Discovery of usage patterns in digital library web logs using Markov modeling

Adrien Nouvellet, Valérie Beaudouin, Florence d'Alché-Buc, Christophe Prieur, François Roueff

Discovery of usage patterns in digital library web logs using Markov modeling

ADRIEN NOUVELLET, LTCI, Télécom ParisTech
VALÉRIE BEAUDOUIN, i3-SES, CNRS, Télécom ParisTech
FLORENCE D'ALCHÉ-BUC, LTCI, Télécom ParisTech
CHRISTOPHE PRIEUR, i3-SES, CNRS, Télécom ParisTech
FRANÇOIS ROUEFF, LTCI, Télécom ParisTech

This paper proposes a family of tools based on Markov modeling to quantitatively analyze how people access the digital collections of the Bibliothèque nationale de France (BnF, the national library of France), through the web platform called Gallica. The aim is to provide the BnF with relevant information about the various usage patterns to help them to better understand their users, improve the mediation efforts and the design of the website, in order to increase the general public use of the 4M-documents collection. For that purpose, the study focuses on the access logs retrieved from the Apache HTTP servers of Gallica that are converted into sequences of actions.

In order to study user navigation behaviors, we propose to model the access log data using Markov Models, whether it be Markov chains when considering sequences of actions without duration, or Markov processes when taking into account duration. Our models are either used to capture an average behavior through meaningful statistics or to cluster the data to exhibit various interpretable types of usage. The numerical results bring new insights on the way the users interact with the platform, 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 information in order to properly interpret the models and the correlations that it highlights, it is able to discover all types of behaviors, including the stealthiest and the most difficult to capture in traditional surveys, giving them their fair weight in terms of audience. We also show how this approach fits into a broader work combining data mining and ethnography.

Additional Key Words and Phrases: digital library, digital uses, log analysis, Markov modeling

ACM Reference Format:
Adrien Nouvellet, Valérie Beaudouin, Florence d'Alché-Buc, Christophe Prieur, and François Roueff. 2018. Discovery of usage patterns in digital library web logs using Markov modeling. 1, 1 (January 2018), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The National Library of France (Bibliothèque nationale de France — BnF) has a long tradition of usage studies on library users, in order to understand their behaviors and preferences and to improve the range of services. Research studies in the reading rooms have been performed for many years, constructing an in-depth representation of readers through quantitative and qualitative surveys. The digital revolution has deeply changed the identity of the library. Since the mid 2000's, a huge program of collection digitalization has been launched. In 2016, the digital

Authors' addresses: Adrien Nouvellet, LTCI, Télécom ParisTech, Université Paris-Saclay, adrien.nouvellet@telecom-paristech.fr; Valérie Beaudouin, i3-SES, CNRS, Télécom ParisTech, valerie.beaudouin@telecom-paristech.fr; Florence d'Alché-Buc, LTCI, Télécom ParisTech, Université Paris-Saclay, florence.dalche@telecom-paristech.fr; Christophe Prieur, i3-SES, CNRS, Télécom ParisTech, christophe.prieur@telecom-paristech.fr; François Roueff, LTCI, Télécom ParisTech, Université Paris-Saclay, francois.roueff@telecom-paristech.fr.

XXXX-XXXX/2018/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

library, Gallica, made more than 4 million documents available online, accessible to everyone. The audience of the library has dramatically changed: online readers compose a much wider audience than users in the reading room. Gallica has $\sim 40,000$ users per day. For analyzing digital use, traditional methods such as interviews and surveys are still useful but new methodologies are needed. To reinforce its academic position on digital use, the BnF has teamed up with Télécom ParisTech, an engineering school and a research center specializing in digital technologies, to found the Bibli-Lab, a laboratory for the study of the uses of digital libraries and of their users.

Many studies have been conducted in this domain. A review of literature identified 200 studies published between 1995 and 2003 [27]. Among the variety of research methodologies used in these papers (ethnographic approaches, surveys, log analysis, etc.), mining web logs that trace the activity of users on the web site (fixed and mobile) offers an interesting vision of library use. Jamali and al. reviewed several papers based on transaction log analysis for studying the use and users of digital libraries [13] to identify the advantages and limits of the method. On the benefits side, all the authors point to what this automatic and non-intrusive approach brings forward: access to the reality of use (including the discrepancies between reported and actual uses), longitudinal studies, performance evaluation of the system, comparison of behavior between different groups. On the negative side, they cite the following points: difficulty in differentiating user performance from system performance, problems of user identification (IP address, session ...), impossibility to assess the motivations and goals of the user. The authors recommend using raw logs rather than logs from proprietary software. Nicholas et al. [20] propose an example of log mining, by extracting some features from the logs (List of issues, Table of Contents, abstract, full text). We take a similar approach defining different sets of features. Mining logs offers an exhaustive vision of all use: the functionalities used, the types of documents consulted, and the sequence of these actions. All these elements are impossible to obtain through interviews or surveys. Moreover, online surveys, which are among the tools used by the BnF to gain insight on the use, tend to be answered by regular users close to the institution and miss newcomers or occasional 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 is twofold : first, web logs are combined with the description of documents to get a better view of the types of documents that are consulted; second, the modeling of sequences of actions is used 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. We first instantiate this approach by considering a mixture of Markov Chains or Hidden Markov Models, a well known approach already successfully applied to log data (see [3, 9, 28]). 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 [17, 18]. In the previously cited papers, sequences are considered time-homogeneous and do not take into account the time spent on each web page. We thus present a second and novel instantiation of the mixture of Markov models, that consists of an extension of [3] for non-homogeneous sequences, i.e. taking into 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.

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[1]. 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

[1]A scientific and technical committee, led by Philippe Chevallier and Emmanuelle Bermès from BnF, met regularly throughout the project

surveys completed by users of Gallica (qualitative interviews, online surveys and videoethnography)[2]. Some of the hypotheses which emerged from the field could be tested on the log data. It is a data-mining process informed by the needs of the institution and by the knowledge acquired on ethnographically observed use.

This paper is organized in 5 sections. Section 2 reports the four consecutive and necessary steps for data-mining from web access logs: 1) raw data gathering and pre-processing for filtering and formatting the log entries; 2) identification of a finite number of actions and pairing with the corresponding HTTP requests; 3) session reconstruction and 4) data enrichment with external document description. Section 3 presents an overview on Gallica's sessions which first shows their high diversity in terms of durations, types of documents and types of actions. Section 4 formally exposes the clustering algorithms that aim at distinguishing Gallica's users browsing patterns and also shows the results of the algorithms applied on 1 month of data. Finally, Section 5 discusses how the results of the analysis provide useful insights for the evolution of Gallica, and suggests possible improvements of the method in the future.

2 DATA

2.1 Log Files

The Apache HTTP Server provides very flexible and comprehensive logging capabilities. Each time a user browses a website, the server access log records all requests processed by the servers. All user queries are then saved as log files which consist of a set of lines corresponding to all queries and in the case of Gallica, we have the following information for each request:

- IP address encrypted with the help of a hash function that preserves the uniqueness of the IPs.
- the geographical origin of the IP, retrieved based on GeoIP databases, ahead of the anonymization step mentioned above.
- the date, 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 referrer, i.e. the address of the web page that linked to the resource being requested.

