Analyse des traces d'usage de Gallica

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

ANALYSE DES TRACES D'USAGE DE GALLICA

Une étude/analyse à partir des logs de connexions au site Gallica

Auteurs :

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

Télécom ParisTech

Novembre 2017

Table des matières

Chapitre 1

Présentation de l'étude

1.1 Contexte

Cette étude s'inscrit dans le cadre de la convention de partenariat entre la Bibliothèque nationale de France et Télécom ParisTech, qui a donné naissance au *Laboratoire d'étude des usages du patrimoine numérique des bibliothèques* (Bibli-Lab). Il a pour but de consolider les fondations épistémologiques du Laboratoire en identifiant les enjeux théoriques et méthodologiques que soulève l'appréhension des usages de Gallica au regard des récents bouleversements dans les usages en ligne. Dans cette optique, ce rapport a pour objectif d'introduire des éléments de statistique décisionnelle permettant la compréhension des usages de Gallica.

Cette étude comprend trois volets :

— une description exhaustive des données sur lesquelles des analyses statistiques sont appliquées,

— la présentation et l'application de méthodes de "datamining" pour la découverte de tendances dans les sessions d'utilisation de Gallica,

— une discussion sur les résultats préliminaires ainsi que la proposition de pistes de recommandations opérationnelles pouvant permettre de faciliter les analyses et d'en améliorer la pertinence.

Elle a été suivie par un comité de projet composé des auteurs de ce rapport et de personnels de différents services de la BnF sous la direction de Philippe Chevallier et d'Emmanuelle Bermès.

1.2 Méthodologie et objectifs

L'ensemble des méthodes présentées dans la suite du rapport repose sur l'exploitation des données de connexion des internautes à Gallica, appelées "logs de connexion". Les données de connexion contiennent principalement, pour chaque internaute (plus précisément pour chaque adresse IP), les pages consultées ainsi que la date et l'heure de consultation. L'utilisation de l'horodatage peut permettre de séparer les différentes sessions sur Gallica (par un identifiant unique) et donc de pouvoir prendre connaissance des documents consultés dans une même session de navigation. Il est à noter qu'une session est construite à partir des données, cette étape est décrite dans la suite du rapport (voir Section 2.2.1). La Bibliothèque nationale de France conserve les données de connexion et les anonymise pour les besoins de la présente recherche. Il est à noter que la méthodologie choisie s'attache à l'étude globale/quantitative des usages de Gallica. De ce fait, il est recherché des *tendances* présentes dans l'ensemble ou un sous-ensemble d'internautes. Cette approche est complémentaire des études qualitatives qui s'appuient sur des entretiens individuels et des observations ethnographiques, restreintes à un panel limité d'usagers. Pour de plus amples informations sur les études qualitatives menées à la BnF le lecteur peut se référer à Beaudouin and Denis [2014].

La présente étude quantitative a pour but final de catégoriser les usages et donc de permettre le regroupement d'utilisateurs de Gallica en fonction de similitudes dans leurs comportements sur Gallica. Identifier des *types d'usages* et connaître leurs proportions chez les utilisateurs n'a, à notre

connaissance, jamais été réalisé par une bibliothèque patrimoniale numérique. On notera cependant une étude basée sur les logs de connexion proposée par Ceccarelli et al. [2011] dans le cadre du projet Europeana. Le présent rapport présente des méthodes statistiques permettant une cartographie des usages de Gallica (type de documents consultés, parcours des utilisateurs sur le site Gallica). Ces méthodes algorithmiques pourraient dans l'avenir servir d'aide à la décision concernant :

— l'amélioration de l'architecture du site,

— la personnalisation avec des recommandations d'ouvrages,

— le "business intelligence" particulièrement pour orienter la numérisation de nouveaux documents potentiellement plus appréciés par les usagers de Gallica,

— l'amélioration des activités de médiation.

1.3 Plan du rapport

Ce rapport se décompose en cinq chapitres et une annexe. Il est important de préciser que l'ordre des chapitres correspond à la chronologie du travail réalisé. Ainsi après un premier chapitre d'introduction, l'étude est décrite à travers les chapitre 2, 3 et 4 puis des recommandations sont proposées dans le chapitre 5. Le chapitre 2 introduit la première partie de l'étude et se concentre sur la description des logs de connexion, support principal de cette étude. Elle se divise en deux sous-parties discutant respectivement de l'information contenue dans les logs et de l'*enrichissement* de ces derniers. L'*enrichissement* consiste à ajouter à chaque document consulté les métadonnées issues de l'entrepôt OAI de Gallica (voir Section 2.1.3), ce qui permet de rendre lisibles et interprétables par l'analyste les logs. Le volume important des données produites par les connexions au site web impose la mise en place d'une infrastructure de stockage qui fait ainsi l'objet d'une sous-partie.

Les deux chapitres suivants forment la deuxième partie de l'étude et s'intéressent à la description de deux analyses quantitatives. Deux axes de travail ont été choisis :

(i) la découverte de parcours utilisateurs (chapitre 3),

(ii) la découverte de sujets privilégiés chez les utilisateurs (chapitre 4).

L'axe (i) a pour but de découvrir des motifs communs dans les sessions de consultation de Gallica. Dans ce chapitre, nous nous sommes attardés à modéliser l'influence d'une action à l'instant t au sein de l'interface Gallica sur l'action à l'instant $t+1$. L'axe (ii) se concentre sur les documents consultés au cours des sessions, et permet la découverte des sujets les plus populaires. La proposition (ii) a également conduit à la recherche de sujets plus *fins* pour décrire les documents contenus dans Gallica.

Enfin, le chapitre 7 regroupe l'ensemble des recommandations déjà prises en compte ou à instruire dans le but de rendre possibles des analyses futures ou d'améliorer et de perfectionner les analyses réalisées dans ce rapport. Nous proposons, en guise de conclusion, des pistes de futurs travaux.

Chapitre 2

Description/Enrichissement des logs de connexions

2.1 Description des logs

2.1.1 Nature des logs

Lorsqu'un internaute navigue sur un site internet, le serveur reçoit une requête lui indiquant quelles informations/pages doivent être renvoyées à l'internaute. L'ensemble des requêtes est enregistré sous la forme de logs. Ces logs forment un ensemble de lignes propre à chaque requête et possèdent les informations suivantes (dans le cas de Gallica) :

- l'adresse IP (identifiant unique d'une connexion interne) anonymisée par application d'une fonction de hachage préservant l'individualité. Il est également extrait la provenance géographique (ville et pays),
- la date, l'heure, la minute et la seconde de la requête,
- la requête HTTP,
- la réponse du serveur (erreur/succès),
- la taille du fichier renvoyé,
- le site référent (la provenance de l'utilisateur).

Les requêtes HTTP peuvent être très variées. Les différentes requêtes rencontrées pour le site Gallica sont décrites dans la sous-section suivante.

2.1.2 Requêtes HTTP

Requêtes HTML

Definition 2.1.1. Une page web statique est une page web dont le contenu ne varie pas en fonction des caractéristiques de la demande, c'est-à-dire qu'à un moment donné, tous les internautes qui demandent la page reçoivent le même contenu.

FIGURE 2.1 – Exemple d'une ligne de logs

Les requêtes HTML interviennent lorsque l'utilisateur souhaite accéder à une page web statique Def. 2.1.1. Au sein de Gallica, les pages statiques sont principalement la page d'accueil et les pages de présentation de collections (par exemple `http://gallica.bnf.fr/html/und/livres/livres`).

Requêtes liées au *web-design*

Une requête liée au *web-design* du site Gallica comprend :

— des scripts javascript : ex. *GET /js/jScrollPane_ min.js*,

— des scripts css : ex. *GET /assets/static/stylesheets/gallica-accueil.css*,

— des images propres au "design" comme un logo : ex. *GET html/sites/default/files/visuel_ applications.png*.

Dans le cadre de cette étude, ce type de requête n'est généralement pas porteur d'information utile ou exploitable.

Requête SRU

Le protocole SRU Morgan [2004] (Search and Retrieve via URL) permet la communication entre un serveur et un client via le protocole HTTP. Ce type de requête intervient lorsqu'un usager fait appel au moteur de recherche de Gallica. Une requête SRU peut utiliser l'une des deux méthodes suivantes :

— *GET* : requête envoyée par le serveur au client pour la récupération d'une recherche.

— *POST* : requête envoyée par le client au serveur pour une recherche.

Une requête de type SRU permet donc d'identifier l'utilisation du moteur de recherche interne à Gallica. Nous lui attachons une importance particulière puisque le contenu d'une telle requête par l'usager nous permet de récupérer les mots clés d'une recherche d'un usager.

Requêtes Ark

Le texte de cette sous-section est issu du site officiel de la BnF : BnF - ARK (Archival Resource Key). ARK (Archival Resource Key) Kunze [2003] est un système d'identifiants mis en place par la California Digital Library (CDL), et qui a vocation à identifier des objets de manière pérenne. Il peut s'agir d'objets de tous types, physiques (table, livre), numériques (livre numérisé...) ou même immatériels (concepts, ...).

L'autorité d'adressage (NMAH :Name Mapping Authority Host) est le service qui se charge de résoudre les identifiants ARK, c'est-à-dire de rendre un identifiant ARK *actionnable* en permettant l'accès à l'objet qu'il identifie et/ou à sa description.

La partie pérenne de l'identifiant est constituée du type d'identifiant, du numéro d'autorité nommante (NAAN) et du nom ARK proprement dit, dans notre exemple il s'agit de la chaîne de caractères "ark :/12148/bpt6k107371t" :

— Le type, ou schème, d'identifiant "ark :" déclare qu'il s'agit d'un identifiant ARK

— Le numéro d'autorité nommante (NAAN : Name Assigning Authority Number) identifie une institution habilitée à attribuer des ARK. Ce numéro sur 5 caractères, unique au sein du schème "ark :", est attribué gratuitement à toute institution qui en fait la demande par la California Digital Library, qui en assure la maintenance et l'unicité. Il est consigné dans le répertoire des autorités nommantes ("NAAN registry")

— Le nom ARK est un identifiant non signifiant attribué par l'autorité nommante. Il peut être composé de préfixes, qui permettent de regrouper de grands ensembles de ressources selon des critères laissés à l'appréciation de l'autorité nommante. Le nom ARK est composé d'une chaîne de caractères alphanumériques à l'exclusion des voyelles. Il est recommandé, bien que non obligatoire, de terminer le nom ARK par un caractère de contrôle.

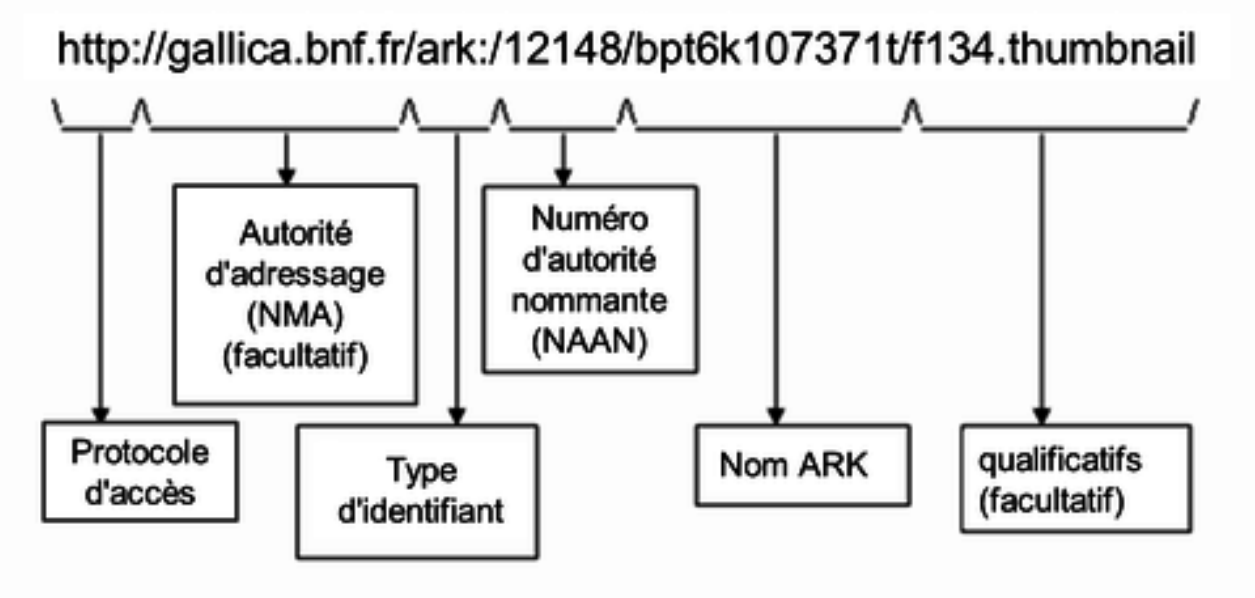

FIGURE 2.2 – Schéma d'une requête Ark

Les qualificatifs sont des suffixes permettant de préciser sa demande sur un document et sont de deux types :

— Les qualificatifs de granularité, commençant par un "/", permettent de demander l'accès à une partie de l'objet (ex. : page d'un document). Ils suivent immédiatement le nom ARK

— Les qualificatifs de service, commençant par un ".", permettent de demander l'accès à une variante particulière d'un document et/ou à un service particulier sur celui-ci (type de dissémination particulier d'un document, version n°1 du document, zoom, téléchargement...).

Ces qualificatifs sont définis et maintenus par l'autorité nommante qui garantit leur cohérence, et sont résolus et interprétés par l'autorité d'adressage. En conjonction avec un qualificatif de granularité, les qualificatifs de services se situent obligatoirement après ce dernier.

2.1.3 Enrichissement avec les métadonnées

Lors de la numérisation des documents, la Bibliothèque nationale de France s'efforce d'accompagner les documents d'une fiche unique contenant des métadonnées. Ces métadonnées encodées au format Dublin Core sont alors accessibles par le biais de l'entrepôt OAI Godet [2015]. Les métadonnées de l'ensemble des documents de Gallica ont donc pu être *moissonnées* et ont permis d'*enrichir* les logs de connexion en associant l'identifiant ARK de chacun des documents à sa notice. L'entrepôt OAI de la Bibliothèque nationale de France comporte les métadonnées suivantes :

— Titre
— Auteur
— Année de publication
— Identifiant ARK
— Source
— Langage
— Type (monographie, fascicules, photographies, ...)
— Thème* (droit, sciences médicales, histoire de France, biographie, ...)
— Corpus* (1914-1918, Rhône-Alpes, voyages en Italie, ...)
— Description*

Les champs annotés d'une étoile * sont facultatifs.

2.2 Pré-traitement

2.2.1 Définition d'une session

Definition 2.2.1. Une session est une séquence de requêtes faite par un unique utilisateur durant la visite d'un site particulier.

Definition 2.2.2. La *sessionisation* est l'étape permettant d'éclater l'ensemble des requêtes en sessions.

Lorsqu'aucune information n'est disponible sur l'appartenance d'une série de requêtes, une étape de *sessionisation* Def. 2.2.2 est nécessaire. Il est cependant difficile à partir des logs de réaliser l'étape de *sessionisation* respectant parfaitement les définitions précédentes Def. 2.2.1 et Def. 2.2.2. On propose alors une heuristique pour l'étape de *sessionisation*. Une *sessionisation* basique s'effectue en regroupant les requêtes partageant la même adresse IP, le même *user-agent*. Une session se termine lorsqu'il existe un intervalle supérieur à X minutes entre deux requêtes. La valeur X est discutable, Cooley et al. [1999] l'estime à 10 minutes, nous lui préférons une valeur supérieure de 60 minutes. Das and Turkoglu [2009], Google [2016] proposent de compléter l'heuristique précédente en limitant la durée maximale à 30 minutes par session. Cette limite est basée sur des observations expérimentales montrant que le temps moyen d'un usage de site internet dépasse rarement ce temps. Nous ne préférons pas incorporer cette limitation dans nos analyses et donc de ne pas fixer une limite haute pour la durée des sessions et ce, parce qu'il nous semble tout à fait probable qu'un usager poursuive une recherche documentaire d'une durée supérieure à 60 minutes. Cette intuition a ensuite été confirmée par les résultats de l'enquête GMV Conseil [2012] dont 45% des sondés affirment avoir des sessions de durée supérieure à 30 minutes.

Il est à noter que la *sessionisation* présente un biais pour des utilisateurs partageant la même adresse IP et la même version de navigateur internet. En effet l'heuristique choisie les considère, à tort, comme une unique session. Ce cas particulier devrait apparaître pour les utilisateurs d'une même institution/entreprise (en particulier les usagers interne à la BnF).

Pour remédier à ce problème, une approche alternative est envisagée et repose sur l'utilisation de "cookies". Un simple regroupement des requêtes associées à un même "cookie" permet d'identifier de manière efficace un utilisateur et une session. Néanmoins, une limitation majeure de cette méthode est l'absence de la première requête (la première requête d'un usager ne possède pas encore de "cookie"). La mise en place de "cookies" sera plus amplement discutée dans le chapitre Chap. 7.

Des méthodes de *sessionisation* stochastiques existent et ont été développées par Sadagopan and Li [2008], Dell et al. [2008] mais ne nous semblent pas adaptées aux logs de Gallica.

Pour résumer, nous modifions la définition précédente de session par Def. 2.2.3 :

Definition 2.2.3. Une session est une séquence de requêtes faite par une unique adresse IP. Une nouvelle session est définie lorsque l'adresse IP devient inactive pendant 60 minutes.

2.2.2 Représentation des sessions

Dans cette sous-section, nous introduisons les notations choisies pour la représentation d'une session. On peut définir une session comme un vecteur $\mathbf{x} = \left[x_1, \ldots, x_M\right]$ où M est le nombre total de pages ou un sous-ensemble de pages. De part sa définition la plus simple donnée par Mobasher et al. [2001], les composantes sont binaires, $x_m \in \{0, 1\}$, en fonction de la visite ou non de la page en question. Mobasher et al. [2002] propose que chaque composante de $\mathbf{x}$ soit normalisée en fonction du temps de visite. Une autre approche consiste à pondérer les composantes du vecteur en fonction d'indicateur d'intérêt pour une page (voir Kelly and Teevan [2003]).

La représentation retenue est due aux travaux de Gündüz and Özsu [2003] qui proposent de conserver la séquentialité d'une session en la représentant comme la succession de requêtes associées à sa date. Dans ce cas précis, une session est représentée par un vecteur $\mathbf{x} = \left[x_1, \ldots, x_n, \ldots, x_N\right]$ avec N le nombre de pages visitées durant la session. Chaque composante $x_n \in \{1, \ldots, M\}$ est donc

l'identifiant de la $n-ieme$ page visitée. Cette représentation sera utilisée pour l'analyse des parcours en chapitre Chap. 3.

Il est à noter que chaque représentation présente des avantages et inconvénients. Le choix de la représentation dépendra du type d'analyse effectuée. Pour plus d'informations sur les représentations de sessions, le lecteur pourra se référer à Demir et al. [2007].

2.2.3 Bots-Crawlers

Definition 2.2.4. Un robot internet ("bot" en anglais) est un logiciel qui réalise des tâches automatisées sur internet.

Definition 2.2.5. Un robot d'indexation (web crawler ou web spider) est un logiciel qui explore automatiquement le Web. Il est généralement conçu pour collecter les ressources (pages Web, images, vidéos, documents Word, PDF ou PostScript, etc.), afin de permettre à un moteur de recherche de les indexer.

Network [2015], Incapsula [2003] affirment que les robots représenteraient presque 50% du trafic internet. Il est donc indispensable de prendre en compte les robots et de les filtrer pour que leur impact soit minimum sur nos analyses. Les robots peuvent être catégorisés de manière manichéenne :

- "bad bots"
  - les bots zombies ou botnets Silva et al. [2013]
  - les bots scrapers : copient/téléchargent le contenu d'un site internet de manière exhaustive Ferrara et al. [2014]
  - les "click bots" : ces bots cliquent intentionnellement et frauduleusement sur les publicités.
- "good bots"
  - les robots d'indexation.
  - les robots liés au droit d'auteur. Ces robots parcourent le web dans le but de trouver du contenu plagié ou comportant un droit d'auteur.

Les "good bots" sont généralement faciles à détecter puisqu'ils s'identifient par le biais du "User-agent". Il suffit donc de filtrer les requêtes dont le "User-agent" contient les mots clés suivants : photon, bingpreview, bot, spider, slurp, linkchecker, parser, google, facebook, yahoo.

La détection des "bad bots" est une tâche compliquée (voir Tan and Kumar [2002]) et nous ne la considérons pas centrale dans ce rapport.

2.3 Moyens techniques

2.3.1 TeraLab

TeraLab est un projet ayant pour but d'apporter à la communauté nationale des chercheurs et enseignants, mais aussi aux entreprises, un environnement de recherche et d'expérimentation pour leurs applications innovantes ou pilotes industriels impliquant de gros volumes de données. La plateforme TeraLab intègre des technologies matérielles, logicielles pour permettre des traitements "batch" ou temps réel et le stockage de données de manière sécurisée.

Definition 2.3.1. Le "cloud computing", ou l'informatique en nuage ou nuagique ou encore l'infonuagique (au Québec), est l'exploitation de la puissance de calcul ou de stockage de serveurs informatiques distants par l'intermédiaire d'un réseau, généralement internet.

TeraLab propose donc une solution de "cloud computing" (Def. 2.3.1) qui a été utilisée pour réaliser l'exploitation des données de la Bibliothèque nationale de France. Un point fort de TeraLab est l'hébergement souverain des données grâce à l'implantation sur le sol français des serveurs. Une politique de TeraLab axée sur la sécurisation des serveurs est également un point essentiel dans le

choix de cette infrastructure au vu du caractère sensible des données exploitées dans le cadre de ce projet.

En résumé, TeraLab a été choisie pour cette étude puisqu'il s'agit d'une plateforme répondant à nos attentes en termes de volume de données, volumes de calcul et de sécurité.

2.3.2 Python

Definition 2.3.2. Python est un langage de programmation objet, multi-paradigme et multiplate-forme. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions. Le langage Python est placé sous une licence libre proche de la licence BSD3 et fonctionne sur la plupart des plates-formes informatiques, des supercalculateurs aux ordinateurs centraux, de Windows à Unix en passant par GNU/Linux, Mac OS, ou encore Android, iOS.

Les algorithmes développés dans le cadre de ce projet ont été intégralement conçus à l'aide du langage Python (Def. 2.3.2). Il est particulièrement adapté à nos besoins puisqu'il offre des outils de haut niveau programmés par la communauté des chercheurs (en *Machine Learning* notamment) et une syntaxe simple à utiliser.

2.3.3 Elasticsearch

Definition 2.3.3. Elasticsearch est un serveur utilisant Lucene pour l'indexation et la recherche des données. Il fournit un moteur de recherche distribué et multi-entité à travers une interface REST. C'est un logiciel libre écrit en Java et publié en open source sous licence Apache.

L'outil Elasticsearch (subsection 2.3.3) a été choisi comme infrastructure pour l'indexation et le stockage des logs de connexion. Ce choix est motivé, en premier lieu, par l'interfaçage efficace et facile entre le langage Python et Elasticsearch. En pratique, une fois indexés, les logs peuvent être aisément récupérés par une simple requête à Elasticsearch.

2.4 Statistiques globales

2.4.1 Solution existante : XiTi

XiTi est un service de mesure d'audience et de statistiques pour les sites web propriétaire et developpé par AT Internet. Cette solution est actuellement employée par la Bibliothèque nationale de France pour l'audit de la fréquentation de Gallica. XiTi regroupe des indicateurs de l'activité des visiteurs tels que : le nombre de visites, le nombre de pages vues ou encore l'origine de la visite. Cependant cet outil reste limité à des statistiques sommaires. Nous avons utilisé les résultats produits par XiTi comme élément de comparaison pour notre approche par logs décrite dans la sous-section suivante.

