

# Optimisation using Natural Language Processing: Personalized Tour Recommendation for Museums

Mayeul Mathias \*, Assema Moussa \*, Fen Zhou \*<sup>†</sup>, Juan-Manuel Torres-Moreno \*<sup>†§</sup>,  
 Marie-Sylvie Poli <sup>†‡</sup>, Didier Josselin \*<sup>†¶</sup>, Marc El-Bèze \*<sup>‡</sup>, Andréa Carneiro Linhares <sup>||</sup>, Françoise Rigat \*\*  
 {mayeul.mathias, assema.moussa}@alumni.univ-avignon.fr  
 {fen.zhou, juan-manuel.torres, marie-sylvie.poli, didier.josselin, marc.elbeze}@univ-avignon.fr  
 andrea.linhares@ufc.br, francoise\_rigat@yahoo.it

\*LIA, Université d'Avignon et des Pays de Vaucluse, France.

<sup>†</sup>CNE, Université d'Avignon et des Pays de Vaucluse, France.

<sup>‡</sup>FR 3621 Agorantic - CNRS Université d'Avignon et des Pays de Vaucluse, France.

<sup>§</sup>École polytechnique de Montréal, (Québec) Canada

<sup>¶</sup>UMR 7300 ESPACE - CNRS, France

<sup>||</sup>Universidade Federal do Ceará, Brazil.

\*\*Università degli Studi di Torino, Italy.

**Abstract**—This paper proposes a new method to provide personalized tour recommendation for museum visits. It combines an optimization of preference criteria of visitors with an automatic extraction of artwork importance from museum information based on Natural Language Processing using textual energy. This project includes researchers from computer and social sciences. Some results are obtained with numerical experiments. They show that our model clearly improves the satisfaction of the visitor who follows the proposed tour. This work foreshadows some interesting outcomes and applications about on-demand personalized visit of museums in a very near future.

## I. INTRODUCTION

MUSEUMS are no longer only institutions that acquire, store and expose our heritage. Going to a museum is a learning activity but also an enjoyment for visitors. With the emergence of the Web, curators and cultural mediators decided to get involved in collaborative and numerical culture to attract a larger public. Today, almost all museums have a website but few of them allow the visitors to prepare their visit in the best conditions.

Some art, science and society museums are collaborating with research laboratories to develop new technologies that improve services in museums in response to the desires of existing and potential visitors.

However, there are still difficulties, epistemological barriers, to study the expectations and the intentions of different publics, including online visitors. Knowing why people want to come and visit museums could allow automatic systems to suggest their tour, save their time and give them the best of the knowledge of the exhibited arts.

Among all possibilities, a recommendation system for personalized routing is by far one of the best improvements. Indeed, some museums exhibit thousands of artworks and it is not conceivable for a visitor to admire all of them because he might spend time in front of artworks which do not match his interests and he might not be able to see other more interesting artworks due to tiredness or a lack of time. A few

museums, as The Louvre, offer a recommendation system<sup>1</sup> but they are limited to the selection of a route in a pre-established set. Moreover, in this particular case, the personalization is restricted to the selection of a theme and the duration of the visit in a set of no more than 10 themes and 4 different durations.

It is essential to propose a personalized route for each visitor or group of visitors according to their interests while taking into account their constraints such as limited schedule, physical handicap or a list of artworks to include of the tour. This operation may also reduce unuseful moves (avoid round trips). But to calculate an optimal tour, we need to assess the visitor interest for each artwork by asking his preferences.

Modeling the preferences with random distributions may not reflect reality because curators take care of the scenography (therefore the coherence) of each room. So we worked on preferred artists (the visitor can select a set of interesting artists) and we propose to use the artworks description to highlight a kind of intrinsic interest from the point of view of the museum. Indeed, the description displayed to the visitor should show how significant is the artwork for the museum. We valueate each item by analyzing their description (with Automatic Text Summarization) and use it as a base value, considering that even without any preference, some artworks are more interesting than others.

### *The Musée de l'Orangerie*

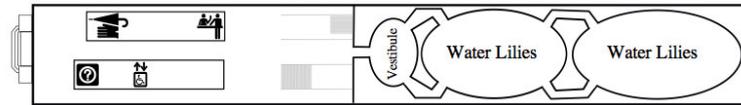
Due to the time needed to extract and check all the data we worked on this small museum to test our model.

The Musée de l'Orangerie (Museum of Orangerie), in Paris (France), regroups 144 artworks from 14 artists in 10 exhibition rooms. The website<sup>2</sup> of the museum supplies a map (shown in Fig. 1) and indexes information about all the

<sup>1</sup><http://www.louvre.fr/en/parcours>

<sup>2</sup><http://musee-orangerie.fr>

## UPPER FLOOR: Water Lilies



## LOWER FLOOR: Jean Walter - Paul Guillaume Collection

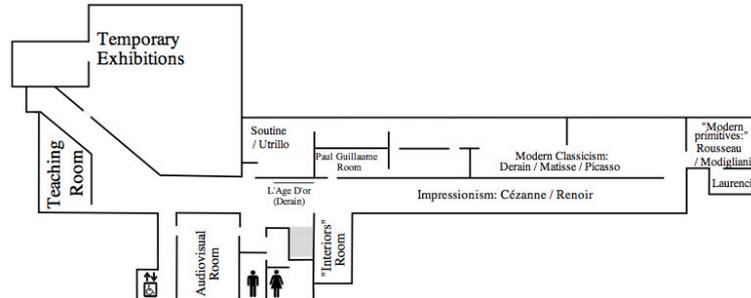


Fig. 1. Map of the Musée de l'Orangerie

artworks including the name of the artist, a description of the artwork and its date of creation.