To perform a large-scale analysis over the access logs, two essential steps have to be considered. First the practitioners need to parse the log file, meaning splitting each line into the distinguished fields listed above. Once the log files are parsed, one must 'clean' the data, which means first remove the failed requests and the 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 higher than 299 and strictly than 200. Web robots, also called spider bots, are software that automatically scan a website to extract its contents. For example, search engines such as Google periodically explore the web in order to update their search indexes. In most cases, web robots are identifiable with their user-agent and can thus be filtered out [8, 23]. However, it may be difficult to identify some some web robots. Eliminating them is out of the scope of this paper but we refer to [26] for more insights into the topic.

2.2 Identification of actions

Data logs record types of actions that are more or less visible and informative about the user. 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 [2]. A series of interviews with Gallica users was conducted. One of the objectives was to understand how the users conduct their research. A major distinction in usage has appeared

[2]See [21], submitted at the current JOCCH special issue

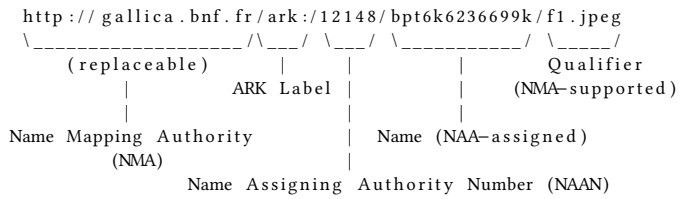

Fig. 1. Scheme of an ARK.

between people who download the documents, and reconstitute personal libraries, and other people who consult the documents online, zooming or turning the pages of a document with the browsing tool provided on the Gallica website. For the latter, a few pages are generally consulted to evaluate the relevance of the document before downloading it or giving up. The same study showed that users do not always follow the navigational path prescribed by the library : starting at the homepage, searching (through the built-in search engine) and consulting or reading documents. Some users access a document directly using a general search engine outside the interface such as Google. Finally, it seemed useful to see whether the important mediation efforts made by the BnF to facilitate the discovery of the collections have an impact on the use. The idea was to explore these assumptions on the data set in order to model and quantify the behaviors. We therefore extracted all the information from the logs that corresponds to these five actions, which reflect the main steps in terms of use :

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 one of the actions. First the homepage is requested by a user, performing **Action-1**, when a log line contains precisely "http://gallica.bnf.fr" as a HTTP request.

The BnF mediation refers to the blog ("http://gallica.bnf.fr/blog") or the collection pages (URLs start with "http://gallica.bnf.fr/html/"). The previous two URLs (Uniform Resource Locator) format allow us to identify the **Action-2**.

The Gallica search engine is built on top of a protocol called Search/Retrieve via URL (SRU). SRU [19] is a standard search protocol for Internet search queries that follows a standardized URL format. Thus **Action-3** can be identified with an URL starting with "http://gallica.bnf.fr/services/engine/search/sru?operation=searchRetrieve".

The **Action-5** is performed when a user requests a document from Gallica through an Archival Resource Key. Archival Resource Key (ARK) [15, 16] 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. As shown in Fig. 1, an ARK first contains the protocol ("http://") followed by the name of the service that provides support for that ARK ("gallica.bnf.fr") also called "Name Mapping Authority" (NMA). Then the previous sequence of characters is preceded by the label "ark:" and an immutable and globally unique identifier that included a "Name Assigning Authority Number" (NAAN) identifying the naming organization and the name that it assigns to the object. An optional "Qualifier" specifies sub-actions such as the requested page of the document, the specification of the layout of the page, a zooming action.

Finally the downloading **Action-4** is made by a user whenever an Ajax script is requested, hence the **Action-4** is identified with a request of type "http://gallica.bnf.fr/services/ajax/action/download" followed by the ARK identifier.

2.3 Web session reconstruction

An important pre-processing step related to the pattern discovery requires identifying and isolating the different users. As explained above, the log files are an unordered sequence of requests. In order to group all requests made by each unique user, a *sessionization* step is necessary. A *sessionization* is defined here as the process of identifying a particular user session from Web data. Sessions can be determined when users log themselves in the application by means of a user-id and a password, or by means of HTTP cookies. An 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 sends his HTTP cookie to the web server. An 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 have to be performed to build sessions out of access logs. In this subsection we describe the chosen method for this study. Following [24] a session is a sequence of requests made by a unique IP. A new session is defined when an IP is inactive for a period of 60 minutes. The inactivity period is often fixed at 30 minutes [4, 7, 14, 24] based on experimental observations. However, in line with recent findings on this specific issue of inactivity period [12], we prefer a higher value of 60 minutes due to the possible long reading process of a document (that is seen as inactivity in the logs). Alternative stochastic sessionization methods have been proposed by [5, 22] but are out of the scope of this paper.

Let us notice that the chosen heuristic for sessionization has the following known limits:

- Users sharing an internet access (through intermediate proxy server) also share the IP address and thus are being consider as a unique user.
- The public IP address that get assigned to the routers of most home and business users by their ISP is a dynamic address. ISPs have a limited range of IP address they could assign to their consumers meaning a specific user can have different public IP addresses for different days. This limitation implies that IP addresses could not be used to track users over days.
- Mobile users that switch between 3G/4G and WiFi exhibit a change in the IP address and thus are considered as multiple sessions.

The session representation used in the next sections are due to [11] 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 $\mathbf{x}_s = \left[(a_{0,s}, t_{0,s}) \quad \dots \quad (a_{n_s-1,s}, t_{n_s-1,s})\right]$ where n_s is the number of actions, $t_{n,s}$ is the timestamp of the n^{th} action for the s^{th} session.

All the log files are stored into an Elasticsearch database to be easily queried for further analysis. In addition, the processing is performed with the help of the cloud research platform TeraLab[3] and Python (Code available with the final submission on GitHub).

2.4 OAI : add descriptive metadata to logs

During the digitization process, the BnF makes an effort to attach metadata to each document. The descriptive metadata, encoded in Dublin Core format, are freely accessible through OAI repositories [10]. It is thus possible to harvest all the metadata and associate the ARKs contained in the logs with the following relevant information:

- title,
- author,
- date of publication,
- source,
- language,

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

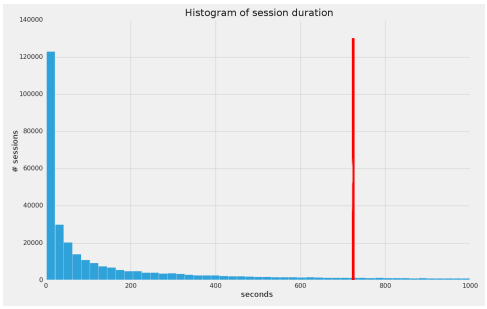

Fig. 2. Distribution of session durations for 1 day (17 May 2016)

- type (booklet, monography, photography, ...),
- theme* (law, medical science, history of France, biography, ...),
- corpus* (1914-1918, Rhône-Alpes, travel in Italy, ...),
- short description*.

The annotated fields are optional. The access to additional information paves the way for performing more in-depth analysis.

3 OVERVIEW OF DIGITAL LIBRARY SESSIONS

This section proposes a series of results describing the sessions: their duration, the types of documents consulted during the sessions, the sequences of actions within a session, and the variety of behaviors depending on the referrers (the websites from which the users come to Gallica).

