A Clustering-based Approach for a Finest Biological Model Generation Describing Visitor Behaviours in a Cultural Heritage Scenario

Salvatore Cuomo[1], Pasquale De Michele[1], Giovanni Ponti[2] and Maria Rosaria Posteraro[1]
[1]Department of Mathematics and Applications, Univeristy of Naples "Federico II", Naples, Italy
[2]UTICT-HPC, ENEA Portici Research Center, Naples, Italy

Keywords: Computational Neural Models, Clustering, Data Mining, User Profiling.

Abstract: We propose a biologically inspired mathematical model to simulate the personalized interactions of users with cultural heritage objects. The main idea is to measure the interests of a spectator w.r.t. an artwork by means of a model able to describe the behaviour dynamics. In this approach, the user is assimilated to a computational neuron, and its interests are deduced by counting potential spike trains, generated by external currents. The main novelty of our approach consists in resorting to clustering task to discover natural groups, which are used in the next step to verify the neuronal response and to tune the computational model. Preliminary experimental results, based on a phantom database and obtained from a real world scenario, are shown. To discuss the obtained results, we report a comparison between the cluster memberships and the spike generation; our approach resulted to perfectly model cluster assignment and spike emission.

1 INTRODUCTION

In the cultural heritage area, the needs of innovative tools and methodologies to enhance the quality of services and to develop smart applications is an increasing requirement. Cultural heritage systems contain a huge amount of interrelated data that are more complex to classify and analyze.

For example, in an art exhibition, it is of great interest to characterize, study, and measure the level of knowledge of a visitor w.r.t. an artwork, and also the dynamics of social interaction on a relationship network. The study of individual interactions with the *tangible culture* (e.g., monuments, works of art, and artifacts) or with the *intangible culture* (e.g., traditions, language, and knowledge) is a very interesting research field.

To understand and to analyze how artworks influence the social behaviours are very hard challenges. Semantic web approaches have been increasingly used to organize different art collections not only to infer information about an opera, but also to browse, visualize, and recommend objects across heterogeneous collections (Middleton et al., 2003). Other methods are based on statistical analysis of user datasets in order to identify common paths (i.e., *patterns*) in the available information. Here, the main difficulty is the management and the retrieval of large databases as well as issues of privacy and professional ethics (Kumar et al., 2010). Finally, models of artificial neural networks, typical of Artificial Intelligence field, are adopted. Unfortunately, these approaches seems to be, in general, too restrictive in describing complex dynamics of social behaviours and interactions in the cultural heritage framework (Kleinberg, 2008).

In this paper, we are interested in analyzing visitor behaviours in cultural assets by means of biological inspired mathematical models (Cuomo et al., 2011; Cuomo et al., 2013; Bianchi et al., 2014). Here, the main novelty w.r.t. previously proposed approaches consists in exploiting unsupervised data groupings to estimate the values characterizing neuron electrical properties that allow to model it as a simple electrical circuit. More specifically, we resorted to a *clustering* task to obtain data groups by employing the well-known K-means algorithm (Jain and Dubes, 1988). This strategy has the main advantage of producing data groups (i.e., *clusters*) that highlight hidden patterns and previously unknown features in the data, without the need of any class labeling or training set.

In the next phase of our approach, we refer to a computational neuroscience terminology for which a cultural asset visitor is *a neuron* and its interests are *the electrical activity* which has been stimulated by appropriate currents. More specifically, the dynamics

Cuomo S., De Michele P., Ponti G. and Posteraro M..
A Clustering-based Approach for a Finest Biological Model Generation Describing Visitor Behaviours in a Cultural Heritage Scenario.
DOI: 10.5220/0005144104270433
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (KomIS-2014), pages 427-433
ISBN: 978-989-758-035-2

of the information flows, which are the social knowledge, are characterized by neural interactions in biological inspired neural networks. Reasoning by similarity, the users are the neurons in a network and its interests are the morphology; the common topics among users are the neuronal synapses; the social knowledge is the electrical activity in terms of quantitative and qualitative neuronal responses (spikes). This lead to produce a characterization of user behaviours in exhibits, starting from a real world scenario.

