A Quantitative analysis of digital library user behaviour based on access logs

Adrien Nouvellet[1], Florence D'Alché-Buc[1], Valérie Baudouin[2], Christophe Prieur[2], François Roueff[1]
[1]LTCI, Télécom ParisTech - Université Paris-Saclay
[2]i3-SES, Télécom ParisTech - Université Paris-Saclay

Abstract: This paper quantitatively analyzes the usage of Gallica, a website platform for accessing the digital library of the Bibliothèque nationale de France (BnF). Our approach relies on the access logs retrieved from the Apache HTTP Servers of Gallica. The server access logs record all requests processed by the server and thus, contain the web pages and the timestamp of the requests along with the corresponding IP of the users. These access logs are augmented with additional structured data via The Open Archives Initiative Protocol for Metadata Harvesting in order to store, when it is possible, the metadata of consulted documents. Beyond straightforward statistics (such as the duration of a session, the number of documents consulted by each session, the most popular type of documents over all the Gallica users), our research aims to model user navigational behaviours by a Mixture of Continuous-Time Markov Chains. This model allows to cluster users into classes of typical paths of navigation on Gallica. The results provide relevant insights on the way the users interact with the interface of Gallica, highlighting the mean duration of some actions such as the interaction with the search engine or the consultation of documents.
Even if our approach requires the use of additional informations in order to properly interpret the models and the correlation that it highlights, it allows the integration of all types of behaviour, including the most stealthy and the most difficult to catch in traditional surveys, giving them their fair weight in terms of audience.

1. Introduction

The French National Library (Bibliothèque nationale de France — BnF) has been studying its users for a long time in order to adapt to new behaviors and preferences. For many years, studies in the reading rooms allowed the BnF to get a in-depth representation of readers through quantitative and qualitative surveys. Since the mid 2000's, the BnF has embraced the digital revolution with an extensive program of collection digitization. Nowadays, the digital library named Gallica offers more than 4 millions of documents accessible by everyone

Received: 14.3.2017 / Accepted: 19.12.2017 ISSN 2241-1925

online. Since, the face of the library audience has dramatically changed and online reading is now the dominant consultation methods for the documents with ~ 40, 000 users per day using Gallica. In this case traditional methods such as interviews and surveys are no longer fully adequate to identify the new behaviors of online users. To reinforce its academic position on digital uses, the BnF has teamed up with Télécom ParisTech, an engineering school and a research center specialized on digital technologies, to found the Bibli-Lab, a laboratory for the study of digital libraries uses and users. Many studies have been conducted in this domain. A review of literature identified 200 studies published between 1995 and 2003 see Tenopir (2003). Among the variety of research methodologies used in these papers mining web logs that trace the activity of users on the web site (fix and mobile) offer an interesting vision of library uses.

Other authors reviewed several papers based on transaction log analysis for studying the use and users of digital libraries Jamali et al. (2005). They identify the advantages and limits of log analysis. All authors point that the automatic and non-intrusive characteristic of log analysis is definitely the best asset for this approach allowing: access to the reality of uses, longitudinal studies, performance evaluation of the system, comparison of behavior between different groups. However on the negative side, they cite the following points: difficulty in of user identification (IP address, session ...), and the impossibility to access the motivations and goals of the user.

We consider the approach of using raw logs rather than logs as recommended by the authors. Mining logs offers an exhaustive vision of all uses: the used functionalities, the types of consulted documents, and the sequence of these actions. All these elements are impossible to obtain through interviews or surveys (which also tend to be answered by regular users close to the institution and miss newcomers or passing-by users). This paper aims at studying the feasibility and usability of data-mining algorithms on web logs to provide a better understanding of the behaviors of digital library users. The originality of our research lies in the modeling of sequences of actions. We focus on clustering algorithms to discover usage patterns in the browsing activity of Gallica's users. Among the various clustering approaches able to deal with sequential data, we have chosen a generative probabilistic approach based on a mixture of Markov models whose main feature is the interpretability of each class, using the parameters of each component of the mixture. Similar approaches have already successfully be applied for the modelisation of internet users see Cadez et al. (2003); Ypma and Heskes (2002). Those methods allow us to keep the sequential properties of a user browsing session in contrast with approaches that see user sessions as fixed-length vectors as in Mobasher et al. (2001). In the previously cited papers, sequences are considered time-homogeneous and do not take into account the time spent on each web page. In this paper, we propose a novel extension of the mixture of discrete-time Markov for non-homogeneous sequences i.e with taking account the duration of actions.