#### Paper organization

The remaining of this paper is organized as follows. We review the related work in Section II. In Section III, we present a Natural Language Processing (NLP) based approach to compute artwork interest. The Personalized Tour Recommendation Problem is presented in Section IV. To solve this problem, we develop an Integer Linear Programming based method to solve the tour optimization problem in Section V and define a model to represent the visitor preferences in Section VI. The simulations are conducted and numerical results are demonstrated in Section VII. Finally, we conclude the paper in Section VIII.

## II. RELATED WORK

A first model developed in 2010 [1] proposes to formulate the visitor routing problem as an extension of the open shop scheduling problem (in which each visitor group is a job and each interesting room is a machine). Each visitor group has to pass through all rooms but it is impossible for two groups of visitors to be simultaneously in the same room. This restriction can lead to non optimal or infeasible solutions if there are more visitor groups than rooms in the museum (which is the case if we consider each single visitor as a group).

Relying on the constraint programming model [2], we propose to reduce the number of used variables. In [2], they generate a route by calculating the smallest number  $K$  of steps required to cross the museum (to visit all the rooms). This model requires that each artwork is represented as  $K$  variables (one per step). Due to the fact that museums often have several thousands of artworks, it leads to a huge number of variables. Moreover they use mathematical distributions to simulate a visitor profile which does not necessary reflect

reality (in museums, artworks are often grouped in a room because they are related to each other, a configuration that a random distribution as they used cannot represent).

In 2013, some works [3] used the visiting style of visitors (the way a visitor go from an artwork to another) but their model requires two matrices of size  $N^2$  (where  $N$  is the number of artworks). The first one indicates the accessibility to an artwork from another (if they are in the same room or in two rooms directly connected) and the second one contains the distance between two artworks. However as the number of artworks is always greater than the number of rooms, most of the museums are modeled as two sparse matrices with duplicated data (in a room, it is often allowed to freely move between artworks). This makes the use of constraint programming expensive.

## III. ARTWORK DESCRIPTION ANALYSIS USING TEXTUAL ENERGY

Our idea is to use the description of each artwork as an independent measure of their interest. Indeed, two similar artworks (same theme, support, artist) will produce the same result but may be very different. By analyzing the description provided by the museum, we tried to differentiate them.

Automatic Text Summarization (ATS) techniques by extraction [4], [5], [6] allow to rank a set of textual segments (sentences, paragraphs etc.) depending on a measure of similarity. Textual Energy algorithm (Enertex) converts a textual document into a physical object and use Statistical Physics to measure its energy [7]. This energy, to which we should refer as Textual Energy, is then computed and apply to summarization. The physical model of Textual Energy gives rise to a single non iterative algorithm of low complexity. Therefore Textual Energy allows to redefine sentence ranking on simple and efficient matrix operations. The resulting algorithms are

much easier to apply to large texts and give better results without using any post-processing.

#### A. Starting point: Hopfield Model

Hopfield's approach [8], [9] was based on magnetic Ising model to build a Neural Network (ANN) with pattern learning capabilities. The capacities and limitations of this ANN (an associative memory), were well established in a theoretical framework [8], [9]: the patterns must not be correlated to obtain free error recovery, the system saturates quickly and only a little fraction of the patterns can be stored correctly. Despite these major drawbacks, Hopfield contributed to ANN theory by introducing the concept of energy by analogy with magnetic systems. A magnetic system is a set of  $N$  spins like small magnets that can adopt several orientations. The simplest model is the dipole one or Ising model where there are only two opposite possible orientations: up ( $\uparrow$  or  $+1$ ) or down ( $\downarrow$  or  $0$ ). Ising magnetic model was used in a large variety of systems that can be completely described by a set  $N$  of binary variables [10] with  $2^N$  possible configurations (patterns). The spins are equivalent to neurons that can interact following Hebb's rule<sup>3</sup>:

$$J_{i,j} = \sum_{\mu=1}^P (s_{\mu,i} \times s_{\mu,j}) \quad (1)$$

$s_{\mu,i}$  and  $s_{\mu,j}$  are the states of neurons  $i$  and  $j$  in the pattern  $\mu$ . The summation concerns the  $P$  patterns to store. This rule of interaction is local, because  $J_{i,j}$  depends only on the states of the connected unities. It has the capacity to store and to recover certain number of configurations of the system, because the Hebb rule transforms these configurations into attractors (minimal local) of the energy function [8]:

$$E_{\mu} = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (s_i \times J_{i,j} \times s_j) \quad (2)$$

The fundamental concept of magnetic energy is a function of the system configuration, that is, of the state of activation or non-activation of its units. The concept of energy induces a type of interaction. If we present a pattern  $\nu$ , every spin will undergo a local field:  $h^i = \sum_{j=1}^N J^{i,j} s^j$  induced by the energy of the others spins. Therefore the total energy of the new system made of the new pattern inserted into the previous system reflects the interaction between the pattern and the initial system.