2.4.2 Approche par logs

Cette section introduit les analyses des chapitres Chap.3, 4 en tentant de résumer des statistiques de base sur les usagers de Gallica. Le tableau Tab. 2.1 compare des métriques de notre approche par logs avec celles proposées par XiTi. Les données sont issues des usagers de Gallica entre le 1er Juin et le 15 Juin 2016. En se concentrant sur le nombres de visiteurs, on remarque une différence entre notre approche et XiTi. XiTi comptabilise $35,000$ visites par jour contre $45,000$ pour notre approche. Cette différence peut dans un premier temps s'expliquer par l'utilisation de bloqueurs de cookies/javascript. On peut également l'expliquer en partie par la non-comptabilisation de visiteurs qui n'accèdent pas à l'interface de Gallica mais qui accèdent à du contenu par hotlink (Def. 2.4.1) voire par l'application IIIF. En effet, XiTi introduit un code dans les pages HTML de Gallica a

	Logs	XiTi
Visites	45,000/jour	35,000/jour
Visites avec au moins 1 document unique	35,000/jour	
Visites avec au moins 5 documents uniques	3,500/jour	
Moyenne des durées de visite	12min14sec	13min
Médiane des durées des visite	12sec	
Nombre de pages/visite moyen	30	20
Nombre de pages/visite médian	4	

TABLE 2.1 – Comparaison approche logs vs XiTi.

partir duquel XiTi incrémente son compteur de visites. L'approche logs nous permet de repérer ces hotlinks et de les comptabiliser. Inclure les usagers "hotlinkers" dans la notion de visiteurs est discutable puisque ces derniers ne sont pas nécessairement conscients qu'ils consultent du contenu de Gallica.

Definition 2.4.1. Le hotlinking (ou liaison automatique ; aussi connu en anglais sous les noms de inline linking, leeching, piggy-backing, direct linking ou offsite image grabs) consiste à utiliser l'adresse d'un fichier publié sur un site web, le plus souvent une image, pour l'afficher sur un autre site, sur un blog, dans un forum. En d'autres termes, au lieu d'enregistrer l'image et de l'installer sur son propre serveur Web, le hotlinkeur crée un lien direct vers le serveur d'origine.

On observe grâce au tableau 2.1 que seulement $\sim$ 8% des visiteurs accèdent à au moins 5 documents uniques. Cette métrique permet de remarquer qu'une grande proportion des usagers de Gallica ont un usage furtif du site. Des résultats similaires pour d'autres sites internet ont été reportés par Attenberg et al. [2009], Qiu and Cui [2010].

Le temps moyen de visite est d'environ 13 minutes, alors que le temps médian est de 12 secondes. Cet écart se justifie par une forte disparité dans la durée des visites. Il existe ainsi des sessions dont le temps de navigation est particulièrement important, ce qui a pour effet d'influencer la moyenne globale. Nous pensons que la moyenne n'est pas une métrique pertinente compte tenu de la distribution de la durée des visites. L'histogramme de la durée des visites donné par la figure 2.3 illustre ces propos. La figure 2.3 décrit l'effet de longue traîne au sein de la durée des visites : il existe une grande diversité dans la durée d'utilisation de Gallica. Cependant, on remarque aussi qu'une grande proportion des usagers de Gallica navigue moins d'une minute (voire même 0 seconde). Comme la durée d'une visite est obtenue en soustrayant le temps de la dernière et de la première requêtes, une durée nulle correspond à une visite ne réalisant qu'une seule requête.

Enfin, il est important de noter une proportion relativement faible de sessions contenant un appel à la recherche. Seulement 27% des sessions sur Gallica ont au minimum une interaction avec le moteur de recherche interne.

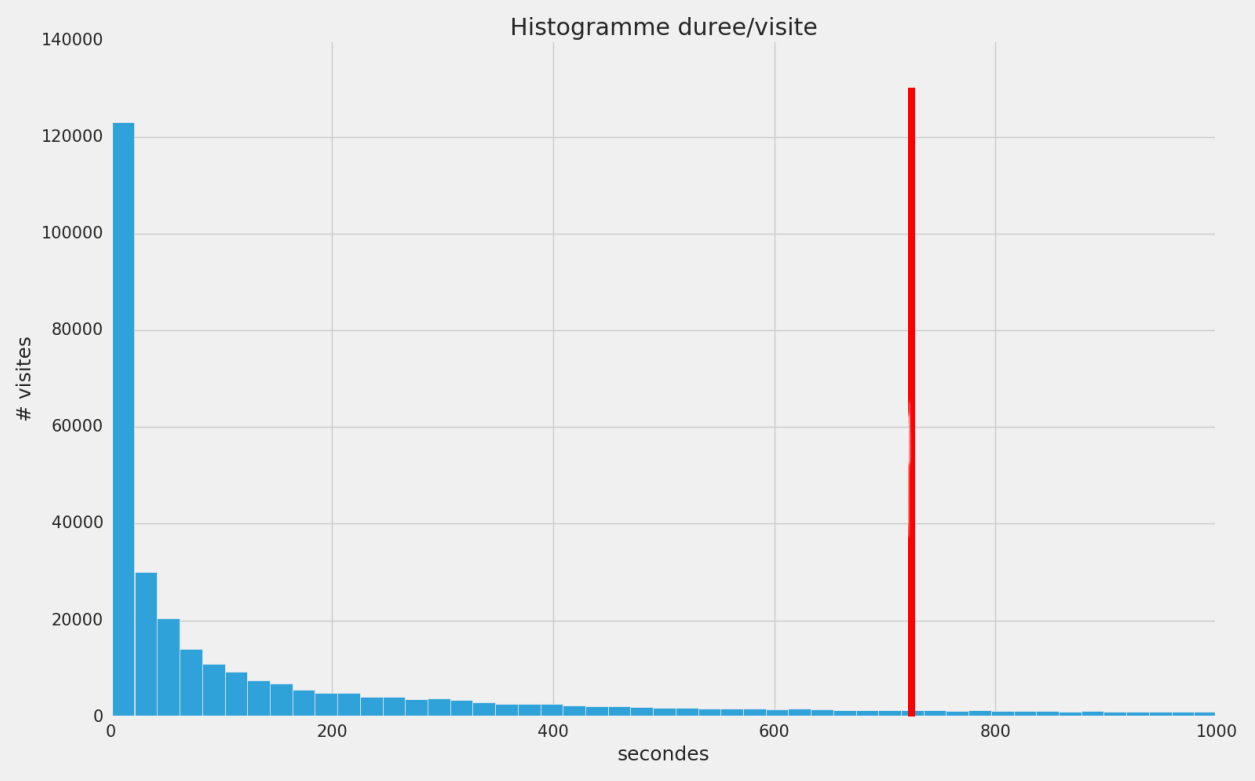

FIGURE 2.3 – Histogramme de la durée par visite en seconde. La barre verticale rouge représente la moyenne. L'histogramme est tronqué par souci de lisibilité à 1000 secondes.

Chapitre 3

Analyse des parcours d'usage

3.1 Classification non supervisée fondée sur un mélange de modèles de Markov

3.1.1 Introduction

Dans l'optique de rechercher des similitudes parmi les sessions, nous proposons un algorithme de *classification non supervisée* (*clustering*, en anglais) fondé sur un modèle de mélange de modèles de Markov. Il nous semble intéressant de regrouper les sessions qui présentent des similitudes dans les chemins parcourus pour en faire ressortir des usages *types* de Gallica "typiques" en termes d'enchaînement des actions au sein d'une session *http*. Cette section présente ainsi un algorithme de clustering qui, dans un premier temps, à partir d'un ensemble de données observées, estime les paramètres d'un modèle probabiliste et dans un deuxième temps, utilise ce modèle afin de catégoriser l'ensemble des sessions considérées en les associant à un comportement "typique". Le modèle que nous avons considéré pour cette tâche est le mélange fini de modèles de Markov d'ordre 1. Ce modèle suppose que chaque session est générée par un modèle tiré au hasard avec des probabilités *a priori* parmi K modèles de chaînes de Markov. Nous souhaitons alors estimer l'ensemble des paramètres de ce modèle :

— Pour chacun des modèles de Markov : les probabilités de l'état initial et final ainsi que les probabilités de passer d'un état à un autre, appelées *probabilités de transition*.

— Les probabilités *a priori* de chacun de ces modèles.

Cet algorithme présente plusieurs avantages :

— il est non supervisé, c'est-à dire qu'il ne requiert pas de données étiquetées.

— il est applicable à des séquences de longueurs arbitraires : il n'est pas nécessaire que les sessions aient le même nombre d'actions.

— il permet la découverte de "motifs" dans les sessions.

— il renvoie la proportion des données présentant les mêmes "motifs".

Dans ce chapitre, nous proposons donc de modéliser les types de navigation sur Gallica à l'aide de chaînes de Markov Rabiner [1990]. Chaque session se présente comme une succession d'actions (listées en nombre fini de manière exhaustive). Pour simplifier la description des sessions, nous repérons dans la séquence cinq types distincts d'actions sans nous soucier de la nature des documents consultés (toutes identifiées par l'URL de la requête `http` enregistrée dans la ligne de log correspondante) :

Action.1 Navigation sur la page d'accueil,

Action.2 Navigation sur les pages de collections,

Action.3 Utilisation du moteur de recherche Gallica,

Action.4 Téléchargement d'un document,

Action.5 Consultation d'un document dans l'interface Gallica (zoom, modification de l'affichage, page suivante...).

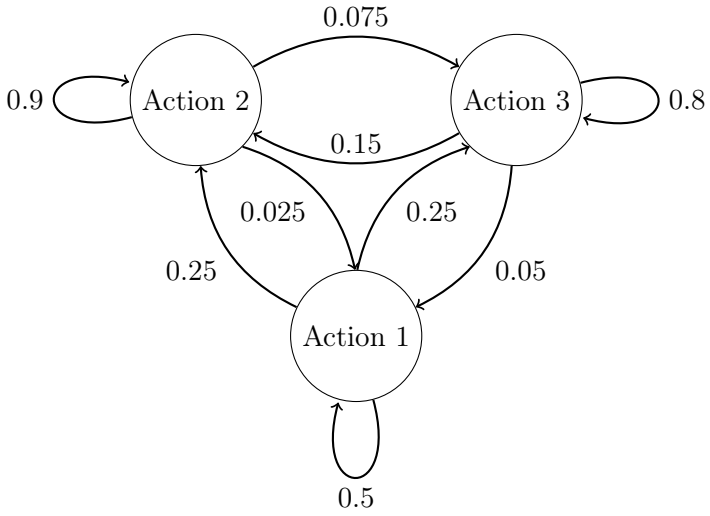

FIGURE 3.1 – Schéma d'une chaîne de Markov pour 3 actions

L'emploi des chaînes de Markov permet de prendre en compte la séquentialité d'une session. La chaîne de Markov est un modèle assez simple à exploiter qui permet de décrire l'influence d'une action à l'instant $n-1$ sur l'action suivante à l'instant n. La motivation derrière cette analyse est de découvrir par exemple quelles actions sont susceptibles d'influencer le téléchargement d'un document.

3.1.2 Descriptions du mélange de modèle de Markov

Soit $\mathbf{x} = \left[x_1, \ldots, x_n, \ldots, x_N\right]^\intercal$ une session produisant une succession de N actions. Le modèle de Markov d'ordre 1 est défini par ses probabilités de transition $\mathbf{A}$, et ses probabilités d'état initial et final :

$$\begin{aligned} A_{i,j} &= p(x_n = i | x_{n-1} = j) \ , \\ \pi_i &= p(x_1 = i) \ , \end{aligned}$$

$A_{i,j}$ est donc la probabilité de passer de l'état/action j à l'état i. Il est important de noter que l'ordre du modèle choisi rend l'état à l'instant n uniquement dépendant de son passé direct $n-1$ à savoir $p(x_n|(x_1, \ldots, x_{n-1})) = p(x_n|x_{n-1})$. Ce modèle ne prend donc pas en compte l'impact des actions antérieures à l'instant $n-1$ sur l'action à l'instant n.

Nous introduisons maintenant la notion de fonction de vraisemblance qui décrit l'adéquation de la distribution des données observées à la loi de probabilité supposée. Cette fonction est particulièrement utilisée en statistique pour l'estimation des paramètres d'une loi par maximisation de cette dernière. Dans le cas du modèle de Markov d'ordre 1 décrit précédemment, la vraisemblance d'observer une séquence $\mathbf{x}$ s'écrit à partir des paramètres du modèle $\boldsymbol{\theta} = \{\mathbf{A}, \pi_{i,\ldots,M}, \rho_{1,\ldots,M}\}$:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}) &= p(x_1)\Pi_{n=2}^{N} p(x_n|x_{n-1}) \ , \\ &= \Pi_{i=1}^{M} \pi_i^{[x_1=i]} \Pi_{n=1}^{N} \Pi_{i=1}^{M} \Pi_{j=1}^{M} A_{i,j}^{[x_t=i][x_{n-1}=j]} \ , \\ &= \Pi_{i=1}^{M} \pi_i^{[x_1=i]} \Pi_{i=1}^{M} \Pi_{j=1}^{M} A_{i,j}^{t_{i,j}} \ , \end{aligned} \tag{3.1}$$

où $[a=b]$ désigne la fonction indicatrice valant 1 lorsque la condition est respectée et 0 dans le cas contraire. Le terme $t_{i,j}$ est le compte des transitions entre l'état i et l'état j. La vraisemblance (3.1)

mesure l'adéquation d'une seule séquence à une seule chaîne de Markov. Maintenant, on considère un ensemble de sessions $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_s, \ldots, \mathbf{x}_S$ généré par un mélange de modèles de Markov. En d'autres termes, on suppose que chaque session $\mathbf{x}_s$ est générée par 1 parmi K modèles de Markov. On définit la probabilité *a priori* que la session $\mathbf{x}_s$ soit générée par le k^{ieme} modèle par :

$$p(z_s = k) = \alpha_k$$

L'ensemble $\boldsymbol{\theta}_k = \{\mathbf{A}^{(k)}, \pi^{(k)}_{1,\ldots,M}\}$ dénote les paramètres du k^{ieme} modèle de Markov, et $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\Theta}_K\}$ la concaténation des paramètres des K modèles. Dans le cas du mélange de modèles de Markov, la vraisemblance d'une séquence $\mathbf{x}_s$ devient :

$$p_{\boldsymbol{\Theta}}(\mathbf{x}_s) = \sum_{k=1}^{K} p(z_s = k) p_{\boldsymbol{\theta}_k}(\mathbf{x}_s | z_s = k) \tag{3.2}$$

On note ici que la variable z_s définit l'appartenance de la session s à la chaîne de Markov k qui n'est pas observé. On dit que cette variable est *latente*.

Si l'on suppose que les séquences sont indépendantes entre elles, la vraisemblance de l'ensemble des sessions $\mathbf{X}$ s'écrit alors :

$$p_{\boldsymbol{\Theta}}(\mathbf{X}) = \Pi_{s=1}^{S} p_{\boldsymbol{\Theta}}(\mathbf{x}_s) \,, \tag{3.3}$$

$$= \Pi_{s=1}^{S} \sum_{k=1}^{K} p(z_s = k) p_{\boldsymbol{\theta}_k}(\mathbf{x}_s | z_s = k) \tag{3.4}$$

3.1.3 Algorithme

Input: sessions, number of cluster : K, threshold
initialization;
$\boldsymbol{\Theta}^{(0)} \leftarrow random$;
$l_0 \leftarrow Equation$;
for $t \leftarrow 1$ **to** *ITER-LIMIT* **do**
  Compute $\boldsymbol{\Theta}^{(t)}$ using Equations Eq.(A.23, A.24, A.25) ;
  $l_t \leftarrow Equation$;
  if $l_t - l_{t-1} < threshold$ **then**
    break;
  end
end
return $\{\alpha_{1,\ldots,K}, \pi_i^{(1,\ldots,K)}, A_{i,j}^{(1,\ldots,K)}, \forall i, j \in \{1, \ldots, M\}^2, l_{1,\ldots}\}$

Algorithm 1: Algorithme Espérance-Maximisation pour apprendre les mélanges de modèles de Markov à partir des données sessions

Dans un premier temps, une vérification de la validité de l'implémentation de l'algorithme 1 est possible en analysant la vraisemblance à chaque itération et en s'assurant que $l_t > l_{t-1}$. Le cas contraire atteste d'une erreur. La convergence de l'algorithme peut se contrôler soit par une limitation du nombre d'itérations possibles, soit en comparant la différence entre les vraisemblances mesurées aux itérations $t-1$ et t à un seuil.

Le choix du nombre de classes K est une tache délicate. K peut être choisi arbitrairement où en fonction de critères comme l'AIC ou le BIC. Lorsque l'on estime un modèle statistique, il est possible d'augmenter la vraisemblance du modèle en ajoutant un ou plusieurs paramètres à ce dernier. Le critère d'information d'Akaike (AIC), tout comme le critère d'information bayésien (BIC), permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie Burnham and Anderson [2004]. Intuitivement plus un modèle sera complexe,

plus il sera en accord avec les données et ce, même s'il ne correspond pas au comportement qu'on aurait pu observer si on avait eu beaucoup plus de données. Le BIC et l'AIC permettent donc d'éviter ce sur-apprentissage ou "overfitting".

Rien n'assure que l'algorithme EM converge vers le maximum global (i.e la solution optimale), il est possible que l'algorithme se retrouve dans un maximum dit local. C'est pourquoi il est recommandé d'entraîner plusieurs mélanges dont les valeurs initiales contenues dans $\boldsymbol{\Theta}^{(0)}$ sont différentes.

3.1.4 Extension : modèle de Markov à temps continu

Nous proposons une extension au modèle précédent Section (3.1.2) prenant en compte le temps entre les différentes actions. La dérivation du modèle est donnée en Appendice B. Cette extension ajoute l'information temporelle qui lui permet d'être plus sensible aux actions qui requièrent un temps important. Il s'agit de joindre à la description Markovienne d'une session par la séquence des types d'action les durées de certaines actions pertinentes. Nous avons, par exemple, remarqué que les usagers qui consultent des pages de médiation ont tendance à réaliser peu d'actions mais restent plus longtemps sur le site web de Gallica. Le modèle de Markov à temps continu permet donc de capturer ce comportement et de lui concéder un poids juste dans les analyses.

3.1.5 Résultats : modèle de Markov à temps discret

Les données entre le 1 et le 5 Juin 2016 (incluant un week-end) et un choix arbitraire de $K = 20$ modèles ont permis d'obtenir les résultats résumés par les figures données dans la suite de cette sous-section. Il est important de rappeler que cette analyse n'a été réalisée qu'avec un sous-ensemble de la totalité des sessions recueillies pendant la période choisie. En effet, nous n'avons conservé que les sessions dont le nombre d'actions est supérieur à 5, ce qui représente $\sim 35\%$ de l'ensemble des sessions.

Nous avons fait le choix d'illustrer les résultats de l'estimation des paramètres du mélange de modèles de Markov en représentant les sessions *stéréotypes* de chacun des modèles de Markov d'ordre 1 estimés. Les sessions *stéréotypes* sont les sessions ayant une probabilité *a posteriori* supérieure à .80 pour ces modèles. Par souci de lisibilité, les sessions *stéréotypes* doivent avoir moins de 80 actions. Ces deux dernières contraintes n'ont permis de représenter que 15 *stéréotypes* parmi les 20.

Pour rappel, on choisit de voir les sessions comme une succession d'actions (de longueur non fixe) :

1. Navigation sur la page d'accueil :
2. Navigation sur les pages de médiations (collection + blog) :
3. Recherche Gallica :
4. Téléchargement d'un document :
5. Consultation *locale* dans l'interface Gallica (zoom, page suivante...) :

FIGURE 3.2 – Exemple de représentation d'une session à 10 actions.

Chacune des sous-figures suivantes est la concaténation verticale de sessions "stéréotypes" reprenant la représentation donnée par la figure Fig. 3.2. L'étiquette en ordonnée de chaque sous-figure contient la probabilité *a priori* des modèles de Markov α_k, soit en d'autres termes la proportion de ces comportements *stéréotypiques* dans les sessions sur Gallica.

