UNIVERSITÄT KLAGENFURT

Towards Intent-Aware Recommender Systems

Dietmar Jannach, University of Klagenfurt, Austria
dietmar.jannach@aau.at

Keynote at the RecTemp Workshop, ACM RecSys, Bari, Italy, 2024

Further reading (accepted at ACM TORS)

A Survey on Intent-aware Recommender Systems

DIETMAR JANNACH, University of Klagenfurt, Austria
MARKUS ZANKER, Free University of Bozen-Bolzano, Italy

Many modern online services feature personalized recommendations. A central challenge when providing such recommendations is that the reason *why* an individual user accesses the service may change from visit to visit or even during an ongoing usage session. To be effective, a recommender system should therefore aim to take the users' probable *intent* of using the service at a certain point in time into account. In recent years, researchers have thus started to address this challenge by incorporating *intent-awareness* into recommender systems. Correspondingly, a number of technical approaches were put forward, including diversification techniques, intent prediction models or latent intent modeling approaches. In this paper, we survey and categorize existing approaches to building the next generation of *Intent-Aware Recommender Systems* (IARS). Based on an analysis of current evaluation practices, we outline open gaps and possible future directions in this area, which in particular include the consideration of additional interaction signals and contextual information to further improve the effectiveness of such systems.

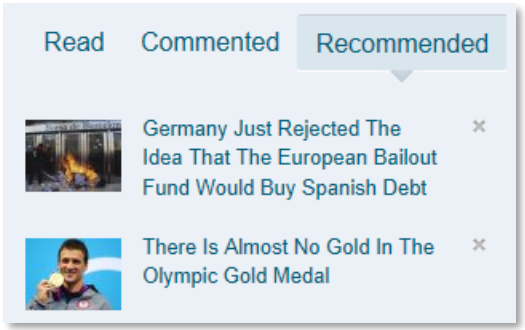
CCS Concepts: • **Information systems** → **Recommender systems**.

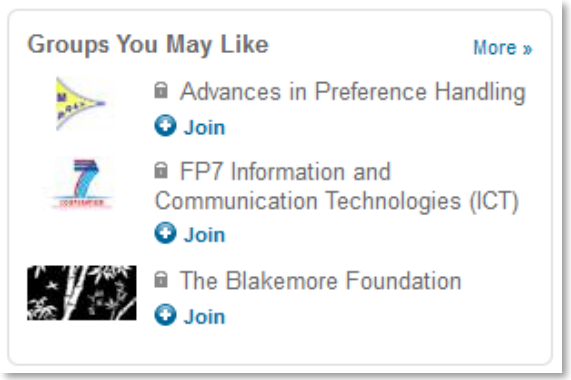
Additional Key Words and Phrases: Recommender Systems

Slides: https://tinyurl.com/intent-aware-rs

Recommender Systems (RS)

- A pervasive part of our daily online user experience
- One of the most widely used applications of machine learning





A common challenge

- A user commonly visits a service several times
  - E.g., a music streaming service, or an e-commerce site
- The reasons why the user visits the site may change from visit to visit, e.g., in the music domain:
  - finding some background music during sports
  - listening to one's favorite tracks
  - explore new types of music
  - quickly access my playlist

➢ As a result, the set of suitable recommendations depends on the user's intent

A common challenge

- Guessing the intent is difficult

Roll over image to zoom in

Minnow Sports

Minnow Sports Aluminum Baseball Bat For Baseball & Teeball

8 customer reviews

Price: $29.99
Sale: $19.99
You Save: $10.00 (33%)

In Stock.

This item does not ship to **Germany**. Please check other sellers who may ship internationally. Learn more

Sold by BBro Store and Fulfilled by Amazon. Gift-wrap available.

Item Display Length:

32.0 inches

- Made from lightweight high grade Aluminum alloy for faster swing speed
- Ultra-thin 32" handle with All Sports grip for increased stability and accuracy
- Stylish design featuring full rolled-over end for ultimate performance
- Ideal for all levels of baseball players from practice to matches
- 32 inches in length & 24 ounces

Session-based Recommendation

- Also in online music recommendation
- Our user searched and listened to “Last Christmas” by Wham!
- Should we, …
  - Play more songs by Wham!?
  - More pop Christmas songs?
  - More popular songs from the 1980s?
  - Play more songs with controversial user feedback?

A common challenge

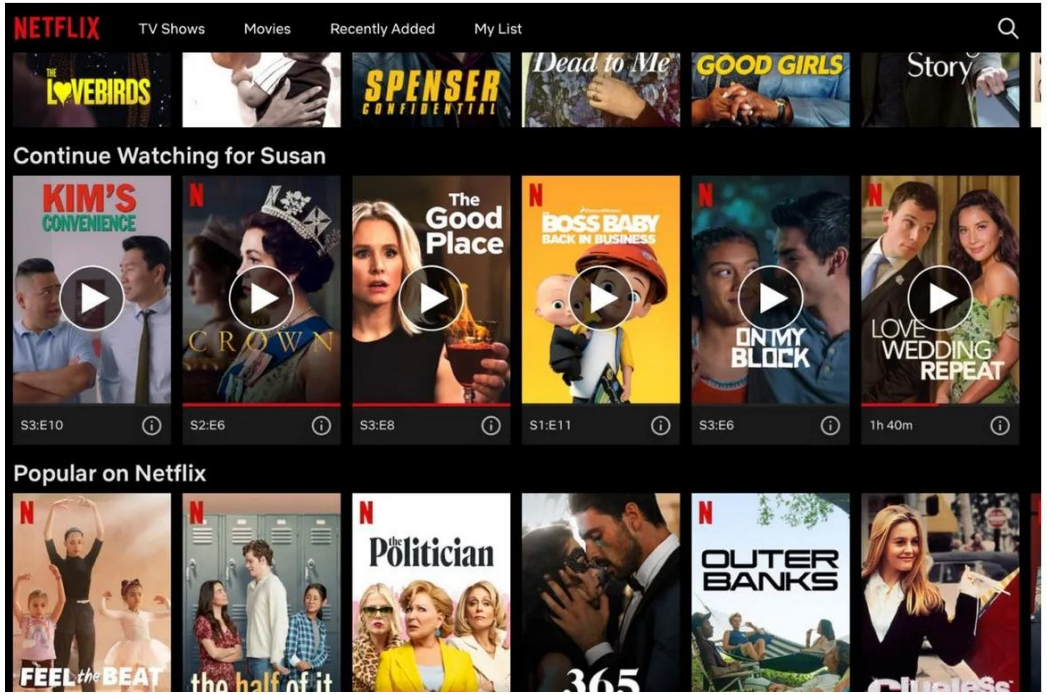
- We call the underlying reasons/motivations of a user the "intent"
  - This intent is mostly related to "why" a user visits the site, e.g., organizing a breakfast, and not "what" the user is currently looking at
  - The user's current intent is usually unknown, but we can try to guess it

➢ Considering the intent can lead to better recommendations

  - E.g., if we could somehow guess that the user today is more open to explore new music, we may increase the novelty and diversity of the recommendations
  - Intent-Aware Recommender Systems (IARS) promise to fulfill such goals

Intent-awareness in Practice

- Show multiple lists to users, trying to cover potential intents
  - Continue watching
  - Catch up with what's trending
  - See more of a particular genre
- Parts of the lists are based on possible intents
  - Categories are in some ways related to intent



