M, Zhou X, Liu Y, Ma S (2021) Legalgnn: legal information enhanced graph neural network for recommendation. ACM Trans Inf Syst. https://doi.org/10.1145/3469887

Zangerle E, Pichl M, Gassler W, Specht G (2014) #nowplaying music dataset: extracting listening behavior from twitter. In: Proceedings of the first international workshop on internet-scale multimedia management. WISMM '14. Association for Computing Machinery, New York, NY, USA, pp 21–26. https://doi.org/10.1145/2661714.2661719

Zheng Y (2019) Multi-stakeholder recommendations: case studies, methods and challenges. In: Proceedings of the 13th ACM conference on recommender systems. RecSys '19. Association for Computing Machinery, New York, NY, USA, pp 578–579. https://doi.org/10.1145/3298689.3346951

Authors and Affiliations

Marcos Aurélio Domingues[1,2,4] · Edleno Silva de Moura[2,4] · Leandro Balby Marinho[3] · Altigran da Silva[2]

✉ Marcos Aurélio Domingues
madomingues@uem.br

Edleno Silva de Moura
edleno@icomp.ufam.edu.br

Leandro Balby Marinho
lbmarinho@computacao.ufcg.edu.br

Altigran da Silva
alti@icomp.ufam.edu.br

[1] State University of Maringá, Maringá, Paraná, Brazil

[2] Federal University of Amazonas, Manaus, Amazonas, Brazil

[3] Federal University of Campina Grande, Campina Grande, Paraiba, Brazil

[4] Jusbrasil, Salvador, Bahia, Brazil