The work is organized as follows. In Section 2 we report the mathematical background of the problem. In Section 3 we discuss a motivation example. The proposed approach based on clustering and the neuron modeling are shown in the Section 4. The Section 5 is devoted to the related works. Finally, the conclusions are drawn in the Section 6.

2 MATHEMATICAL BACKGROUND

A mathematical model, corresponding to a particular physical system S, consists of one or more equations, whose individual solutions, in response to a given input, represent a good approximation of the variables that are measured in S. A biological neuron model consists of a mathematical description of nervous cell properties, more or less accurate, and allows to describe and predict certain biological behaviours. A neuron can be modeled at different levels of complexity: if we consider the propagation effects, then we have compartmental models defined by means of Partial Differential Equations (PDEs); if, instead, we assume that the action potential propagation is almost instantaneous if compared to the time scale of the generation of itself, then we have single compartment models defined by means of Ordinary Differential Equations (ODEs) and algebraic equations.

The *Integrate&Fire* (*I&F*) is a simple ODE model that considers the neuron as an electrical circuit, in which only the effects of the membrane capacitance are evaluated. The circuit is represented by the time derivative of the capacitance law ($Q = CV$), that is

$$\begin{cases} \frac{dV_m}{dt} + \frac{V}{\tau} = \frac{I}{C_m} \\ V_m(0) = V_0 \\ \text{if} \quad \exists t : V_m(t) = \theta \rightarrow V_m(t)^+ = 0 \end{cases}$$

where $t^+ = t + \varepsilon$ with ε very small, V_m is the membrane potential, C_m is the membrane capacitance, $I(t)$ is the ionic current of the neuron m, $\tau = R_m \cdot C_m$ and R_m is the resistance. By the previous definition we have that

$$C_m \frac{dV_m}{dt} = -\frac{V_m}{R_m} + I(t)$$

The application of an external current in input leads a membrane potential increase, until this reaches a threshold value: at this point the neuron emits a spike, after which the potential V_m returns at the rest value. The *I&F* describes simplified biological dynamics able to illustrate only some features of the neuronal activities. Our goal is to apply the discussed model to a case study of an artwork visitor of a cultural heritage asset in an exhibit.

3 MOTIVATION EXAMPLE

We start to analyze data collected from a real scenario. In particular, the key point event was an art exhibition within Maschio Angioino Castle, in Naples (Italy) of sculptures by Francesco Jerace, promoted by DATABENC (Databenc, 2013), a High Technology District for Cultural Heritage management recently founded by Regione Campania (Italy). The sculptures was located in three rooms and each of them was equipped with a sensor, able to “talk” with the users. After the event, the collected data have been organized in a structured knowledge entity, named “booklet” (Chianese et al., 2013b). The booklet contents are necessary to feed the artworks fruition and they require a particular structure to ensure that the artworks start to talk and interact with the people. The Listing 1 shows a XML schema diagram of a simplified model of the booklet entity, characterized by the attributes of an artwork.

In this paper, we analyze the *log file* of a phantom database that was populated with both real and random data. It represents the basic knowledge on which we test the applicability of the proposed biological inspired model.

4 THE BIOLOGICAL INSPIRED MODEL

The *I&F* model can be adopted to characterize the user dynamics w.r.t. the interactions with an artwork. In (Cuomo et al., 2014), this issue has been addressed by proposing a novel approach to find the *I&F* dynamic correlations with the output of a such well-known classification method. Data have been analyzed through a naive Bayesian Classifier, in order to have a comparison metric.