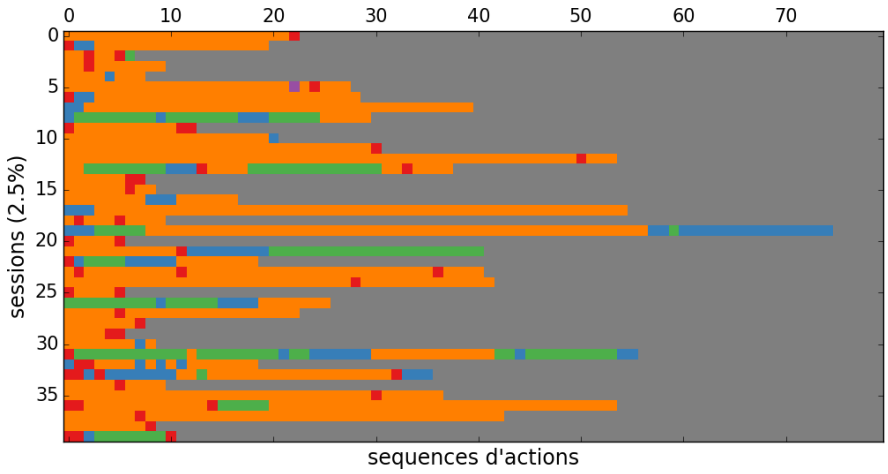

FIGURE 3.3 – cluster 2

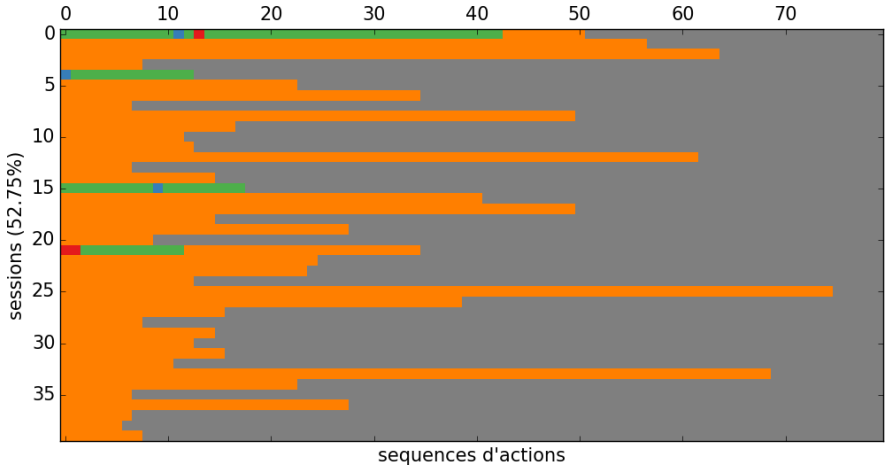

FIGURE 3.4 – cluster 3

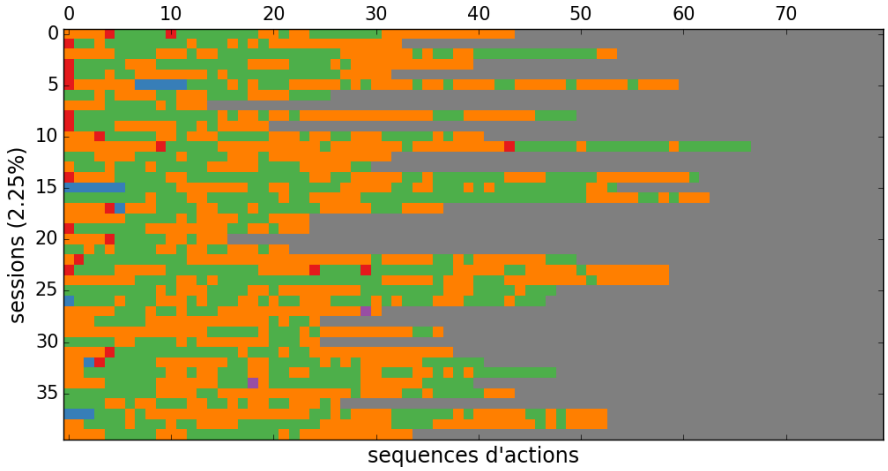

FIGURE 3.5 – cluster 4

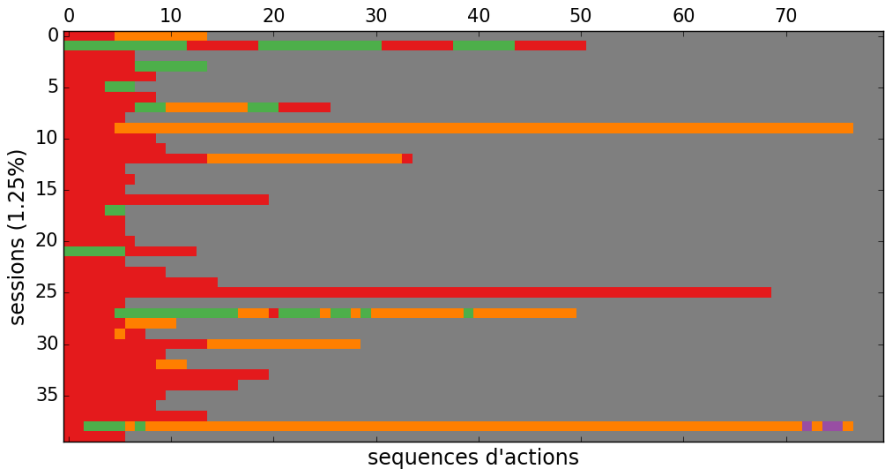

FIGURE 3.6 – cluster 5

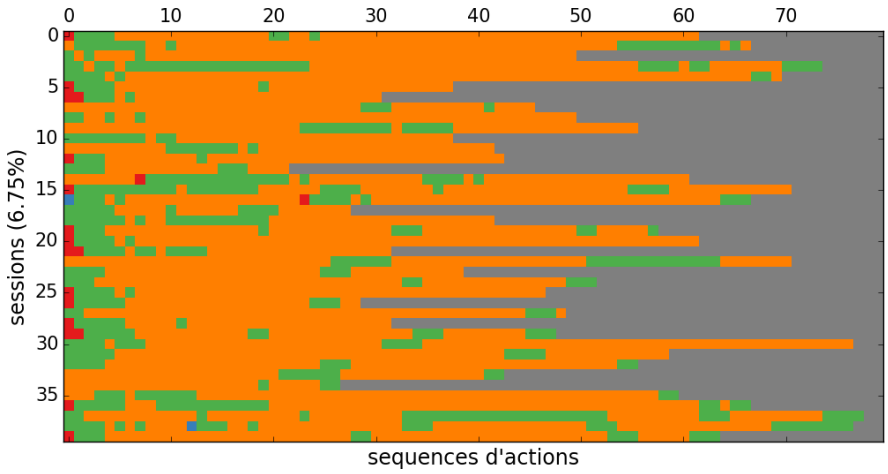

FIGURE 3.7 – cluster 6

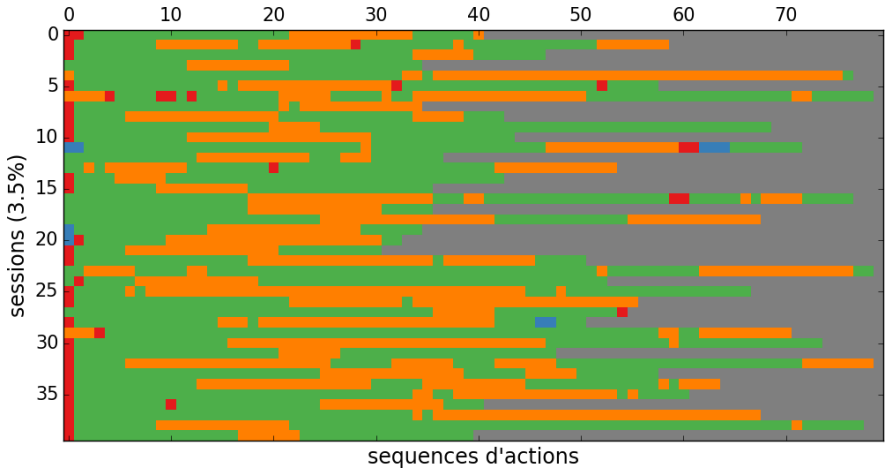

FIGURE 3.8 – cluster 7

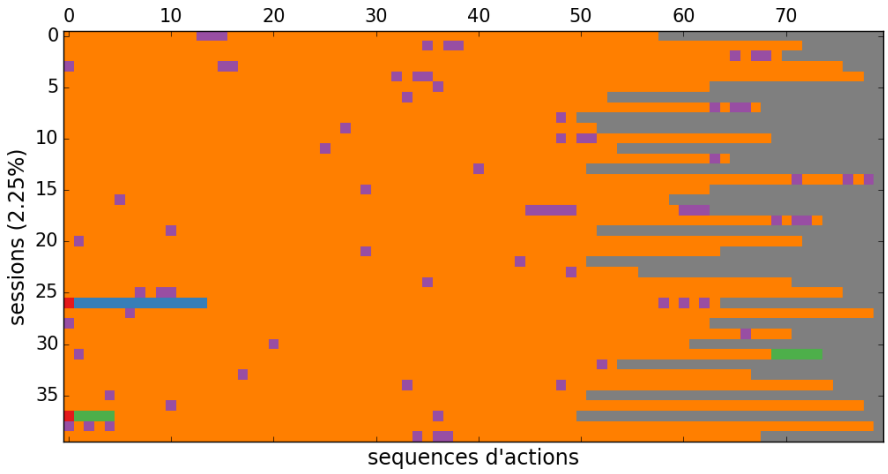

FIGURE 3.9

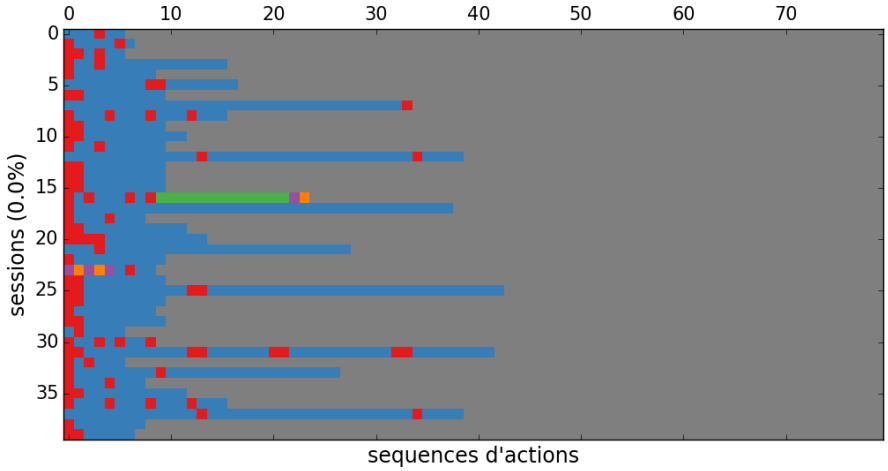

FIGURE 3.10

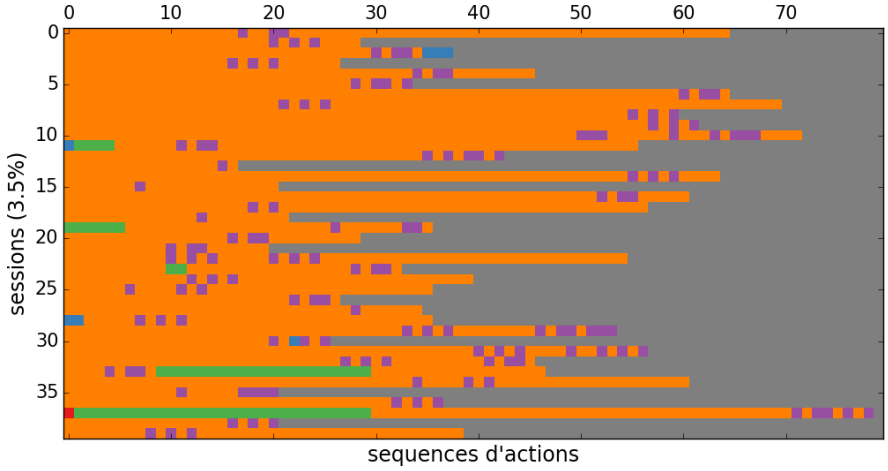

FIGURE 3.11

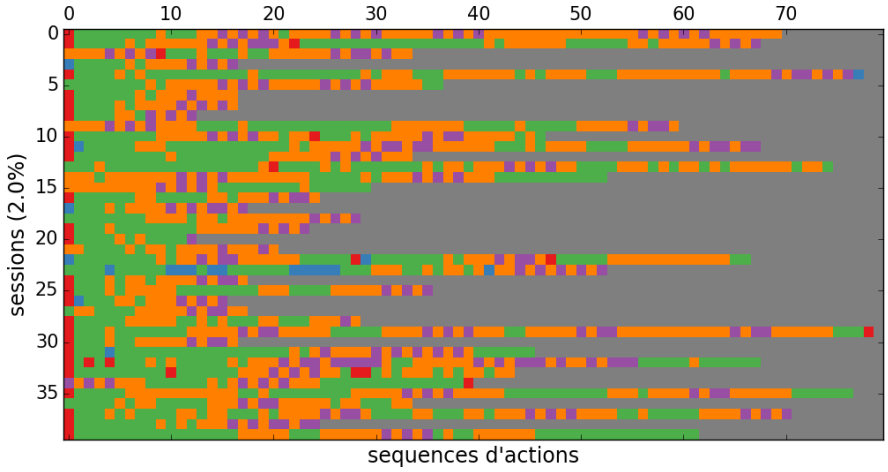

FIGURE 3.12

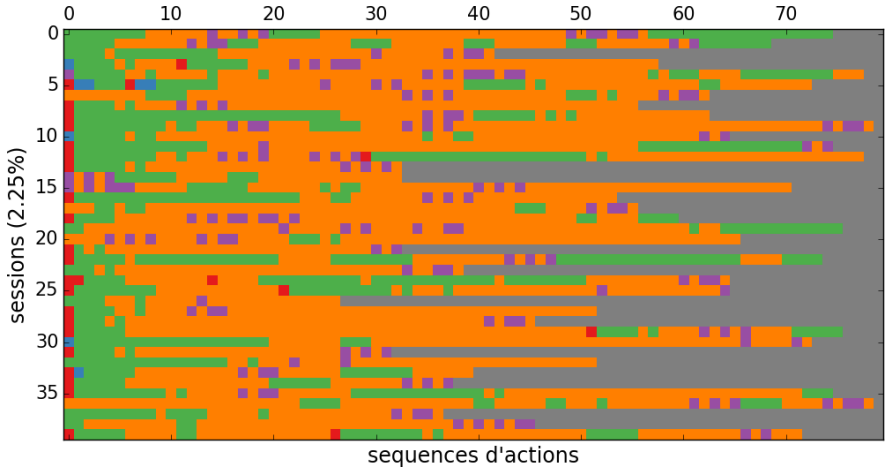

FIGURE 3.13

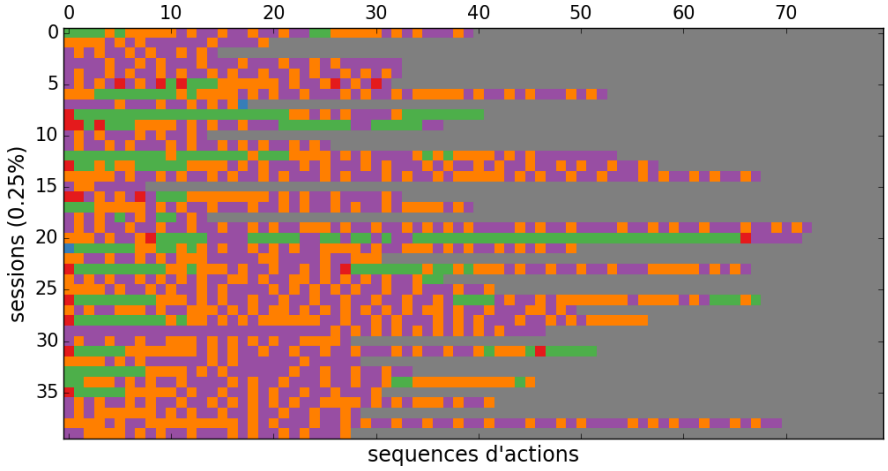

FIGURE 3.14

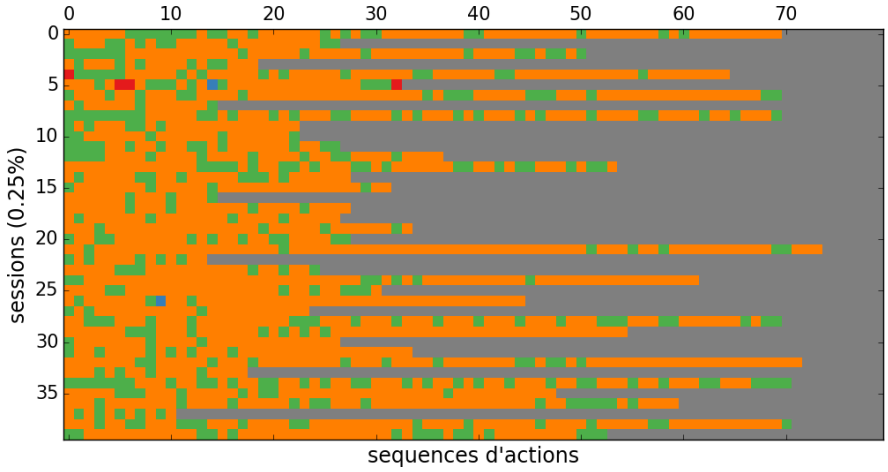

FIGURE 3.15

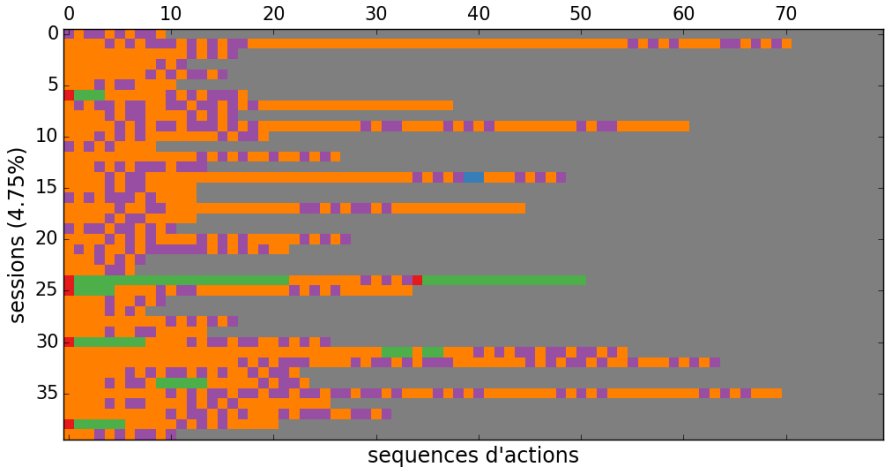

FIGURE 3.16

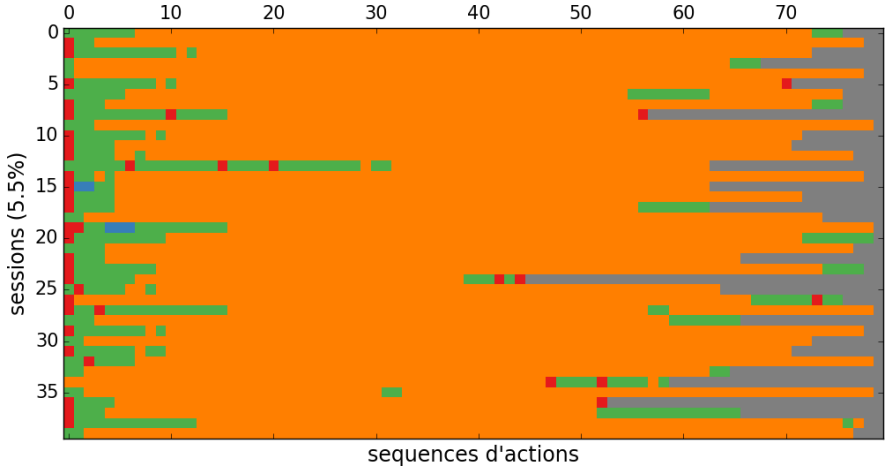

FIGURE 3.17

Cette catégorisation des sessions en fonction des types d'action fait émerger un type dominant qui regroupe 53% [1] des sessions retenues Fig. 3.4 : ces sessions correspondent principalement à des

1. Il est bon de rappeler qu'il s'agit d'un pourcentage relatif aux sessions retenues : à plus de 5 actions

séquences de consultation de documents enchaînant un nombre plus ou moins conséquent d'actions de consultation sans passer par la page d'accueil, ni par le moteur de recherche.

On peut supposer qu'il s'agit de sessions de lecture en ligne de documents précédemment identifiés ou trouvés directement via un moteur de recherche généraliste. Par delà ce type de session très majoritaire, les clusters de sessions ont des effectifs plus faibles, inférieurs à 8%. En procédant à un regroupement manuel des clusters, trois ensembles peuvent être distingués.

Le premier (18% des sessions) (Figs. 3.5, 3.7, 3.8, 3.15, 3.17) est caractérisé par l'alternance entre l'utilisation du moteur de recherche et la consultation de documents. En général, ces sessions commencent par la page d'accueil. A l'intérieur de cet ensemble, les variations tiennent au rythme d'alternance moteur/consultation. Dans certaines sessions, une phase de recherche concentrée est suivie par un long temps de consultations ; dans d'autres l'alternance recherche/consultation est constante tout au long de la session.

Le deuxième (13%) regroupe des sessions marquées par des phases de téléchargement de documents (Figs. 3.9, 3.11, 3.12, 3.13, 3.14, 3.16). On y trouve de la recherche, de la consultation de documents et du téléchargement, avec des variations dans l'ordonnancement de ces actions. La page d'accueil est parfois présente en ouverture.

Enfin, le troisième ensemble regroupe des types de sessions très spécifiques, à faible effectif, comme des sessions (1.25%) où l'utilisateur reste exclusivement sur la page d'accueil Fig. 3.6, ou des sessions (moins de 1%) de navigation exclusive sur les pages de médiation Fig. 3.10. Consulter une page de médiation est donc un comportement spécifique.

3.2 Résultats : modèle de Markov à temps continu (v2)

Dans cette section, nous exploitons le modèle de Markov à temps continu (décrit dans la Section 3.1.4). Les données du 15 Mai 2016 au 15 Juin 2016 ont été utilisées pour illustrer le modèle. Pour des raisons de lisibilité le nombre de clusters a été limité à 15.