3.1 Session lengths

We extracted the sessions from the web access log for a period of 1 month (15 May 2016 to 15 June 2016). Before considering the content of the sessions, we looked at their length in terms of time and number of actions. We have of course no indication of the time actually spent on each recorded url, but the timelapse between the first and last actions of the session is a significant proxy that we will call its *duration*. Defined in this way, the duration is thus zero when the session is reduced to a single action. The distribution of the duration of sessions is plotted in Fig. 2, showing, as is usually the case in this kind of data, that the majority of the sessions have a very short duration (less than 1 minutes): the median is of twelve seconds. The mean duration of the sessions, around 12 minutes (the red line on the figure), is exceptionally high, due to the impact of a few very long sessions.

Quite logically, the number of actions per session follows the same kind of highly heterogeneous distribution, with the following ratios:

(1) 1 action: $\sim$ 30% of the sessions,
(2) $2-4$ actions: $\sim$ 30% of the sessions,
(3) >4 actions: $\sim$ 40% of the sessions.

3.2 Variety of documents within the sessions

In Section 2.4, we have explained how logs were enriched with the metadata describing the library's documents consulted during the sessions. The digital library gathers different kinds of documents: booklets (press), monographs, photographs, manuscripts... Figure 3 plots the proportion of the main types of documents consulted within all the sessions. One can see that booklets (press) and monographs represent the majority of the consultation on

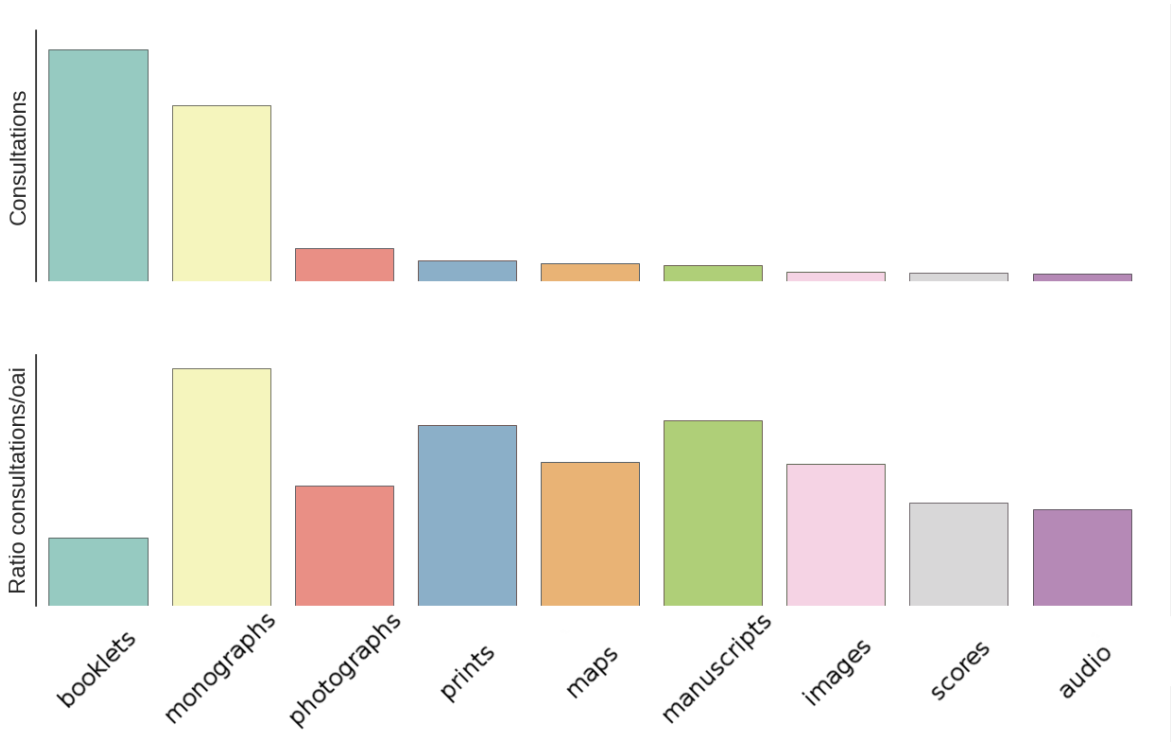

Fig. 3. Up: Proportion of consulted type of documents within all the sessions. Down: Ratio of the number of consulted type of documents within all the sessions and the number of type of documents within the OAI repositories.

Gallica. However, when compared to the available documents of each type in the digital library (the bottom part of the figure), the over-representation of press is much less visible, while other types emerge as highly popular, namely prints (engravings) and manuscripts, for instance.

Now we can see how the types of documents are combined among the sessions. In other words, does the digital library encourage crossing the walls that separate the different reading rooms and departments inside the physical library? To answer this question, we selected the sessions with at least 5 different documents consulted, which represent 8% of the sessions, and performed a clustering (with the K-means method) based on the types of documents consulted within each session. On Fig. 3, each line represents one of the 8 estimated clusters, with its size reported on the left. The heights of the colored rectangles correspond to the average frequences of the types of documents in the cluster. In 45% of the sessions (three clusters), a single type of document was consulted (booklets in 27% of the sessions, monographs 12%, prints/engravings in 6%). In 39% of the sessions, the user consulted two types of documents (booklets and monographs in 31%; monographs and manuscripts in 8%). Only three clusters (roughly $\sim$ 14% of the sessions) present a high diversity of type of documents. This clustering thus reveals a high level of homogeneity in the types of documents when consulting digital library.

3.3 Average pattern of successive actions

We now explore how actions sequentially occur within sessions, without considering the time spent between two consecutive actions. Hence, we here use the observations $\mathbf{y}_s = \begin{bmatrix} a_{0,s} & \dots & a_{n_s-1,s} \end{bmatrix}$, for all sessions s, extracted

