DEvIR: Data Collection and Analysis for the Recommendation of Events and Itineraries

Diana Nurbakova, Léa Laporte, Sylvie Calabretto, Jérôme Gensel

DEvIR: Data Collection and Analysis for the Recommendation of Events and Itineraries

Diana Nurbakova and Léa Laporte and Sylvie Calabretto
LIRIS – INSA Lyon
University of Lyon
Villeurbanne, France
{diana.nurbakova, lea.laporte, sylvie.calabretto}@insa-lyon.fr

Jérôme Gensel
Univ. Grenoble Alpes, CNRS,
Grenoble INP, LIG
Grenoble, France
jerome.gensel@imag.fr

Abstract

Distributed events such as multi-day festivals and conventions attract thousands of attendees. Their programs are usually very dense, which makes it difficult for users to select activities to perform. Recent works have proposed event and itinerary recommendation algorithms to solve this problem. Although several datasets have been made available for the evaluation of event recommendation algorithms, they do not suit well for the case of distributed events or itinerary recommendation. Based on the study of available online resources, we define dataset attributes required to perform event and itinerary recommendations in the context of distributed events, and discuss the compliance of existing datasets to these requirements. Revealing the lack of publicly available datasets with desired features, we describe a data collection process to acquire the publicly available data from a major comic book convention website. We present the characteristics of the collected data and discuss its usability for evaluating recommendation algorithms.

1. Introduction

Multi-days conventions, festivals and congresses attract thousands of attendees, providing venues for like-minded people to get absorbed in the universe of the ideas, concepts, and artforms to the promotion of which such events are dedicated. As they unite under one umbrella hundreds or thousands of sub-events distributed in space and time, they are usually referred to in the literature as *distributed events* [1]. Their programs are usually very dense, consisting of multiple short-lived events happening in parallel. Examples of such distributed events are comic book conventions like Comic-Con International: San Diego (comic-con.org), Nights of Museums, *e.g.* Lange Nacht der Museen in Berlin (lange-nacht-der-museen.de), music festivals *e.g.* Coachella (coachella.com), Pinkpop (pinkpop.nl), etc.

Looking for a unique experience, attendees of such distributed events are overwhelmed with the amount of parallel options and have troubles with organising their time. Recommender systems have appeared as the most accurate solution to handle these issues. Recent work have thus focused on proposing novel approaches for events [2, 3] and itineraries recommendation [1, 4].

Event recommendation aims at providing a user with a list of future events that may represent the highest interest for him/her. In this context, an event is generally defined as a planned activity that is valid for a limited time and takes place in a specific location. Event recommendation is usually considered as a more complex problem than item recommendation (such as book or movie) [2]. Indeed, as events have a short lifetime and must be recommended to users before they actually occur, no explicit relevance judgments are available, contrary to item or POI for which one may have ratings of users that have already buy the item or visit the POI. To handle this issue, in Event Based Social Networks, researchers have proposed to consider RSVP as binary indicators of interest. RSVP is the French acronym for "Répondez s'il vous plaît", meaning "Please, answer". On EBSN, RSVP are used by user to indicate their intent to attend an event. In the literature, event recommendation is usually considered as a top-k recommendation problem. It can be formulated as a list-wise [2, 5] or pair-wise [3] ranking problem.

Contrary to event recommendation that considers events independently, *itinerary recommendation* during distributed events aims at recommending a whole sequence of consecutive events [4, 1]. Taking into account temporal constraints is crucial in this context. The problem consists in finding an event sequence (or itinerary) that maximises the user's satisfaction with attended events while taking into account the spatio-temporal constraints that guarantee the feasibility of the undertaken sequence (*e.g.* limited availability of events, simultaneous events, travelling time, etc.).

In the context of a distributed event, the process to decide which sub-events or activities to undertake becomes more constrained than in the case of a

traditional event recommendation. Indeed, the amount of activities may be higher than usual, while activities occur in parallel. Moreover, the sub-events are unique, short-lived and gathered under the umbrella of a general theme of the event. Itinerary recommendation thus differs from single event recommendation [2, 3]. Therefore, the existing datasets created based on the crawling of event-based social networks (EBSNs, *e.g.* Meetup, Plancast, Eventbrite) do not completely reflect this scenario. Moreover, most of them are missing crucial attributes for the evaluation of itineraries recommendation, making them inaccurate for a realistic evaluation of events and itineraries recommender systems. However, more and more distributed events provide their program online, together with some scheduling/agenda applications. This not only allows the attendees to discover the program and select the activities they wish to attend, but also to share their selection publicly with the other attendees on the event website. One typical example of this kind is the International Comic-Con: San Diego, a comic book convention taking place every year since 1970 in San Diego, California. Since 2013, they provide their program online using an agenda application, allowing their users to manage their agenda and to indicate on the website if they plan to attend an activity.