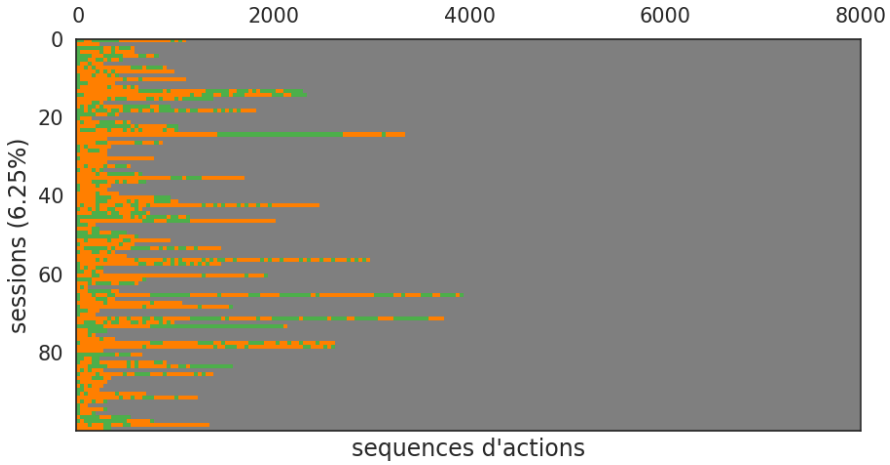

FIGURE 3.18 – Cluster 1 v2

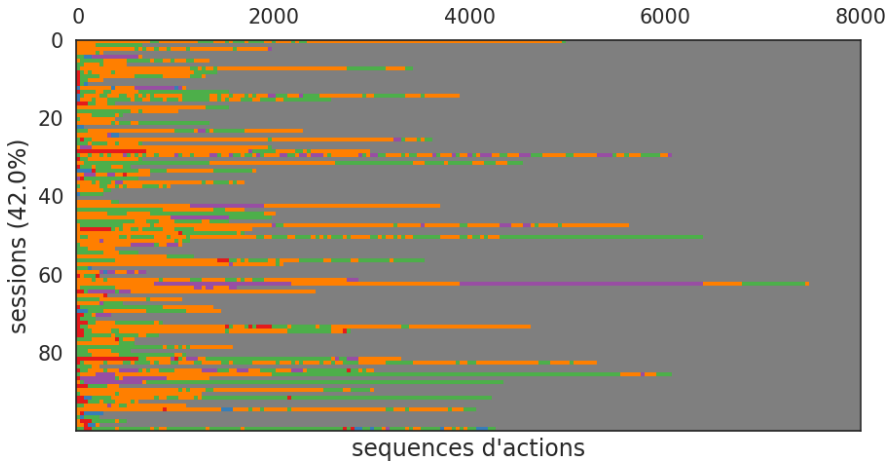

FIGURE 3.19 – Cluster 2 v2

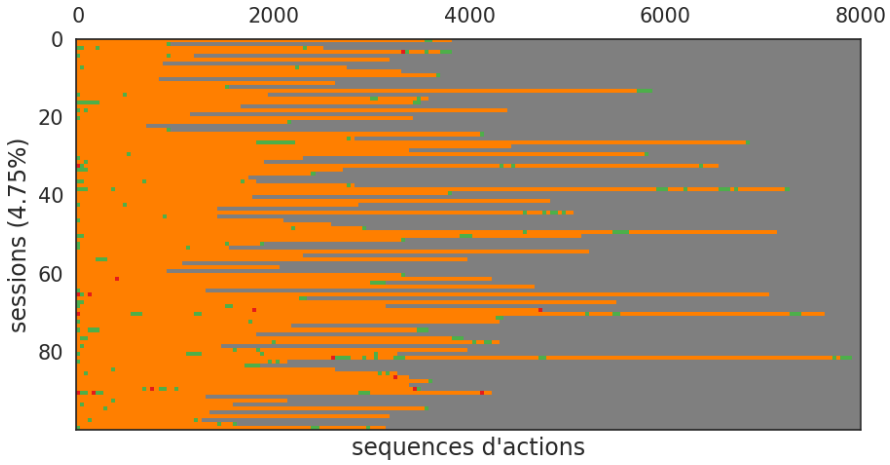

FIGURE 3.20 – Cluster 3 v2

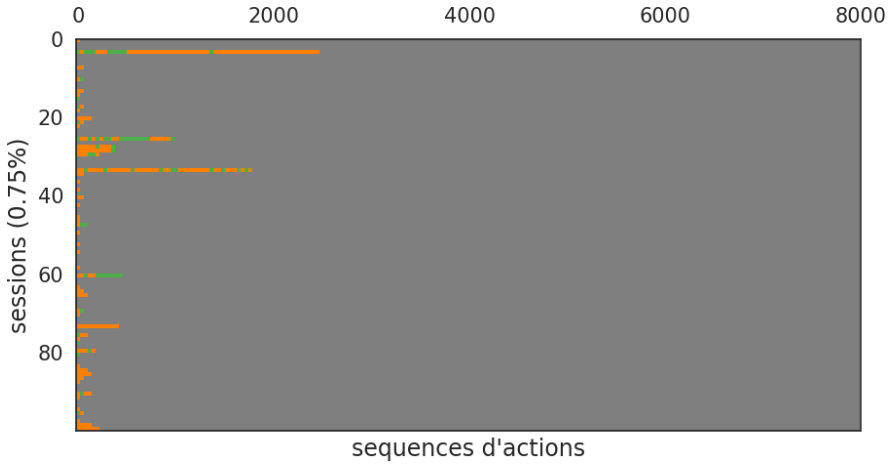

FIGURE 3.21 – Cluster 4 v2

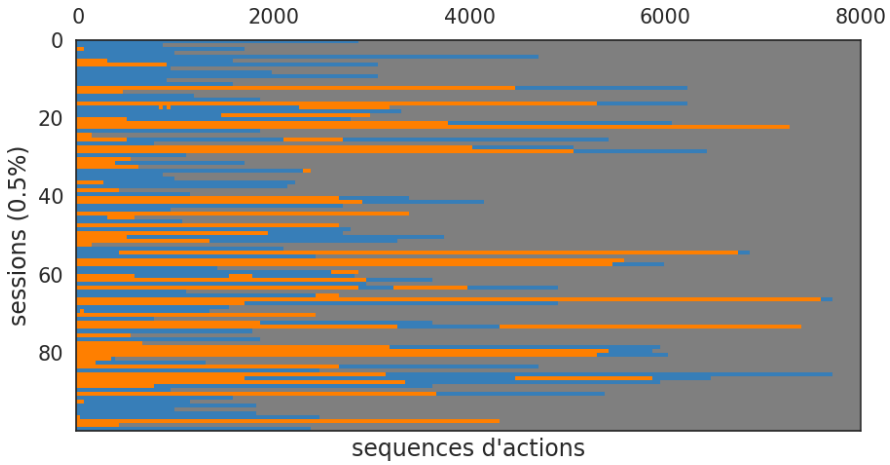

FIGURE 3.22 – Cluster 5 v2

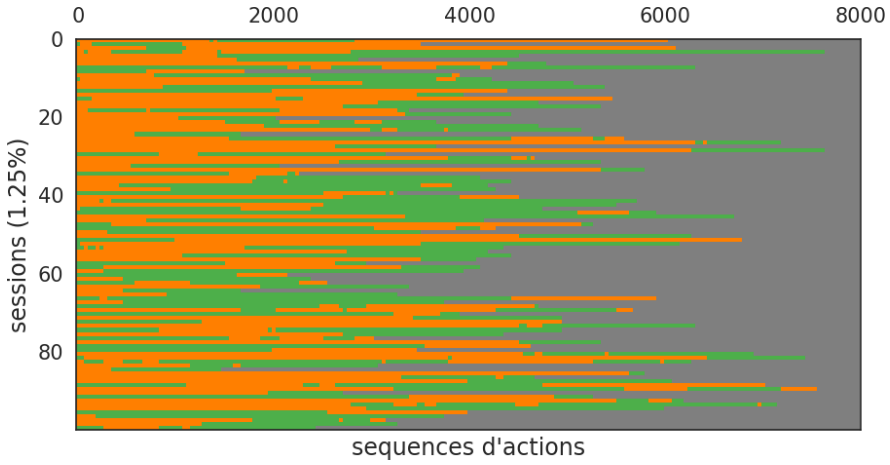

FIGURE 3.23 – Cluster 6 v2

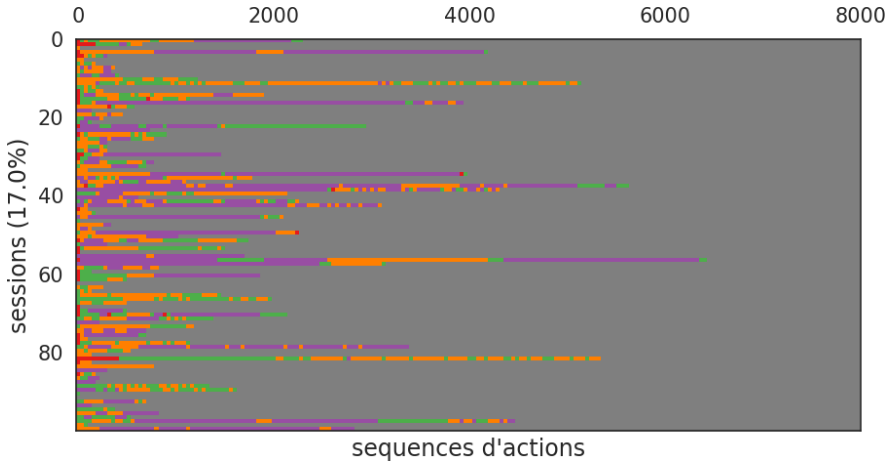

FIGURE 3.24 – Cluster 7 v2

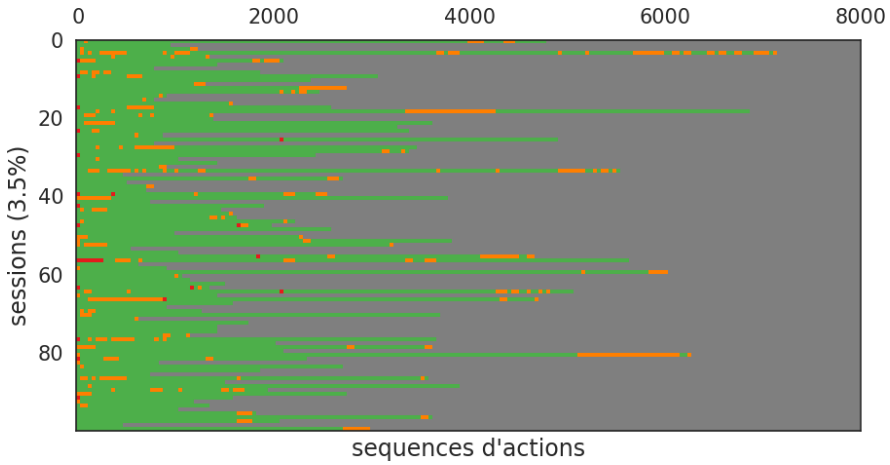

FIGURE 3.25 – Cluster 8 v2

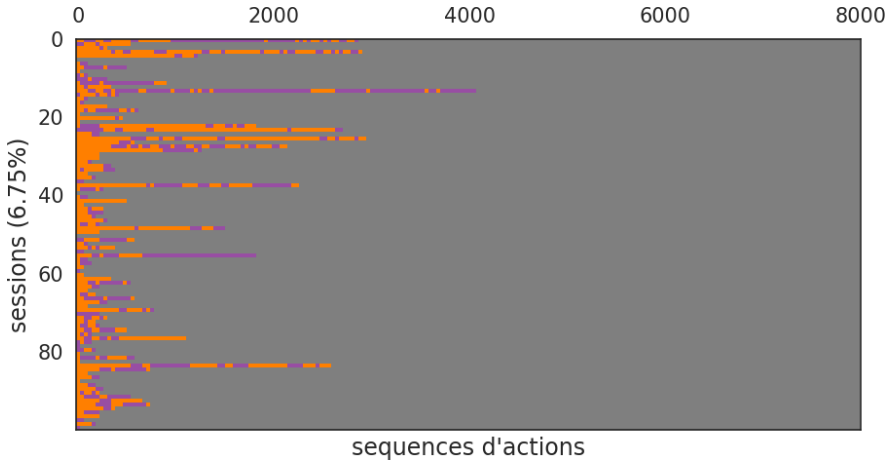

FIGURE 3.26 – Cluster 9 v2

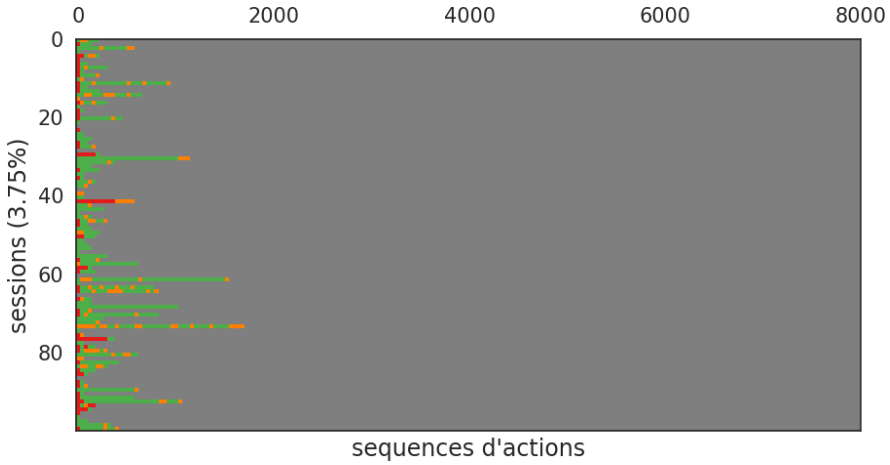

FIGURE 3.27 – Cluster 10 v2

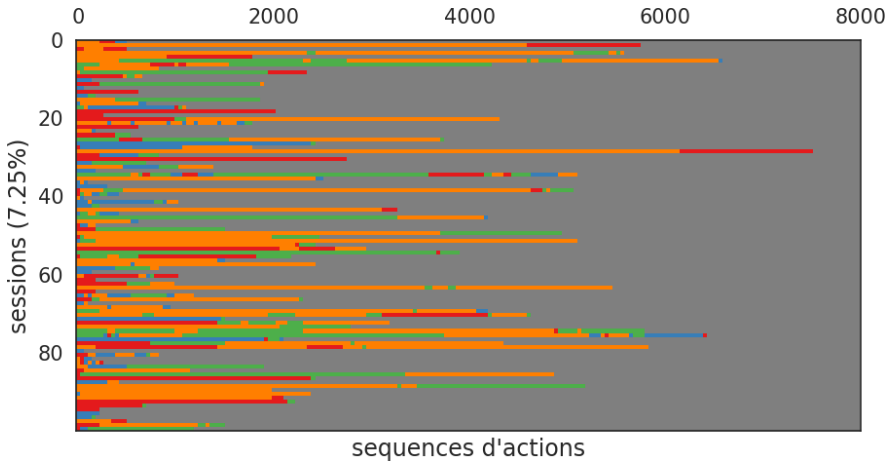

FIGURE 3.28 – Cluster 11 v2

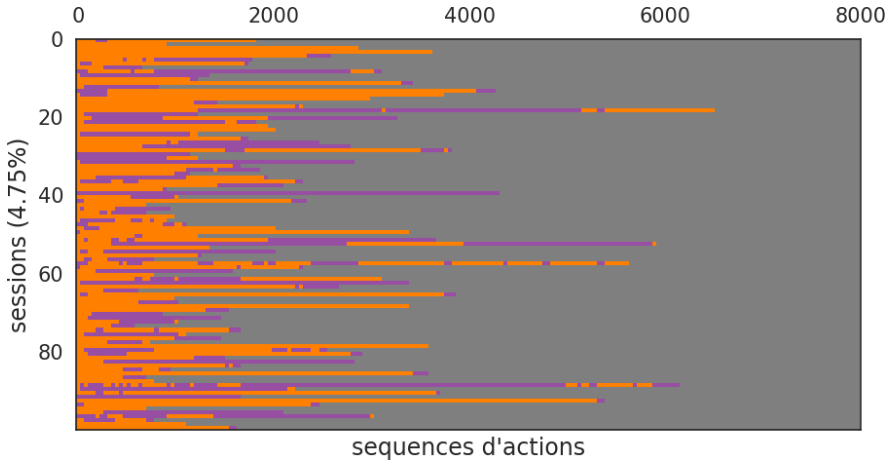

FIGURE 3.29 – Cluster 13 v2

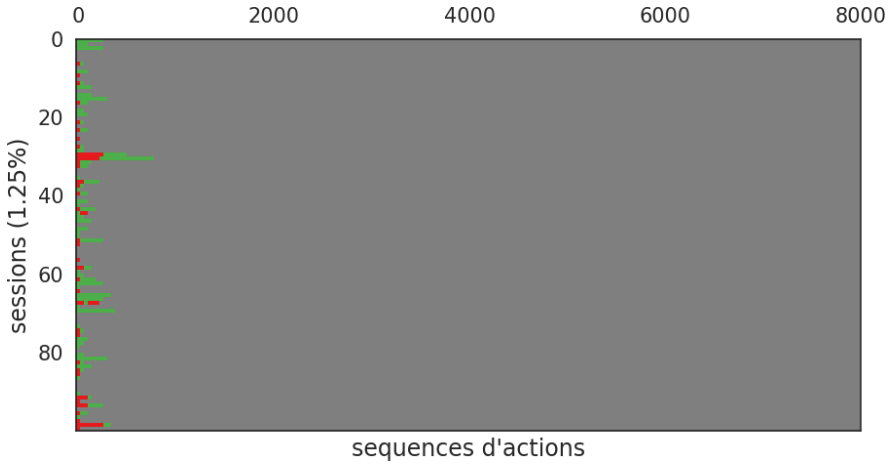

FIGURE 3.30 – Cluster 14 v2

cluster	1	2	3	4	5	6	7	8	9	10	11	13	14
proportion en %	6	42	5	.7	.5	1	17	4	7	4	7	5	1
accueil	0	32	6	0	0	0	12	13	0	26	270	0	25
médiation	0	15	2	0	870	0	0	0	0	0	125	0	0
recherche	30	110	23	6	0	920	30	650	0	40	200	0	50
téléchargement	0	56	2	0	0	0	130	0	33	0	0	380	0
consultation	46	180	980	9	3100	2370	33	35	60	9	940	700	0

TABLE 3.1 – Temps moyen (sec) pour chaque action et chaque cluster. Les clusters 12 et 15 ont été omis puisque le premier est proche du cluster 14 et le second est caractérisé par des usages très courts.

Les figures (3.18-3.30) représentent les clusters de parcours d'usagers modélisés par modèles de Markov à temps continu. On note que le cluster 12 a été omis puisqu'il est très similaire au cluster 14 et n'apporte pas d'information supplémentaire sur les usages de Gallica.

Dans les sessions contenus dans les clusters reportés en figures (3.18, 3.21, 3.23, 3.27) (environ $\sim 12\%$ des sessions) les utilisateurs interagissent uniquement avec les fonctionnalités de recherche et de consultation. Cependant on remarque que les usagers consacrent un temps différent à ces deux actions.

On peut également extraire des usages différents parmi ces clusters. En effet, ces usagers investissent des temps dissemblables pour ces deux actions. Par exemple, pour le cluster figure (3.20) la majorité du temps est captée par la consultation tandis que pour le cluster figure (3.25) la majorité du temps est captée par la recherche. Ces usagers ne passent donc pas par la page d'accueil pour se diriger dans Gallica.

A l'inverse, les sessions formant les clusters, retranscrits en figures (3.19, 3.20, 3.24, 3.25, 3.27, 3.28, 3.30), incluent la page d'accueil au cours de leur parcours sur Gallica. On note cependant, d'après le tableau (3.1) que le temps passé sur la page d'accueil est très court (moins de 30 secondes) à l'exception du cluster figure (3.28) regroupant $\sim 7\%$ des usagers.

Les sessions des clusters en figures 3.24, 3.26, 3.29 incluent en plus de la recherche et de la consultation, le téléchargement de documents.

Les sessions du cluster 3.22 sont centrées sur la médiation ($\sim .5\%$). Le temps moyen de consultation est le plus élevé dans les sessions et peuvent correspondre à des utilisateurs passionnés.

La prise en compte du temps valorise en proportion les sessions commençant par la page d'accueil auparavant *écrasées* par des sessions courtes dédiées à la consultation *directe*. Il en va de même pour la médiation, même si celle-ci reste liée à un unique type de comportement.

3.3 Résultats : modèle de Markov à temps continu par types de documents consultés (v3)

Dans cette section, nous investiguons les parcours d'usagers en termes de types de documents (voir Section 2.1.3). Les actions sont alors redéfinies à partir des types de documents que les usagers peuvent successivement consulter. Les types de documents et leurs couleurs pour la visualisation sont les suivants :

— fascicules :
— monographies :
— objets :
— photographies :
— estampes :
— manuscrits :
— cartes :
— images :
— partitions :
— audio :
— dessins :
— titres fascicules :
— autres :

