HAL
open science

Sequence Mining within Formal Concept Analysis for Analyzing Visitor Trajectories

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli, Chedy Raïssi

To cite this version:

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli, Chedy Raïssi. Sequence Mining within Formal Concept Analysis for Analyzing Visitor Trajectories. SMAP 2018 - 13th International Workshop on Semantic and Social Media Adaptation and Personalization, Sep 2018, Zaragoza, Spain. hal-01887927

HAL Id: hal-01887927

https://inria.hal.science/hal-01887927

Submitted on 4 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sequence Mining within Formal Concept Analysis for Analyzing Visitor Trajectories

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli, Chedy Raïssi
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Email: firstname.lastname@loria.fr

***Abstract*—This paper presents our work about mining visitor trajectories, within the framework of CrossCult European Project about cultural heritage. We present a theoretical and practical research work about the characterization of visitor trajectories and the mining of these trajectories as sequences. The mining process is based on two approaches, namely the mining of subsequences without any constraint and the mining of frequent contiguous subsequences. Both approaches are defined within Formal Concept Analysis and its extension pattern structures. In parallel, a similarity measure allows us to build a hierarchical classification which is used for interpretation and characterization of the trajectories w.r.t. four well-known visiting styles in museum studies.**

I. Introduction

This paper is related to the CrossCult European Project about cultural heritage (http://www.crosscult.eu/). The general idea of CrossCult is to support the emergence of a European cultural heritage by allowing visitors in different locations (e.g. museum, city, archaeological site) to consider their visit at a European level by using adapted computer-based devices.

In this project, we are mainly interested in the analysis of visitor trajectories and recommendation. The trajectory of a visitor in a specific location is considered as a multi-dimensional sequence depending on a number of variables, such as space (e.g. paths, rooms, environment), time (e.g. hour, day, season, news), history and geography (e.g. town, region, country...). Moreover, additional domain knowledge and general knowledge bases such as DBpedia, Freebase, or YAGO can be reused to draw inferences and improve the quality of both analysis and recommendation.

Here, we have two main objectives, (i) the mining of visitor trajectories based on sequence mining, and (ii) the characterization of a trajectory in terms of the subsequences which are mined. We assume that the subsequences can be related to the visiting styles, the visit content, and the environment. Thus subsequences can be used for analyzing the trajectory of a visitor and for making recommendations all along the visit. Moreover, the occurrences of some subsequences at a given moment within a trajectory can witness a change of behavior –which in turn triggers a change in the recommendations.

In the present paper, we discuss theoretical and practical work about the definition of visitor trajectories and the mining of these trajectories as sequences. The mining process is based on two approaches about sequence mining in Formal Concept Analysis (FCA [1]): `MRGS` for "Mining Rare General Subsequences" [2] and `MFCS` for "Mining Frequent Contiguous Subsequences" [3]. The first approach mines rare subsequences in a general way, i.e. gaps may appear in the subsequences, while the second approach searches for frequent subsequences without any gap (a kind of substrings). We also reuse the similarity measure sim_{ACS} developed for analyzing the trajectories of patients between hospitals [4], [5]. If the original paper about `MRGS` [2] was interested in rare subsequences, this is no more the case here and we work on frequent subsequences as well. This similarity measure allows us to build a hierarchical classification that will play a role of "reference classification". For analyzing and interpreting the trajectories of visitors, it is interesting to compare the outputs of `MRGS` and `MFCS` algorithms w.r.t. the clustering produced by sim_{ACS}. Moreover, these outputs and the clustering are analyzed thanks to four theoretical visiting styles, namely "ant", "butterfly", "fish" and "grasshopper" [6].

Several challenges are faced in this research work in the FCA framework: the mining of complex sequential data and dynamics in adapting two algorithms based on pattern structures, the analysis of the trajectories based on jumping emerging patterns and clustering. Here, data are not necessarily big but are rather complex and multidimensional, and this should be taken into account.

The paper is organized as follows. Section II recalls the basic definitions about sequence mining that are useful for understanding the present work. Then, Section III presents the characteristics of the dataset that was used as a basis for the current work. In Section V and Section VI, we present respectively the application of clustering on data enabling to build classes of visitors, and the application of two algorithms for mining interesting subsequences. Finally, an interpretation of the results and a discussion on the characterization of the visitor trajectories is given in Section VII.