In this paper, we investigate the use of such kind of resources to collect data for a realistic evaluation of event and itinerary recommendation during distributed events. Our main contributions are as follows:

1. The definition and description of the required and most desired attributes of datasets for the evaluation of event and itineraries recommender systems, when users' attendances to past events are unknown (only RSVP are available).

2. The description of the data collection process and the dataset construction based on a well-known large distributed event, the comic book convention (Comic-Con International: San Diego).

3. The demonstration of the usability and limits of this kind of data for event and itinerary recommendation through its deep analysis.

The remainder of the paper is organised as follows. In Section 2, based on a review of related work and the reminder of some definitions and formulations of the related problems, we propose a set of attributes required to construct a dataset for evaluating event and itinerary recommendation approaches. In Section 3, we describe the distributed event, the data collection process and the main statistics of the data. Section 4 provides some insights about the use of the collected data for recommendation of events and discusses its advantages and limitations. Section 5 concludes the paper.

2. Definition of dataset requirements

In this section, we define the attributes a dataset should possess for a proper evaluation of events and itineraries recommendation approaches, in the context of distributed events. Based on a review of related work, we extract some required attributes and some additional attributes. We first present the background and related work, then discuss about the choice of the attributes. Finally, we present the existing datasets and discuss their usability and limits for the evaluation of events and itineraries recommendation.

2.1. Background

The background of this work can be defined with respect to two main axes. The first axis refers to the problems of recommending single spatial items such as *Point-of-Interest* and *events*, while the second axis refers to the recommendation of sequences of POIs or events, namely the Trip recommendation and the itineraries recommendation problems.

POI and Event recommendation. Several kinds of recommendation problems exist in the literature (refer to [6] for a comprehensive review). In this paper, we focus on the recommendation of spatial items, namely Point-of-Interests (POIs) and Events, which has attracted a lot of interest in the industry and research community during the past decades [7, 8, 9, 3, 2, 5]. POI and event recommendation are two closely related problems, whose differences rely on the intrinsic nature of the items to be recommended. A *Point of Interest* can be considered, in a generic way, as "a human construct, describing what can be found at a location"[1]. Thus, in the literature [7, 10, 11, 12], a POI is generally defined by its name and location (geo-coordinates), and additional attributes such as keywords, opening and closing hours, temporal constraints, etc. POI recommendation is usually seen as a top-k recommendation problem, whose goal is to provide a given user with a ranked list of not already visited POIs [9] that he/she may be interested in at a specific time [8], based on his/her current location and the POIs visited before. Such recommender systems usually use check-in information data or POI ratings from users in order to recommend POIs to other users. *Events*, such as festivals, concerts or talks, intrinsically differ from POI due to their ephemeral

[1] http://www.w3.org/2010/POI/wiki/Main_Page

nature. Definition 1 provides a formal definition of an event.

Definition 1. *An **event** e is a unique social occasion of a limited duration that takes place in a specific location during the specific time window. Thus, it can be represented as a tuple $e =< id, n, desc, l, \delta, start\_time, end\_time, category >$, where id is the event identifier, n is its title, $desc$ denotes its textual description, l indicated the geographical location the event takes place, δ stands for its duration, $start_time$ and end_time define the time window of its availability, and $category$ is the list of categories associated with the event.*

The event recommendation problem aims at providing a user with a list of future events she may be interested in, taking into account her preferences and constraints (location, time, etc).

Due to their short lifetime, events have to be recommended before being attended, meaning that no ratings are available, contrary to more classical recommendation scenarios such as POI recommendation. In the literature, some authors have proposed to counter this drawback by considering RSVP and context information [2]. RSVP are special actions on EBSN that users can perform to indicate their intent to attend an event or not. RSVPs can be provided by users once the event has been made publicly available on the social network or a website and until its actual occurrence. Thus, they can be used as binary indicators of interest by the recommendation approaches. The underlying assumption is that there exists a binary matrix $\mathcal{M}_{|U|\times|E|}$ of the users' historical data, which reflects the users' intentions to join the events, *i.e.* $\mathcal{M}_{ue} = \begin{cases} 1, & \text{if } u \text{ intents to join } e \\ 0, & \text{, otherwise} \end{cases}$. The ground truth (*i.e.* real-life users' traces) about people attending the events usually is not available at the time of recommendation. Thus, it is assumed that a user who saved an event to his/her agenda is more likely to attend this event than a user who did not save it. Therefore, a positive intention to attend an event expressed by a user via saving it is considered to be a proxy value of his/her actual attendance of this event.

Trip and Itinerary recommendation The Trip recommendation problem has recently attracted vivid interest from the industry and the research community in the tourism domain. It aims at finding an optimal trip route (a sequence of POIs to visit), which maximises the score of user's happiness, collected by visiting the POIs [13]. To the best of our knowledge, there are two main ways to address the trip recommendation problem in the literature. The first group of approaches divides the problem into two steps [13, 1]. First, the estimation of the user's interest in a POI is performed. Second, the problem is formulated as an instance of the Orienteering problem [14] or an extension of the Traveling Salesman problem [15] in order to find an optimal path. The second kind of approaches consists in estimating the transition probabilities from one POI to another, and chaining the most probable transitions in order to make a suitable sequence [16]. While trip recommendation aims at providing the user with a sequence of POIs, the itinerary recommendation problem considers sequences of events. For sake of clarity, we reproduce in definitions 2 and 3, the definitions of an itinerary and the itinerary recommendation problem respectively, as initially proposed in [4].