We shall focus on theoretical objects that are usually considered in Statistical Physics. In magnetic system analysis, these are energy function distributions [11]. Hopfield himself used these functions to show that the recovery is convergent. Our Enextex system is entirely grounded on them.

#### B. Energy as a document similarity measure

The Vector Space Model (VSM) has also been applied to texts since [12] following a bag of word representation of sentences. In this model vectors represent sentences and

a document gives rise to a matrix. We have used VSM to represent documents in our model magnetic system: a sentence (a row vector) is equivalent to a Ising spin chain and a document (a magnetic system) is represented by a matrix of  $P$  rows  $\times$   $N$  columns. Therefore, sentences can be studied as Ising spin chains. More formally, with a vocabulary of  $N$  words (terms) in a document, it is possible to represent a sentence as a chain of  $N$  spins,  $i = 1, \dots, N$ . A document with  $P$  sentences is formed of  $P$  chains in the vector space  $\Xi$  of dimension  $N$ . In this paper, the description of each artwork is assimilated as a long pseudo-sentence. Therefore, a document (the collection of a museum) is constituted of a set of  $P$  (pseudo-)sentences.

Documents are preprocessed by removing functional words (by using a stop list), normalized and lemmatized [13], [14]. This preprocessing reduces considerably the document dimensionality. Let be  $T = \{t_1, \dots, t_N\}$  the set of remaining terms after this preprocessing. Once segmented into units, usually sentences, the text is represented by a set  $S = \{\vec{s}_1, \dots, \vec{s}_{\mu}, \dots, \vec{s}_P\}$  where each  $\vec{s}_{\mu}$  is the bag of words in segment  $\mu$ . As usual in text vector model, we consider the matrix  $S_{[P \times N]} = (s_{\mu,j})_{1 \leq \mu \leq P, 1 \leq j \leq N}$  of frequency/absence associated to  $H$  by:

$$s_{\mu,j} = \begin{cases} tf_{\mu,j} & \text{if } t_j \in \text{sentence } \mu \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where  $tf_{\mu,j}$  is the term frequency of  $t_j$  in the sentence  $\mu$ .

We therefore consider the presence of term  $t_j$  as a spin  $s_j$   $\uparrow$  with magnitude  $tf_{\mu,j}$  (its absence by a  $\downarrow$  respectively), and a description of each artwork (text segment) by a chain of  $N$  spins.

It is common to consider that these vectors are correlated according to the shared words. Here the introduction of the magnetic model induces moreover indirect interactions. In this model sentences that do not share any word could however interact because of the magnetic field generated by the other sentences of the document that form the global magnetic system.

We have studied the interactions between the terms and the sentences using Hebb's rule and Ising energy respectively. To obtain the matrix  $J$  of interactions between the  $N$  terms, we apply Hebb's rule (equation 1) in its matrix form:

$$J = S^T \times S \quad (4)$$

where  $J_{i,j}$  is the number of co-occurrences of terms in sentences. The energy function (equation 2) of a (magnetic) system  $S$  is:

$$E = S \times J \times S^T = (S \times S^T)^2 \quad (5)$$

Each element  $E_{\mu,\nu}$  represents the energy between sentences  $\mu$  and  $\nu$ . The values in the first matrix diagonal quantify the interaction energy between words into a sentence meanwhile the other values in the rest of the matrix show the interactions between distinct sentences. The sum of absolute values in one

<sup>3</sup>Hebb [9] suggested that synaptic connections change according to the correlation between neuronal states.

row gives the total energy of interaction of the corresponding sentence  $\mu$  with the document:

$$E_{\mu} = \sum_{\nu} |e_{\mu,\nu}| \quad (6)$$

We use this energy value to rank sentences (description of artwork) by order of decreasing importance. The most energetic will be considered as the most important.

#### IV. PERSONALIZED TOUR RECOMMENDATION PROBLEM

The Personalized Tour Recommendation Problem (PTRP) can be viewed as an optimization problem and solved by optimization techniques. For this purpose, we first model the museum topology as a graph and then formulate the studied problem as an Integer Linear Programming (ILP) instance. Therefore, the optimal personalized tour can be obtained by solving the ILP model we propose.