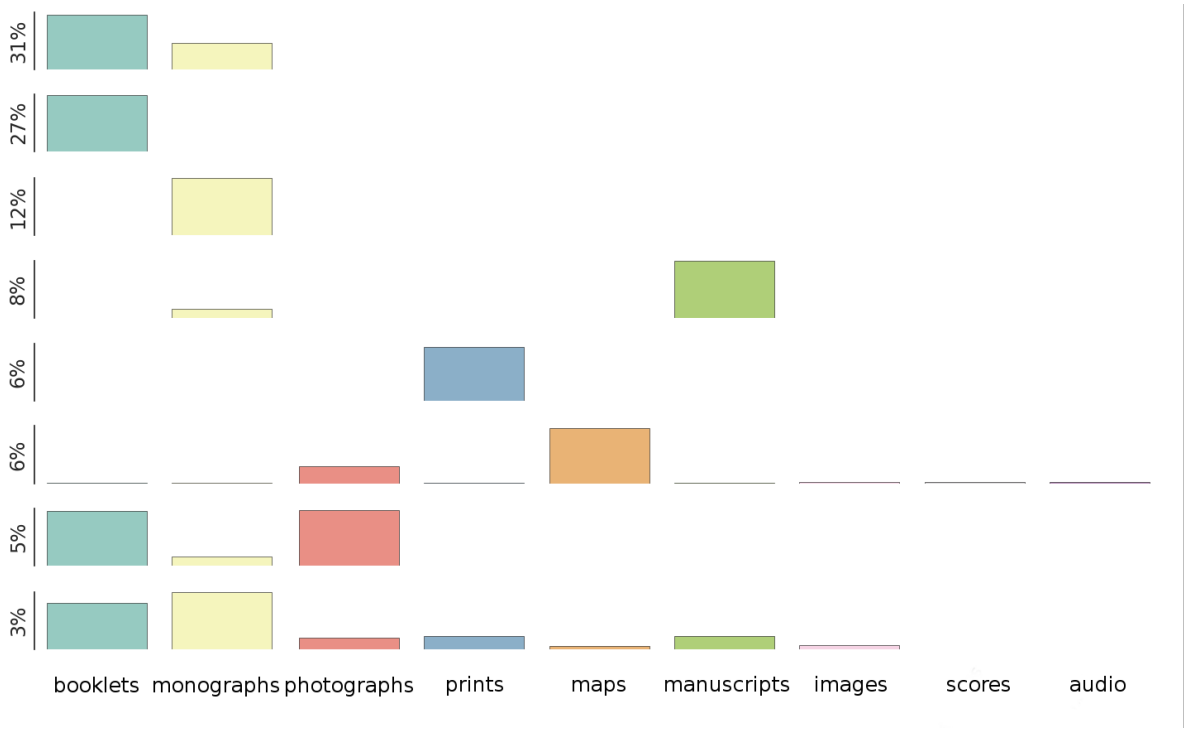

Fig. 4. K-means clustering of the types of documents consulted within sessions.

from the input data $\mathbf{x}_s$ described in Section 2.3. Our goal is to have a first snapshot providing a global view of the mean browsing behavior in terms of successive actions within a session.

As we have seen above in Section 3.1, the number of actions n_s varies a lot from one session to an other. To take this number into account in our model, we add a 6th action named *End* that closes each session, e.g. we set $a_{n_s,s} = 6$ for all s. Now each observed session, $\mathbf{y}_s$, is modeled as a Markov chain defined on the state space $\{1, \ldots, 6\}$ with a transition matrix $\mathbf{A}$ and an initial distribution $\boldsymbol{\pi}$. This means that each session starts with action i with probability π_i and that, within an ongoing session, the probability of taking action j given the knowledge of all previous actions only depends on the previous action, and takes value $A_{i,j}$ if this previous action is i.

We assume here that all sessions share the same transition matrix and initial distribution. This assumption is of course disputable, as one would expect each individual to behave differently. However a session alone is not sufficient to estimate these parameters and we thus need to look for *average behaviors*. Presently, we are interested in a global view, hence a behavior averaged over all observed sessions. An alternative and more precise approach is to cluster the data into groups of sessions that share the same behavior. This will be investigated in Section 4, using Markov processes that take into account the time intervals between successive actions.

In order to get an idea of the average navigational behavior on Gallica, we display the transition matrix empirically estimated from the whole dataset in Fig. 5. In the 6×6 image, the color of the square at row i and column j represents the value of $A_{i,j}$ for $i = 1, \ldots, 5$ and $j = 1, \ldots, 6$, or of π_i if $i = 6$. For instance the deeper red the pixel at coordinate $(1, 3)$ is, the closer the probability of a *Homepage* action to be followed by a *Search* action is to 1. All lines sum to 1. The last column represents the probabilities for the action of the corresponding row to finish the session while the last row represents the probabilities for the action of the corresponding column to

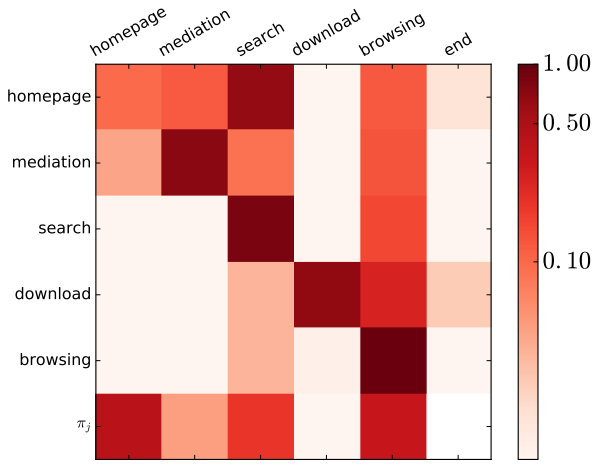

Fig. 5. Global transition matrix $\mathbf{A}$ and initial probabilities $\boldsymbol{\pi}$.

start the session. Note that the lower right square corresponding to the coordinate $(6, 6)$ is zero, as we kept only the sessions having at least 5 actions and thus an action cannot simultaneously be first and last.

This transition matrix reflects the global behavior of Gallica's users. While it could have been expected that all sessions start with the homepage, the matrix given in Fig. 5 shows that the probability of starting a session directly by consulting a document is just as high. In this case, direct access to the document is made outside of Gallica, for instance via Google, the catalog of the library[4], or social media[5]. When a user goes through the homepage, the most likely action that follows is the use of the search engine. Fig. 5 also exhibits strong diagonal values for the transition matrix telling us that users often perform the same action several times. It is of course not surprising due to the design of Gallica and the definition of the actions. For example, reading multiple pages of a document results in a series of *Browsing* actions.

3.4 Where do users come from?

The web logs record the browsing activity exclusively on Gallica. But it is interesting to know the origin of Gallica' visits: where do the users come from? The logs provide a partial answer to this question: the " referrer " field identifies the url of the web page where the user clicked to get to Gallica. This field can therefore enable the identification of refering sites that redirect part of their traffic to the gallica.bnf.fr domain.

To identify the origin of a visit on Gallica, we followed this 4-steps protocol:

- extraction of the field "referrer" in each log line,
- selection of the first "referrer" in each session (based on the timestamps),
- identification of the website referrer : the first "referrer" that does not contain gallica.bnf.fr
- removal of subdomains and extensions from the url (scholar.google.fr -> google)

The distribution of the referrer websites shows that in about 40% of the sessions, there is no referrer. If the user starts his/her web navigation by directly typing the address of Gallica, no referrer will be indicated. Some technical problems lead to an overestimation of the number of sessions. The most notable result is the critical role played by Google as referrer: nearly 40% of the sessions come from Google.