Listing 1: An example of the structured LOG file.

```xml
<?xml version="1.0" encoding="UTF-8"?>
 <USER ID='UI001'>
        <STEREOTYPE_USER>2</
            STEREOTYPE_USER>
        <START_SESSION></START_SESSION>
        <END_SESSION></END_SESSION>
   <TRANSACTION>
    <REQUEST>
      <HTTP_METHOD>GET</HTTP_METHOD>
      <PATH_INFO>/opera</PATH_INFO>
      <REQUEST_PARAMETERS>
        <CODEARTWORK>ART0224VICTA</
            CODEARTWORK>
        <DATE>13/05/2013</DATE>
          </REQUEST_PARAMETERS>
        <REMOTE_ADDRESS>192.168.1.6</
            REMOTE_ADDRESS>
   </REQUEST>
     <PARAMETERS_LOG>
        <HOUR_LISTEN_START>13/05/2013
            13:58:12</HOUR_LISTEN_START
            >
        <HOUR_LISTEN_END>13/05/2013 14
            :05:42</HOUR_LISTEN_END>
        <AUDIOS>
     <TOT_NUMBER>3</TOT_NUMBER>
     <AUDIO ID='AU1111'>
           <HOUR_END>13/05/2013 14
               :00:42</HOUR_END>
      <LENGTH>180</LENGTH>
   </AUDIO>
      </AUDIOS>
        <IMAGES>
     <TOT_NUMBER>11</TOT_NUMBER>
     <IMAGE  ID='IM1122' />
     <IMAGE  ID='IM1134' />
     <IMAGE  ID='IM1135' />
      </IMAGES>
        <VIDEOS>
     <TOT_NUMBER>2</TOT_NUMBER>
     <VIDEO ID='VI3333'>
           <HOUR_END>13/05/2013 14
               :20:12</HOUR_END>
      <LENGTH>180</LENGTH>
   </VIDEO>
      </VIDEOS>
        <TEXTS>
     <TOT_NUMBER>4</TOT_NUMBER>
     <TEXT ID='TX4455' />
          <TEXT ID='TX4456' />
        <TEXT ID='TX4457' />
      <TEXT ID='TX4458' />
</TEXTS>
    </PARAMETERS_LOG>
  </TRANSACTION>
</USER>
```

In this work, we propose a new strategy to discover classes in the data which can be used for the next modeling step, that is the tuning of the electrical parameters for the circuit model characterizing the neuron. In fact, classification algorithms have the major limitation of labeling data according to a yet-known training set, as they are supervised approaches. In many real world datasets, data objects do not typically have assigned class membership, and this may lead to have accuracy issues in the whole classification process.

For this reason, we propose to address such an issue by introducing a *clustering*-based approach (Hartigan, 1975; Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990) to discover data groups. Clustering is an *unsupervised* task, since it can be applied to unclassified data (i.e., unlabeled) to obtain homogeneous object groupings. In this approach, groups are more representative w.r.t. single object as they summarize their common features and/or patterns; indeed, objects belonging to the same group are quite similar each other, whereas objects in different groups are quite dissimilar.

In our context, data to be clustered are *tuples* representing visitor's behaviours related to an artwork. Note that now "spike" has a more informative role in the dataset, as it is not seen as a class but as a further information about visitor's behaviour. In our experiments, we assume the following criteria for spike generation. A visitor enjoyed an artwork if he benefits from the whole content of at least one of the available services, or if he exploits more than the 66% of the total contents.

This new clustering-based approach allows us to produce a more general dataset, in which we do not need to assign object classes, and also attributes can take values in a continuous range, instead of in a discrete one. Therefore, the clustering phase produces groups according to visitor's preferences, which are not necessary driven by spike generation.