##### A. Museum modeling

A museum is modeled as a 7-tuple  $G = \langle V, A, E, X, P, r \rangle$  where:

- $V$  is the set of vertices, each vertex is an exhibition room, an entrance or an exit of the museum.
- $A$  is the set of arcs which connect different rooms. There is an arc  $a_{ij} \in A$  between two vertices  $i$  and  $j$ , if we can go from room  $i$  to room  $j$  directly without passing through other rooms.
- $E$  is the set of entrances of a museum, which is a subset of  $V$ , i.e.  $E \subset V$ .
- $X$  is the set of exits of a museum, which is also a subset of  $V$ , i.e.  $X \subset V$ .
- $P$  is the set of all artworks in the museum.
- $r$  is a mapping function  $P \rightarrow V$ . For each artwork  $p \in P$ ,  $r(p)$  is the room containing  $p$ .

Some large museums may have several entrances and exits, that is why  $E$  and  $X$  are two subsets of  $V$ . We also admit that there is always a path from any entrance to an exit. We consider directed arcs and not edges because some museums may impose a flow direction for several reasons (minimizing congestion, pedagogical tour). Note that by definition of  $A$ , there is no incoming arc to any entrance and there is no any outgoing arc from any exit neither.

*Application to the Musée de l'Orangerie:* The Musée de l'Orangerie can be represented as the graph presented in Figure 2. We can see that there is only one entrance and one exit in the museum and they are located at the same place. Therefore, we consider the entrance and the exit as two different vertices in the graph to facilitate the model. The mapping between vertices and rooms is shown in Table I.

##### B. Personalized tour problem modeling

For the sake of satisfying the visitor maximally, a visit tour should be proposed according to the visitor's preferences and constraints.

A personalized tour problem can be defined as a 6-tuple  $\langle G, I, R, u, t, T_{MAX} \rangle$  where:

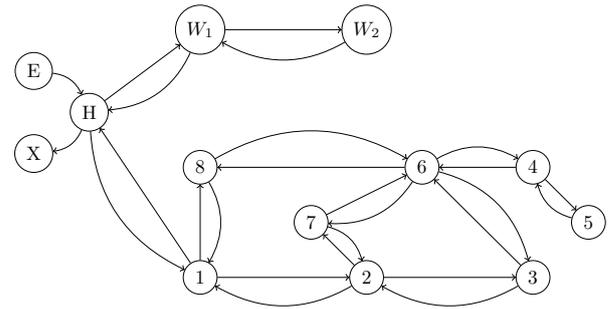


Fig. 2. A possible graph for the Musée de l'Orangerie

TABLE I  
THE MUSÉE DE L'ORANGERIE: VERTICES AND ROOMS

Vertex	Room
E	Entrance
X	Exit
H	Hall
$W_1$	Water Lilies (first part)
$W_2$	Water Lilies (second part)
1	L'Age d'Or
2	Paul Guillaume Room
3	Impressionism
4	Modern primitives
5	Laurencin room
6	Modern classicism
7	Derain room
8	Soutine / Utrillo room

- $G$  is the museum graph representing the museum topology as defined above.
- $I$  is the set of artworks which have to be included in the tour.
- $R$  is the set of artworks which have to be excluded of the tour.
- $u$  is a mapping  $u : P \rightarrow \mathbb{R}^+$ . For each artwork  $p \in P$ ,  $u(p)$  denotes the interest of the visitor for the artwork  $p$ .
- $t$  is a mapping  $t : A \cup V \cup P \rightarrow \mathbb{R}^+$ . For each room, arc and artwork, we have a time to spend. It can be the time needed to cross a room or an arc. It can also be the time to see an artwork.
- $T_{MAX}$  is the maximum time that a visitor wants to spend in the museum.

A visit tour may be a simple path without any cycles (an elementary path) or a sophisticate path including cycles (a non-elementary path).

We define a tour as a sequence of pairs  $\langle v, P_v \rangle$  where  $v \in V$  and  $P_v \subseteq r(v)$  (note that  $P_v$  may be  $\emptyset$  because we can cross a room without seeing any artwork). A tour  $T = (\langle v_1, P_{v_1} \rangle, \dots, \langle v_n, P_{v_n} \rangle)$  is a solution to the personalized tour recommendation problem when:

- 1) The vertex  $v$  in the first element of  $T$  is an element of  $E$  (the tour starts by an entrance).
- 2) The vertex  $v$  in the last element of  $T$  is an element of  $X$  (the tour ends by an exit).
- 3) All consecutive elements  $T_1$  and  $T_2$  of  $T$  share the same vertex or an arc  $a_{ij} \in A$  must exist from the vertex  $i$  of

$T_1$  to the vertex  $j$  of  $T_2$ .

- 4) The total time required to see all the artworks and pass through all the rooms (and ways) is not bigger than  $T_{MAX}$ .

*Application to the Musée de l'Orangerie:* we may have a visit tour like the following:

$$\begin{aligned} & (\langle E, \emptyset \rangle, \langle H, \emptyset \rangle, \langle W_1, p_1 \rangle, \langle W_2, p_5 \rangle, \langle W_1, \emptyset \rangle, \langle H, \emptyset \rangle, \\ & \langle 1, p_{10} \rangle, \langle 8, p_{103} \rangle, \langle 1, \emptyset \rangle, \langle H, \emptyset \rangle, \langle X, \emptyset \rangle) \end{aligned}$$

In this tour, the visitor should cross the receiving hall  $H$  three times, exhibition room  $W_1$  and 1 twice respectively. Although we may traverse a room several times, the visitor is supposed to visit the room only once. Consider for instance the exhibition room  $W_1$ , we may visit the selected artworks when we reach this room for the first time. The second time, we would just cross the room to visit another one.