[4]http://catalogue.bnf.fr/

[5]https://twitter.com/gallicabnf and https://facebook.com/GallicaBnF{}

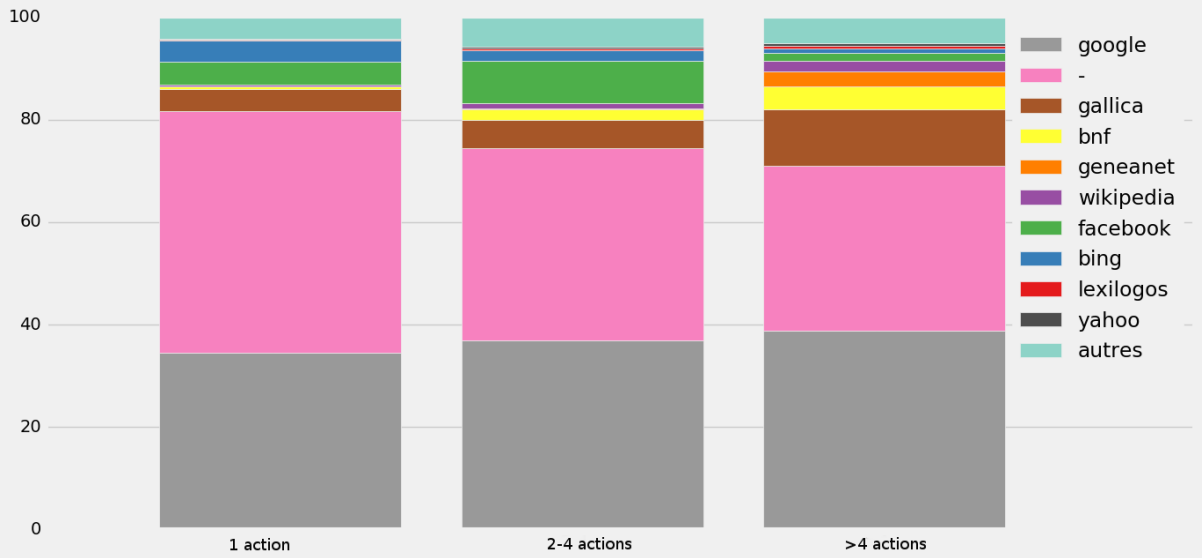

Fig. 6. referrers among short or longer sessions.

We compared the distribution of referrers according to the length of the sessions, more precisely according to the number of actions in the session (see figure 6). Facebook plays a more significant role in relatively short sessions (less than five actions but more than one). This indicates that a significant part of the sessions coming from the social network do not lead to long explorations of Gallica. As for sessions with a single action, which represent 30% of the sessions, it appears that in 80% of them, it is a *Browsing* action, so the user simply consults a document. In half of these cases, the user reached the document through Google. The importance of Facebook and Google in very short sessions raises the question of the retention of new users who reach Gallica and might find new resources making them want to explore it further.

Conversely, for long sessions the role of `bnf.fr` and its subdomains (`catalog.bnf.fr` or `data.bnf.fr`) is more important. We also note in the long sessions the significant presence of Geneanet, one of the main genealogy platforms, which reveals a particular and numerically important type of use of Gallica: genealogical research.

We also identified the presence of external referrers within the sessions (and not only on the actions that start the sessions). The qualitative survey showed that the use of the Google search engine is intertwined with the consultation of Gallica [21], and we wanted to quantify this phenomenon. After filtering the `gallica.bnf.fr` referrers, we obtained the following:

- 23% of sessions with at least 2 external referrers,
- 30% of sessions with more than 4 actions with at least 2 external referrers,
- 18% of sessions with more than 4 actions with `google` as external referrer for at least 2 of them.

This confirms the significant presence of Google in the workflow of some users.

4 DISCOVERING USAGE PATTERNS: CLUSTERING THE LOG DATA

In order to identify a variety of behaviors among the users of the digital library Gallica, we now use a clustering method to segment the log data in groups (or *clusters*) of sessions sharing the same statistical patterns. To achieve this, our observation data is the set $\mathbf{x}_s$, over all sessions s, introduced in Section 2.3 (each session being composed

of a sequence of timestamped actions, each action being one among the five defined in Section 2.2). Clustering such sequences provides a coarse-grained segmentation of the usage patterns, highlighting the nature of activity performed within a session, assumed to belong to a cluster of similarly behaving sessions.

We have opted for mixture models to provide a segmentation of the data at hand. This choice is motivated by the interpretability that a probabilistic graphical model offers and the versatility of the estimation algorithm, the well-known Expectation-Maximization algorithm. The statistical inference for such clustering models is classically performed using the celebrated EM-algorithm of [6]. Suppose that the sessions can be segmented into K classes. The choice of the number of classes K is a compromise between the wish to encompass the variety of all behaviors among the observed sessions and the difficulty of dealing with this variety both from a statistical inference point of view (yielding estimators with high variances) and from a practical point of view (making it difficult to interpret a continuum of numerous classes rather than a small number of well separated ones).

The idea behind the mixture model is to interpret the class index $k_s \in \{1, \ldots, K\}$ of the session s as a hidden variable with *a priori* probabilities $\alpha_1, \ldots, \alpha_K$ to belong in each class, corresponding to the proportions of the clusters within the overall population. The log-likelihood function of the complete model is then derived from the likelihood associated to each class but cannot be computed since the hidden class index of each session is unknown. Nevertheless, the E and M steps of the EM-algorithm can be used to provide successive approximations of the complete log-likelihood using its conditional expectation given the observations only. We refer to [1] for more details.

4.1 Clustering sessions using Markov processes

To complete the clustering method, we now need to come up with a convenient model for the clusters and derive the corresponding likelihood that will lead to the previously mentioned E and M steps.

We already proposed in Section 3.3 a Markov chain model to describe the successive occurrences of actions within sessions. We would embed this model into a mixture approach in order to cluster the sessions in several classes sharing similar initial distributions and transition matrices. However, a drawback of this model is that it does not take into account the time spent between two successive actions. Since this duration could convey valuable information about the level of expertise of a user in managing documents and requests, we prefer to use a more complete model to describe the user behavior within a session. Hence, in contrast to the modeling approach of Section 3.3, we now use all the observations available in the session data, namely $\mathbf{x}_s = \left[(a_{0,s}, t_{0,s}) \quad \ldots \quad (a_{n_s-1,s}, t_{n_s-1,s})\right]$ for all sessions s. Recall that n_s denotes the number of actions and that $a_{n,s}$ and $t_{n,s}$ are the action label and timestamp of the n^{th} action for the s^{th} session. 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 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 a Markov process (see [25]), valued in the state space of action labels. Markov processes in finite state spaces constitute the continuous-time analogue of the Markov chains used previously.

Since we now consider a mixture model, we assume that each cluster is assigned a Markov process model and thus, the parameter of the Markov model governing the session s has parameters only depending on the cluster k_s to which s belongs. To our knowledge, this model is new and does not correspond to previous works on variants of Markov Processes. The Markov process model of cluster k is parameterized by a so-called *transition rate matrix* $\mathbf{Q}^{(k)}$ whose non-diagonal entries (i, j) contain 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 (instead of 1 for the transition matrix $\mathbf{A}^{(k)}$ of a Markov chain), hence the diagonal contains the negated value of the overall rate of arrival of a new action. As a consequence, given that action $a_{n,s} = i \in \{1, \ldots, 5\}$ took place at time $t_{n,s}$, the delay before the occurrence of the next action follows an exponential distribution with intensity $-Q_{i,i}^{(k_s)}$, and the probability for

this action to be of type $j \neq i$ is $-Q_{i,j}^{(k_s)}/Q_{i,i}^{(k_s)}$. As in the Markov chain case, 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. Recall that, for all $k \in \{1, \ldots, K\}$, we denote by α_k the *a priori* probabilities to the classes and, for all $i \in \{1, \ldots, 6\}$, we denote by $\pi_i^{(k)}$ the probability to start the session with action i given that the class index is k. Hence, assuming a mixture of K clusters of Markov processes, the complete set of unknown parameters is $(\mathbf{Q}^{(k)}, \boldsymbol{\pi}^{(k)}, \alpha_k)$ with $k = 1, \ldots, K$ and these parameters have to be inferred from the observations $\mathbf{x}_s$ over all sessions s.