Definition 2. *An **itinerary** $\xi(u) = e_i \rightarrow e_j \rightarrow ... \rightarrow e_m$ is a chronologically ordered series of events of the user* $\mathtt{u} \in \mathtt{U}$ *that satisfies the set of constraints:*
(1) Event availability: an event can be joined only within the time window of its availability.
(2) Time budget: the total time needed to attend all the events within an itinerary should not exceed the time budget.
(3) Activity completion: a user may join an event if there is enough time to perform it completely.

Definition 3. *The **itinerary recommendation problem** consists in providing a given user $u \in U$ with a feasible **itinerary** ξ that maximises his/her **satisfaction** with all attended events: $Max\ \sigma(\xi, u)$, subject to the set of **constraints**.*

Note that in definition 3, the overall satisfaction over a sequence of events is defined as the sum of scores for all the events that compose the sequence. Similarly to trip recommendation, the itinerary recommendation problem can be addressed through a two-step process. Thus, a recent approach by [4] proposes in the first phase to consider several influences (content, category, time of event) in order to estimate the satisfaction score of an event for a given user. In the second phase, they build a sequence of events based on an iterative local search algorithm for Orienteering Problem with Time Windows (OP-TW), enhanced with the transition probabilities between the categories of events.

2.2. Definition of dataset attributes

Based on the definition of events, itineraries and the related problems, we propose to extract a list of dataset attributes that are required for the evaluation of event and itinerary recommender systems, in particular in the context of distributed events. We distinguish between attributes that are required for both event and itinerary

recommendation, attributes required only for itinerary recommendation and additional attributes that can be used, but are not required, for both tasks.

Common required attributes. As event and itinerary recommendation problems are both dealing with events, some common attributes related to events are required for both tasks. Thus, the name, location, date and starting time of the event must be included in the dataset for evaluation. Regarding the location of an event, geographical coordinates, if not initially available, will have to be extracted based on the postal addresses, since they are used to compute the distance between the user and the place where the event takes place. As most of existing approaches make use of textual content to recommend the events, textual resources about the events, such as its description and its category, are required. Historical user-event data such as RSVP is also required for recommender systems in order to estimate the users' interest in new events and to evaluate the performance of recommendation algorithms.

Attributes for itinerary recommendation mainly. As itinerary recommendation focuses on recommending a sequence of events, some attributes related to specific sequential constraints are required. If we consider the three constraints in definition 3 that have to be fulfilled in order to create a feasible sequence of events, it appears that the temporal constraints are crucial in this context. Indeed, not only the date and starting time of an event are required, but also its end time and duration. Note that the literature [4, 1] distinguishes between the time window of an event and its duration: the time windows refers to the time interval of event availability while the duration refers to the actual time needed to perform the activity during the event. For example, if we consider a book signing event during a convention, the time window may be from 2 p.m. to 5 p.m. (3 hours in total), but the time needed for an attendee to have his/her own book signed by the author (the duration) may be only 5 minutes. Duration is often referred to as *service time* when dealing with itinerary construction, in particular when considering Orienteering Problems [14].

Additional attributes. Some additional attributes can also be used to enhance the recommendation process. Among the most desired attributes, we can list the users-users relations. Indeed, recent works on event recommendation [2] have shown that the group a user belongs to or the relatives and friends he is spending time with may have an impact on his choice of events to attend. Other works have also shown that the personality of users have an impact on the items (movies, books, etc) they consumed [17]. Other attributes, such as the price of an event or age limitation may also be considered.

Table 1. Comparison of available datasets.

Dataset	TW	(x, y)	s	Cat	$Hist$
Meetup [2]		✓		✓	✓
Flickr [18]		✓			✓
Foursquare [19]		✓			✓
TripBuilder [20]		✓		✓	✓
OP-TW [14]	✓	✓	✓		

Notations: TW - time Windows, (x, y) - location, s - service time, Cat - categories, $Hist$ - historical user-item data

2.3. Usability of existing datasets

In recent years, a number of datasets has been made available for researchers with the purpose of reproducibility of recommendation algorithms. A short summary of the datasets with respect to the data attributes is given in table 2.3. Most of them are issued from logs of Location-Based Social Networks, *e.g.* Meetup [2], Flickr [18], Foursquare [19]. Though these datasets reflect real-world user behaviour and can be used for event recommendation, they are not fully adapted to benchmark algorithms for itinerary recommendation, especially in the context of distributed events. Indeed, as pointed out on Tab. 2.3, they do not contain the time window nor the service time of events, which are crucial attributes to construct feasible sequences. Another group of datasets consists of synthetic instances to benchmark algorithms for solving optimisation problems, *e.g.* OP-TW and its extensions [14]. Such datasets lack user and item-related attributes, therefore could be hardly solely used for personalised recommendation purpose. There is thus a need for new datasets that possess all the required attributes.