We have organized the log file structure, discussed in the previous section, in a Weka's `ARFF` file format (Weka) and we have used it as an input of the clustering task. In the following, we show the `ARFF` file suitable for clustering process.

```
@RELATION ARTWORK
@ATTRIBUTE audios NUMERIC [0..1]
@ATTRIBUTE images NUMERIC [0..1]
@ATTRIBUTE texts NUMERIC [0..1]
@ATTRIBUTE spike {0,1}
@DATA
0.1,0.4,1.0,1
0.3,0.6,0.4,0
...
0.5,1.0,0.7,1
...
```

In the proposed scheme, data values represent the amount of information that the visitor has exploited for an artwork for each attribute of the dataset, and the last attribute describes the spike generation according to the algorithm previously described. In this way, combining the values of the attributes `audios`, `images` and `texts`, it is possible to obtain a total of $N = 1,331$ different data objects (i.e., tuples) — for simplicity, we take into account just real values rounded at the first decimal value.

As regards the clustering task, we can employ any algorithm to discover groups. However, in this paper, we resorted to the well-known K-means clustering algorithm (Jain and Dubes, 1988). K-means requires only one parameter, that is the number K of clusters (i.e., groups) to be discovered. Algorithm 1 shows the outline of the K-means clustering algorithm.

Algorithm 1: K-means.

Require: a dataset objects $\mathcal{D} = \{o_1, \ldots, o_N\}$; the number of output clusters K
Ensure: a set of clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$

for $i = 1$ to K **do**
  $c_i \leftarrow$randomInitialize($\mathcal{D}$)
end for
repeat
  for all $C_i \in \mathcal{C}$ **do**
    $C_i \leftarrow \emptyset$
  end for
  for all $o_u \in \mathcal{D}$ **do**
    $j \leftarrow argmin_{i \in [1..K]} dist(o_u, c_i)$
    $C_j \leftarrow C_j \cup \{o_u\}$
  end for
  for all $C_i \in \mathcal{C}$ **do**
    $c_i \leftarrow$ updateCentroid(C_i)
  end for
until centroids do not change or a certain termination criterion is reached

In our experiments, we first started with $K = 2$, which is the natural starting choice to model a classification-like approach (i.e, "spike" or "no-spike"). Nevertheless, we can also perform further experiments by setting higher values for K to capture finest similarities and/or hidden patterns in the data. Figure 1 shows the output of the clustering phase with $K = 2$. Note that we do not take into account the "spike" attribute in the clustering process, as it could clearly bias the entire process. However, we exploited it at the end of the clustering phase to assess the result accuracy. We resorted to Weka "simpleKMeans" implementation, and the plot is also obtained employing Weka clustering visualization facilities.

The plot represents tuples in terms of cluster membership (x-axis) and spike emission (y-axis). It is easy to note that all the data in *cluster0* refer to tuples that

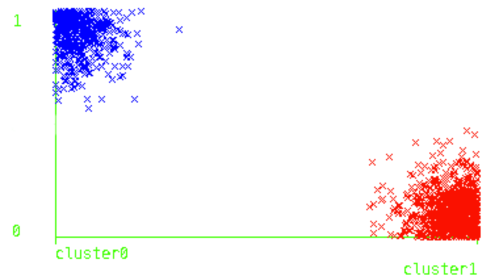

Figure 1: Clustering results for K-means ($K = 2$).

produce spikes (i.e., with value 1), whereas all the ones in *cluster1* identify tuples that do not emit spike (i.e., with value 0). Therefore, evaluating clustering results in terms of well-separation of the data w.r.t. the spike emission issue, we achieved a high-quality clustering as all the data have been correctly separated.

Starting from the clustering output, in a second approach, we have integrated the *I&F* computational model in order to find some correlations with the clustering results. In particular, the couple (R_m, C_m) represents the visitor sensitivity to the artwork. We have exploited the clustering results in order to tune the values of the resistance R_m and conductance C_m of the circuit that represents the model. In a first experiment, a good choice for the couple (R_m, C_m) is

$$(R_m, C_m) = (0.51kOhm, 30\mu F)$$

The current is a linear combination of the values of the attributes in the dataset. The Figure 2 gives the dynamic response of the neuron.