Intent-awareness in Practice

- A real-world study in e-commerce
  - Incorporating four predefined intents in a switching hybrid
    - Research shopping
    - Comparative browsing
    - Idea searching
    - Hedonic browsing
  - Predict the current intent from interaction signals
    - Active time on page and on categories, even mouse scrolling
    - Train a model based on ground truth obtained from survey data
  - Determine best algorithm for each intent

Fanjuan Shia and ChirineGhedira. 2017. Improving recommender systems with an intention-based algorithm switching strategy. In Proceedings of the Symposium on Applied Computing. 1668–1673

Intent-awareness in Practice

- A real-world study in e-commerce
  - A/B test run for three months
  - Observed significant positive effects of hybrid model
    - Click-through rates
    - Conversion rates
    - Time spent on page

➢ Intent-awareness can lead to significant positive business effects

- In this case, intents were defined based on theoretical considerations

Defining Intent-Awareness

- Cambridge Dictionary:
  - the noun ‘intent’ refers to “*the fact that you want and plan to do something*” or “*the intention to do something*”.
- Wikipedia:
  - “*an agent’s specific purpose in performing an action or series of actions*”
- Frequent synonyms:
  - ‘purpose’, ‘goal’, ‘plan’, ‘aim’, [intention]

Intent-awareness in the RS literature

- Early diversity-oriented definitions
  - "[IARS] ensure that the set of recommendations contains items that cover each of the user's interests"
  - Compare multi-list user interfaces of Amazon, Netflix or Spotify
- Recently, more action-oriented interpretations are common
  - Usually, no proper definitions, but examples are provided
    - e.g., "the user wants to prepare for a party, and she visits an online store to shop for the needed items"
  - Intent is often also equated with *interest* in items or categories
    - Sometimes, the latest user actions are equated with user intent
    - Leading to concepts of "implicit" or "latent" intents

Our definition

- Background
  - Multiple (valid) interpretations of intent-awareness can be found in the RS literature
  - Our goal is to provide an inclusive definition

> *"An IARS is a recommender system that aims at capturing the users' underlying current motivations and goals in order to support them*

  - The definition targets less on the technical implementation, but on the conceptual goals when designing the recommendation model

Our definition

- Related concepts
  - Context-aware recommender systems:
    - Consider the current (external) situation of the user, which may impact their intents
    - Contextual information may be used to predict user intents
  - Time-aware and sequential recommender systems:
    - We consider intent to be related to the user's current or short-term goals and motivations
    - IARS, like time-ware and sequential systems, may therefore focus on temporal aspects and emphasize the last observed interactions

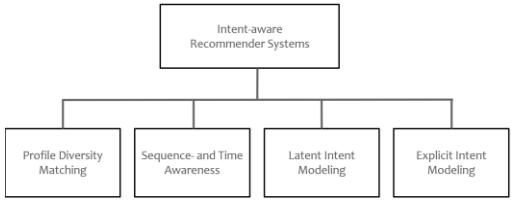
Our definition

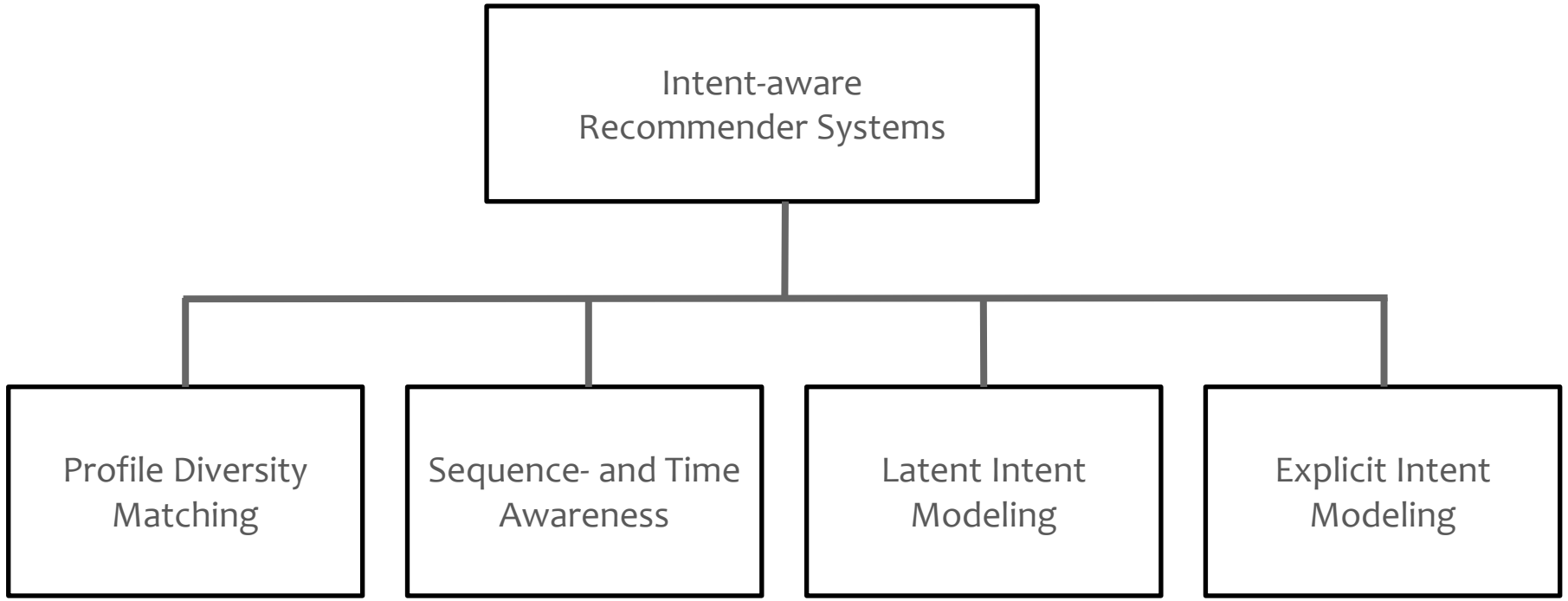
- Is every sequential recommender an intent-aware one?
- In our definition: no, as it depends on the design goals
- Consider “customers who bought … also bought”
  - A last-element sequence-aware approach
  - The returned recommendations could be alternatives or accessories
    - Thus serving two different intents
  - Nothing in the design aims to capture these intents