3. Data Collection and Description

In the following, we describe the data collection process undertaken in order to acquire a new **d**ataset for **ev**ent and **i**tinerary **r**ecommendation (**DEvIR**) and some statistics. Our main focus is recommendation during distributed events. Therefore, we have chosen one of the biggest conventions, namely Comic-Con International: San Diego, further referred to as *Convention*, which provides an online schedule of all programmed events together with the RSVPs of users. Thus, DEvIR aims at fulfilling the lack of datasets for the recommendation of single events and itineraries during distributed events.

3.1. Data Collection

The procedure of the data collection is mainly based on the crawl of the official website of the Convention[2]. It has undergone the following process. In August 2017, we crawled the official website of the convention in order to retrieve the 2013-2017 programs of events and available data about event attendance, namely the lists of events pre-selected by users. By 'users' we denote the users of the scheduling application *Sched* (https://sched.com) used by the Convention organisers. In order to mark the events the users would like to attend, they are invited to create their Sched accounts or use their Facebook account to sign up. The users may restrict the access to their profile making it '*private*' in order to hide their identity. The users that have marked an event in their custom schedule appear in the '*Attendees*' section of the corresponding event page. Private users are displayed with 'Private' icon with no further information provided. It should be noted that the users do not rate the events.

The Convention website gives access to all programs of the past and ongoing events starting from 2013. We iteratively crawled the program pages for editions 2013-2017, as well as all the corresponding event and user pages. Following the described procedure, we could create a dataset that consists of the following entities[3]:

- *event*: a core entity, containing a list of events from the Convention programs.
- user: a core entity, containing a list of users registered at Sched who expressed their intentions to attend events. Please, note that for privacy concerns, we have anonymised user names and ids in the dataset.
- *location*: a list of venues where the events take place. We enrich the crawled data with X and Y coordinates, and the address of the corresponding buildings queried from Google Maps (maps.google.com).
- *tag*: a list of event custom-based tags.
- *category*: a hierarchical list of event types (categories). The categories are organised into a two-level structure, where the parent elements represent the main categories (tracks) of the convention or the service categories (*e.g.* 1: Programs, 2: Anime, or U: Updated).
- *event-user*: users RSVPs, indicating users' intentions to attend the events, expressed by a binary attribute `value`.
- *user-user*: a list of user-user pairs who appear on the user's pages in the friends list, where `value` is a binary relation value.
- *event-category*: a list of event-category pairs `[event_id,event_name,category_link,value]`.
- *event-tag*: a list of event-tag pairs `[event_id, event_name, tag_link, value]`.
- *location-location*: distance matrix between locations, reflecting travelling (walking) time between the buildings of the corresponding buildings queried using Google Maps API.

Figure 3.1 depicts a class diagram of dataset entities. Table 2 provides a detailed description of attributes of the entities with examples. We did not include `event-category` and `event-tag`, to Table 2, as these entities can be considered secondary, being derivatives from the lists of categories and tags of events. We mention them separately in order to provide a better representation of relations between entities. Similarly, `location-location` entity can be considered secondary, as it is not issued from the original crawl of the website.

Note on event duration. In Section 2.2, we have stated that event *duration* (also referred to as *service time*) is an attribute required for an accurate itinerary recommendation. Note that the actual event duration is often empirical and may vary. Event organisers may provide approximate event duration, or only indicate the time window of event availability. The Convention program does not explicitly indicate event duration, except for a few events. Therefore, when not explicitly indicated in the event description, the duration has been assigned as follows:

- *External source based*: For the events with known or approximately known duration (*e.g.* film, series, board games, etc.), we have queried external sources to obtain the time length, *e.g.* IMDB (imdb.com), MyAnimeList (myanimelist.net), BoardGameGeek (boardgamegeek.com), YouTube (youtube.com), etc.
- *Default value*: We have assumed the default value to be equal to the time difference between `end_time` and `start_time`. This value has

[2]Mind, that the website content can be used only in non-commercial purpose, as the copyright is hold by the Convention. Here, we are presenting the data collection process to undertake.

[3]The dataset is available at: https://github.com/ecafidid/DEvIR

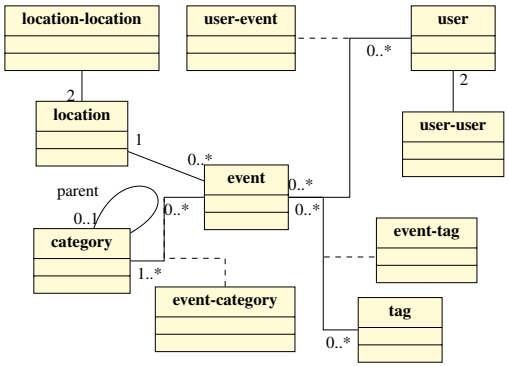

Figure 1. Diagram of DEvIR.

been assigned to the panel sessions, and most of the events of the type '1: Program'.