In the first case (top of the Figure 2) the current $I(t)$ is not sufficient to trigger a potential difference which gives a spike. In the second one (bottom of the Figure 2) the neuron that has received stimuli is able to produce an interesting dynamic.

In these experiments, we show how the computational model and the clustering give information about the interest of a visitor about an artwork. In the Table 1, experimental results for the clustering and our model are reported. M.C.F. represents the Media Content Fruition w.r.t. the overall media contents. With the symbol (*) we have labeled the tuple combinations that contain the information about the fully fruition of at least one media content. Note that the last column of the table indicates the degree of the visitor interest for an artwork. Thus, in this respect, such an information is obtained by the proposed *I&F* neuron model to achieve a fine-grained indication for spikes.

Let us suppose that we have two users with different sensitivity (R_m, C_m) respect to a fixed artwork. The question is *how is the behaviour of the users in presence of the same combination of stimuli repre-*

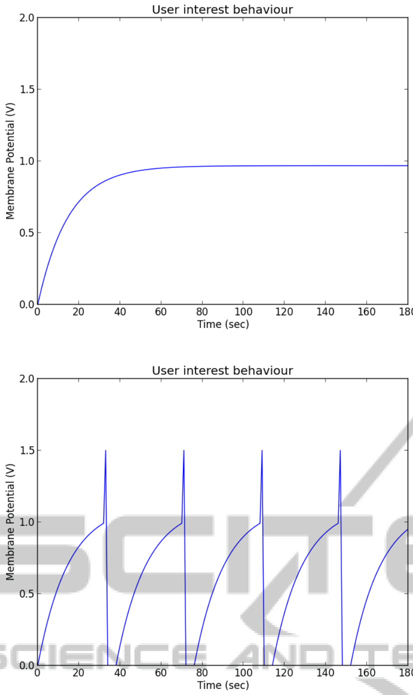

Figure 2: **Top**. With a current $I(t) = 0.6 + 0.6 + 0.7$, we observe no spike presence. **Bottom** With a current $I(t) = 0.6 + 0.8 + 0.8$ we observe 4 spikes.

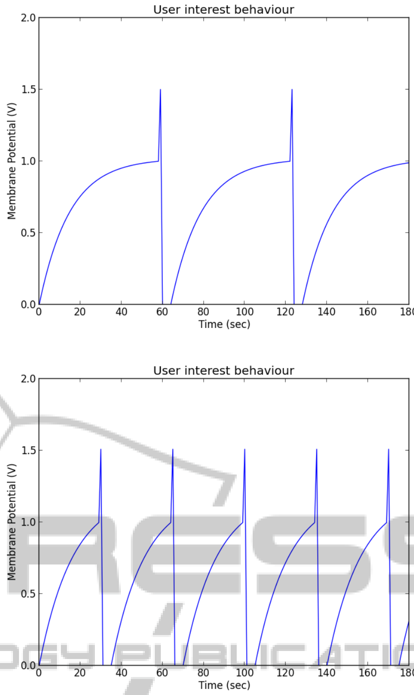

Figure 3: **Top**. With the couple $(R_m, C_m) = (0.51, 30)$ the neuron has 2 spikes. **Bottom** With the couple $(R_m, C_m) = (0.6, 28)$ the neuron has 5 spikes.

Table 1: Spike response for clustering and $I\&F$ neuron with $(R_m, C_M) = (0.51kOhm, 30\mu F)$.