We now derive the EM-algorithm that we used to perform the statistical inference of the parameters. Let us denote by $\delta_{n,s} = t_{n,s} - t_{n-1,s}$ the delay between the n^{th} action and the previous one. Then, the complete log-likelihood $\ell(\boldsymbol{\theta})$ is obtained by summing

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

over all observed sessions indices s. For any parameter $\boldsymbol{\theta}' = (\mathbf{Q}'^{(k)}, \pi'^{(k)}, \alpha'_k)_{k=1,\ldots,K}$, the posterior probability for session s to belong to cluster k given the actions and delays of that session is then given by the fact that $\sum_k \beta_{k,s} = 1$ and

$$\beta_{k,s}(\boldsymbol{\theta}') \propto_k \pi'^{(k)}_{a_{0,s}} + \prod_{n=1}^{n_s} \left(Q'^{(k)}_{a_{n-1,s},a_{n,s}} \mathrm{e}^{Q'^{(k)}_{a_{n-1,s},a_{n-1,s}} \delta_{n,s}} \right) \alpha'_k \ ,$$

where $\propto_k$ means equal to up to a multiplicative factor not depending on k. Step E of the EM algorithm is obtained by taking the expectation of $\ell(\boldsymbol{\theta})$ given the actions and delays of all sessions under parameter $\boldsymbol{\theta}'$. This reads as

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

Next the M step, that is, maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ with respect to $\boldsymbol{\theta}$, under the usual constraints on the parameters, yields, for all $k = 1, \ldots, K$ and $i \neq j \in \{1, \ldots, 6\}$,

$$\begin{aligned}
\alpha_k &\propto_k \sum_s \beta_{k,s}(\boldsymbol{\theta}') \quad \text{with} \quad \sum_k \alpha_k = 1 \ , \\
\pi_i^{(k)} &\propto_i \sum_s \mathbb{1}_{\{a_{0,s}=i\}} \beta_{k,s}(\boldsymbol{\theta}') \quad \text{with} \quad \sum_i \pi_i^{(k)} = 1 \ , \\
Q_{i,i}^{(k)} &= -\frac{\sum_s \beta_{k,s}(\boldsymbol{\theta}')}{\sum_s \frac{T_s(i)}{\sum_{w \neq i} M_s(i,w)} \beta_{k,s}(\boldsymbol{\theta}')} \\
Q_{i,j}^{(k)} &= -Q_{i,i}^{(k)} \frac{\sum_s M_s(i,j) \beta_{k,s}(\boldsymbol{\theta}')}{\sum_{w \neq i} \sum_s M_s(i,w) \beta_{k,s}(\boldsymbol{\theta}')} \quad \text{with} \quad i \neq j
\end{aligned}$$

where $M_s(i,j) = \sum_{n=1}^{n_s} \mathbb{1}_{\{a_{n-1,s}=i, a_{n,s}=j\}}$ is the number of transitions from i to j in session s and $T_s(i) = \sum_{n=1}^{n_s} \mathbb{1}_{\{a_{n-1,s}=i\}} \delta_{n,s}$ is the time spent in state i in session s.

4.2 Usage patterns on one month of data

To illustrate the model given in Section 4.1, we have applied it to the sessions extracted from the web access log for a period of 1 month (15 May 2016 to 15 June 2016). We selected the sessions with at least 5 actions, since they reflect more complex behaviors that are not straightforward to analyze. As mentioned above (Section 3.1), they represent 40% of the sessions.

We have chosen a decomposition of the sessions into ten clusters ($K = 10$) to get a good balance between the precision of the model and its complexity. Figure 7 shows, for each of the ten estimated clusters, a sample of 80

randomly chosen sessions among those having a high posterior probability (> 0.90) to belong to the considered cluster. Each line represents a session, with colored pixels indicating 20-second actions (since the duration of the last action of a session is unknown, it is always represented as a single pixel, arbitrarily standing for 20 seconds). This visualization has proven very efficient to help interpret the clusters, both during the iterative process of designing the model, as well as when discussing the results with the Gallica team.

The clustering presented in Fig. 7 makes it possible to distinguish and quantify well-differentiated profiles. Some sessions are composed mainly of one type of action: Search, for Cluster (c), Browsing, for (f). Note that even though a long green line may stand for only one search action followed by a browsing action several minutes later, the sessions have at least five actions, so most of the sessions represented in Cluster (c) do have several search actions. In a similar way, cluster (j) gathers sessions mainly focused on the resources put forward by the editorial team (**Action-2**: *Mediation*), while Cluster (i) shows a very simple pattern: Homepage, short Search, then Browse and leave. Even though they represent together only 8% of the sessions, each one of these behaviors deserves a more thorough investigation, picking some sessions to examine the actual sequence of timestamped urls to try and figure out what the user's aim was, and overall, whether it was fulfilled. Long sequences of search, for instance, are of interest to understand whether they are successful or not, and what strategies the users implement to find the documents they are looking for. This question has been addressed in the ethnographic study [21] and the sessions in this cluster provide very useful data to push it forward. In the same way, short sessions as those shown in Cluster (i) (home, search, browse, stop) raise the question of the satisfaction of the users: were the sessions successful or not? And more broadly, retaining users so that they carry on browsing once they have found what they were looking for, is a question this cluster can help address.

The other clusters show sessions alternating *Search*, *Browsing* and *Download* in various proportions (and also *Mediation*for Clusters (a) and (e)): high proportion of *Download* in Clusters (d) and (b), and also of *Search* in the latter, while no *Download* at all in Cluster (h), where the sessions constantly alternate between *Search* and *Browsing*. These advanced uses of the platform also come as a significant complement to the ethnographic study [21].

These results also highlight the interest of the representation taking into account the duration of each action. Indeed, the temporality of each of the actions can vary significantly (see Tab. 1 for expected times). The user may stay a few seconds on the homepage or on the contrary deeply explore it. The user may spend a very short time on the results of a search, if he considers it irrelevant, or on the contrary may dwell on the results for a long time, as shown in the video-ethnographic analysis carried out in the other project [21]. 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.

A Principal Component Analysis (PCA, see Fig. 8) can be used to visualize the distances between clusters. The PCA is performed on the concatenation of the initial distributions (of the actions per session) and the vectorized transition rate matrices for the 10 clusters. The size of the circles representing each cluster in Fig. 8 is proportional to their *a priori* probability. We kept the first two principal components, accounting for more than 66% of the variance. Cluster (j), with an *a priori* probability of 0.3%, is distant from the others: users explore almost exclusively the resources proposed by Gallica (*Mediation*).

In order to explore more diverse behaviors, which is useful for the Gallica team, we have also used higher values of K, clustering sessions into 15, 25, then 35 groups. Those behaviors are not exposed in this paper but a comparison of a global PCA analysis is reported in Fig. 9. In contrast with Fig. 8 where the PCA is computed uniquely with the parameters of 10 clusters, the PCA reported in Fig. 9 is computed considering all the parameters of the $10 + 15 + 25 + 35 = 85$ clusters: the parameters obtained for each of the 4 different values of K are concatenated to perform one PCA on these 85 clusters, whose projection on the first two components is split into four pictures in Fig. 9. As the value of K grows, smaller classes emerge, which may help identify more specific behaviors.