- *Approximation*: We have assigned approximate value based on the main type of the event: '3: Autograph'[4] - 60, '8: Retailers' - 30, '7: Portfolio Review'[5] - 20. The value selection Was performed so that it fits the minimum time window of a given event type and was motivated by attendance rules and procedures described on the website and attendees reports[6].

3.2. Dataset Analytics

The general statistics of DEvIR are displayed in Tab. 3. While calculating the number of users and event-user pairs, we removed 'private' users, as we cannot distinguish between them. In the following we investigate some of the characteristics of the data.

The attendance of events is not uniform over different events. Thus, the variations of the number of RSVPs provided by users for events are depicted in Fig. 3. Moreover, one can note that among all the possible options, the users select 23 events in average for the whole duration of the Convention (see Tab. 3).

The Convention is a multi-day event with numerous activities each day (see Tab. 4). Thus, the maximum number of events (640) in 2013-2017 was achieved on the 3rd day of 2016 edition. Such an amount of options makes it very hard for attendees to select events of their interest and attend them. The number of events per user follows the power-law distribution with $\alpha = 3.841$ (see Fig. 2)[7]. In addition, Table 5 shows the average number of events selected by users per day. Note that the distribution of the event selection by users is not uniform

[4]https://www.comic-con.org/cci/2017/autographs
[5]https://www.comic-con.org/cci/2018/portfolio-review
[6]https://www.wired.com/2015/07/nerdist-comic-con-guide/
[7]For fitting, we used `powerlaw` Python package [21]

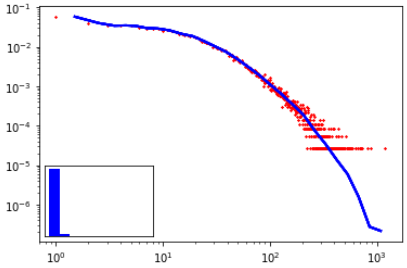

Figure 2. Distribution of the number of events per user.

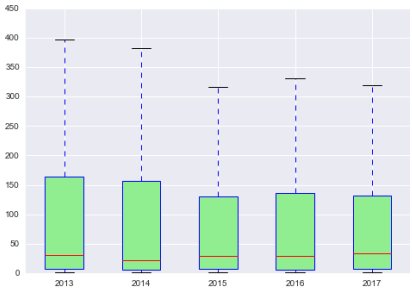

Figure 3. Number of RSVPs per event per year.

over the days. Thus, the average attendance of the events reaches its peak on the 3rd day, while the events of the first day lack participants. This is due to the fact that the first day hosts only few events (see Tab. 4).

Due to the high popularity of the Convention, many attendees return to the next editions of the convention. Table 7 summarises the number of users participation in different editions of the Convention. It can be noted that more than 700 users have taken part in all the editions of the Convention, according to the collected data .

Another characteristic of DEvIR describes the range of overlapping events within the program of the convention. By the overlapping range of events we understand the number of simultaneously happening events. In order to estimate it, we divide a day into 15-minute timeslots. An event is considered to occur within a timeslot, if the time window of its availability is partially or fully covered by the timeslot (see Fig. 4).

Similarly, for each timeslot we analyse the number of simultaneous events the users intent to take part in. We exclude from the analysis the timeslots with no scheduled events. Thus, the average number of parallel activities is 37. The maximum number of parallel events is 112 and was attained on the 3rd day of 2016 edition during the timeslot 16h15-16h30. The average number of activities selected by a user in a given timeslot is 1.5. These characteristics emphasise the selection problem faced by the users, since they tend to select events that have a high probability to occur

Table 2. Dataset description.