This allows us to highlight web pages that require a long reading process such as the consultation of a digital document. We illustrate the relevance of this model on web access logs from the website Gallica of the BnF. This paper is organized in 4 sections. Section 2 reports the pre-processing steps necessary for the exploitation of raw web access logs. Section 3 exposes formally the clustering algorithm that aim at distinguishing Gallica's user browsing patterns. This section also shows the results of the algorithms applied on 1 month of data. Finally Section 4 discusses how the results of the analysis may impact the services of Gallica and suggests new ways of possible improvements of the method in the next future.

2. Data

2.1. Log files:

The Apache HTTP Server provides very flexible and comprehensive logging capabilities. The server access log records all requests made when users browses a website. It allows to saved all users' queries as log files which consist of a set of lines corresponding to each queries and in the case of Gallica, we have the following information for each request

- encrypted IP address that preserves the uniqueness of the IPs.
- the geographical origin of the IP
- the day, hour, minute and second of the request.
- the HTTP request from the client.
- the status code that the server sends back to the client.
- the user-agent of the client (web browser version).
- the HTTP referer, i.e. the address of the web page that linked to the resource being requested.

Quantitative analysis with the access logs require two steps. First, one has to parse the log file i.e splitting each line into the distinguished fields listed above. Then, one must perform data cleansing which mainly consists of removing the failed requests and requests made by web robots. The deletion of unnecessary requests is fundamental to reduce the volume of data to be analyzed. The failed requests are easily identified with the status codes that are strictly higher than 299 and strictly lower than 200. In most cases, web robots are identifiable with their user-agent and can easily be filtered out Silva et al. (2013). However, some web robots could not be easily identified. Eliminating them is out of the scope of this paper but we refer to Tan and Kumar (2004) for more insights into the topic.

2.2. Identification of requests type and metadata:

As explained, data logs records all the requests made by users and thus contains a wide varieties of queries. Instead of conducting an exhaustive analysis of everything recorded in the logs, we have adopted a user-centered approach, based on the results of a qualitative study conducted in Beaudoin et al. (2016).

Interviews with Gallica users allowed us to identify a finite and reasonable number of requests type made by "Gallicanautes'' during a session. We therefore extracted in the logs all the information that corresponds to these five actions, which reflect the main steps of a session in terms of usage:

- Action-1 Homepage : Browsing the homepage
- Action-2 Mediation : Browsing the mediation area
- Action-3 Search : Using the built-in search engine
- Action-4 Download : Downloading a document
- Action-5 Browsing : Browsing/Consulting documents

These actions are clearly identifiable with a set of requests that uniquely correspond to each of the actions. Action-1 is performed whenever a user requests the homepage of Gallica (http://gallica.bnf.fr/). The Action-2 refers to the BnF mediation including the blog (http://gallica.bnf.fr/blog) and the collection pages (URLs starting with http://gallica.bnf.fr/html/). Action-3 corresponds to the use of the search engine built on top of a protocol called Search/Retrieve via URL (SRU). SRU (see Morgan (2004)) follows a standard URL format and contains the following substring: /services/engine/search/sru?operation=searchRetrieve. For Action-4, downloading a document is identified by a AJAX request and is visible in the log with a HTTP request starting with http://gallica.bnf.fr/services/ajax/action/download. The consultation of a document (Action-5) is performed when a user requests a document from Gallica through an Archival Resource Key. Archival Resource Key (ARK) describes in Peyrard et al. (2014) is a URL defining a persistent identifier for information objects of any type. Gallica uses the ARK to support long-term access to its digital content. It allows any document of the Gallica's catalogue to have a unique identifier.

In addition to ARK, Gallica uses a framework called OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) that allows to store and harvest metadata. Thus one can retrieve the metadata of each document consulted by a user namely the type (periodicals, monographs, images, photographies…), the authors and the date of publications.