## V. INTEGER LINEAR PROGRAMMING APPROACH TO SOLVE THE OPTIMAL PERSONALIZED TOUR RECOMMENDATION PROBLEM

Before introducing an ILP model to solve the Personalized Tour Recommendation Problem, we define several decision variables:

- $x_p$  equals 1 if the artwork  $p \in P$  is included in the proposed tour, 0 otherwise.
- $y_a$  equals 1 if arc  $a \in A$  is crossed in the proposed tour, 0 otherwise.
- $f_a$  denotes the number of rooms crossed when we arrive at arc  $a \in A$  in the visit walk.
- $z_v$  equals 1 if room  $v \in V$  is traversed in the proposed tour (no matter whether we visit an artwork of this room or not), 0 otherwise.

Given a personalized tour problem (as defined in section IV), the objective function of the Optimal Personalized Tour Recommendation Problem (OPTRP) is to maximize the overall satisfaction of the proposed visit tour for the visitor:

$$\max \sum_{p \in P} x_p \times u(p) \quad \text{OPTRP} \quad (7)$$

s.t.

$$\sum_{v \in E} \sum_{a \in \delta^+(v)} y_a = 1 \quad (8)$$

$$\sum_{v \in X} \sum_{a \in \delta^-(v)} y_a = 1 \quad (9)$$

$$\sum_{a \in \delta^+(v)} y_a = \sum_{a \in \delta^-(v)} y_a, \quad \forall v \in V \setminus (E \cup X) \quad (10)$$

$$f_a \geq y_a, \quad \forall a \in A \quad (11)$$

$$f_a \leq |V| \times y_a, \quad \forall a \in A \quad (12)$$

$$\sum_{a \in \delta^-(v)} f_a = \sum_{a \in \delta^+(v)} f_a - z_v, \quad \forall v \in V \setminus (E \cup X) \quad (13)$$

$$\sum_{a \in \delta^+(v)} y_a + \sum_{a \in \delta^-(v)} y_a \geq z_v, \quad \forall v \in V \quad (14)$$

$$y_a \leq z_v, \quad \forall v \in V, \forall a \in \delta^+(v) \cup \delta^-(v) \quad (15)$$

$$x_p \leq z_v, \quad \forall v \in V, \forall p \in \{p : r(p) = v\} \quad (16)$$

$$x_p = 1, \quad \forall p \in I \quad (17)$$

$$x_p = 0, \quad \forall p \in R \quad (18)$$

$$\sum_{v \in V} z_v \times t_v + \sum_{a \in A} y_a \times t_a + \sum_{p \in P} x_p \times t_p \leq T_{MAX} \quad (19)$$

Constraints (8) and (9) ensure that the visitor should enter a museum from a unique entrance and finish the visit by a unique exit respectively (this model considers the case of multiple entrances and exits). Constraint (10) makes sure that a visitor should exit a room  $v$  after crossing or visiting it. Constraint (11) expresses that a visitor should have crossed at least a room before arriving at an arc  $a$ , while constraint (12) imposes that no flow is moving on the arc  $a$ , if it is not crossed in the visit tour. Constraint (13) means that if a room  $v$  is crossed in the tour, then the number of rooms crossed before arriving at this room equals the number of rooms crossed after leaving  $v$  minus one. Otherwise, they should be equal, since the room will not appear in the tour. Constraint (15) imposes that a room  $v$  should be crossed as long as one input arc or one outgoing arc is crossed. Constraint (14) ensures that a room  $v$  should not be crossed if none of the input arc or output link is used during the visit. Constraint (16) indicates that if a room  $v$  is not crossed, none of its artworks will be proposed for visiting. Constraints (17) and (18) ensure that an artwork should be included or excluded from the proposed tour if the visitor asks for it. The last constraint (19) guarantees that the time spent in front of the artworks and the time required to pass through rooms (and ways) does not exceed the available time for the visitor.

The ILP model we propose provides a visit tour starting from an entrance and terminating at an exit. In [2], authors also proposed an ILP model to plan the personalized visit. They divided the studied proposed into two sub-problems: first determine the number of moves (denoted as  $K$ ) for a complete walk in the museum graph, and then solve the museum routing problem while maximizing visitor satisfaction. Since both of these sub-problems are NP-Hard, authors of [2] proposed to solve both of them by constraint programming. The complexity of their model depends a lot on  $K$ , which is generally large (at least equals to  $|V|$ ). To compare the complexity of our model with the ILP mode in [2], the number of variables and constraints are listed in Table II.

## VI. VISITOR PREFERENCES MODELING

The interest function  $u$  should reflect the satisfaction of the visitor for each artwork  $p \in P$ . The nearer to his preferences is an artwork  $p$ , the greater is  $u(p)$ .