entity	attributes	attribute description	example
event	year	year of the event occurrence	2017
	day	day of the event	3
	id	identifier	550b10edc277c7477eef06d4a6c76c5f
	name	name	Fata Morgana
	link	link of the event	/event/BSRu/fata-morgana
	description	textual description	Held in AA26: Fata Morgana Steven Boyett and Ken Mitchroney
	time	scheduling time of the event (string)	Friday July 21, 2017 10:00am - 2:30pm
	start_time	start time	2017-07-21 10:00:00
	end_time	end time	2017-07-21 14:30:00
	duration	service time in min	60
	location	venue of the event	Sails Pavilion - Autographs
	location_link	link to the venue	/venue/Sails+Pavilion+-+Autographs
	event_type	list of categories associated with the event	[['3: Autographs', '/type/3%3A+autographs'], ['Group Signing', '/type/3%3A+autographs/group+signing']]
	event_tag	list of event tags	[['Held in: AA26', '/tag/Held+in%3A+AA26']]
user	year	list of participation years	[2013, 2017]
	id	identifier	8
	user_name	user name (anonymised)	user000008
	user_link	link to the user page	/user000008
	page_name	name of user page	user000008
	about	personal description	
location	year	list of years of venue use	[2013, 2014]
	location	venue name	Marriott Hall 6, Marriott Marquis & Marina
	location_link	link to the venue	/venue/Marriott+Hall+6%2C+Marriott+Marquis+%26+Marina
	address	address	333 W Harbor Dr, San Diego, CA 92101, USA
	x_coordinate	x-coordinate (latitude)	32.7084733
	y_coordinate	y-coordinate (longitude)	-117.16742250000001
category	year	list of years of category use	[2013, 2014, 2015, 2016, 2017]
	category_name	category name	Action Figures - Toys - Collectibles
	category_link	link to the category	/type/1%3A+programs/action+figures+-+toys+-+collectibles
	parent_link	link to the parent category	/type/1%3A+programs/
tag	year	list of years the tag was used	[2014.0, 2015.0, 2017.0]
	tag_name	tag name	Ticketed Events
	tag_link	link of the tag	/tag/Ticketed+Events
event-user	year	year of relation	2017
	event_id	id of the event	550b10edc277c7477eef06d4a6c76c5f
	user_id	user_link of the user	/user000008
	user_name	user_name of the user	user000008
	value	binary RSVP value	1
user-user	year	year of relation	2017
	user_1	user_link of the 1st user	/user000008
	user_2	user_link of the 2nd user	/user000268
	value	binary relation value	1

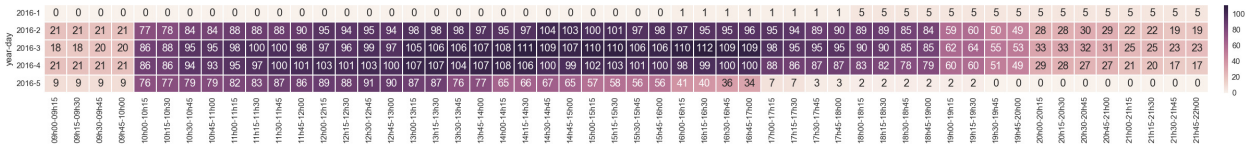

Figure 4. Heatmap of the overlapping events at 2016 edition with respect to 15min timeslots from 6am to 10pm.

Table 3. General statistics of the DEvIR dataset.

	2013	2014	2015	2016	2017
#events	1,760	1,880	2,038	2,184	1,909
#locations	38	38	47	45	47
#categories	115	179	210	213	179
#tags	50	164	191	197	235
#users	11,147	10,945	9,033	10,697	9,001
#user-user	7,818	6,710	4,119	5,220	4,388
#user-event	249,439	247,003	220,565	250,396	202,244
avg. duration	53.81	52.87	55.88	50.4	51.26

Table 4. Number of events per day.

	1	2	3	4	5	6
2013	15	465	505	500	275	–
2014	10	507	542	531	290	–
2015	1	24	555	588	574	296
2016	16	602	640	624	302	–
2017	16	513	567	533	280	–

Table 5. Mean and standard deviation of the number of events per user per day.

	1	2	3	4	5	6
2013	1.07 (0.43)	7.94 (7.99)	8.25 (8.19)	8.1 (7.89)	4.55 (4.51)	–
2014	1.05 (0.44)	9.01 (9.25)	8.82 (8.67)	8.22 (7.93)	4.52 (4.54)	–
2015	1.0 (0.0)	1.19 (0.7)	9.6 (9.66)	9.51 (9.02)	9.16 (8.8)	4.46 (4.3)
2016	1.14 (0.71)	9.12 (9.12)	9.34 (9.17)	9.12 (9.05)	4.6 (4.77)	–
2017	1.23 (0.65)	9.19 (9.06)	9.49 (9.16)	8.32 (8.07)	4.41 (4.29)	–

Table 6. Ratio of the user's events shared with friends to the total amount of the user's events.

2013	2014	2015	2016	2017
0.161	0.140	0.159	0.145	0.126

Table 7. Number of the users' participation.

# editions	1	2	3	4	5
# users	27,451	5,007	1,923	1,001	714

simultaneously. Moreover, we note the presence of the attendance bias, as we do not have access to the users' real-life attendance of the events at the Convention, and the users do not rate events, so that the user's explicit preference of one event over another is unknown.

The users may indicate their '*friends*'. In DEvIR, 5,150 users out of 36,100 have listed at least one friend (referred to as 'users with friends'). The maximum number of friends is 108, while the average number of friends per user is 3.54 (among users with friends). For the users with friends, we have estimated the ratio of events shared with their friends to the total amount of the user's events (Tab 6). Note that only 15% of the users events get RSVP from the user's friends.