2.3. Web session reconstruction

To perform, a pattern discovery analysis we first need to identify and isolate the requests made by each different users. The pre-processing that group all requests made by each unique user is called sessionization. More precisely a sessionization is defined here as the process of identifying a particular user session from Web data. Typically a sessionization is performed with the help of a user-id and a password (identification to a website) or with the help of HTTP cookies. A HTTP cookie is a unique identifier sent from a website and stored on the user's computer by the web browser. Each time a user browses the web site

again, his web browser send his HTTP cookie to the web server. A HTTP cookie can therefore be used to perform the sessionization step. However Gallica's web server does not send HTTP cookies and the logging capabilities of Gallica is not really popular among its users. In this case, heuristic methods based on the IP addresses have to be used to build sessions out of access logs. In this subsection we describe the chosen method for this study. Following Spiliopoulou et al. (2003) a session is defined as a sequence of requests made by an unique IP. However a new session is defined whenever an IP stays inactive for a period of 60 minutes. The inactivity period is usually fixed at 30 minutes (see Spiliopoulou et al. (2003)), but based on the qualitative interviews we prefer a higher value of 60 minutes due to the possible long reading process of a document (seen as inactivity in the logs). The session representation used in the next sections are due to Gündüz and Özsu (2003) that proposes to represent a session as a sequence of requests associated with their timestamps.

More precisely, the s^{th} session is represented as a vector $x_s = ((a_{0,s}, t_{0,s}), \dots, (a_{n-1,s}, t_{n_s-1,s}))$where $n_s$the number of actions is, and $t_{n,s}$ is the timestamp of the n^{th}action for the s^{th}session.

All the log files are stored in an Elasticsearch database to be easily queried for the furthers analysis. In addition, the processings are performed with the help of the cloud research platform TeraLab[1] and Python.

3. Model & Results

This section aims to formally expose our model. Our goal is to use the observations (sequence of actions with their timestamps) and grouped them into K different classes of usage. We then use the obtained classes and interpret them in term of users' behaviors. First we need to redefine a session in order to take into account the duration of the actions. If the n^{th} and $(n+1)^{th}$ actions $a_{n,s}$ and $a_{n+1,s}$ differ, action $a_{n,s}$ is assigned the duration $t_{n+1,s} - t_{n,s}$. Otherwise action $a_{n+1,s}$ is discarded and only the first next action that differs from $a_{n,s}$ is kept to compute its duration. This new representation of a session cannot be considered as a realization of a simple Markov chain but of Markov process (see Stroock (2014)). Markov processes in finite state spaces constitute the continuous-time analogue of the Markov chains. Let us note that the number of actions of each sessions n_s may vary a lot from one session to another. To take this into account, we add a $i = 6^{th}$action corresponding to the end of the session by setting $a_{n_s,s} = 6$and $t_{n,s} = t_{n_s-1}$ for all s. Thus each session is modeled as a Markov process defined on the state space $\{1, \dots, 6\}$with a transition rate matrix $Q^{(k_s)}$and initial probability $\pi^{(k_s)} = (\pi_i{}^{(k_s)})_{i\in\{1,\dots,6\}}$ only depending on the class index of the session $k_s \in \{1, \dots, K\}$. Since the session (containing a reasonable amount of actions) is only observed once these parameters could not be inferred

[1] https://www.teralab-datascience.fr/en/home

precisely for each session. Mixture models supposed that sessions can be grouped in classes (also called clusters) sharing the same model parameters. The choice of the number of classes,K, is a in this case a compromise between the number of distinct behaviors wished to be discovered and the difficulty to deal with this variety. Indeed a high K may result both in high variance estimators and in difficulties to interpret the numerous classes. To our knowledge, this model is new and does not correspond to previous works on variants of Markov Processes. The model is parametrized by a set of transition rate matrices $Q^{(k_s)}(k_s \in \{1,\dots,K\})$ whose non-diagonal entries (i,j) contains the rate of arrival of an action of type j given that the current state is i. By convention, the lines of the transition rate matrix sum to 0, hence the diagonal contains the negated value of the overall rate of arrival of a new action. As a consequence, given that action $a_{n,s} = i \in \{1,\dots,6\}$ arrived at time $t_{n,s}$, the delay before the next action arrives follows an exponential distribution with intensity $-Q_{i,j}^{(k_s)}/Q_{i,i}^{(k_s)}$, and the probability for this action to be of type $j \neq i$is $-Q_{i,j}^{(k_s)}/Q_{i,i}^{(k_s)}$.