### *Representation of artworks and the visitor preferences*

We define  $C_p$  as the set of all characteristics of an artwork  $p$  and  $C$  as the union of all these sets (i.e.  $C =$

TABLE II  
COMPARISON OF ILP MODELS

Terms	OPTRP ILP	ILP [2]
Variables	$x_p, y_a, f_a, z_v$	$x_p, x_{p,k}, c_{i,j,k}$
Number of variables	$ P  + 2 A  +  V $	$( P  +  A ) \times K +  P $
Constraints	(8)-(19)	(2)-(8) in [2]
Number of constraints	$3 + 4 A  +  P  + 3 V  +  I  - 2( E  +  X )$	$ A  \times (K - 1) +  P  \times (2K - 1) + K + 3$

$\{c|c \in C_p \forall p \in P\}$ ). A characteristic may be the theme, the type of support, the date of creation, the name of the artist or anything that identify an artwork.

We can represent any artwork  $p \in P$  as a characteristics vector  $v_p = (v_{p_1}, \dots, v_{p_n})$  in a vector-space of  $|C|$  dimensions. Each element  $v_{p_i} \in \mathbb{R}^+$  in the vector is a numerical value measuring the relevance of the artwork  $p$  to the associated characteristic. Additionally we define a vector  $v$  in the same vector-space as the vector representing the visitor preferences (where each element of  $v$  measures the interest of the visitor for the associated characteristic).

#### Measuring the interest for an artwork

To identify the interest for the visitor to an artwork, we compare  $v$  and  $v_p$  with the cosine similarity which calculate the angle between two vectors. The formula is the following:

$$\text{similarity}(v_p, v) = \frac{\sum_i^n v_{p_i} \times v_i}{\sqrt{\sum_i^n v_{p_i}^2} \times \sqrt{\sum_i^n v_i^2}} \quad (20)$$

The resulting similarity ranges from 0, meaning that the visitor is not interested at all by the artwork, to 1 meaning that the artwork exactly matches his preferences.

In our model, we used  $u(p) = \text{similarity}(v_p, v)$  where  $v_p$  and  $v$  are the vectors of the artwork  $p$  and the visitor preferences respectively.

## VII. SIMULATIONS AND NUMERICAL RESULTS

We implemented the ILP model described in section V by using the IBM CPLEX 12 library<sup>4</sup>.

The program takes several input parameters:

- The graph modeling the museum as defined in section IV
- The interest function  $f$  to use. This function produces interest vectors as defined in section VI
- The maximum duration that a visitor can spend in the museum

It outputs the proposed tour (as defined in section IV).

### A. Intrinsic interest

As we saw in section III, the Enertex algorithm ranks the sentences of a document. We used Enertex as the following:

- 1) From the website of the museum, we created an XML file containing the following information for each artwork: the title, the artist name, the year and the description of the artwork.

<sup>4</sup><http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/>

- 2) We extracted data from the XML to produce a file where each pseudo sentence is a concatenation of title, artist and description.
- 3) The latter file is used as an input to Enertex with the query "musée orangerie peinture impressionniste postimpressionniste" to drive the balancing process of the system.

It produces a ranking for artworks depending on the information displayed by the museum (for each artwork, the result is a value ranging from 0 to 1).

Fig. 3 shows the ranking of the artworks in the *Musée de l'Orangerie* provided by Enertex.

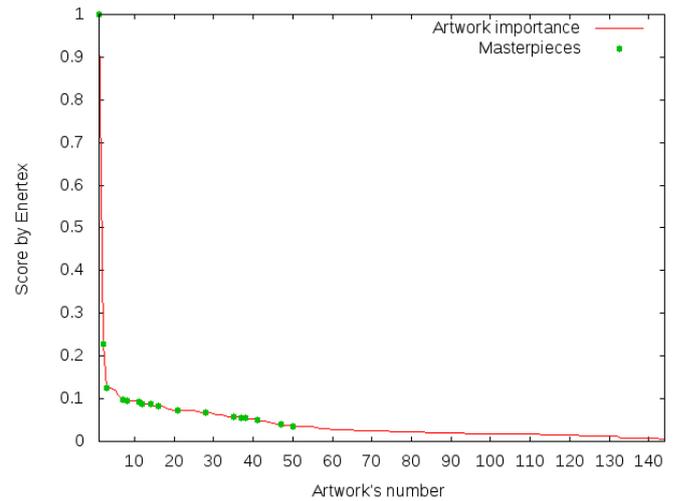


Fig. 3. Ranking by Enertex of the artworks in the Musée de l'Orangerie

As we can see, the resulting ranking is in agreement with the information provided by the museum. Indeed the masterpieces (according to the website of the museum) represent the most important artworks (which have the highest scores).

### B. Interest functions

Four different interest function were designed to simulate the visitor preferences.