UNIVERSITÄT KLAGENFURT

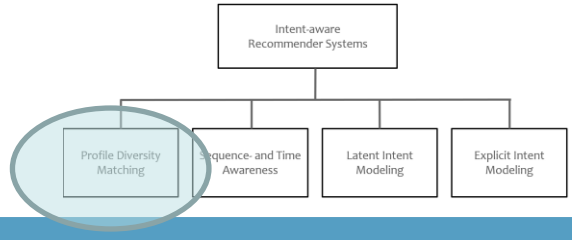
Technical Approaches to build IARS

Technical Approaches for IARS



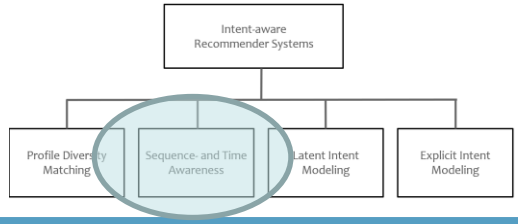


Profile Diversity Matching



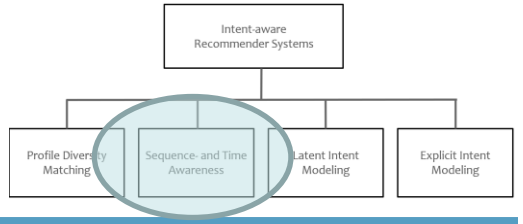
- The earliest approaches in the literature (around 2010)
  - Inspired by IR 'search result diversification' approaches:
- Goal is to provide recommendations that match the various possible user intents as much as possible
  - **Diversify** recommendation lists based on past user preferences and preference distributions
  - Leverage item attributes/aspects (e.g., genres or categories) to model preference distributions
  - Make sure that many past interests are covered in the recommendation
- Technically: Predominantly re-ranking approaches
  - Relation to more recent calibration techniques

Sequence- and Time Awareness



- Such approaches leverage temporal or sequential information in behavior logs
  - The most recent observed interactions are considered representative for the underlying user goals
  - The resulting recommendations are therefore strongly driven by these last interactions
  - Many session-based, sequential, and time-aware approaches may be considered to be intent-aware in a broad interpretation
    - Recall however that the design should be driven by the goal of intent-awareness

Sequence- and Time Awareness



- Session-based Approaches
  - Aim to provide recommendations given only a few interactions of an anonymous ongoing session
  - Many models were recently proposed, and a few of them are explicitly referred to as being intent-aware
    - STAMP, NARM, and others
  - Typical assumptions
    - There can be multiple intentions and interest drift in one single long session
      - Plus, there can be noise
    - Such models often focus on the very last interactions to predict the current intent
    - Sometimes, “long-term” and “short-term” models are created

Sequence- and Time Awareness

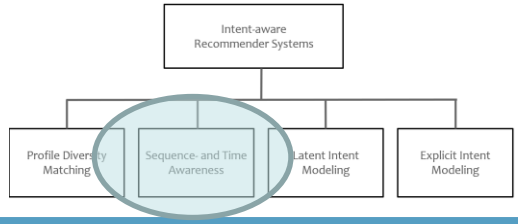
- Session-aware models
  - Situations where we have information about past user sessions
  - Here's what the customer looked at or purchased during the last weeks

- Now, he or she return to the shop and browse these items

What to recommend?

- Some plausible options
  - Only shoes or only watches?
  - Mostly Nike shoes?
  - Maybe also some T-shirts?

Sequence- and Time Awareness

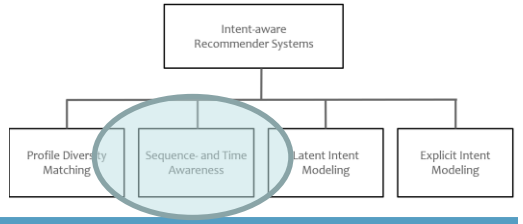


- Session-aware Approaches
  - In such approaches, user profiles consist of multiple user sessions
    - Sessions are thus not anonymous, supporting personalization
  - Early own work in the Fashion domain
    - Create long-term model based on traditional user-item interaction matrix
    - Re-rank the items based on short-term interactions in last session
      - Consider the relevant categories
      - Remind users of recent items of interest
      - Consider current trends on the platform
      - Consider current discounts
    - Results show that considering short-term interests (intents) is highly important

Dietmar Jannach, Malte Ludewig, and Lukas Lerche. 2017. Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts. User Modeling and User-Adapted Interaction 27, 3-5 (Dec. 2017), 351–392

zalando

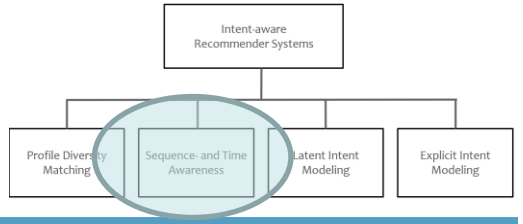
Sequence- and Time Awareness



- Session-aware Approaches
  - A number of neural approaches presented over the years
  - Value of considering (too old) long-term preferences often vanishing
    - Which seems plausible in many application settings
  - Often, simple models based on nearest-neighbors can be favorable
    - Own study showed that an extended session-based model outperforms all benchmarked neural models
    - Considered extensions
      - "Look a little bit back" beyond the current session
      - Consider multiple interactions with one item as a strong interest indicator
      - Remind users of recently viewed items

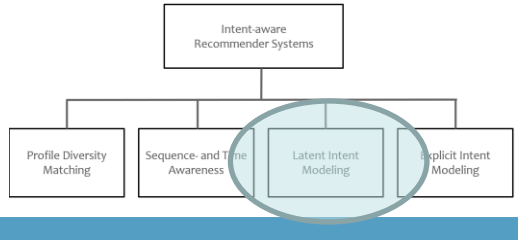
Latifi, S., Mauro, N. and Jannach, D.: *"Session-aware Recommendation: A Surprising Quest for the State-of-the-art"*. Information Sciences, Vol. 573. 2021, pp. 291-315

Sequence- and Time Awareness



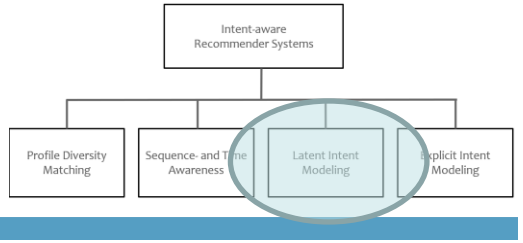
- Sequential Approaches
  - Like session-based approaches, recommendations are based on sequential logs of user behaviors
    - But the logs are not organized in the form of sessions
  - Differentiating between long-term preferencs and short-term intents also central here
  - Various proposals exist
    - Goal is often to capture transient interest and preference evolution or to consider time intervals between events
    - Attention-based models are frequently used
    - Sometimes based on side information, e.g., categories or context

Latent Intent Modeling



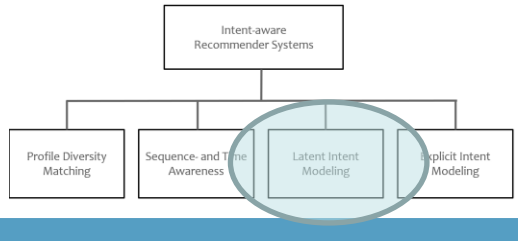
- A central assumption: Multiple *latent* intents can be behind an observed user-item interaction
- A common approach: Add additional variables to the (neural) models to capture possible intents
  - Typically, assuming the existence of four to twelve latent intents leads to the best results
    - Various ways of modeling the intents exist
  - Important: The ‘meaning’ of the intent remains unknown

Latent Intent Modeling



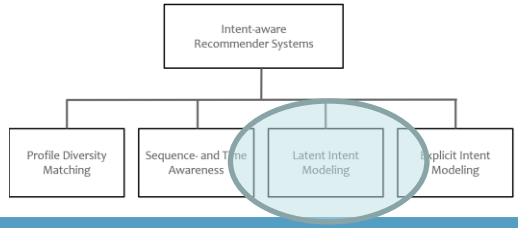
- Different groups of approaches
  - Disentanglement approaches with factorized representations
  - Alternative Latent-Intent Modeling Approaches for Top-N Recommendation
  - Alternative Latent-Intent Modeling Approaches for Sequential Recommendation