The inference for mixture models is classically performed with the well-known EM-algorithm (see Dempster et al. (1977)). Let us detail the E and M steps in this context. Recall that, for all $k \in \{1,\dots,K\}$, we denote by $\alpha_k$the *a priori* probabilities to the classes and, for all $i \in \{1,\dots,6\}$, we denote by $\pi_i^{(k)}$ the probability to start the session with action i given the class index is k. We further denote by $\delta_{n,s} = t_{n,s} - t_{n-1,s}$ the delay between the n^{th} action and the previous one and by Θ the complete set of parameters $Q^{(k)}, \pi^{(k)}, \alpha_k$; with $k = 1,\dots,K$.
Then, the complete log-likelihood $l(\Theta)$is obtained by summing

$$l_s^{\Theta}(k_s) := ln(\pi_{a_{0,s}}^{(k_s)}) + \sum_{n=1}^{n_s} \left(ln(Q_{a_{n-1,s},a_{n,s}}^{(k_s)}) + ln(Q_{a_{n-1,s},a_{n-1,s}}^{(k_s)} \delta_{n,s}) \right) + ln(\alpha_{k_s})$$

over all observed sessions indices s.

For any parameter $\Theta' = (Q^{'(k)}, \pi^{'(k)}, \alpha_{a_{0,s}}^{'(k)})_{k=1,\dots,K}$, the posterior probability for session s belong to cluster k given the actions and delays of that session is then given by the fact that and $\sum_k \beta_{k,s} = 1$and

$$\beta_{k,s}(\Theta') \propto_k \pi_{a_{0,s}}^{'(k_s)} + \prod_{n=1}^{n_s} \left(Q_{a_{n-1,s},a_{n,s}}^{'(k)} exp(Q_{a_{n-1,s},a_{n-1,s}}^{'(k)} \delta_{n,s}) \right) \alpha_k',$$

where $\propto_k$ means equal to up to a multiplicative factor not depending on

k. Step E of the EM algorithm is then obtained by taking the expectation of $l(\theta)$given the actions and delays of all sessions under parameter θ'. This reads as:

$$H(\theta,\theta') = \sum_s \quad \sum_k \quad l_s^{\theta}(k)\beta_{k,s}(\theta').$$

Next the M step, that is, maximizing $H(\theta,\theta')$ with respect to θ', under the usual constraints on the parameters allow us to obtain the update equations for each parameters.

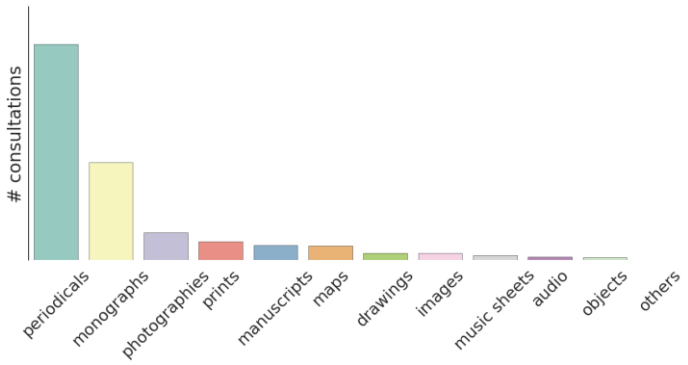


Fig 1. Histogram of type of consulted documents by Gallica's users.

We illustrate the model previously exposed with sessions extracted from web access log for a period of 1 month from 15 May 2016 to 15 June 2016.
First, Fig 1. shows the ranking of the type of consulted documents by Gallica's users. The most consulted type of document is *periodicals*, *monographs*, and *photographies*. Let us note that the *periodicals* and monographs count for respectively ~60% and ~14% of all available unique documents available on Gallica.