- $f_1$ : produces the same vector  $v = (1)$  for each artwork and visitor preferences
- $f_2$ : produces a vector  $v_p = (s_p)$  where  $s_p$  is the score given by Enertex for the artwork  $p$  and produces a vector  $v = (1)$  as the visitor preferences.
- $f_3$ : produces vectors  $v = (v_1, \dots, v_n)$  of size  $n$  equals to the number of artists. Each artwork is represented as a

vector where  $v_i = 1$  if the artwork is created by the artist  $i$ , 0 otherwise. The visitor preferences are represented as a vector where  $v_i = 1$  if the visitor is interested by the artist  $i$ , 0 otherwise.

- $f_4$ : produces vectors  $v = v_{f_2} \parallel 10 \times v_{f_3}$  where  $v_{f_2}$  and  $v_{f_3}$  are the vectors produced by  $f_2$  and  $f_3$  respectively.

The first function defines the baseline: the visitor has no interest at all. The second uses the ranking provided by Enertex: the visitor wants to discover the most important artworks of the museum. The third uses the visitor preferences (a set of interesting artists). The last combines both visitor preferences and museum point of view (we multiply by 10 because we want to give more importance to the visitor preferences than to the museum point of view).

### C. Evaluation

To evaluate the output tour, we measure the relevance percentage defined as :

$$rp = \frac{100 * \sum_{p \in T} f(p)}{\sum_{p \in P} f(p)} \quad (21)$$

where  $T$  is the set of artworks proposed in the tour (i.e.  $T \subseteq P$ ) and  $f$  the interest function used (as saw above). The relevance percentage  $rp$  denotes a satisfaction rate of the visitor.

### D. Results

For each function  $f$ , we ran the program with different time limits from 30 to 330 minutes (the time required to visit the entire collection) by steps of 15 minutes. For  $f_2$  and  $f_3$ , we randomly pre-generated 5,000 combinations of 2, 3, 4 and 5 artists (i.e. the same combinations are used with  $f_2$  and  $f_3$ ) and calculated the arithmetic mean relevance percentage for each duration.

Figure 4 shows the evolution of  $rp$  for each function  $f$ . As expected, the first interest function  $f_1$  produces a linear evolution of the relevance percentage (given that all artworks have the same interest, the tour includes the greatest number of artworks). With Enertex we are able to propose efficient tours to visitors who want to discover the museum (without particular preferences). The combination of both visitor preferences and intrinsic preferences produces the best results up to 49% of relevance improvement. It also appears that after 150 minutes, the improvement is less significant, we could assume that, from the visitor point of view, the optimal tour duration is about 2 hours and a half.

## VIII. CONCLUSION AND FUTURE WORKS

This research tackles the problem of optimizing museum visits according to visitors preference and artwork importance. As a first milestone for next works taking into account the individual behavior in museum visits, it sets an original model combining computational optimisation and automatic learning via artificial intelligence. We first drew the optimization framework based on graph theory to depict the spatial organization of the museum (including rooms and paths), that requires an

Integer Linear Programming to maximize the visitor overall satisfaction and to generate an optimal path, that is to say a series of rooms and artworks to be seen by the visitor. In complement, we compute an artwork description analysis by a natural language processing based on textual energy (using an algorithm called Enertex). This leads to ranking the different artworks according to the descriptions given by the museum, related to their artistic importance. Associating those two complementary approaches, we are then able to design optimal paths for visitors according to different interest functions based on artwork objective values assigned by museums.

Future works concern more subjective behavior of visitors depending on their profiles and leisure practices. Indeed, the project aims at finding relevant recommendations for optimal visit tours that rise a better fitness between the visitor wishes and the museum artistic supply. We can think about using natural language processing to generate the set of characteristics for all the artworks in a museum and calculate better interest vectors but also produce a summary of the proposed tour.

This information may advantageously be used by existing and potential visitors to refine the way they get involved in their cultural practices. Indeed, it is admitted that the museum connoisseurs use to develop a critical mind about new services in a numerical society. Thence, aware visitors become able to appreciate the personalized routing recommendation system provided by their preferred museums.

### ACKNOWLEDGMENTS

The authors would like to thank the *Departement du Vaucluse* (France) and the *FR Agorantic* for the financial supports (projects @MUSE and InfoMuse).

### REFERENCES

- [1] V. F. Yu, S.-W. Lin, and S.-Y. Chou, "The museum visitor routing problem," *Applied Mathematics and Computation*, vol. 216, no. 3, pp. 719–729, 2010.
- [2] D. L. Berre, P. Marquis, and S. Roussel, "Planning personalised museum visits," in *ICAPS*, D. Borrajo, S. Kambhampati, A. Oddi, and S. Fratini, Eds. AAAI, 2013.
- [3] I. Lykourantzou, X. Claude, Y. Naudet, E. Tobias, A. Antoniou, G. Lepouras, and C. Vassilakis, "Improving museum visitors' quality of experience through intelligent recommendations: A visiting style-based approach," in *Intelligent Environments (Workshops)*, ser. Ambient Intelligence and Smart Environments, J. A. Botía and D. Charitos, Eds., vol. 17. IOS Press, 2013, pp. 507–518.
- [4] H. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.
- [5] T. Sakai and K. Spärck-Jones, "Generic summaries for indexing in Information Retrieval," in *ACM Special Interest Group on Information Retrieval (SIGIR'01): 24th International Conference on Research and Development in Information Retrieval*. New Orleans, LA, USA: ACM, 2001, pp. 190–198.
- [6] J.-M. Torres-Moreno, *Résumé automatique de documents : une approche statistique*. Hermes-Lavoisier (Paris), 2011.
- [7] S. Fernández, E. SanJuan, and J. M. Torres-Moreno, "Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation," in *LNAI 4287, MICAI'07, Mexico*, 2007, pp. 861–871.
- [8] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the USA*, vol. 9, pp. 2554–2558, 1982.
- [9] J. Hertz, A. Krogh, and G. Palmer, *Introduction to the theory of Neural Computation*. Redwood City, CA: Addison Wesley, 1991.
- [10] S. Ma, *Statistical Mechanics*. Philadelphia, CA: World Scientific, 1985.