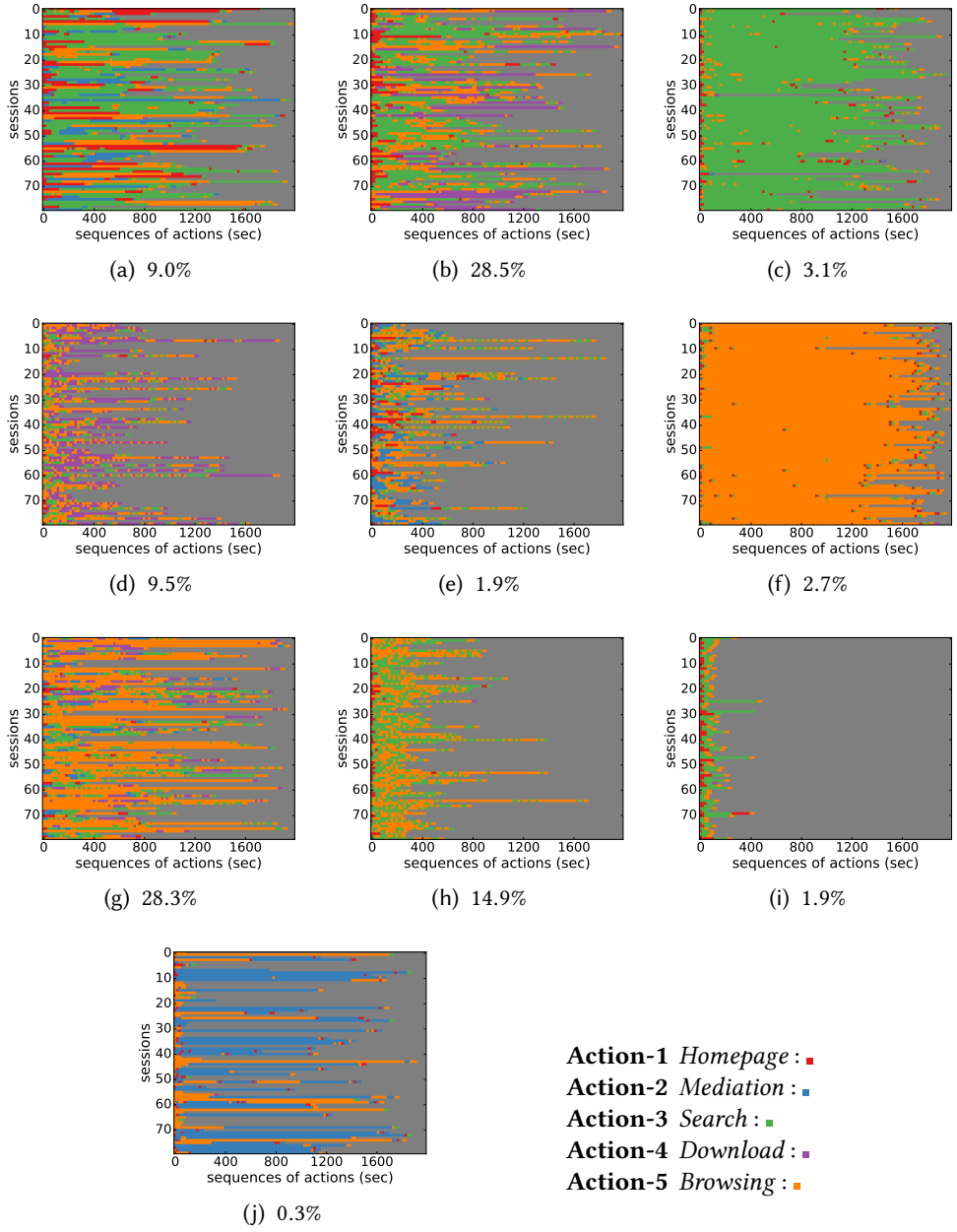

Fig. 7. "Typical" sequences of actions (in seconds) for 10 clusters estimated with Markov processes.

cluster	**Action-1** homepage	**Action-2** mediation	**Action-3** search	**Action-4** download	**Action-5** browse
(a)	194	119	393	0	786
(b)	39	0	138	192	97
(c)	15	0	611	0	23
(d)	08	0	18	65	26
(e)	38	68	12	0	41
(f)	10	0	35	0	1468
(g)	16	28	59	60	311
(h)	11	0	29	0	33
(i)	19	0	36	0	0
(j)	04	630	00	0	1757

Table 1. Expected time for an individual in action i to remain in that action for each cluster: $-1/Q_{i,i}^{(k)}$.

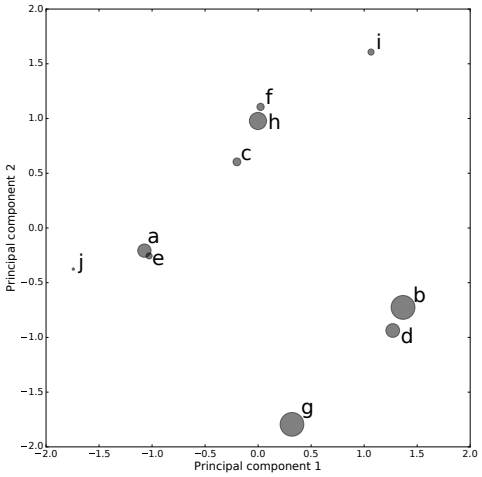

Fig. 8. Principal component analysis of the 10 estimated clusters.

5 DISCUSSION

Web access logs from the Gallica website contain a wealth of information that the BnF is now in a position to exploit. Computing the global transition matrix has put into evidence alternations in the types of actions with rates that were not expected. Then clustering sequences of actions has highlighted the diversity of typical navigational paths. None of these could have been observed over the surveys and audience measurements conducted so far. This work thus raised awareness of the amount of choices hidden behind usual audience metrics, and on the benefits of building a continuous dialogue between data-mining and decision making.

The clustering of action sequences has isolated simple patterns with mainly one type of action (*Search*, *Browsing*, or *Mediation*) or simple sequences (*Homepage-Search-Browsing*, or *Homepage* then alternation of *Search* and

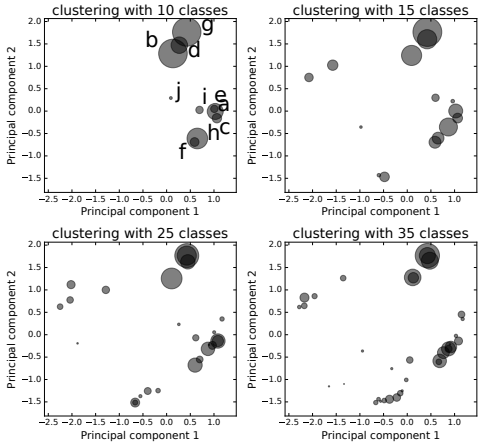

Fig. 9. Principal component analysis of clustering parameters for various values of K.