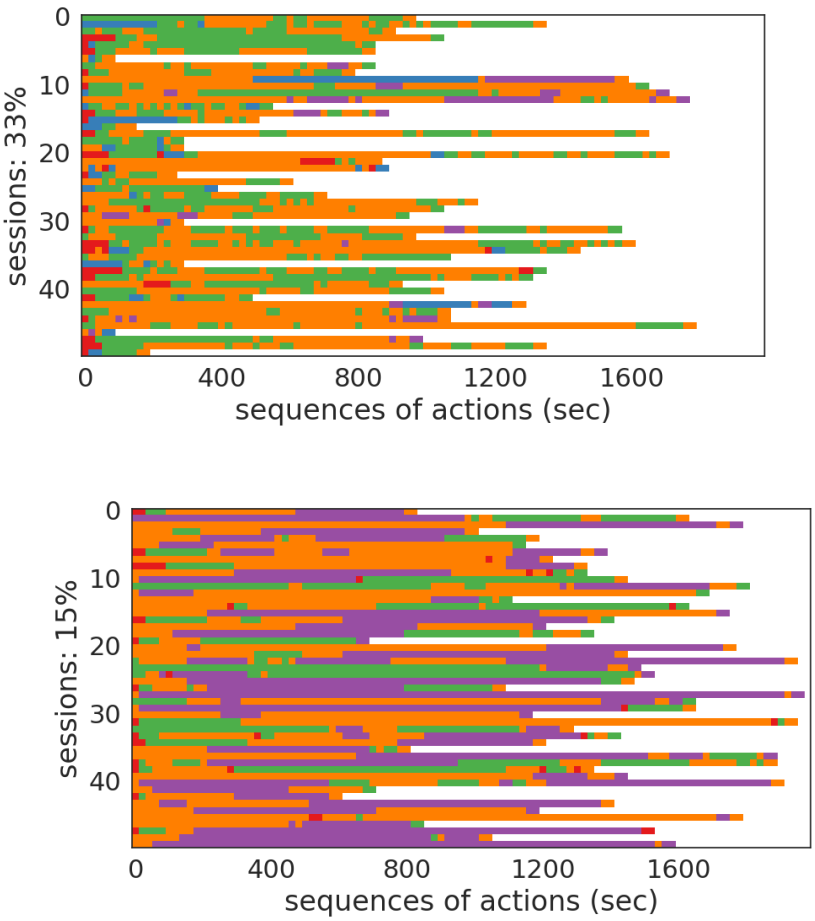
Fig 2. are vertical concatenation of 50 randomly chosen "typical" sequence for each of the 10 clusters. "Typical" sequences correspond to sequences with high posteriors (>.90), in other terms, sequences that have a high probability of belonging to a specific cluster. For each subfigures the number given in y-axis is the percentage of sessions for each clusters.

We recall here the set of action with their corresponding color:

- Action-1 Homepage :
- Action-2 Mediation :
- Action-3 Search :

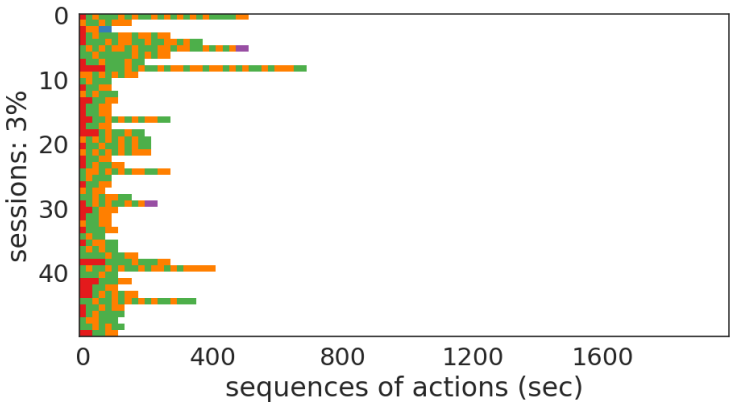
- Action-4 Download :
- Action-5 Browsing