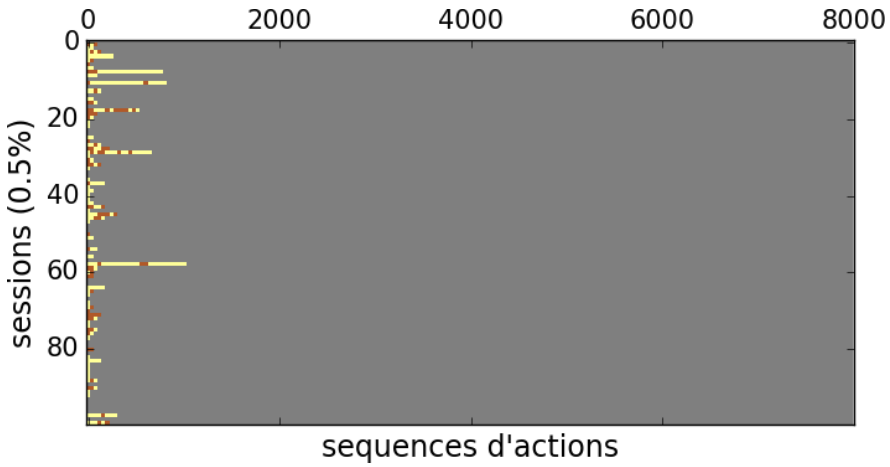

FIGURE 3.31 – Cluster 0 v3

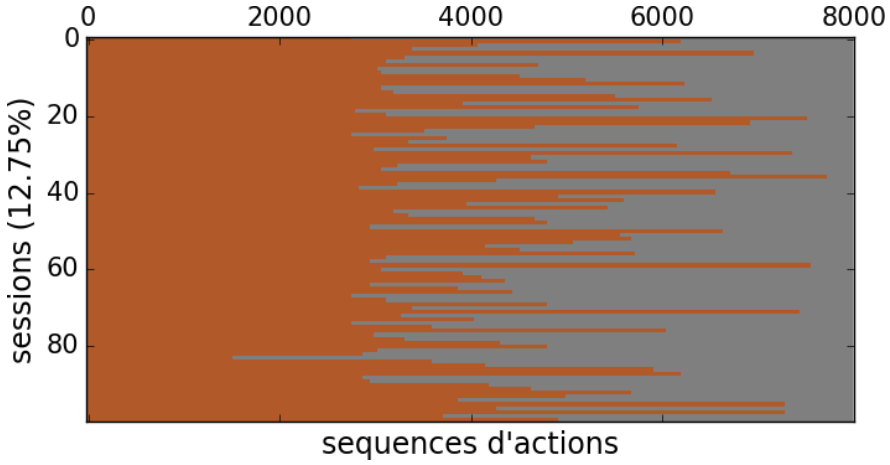

FIGURE 3.32 – Cluster 3 v3

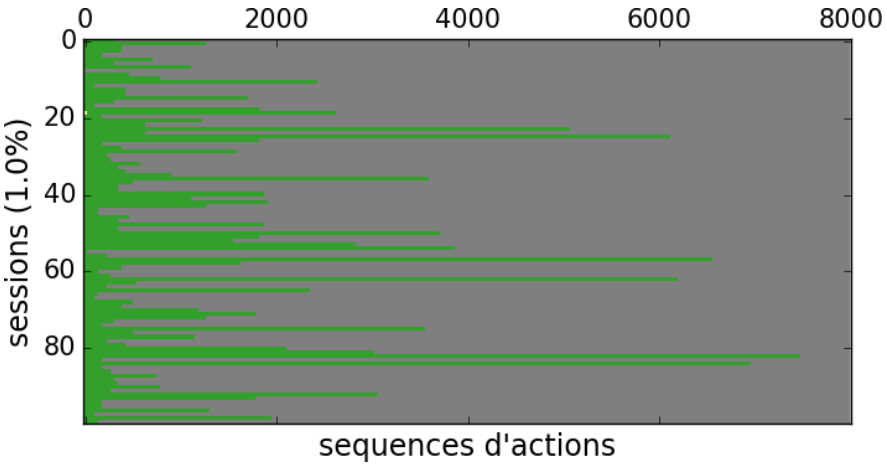

FIGURE 3.33 – Cluster 4 v3

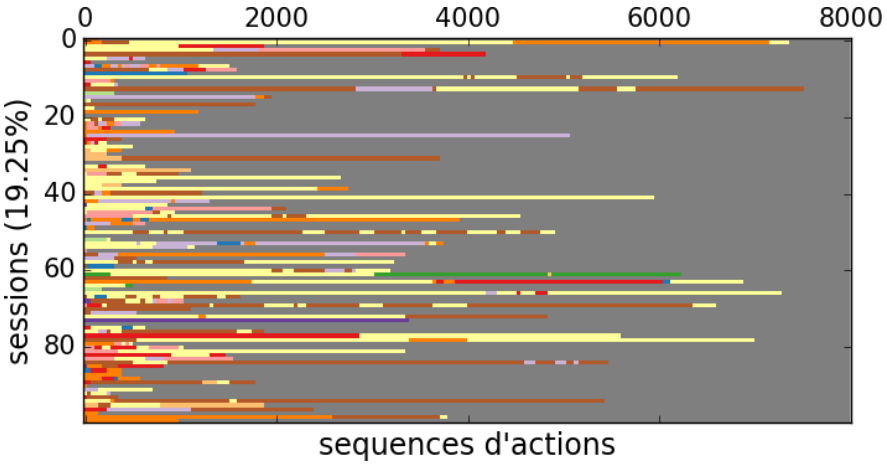

FIGURE 3.34 – Cluster 5 v3

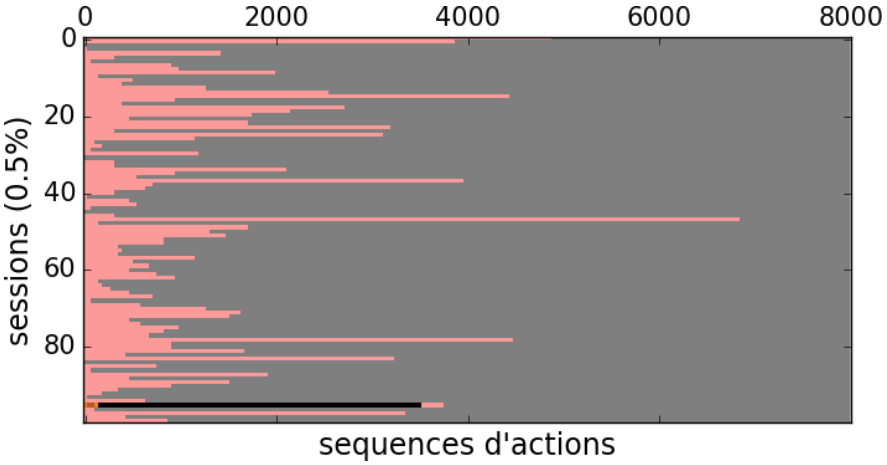

FIGURE 3.35 – Cluster 8 v3

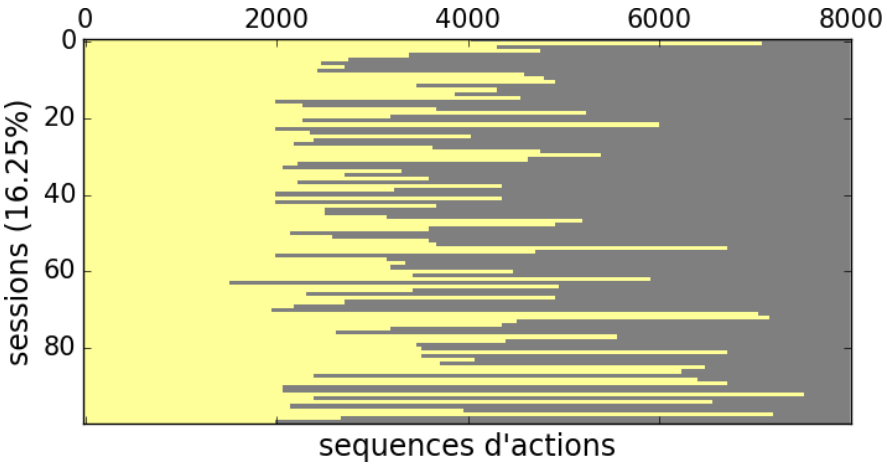

FIGURE 3.36 – Cluster 10 v3

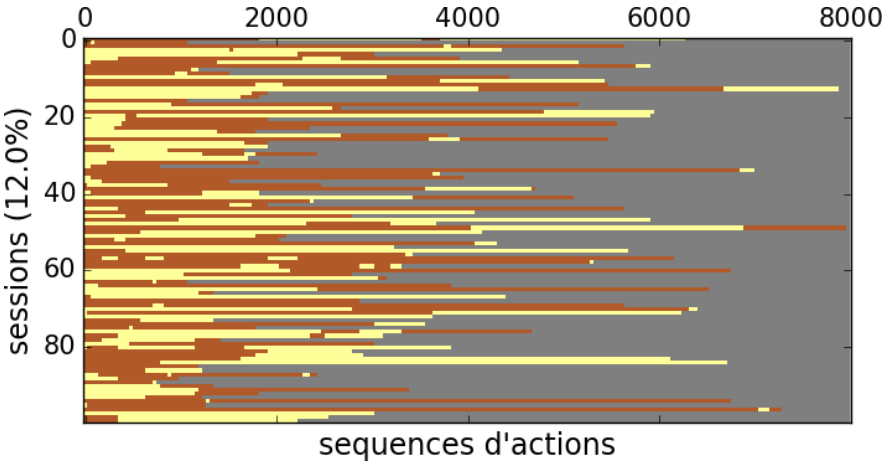

FIGURE 3.37 – Cluster 11 v3

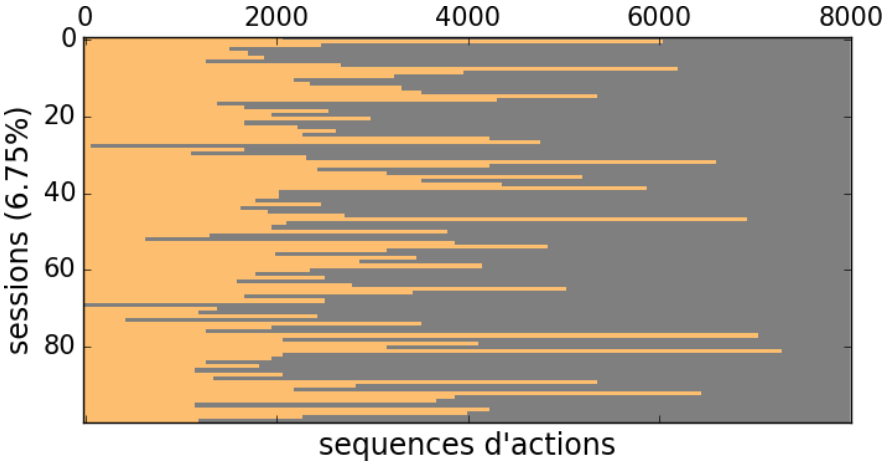

FIGURE 3.38 – Cluster 12 v3

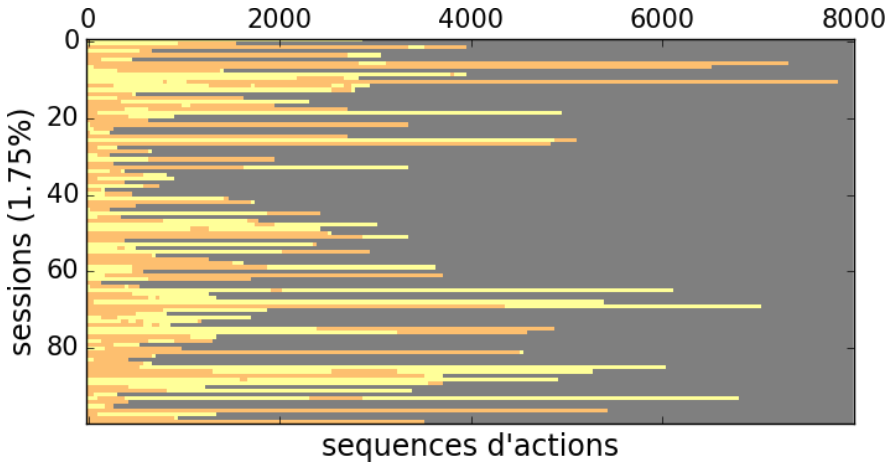

FIGURE 3.39 – Cluster 13 v3

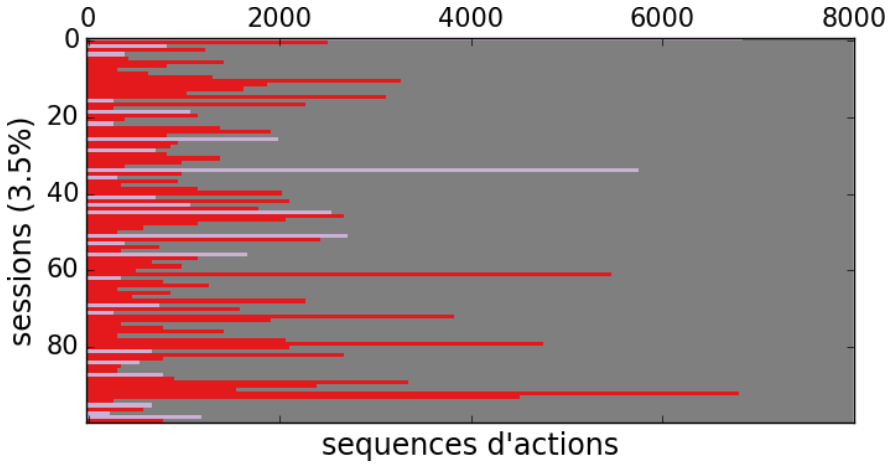

FIGURE 3.40 – Cluster 14 v3

L'ensemble de figures de 3.31 à 3.40 représente les clusters de parcours documentaires dans les sessions d'accès à Gallica. Une majorité de clusters (représentant $\sim 37\%$ des sessions) contiennent un unique type de document (voir figures (3.32, 3.33, 3.36, 3.38)), attestant qu'une grande partie des usagers de Gallica n'ont pas au cours d'une même session une consultation diversifiée en terme de type de document. $\sim 18\%$ des sessions des consultations documentaires, figure (3.31, 3.37, 3.39, 3.40) alternent entre deux types de documents (ex : fascicules-monographies ou monographies-manuscrits). Le cluster représenté en figure (3.34), regroupant $\sim 20\%$ des sessions, est le seul cluster qui inclut des consultations de documents diversifiés en termes de type.

clusters	accueil	médiation	recherche	téléchargement	consultation
2 : monographies	22	32	43	26	32
3 : fascicules	28	78	451	603	1595
4 : partitions	20	39	121	177	224
5 : tout	31	92	319	265	609
10 : monographies bis	30	49	188	518	983
11 : fascicules/mono	29	63	519	365	937
12 : manuscrits	22	46	70	380	65
13 : manuscrits/mono	28	44	162	308	413
14 : photo/cartes	24	61	140	172	296
Ensemble des usagers	28	70	234	332	552

TABLE 3.2 – Temps médian de chaque action pour chaque cluster par types

clusters	acceuil	médiation	recherche	téléchargement	consultation
2 : monographies	14	21	28	17	21
3 : fascicules	1	3	16	22	58
4 : partitions	3	7	21	30	39
5 : tout	2	7	24	20	46
10 : monographies bis	2	3	11	29	56
11 : fascicules/mono	2	3	27	19	49
12 : manuscrits	4	8	12	65	11
13 : manuscrits/mono	3	5	17	32	43
14 : photo/cartes	3	9	20	25	43
Ensemble des usagers	2	6	19	27	45

TABLE 3.3 – Pourcentage du temps médian de chaque action pour chaque cluster par types

Les tableaux (3.2, 3.3) mettent en parallèle la catégorisation des sessions par type de document avec le temps passé à réaliser chaque action. Les temps médians des actions réalisées par les sessions de chacun des clusters (v3) sont reportés dans le tableau (3.2). On remarque par exemple que les usagers qui consultent exclusivement des fascicules (cluster 3 : “fascicules”) ont en moyenne des consultations plus longues que l’ensemble des utilisateurs de Gallica.

Notons qu’il est plus confortable de considérer ces temps en pourcentage. C’est pourquoi le tableau (3.3) contient donc le pourcentage du temps médian de chaque action pour chaque cluster par type. Dans ce cas, chaque ligne somme à 100%. Les usagers appartenant aux clusters de “fascicules” et de “monographies” (clusters 3 et 10) ont tendance à avoir des sessions majoritairement centrées sur la consultation. A l’inverse, les sessions qui ne consultent que des manuscrits (cluster 12) consacrent très peu de temps à la consultation et font largement appel à la fonctionnalité de téléchargement des documents.

Chapitre 4

Analyse des usages documentaires

4.1 Vue d'ensemble des usages

Pour décrire les usages documentaires de Gallica, nous proposons d'extraire et de présenter quelques analyses globales sur les consultations documentaires. Nous nous sommes d'abord attachés à extraire les types de documents les plus consultés. Dans ce chapitre, un type fait référence au champ Type des métadonnées (voir Section 2.1.3). Pour l'ensemble des trois analyses conduites dans ce chapitre, nous avons appliqué le traitement suivant :

1. Récupération des logs de connexion entre le 5 et le 30 Mai 2016 inclus,
2. Sessionisation,
3. Suppression des ARKs identiques dans une même session,
4. Récupération des métadonnées pour les ARKs restants.

L'étape 3 est indispensable pour s'assurer qu'un même document n'est comptabilisé qu'une fois dans une session même si plusieurs pages d'un document ont été consultées.

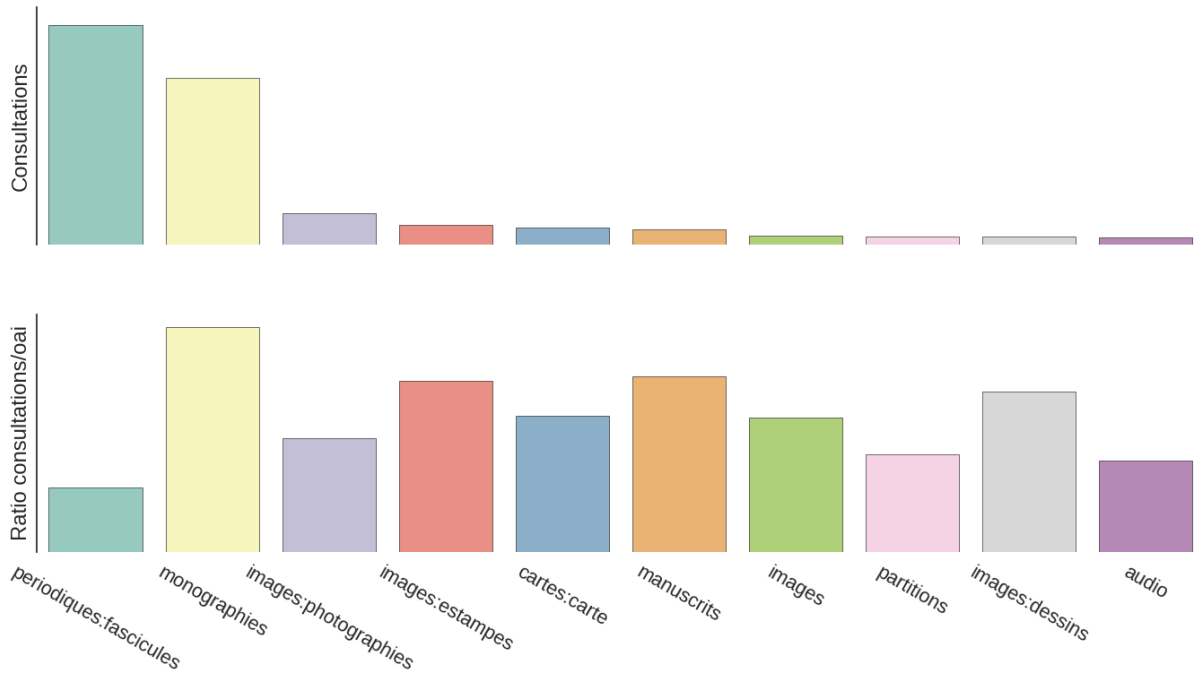

FIGURE 4.1 – Proportion (sans métrique verticale) de consultations pour les 10 principaux types de document.

La figure supérieure de la figure 4.1 représente la proportion dans les consultations des 10 principaux types de documents. On remarque clairement que la majorité des consultations se concentre sur deux types : les périodiques-fascicules et les monographies. Parmi les types de documents populaires, on retrouve les photographies, estampes, cartes et manuscrits. Pour la partie inférieure de la figure 4.1, la consultation des 10 principaux types est rapportée au nombre de documents de ces différents types dans l'entrepôt OAI.

On se rend donc compte que rapportée à la quantité présente dans le catalogue, la consultation est relativement homogène.

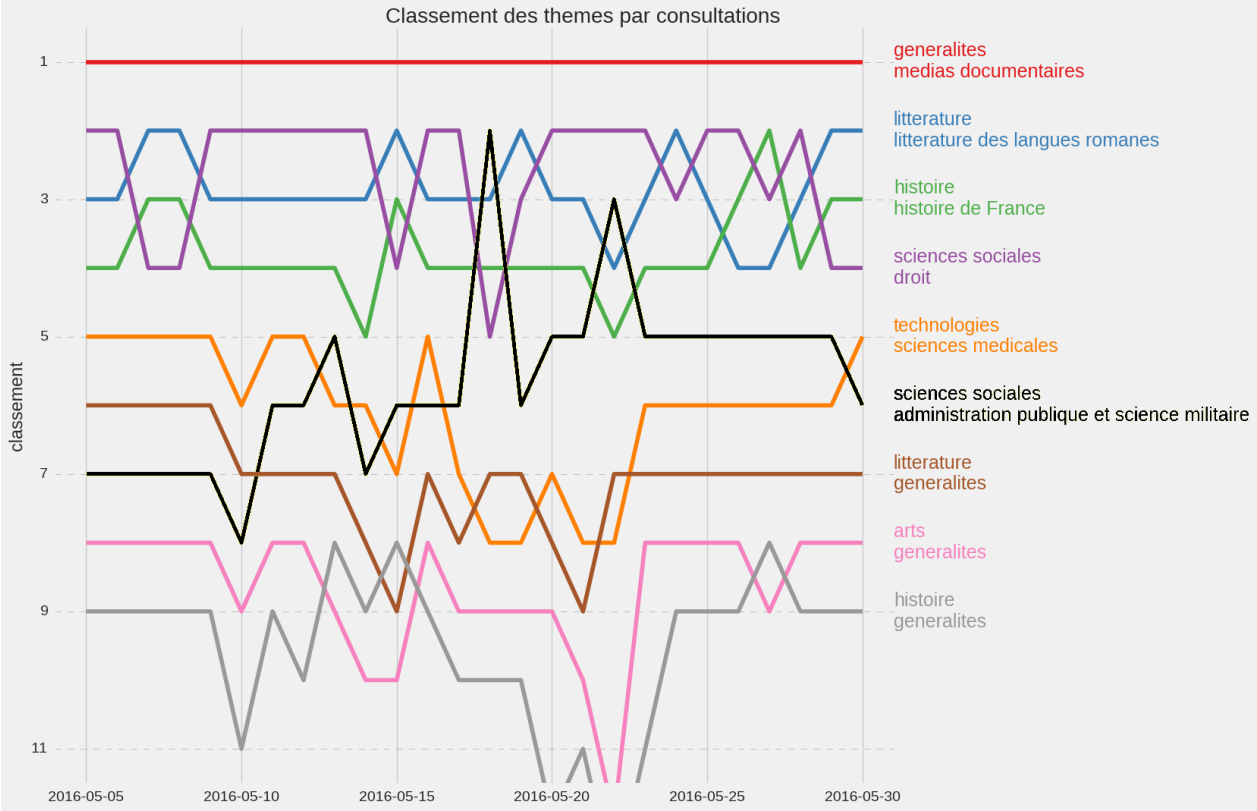

FIGURE 4.2 – Classement journalier des thèmes les plus consultés.

La figure 4.2 est une représentation du classement journalier des thèmes les plus consultés. On observe une variabilité notable dans le classement même si le thème "généralités : médias documentaires" (contenant de la presse) conserve la première place du classement durant tout le mois de Mai. "Littérature des langues romanes", "histoire de France" et "droit" viennent juste après et leurs positions oscillent entre la 2ème et la 5ème place. D'autres thèmes connaissent des fluctuations temporelles plus fortes, comme "sciences sociales : administration publique et science militaire", en jaune, qui varie de la 2ème à la 8ème place.

Pour la liste complète des thèmes de Gallica, nous renvoyons le lecteur à la page web suivante `http://gallica.bnf.fr/FromHomeToThemes`.

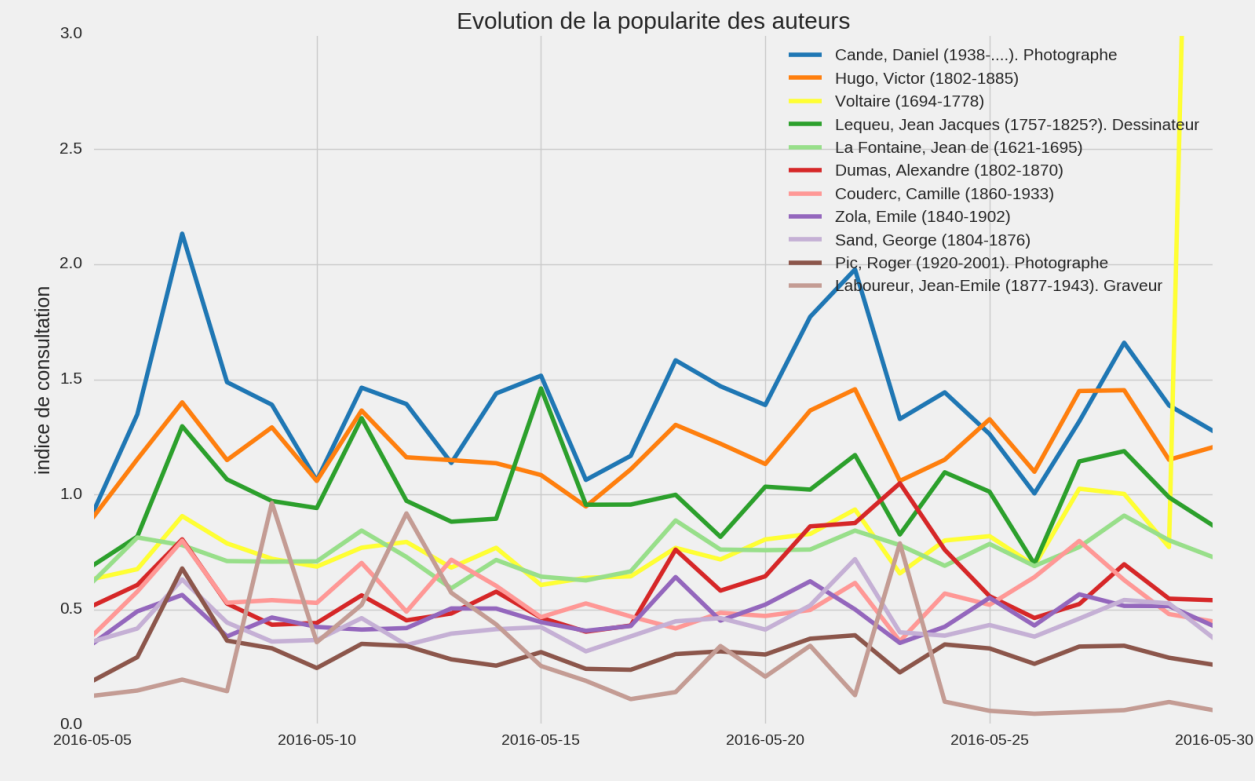

FIGURE 4.3 – Évolution de la popularité d'auteurs au cours du mois de Mai.

La figure 4.3 décrit l'évolution de la popularité d'auteurs pris parmi les auteurs les plus consultés de Gallica au cours du mois de mai, dont on a retiré les auteurs institutionnels tels que "La ville de Paris", "France, sénat" ou les agences de presse telle que l'agence "Meurisse". L'indice de consultation figurant en ordonnée de la figure Fig. 4.3 est le ratio du nombre de consultations d'œuvres de l'auteur rapporté au nombre de sessions uniques.

La popularité surprenante de Roger Pic et de Jean-Emile Laboureur s'explique par la numérisation récente des œuvres de ces derniers qui se retrouvent donc mise en avant sur la page d'accueil de Gallica. La page d'accueil de Gallica est donc un facteur qui influence les consultations des usagers.

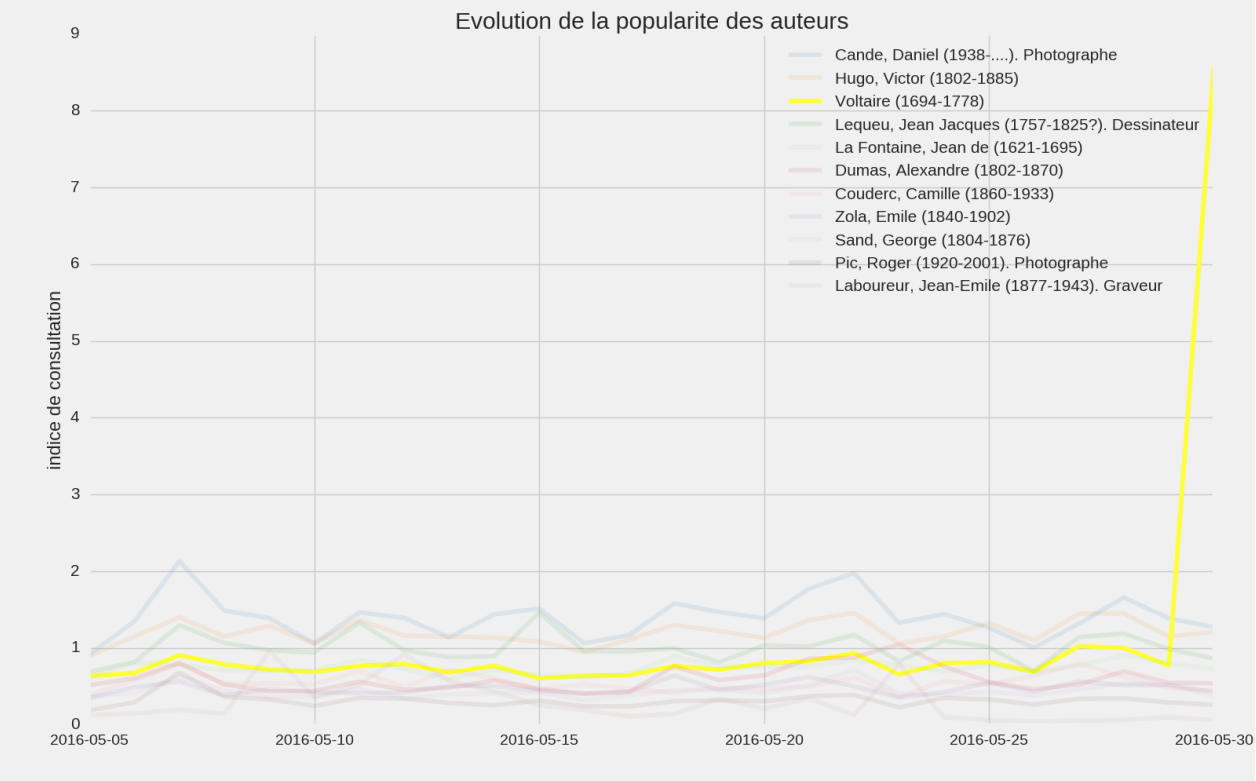

FIGURE 4.4 – Évolution de la popularité de Voltaire au courant du mois de Mai

La figure 4.4 est similaire à la figure 4.3 mais elle est centrée sur Voltaire : elle montre un pic de consultation le 30 mai, date anniversaire de sa mort (30 Mai 1778). Cette figure illustre l'effet d'événements particuliers sur les consultations au sein de l'interface de Gallica. Notons qu'un post sur la page Facebook de Gallica le 30 Mai 2016 a eu un grand succès et pourrait expliquer ce pic.

4.2 Diversité dans les consultations

Une première piste pour étudier la diversité des sessions en fonction des types de documents consultés consiste à s'appuyer sur une classification des documents existante, celle de l'entrepôt OAI. L'idée est d'associer le type du document (monographie, fascicules...) à chaque document consulté dans la session. Pour cette étude, nous nous sommes limités aux sessions ayant accédé à au moins 5 documents uniques (soit $\sim 8\%$ des sessions voir Section 2.4). En effet, il nous semble difficile de conclure sur la diversité des usages à partir de sessions qui consultent peu de documents. Même si ces sessions ont été écartées de l'analyse, nous avons remarqué que parmi les sessions comportant entre 2 et 4 documents uniques consultés 87% sont mono-types à savoir que les documents consultés sont tous du même type.

Afin d'observer la diversité dans les consultations, nous proposons la procédure suivante :

1. Récupération des logs de connexion entre le 1 et le 15 Juin 2016 inclus.
2. Sessionisation
3. Suppression des ARKs identiques dans une même session
4. Récupération des métadonnées pour les ARKs restants (et donc des types)
5. Calcul de la fréquence des types pour chaque session
6. "Clustering" par "k-means" (partitionnement n observations en k clusters en affectant à chaque observation le cluster dont le centre est le plus proche) sur une échelle logarithmique.

L'algorithme "k-means" ou k-moyennes est une méthode de partitionnement de données ("clustering") et un problème d'optimisation combinatoire. Étant donnés des données et un entier k, le

"k-means" divise les points en k groupes distincts ("clusters"). Il est affecté à chaque donnée le "cluster" dont la moyenne est la plus proche.