II. The Mining of Sequences

A. Basic Definitions

Pattern mining is the task of finding repeated occurrences in a dataset. For example, in a data about customer transactions, an objective can be to find a set of items that are frequently ordered in a single transaction. This specific task in pattern mining is related to sequential pattern mining. We recall below the basic definitions that we will need.

Definition 1. A sequence is an ordered list $\langle s_1 s_2 \dots s_m \rangle$, where s_i is an itemset $\{i_1, \dots, i_n\}$, and m is the *size* of a sequence. The *length* of a sequence is the total number of items, i.e. $\sum |s_i|$.

978-1-5386-8225-8/18/$31.00 ©2018 European Union

TABLE I: A formal context for four visitors, with four items: `102`, `302`, `402`, and `704` as an example. An $\times$ indicates that the visitor visits the item.

	`102`	`302`	`402`	`704`
V_1				$\times$
V_2	$\times$		$\times$	
V_3	$\times$	$\times$	$\times$	
V_4		$\times$	$\times$	$\times$

Definition 2. A sequence $s = \langle s_1 s_2 \ldots s_m \rangle$ is a subsequence of sequence $s' = \langle s'_1 s'_2 \ldots s'_n \rangle$, denoted by $s \preceq s'$, if there exist indices $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ such that $s_j \subseteq s'_{i_j}$ for all $j = 1 \ldots m$ and $m \leq n$.

Therefore, the sequence $\langle \{a\}\{d\} \rangle$ is a subsequence of $\langle \{a,b\}\{a,c,d\} \rangle$, while sequence $\langle \{c\}\{d\} \rangle$ is not.

One way of evaluating the quality of a subsequence is to compute the support of the subsequence. Given a user-defined threshold, the subsequence can be "frequent", i.e. the support is above the threshold.

Definition 3. Let $\mathcal{S}$ be a database of sequences. The *support* of a sequence s in $\mathcal{S}$ is: $support(s, \mathcal{S}) = |\{s_i \in \mathcal{S}; s \preceq s_i\}|$

There exist algorithms that can retrieve all frequent sequences [7], [8]. Beside mining frequent sequences, another complex task is finding homogeneous sequence groups (clustering). To achieve such a task, a distance or a similarity measure between two sequences has to be defined. The similarity measure sim_{ACS} was proposed in [5], which counts the number of all common subsequences (ACS), formulated as:

$$sim_{ACS}(S_i, S_j) = \frac{\phi_C(S_i, S_j)}{\max\{\phi_D(S_i), \phi_D(S_j)\}}$$

where $\phi_C(S_i, S_j)$ is the number of all common distinct subsequences between S_i and S_j, while $\phi_D(S_i)$ is the number of all distinct subsequences of S_i.

B. FCA and Pattern Structures

Formal concept analysis (FCA) is a mathematical framework based on lattice theory and used for classification, data analysis, and knowledge discovery [1]. From a formal context, FCA detects all formal concepts, and arranges them in a concept lattice.

Definition 4. A formal context is a triple (G, M, I), where G is a set of objects, M is a set of attributes, and I is a binary relation between G and M, i.e. $I \subseteq G \times M$.

If an object g has an attribute m, then $(g, m) \in I$. An example of a formal context is shown in Table I. This table shows whether a visitor (V_1–V_4) visits an item (`102`, `302`, `402`, or `704`).

The Galois connection for a formal context (G, M, I) is defined as follows:

Definition 5. For a subset of objects $A \subseteq G$, A' is the set of attributes that are possessed by all objects in A, i.e.:

$$A' = \{m \in M | \forall g \in A, (g, m) \in I\}, \qquad A \subseteq G$$

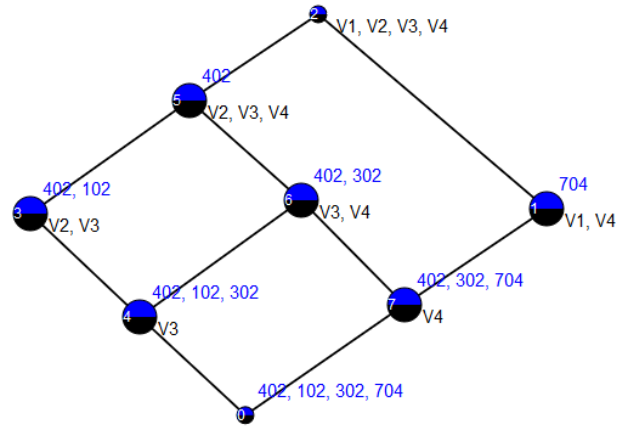

Fig. 1: Concept lattice for the formal context in Table I

Dually, for a subset of attributes $B \subseteq M$, B' is the set of objects that have all attributes in B, i.e.:

$$B' = \{g \in G | \forall m \in B, (g, m) \in I\}, \qquad B \subseteq M$$

Definition 6. A formal concept is a pair (A, B), where $A \subseteq G$ and $B \subseteq M$, and such that $A' = B$ and $B' = A$.

A formal concept (A, B) is a *subconcept* of (C, D) – denoted by $(A, B) \leq (C, D)$ – if $A \subseteq C$ (or equivalently $D \subseteq B$). A concept lattice can be formed using the $\leq$ relation which defines the order among concepts. For the context in Table I, the formal concepts and their corresponding lattice are shown in Fig. 1.

FCA is restricted to specific datasets where each attribute is binary (e.g. has only yes/no value). For more complex values (e.g. numbers, strings, trees, graphs...), FCA is then generalized into pattern structures [12].

Definition 7. A pattern structure is a triple $(G, (D, \sqcap), \delta)$, where G is a set of objects, $(D, \sqcap)$ is a complete meet-semilattice of descriptions, and $\delta : G \to D$ maps an object to a description.

The operator $\sqcap$ is a similarity operation that returns the common elements between any two descriptions. A description can be a set, a sequence, or other complex structure. In the case of set as a description, $\sqcap$ corresponds to set intersection ($\cap$), i.e. $\{a, b, c\} \sqcap \{a, b, d\} = \{a, b\}$, and $\sqsubseteq$ corresponds to subset inclusion ($\subseteq$). In the case of sequence as a description, $\sqcap$ is a set of common closed subsequences (SCCS) [2]. Similarly, $\sqsubseteq$ corresponds to subsequence inclusion ($\preceq$).

Definition 8. The Galois connection for a pattern structure $(G, (D, \sqcap), \delta)$ is defined as:

$$A^\diamond = \bigsqcap_{g \in A} \delta(g), \qquad A \subseteq G$$

$$d^\diamond = \{g \in G | d \sqsubseteq \delta(g)\}, \qquad d \in D$$

Finally, a pattern concept is similar to a standard formal concept:

Definition 9. A pattern concept is a pair (A, d), $A \subseteq G$ and $d \in D$, where $A^\diamond = d$ and $d^\diamond = A$.

TABLE II: An example of one visitor trajectory.

Start time	End time	Item name
12:55:39	12:58:05	Crafts and Arts
12:58:06	12:58:22	Religion and Cult
12:58:22	12:58:27	Building Methods and Facilities
12:58:29	13:05:09	Wooden Tools

Examples of pattern concepts are shown in Table V. There are two definitions of $\sqcap$ for sequence, and will be explained in Section VI-A.

C. Sequence Mining in FCA

In this section we briefly present the two algorithms that are adapted for mining the trajectories of visitors in a museum, namely MFCS [3] and MRGS [2]. The names of the algorithms are not used as such in the papers but here we use them by commodity. Both algorithms are original and very efficient, and among the few algorithms performing sequence mining in the framework of FCA.

MFCS was originally introduced for mining trajectories of patients in hospitals. One important characteristic of MFCS is that it mines contiguous subsequences, or stated differently, subsequences without any gap between items. This is due to the fact that physicians are mainly interested in consecutive events when analyzing healthcare trajectories. In addition, but this is not needed in our framework, MFCS is able to take into account a partial ordering – given by domain knowledge for example – defined on the items composing the sequences.