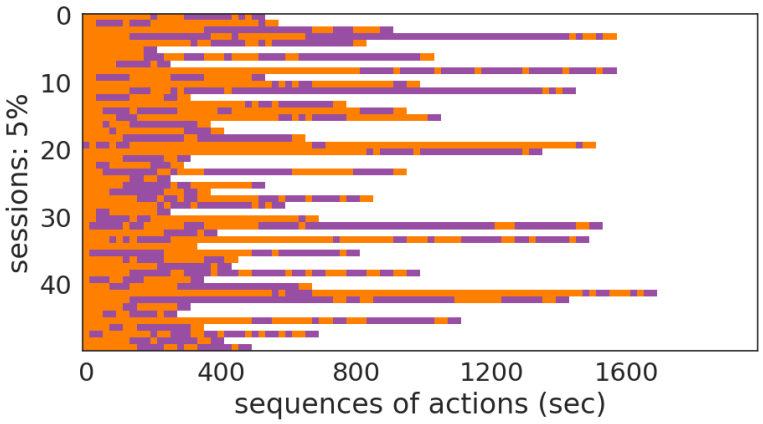
The representation given in Fig 2. makes it possible to distinguish and quantify well-differentiated behaviors. For example sessions belonging to the clusters represented in Fig 2.(a) and (b) almost only consult (Action-5) or use the built in search engine (Action-3). This profile of users corresponds to around 18% of Gallica's users. The behavior represented by Fig 2.(g) (around 6%) distinguishes itself with a high time spending on searching. Cluster (d) (3%) and (e) (1%) are short sessions. Cluster given in Fig 2.(c) and (f) are hard to interpret and does not exhibit a clear homogeneity between "typical" sessions. It is due to the chosen number of estimated cluster (K=10) that is too low and multiple behaviors are grouped together in one cluster. Finally the use of Action-4 (Downloading) distinguishes the remaining clusters, (h), (i), (j).

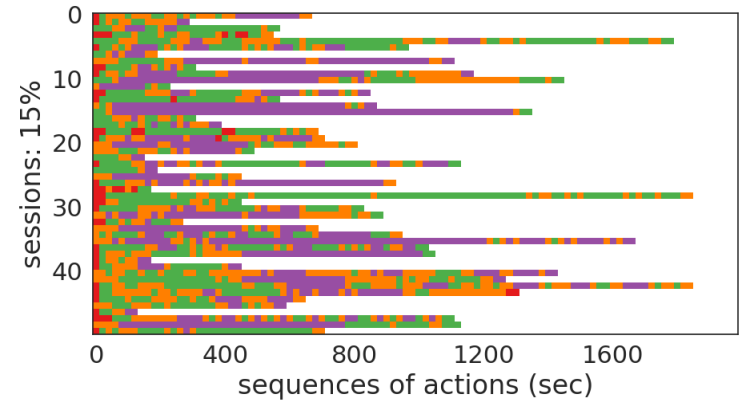


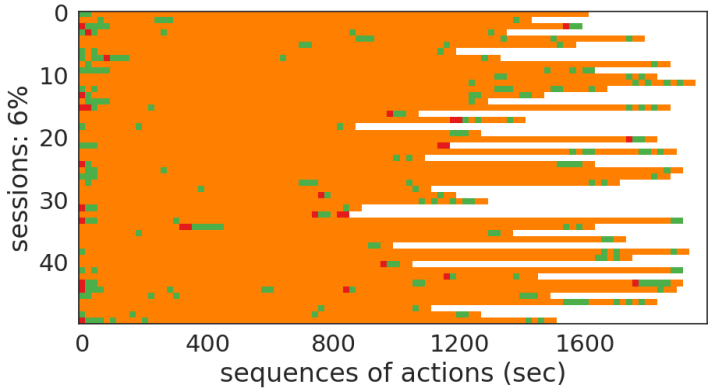
(a) (b) (c)

