Web Mining Techniques Applicable for Cultural Heritage Observations

Emanuela Mitreva[1], Alexandra Nikolova[2], Vladimir Georgiev[3]

[1] Sofia University St. Kliment Ohridski, Sofia, Bulgaria
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[3] Department of Computer Science, American University in Bulgaria, Blagoevgrad, Bulgaria
emitreva@gmail.com, alxnikolova@gmail.com, vgeorgiev@aubg.edu

Abstract. The rapid digitalization in the cultural domain generated enormous amounts of data. But that data is not personalized or processed, thus the benefit for different users is limited, due to the variety of users – people learn through different methods. However if techniques used in the last decade to personalize web content are applied to cultural content more users could easily absorb and process the cultural heritage knowledge. Also techniques could be applied to cluster and extract useful knowledge.

Keywords: Data Mining, Recommendation Systems, Cultural Heritage, Data Collection.

1 Introduction

Nowadays, a lot of the cultural content has been digitized (Díaz-Corona et al., 2018) in the efforts to save the cultural heritage. As a result, the cultural content can be disseminated through websites, digital libraries, and virtual museums. However, the heterogeneity and disorganization of the information make it difficult to properly and easily process it through an uniform method.

In the cultural domain we could differentiated two areas of improvement – organizing the documents – using clustering and other data mining methods to extract knowledge from the content (Rehm et al., 2019) – or using data mining methods to personalize the way users perceive and interact with the cultural content (Konstantakis & Caridakis, 2020).

In the last decade, different methods were used to improve the user experience. Personalization and recommendation systems have been applied in some fields for a long time – movies, shopping items, but for cultural content, it is yet to be applied effectively (Hong et al., 2017).

Some of the techniques used in other areas can be applied to cultural domain to improve the experience of the tourists and we are highlighting some of them in this research.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 11, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2021. ISSN: 1314-4006, eISSN: 2535-0366

2 Techniques for Personalization and Mining

The amount of data in the cultural domain is extensive because of the rapid digitalization of all the data in the domain. Digitalization of the cultural content is much needed to preserve our cultural heritage, but the process is done so promptly, that it also generates problems – too much data is unavailing if it cannot be transformed to knowledge or it cannot be properly used. To solve those problems, diverse techniques should be used to transform that data into knowledge. Also, technologies can really improve the way cultural content is processed by people – the cultural heritage can no longer be static and disorganized content. The data in the domain is transformed in a more interactive way that has an impact on people (Hong et al., 2017).

However, to effect people, the correct approach or approaches need to be found. One thing to recon is that the consumers of cultural content are very heterogeneous - some people process data through visual aids, others through written data (Raptis & Avouris, 2019). If we also include the factor of learning from the cultural content and the goal to change the content from boring to interesting and interactive (Stanchev et al., 2017), then personalization is to be contemplated. But the amount of cultural data – both written (documents, pdfs, etc.) and visual (images) is so diverse and immeasurable that it is becoming more and more challenging to organize it. Organizing it can also facilitate better analysis and understanding of the data (Santos et al., 2019).

It is clear that improvements need to be done in the cultural domain so that the barrier between the people and data is destroyed. Regardless of the domain, the first two steps that need to be taken when dealing with data, are data collection and data transformation (Santos et al., 2019). In this research, we are covering explicit and implicit ways (Konstantakis & Caridakis, 2020) to gather the information needed to perform personalization or recommendation and approaches that could be applied after the data is collected.

In the cultural domain, the problem of organizing the data can be solved with data processing - personalization of how cultural content is displayed, extracting information from the content, and enriching it with knowledge. The explicit methods require input from the users - through questionnaires or by configuring the content to be displayed in certain way. The implicit methods require more effort - identifying what information could be used for the techniques (visitors' logs, other logs, social media information, etc.) and applying the proper method to transform the data to knowledge. In the next two subsections, we are going into more details for both the explicit and implicit methods.

2.1 Explicit Methods

In our previous research (Christozov & Mitreva, 2020) we have proposed that the user should be able to explicitly configure how a piece of certain information is displayed and to setup the details of the explanation. This method is also acceptable for displaying cultural content – the users of a certain website showing cultural content should be able to configure what and how the information should be displayed.

We have researched the ways Drupal as a content management system can offer certain extent of personalization based on the preferences of the users. What should be the level of details in the properties, is also something to consider, because some of the users might not be technology savvy. The idea is that Drupal offers ways to show/hide/prioritize menus and with such capabilities the system should be flexible enough for most users. Nevertheless, there will be a default settings for users without specific preference or without the knowledge to configure it. How to determinate what will be default setting is something that should be researched or discovered through feedback from users, surveys, etc. It will be easier for the users that the personalized view to be done automatically through the system that is being implemented. So after the initial web mining filtering and processing is done, the changes could be applied through the functionalities that the Drupal modules offer.