MRGS is also a sequence miner based on pattern structures but with a different purpose. The objective of MRGS is to mine rare rather than frequent subsequences, and in particular long subsequences with special characteristics. The algorithm is based on a specific pattern structure of subsequences, where the similarity operation is based on the discovery of common close subsequences (SCCS operation is illustrated in a next section). The SCCS operation is based on a directed graph of alignments (DAG of alignments) which guides the mining of common subsequences. The algorithm shows very good performances and is most probably one of the few algorithms whose objective is the mining of rare subsequences. In our framework, we adapted MRGS and the support threshold for comparison purposes with frequent subsequences. However, in our context we will use MRGS as a standard sequence miner and we will be interested in frequent subsequences.

III. The Dataset of Museum Visitors

A. The Museum

In the framework of the CrossCult project, we are working on a specific dataset about the trajectories of 254 visitors in Hecht Museum in Haifa, Israel [13]. In the raw dataset, a visitor trajectory contains a list of visited items, where each visit is composed of three elements: "start time", "end time", and "item name". An example is presented in Table II. When modeling trajectories into sequences, in this paper we consider only the "item name", so every itemset contains only one item. For simplicity, we omit the curly brackets to describe an itemset. Therefore we will write $\langle\{a\}\{d\}\{e\}\rangle$ as $\langle a, d, e\rangle$.

TABLE III: Grouping of museum items.

Category	Items and their ID
1	Entrance Reuben Hecht (101), Symbols Jewish Menorah (102), Persian Cult (103), Jerusalem Photo (104)
2	Religion and Cult (201), Everyday Pottery (202), Phoenician Writing (203), Burial Tradition (204), Building Methods and Facilities (205), Maritime Commerce (206), Imported Pottery (207), Crafts and Arts (208)

TABLE IV: Examples of visitor trajectories.

Visitor	Trajectory
V_1	$\langle$101, 101, 401, 704$\rangle$
V_2	$\langle$102, 402, 808, 206, 808$\rangle$
V_3	$\langle$302, 102, 201, 302, 705, 402, 802$\rangle$
V_4	$\langle$104, 704, 602, 302, 402, 103$\rangle$

A visitor can have visits with various time lengths. In order to obtain more meaningful results and to reduce the complexity, we only consider visits lasting at least 90 seconds, but this is a parameter than can be relaxed or more constrained. Thirty-eight trajectories have no visit more than this threshold, so they are ignored, leaving us with 216 trajectories. Moreover, we model each trajectory as a sequence of visited items. Therefore, for trajectory in Table II, the corresponding sequence is $\langle$Crafts and Arts,Wooden Tools$\rangle$. This preprocessing results in sequences of various size. Forty-five sequences have only one itemset, while three sequences have more than 15 itemsets.

We group the museum items according to their location, so that we obtain 8 categories of items. To illustrate the numbering of items, the first two categories and their items are listed in Table III. We convert the raw dataset into sequences of items, where each item is represented by its ID. We define the IDs such that we can infer the category of an item by its first digit. Therefore, we obtain a dataset of 216 sequences of visitor trajectories (named V_1–V_{216}) where each sequence is composed by a list of IDs, as illustrated in Table IV.

MFCS and MRGS algorithms are suitable for our dataset, since we want to include the hierarchical information in the mining process. For example, given two sequences $\langle 102, 203\rangle$ and $\langle 103, 204\rangle$, it is interesting to mine subsequence $\langle 1, 2\rangle$.

B. The Four Visiting Styles

In a seminal work about the typing of visitor styles in a museum [6], four main behaviors have been detected and described, leading to different recommendations all along a visit [14], [15]. These four styles are summarized below:

- The *ant* is a visitor who will surely see all the works following their location order in the museum. Then the recommendation can be the following item, but depending also on some environmental factors such as the crowd in the museum, the accessibility of the item and the fatigue of the visitor.
- The *grasshopper* is a visitor who will see only certain artworks, jumping from one to another. Then, to encourage such a person to visit more items, the

recommendation can be to visit items having a content similar to items already visited.

- The *butterfly* is a visitor wanting to discover some and not all artworks, without having any exact preferences. Then, the recommendation is open and can be based on surprise (items which are very different one from the other).
- The *fish* is a visitor who does not feel that much interested in the artworks and stays most of the time in the center of the rooms without any precise objective. Then the recommendation can be to visit the most famous items in the museum which are the closer to the current visitor location, for encouraging the visitor to continue the visit and gain more interest.

Indeed, a visitor can change his/her style during a visit and other elements may be of importance, e.g. crowd or fatigue of the visitor.