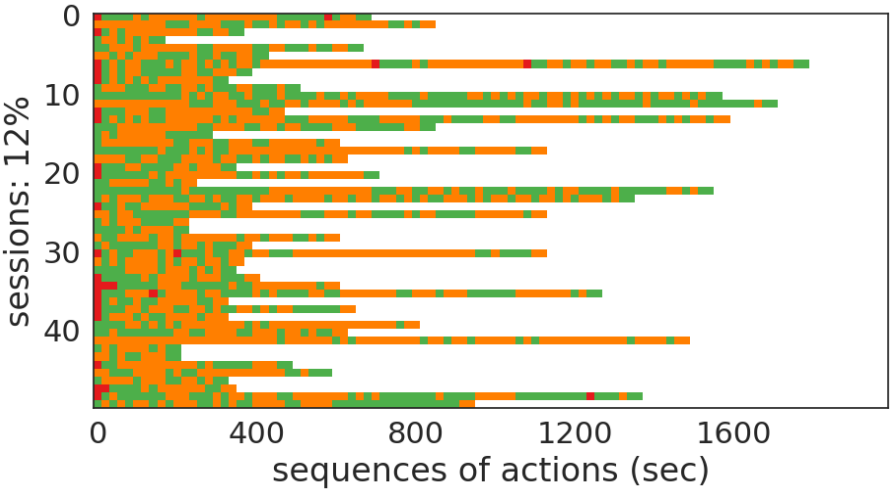
(d) (e) (f)


0
10
20
30
40
sessions: 3%
0
400
800
1200
1600
sequences of actions (sec)


0
10
20
30
40
sessions: 5%
0
400
800
1200
1600
sequences of actions (sec)


sessions: 15%
0
10
20
30
40
0
400
800
1200
1600
sequences of actions (sec)


sessions: 6%
0
10
20
30
40
0
400
800
1200
1600
sequences of actions (sec)


sessions: 12%
0
10
20
30
40
0
400
800
1200
1600
sequences of actions (sec)

(j)

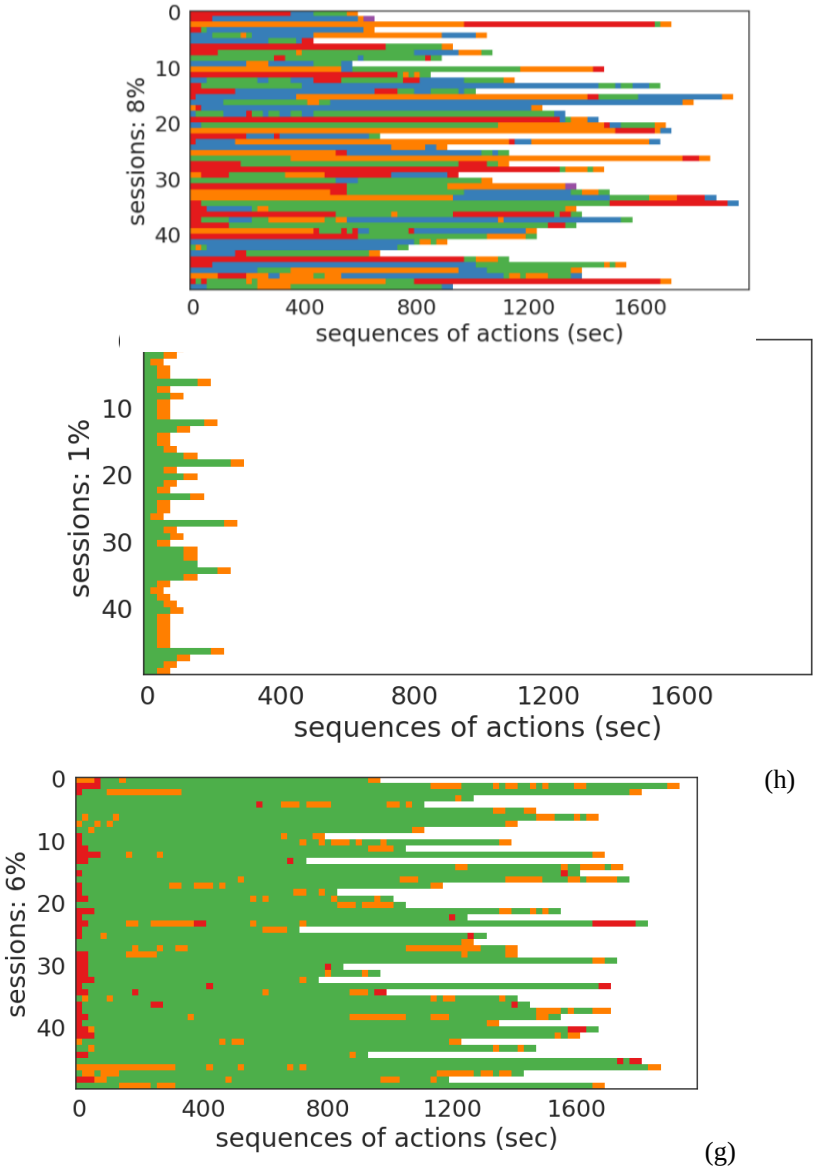


Fig 2. "Typical" sequences of actions for 10 clusters estimated with Markov processes.