Browsing), and more complex sequences of *Search*, *Browsing* and *Download*, with various proportions. This suggests that the users of Gallica are much more diverse than researchers or students who come into reading rooms. Now each of these navigational patterns deserves a thorough investigation, which would consist of picking random sessions among a cluster, examining the sequence of urls consulted, the queries submitted to the search engine, the documents browsed, along with the time (all this information is present in the logs), and trying to figure out the purpose of the chosen user.

Some of these clusters will for instance provide insight for the mediation efforts to attract new types of users, or taking into account the diversity of uses in the design of the site. To this end, our results have successfully convinced the BnF to install a cookie in order to strengthen the analysis of Gallica users in a longer time frame instead of inferred sessions. Studying usage patterns might also enable design recommendations to increase browsing possibilities in order to retain occasional users who come from mediation pages, external search engines or social media, as was suggested in Section 3.4. Such findings contribute to ease the access to culture, which is one of the main missions of the BnF.

Now the data mining process itself could be extended in several directions. As mentioned in Section 3.4 devoted to referrers, the study of logs has shown a significant presence of Google referrers *within* sessions, implying that users tend to use Google as a search engine instead of the built-in search engine of Gallica. We could thus add a *Google Search* action in the modeling to enable it in the navigation patterns. Among other ways to exploit the referrer field, we could focus on specific referrers such as Google or Facebook, to further quantitatively analyze how the sessions coming from these websites differ from other sessions. Internal referrers are of interest as well: for instance when a user opens several urls from the same page to read them later, they all have that page as a referrer in the server logs. We could thus have a more precise indication of the actual sequence of pages in the user's attention flow by taking this information into account. Among the limitations of our approach, we are also aware that the estimated duration of an action does not account for various events that may keep the reader away from Gallica. It is also potentially interesting to take into account the time and the date when a session occurs and include this parameter into the models. Furthermore, all the clustering analyses can be enriched by incorporating information about the discipline, the type and the content of the documents accessed by the reader, calling for a combination of time series modeling and natural language processing.

Our overall objective is to understand, describe and quantify the many diverse ways in which the users of a digital library, access, search, navigate through, and gather material from, a large collection of digital resources. The work described here is a good first step in a collaborative and iterative process involving complementary and interdisciplinary approaches. Once clusters of action sequences have been computed, one needs to explore these clusters by "reading" sessions in them. This is an ethnographic work of a new kind, enabled by data mining: much less informed than interviews with the users (like a historian in front of historical documents), but with the possibility to observe many sessions with similar patterns. This work is to be continued with the teams at the BnF, as a follow up of the data mining process, which has itself been conducted in conjunction with the ethnographic study done with interviews and camera-aided observation [21]. Such an iterative process is, in our point of view, a necessary workflow to properly use a large amount of data to study social processes, keeping a focus on actual behaviors and motives of the studied actors.

ACKNOWLEDGMENTS

We thank the BnF teams for their active contribution to this project, from providing the data to constant feedback on the analyses. The data was processed on the Teralab big data infrastructure[6].

REFERENCES

[1] D. Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press. 411–416 pages.

[2] V. Beaudouin, I. Garron, and N. Rollet. 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.

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

[4] R. Cooley, B. Mobasher, and J. Srivastava. 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems* 1, 1 (1999), 5–32.

[5] R. F. Dell, P. E. Román, and J. D. Velásquez. 2008. Web user session reconstruction using integer programming. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 385–388.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1 (1977), 1–38 (with discussion).

[7] C. Eickhoff, J. Teevan, R. White, and S. Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2556195.2556217

[8] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner. 2014. Web data extraction, applications and techniques: a survey. *Knowledge-based systems* 70 (2014), 301–323.

[9] Mark Girolami and Ata Kabán. 2003. Simplicial Mixtures of Markov Chains: Distributed Modelling of Dynamic User Profiles. In *Advances in Neural Information Processing Systems 16*, Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf (Eds.). MIT Press, Cambridge, MA, None. http://books.nips.cc/papers/files/nips16/NIPS2003_AA02.pdf

[10] Guillaume Godet. 2015. *Guide d'interopérabilité OAI-PMH pour un référencement des documents numériques dans Gallica*. Technical Report. BnF. http://www.bnf.fr/documents/Guide_oaipmh.pdf

[11] S. Gündüz and M. T. Özsu. 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*. ACM, 535–540.

[12] Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *WWW 2015*. ACM Press, 410–418. https://doi.org/10.1145/2736277.2741117

[13] H. R. Jamali, D. Nicholas, and P. Huntington. 2005. The use and users of scholarly e-journals: a review of log analysis studies. *Aslib Proceedings* 57, 6 (2005), 554–571. https://doi.org/10.1108/00012530510634271

[14] R. Jones and K. L. Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 699–708.

[15] J. Kunze. 2003. Towards electronic persistence using ARK identifiers. In *Proceedings of the 3rd ECDL Workshop on Web Archives*.

[6] https://www.teralab-datascience.fr

[16] J. Kunze. 2003. Towards Electronic Persistence Using ARK Identifiers, ARK motivation and overview. (2003). http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf
[17] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. 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 (WIDM '01)*. ACM, New York, NY, USA, 9–15. https://doi.org/10.1145/502932.502935
[18] D. Mobasher, H. Dai, T. Luo, and M. Nakagawa. 2002. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery* 6, 1 (2002), 61–82. https://doi.org/10.1023/A:1013232803866
[19] E. L. Morgan. 2004. An Introduction to the Search/Retrieve URL Service (SRU). (2004). http://www.ariadne.ac.uk/issue40/morgan/
[20] D. Nicholas, P. Huntington, and A. Watkinson. 2005. Scholarly journal usage: the results of deep log analysis. *Journal of Documentation* 61, 2 (2005), 248–280. https://doi.org/10.1108/00220410510585214
[21] N. Rollet, I. Garron, and V. Beaudouin. 2017. *Vidéo-ethnographie des usages de Gallica : une exploration au plus près de l'activité.* Technical Report. BnF - Labex Obvil - Telecom ParisTech, Paris.
[22] N. Sadagopan and J. Li. 2008. Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 885–894.
[23] S. S. C. Silva, R. M. P. Silva, R. C. G. Pinto, and R. M. Salles. 2013. Botnets: A Survey. *Comput. Netw.* 57, 2 (Feb. 2013), 378–403. https://doi.org/10.1016/j.comnet.2012.07.021
[24] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. 2003. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs journal on computing* 15, 2 (2003), 171–190.
[25] D. W. Stroock. 2014. *An introduction to Markov processes* (second ed.). Graduate Texts in Mathematics, Vol. 230. Springer, Heidelberg. xviii+203 pages. https://doi.org/10.1007/978-3-642-40523-5
[26] P-N. Tan and V. Kumar. 2002. Discovery of Web Robot Sessions Based on their Navigational Patterns. *Data Mining and Knowledge Discovery* 6, 1 (2002), 9–35. https://doi.org/10.1023/A:1013228602957
[27] C. Tenopir. 2003. Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies. *Council on Library and Information Resources* (August 2003), 72. http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED499383
[28] A. Ypma and T. Heskes. 2003. *Automatic Categorization of Web Pages and User Clustering with Mixtures of Hidden Markov Models.* Springer Berlin Heidelberg, 35–49.