IV. Workflow for Analyzing the Trajectories

In the following, one objective is to map specific subsequences included in the visitor trajectories to each visiting style for characterizing more precisely the style and then making smarter recommendations. To identify the behavior of each visitor, we propose the following workflow:

1) Cluster the visitor trajectories and assign a label for each visitor (Section V).
2) Create two concept lattices using `MFCS` and `MRGS` over the whole dataset (Section VI-A).
3) From the two lattices, find jumping emerging patterns (JEPs) for each label (Section VII-A).
4) Based on their JEPs, these labels are then mapped into four visiting styles as explained in Section III-B.

V. Clustering of Trajectories

In this first experiment, we reuse the sim_{ACS} similarity measure for clustering the visitor trajectories. The idea is to check whether it is possible to distinguish the four visiting styles introduced above. We apply hierarchical clustering[1] based on sim_{ACS} to build a distance matrix between individuals. From the resulting dendrogram, we retained 5 clusters denoted by "A", "B", "C", "D", and "E". Four of them are expected to match the four visiting patterns, namely *ant*, *butterfly*, *fish*, and *grasshopper*. The last cluster will gather all non-classified trajectories. These five clusters have various sizes. Cluster "A", "B", "C", "D", and "E" have 11, 11, 59, 102, and 33 visitors respectively.

Actually, it is not easy to directly match the five clusters to corresponding visiting styles. For doing so, we will analyze the subsequences that can be attached to each cluster of trajectories. The benefit of the clustering is actually to provide a label among "A", "B", "C", "D", and "E" to the visitors. Thanks to these labels, we can search the so-called "jumping emerging patterns" and attach a characterization to the clusters based on the mined subsequences.

[1]We use the *hclust* method from the R software [16].

VI. The Mining of Trajectories Considered as Sequences

A. Mining Subsequences with `MFCS` and `MRGS`

Below, we explain the application of the `MFCS` and `MRGS` algorithms to the museum dataset and the building of an associated concept lattice. Moreover, as discussed in Section VI-B, the mining of jumping sequential patterns will help us to characterize the visitor trajectories.

In `MFCS` and `MRGS`, pattern structures are used for mining sequences. The similarity operator ($\sqcap$) between any two sets of sequences is defined as the set of closed common subsequences (SCCS) in the two input sequences. Then, given two sequences, say $S_1 = \langle \texttt{401,502,503} \rangle$ and $S_2 = \langle \texttt{401,503,502} \rangle$, the similarity between these descriptions is:

$$\begin{aligned}\delta(S_1) \sqcap \delta(S_2) &= \{\langle \texttt{401,502,503} \rangle\} \sqcap \{\langle \texttt{401,503,502} \rangle\} \\ &= \{\langle \texttt{401,502} \rangle, \langle \texttt{401,503} \rangle\}\end{aligned}$$

In the dataset, the items are grouped into categories (indicated by their first digit) and the SCCS calculation is performed, checking whether two items belong to the same category. Using the `MFCS` algorithm it becomes:

$$\begin{aligned}\delta(S_1) \sqcap \delta(S_2) &= \{\langle \texttt{401,502,503} \rangle\} \sqcap \{\langle \texttt{401,503,502} \rangle\} \\ &= \{\langle \texttt{502} \rangle, \langle \texttt{503} \rangle, \langle \texttt{401,5,5} \rangle\}\end{aligned}$$

It should be noticed that `MFCS` mines contiguous subsequences, i.e. in Definition 2, $i_k = i_{k-1} + 1$ for all $k \in \{2, 3, \ldots, m\}$. Furthermore, subsequence $\langle \texttt{401,5,5} \rangle$ can be regarded as a generalization, meaning that after item `401`, the next two visited items are something in category `5`.

In parallel, the default similarity operator of `MRGS` algorithm can be modified to accommodate our needs, such that non-contiguous common subsequences can be mined:

$$\begin{aligned}\delta(S_1) \sqcap \delta(S_2) &= \{\langle \texttt{401,502,503} \rangle\} \sqcap \{\langle \texttt{401,503,502} \rangle\} \\ &= \{\langle \texttt{401,502} \rangle, \langle \texttt{401,503} \rangle, \langle \texttt{401,5,5} \rangle\}\end{aligned}$$

Then, based either on `MFCS` or `MRGS`, a concept has an extent including a set of trajectories and an intent including a set of common subsequences. Again, it should be noticed that, based on whether a subsequence is contiguous or not, the obtained concepts are different.