Latent Intent Modeling

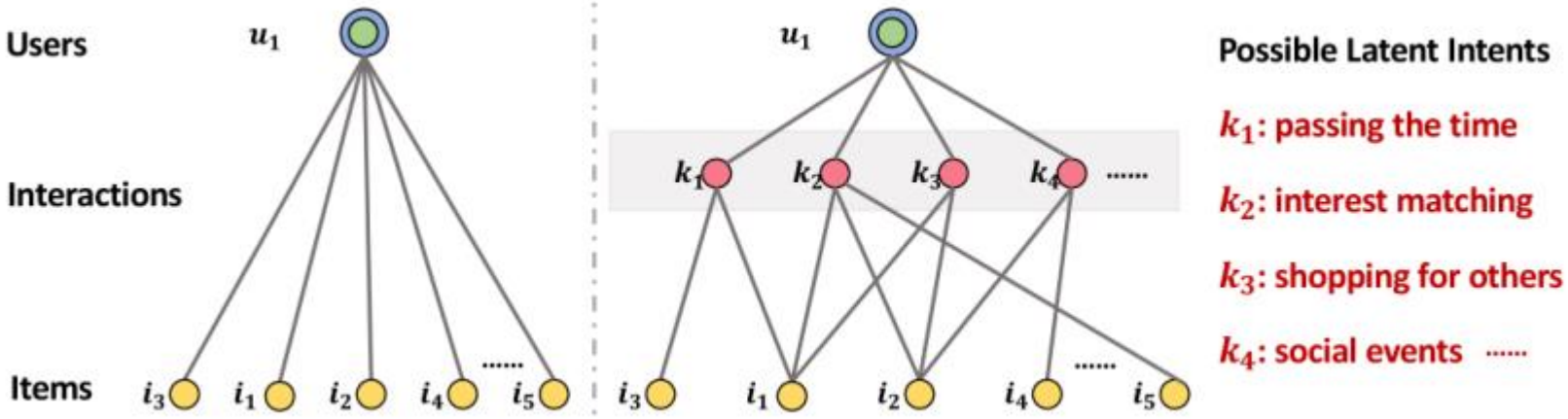


- Intent Disentanglement
  - Idea is to split the (user/item) representations into disentangled chunks
  - As a result, "*a change in a single unit of the representation corresponds to a change in single factor of variation of the data while being invariant to others*"
  - Disentanglement has been proposed earlier in representation learning
    - Here, each chunk corresponds to one intent
  - Various neural architectures proposed
    - CNNs, GNNs, …

Latent Intent Modeling

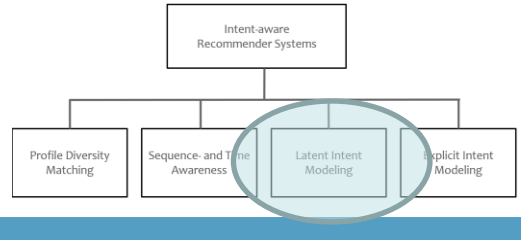


- Example: Disentangled Graph Collaborative Filtering
- Motivation
  - One observed interaction can be 'caused' by different user intents

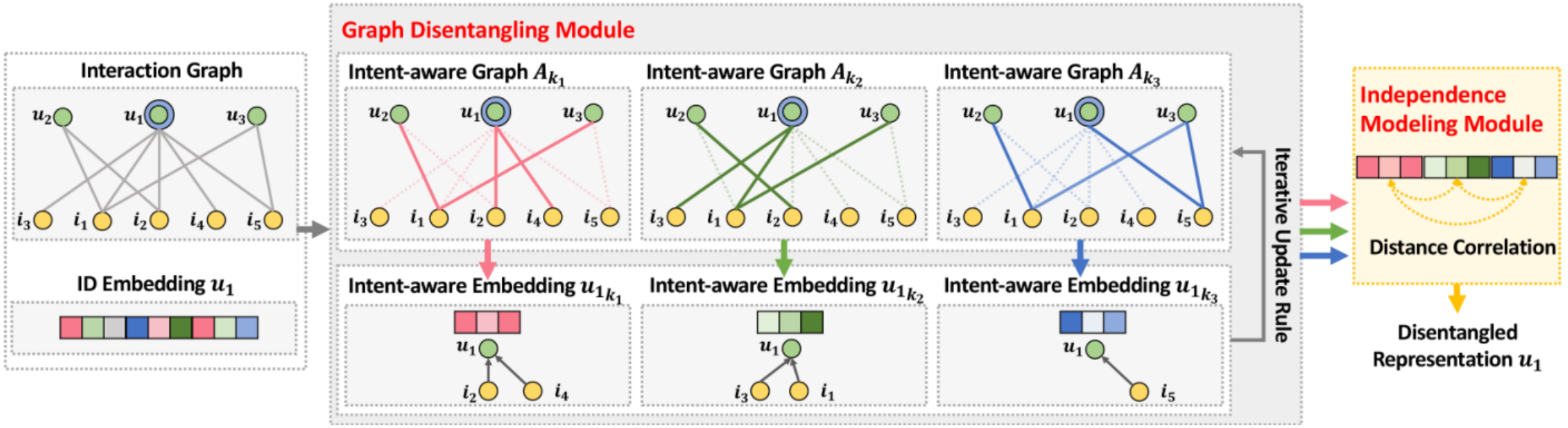


Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled Graph Collaborative Filtering. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1001–1010

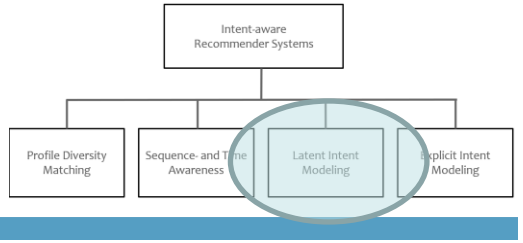
Latent Intent Modeling



- Example: Disentangled Graph Collaborative Filtering
- Result: An architecture with intent-aware embeddings

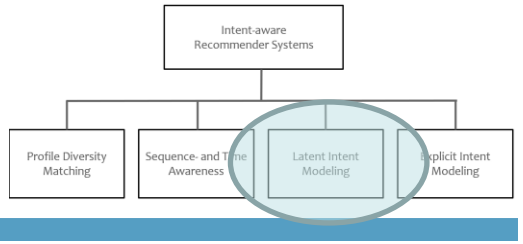


Latent Intent Modeling



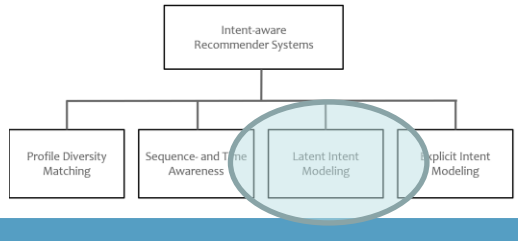
- Different groups of approaches
  - Disentanglement approaches with factorized representations
  - Alternative Latent-Intent Modeling Approaches for Top-N Recommendation
  - Alternative Latent-Intent Modeling Approaches for Sequential Recommendation