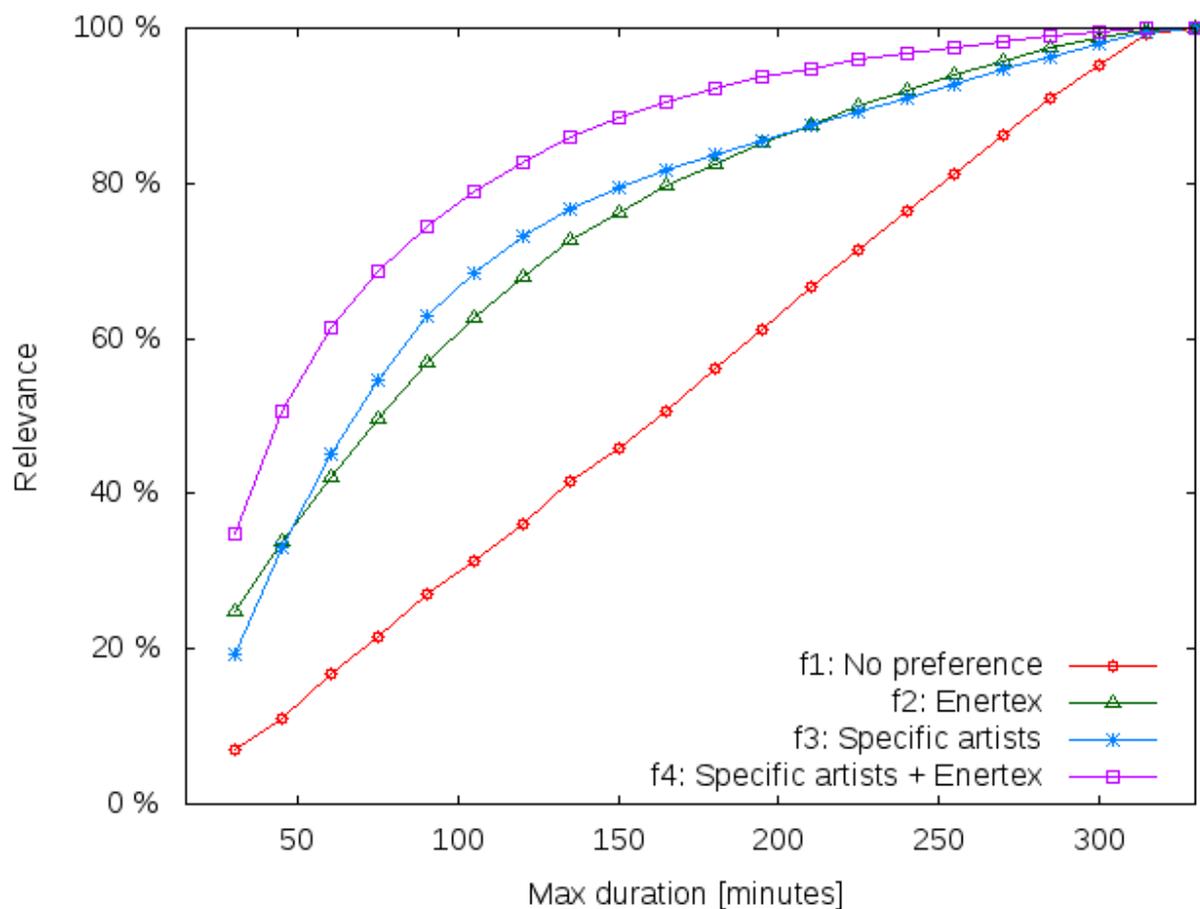


Fig. 4. Evolution of the relevance percentage

- [11] A. Molina, J.-M. Torres-Moreno, E. SanJuan, G. Sierra, and J. Rojas-Mora, "Analysis and Transformation of Textual Energy Distribution," in *(MICAI), 2013 12th Mexican International Conference on Artificial Intelligence*. IEEE, 2013, p. 203–208.
- [12] G. Salton and M. McGill, *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company, 1983.

- [13] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, July 1980.
- [14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999. [Online]. Available: [citeseer.ist.psu.edu/635422.html](http://citeseer.ist.psu.edu/635422.html)