4. Conclusions

In this project, the exploration of data benefits from a double contribution. On the one hand, it is conducted in close relationship with the BnF which brings its expertise and expresses its questions. Indeed, the analysis of the logs is intended to guide decisions on the evolution of the interface, in order to improve the user experience. On the other hand, the project benefits from the work carried out in another project of the Bibli-Lab based on surveys with users of Gallica (qualitative interviews, online survey and video-ethnography). The log-mining of allowed us to get an insight of users' behaviors. Even though we have of course no information on what the user actually does between two consecutive actions, the introduction of the temporal dimension is expected to provide a real improvement for analyzing the use of the digital library. For example, we found out that the users mostly arrived on Gallica directly on a document and not through the homepage. Thus, it shows the importance of a good SEO (Search engine optimization) that can bring a significant number of newcomers that are not familiar with Gallica. Moreover, some sessions end with a "Search" action that could be due to the frustration of not finding the desired document through the built in search engine.

Our approach has a few limitations we plan to address in the near future. Indeed, the limitations of IPs for the definition of a session (IPs changing for a day to another) have convinced the BnF to install a cookie that is more reliable to study usage through multiple days. The study of logs also exhibits a significant amount of Google within sessions as a HTTP referrer implying that some users tend to use Google as a search engine instead of the built in search engine of Gallica. This observation motivates the addition of a new action called "Google Search" to study the use of Google within sessions. We are aware that estimated duration of each action does not account for various events that may keep the reader away from Gallica and thus be falsely interpreted (coffee break, browsing another website…). Finally we plan to incorporate additional information in the clustering analysis such as the type the browsed content that will lead to a combination of time-series modeling and natural language processing.

References

Beaudoin V., Garron I. and Rollet, N., (2016) « Je pars d'un sujet, je rebondis sur un autre » Pratiques et usages des publics de Gallica. Technical Report. BnF – Labex Obvil – Telecom ParisTech, Paris

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S., (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, *7*(4), 399-424.

Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Gündüz, Ş., and Özsu, M. T., (2003). A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-540). ACM.
Jamali, H. R., Nicholas, D., & Huntington, P., (2005). The use and users of scholarly e-journals: a review of log analysis studies. In Aslib Proceedings (Vol. 57, No. 6, pp. 554-571).

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., (2001). Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management* (pp. 9-15). ACM
Stroock, D. W., (2014). *An introduction to Markov processes*. Vol. 230. Springer Science & Business Media

Morgan, E. L., (2004). An introduction to the Search/Retrieve URL service (SRU). *Ariadne*, (40).

Peyrard, S., Kunze, J. A., and Tramoni, J. P., (2014). The ARK identifier scheme: lessons learnt at the BnF and questions yet unanswered. In *International Conference on Dublin Core and Metadata Applications* (pp. 83-94).

Silva, S. S., Silva, R. M., Pinto, R. C., & Salles, R. M., (2013). Botnets: A survey. *Computer Networks*, *57*(2), 378-403.

Spiliopoulou, M., Mobasher, B., Berendt, B., and Nakagawa, M., (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs journal on computing*, *15*(2), 171-190.

Tan, P. N., and Kumar, V., (2004). Discovery of web robot sessions based on their navigational patterns. In *Intelligent Technologies for Information Analysis* (pp. 193-222). Springer Berlin Heidelberg.

Tenopir C., (2003). Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies. Washington, D.C.: Council on Library and Information Resources.

Ypma, A., and Heskes, T., (2002). Automatic categorization of web pages and user clustering with mixtures of hidden markov models. In *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles* (pp. 35-49). Springer Berlin Heidelberg.

.