Plus formellement, étant donné un ensemble de point $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, on cherche à partitionner les n points en k ensembles $\mathbf{C} = \{C_1, \ldots, C_k\}$ $(k \leq n)$ en minimisant la distance entre les points à l'intérieur de chaque "cluster" :

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 , \tag{4.1}$$

où $\boldsymbol{\mu}_i$ est la moyenne des points appartenant au "cluster" S_i. En fixant le nombre k de "clusters" on peut itérativement minimiser le problème Equation 4.1 grâce à un algorithme de type EM Jain et al. [1999].

L'idée derrière l'approche proposée est donc de découvrir des "clusters" de sessions dont les fréquences de consultation de chacun des types de documents sont similaires.

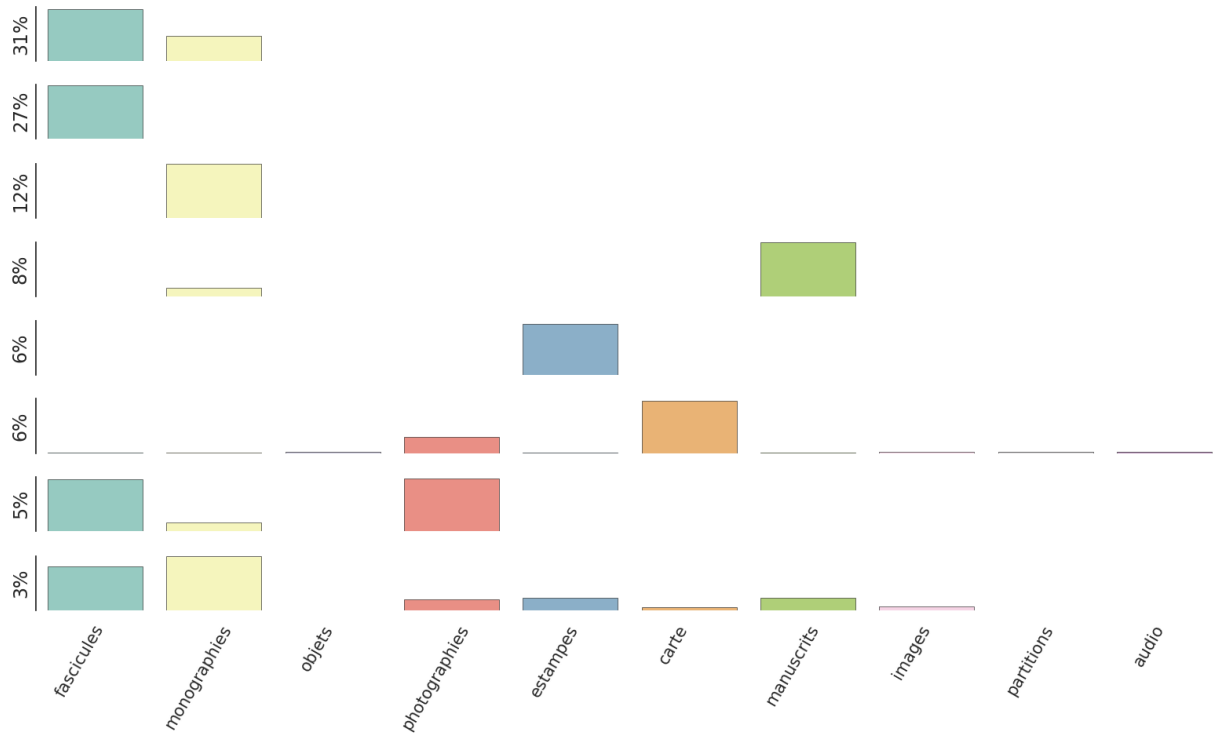

FIGURE 4.5 – Représentation des centroides des 8 "clusters" obtenus par "k-means"

Une fois la procédure appliquée sur des données entre le 1 et le 15 Juin 2016, les centroides des "clusters" ont été représentés en figure 4.5. Les centroides sont des vecteurs de taille 10 contenant les fréquences moyennes de consultation des 10 types qui sont résumées par les *barres*. Le pourcentage fourni en ordonnée correspond à la proportion des sessions sélectionnées appartenant à chacun des "clusters". On remarque d'abord que 3 des 8 "clusters", soit $\sim 45\%$ des sessions à plus de 5 documents, sont mono-types : dans 27% des sessions, les consultations ne portent que sur des fascicules, dans 12% que sur des monographies, dans 6% que sur des estampes. Globalement, on note que les sessions sont peu diversifiées au niveau des types de documents consultés. Il y a cependant des associations privilégiées : les sessions qui mêlent fascicules et monographies représentent 31% des sessions ; les manuscrits sont associés à des monographies dans 8% des sessions et les cartes aux photographies dans 6%. Seul un "cluster" qui représente 3% des sessions fait apparaître des consultations réparties entre de multiples types.

4.3 Vers une classification plus fine des documents

Les métadonnées décrites en Section. 2.1.3 ont des champs facultatifs. Certains champs comme le Thème ou le Corpus auraient été des candidats idéaux pour une classification plus fine des documents du corpus de Gallica. Cependant comme certains documents ne possèdent pas ces informations, nous proposons une approche fondée sur le modèle de "Bag of Words" dans le but de découvrir une classification plus fine des documents.

4.3.1 Modèle "Bag of Words"

Aussi appelé sac de mots, le modèle "Bag of Words" est une représentation de documents (image, texte ...) très populaire en traitement automatique du langage naturel (voir Def. 4.3.1).

Definition 4.3.1. Le traitement automatique du langage naturel est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain Charniak [1985].

Considérons un dictionnaire de mots, la version la plus simple du modèle "Bag of Words" représente un document particulier par l'histogramme des occurrences des mots le composant (ainsi chaque mot du document se voit affecté le nombre de fois qu'il apparaît). On associe alors à un document un vecteur de la même taille que le dictionnaire, dont la composante i indique le nombre d'occurrences du $i-eme$ mot du dictionnaire dans le document.

4.3.2 Allocation de Dirichlet Latente (LDA)

L'Allocation de Dirichlet Latente dont l'acronyme est LDA (de l'anglais Latente Dirichlet Allocation) est un modèle probabiliste génératif qui permet la catégorisation/classification automatique et non-supervisée de *documents*. L'idée du LDA consiste à considérer les *documents* comme des mélanges aléatoires de *topics*, où chaque *topic* est caractérisé par une distribution sur les mots. Les mots sont ici considérés comme des réalisations de variables aléatoires observées tandis que les *topics* sont vus comme des variables latentes et donc non observables. Le but est donc à partir d'un *corpus* de *documents* de découvrir des *topics* cachés.

Dans un premier temps, le LDA consiste à tirer un vecteur de poids des *topics*, selon une distribution de Dirichlet, qui décrit quels *topics* ont la plus forte probabilité d'apparaitre dans un *document*. Puis, pour chaque mot appartenant au *document*, on choisit un unique *topic* à partir du vecteur de poids. La génération du mot est réalisée en "tirant" depuis la distribution conditionnée par le *topic* choisi. On note que, d'après cette procédure, les mots d'un *document* sont donc générés par des *topics* différents et choisis de manière aléatoire.

En faisant l'hypothèse de ce modèle génératif pour une collection de *documents*, le LDA essaie donc, à partir des *documents*, de trouver un ensemble de *topics* susceptibles d'avoir généré la collection.

Le formalisme mathématique, notamment l'estimation des différents paramètres du LDA n'a pas été développé dans ce rapport. Nous invitons le lecteur intéressé à consulter les travaux de références suivants Steyvers and Griffiths, Blei et al. [2003], Minka and Lafferty [2002], Heinrich [2004].

La figure 4.6 illustre la modélisation des *topics* et *documents* par le LDA. Ici le *topics* k est une distribution sur les mots tandis que le *document* est une distribution sur les *topics*. Dans cette illustration, les *topics* pour le *document* d sont nommés, ce que ne produit pas directement l'algorithme LDA. Les noms (politics, sports, news, economy, war) sont déduits à partir de la distribution sur les mots et requiert donc l'intervention humaine pour nommer les *topics*.

4.3.3 Clustering des documents du catalogue de Gallica

Pour la classification non supervisée des documents référencés dans l'entrepôt OAI de Gallica, nous nous appuyons sur les métadonnées de ces derniers. Ainsi les *documents*, au sens du LDA, décrits

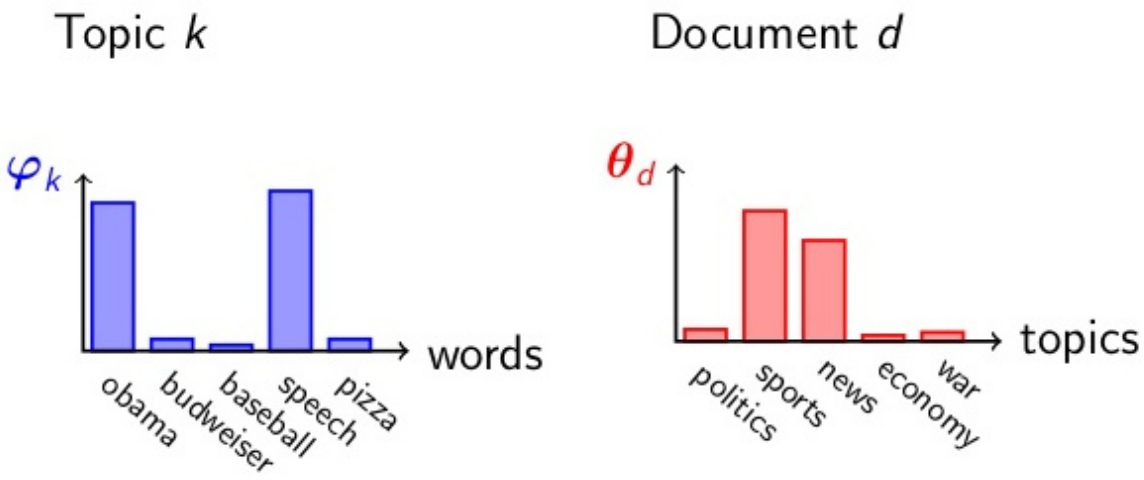

FIGURE 4.6 – Illustration de la modélisation des *topics* et *documents* par le LDA

dans la sous-section précédente sont ici la concaténation des différents champs des métadonnées présents dans l'entrepôt OAI. Un pré-traitement de ces métadonnées est nécessaire pour rendre pertinents les résultats de l'algorithme LDA. Dans un premier temps une étape aboutissant à la suppression des ponctuations et à une segmentation des notices OAI en vecteurs de mots est requise. Ensuite les mots les plus communs de la langue française comme les prépositions, pronoms ou autres verbes courants (être, avoir...) sont retirés des métadonnées. Cette étape est nécessaire pour une meilleure interprétation des *topics* estimés. En effet, ces mots communs ne sont pas informatifs et ne permettent donc pas de déduire une nomenclature pour les *topics*.

Comme expliqué dans la sous-section précédente, les *documents* sont donc décrits par un ensemble de *topics* avec une certaine proportion. Cependant un *document* possède un *topic* principal (avec la plus forte probabilité) à partir desquels nous avons classifié les documents. Les *documents* possédant le même *topic* principal sont regroupés dans un classe commune.

Une illustration de cette méthode appliquée aux notices en utilisant tous les champs disponibles est détaillée à l'appendice D, pour une segmentation en 20 classes. Nous avons pris ici ce cas d'un nombre réduit de classes (pour $800,000$ notices tirées au hasard pour des raisons de faisabilité numérique) pour permettre un affichage lisible des résultats.

Il est intéressant de constater que les classes reproduites se regroupent principalement sur la présence de termes "techniques" (type de document, présence de mots relatifs au type de média : "cm" pour les cartes/estampes/disques..., "agence" pour les photographies, "vol." pour les monographies, "presse"). Il est en effet inévitable que ces mots apparaissent en bonne place puisqu'ils discriminent les documents les uns des autres dès lors qu'ils sont fréquents dans une proportion significative de notices et absents des autres.

Cela peut apparaître comme une limite de la méthode appliquée à ce corpus. Nous verrons dans le paragraphe 7.1.2 plusieurs directions possibles permettant d'améliorer ces résultats.

Chapitre 5

Provenance des usagers

5.1 Protocole

La provenance des usagers de Gallica est une information essentielle pour comprendre les utilisateurs de la plateforme proposée par la BnF. Dans les logs, le champ "referer" identifie l'adresse de la page web qui est à l'origine de la requête. Ce champ peut donc permettre l'identification de sites référents qui redirigent une partie de leur trafic vers le domaine gallica.bnf.fr. La connaissance de ces sites offre une indication sur la manière dont les Gallicanautes découvrent/arrivent sur Gallica.

Le protocole pour déterminer le site de provenance d'une session est le suivant :

— extraction du champ "referer" de chaque ligne de logs,

— extraction des "referers" des lignes de logs dont la date correspond aux débuts de la session,

— le site de provenance correspond au premier "referer" ne contenant pas gallica.bnf.fr

— suppression des sous-domaines [1] et des extensions des "referers".

La dernière étape de filtrage est essentielle puisque les données de logs sont horodatées à la seconde. Dans ce cas précis, un ensemble de requêtes présentent la même date et est alors ordonnancé aléatoirement. Par exemple lorsqu'un utilisateur fait une requête pour obtenir la page d'accueil son navigateur génère également des requêtes supplémentaires liées au design (css, documents présents sur la page d'acceuil...) qui ont alors gallica.bnf.fr en tant que "referers".

5.2 Présentation des données

Pour étudier la provenance des usagers de Gallica, nous nous sommes basés sur les données comprises entre le 1 Juin 2016 et le 16 Juin 2016 comptabilisant environ $\sim 460,000$ sessions. Les sessions ont été divisées en 3 catégories en fonction de leur nombre d'actions :

1. 1 actions : $\sim 30\%$ des sessions,
2. $2-4$ actions : $\sim 30\%$ des sessions,
3. > 4 actions : $\sim 40\%$ des sessions.

Cette classification a pour but d'isoler les sites référents susceptibles d'être caractéristiques de sessions courtes ou à l'inverse de sessions longues.

1. Un sous-domaine est la partie de l'adresse internet d'un site qui précède le nom de domaine. L'extension d'un nom de domaine est le suffixe situé à droite du nom de domaine après le point.

5.3 Résultats

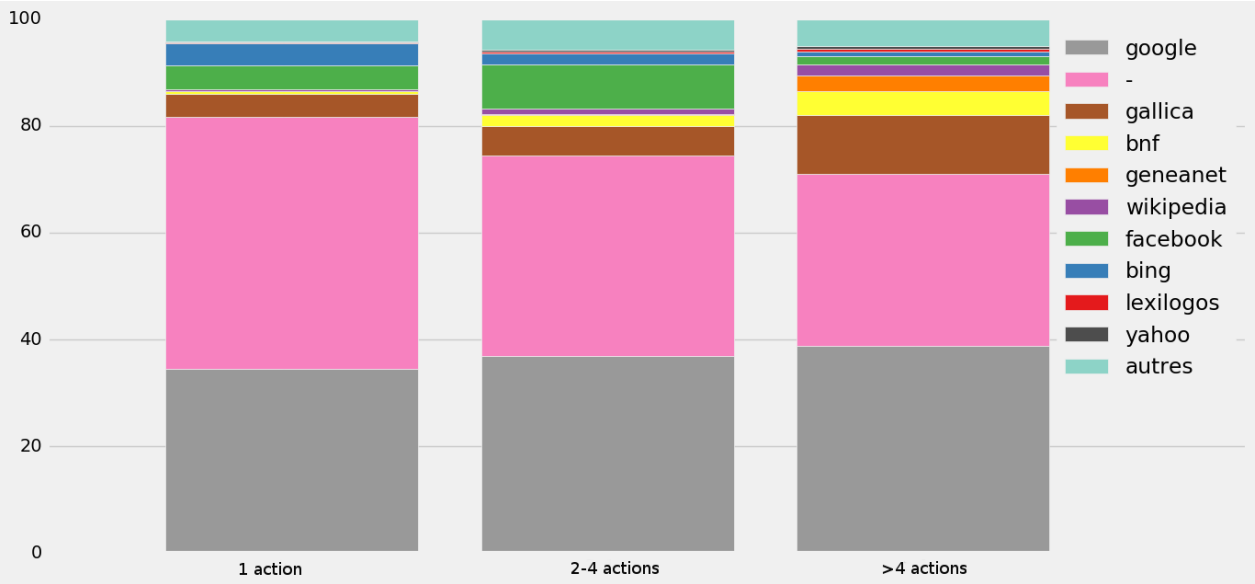

FIGURE 5.1 – Histogramme empilé des sites référents des sessions en fonction de leur nombre d'actions.

La figure (5.1) indique que les 2 principaux sites référents sont "google" et "-" quel que soit le nombre d'actions. Lorsqu'un "referers" n'est pas renseigné par le navigateur de l'utilisateur, Apache inscrit "-" dans les logs. C'est généralement le cas lorsqu'un utilisateur arrive directement sur Gallica en renseignant son URL (gallica.bnf.fr) dans le navigateur. On notera que la proportion de "-" en temps que "referers" ($\sim$ 40%) nous semble particulièrement élevée. Une discussion avec l'équipe technique de Gallica a permis d'identifier : une *perte* du "referer" lorsqu'un usager débute une session par la page d'accueil à cause d'une redirection automatique. Nous avons également remarqué que les générateurs de liens courts (ex : bit.ly), utilisés systématiquement par l'équipe de la médiation sur twitter, offusquent le champ "referer".

La part verte des histogrammes, correspondant au réseau social Facebook, est plus importante dans les sessions relativement courtes (nombre d'actions inférieur à 4 mais supérieur à 1). Cela semble indiquer que les sessions en provenance des réseaux sociaux ne réalisent que peu d'actions sur Gallica. Une étude plus approfondie sur les réseaux sociaux est proposée dans le chapitre 6. A l'inverse, la part jaune liée aux sous-domaines de bnf.fr (catalogue.bnf.fr, data.bnf.fr ou bnf.fr) est plus importante chez les sessions longues. On note également une proportion significative de geneanet qui indique la popularité de Gallica chez les passionnés de généalogie. On peut dire que geneanet est le premier "referer" thématique.

Nous avons également cherché à voir dans les sessions la part des "referers" externes à Gallica. Après filtrage des referers "gallica.bnf.fr" il y a :

— 23% des sessions avec au moins 2 referers externes,

— 30% des sessions avec plus de 5 actions qui ont au moins 2 referers externes,

— 18% des sessions avec plus de 5 actions qui ont au moins 2 "google" comme referers.

Dans la suite de cette section, nous nous sommes concentrés sur les sessions réalisant une unique action. Ces sessions méritent une attention toute particulière puisqu'elles représentent 30% des usagers. Il est difficile d'appréhender ces usages et nous lui avons donc dédié une analyse portant sur les actions et les sites référents. Dans un premier temps, le tableau (5.1) permet d'affirmer qu'une majorité de ces sessions à action unique sont liées soit à la page d'accueil soit à la consultation d'un document.

actions	%
accueil	13
mediation	5
recherche	3
téléchargement	0.3
consultation	79

TABLE 5.1 – Pourcentage des actions pour les sessions à 1 actions.

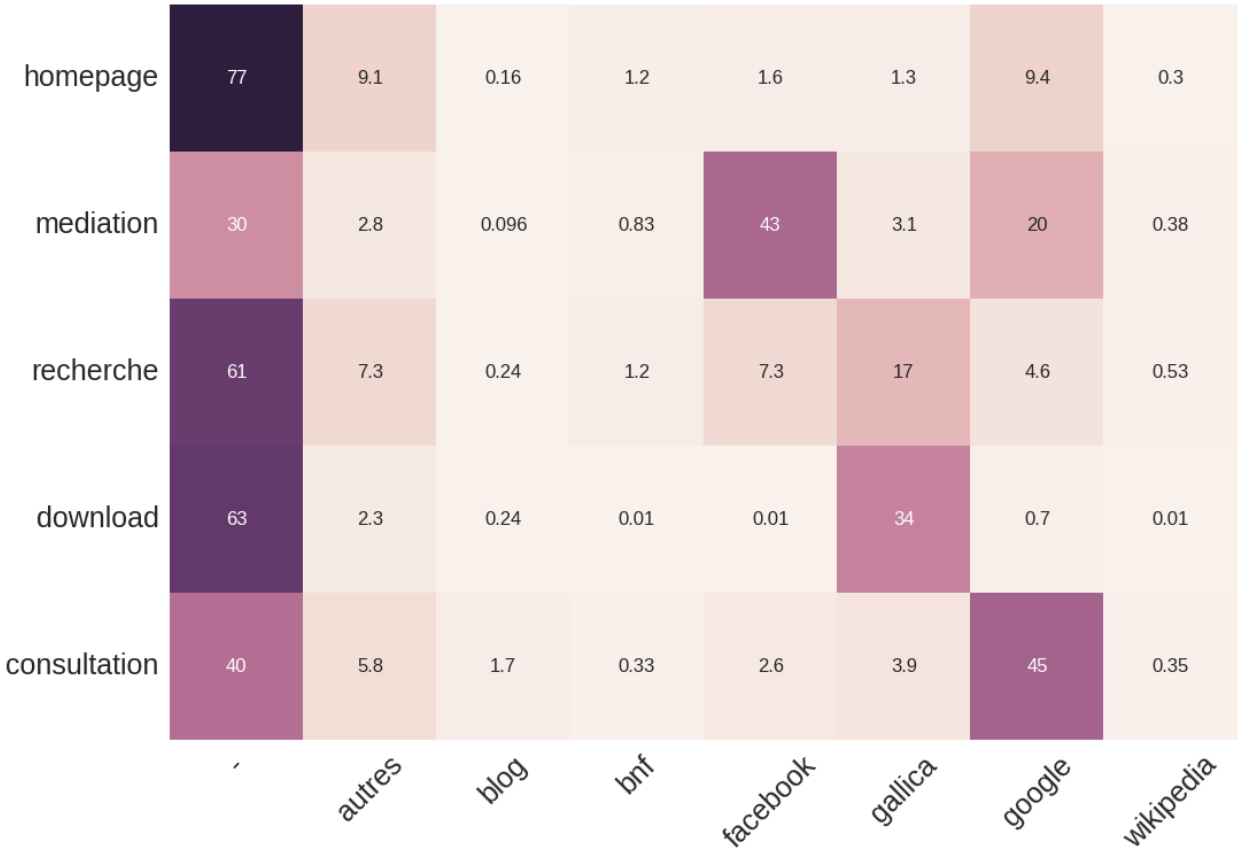

FIGURE 5.2 – Matrice de corrélation actions/référents.

Enfin la matrice de correlation actions/référents figure (5.2) illustre quel site référent est à l'origine de l'unique action réalisée dans ces sessions. On peut donc voir qu'en grande majorité (43%) les usagers ayant une session ne contenant qu'une action de médiation arrivent sur Gallica depuis le réseau social Facebook. De plus, lorqu'une session est définie par une unique consultation cette dernière a 45% de chance d'être en provenance de Google.

Chapitre 6

Impact de la médiation

6.1 Motivation et méthodologie

Le chapitre (5) a montré une proportion significative des réseaux sociaux et particulièrement de Facebook parmi les sites référents. Il nous semble donc intéressant d'approfondir l'impact de la médiation sur le trafic de Gallica. Ce chapitre a pour but d'illustrer et de quantifier l'impact des différentes méthodes de médiation :

— le blog qui regroupe un ensemble d'articles centrés sur un thème précis,

— la page Facebook qui met en avant un unique document pour chaque publication.

Une étape de *moissonnage* est nécessaire pour être capable d'agréger, pour chaque méthode de médiation, l'ensemble des liens renvoyant vers Gallica. A l'aide d'un script Python l'ensemble du contenu des articles du blog a été téléchargé puis seuls les liens contenant "gallica.bnf.fr" ont été extraits. De manière similaire, Facebook met à disposition une API permettant de moissonner la totalité des publications (liens associés, commentaires, nombre de partages...) de la page de Gallica.

Une fois les liens vers les documents de Gallica extraits, il est facile à l'aide des logs de connaître le nombre de consultations de chacun de ces documents.

6.2 Résultats

6.2.1 Blog

L'ensemble des publications du blog de Gallica (entre le 6 Mai et le 25 Juin 2016) contient un total de 142 liens vers un document. Ces documents ont été visités en moyenne par 35 visiteurs uniques.

6.2.2 Facebook

Le *moissonnage* des données de la page Facebook jusqu'au 14 Septembre 2016 a permis de récupérer les informations suivantes pour chaque publication :

— la date de publication,

— le contenu,

— le lien associé,

— le nombre de réactions,

— le nombre de commentaires,

— le nombre de partages.