4. Use for Recommendation

4.1. Evaluation Protocol

We aim at mirroring a realistic scenario where for each user we generate an ordered list of events. Recommendations are calculated on a daily basis, *i.e.* for each day of a convention edition. We use the *Precision at rank k* (P@k) as the evaluation metric.

In the experiments, we divide the data into train and test sets by adapting the evaluation protocol suggested in [3] as follows. The train set includes the 4 editions of 2013-2016 available in DEvIR. The test set includes the data of 2017. As the recommendations are calculated on the daily basis, we gradually extend the training set with the data from previous days, *i.e.* for the recommendation for the n^{th} day, the users profiles are modelled based on the users past events from the train set and the days ranging in $(1, n-1)$. It has to be noted that 3,785 out of 9,000 users who expressed their interest in taking part in 2017 edition have taken part in at least one previous edition of the convention. For the users who have previously attended the Convention, there exist historical data that can be used to create their profiles on the train set. In this paper, we focus on these users.

4.2. Event Recommendation

In our experiments we use one non-personalised and two personalised recommendation algorithms.

Popularity-based (Pop). Similar to [2], we rank the candidate events in the descending order of their popularity, *i.e.* the number of users who expressed their intention to join the event.

Content-based (CB). We represent each event using bag-of-words TF-IDF of their description, then we compute the cosine similarity to estimate the similarity between upcoming events and a user's profile. We model a user's profile $\vec{u}$ similar to [2], *i.e.*: $\vec{u} := \sum_{e \in E_u} \frac{1}{(1+\alpha)^{\tau(e)}} \times \vec{e}$, where E_u is the set of user's past events, e is an event representation using TF-IDF, α is a time decay factor (we set $\alpha = 0.01$ similar to [2]), and $\tau(e)$ returns the number of years between the current events and the user's past events. The content-based score is therefore calculated based on the cosine similarity between the current event and the user's profile, *i.e.*: $\hat{s}_{cb}(u, e) = cos(\vec{u}, \vec{e})$.

Category-based (Cat). Each event is associated with a list of categories. The categories are organised into a 2-level hierarchy. Thus, we distinguish between 12 *main categories* (*i.e.* the categories that are on the top of the hierarchy, the attribute `parent_link` is null) that we denote $\mathcal{C}_{main}$ and 453 *child categories*, denoted $\mathcal{C}_{child}$. We represent each event as a 1×465-dimension binary vector of categories $cat(e)$. We model a category-based user's profile as follows: $\vec{u} := agg_{e \in E_u}\big(\frac{1}{(1+\alpha)^{\tau(e)}} \times cat(e)\big)$, where agg denotes an aggregation function (we used the mean in our experiments), $cat(e)$ is a vector composed of (1) $cat_{main}(e)$, *i.e.* the elements of the main-category vector of the event e, and (2) $\frac{cat_{child}(e)}{|\mathcal{C}_{child}(e)|}$, where $cat_{child}(e)$ denotes the elements of the child-category vector of the event e, $|\mathcal{C}_{child}(e)|$ denotes the number of child categories that the event e is assigned to. Generally speaking, such representation of the user's profile reflects the weighted frequency of the categories of the events attended by the user in the past. The category-based score is then calculated based on the cosine similarity between the current event and the user's profile, *i.e.*: $\hat{s}_{cat}(u, e) = cos(\vec{u}, \vec{e})$.

Table 8 presents the precision P@10 of the recommendation algorithms on the DEvIR dataset. The rank 10 has been selected based on the average number of events selected by users per day (see Table 5). As it can be seen, the precision is rather low, but there exist a positive trend with the increase of the number of historical days. This can be explained by the sparsity of data. Moreover, the low precision of the Day 1 can be explained by a low number of events and weak attendance on the first day, which is coherent with the statistics given in Table 4-5. In contrast, the highest precision is reached on Day 3. It can be explained by the highest number of the events available on this day, and the highest attendance of the events (see Table 4-5). Another surprising finding is the importance of the popularity factor, as popularity-based method outperforms the others for Days 2-4.

Table 8. Results of the three considered recommendation techniques in terms of Precision@10.

	Day 1	Day 2	Day 3	Day 4	Day 5
Pop	0.0118	0.0568	0.0767	0.0756	0.0428
Cat	0.0138	0.0376	0.0551	0.0531	0.0489
CB	0.0095	0.0312	0.0557	0.0543	0.0495

However, the top-k results retrieved by the algorithms mentioned above do not take into account the time availability constraint, *i.e.* they do not deal with conflicts when two or more recommended events are happening simultaneously. In the context of a distributed event, time constraints and limited availability of events become crucial. Therefore, it is relevant to construct personalised itineraries of events. We leave this question for future work.