The second option to gather the information and explicitly states how the content should be exhibited is to use questionnaires that can be distributed among users. This approach has a couple of drawbacks - firstly, the questionnaire needs to be carefully created, otherwise, the result from it might not be just not useful, it might be misleading. Secondly, this should be done on iterations, which means it will take time to generate some helpful conclusions. The feedback from different iterations should be processed and applied. Another thing to consider with this approach is whether the questionnaire is distributed to diversely enough groups so that the proper conclusions can be drawn from the results. Also are people willing to participate in such surveys, so that enough data is collected from the answers. If not enough data is generated, then the results might not be helpful, or might not correspond to various users.

Both ways, though, require time to gather the necessary information or feedback from the users so that the proper setting or preference is applied. And both methods require the people to be willing to provide the needed information.

2.2 Web Mining Methods

In the previous section, we have discussed the explicit ways to collect the data to personalize the way the cultural content is displayed. The previous methods require people to voluntarily to either participate in surveys or to provide their preference of setting. In this section, we are presenting different approaches to collecting useful data for users without the need of the users to participate in surveys or to provide any data.

The first approach that we are going to conder is web usage mining – a method that we have discussed in previous research (Christozov & Mitreva, 2020). This method could be used in case we have a way to obtain the web access logs of a UI that is presenting cultural content. To apply that method the logs need to be collected and filtered – a lot of the logs correspond to wrong requests, crawlers that are gathering information about the tree of websites, etc. In the system that we have implemented, such records are filtered and then only the rest of the records are used to generate several profiles. We are considering the generation of different profile per user is involving too much complexity and unneeded diversity. It is better to create a few general profiles to cover most of the needs and preferences of the users. Moreover, if the web sources that are providing the cultural content don't require authentication of any kind, then the task

to separate the different users becomes almost an impossible to handle. We are reconning that although the diversity of the users is enormous in this area, the different types of profiles is finite. After the profiles are created, we could direct our efforts and research into two directions – generating some recommendations for the designers of the web resource or trying to generate some automation in the application of those recommendations. However, not much is done on the latter task. In the previous section we have mentioned our efforts to use Drupal to automatically apply rules generated from the profiles.

The second method is proposed in (De Angelis et al., 2017; Sansonetti et al., 2019). It is suggesting using social media and the user's activities to enrich that information with knowledge, provided by lined-open data (LOD). They are stating that the recommender system should take into account the lifestyle and the personal preference of the visitors. But that recommender system would also need additional data that can be obtained by some of the explicit way we have described – surveys or/and questionnaires (Hong et al., 2017). The method is to be combined with the first approach - the history from the previous visits of the user and their social media information about their visits can be used to generate a profile or to find the proper already existing profile. That profile could be used the next time the user accesses the cultural content. It will be even more beneficial to use custom Drupal module to automatically apply the settings gather in the previous step. Using the logs and social data and the conclusions that were drawn from that data, might help revise the way the content is displayed.

The next source of data that could help with the personalization or can create cultural content recommendations is through a search history (Sumikawa et al., 2019) of digital libraries. This method uses something similar to web usage mining – it uses what searches the users have performed to generate profiles. Thus, if a user makes similar searches as a certain group or profile, then they could be offered the same content as the one that the group was interested in.

2.3 Comparison of the Methods

In this section, we provide a comparison of the methods we have discussed in the previous two sections.

Table 1. Comparison table of the different methods

Method/ Source	Description	User input is needed	Processing is needed
Surveys/Questionnaires	The data is retrieved through surveys distributed to the users.	x	x
Predefined configurations	The UI interface provides predefined settings of how the content should be displayed	x	x
Web usage mining	Based on the collected access logs and other logs a user profile is created with a sample preference for personalization		x

Using social media information		x
Search history	Using the search history of the user to generate an area of interests	x

As Table 1 shows we have different methods of collecting the data, a few of them require some user input – it is either through a form of answering questions that are part of the survey or by directly allowing them to configure what they see. However, all of the methods require some additional processing. We have briefly mentioned this, but the data that is gathered would first need some filtered or cleaning, before we could proceed to the next step of processing. For example, for web usage mining it is advisable to filter all unnecessary access records – all data that is generated from wrong requests will only our results unhelpful and even wrong. Also, the actual processing might require human feedback or intervention due to several reasons – to fine grain the algorithm, to remove outliers, to prevent user privacy violation (Ferrara et al., 2014).