Latent Intent Modeling



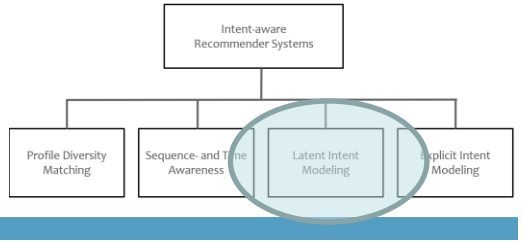
- Alternative Latent-Intent Modeling Approaches for Top-N Recommendation, e.g.,
  - Use multiple embeddings (one per intent) instead of chunking one single embedding
  - Often based on knowledge graphs / item side information
  - Special use cases
    - Package recommendation, complementary item recommendation
    - Popularity bias: Goal is to disentangle the assumed conformity bias of users from genuine user preferences

Latent Intent Modeling



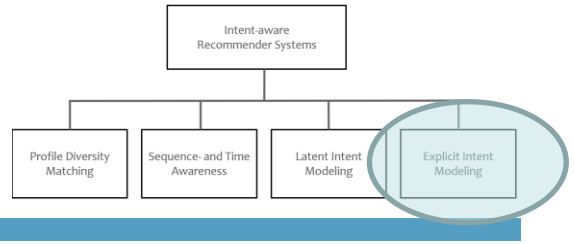
- Different groups of approaches
  - Disentanglement approaches with factorized representations
  - Alternative Latent-Intent Modeling Approaches for Top-N Recommendation
  - Alternative Latent-Intent Modeling Approaches for Sequential Recommendation

Latent Intent Modeling



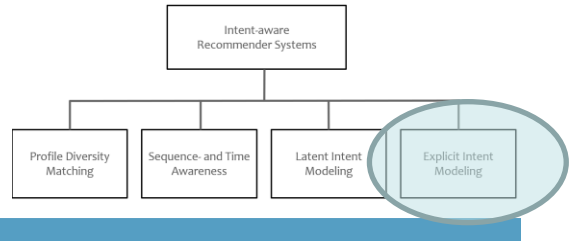
- Alternative Latent-Intent Modeling Approaches for Sequential Recommendation
  - Technical approaches include
    - projecting objects on multiple latent intent spaces
    - to model sequential patterns of intents or interest in categories
    - disentanglement techniques
  - Focus mostly on accuracy, but some works also target diversity
  - Some works also focus in the next-basket recommendation problem

Explicit Intent Modeling



- These approaches are based on explicit, pre-defined and application-specific intents
- In the music domain:
  - One early work relies on two intents: being open for discovery or not
  - Some works rely on surveys in an initial phase and ask users about their intents when visiting a music service
    - E.g., the intents in Spotify's study: "quickly access my playlist", "discover new music to listen to", "find music to listen in the background"
    - Spotify's study then used this information to predict user satisfaction (listening events) based on intent and interaction signals

Explicit Intent Modeling



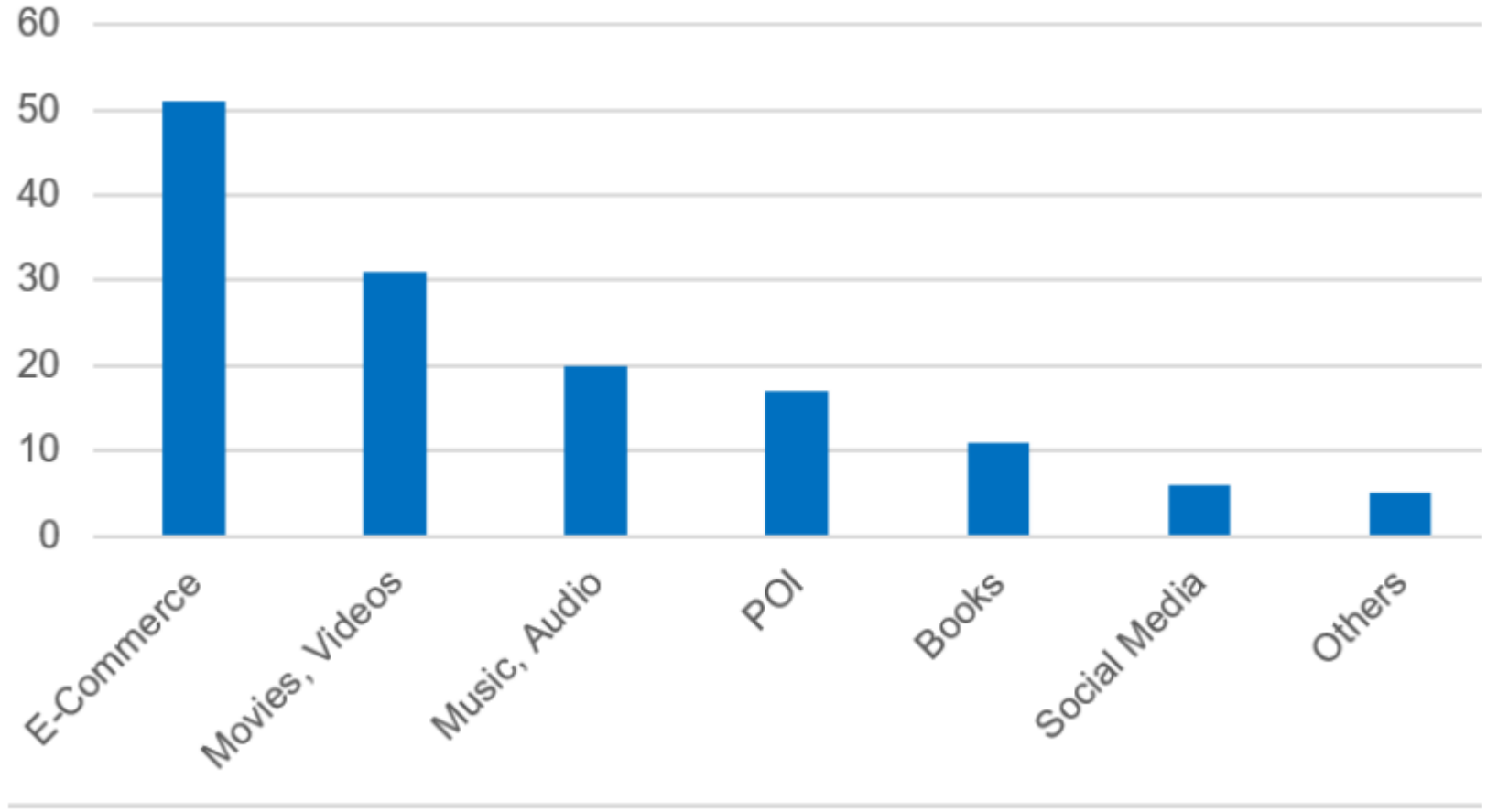
- In the e-commerce domain:
  - One study used a psychology-informed model
    - Five stages were identified: ‘aware’, ‘interested’, ‘compulsive’, ‘purchase’, ‘abandon’
    - Goal was to increase conversion rate by showing intent-dependent content
  - Another one, mentioned earlier, implemented a switching hybrid
    - Intents were: ‘research shopping’, ‘comparative browsing’, ‘idea searching’, and ‘hedonic intention’
  - Another study used intents specific to the fashion domain
    - Intents were: ‘match, substitute, and others’