Tuples	M.C.F. (%)	Cluster	# spikes
$0.2, 0.2, 0.2$	20%	cluster1	0
$0.2, 0.2, 0.4$	27%	cluster1	0
$0.4, 0.2, 0.2$	27%	cluster1	0
$0.6, 0.6, 0.7$	63%	cluster1	0
$0.6, 0.8, 0.8$	73%	cluster0	4
$0.7, 0.9, 0.5$	70%	cluster0	4
$0.8, 0.9, 0.3$	67%	cluster0	2
$0.8, 0.9, 0.6$	76%	cluster0	5
$1.0, 0.2, 0.1$	$43\%^{(*)}$	cluster0	5
$1.0, 0.8, 0.9$	$90\%^{(*)}$	cluster0	10
$1.0, 1.0, 0.6$	$86\%^{(*)}$	cluster0	13
$1.0, 1.0, 1.0$	$100\%^{(*)}$	cluster0	16

sented by tuple interest values? The clustering-based model can not answer to this question in a simple way. In fact, taking into account only $K = 2$ clusters, we just distinguish between two behaviours, that are "spike" and "no-spike". For this reason, here we highlight the feature of $I\&F$ model to address the problem.

In the Figure 3, we have fixed

$$I(t) = 0.8 + 0.9 + 0.3$$

as a stimulus and we have compared two users U_1 with $(R_m, C_m) = (0.51, 30)$ and U_2 with $(R_m, C_m) = (0.6, 28)$.

We can observe the different number of spikes between U_1 and U_2 respect to the same artwork. If the spike are related to the the interests that a cultural asset has aroused in a viewer, the $I\&F$ is able to emerge this features. The choice of the pair (R_m, C_m) suitable for a established user is the real challenge of the model. More in general, it may be multiple scenarios to apply these dynamics. An example is the case of a cultural asset exhibition in which the target is how to place artworks. A possible choice is to select the operas that have attracted the visitors with common interests, i.e., users with similar (R_m, C_m). In the context-aware profiling instead the aim is how to change (R_m, C_m) in such a way to predict the user behaviours in terms of spikes that represent its cultural assets.

5 RELATED WORK

The studying of efficient methods for learning and classifying the user behaviours and dynamics in the real or digital life is a very large and fascinating research area. The challenge is to have automatic frameworks based on sensor networks, semantic web models, reputation systems and classifiers able to map human activity and social user interactions. More in details a smart system should be have the ability to automatically infer interests of users and track the propagation of the information. For real life applications, in (Amato et al., 2013; Chianese et al., 2013a) a wireless sensor network, using bluetooth technology, able to sense the surrounding area for detecting user devices' presence in a museum is discussed. About the digital user behaviours a study of the relevance of feedbacks, typically adopted for the profiling during long-term modeling is given in (Kelly and Teevan, 2003). In (Widyantoro et al., 2001) an algorithm based on the descriptors representation is developed to acquire high accuracy of recognition for long-term interests, and to adapt quickly to changing interests in the learning user activity. Other methodologies using computational approaches are based on machine-learning (Domingos, 2012). Here, the focus is to estimate the dynamics of the users' group membership and to characterize the social relationships by means of behaviour patterns with statistical learning methods. In (Pentland, 2007), using the users data to model an individual behaviour as a stochastic process, the authors show a framework that predicts the future activity, obtained by modeling the interactions between individual processes. Ontological methodologies for user profiling in recommender systems are described in (Middleton et al., 2003). Finally, a multimedia recommender system based on the social choice problem has been recently proposed in (Albanese et al., 2013).

6 CONCLUSIONS

In this paper, we describe a framework that is closed to the computational methodology, adopted to infer information about visitors in a cultural heritage context. The challenge is to map, in a realistic way, the biological morphology of a neuron in this application scenario. We deal with a model where the (R_m, C_m) couple represents the sensitivity of the user respect to an artwork. The main novelty of our work has been to employ a clustering algorithm methodology to obtain starting groups from which these electrical parameters can be tuned.

A very nice issue is to adapt, in a smart way, this computational framework to many different application issues such as the context-aware profiling, feedback based system or recommendation systems. In future research lines, we will study more complex neuronal dynamics by morphology point of view with the aim to develop models that are more close to the real users. Other research tracks will be the building of computational neural networks able to reproduce the interactions in social cultural heritage networks. In addition, regarding the preliminary clustering phase, we will tune our model with more than two clusters, with the aim of obtaining fine-grainer clustering solutions that are able to capture and to highlight other neuron aspects, apart from spike generation.