La page Facebook de Gallica (au 14 Septembre 2016) compte donc :

— 2,237 publications,

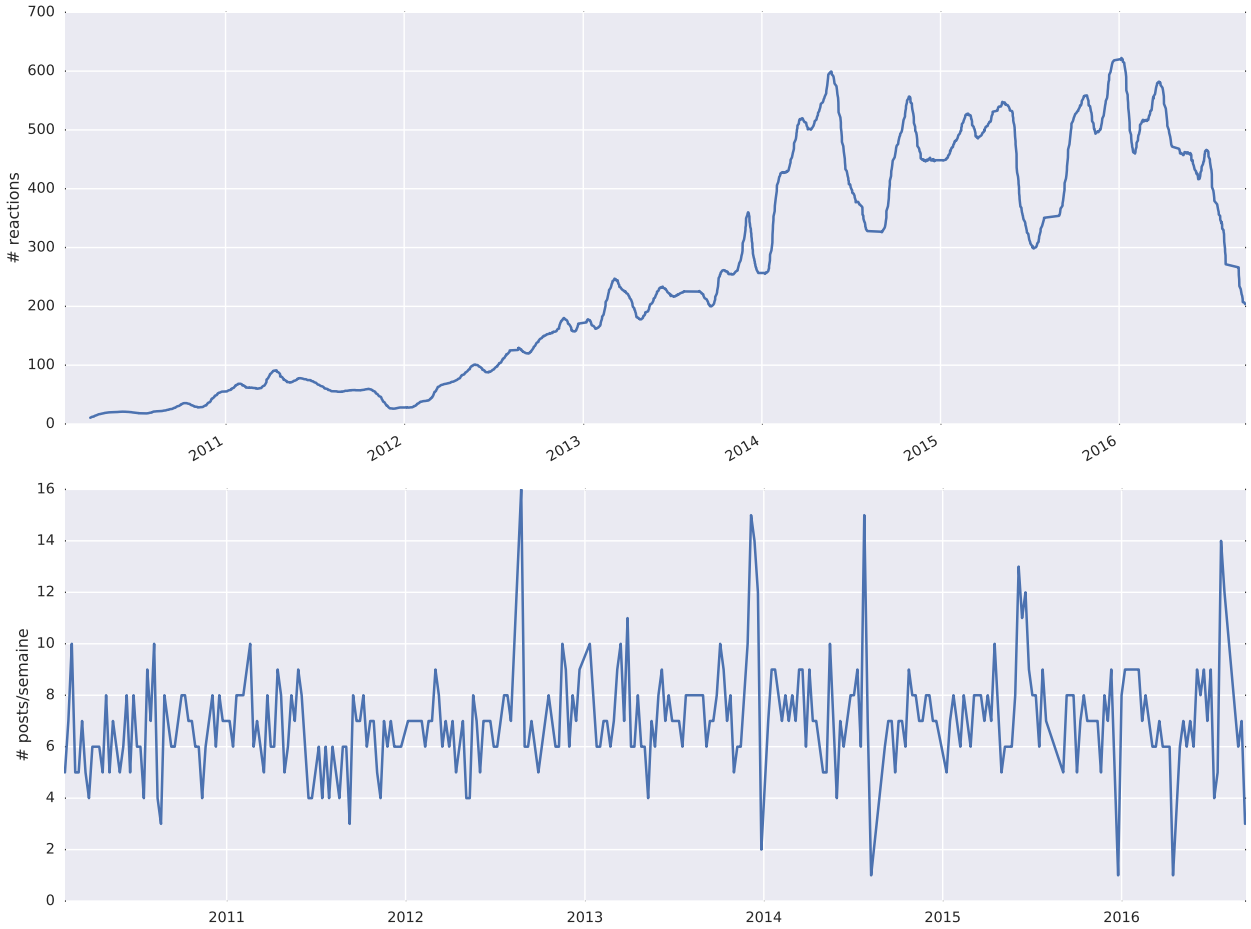

FIGURE 6.1 – Haut : Nombre de réactions en fonction du temps. Bas : Nombre de publications par semaine en fonction du temps.

— 17,900 commentaires ; (soit ~ 8 commentaires/publication),

— 7,236 auteurs uniques de commentaires,

— 571,198 réactions ; (soit ~ 250 réactions/publication),

— 206,486 partages ; (soit ~ 90 partages/publication).

La figure (6.1) montre que le nombre de publications par semaine reste stable (~ 7/semaine) depuis la création de la page en 2011. A l'inverse, le nombre de réactions par publication a significativement augmenté passant de moins de 100 réactions/publication avant 2013 à plus de 300 après 2014. Ce résultat illustre le gain de popularité de la page Facebook ces dernières années (alors que la fréquence de publications reste stable).

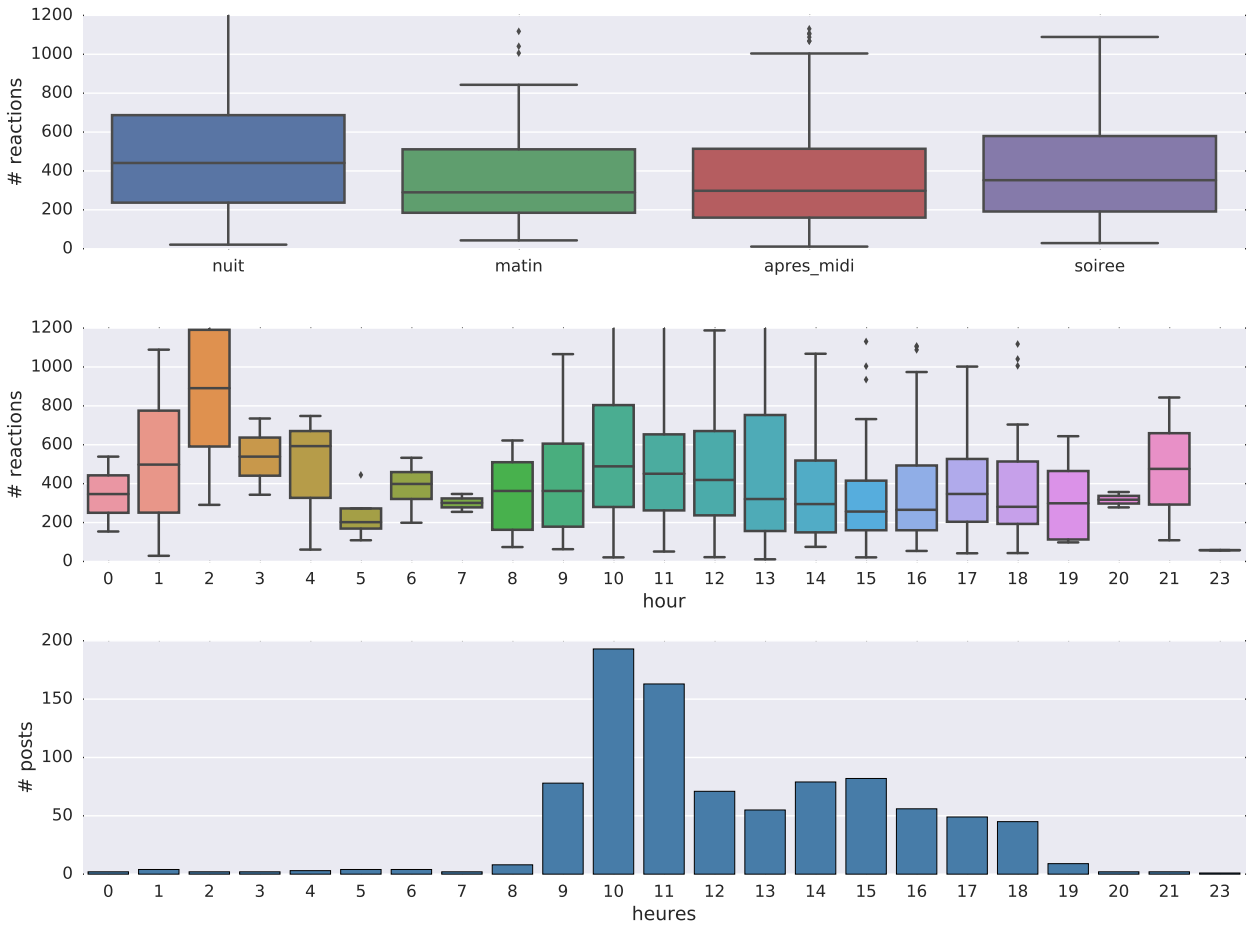

FIGURE 6.2 – Bas : Histogramme du nombre de publications en fonction de l'heure. Milieu : Boxplot du nombre de réactions par publication en fonction du temps. Haut : Boxplot du nombre de réactions par publication en fonction de la période de publication.

La figure (6.2) illustre la répartition de l'heure des publications ainsi que le nombre de réactions qu'elles génèrent. Cette figure permet de conclure que l'heure de publication n'a pas de réel impact sur le nombre de réactions.

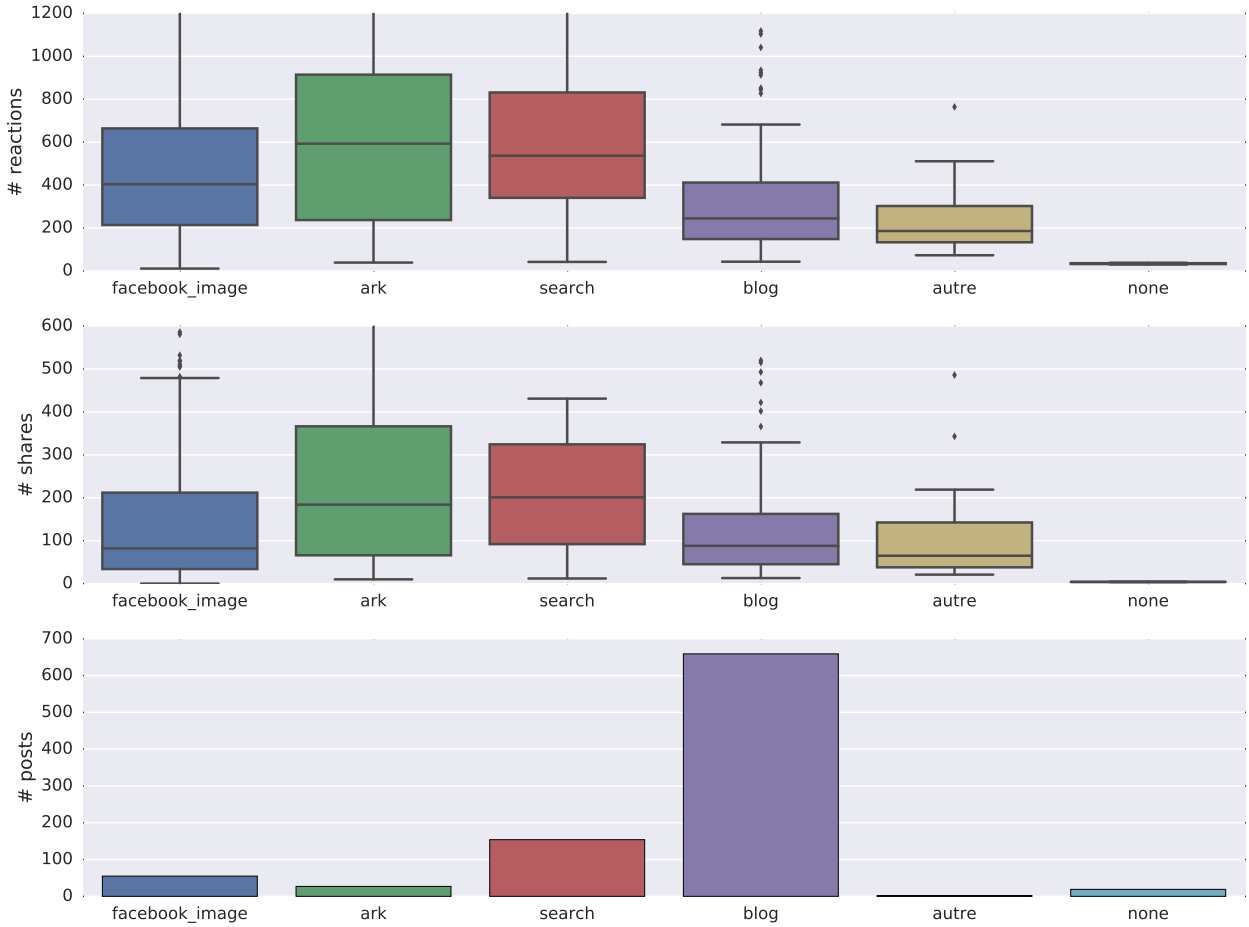

FIGURE 6.3 – Nombre de réactions/partages en fonction du type de liens associé aux publications.

Ensuite, nous avons souhaité vérifier l'impact du type de lien associé aux publications sur le nombre de réactions et de partages.

Il est possible d'associer à une publication Facebook différents types de liens (dénomination employée dans la figure (6.3) :

— "facebook_image" : est une illustration. Lorsque l'utilisateur clique sur ce type de lien, une nouvelle fenêtre du navigateur s'ouvre mais l'utilisateur reste sur le domaine "facebook.com".

— "ark" : lien qui renvoie vers un document spécifique de Gallica. L'utilisateur accède ainsi au document depuis l'interface Gallica et quitte Facebook.

— "search" : lien qui renvoie vers une recherche Gallica.

— "blog" : lien qui renvoie vers une publication dans le blog de Gallica (http ://gallica.bnf.fr/blog)

— "autre" : lien qui renvoie vers une URL sans rapport avec Gallica.

— "none" : aucun lien n'est associé à la publication.

La figure (6.3) montre que les utilisateurs ont tendance à moins partager une publication dont le lien est du type "facebook_image", "blog" ou "autre". A l'inverse, lorsque le lien est de type "search" ou "ark", la publication est plus largement partagée.

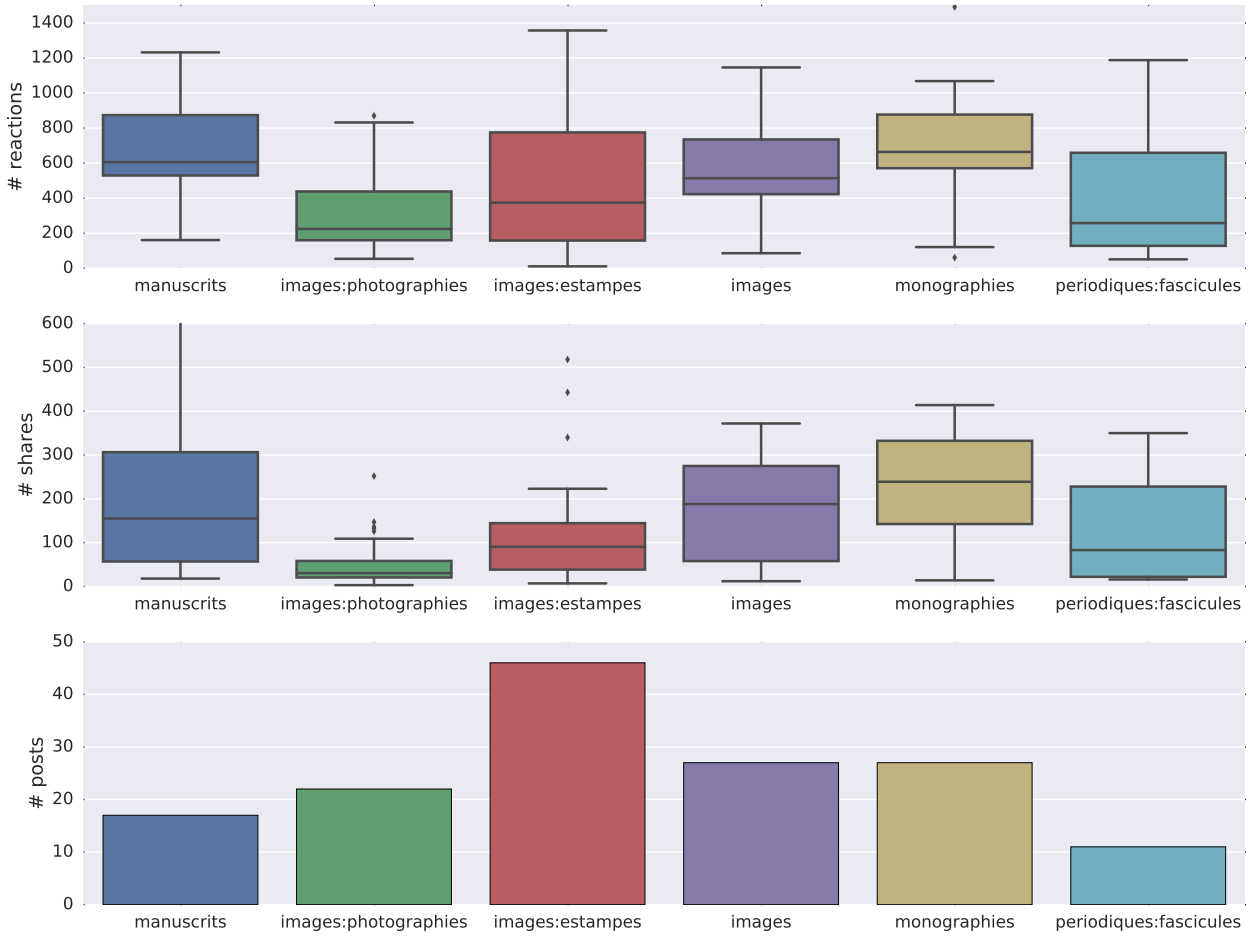

FIGURE 6.4 – Nombre de réactions/partages en fonction du type/catégorie de documents associé aux publications.

Enfin, lorsque l'on se concentre sur les publications de type "ark", il est possible de connaître le type (voir Sec. 2.1.3) du document partagé par l'équipe de médiation de la page Facebook de Gallica. Dans ce cas précis, la figure (6.4) montre que les manuscrits, les images et les monographies sont les documents les plus populaires sur Facebook. Il reste cependant important de noter que ces résultats sont extrapolés à partir de peu de données et qu'il sera nécessaire de confirmer cette conclusion avec une quantité plus importante de données.

Notons maintenant qu'il existe deux façons distinctes de publier un document renvoyant vers Gallica ("ark"). La figure (6.5) est un exemple de publication Facebook. Le rédacteur de la publication peut choisir de renvoyer l'utilisateur sur Gallica soit depuis un lien dans le texte (en bleu), soit par le biais d'un lien en vignette (en rouge).

FIGURE 6.5 – Exemple de publication Facebook.

Pour découvrir l'impact de la médiation Facebook et plus particulièrement de la méthode de publication, l'équipe de médiation a publié, entre le 1er Janvier 2017 et le 1er Mars 2017, 32 documents dont ;

— 16 utilisent un "lien dans le texte" (url dans le texte),

— 16 utilisent un "lien en vignette" (url dans l'image).

A partir des logs, il est alors possible de déterminer le nombre d'utilisateur en provenance de Facebook qui a consulté ces documents précis.

type	# hits	# hits moyen	# hits median
"lien dans le texte"	4,550	285	108
"lien en vignette"	126,889	7,930	2,417

TABLE 6.1 – Comparaison type de publications Facebook sur le nombre de visiteurs.

En conclusion, l'utilisation d'un "lien en vignette" a pour effet d'engendrer 25 fois plus de visites que l'utilisation d'un "lien dans le texte". Ce résultat a conduit la BnF à modifier sa manière de publier sur Facebook.

Chapitre 7

Recommandations et perspectives

7.1 Perspectives

Dans ce rapport, nous avons principalement décrit deux traitements : l'analyse des parcours (chapitre 3) et des usages documentaires (chapitre 4). Cependant, au cours de l'analyse des logs de connexions différentes pistes d'études supplémentaires ont émergé. Ces perspectives seront brièvement expliquées dans la suite de cette section.

7.1.1 "Clustering" de documents basé sur les usages documentaires

Au cours de la Section 4.3, nous avons proposé un "clustering" des documents en nous appuyant sur les métadonnées et l'algorithme LDA. Cette approche prend l'ensemble du catalogue des œuvres numérisées de Gallica *à plat* sans tenir compte des usagers.

Une approche similaire, également inspirée du LDA, peut être dérivée en posant l'hypothèse suivante : les documents consultés dans une même session appartiennent à un même sujet/domaine et peuvent donc être rapprochés.

7.1.2 Amélioration de la classification des documents par LDA

Les premières expériences de classifications des documents par une application de la méthode LDA aux notices ont soulevé de nombreuses questions telles que :

1. la pertinence relative de cette approche par rapport à la segmentation déjà fournie par le type de document (pour un nombre de classes limité, l'algorithme retombe sur les catégories du catalogue),
2. l'utilisation de l'ensemble des champs qui comportent des mots liés au type de média du document plutôt qu'à son contenu sémantique,
3. la difficulté d'interprétation des classes obtenues (qui peuvent devenir pertinentes s'il on en accepte un grand nombre).

Les deux premières n'ont pas de réponse simple par la nature même du corpus des notices. On pourrait restreindre l'analyse à des champs ou à des éléments du document au contenu plus "sémantique" tels que le titre, ou la description, ou, si l'on revient aux documents eux-même, à leur table de matière ou à leur légende. Néanmoins ces approches ne sont valables qu'à l'intérieur de certains types de documents, ceux pour lesquels ces informations sont disponibles.

Sans préjuger des résultats de ces approches complémentaires, il est important de ne pas perdre de vue l'objectif initial d'analyse des usages du site Gallica, l'étape de classification des notices permettant de l'aborder sous l'angle de la diversité des documents rencontrés au cours d'une même session. À ce titre, une approche plus "aveugle" consisterait à effectuer une classification massive, autorisée par la masse de notices disponibles, de l'ordre d'une centaine de "topics". Il est clair que les résultats d'une telle classification seront difficiles à interpréter mais pourront servir pour l'analyse de la diversité documentaire des sessions, dont l'interprétation nous importe au premier chef.

7.1.3 "Clustering" de documents basé sur l'approche Word2Vec

Des travaux récents Mikolov et al. [2013], Goldberg and Levy [2014], Bojanowski et al. [2016], Le and Mikolov [2014] se sont concentrés sur le "word embedding" (Def. 7.1.1).

Definition 7.1.1. Le "word embedding" est une méthode d'apprentissage automatique issue de l'apprentissage profond ("deep learning") qui permet d'obtenir une représentation vectorielle de faible dimension de mots ou de phrases.

Le "word embedding" part de l'hypothèse que des mots contenus dans une même phrase sont sémantiquement proches. Ainsi, le "word embedding" cherche à attribuer des vecteurs *proches* pour des mots de sens proches. Inversement, les vecteurs de mots dont les sens sont éloignés seront potentiellement différents. Le point fort de cette approche est l'utilisation d'une base de données de type Wikipédia lors de l'apprentissage des représentations vectorielles.

A terme, nous espérons que le "word embedding" soit capable de rapprocher des documents sémantiquement proches et ce même s'ils ne partagent pas de mots communs (ce qui n'est pas le cas du LDA).

Un modèle a été appris à partir des sauvegardes de la base de données de Wikipedia pour vérifier l'efficacité de l'approche "word embedding" sur des données qui pourraient être celles de à Gallica.

Dans un premier temps, nous illustrons le modèle appris à partir d'une sélection de personnalités célèbres (scientifiques, photographes, empereurs, présidents, écrivains et poètes). Les vecteurs associés à chacune de ces personnalités sont de dimension 300. Pour pouvoir les représenter sous forme graphique, il est nécessaire de réduire cette dimension à 2 (par le biais d'une Analyse en Composantes Principales). La figure 7.1 permet de confirmer la performance de la méthode : des personnalités en provenance d'une même discipline sont proches les unes des autres et éloignées des personnages en provenance d'une discipline différente.