5. Discussion and Conclusion

In this paper, we have presented a data collection process and a case study performed on a new dataset for the recommendation of events and itineraries during big distributed events, such as big conventions, festivals, etc. We have provided the details about the data collection and have described the general characteristics. The first insight we can make consists in the evidence of the conflicts of the users' interest in events. Thus, we have seen that the number of overlapping events is rather high and the users select a couple of events happening at the same time. Moreover, the lack of data about the users' real-life attendance contributes to the attendance bias. Such settings are common for EBSNs, (Meetup, Facebook Events etc.), where users express their intention to join proposed events by providing their RSVP, and therefore, the data available for analysis and recommendation. An assumption commonly made in the field of event recommendation is that the users' RSVP indicating the intentions in joining events may be considered as a proxy value of attendance [2, 3].

The second characteristic that we have remarked in DEvIR implies the lack of relevant graded preference of

one event over another for a given user. It is common to the datasets with implicit feedback [22]. Moreover, we have applied a few recommendation techniques to the dataset to show its use for the recommendation purpose. We hope that a dataset collected in the way described in the paper will serve as a test dataset for the new approaches to the recommendation of events and itineraries during distributed events. Our future work mainly consists in a proposal and implementation of novel prediction techniques for both, event and itinerary recommendation, that could improve the prediction quality of the models. One of the possible methodologies is to use sequence pattern mining.

6. Acknowledgments

D. Nurbakova held a doctoral fellowship from Région Auvergne-Rhône-Alpes. Additional support was received from the Franco-German University.

References

[1] R. Schaller, M. Harvey, and D. Elsweiler, "Recsys for distributed events: Investigating the influence of recommendations on visitor plans," in *SIGIR'13*, pp. 953–956, 2013.

[2] A. Q. Macedo, L. B. Marinho, and R. L. Santos, "Context-aware event recommendation in event-based social networks," in *RecSys '15*, pp. 123–130, 2015.

[3] E. Minkov, B. Charrow, J. Ledlie, S. Teller, and T. Jaakkola, "Collaborative future event recommendation," in *CIKM '10*, pp. 819–828, 2010.

[4] D. Nurbakova, L. Laporte, S. Calabretto, and J. Gensel, "Recommendation of short-term activity sequences during distributed events," *Procedia Computer Science*, vol. 108, no. ICCS'17, Supplement C, pp. 2069 – 2078, 2017.

[5] Z. Qiao, P. Zhang, Y. Cao, C. Zhou, L. Guo, and B. Fang, "Combining heterogenous social and geographical information for event recommendation," in *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pp. 145–151, 2014.

[6] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*. Springer Publishing Company, Incorporated, 2nd ed., 2015.

[7] H. Yin, X. Zhou, Y. Shao, H. Wang, and S. Sadiq, "Joint modeling of user check-in behaviors for point-of-interest recommendation," in *Proc. of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pp. 1631–1640, 2015.

[8] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proc. of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 659–668, 2014.

[9] Y. Liu, W. Wei, A. Sun, and C. Miao, "Exploiting geographical neighborhood characteristics for location recommendation," in *Proc. of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 739–748, 2014.

[10] I. R. Brilhante, J. A. F. de Macêdo, F. M. Nardini, R. Perego, and C. Renso, "On planning sightseeing tours with tripbuilder," *Inf. Process. Manage.*, vol. 51, no. 2, pp. 1–15, 2015.

[11] W. Chen, L. Zhao, X. Jiajie, K. Zheng, and X. Zhou, *WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I*, ch. Ranking Based Activity Trajectory Search, pp. 170–185. Cham: Springer International Publishing, 2014.

[12] A. Rae, V. Murdock, A. Popescu, and H. Bouchard, "Mining the web for points of interest," in *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pp. 711–720, ACM, 2012.

[13] C. Zhang, H. Liang, K. Wang, and J. Sun, "Personalized trip recommendation with poi availability and uncertain traveling time," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pp. 911–920, 2015.

[14] P. Vansteenwegen, W. Souffriau, and D. V. Oudheusden, "The orienteering problem: A survey," *European Journal of Operational Research*, vol. 209, no. 1, pp. 1 – 10, 2011.

[15] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 151–158, 2016.

[16] J. Sang, T. Mei, and C. Xu, "Activity sensor: Check-in usage mining for local recommendation," *ACM Trans. Intell. Syst. Technol.*, vol. 6, pp. 41:1–41:24, Apr. 2015.

[17] B. Ferwerda, M. Schedl, and M. Tkalcic, "Personality traits and the relationship with (non-) disclosure behavior on facebook," in *Proc. of the 25th International Conference Companion on World Wide Web*, pp. 565–568, 2016.

[18] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[19] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM TIST*, vol. 7, no. 3, p. 30, 2016.

[20] I. Brilhante, J. A. Macedo, F. M. Nardini, R. Perego, and C. Renso, "Where shall we go today?: Planning touristic tours with tripbuilder," in *Proc. of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pp. 757–762, 2013.

[21] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: A python package for analysis of heavy-tailed distributions," *PLOS ONE*, vol. 9, pp. 1–11, 01 2014.

[22] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *ICDM'08*, pp. 263–272, 2008.