For example, the concepts corresponding to Table IV are shown in Table V. Notice that both algorithms obtain a concept whose extent is V_2, V_3, V_4, albeit with different intent. Based on `MRGS`, the common subsequence of V_2, V_3, V_4 is $\langle 1, 402 \rangle$, while according to `MFCS`, their common subsequences are $\langle 1 \rangle$ and $\langle 402 \rangle$. This is because items 1 and 402 are not contiguous in V_3 and V_4.

B. Jumping Emerging Patterns

FCA is a non supervised classification process that can be turned into a supervised process thanks to the adding of a target attribute in the context, generally corresponding to a target class. Then the idea is to search for the so-called "Jumping Emerging Patterns" (JEPs) [17]. We have already

TABLE V: The concepts that are computed by of MFCS and MRGS from four visitors in Table IV.

Extent	Intent (MFCS)	Intent (MRGS)
V_1	$\langle$101,101,401,704$\rangle$	
V_2	$\langle$102,402,808,206,808$\rangle$	
V_3	$\langle$302,102,201,302,705,402,802$\rangle$	
V_4	$\langle$104,704,602,302,402,103$\rangle$	
$V_{1,2}$	$\langle$1,4$\rangle$	*not present*
$V_{1,4}$	$\langle$1$\rangle$, $\langle$4$\rangle$, $\langle$704$\rangle$	$\langle$1,1$\rangle$, $\langle$1,4$\rangle$, $\langle$1,704$\rangle$
$V_{2,3}$	$\langle$2$\rangle$, $\langle$102$\rangle$, $\langle$402,8$\rangle$	$\langle$102,402,8$\rangle$, $\langle$102,2,8$\rangle$
$V_{3,4}$	$\langle$1$\rangle$, $\langle$302$\rangle$, $\langle$402$\rangle$, $\langle$7$\rangle$	$\langle$1,302,402$\rangle$, $\langle$302,1$\rangle$, $\langle$1,7,402$\rangle$
$V_{1,3,4}$	$\langle$1$\rangle$, $\langle$4$\rangle$, $\langle$7$\rangle$	$\langle$1,4$\rangle$, $\langle$1,7$\rangle$
$V_{2,3,4}$	$\langle$1$\rangle$, $\langle$402$\rangle$	$\langle$1,402$\rangle$
$V_{1,2,3,4}$	$\langle$1$\rangle$, $\langle$4$\rangle$	$\langle$1,4$\rangle$

applied this approach in [18] for analyzing and characterizing clusters of biological inhibitors. Here we adapt the same idea for characterizing this time the clusters of visitors discovered with the similarity measure sim_{ACS}.

More precisely, five clusters are discovered by classifying visitor trajectories with sim_{ACS}. These same trajectories are then considered as sequences composed of subsequences. Then a set of characteristic subsequences is extracted and these subsequences are used as "attributes" in a formal context where objects are visitor trajectories. The resulting formal context is completed with an extra attribute corresponding to the "cluster information", i.e. the cluster in which the trajectory is classified according to sim_{ACS}. A concept lattice can then be built from this completed context.

More interestingly, the cluster information is used for characterizing the concepts whose extents include trajectories of a single cluster. The intents – made of subsequences – of these particular concepts are JEPs, and as such they can be used to characterize the corresponding clusters. For example, if the extent of the concept $(\{V_{103}, V_{165}, V_{188}\}, \{\langle 4\rangle, \langle 1\rangle, \langle 306\rangle, \langle 701,707\rangle\})$ includes visitors from cluster B only, then its intent is a JEP for that cluster.

VII. Discussion

A. Cluster Characterization

Now we are interested in characterizing the five clusters that were introduced in the previous section. For doing so, JEPs are searched in the two concept lattices obtained with MFCS and MRGS algorithms. Some of these concepts are listed in Table VI and Table VII.