3 Future Work

In this research, we have discussed the different methods to collect data and we have concluded that regardless of the method some pre-processing (cleaning of the data) and additional processing will be used. Instead of trying to figure out what approach is better to use, what we could do is use all of the methods – to collect data from all the different sources. However, for all of the data, we need first to clean the data from the outliers or wrong entries so that the process of transforming the data to knowledge works on only the meaningful data. We have already mentioned some ways that we used in our system of filtering that data – cleaning all errors or automatic requests. The same can be applied on searches - clean all searches that results in no result or if possible, identify and filter out searches that were done by some automatic crawlers. For the social media data, it will be a little more difficult to determine what could be useful and what should be filtered out, not to mentioned that we should be careful about violating some privacy laws.

After the data is filtered, we could proceed to the next step and extract useful rules to be applied in a Drupal module. Finally, the two outcomes will be to generate the profiles and to offer the user either the necessary personalization and recommendation, based on the profile we have determined that they belong, or to be able to automatically change the UI according to that profile. One could argue that automatic changes might not be what some users would like, as for other users the seamless change to what might be what they need or like is the best solution. One enhancement to be done is to just offer the user some changes and only after approval to be applied.

4 Conclusion

In this paper, we have conferred the importance of the user experience and how it was improved and changed in the last decade. How the impact of the cultural content on the users, can be profoundly changed if some of the methods, used in other areas, are used in this domain. However, to accomplished that goal we have outlined that we have a few steps to take - to collect the data, to filter it, and to transform that data into knowledge. We have mentioned that we could have different sources of data or information – search logs, access logs, social media that could help us generate groups of profile. Based on the generated profiles we could offer the user a certain way of displaying the information, so that fits their learning way better or it just makes the information more interesting, but all approached would require to have the necessary time to collect and process the data.

References

Christozov, D., & Mitreva, E. (2020). TRUST IN LEARNING FROM BIG DATA: THE TWO SIDES OF THE SAME COIN. *Issues In Information Systems*. https://doi.org/10.48009/1_iis_2020_147-152

De Angelis, A., Gasparetti, F., Micarelli, A., & Sansonetti, G. (2017). A social cultural recommender based on linked open data. *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, 329–332. https://doi.org/10.1145/3099023.3099092

Díaz-Corona, D., Lacasta, J., & Nogueras-Iso, J. (2018). *Barriers for the access to knowledge models in Linked Data cultural heritage collections*. https://doi.org/10.1145/3230599.3230615_15

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). *Web Data Extraction, Applications and Techniques: A Survey*. https://doi.org/10.1016/j.knosys.2014.07.007

Hong, M., Jung, J. J., Piccialli, F., & Chianese, A. (2017). Social recommendation service for cultural heritage. *Personal and Ubiquitous Computing*, *21*(2), 191–201. https://doi.org/10.1007/s00779-016-0985-x

Konstantakis, M., & Caridakis, G. (2020). Adding culture to UX: UX research methodologies and applications in cultural heritage. *Journal on Computing and Cultural Heritage*, *13*(1). https://doi.org/10.1145/3354002

Raptis, G. E., & Avouris, N. M. (2019). Supporting Designers in Creating Cognition-centered Personalized Cultural Heritage Activities. *ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 407–411. https://doi.org/10.1145/3314183.3323868

Rehm, G., Lee, M., Moreno-Schneider, J., & Bourgonje, P. (2019). Curation technologies for cultural heritage archives analysing and transforming a heterogeneous data set into an interactive curation workbench. *ACM International Conference Proceeding Series*, 117–122. https://doi.org/10.1145/3322905.3322909

Sansonetti, G., Gasparetti, F., & Micarelli, A. (2019). Cross-domain recommendation

for enhancing cultural heritage experience. *ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 413–415. https://doi.org/10.1145/3314183.3323869

Santos, B. T., Tolentino, J., Aquino, D. N., Malibiran, R., Lim-Ramos, C. D., Cheng, C., & Ngo, C. A. (2019, December 2). A museum information system for sustaining and analyzing national cultural expressions. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3366030.3366132

Stanchev, P. L., Paneva-Marinova, D., & Iliev, A. (2017). Enhanced user experience and behavioral patterns for digital cultural ecosystems. *9th International Conference on Management of Digital EcoSystems, MEDES 2017*, *2017-January*, 287–292. https://doi.org/10.1145/3167020.3167063

Sumikawa, Y., Jatowt, A., Doucet, A., & Moreux, J. P. (2019). Large scale analysis of semantic and temporal aspects in cultural heritage collection's search. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, *2019-June*, 77–86. https://doi.org/10.1109/JCDL.2019.00021

Received: June 14, 2021
Reviewed: July 10, 2021
Finally Accepted: July 21, 2021