ACKNOWLEDGEMENTS

Authors thank DATABENC, a High Technology District for Cultural Heritage management of Regione Campania (Italy), and ENEA Portici Research Center, UTICT-HPC Department, for supporting the paper.

REFERENCES

Albanese, M., d'Acierno, A., Moscato, V., Persia, F., Picariello, A.: A Multimedia Recommender System. ACM Trans. Internet Technol. (2013) 13(1) 3:1–3:32.

Amato, F., Chianese, A., Mazzeo, A., Moscato, V., Picariello, A., Piccialli, F.: The Talking Museum Project. Procedia Computer Science. (2013) 21(0) 114–121.

Bianchi, D., De Michele, P., Marchetti, C., Tirozzi, B., Cuomo, S., Marie, H., Migliore, M.: Effects of increasing CREB-dependent transcription on the storage and recall processes in a hippocampal CA1 microcircuit. HIPPOCAMPUS. 24(2) (2014) 165–177.

Chianese, A., Marulli, F., Moscato, V., Piccialli, F.: SmARTweet: A location-based smart application for exhibits and museums. Proceedings - 2013 International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2013. (2013) 408–415.

Chianese, A., Marulli, F., Piccialli, F., Valente, I.: A novel challenge into multimedia cultural heritage: An integrated approach to support cultural information enrichment. Proceedings - 2013 International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2013. (2013) 217–224.

Cuomo, S., De Michele, P., Chinnici, M.: Parallel tools and techniques for biological cells modelling. Buletinul Institutului Politehnic DIN IASI, Automatic Control and Computer Science Section. LXI (2011) 61–75.

Cuomo, S., De Michele, P., Piccialli, F.: A performance evaluation of a parallel biological network microcircuit in neuron. International Journal of Distributed & Parallel Systems. 4(1) (2013) 15–31.

Cuomo, S., De Michele, P., Posteraro, M. (2014): A biologically inspired model for describing the user behaviors in a Cultural Heritage environment. SEBD2014, 22nd Italian Symposium on Advanced Database Systems, June 16th - June 18th 2014, Sorrento Coast.

DATABENC, High Technology District for Cultural Heritage, http://www.databenc.it

Domingos, P.: A Few Useful Things to Know About Machine Learning. Commun. ACM. 55(10) (2012) 78–87.

Hartigan, J. A.: *Clustering Algorithms*. Applied Statistics. John Wiley & Sons, 1975.

Jain, A. K., and Dubes R. C.. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

Kaufman, L., and Rousseeuw, P. J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum. 37(2) (2003) 18–28.

Kleinberg, J.: The convergence of social and technological networks. Commun. ACM 51, 11, 2008 66–72.

Kumar, R., Novak, J., Tomkins, A.: Structure and Evolution of Online Social Network. Link Mining: Models, Algorithms, and Applications J. Am. Soc. Inf. Sci. Technol. 978-1-4419-6515-8 (2010) 337–357.

Middleton, S. E., Shadbolt, N. R., De Roure, D. C.: Capturing Interest Through Inference and Visualization: Ontological User Profiling in Recommender Systems. Proceedings of the 2Nd International Conference on Knowledge Capture. (2003) 1-58113-583-1 62–69.

Pentland, A. S.: Automatic mapping and modeling of human networks. Physica A. (2007) 378(1) 59–67.

Roderick J. A. Little, Donald B. Rubin : Statistical Analysis with Missing Data. Wiley Editor. 978-0-471-18386-0 (2002).

Weka, Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

Widyantoro, D. H., Ioerger, T. R., Yen, J.: Learning User Interest Dynamics with a Three-descriptor Representation. J. Am. Soc. Inf. Sci. Technol. 52(3) (2001) 212–225.