First, from both MFCS and MRGS, we cannot find any satisfying concept for JEP of cluster "E". This is because among all the concepts whose extent is exclusively from cluster "E", none of them has more than one visitor. If we consider the dataset, among 33 members of cluster "E", 32 of them visit less than 2 items during their whole visit. We can assume that these visitors are not really interested in visiting the museum. Therefore, we can "safely" label this cluster as *fish*.

Cluster "D" is more easily distinguishable. Based on subsequences of concept FD2–FD4, many visitors in this class skip some items. Also, in concept RD1 and RD2, some of them visit other items after item 701. This is not a natural direction, because items in category 7 are located farther from the entrance than items in category 4 or 5. We can interpret the visitors of this cluster as *grasshopper*, since they "jump" from one item to another.

Clusters "A", "B", and "C" are relatively similar to each other. The visitors associated to these clusters follow an *ant* behavior: a natural flow (based on RA1–RC1) and no "jump" (based on FA1–FC2). However, in FC3, three visitors visit 101, then 102, then back again to 101, indicating rather a *butterfly* behavior.

B. Conclusion

In this article, we have presented our experiments in mining visitor trajectories that are modeled as sequences of items. First, we clustered the trajectories according to their common subsequences. Then, we tried to validate the clusters as behaviors (ant, butterfly, fish, and grasshopper). This is done by applying two sequence miners based on FCA to the visitor trajectories, namely MFCS and MRGS, to discover characterizing contiguous and general subsequences in each cluster.

Our result highlights some interesting patterns that may define visitor behaviors. This can help museum researchers to analyze and evaluate the placement of items and the visiting styles. Moreover, we have also studied the possibility of clustering the visitors based on a concept lattice. These clusters can be analyzed to build a recommendation system for future visitors, but we did not yet study this aspect until now.

In this paper, we only included partial information about the museum in the sequences. More interesting results are expected if other elements are taken into account, such as more general knowledge about history and geography, as well as the duration and time of the visit. Furthermore, the selection of interesting concepts can be also guided by the "stability" of concepts [19]. Finally, from a more dynamic point of view, ongoing information such as comments and state of the visitor during the visit could be also considered for analysis and online recommendation.

References

[1] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, 2nd ed. Springer-Verlag, 1999.

[2] V. Codocedo, G. Bosc, M. Kaytoue, J.-F. Boulicaut, and A. Napoli, "A proposition for sequence mining using pattern structures," in *International Conference on Formal Concept Analysis*. Springer, 2017, pp. 106–121.

[3] A. Buzmakov, E. Egho, N. Jay, S. O. Kuznetsov, A. Napoli, and C. Raïssi, "On mining complex sequential data by means of FCA and pattern structures," *International Journal of General Systems*, vol. 45, no. 2, pp. 135–159, 2016.

[4] E. Egho, N. Jay, C. Raïssi, D. Ienco, P. Poncelet, M. Teisseire, and A. Napoli, "A contribution to the discovery of multidimensional patterns in healthcare trajectories," *Journal of Intelligent Information Systems*, vol. 42, no. 2, pp. 283–305, 2014.

[5] E. Egho, C. Raïssi, T. Calders, N. Jay, and A. Napoli, "On measuring similarity for sequences of itemsets," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 732–764, May 2015.

[6] E. Véron and M. Levasseur, *Ethnographie de l'exposition*. Paris: Bibliothèque Publique d'Information, Centre Georges Pompidou, 1983.

[7] J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proceedings of the 17th International Conference on Data Engineering*, 2001, pp. 215–224.

TABLE VI: Interesting concepts discovered by the MFCS algorithm.