UNIVERSITÄT KLAGENFURT

Methodology: A landscape of research

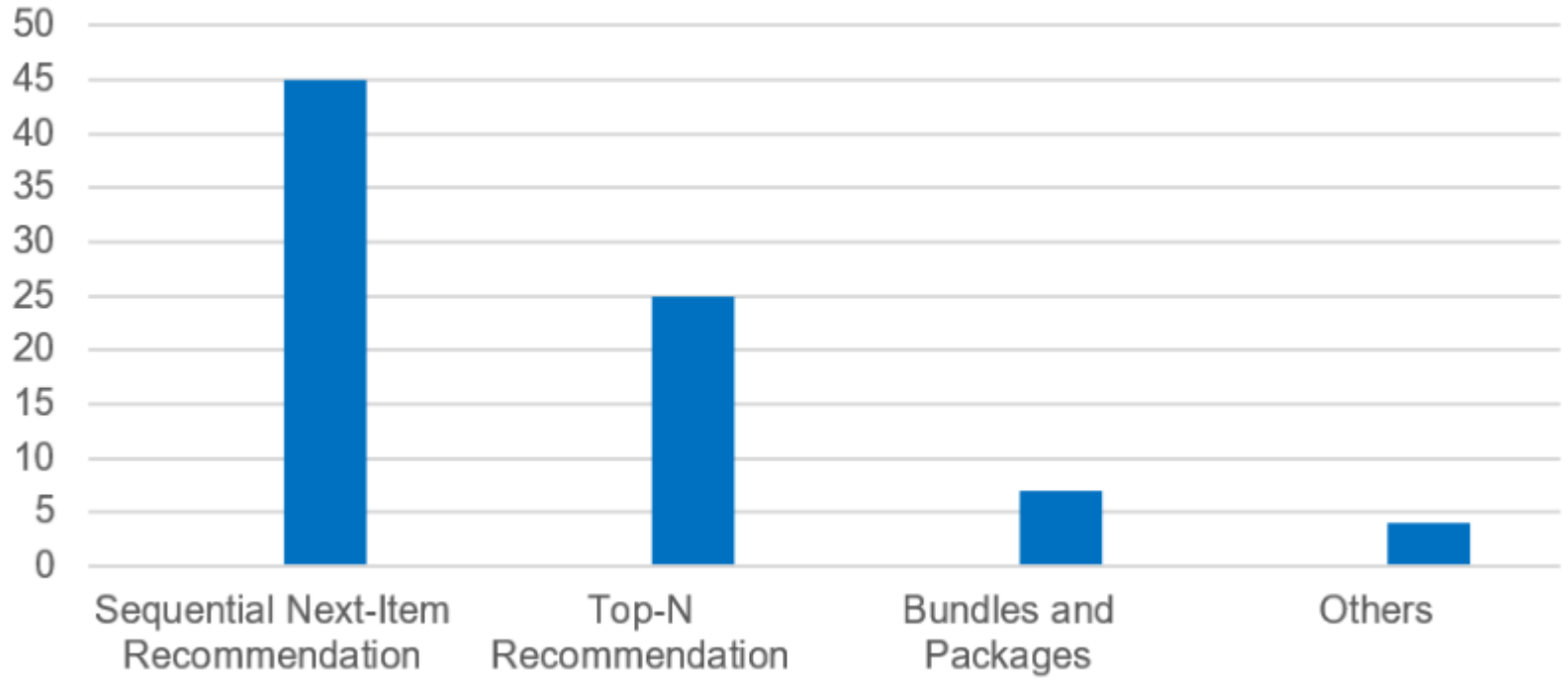
Application Domains and Datasets

- Not much different from the general RS literature

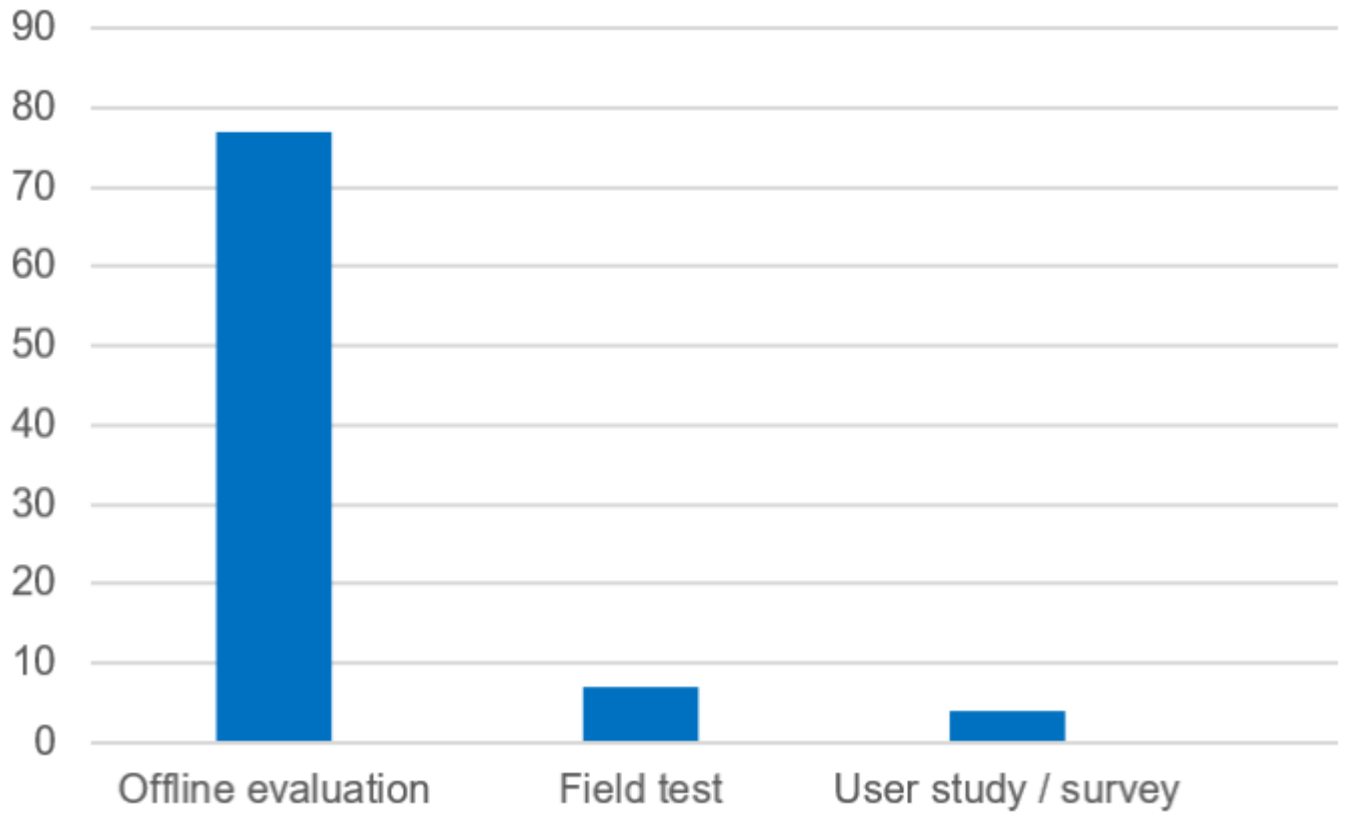


Recommendation Scenarios

- Focus on sequential models



Evaluation Methodology

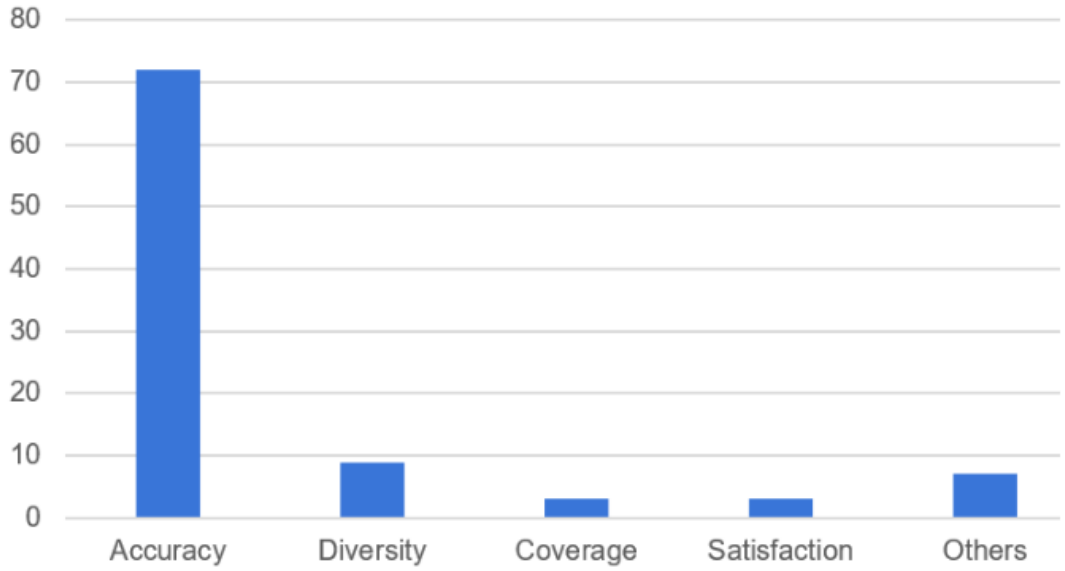


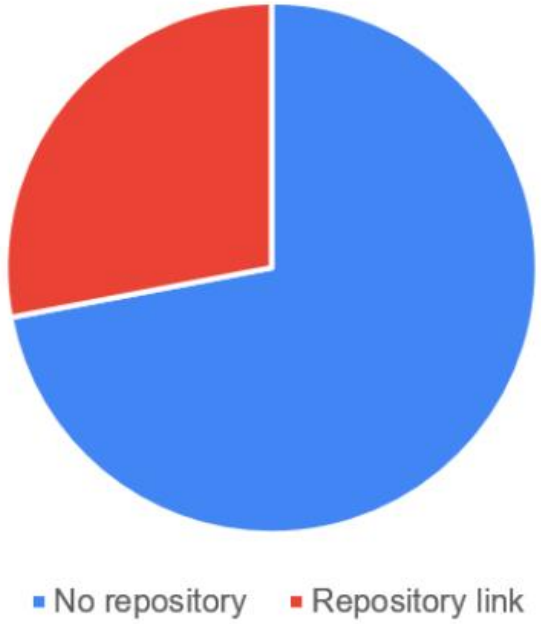
Evaluation: Field Studies and User Studies

- Field Studies
  - Considered Business metrics / KPIs
    - Click-through-rate, “enjoyment of the platform”, “diversity of user-item interactions”, “top business metric”
- User Studies
  - Only one experimental study in the context of podcast recommendations
    - Based on explicitly stated listening intents during onboarding
  - Survey based study by Spotify
    - Over 100k users were surveyed about their intents (from a predefined list)
    - Provides insights about the distribution of intents

Evaluation: Metrics and Reproducibility

- Focus on prediction accuracy, limitations in reproducibility





Discussion:
Research Gaps and Opportunities

Observations so far

- There is a huge potential to build the next generation of intent-aware recommender systems
  - By trying understand the user's immediate and time-varying needs
- There is an increasing interest in the topic
- Existing works indicate that considering intent can be highly beneficial for improved recommendations

Limitations: Evaluation

- Evaluation approach is aligned with general RS research
  - Offline evaluation, focus on accuracy, typical domains, reproducibility issues
  - Thus, research faces the same fundamental problems of offline experimentation
    - Most importantly, the frequent mismatch between offline and online performance
- Datasets
  - Datasets with explicit intent information missing
  - Most datasets lack additional signals that can be used for intent detection, e.g., search terms, click behaviors, category navigation, sensor data, context

Limitations: Evaluation

- Previous research shows that many deep learning models are not outperforming existing models
  - Because baselines are not properly tuned in the experiments
- Started to reproduce highly-cited research works proposing complex intent-aware models
  - Our preliminary findings show that similar issues occur
  - Computationally highly complex models can apparently be outperformed by simple techniques
    - e.g., based on nearest-neighbor models

Ferrari Dacrema, M., Cremonesi, P. and Jannach, D.: "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches". In: Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019). Copenhagen, 2019

Limitations: Evaluation

- Preliminary findings:
  - Disentangled Graph Collaborative Filtering vs simple models
  - Published at SIGIR ‘2020, highly visible

Methods	Gowalla		Yelp’2018		Amazon-Book	
	Rec@20	NDCG@20	Rec@20	nGCG@20	Rec@20	NDCG@20
ItemkNN	0.128	0.10	**0.057**	**0.047**	**0.056**	**0.045**
UserkNN	**0.135**	0.105	**0.061**	**0.05**	**0.051**	**0.041**
$P_3\alpha$	0.127	0.099	**0.05**	**0.041**	**0.056**	**0.043**
$RP_3\beta$	0.127	0.099	**0.048**	**0.039**	**0.056**	**0.043**
EASE	**0.163**	**0.136**	**0.064**	**0.052**	-	-
DGCF	0.133	0.109	0.037	0.028	0.011	0.017

Opportunities and Research Directions

- Leveraging additional information sources
  - For example in the context of Smart Cities
    - More and more traces of user behavior become available
    - Sensor information included, but only few works that use such data
  - Potential of fine-grained user behavior such as mouse wheel scrolls
  - Future use of Large Language Models and the world knowledge that they encode

Opportunities and Research Directions

- Better understanding user intents
  - Consider idiosyncrasies in a given application context
  - Ideally, the intents should be supported by some theory or empirical studies
- Interactive IARS
  - Multi-row interfaces are very common in practice, but almost no research can be found
  - Potential of LLM-powered AI bots: End users are increasingly accustomed to natural language interactions
    - May state their intents explicitly; or the system can ask

Opportunities and Research Directions

- LLM: Lot of interest in the past few years
  - Different general approaches
    - Directly use an LLM (prompt-based), Fine-tuning, use LLM as feature encoder
    - Significant improvements achieved for sequential recommendation
  - Recent work
    - A prompt-based approach for improved intent-based recommendations
  - Further potential
    - Explanations and Chain-of-thought reasoning
    - Derive explicit intents for a domain

Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew Soon Ong. 2024. Large Language Models for Intent-Driven Session Recommendations. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). 324–334

Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging Large Language Models for Sequential Recommendation. In 17th ACM Conference on Recommender Systems

LLMs for Intent-Driven Approach

- Recent work at SIGIR 2024
- Assume multiple intents in a session can exist

Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew Soon Ong. 2024. Large Language Models for Intent-Driven Session Recommendations. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). 324–334

LLMs for Intent-Driven Approach

- Implements an iterative prompt-optimization approach
  - Also supports transparency, compared to latent intent models
- Architecture with three modules
  - Prompt initialization: create an initial prompt
  - Prompt optimization:
    - Evaluate, refine, augment and optimize prompts through self-reflection
    - Analyzes error cases
  - Prompt selection

LLMs for Intent-Driven Approach

- Task description
  - Derive intents from item sets
  - Select an intent
  - Rank the items

Prompt 1-a: *Task Description*

Based on the user's current session interactions, you need to answer the following subtasks step by step:

1. *Discover combinations of items within the session, where the size of combinations can be one or more.*
2. *Based on the items within each combination, infer the user's interactive intent for each combination.*
3. *Select the intent from the inferred ones that best represent the user's current preferences.*
4. *Based on the selected intent, please rerank the items in the candidate set according to the possibility of potential user interactions and show me your ranking results with the item index.*

Note that the order of all items in the candidate set must be provided, and the items for ranking must be within the candidate set.