Concept ID	Extent	Intent	Support	Cluster
FA1	$\{V_{70}, V_{107}, V_{121}, V_{133}, V_{201}, V_{202}\}$	$\{\langle 1,1,402\rangle, \langle 103\rangle, \langle 2\rangle\}$	6	A
FA2	$\{V_{70}, V_{93}, V_{107}, V_{121}\}$	$\{\langle 402\rangle, \langle 103,104\rangle\}$	4	A
FB1	$\{V_{103}, V_{165}, V_{188}\}$	$\{\langle 4\rangle, \langle 1\rangle, \langle 306\rangle, \langle 701,707\rangle\}$	3	B
FC1	$\{V_{4}, V_{8}, V_{28}, V_{32}, V_{84}, V_{152}\}$	$\{\langle 102\rangle, \langle 101,1,101\rangle\}$	6	C
FC2	$\{V_{53}, V_{152}, V_{169}, V_{189}, V_{190}, V_{203}\}$	$\{\langle 7\rangle, \langle 102,4\rangle\}$	6	C
FC3	$\{V_{4}, V_{8}, V_{32}\}$	$\{\langle 101,102,101\rangle\}$	3	C
FD1	$\{V_{54}, V_{105}, V_{139}, V_{168}\}$	$\{\langle 202,4\rangle\}$	4	D
FD2	$\{V_{139}, V_{168}\}$	$\{\langle 202,405,701\rangle\}$	2	D
FD3	$\{V_{46}, V_{47}\}$	$\{\langle 101,602\rangle\}$	2	D
FD4	$\{V_{89}, V_{163}\}$	$\{\langle 602,203\rangle\}$	2	D

TABLE VII: Interesting concepts discovered by the MRGS algorithm.

Concept ID	Extent	Intent	Support	Cluster
RA1	$\{V_{70}, V_{107}, V_{121}, V_{133}, V_{201}, V_{202}\}$	$\{\langle 1,1,402,2\rangle, \langle 1,1,4\rangle, \langle 103,402,2\rangle, \langle 103,4\rangle\}$	6	A
RB1	$\{V_{142}, V_{183}, V_{192}\}$	$\{\langle 102,1,1,1,1\rangle, \langle 102,103,1,1\rangle, \langle 1,1,1,1,1\rangle, \langle 1,103,1,1\rangle\}$	3	B
RC1	$\{V_{4}, V_{8}, V_{28}, V_{84}, V_{152}\}$	$\{\langle 1,1,1,101\rangle, \langle 1,101,1,101\rangle, \langle 1,1,1,1\rangle, \langle 1,101,1,1\rangle,$ $\langle 101,1,1,1\rangle, \langle 101,101,1,1\rangle, \langle 101,101,101\rangle, \langle 102,101\rangle, \langle 102,1\rangle\}$	5	C
RD1	$\{V_{71}, V_{79}\}$	$\{\langle 701,504\rangle\}$	2	D
RD2	$\{V_{97}, V_{98}\}$	$\{\langle 701,406\rangle\}$	2	D

[8] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2002, pp. 429–435.

[9] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining: Closed sequential patterns in large datasets," in *Proceedings of the 2003 SIAM International Conference on Data Mining*. SIAM, 2003, pp. 166–177.

[10] J. Wang and J. Han, "BIDE: Efficient mining of frequent closed sequences," in *Proceedings of 20th International Conference on Data Engineering*. IEEE, 2004, pp. 79–90.

[11] A. Gomariz, M. Campos, R. Marin, and B. Goethals, "ClaSP: an efficient algorithm for mining frequent closed sequences," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 50–61.

[12] B. Ganter and S. O. Kuznetsov, "Pattern structures and their projections," in *International Conference on Conceptual Structures*. Springer, 2001, pp. 129–142.

[13] J. Lanir, T. Kuflik, E. Dim, A. J. Wecker, and O. Stock, "The influence of a location-aware mobile guide on museum visitors' behavior," *Interacting with Computers*, vol. 25, no. 6, pp. 443–460, 2013.

[14] M. Zancanaro, T. Kuflik, Z. Boger, D. Goren-Bar, and D. Goldwasser, "Analyzing museum visitors' behavior patterns," in *International Conference on User Modeling*. Springer, 2007, pp. 238–246.

[15] T. Kuflik, Z. Boger, and M. Zancanaro, "Analysis and prediction of museum visitors' behavioral pattern types," in *Ubiquitous Display Environments*. Springer, 2012, pp. 161–176.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2014.

[17] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 43–52.

[18] Y. Asses, A. Buzmakov, T. Bourquard, S. O. Kuznetsov, and A. Napoli, "A Hybrid Classification Approach based on FCA and Emerging Patterns - An application for the classification of biological inhibitors," in *Proceedings of CLA*, ser. CEUR Workshop Proceedings, vol. 972, 2012, pp. 211–222.

[19] S. O. Kuznetsov and D. I. Ignatov, "Concept stability for constructing taxonomies of web-site users," *arXiv preprint arXiv:0905.1424*, 2009.