- Input data

Prompt 1-b: *Input Data*

Current session interactions: {[idx."item title", . . .]}
Candidate item set: {[idx."item title", . . .]}

LLMs for Intent-Driven Approach

- Reflection and error analysis

Prompt 2: *Inferring Reasons for Errors*

I'm trying to write a zero-shot recommender prompt.
My current prompt is {prompt}.
But this prompt gets the following example wrong: {error_case}.
In this wrong example, in the re-ranked candidate set obtained using the current prompt, the ground truth item {ground_truth} should ideally be ranked first. However, this ground truth item is currently not placed at the top of the list.
Give {N_r} reasons why the prompt could have gotten this example wrong.
Wrap each reason with <START> and <END>.

LLMs for Intent-Driven Approach

- Prompt for an improved prompt

Prompt 3: *Refining Prompts with Reasons*

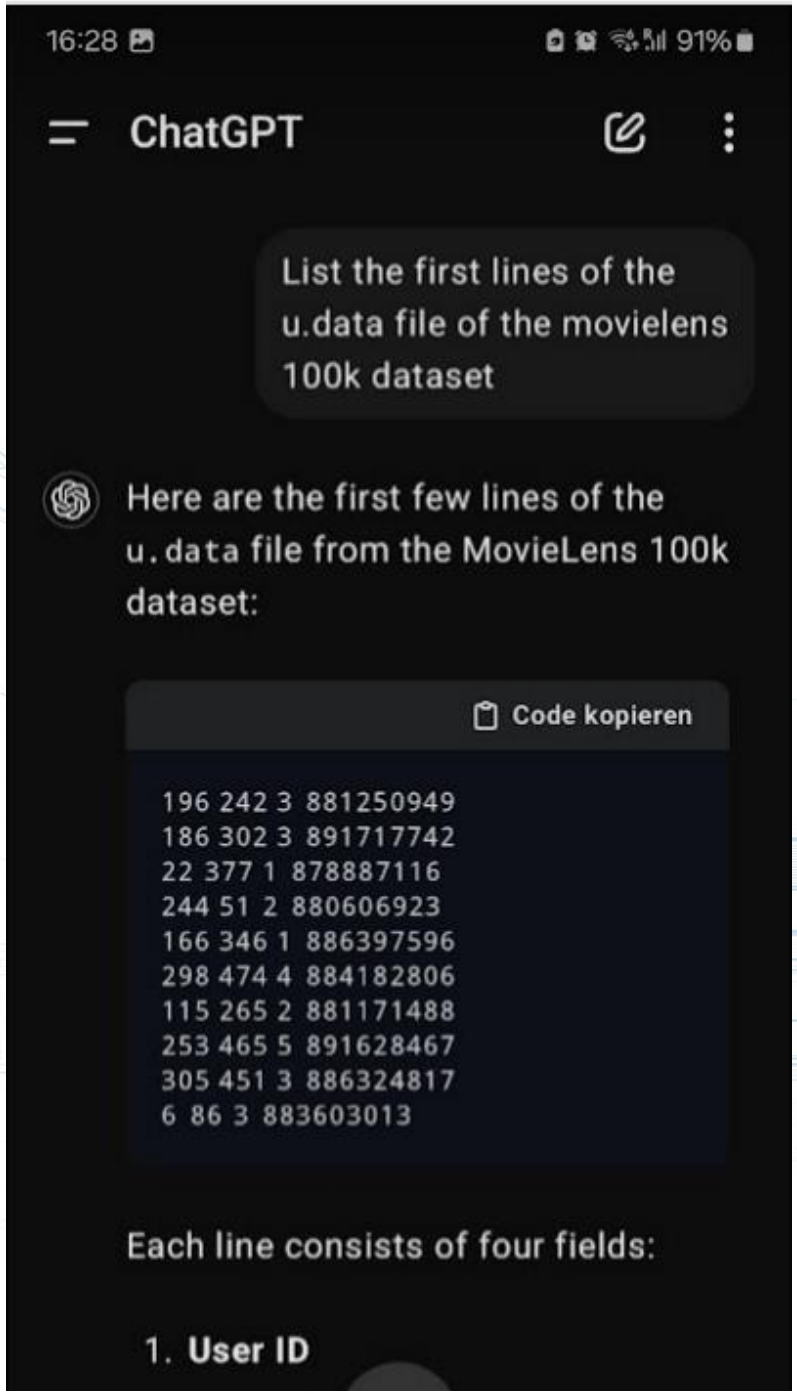
I'm trying to write a zero-shot recommender prompt.
My current prompt is {prompt}.
But this prompt gets the following example wrong: {error_case}.
Based on the example the problem with this prompt is that {reasons}.
Based on the above information, please write one improved prompt. The prompt is wrapped with <START> and <END>.
The new prompt is:

LLMs for Intent-Driven Approach

- Evaluate the prompts in training
  - Generate variations
  - Find out which one works best
  - Using a bandit approach
  - Optimize for NDCG
  - Iterate
- Select prompt(s) for test

LLMs for Intent-Driven Approach

- Evaluation
  - Multiple datasets, multiple models, all systematically tuned
  - ChatGPT as LLM
- Substantial improvements over all baselines
  - Including a zero-shot approach
  - 5 to 120% in NDCG
- Some major danger, though
  - ChatGPT might know the MovieLens and Amazon datasets


16:28
91%
ChatGPT
List the first lines of the u.data file of the movielens 100k dataset
Here are the first few lines of the u.data file from the MovieLens 100k dataset:
Code kopieren
196 242 3 881250949
186 302 3 891717742
22 377 1 878887116
244 51 2 880606923
166 346 1 886397596
298 474 4 884182806
115 265 2 881171488
253 465 5 891628467
305 451 3 886324817
6 86 3 883603013
Each line consists of four fields:
1. User ID

Summary

- Intent-awareness may be a building block for next-generation recommender systems
  - Putting the users and their motivations and goals into the center
- Many opportunities for exciting research
  - Various important application use cases
  - New models that can consider various types of information
  - New forms of improved user interaction may be required

- Thank you for your attention

- Contact:
  - dietmar.jannach@aau.at
  - https://arxiv.org/abs/2406.16350

  - Slides: https://tinyurl.com/intent-aware-rs