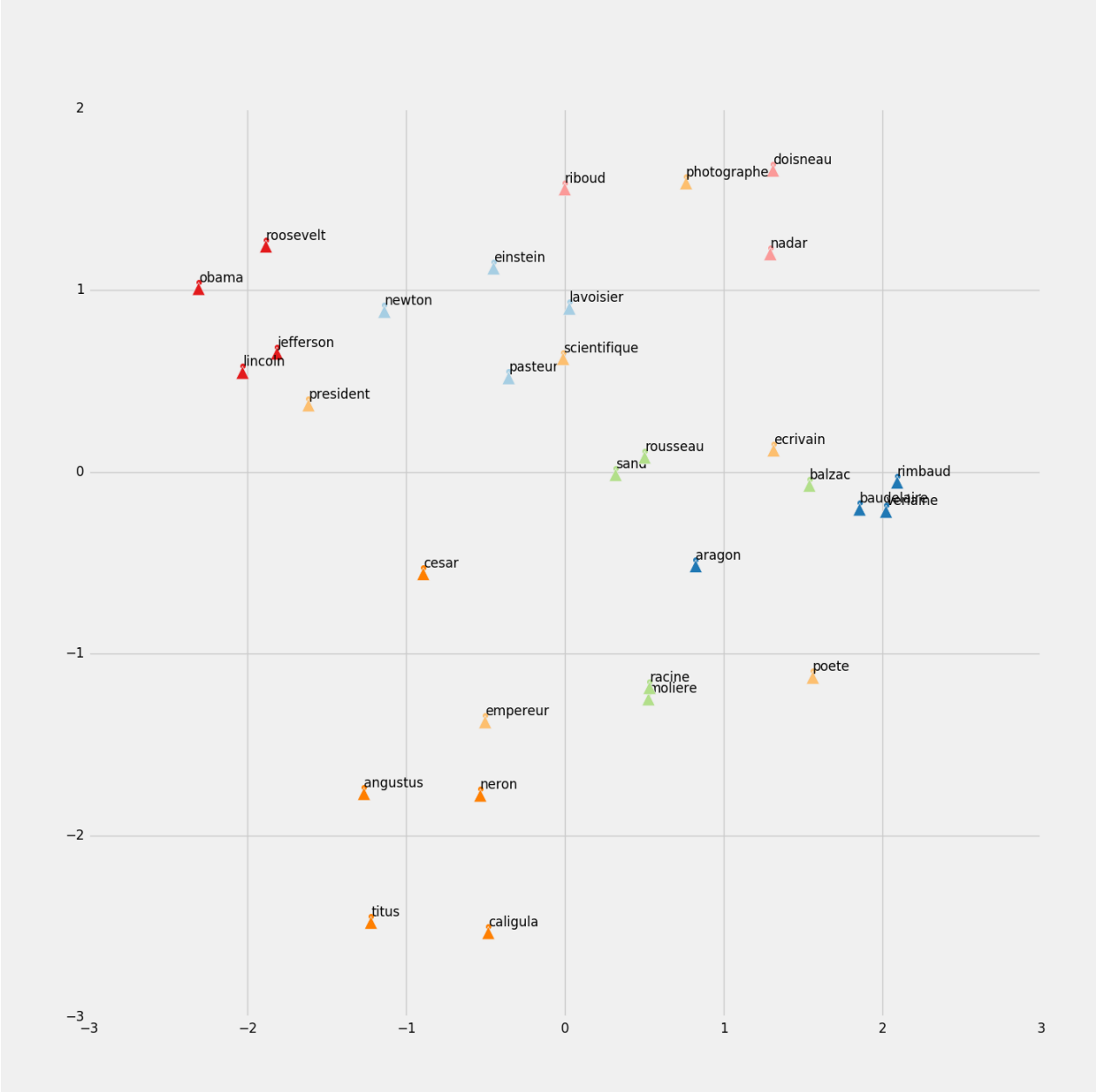

FIGURE 7.1 – Représentation en 2 dimensions de vecteurs issus du modèle de "word embedding" de personnalités célèbres.

7.2 Recommandations

Cette section a pour but de proposer des recommandations à la Bibliothèque nationale de France, susceptibles d'améliorer les analyses présentées voire d'ouvrir la porte à de nouvelles analyses impossibles à réaliser actuellement.

Auparavant conservés pour des raisons de sécurité, les logs ont été pour la première fois utilisés pour une autre finalité : connaître les usages.

Dans un premier temps, il nous semble important de rappeler les recommandations qui ont déjà fait l'objet de modifications de la part de la BnF. Une analyse rapide des logs a montré que les logs étaient incomplets. A savoir qu'il existait des *trous* dans les logs. Le problème provenait des serveurs de Gallica qui ne sauvegardaient pas les logs à la suite d'un *plantage*. Ces *plantages* nous ont été décrits comme fréquents mais normaux. Un correctif résolvant ce problème a été appliqué au début du mois de mai 2016. Dans un deuxième temps, le format initial des logs ne comportait pas le

champ “User-agent”. Il nous a semblé important de le rajouter dans le but de détecter efficacement les “good bots” subsection 2.2.3 et donc de pouvoir exclure les logs relatifs à ces derniers. Enfin, le champ “referer” était lui aussi absent des logs ne permettant pas d’analyser la provenance des usagers de Gallica. Nous tenons à souligner la réactivité du service technique de Gallica qui a su répondre positivement et rapidement à ces recommandations.

Definition 7.2.1. Un cookie (parfois traduit *témoin*) est défini par le protocole de communication HTTP comme étant une suite d’informations envoyée par un serveur HTTP à un client HTTP, que ce dernier retourne lors de chaque interrogation du même serveur HTTP sous certaines conditions. Le cookie est l’équivalent d’un fichier texte de petite taille, stocké sur le terminal de l’internaute. Existant depuis les années 1990, ils permettent aux développeurs de sites web de conserver des données utilisateur afin de faciliter la navigation et de permettre certaines fonctionnalités. Les cookies ont toujours été plus ou moins controversés car contenant des informations personnelles résiduelles pouvant potentiellement être exploitées par des tiers. [1]

Aujourd’hui, derrière une même adresse IP peuvent se cacher plusieurs individus différents et un même utilisateur peut changer d’adresse d’une session à l’autre. Le cookie (Def. 7.2.1) permettrait d’identifier un même utilisateur sur différentes sessions à condition qu’il conserve le même outil de navigation et que celui-ci autorise l’utilisation de cookies. Par conséquent, l’implantation d’un cookie pourrait servir à identifier les utilisateurs de manière plus précise et plus durable (au delà de la session) puisqu’il est possible de configurer une *durée de vie* illimité d’un “cookie”. Cependant, rien ne garantit que les navigateurs conservent ces derniers durant une durée illimitée. Un “cookie” peut être supprimé parce qu’il n’a pas été utilisé depuis longtemps, ou par l’intervention de l’utilisateur.

Il nous semble raisonnable de proposer une durée limite de 1 à 2 mois sur le site de Gallica. La mise en place d’un “cookie” sur Gallica pourrait permettre une étude plus approfondie des utilisateurs. Actuellement, nos analyses se fondent sur les sessions et ne permettent pas de voir la globalité des usages d’un utilisateur. Il serait intéressant d’analyser l’évolution des sessions d’un même utilisateur au cours du temps, ce qui n’est pas possible avec les logs de connexion actuels.

Enfin, un intérêt supplémentaire de la création d’un cookie est de pouvoir l’utiliser afin de proposer un système de recommandation de documents personnalisé.

1. Il est à noter que la méthode de mesure d’audience mise en place par XiTi (voir section 2.4) repose sur l’utilisation d’un cookie.

Bibliographie

Josh Attenberg, Sandeep Pandey, and Torsten Suel. Modeling and predicting user behavior in sponsored search. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1067–1076. ACM, 2009.

Valérie Beaudouin and Jérôme Denis. Observer et évaluer les usages de Gallica. Réflexion épistémologique et stratégique. Research report, BnF ; Telecom ParisTech, September 2014. URL https://halshs.archives-ouvertes.fr/halshs-01078530.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv :1607.04606*, 2016.

Kenneth P Burnham and David R Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2) :261–304, 2004.

Diego Ceccarelli, Sergiu Gordea, Claudio Lucchese, Franco Maria Nardini, and Gabriele Tolomei. Improving europeana search experience using query logs. In *International Conference on Theory and Practice of Digital Libraries*, pages 384–395. Springer, 2011.

Eugene Charniak. *Introduction to artificial intelligence*. Pearson Education India, 1985.

Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, 1(1) :5–32, 1999.

Resul Das and Ibrahim Turkoglu. Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Syst. Appl.*, 36(3) :6635–6644, April 2009. ISSN 0957-4174. doi : 10.1016/j.eswa.2008.08.067. URL http://dx.doi.org/10.1016/j.eswa.2008.08.067.

Robert F Dell, Pablo E Román, and Juan D Velásquez. 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*, pages 385–388. IEEE Computer Society, 2008.

Gul Nildem Demir, Murat Goksedef, and A. Sima Etaner-Uyar. Effects of session representation models on the performance of web recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW '07, pages 931–936, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 978-1-4244-0831-3. doi : 10.1109/ICDEW.2007.4401087. URL http://dx.doi.org/10.1109/ICDEW.2007.4401087.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques : a survey. *Knowledge-based systems*, 70 :301–323, 2014.

GMV Conseil. Evaluation de l'usage et de la satisfaction de la bibliothèque numérique Gallica et perspectives d'évolution. Consulting report, BnF, 2012.

Guillaume Godet. Guide d'interopérabilité OAI-PMH pour un référencement des documents numériques dans Gallica. Technical report, BnF, June 2015. URL http://www.bnf.fr/documents/Guide_oaipmh.pdf.

Yoav Goldberg and Omer Levy. word2vec explained : deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv :1402.3722*, 2014.

Google. How a session is defined in analytics - analytics help, 2016. URL https://support.google.com/analytics/answer/2731565?hl=en.

Şule Gündüz and M Tamer Özsu. 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*, pages 535–540. ACM, 2003.

Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

Incapsula. Bot traffic report 2013, 2003. URL https://www.incapsula.com/blog/bot-traffic-report-2013.html.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering : a review. *ACM computing surveys (CSUR)*, 31(3) :264–323, 1999.

Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference : A bibliography. *SIGIR Forum*, 37(2) :18–28, September 2003. ISSN 0163-5840. doi : 10.1145/959258.959260. URL http://doi.acm.org/10.1145/959258.959260.

John A. Kunze. Towards electronic persistence using ark identifiers, ark motivation and overview, 2003.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc., 2002.

Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management*, pages 9–15. ACM, 2001.

Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data mining and knowledge discovery*, 6(1) : 61–82, 2002.

Eric L. Morgan. An Introduction to the Search/Retrieve URL Service (SRU), 2004. URL http://www.ariadne.ac.uk/issue40/morgan/.

Distil Network. The 2015 bad bot landscape report, 2015. URL https://cdn2.hubspot.net/hubfs/258389/WP__2015_Bad_Bot_Landscape_Report.pdf.

Fan Qiu and Yi Cui. An analysis of user behavior in online video streaming. In *Proceedings of the International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval*, VLS-MCMR '10, pages 49–54, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0166-4. doi : 10.1145/1878137.1878149. URL http://doi.acm.org/10.1145/1878137.1878149.

Lawrence R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4. URL http://dl.acm.org/citation.cfm?id=108235.108253.

Narayanan Sadagopan and Jie Li. Characterizing typical and atypical user sessions in clickstreams. In *Proceedings of the 17th international conference on World Wide Web*, pages 885–894. ACM, 2008.

SéRgio S. C. Silva, Rodrigo M. P. Silva, Raquel C. G. Pinto, and Ronaldo M. Salles. Botnets : A survey. *Comput. Netw.*, 57(2) :378–403, February 2013. ISSN 1389-1286. doi : 10.1016/j.comnet.2012.07.021. URL http://dx.doi.org/10.1016/j.comnet.2012.07.021.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7) :424–440.

Pang-Ning Tan and Vipin Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6(1) :9–35, 2002. ISSN 1573-756X. doi : 10.1023/A:1013228602957. URL http://dx.doi.org/10.1023/A:1013228602957.

Annexe A

Algorithme d'Espérance-Maximisation (EM)

Dans un modèle $p_{\mathbf{\Theta}}(\mathbf{X})$ à données latentes $\mathbf{z} = \Big[z_1, \ldots, z_S\Big]$, le paramètre $\mathbf{\Theta}$ peut s'estimer par le maximum de vraisemblance :

$$\hat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \log p_{\mathbf{\Theta}}(\mathbf{X}) \,, \tag{A.1}$$

$$= \arg\max \log p_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{z}) - \log p_{\mathbf{\Theta}}(\mathbf{z}|\mathbf{X}) \,, \tag{A.2}$$

La vraisemblance $p_{\mathbf{\Theta}}(\mathbf{X})$ est malheureusement non concave et sa maximisation directe devient difficile. C'est pourquoi Dempster et al. [1977] propose de rendre la maximisation de (A.2) tractable par la maximisation itérative d'une borne inférieure de cette vraisemblance. La vraisemblance est redéfinie avec l'espérance des données complétées conditionnellement au paramètre courant $\mathbf{\Theta}^{(t)}$:

$$\log p_{\mathbf{\Theta}}(\mathbf{X}) = \sum_{\mathbf{z}} p_{\mathbf{\Theta}^{(t)}}(\mathbf{z}|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{z}) - \sum_{\mathbf{z}} p_{\mathbf{\Theta}^{(t)}}(\mathbf{z}|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{z}|\mathbf{X}) \,, \tag{A.3}$$

$$= Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) + H(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) \,. \tag{A.4}$$

où $\mathbf{\Theta}^{(t)}$ est la valeur de l'algorithme à la t^{ieme} itération.
Il a été montré que (voir Dempster et al. [1977])

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) \tag{A.5}$$

garantit une suite à vraisemblance croissante. On distingue deux étapes dans cette itération, dite "E step" et "M step".
E step : $p_{\mathbf{\Theta}^{(t)}}(z|\mathbf{x}) = \Pi_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s|\mathbf{x}_s)$, on calcule chaque espérance de z_n par,

$$p_{\mathbf{\Theta}^{(t)}}(z_s|\mathbf{x}_s) \propto p(z_s) p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) \tag{A.6}$$

M step :

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) \tag{A.7}$$

On réécrit la vraisemblance jointe du modèle de mélange de chaînes de Markov dans lequel$\Theta =$

$(\alpha_k, (\pi_i^k)_{i=1,\dots,M}, (A_{i,j}^k)_{i,j=1,\dots,M})_{k=1,\dots,K}$:

$$\log p_{\mathbf{\Theta}}(\mathbf{X}, z) = \log p_{\mathbf{\Theta}}(\mathbf{X}|z) + \log p_{\mathbf{\Theta}}(z) \ , \tag{A.8}$$

$$= \log \Pi_{s=1}^{S} p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) + \log \Pi_{s=1}^{S} p_{\mathbf{\Theta}}(z_n) \ , \tag{A.9}$$

$$= \sum_s \log p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) + \sum_s \log p_{\mathbf{\Theta}}(z_s) \ , \tag{A.10}$$

$$= \sum_s \sum_k [z_s = k] \tag{A.11}$$

$$\times \log \left(\sum_{i=1}^{M} ([x_{s,1} = i] \log \pi_i^{(k)}) + \sum_i^M \sum_j^M t_{s,i,j} \log A_{i,j}^{(k)} \right) \tag{A.12}$$

$$+ \sum_s \sum_k [z_s = k] \log \alpha_k \sum_k [z_s = k] \log \alpha_k \tag{A.13}$$

Finalement,

$$Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) = \sum_z p_{\mathbf{\Theta}^{(t)}}(z|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{X}, z) \ , \tag{A.14}$$

$$= \sum_{s=1}^{S} \sum_{k=1}^{K} p_{\mathbf{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) \tag{A.15}$$

$$\times \left(\sum_{i=1}^{M} ([x_{s,1} = i] \log \pi_i^{(k)}) + \sum_{i,j=1}^{M} t_{s,i,j} \log A_{i,j}^{(k)} \right) \tag{A.16}$$

$$+ \sum_{s=1}^{S} \sum_{k=1}^{K} p_{\mathbf{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) \log \alpha_k \tag{A.17}$$

L'estimation des paramètres se réalise en $\sum_k \alpha_k = 1$, d'où l'introduction des dérivées du Lagrangien :

$$\frac{\partial}{\partial \alpha_k} \left(Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) + \lambda (\sum_{\ell=1}^{K} \alpha_\ell - 1) \right) = 0 \ , \tag{A.18}$$

$$\frac{1}{\alpha_k} \sum_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) + \lambda = 0 \ , \tag{A.19}$$

On a alors,

$$\alpha_k = -\frac{1}{\lambda} \sum_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) \tag{A.20}$$

D'après la contrainte on a,

$$\sum_{\ell=1}^{K} \alpha_\ell = 1 \ , \tag{A.21}$$

$$-\frac{1}{\lambda} \sum_{\ell=1}^{K} \sum_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s = \ell|\mathbf{x}_s) = 1 \ . \tag{A.22}$$

Il s'en suit que $\lambda = -\sum_{\ell=1}^{K} \sum_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s = \ell|\mathbf{x}_s)$ permettant d'en déduire la valeur mise à jour,

$$\alpha_k^{(t+1)} = \frac{\sum_{s=1}^{S} p_{\boldsymbol{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s)}{\sum_{\ell=1}^{K}\sum_{s=1}^{S} p_{\boldsymbol{\Theta}^{(t)}}(z_s = \ell|\mathbf{x}_s)} \ . \tag{A.23}$$

Les autres mise à jour sont énoncées dans la suite mais le détail des calculs n'est pas exposé.

$$\pi_i^{k,(t+1)} = \frac{\sum_s [x_{s,1} = i] p_{\boldsymbol{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s)}{\sum_{j=1}^{M}\sum_s [x_{s,1} = j] p_{\boldsymbol{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s)} \ , \tag{A.24}$$

$$A_{i,j}^{k,(t+1)} = \frac{\sum_s p_{\boldsymbol{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) t_{s,i,j}}{\sum_{\ell=1}^{M}\sum_{s=1}^{S} p_{\boldsymbol{\Theta}^{(t)}}(z_s = k|\mathbf{x}_s) t_{s,\ell,j}} \ . \tag{A.25}$$

Annexe B

Mixture of Time-Continuous Markov Chains

Soit $\mathbf{x}_s = \{(s_0, t_0), \ldots, (s_n, t_n)\}$ une séquence de longueur finie générée par une Continuous-Time Markov Chaîne (CTMC) est définie comme une séquence d'états et de temps de transition. Paramètres de la CTMC : $\theta_m = \{\pi_m(i), q_m(i,j) \forall i,j \in \{1, \ldots, D\}^2\}$ avec D le nombre d'état possible.

La vraisemblance de $\mathbf{x}$ est le produit des probabilités conditionnelles de chaque évènement :

$$p_{\theta_m}(\mathbf{x}_s) = \Pi_{i=1}^{D} \pi_m(i)^{[\mathbf{x}_{s,1}=i]} \Pi_{i \neq j} q_m(i,j)^{n_s(i,j)} \Pi_{i=1}^{D} \mathrm{e}^{-q_m(i,i)\tau_s(i)} , \tag{B.1}$$

où $\tau(i)$ le temps passé à l'état i et n_{ij} le nombre total de transition de l'état i à l'état j.

Dans le cas où $\mathbf{x}$ est générée par M CTCM paramétrisées par $\mathbf{\Theta} = \{\theta_1, \ldots, \theta_M\}$, la vraisemblance devient :

$$p_{\mathbf{\Theta}}(\mathbf{x}_s) = \sum_{m=1}^{M} \alpha_m p_{\theta_m}(\mathbf{x}_s) . \tag{B.2}$$

La vraisemblance de $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_S\}$

$$p_{\mathbf{\Theta}}(\mathbf{X}) = \Pi_{s=1}^{S} p_{\mathbf{\Theta}}(\mathbf{x}_s) . \tag{B.3}$$

On introduit $z_s = \{1, \ldots, M\}$ la variable cachée qui caractérisent quelle CTMC est responsable de la génération de la séquence $\mathbf{x}_s$.

Dans un modèle $p_{\mathbf{\Theta}}(\mathbf{X})$ à données latentes $\mathbf{z} = \{z_1, \ldots, z_S\}$, le paramètre $\mathbf{\Theta}$ peut s'estimer par le maximum de vraisemblance :

$$\log p_{\mathbf{\Theta}}(\mathbf{X}) = \sum_{\mathbf{z}} p_{\mathbf{\Theta}^{(t)}}(\mathbf{z}|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{X}, \mathbf{z}) - \sum_{\mathbf{z}} p_{\mathbf{\Theta}^{(t)}}(\mathbf{z}|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{z}|\mathbf{X}) , \tag{B.4}$$

$$= Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) + H(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) . \tag{B.5}$$

où $\mathbf{\Theta}^{(t)}$ est la valeur de l'algorithme à la t^{ieme} itération.

Il a été montré que (voir Dempster et al. [1977])

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) \tag{B.6}$$

garantit une suite à vraisemblance croissante. On distingue deux étapes dans cette itération, dite "E step" et "M step".

E step : $p_{\mathbf{\Theta}^{(t)}}(z|\mathbf{x}) = \Pi_{s=1}^{S} p_{\mathbf{\Theta}^{(t)}}(z_s|\mathbf{x}_s)$, on calcule chaque espérance de z_n par,

$$p_{\mathbf{\Theta}^{(t)}}(z_s|\mathbf{x}_s) \propto p(z_s) p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) \tag{B.7}$$

M step :

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) \tag{B.8}$$

On réécrit la vraisemblance jointe du modèle de mélange de chaînes de Markov dans lequel$\Theta = (\alpha_k, (\pi_i^k)_{i=1,\dots,M}, (A_{i,j}^k)_{i,j=1,\dots,M})_{k=1,\dots,K}$:

$$\log p_{\mathbf{\Theta}}(\mathbf{X}, z) = \log p_{\mathbf{\Theta}}(\mathbf{X}|z) + \log p_{\mathbf{\Theta}}(z) \ , \tag{B.9}$$

$$= \log \Pi_{s=1}^{S} p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) + \log \Pi_{s=1}^{S} p_{\mathbf{\Theta}}(z_n) \ , \tag{B.10}$$

$$= \sum_s \log p_{\mathbf{\Theta}}(\mathbf{x}_s|z_s) + \sum_s \log p_{\mathbf{\Theta}}(z_s) \ , \tag{B.11}$$

$$= \sum_s \sum_m [z_s = m] \tag{B.12}$$

$$\times \log \left(\sum_{i=1}^{D} ([x_{s,1} = i] \log \pi_m(i) + \sum_{i \neq j} n_s(i,j) \log q_m(i,j) - \sum_{i=1}^{D} q_m(i) \tau_s(i) \right) \tag{B.13}$$

$$+ \sum_s \sum_k [z_s = m] \log \alpha_m \tag{B.14}$$

Finalement,

$$Q(\mathbf{\Theta}|\mathbf{\Theta}^{(t)}) = \sum_z p_{\mathbf{\Theta}^{(t)}}(z|\mathbf{X}) \log p_{\mathbf{\Theta}}(\mathbf{X}, z) \tag{B.15}$$

Les contraintes :

— $\sum_{j \neq i} q_m(i,j) = -q_i, \forall i$

— $\sum_m \alpha_m = 1$

— $\sum_i \pi_m(i) = 1, \forall m$

$$\alpha^{t+1} = \frac{\sum_s p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)}{\sum_{w=1}^{M} \sum_s p_{\mathbf{\Theta}^t}(z_s = w|\mathbf{x}_s)} \tag{B.16}$$

$$\pi_m^{t+1}(i) = \frac{\sum_s p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)[\mathbf{x}_{s,1} = i]}{\sum_{w=1}^{M} \sum_s p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)[\mathbf{x}_{s,1} = w]} \tag{B.17}$$

$q_m(i)$ est obtenue par sa définition : $-1/q_m(i)$ est the expected time of staying in state i,

$$-\frac{1}{q_m(i)} = \mathbb{E}\left[\text{total time on i}|\mathbf{X} \sim p_{\mathbf{\Theta}_m}\right] \tag{B.18}$$

$$-\frac{1}{q_m(i)^{t+1}} = \frac{\sum_s \tau_s(i) p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)}{\sum_s \sum_{w \ \neq i} n_s(i,w) p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)} \tag{B.19}$$

$$-\frac{1}{q_m(i)^{t+1}} = \frac{\sum_s \frac{\tau_s(i)}{\sum_{w \neq i} n_s(i,w)} p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)}{\sum_s p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)} \tag{B.20}$$

$$q_m(i,j)^{t+1} = -q_m(i)^{t+1} \frac{\sum_s n_s(i,j) p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)}{\sum_{w \neq i} \sum_s n_s(i,w) p_{\mathbf{\Theta}^t}(z_s = m|\mathbf{x}_s)} \tag{B.21}$$

Annexe C

Cluster avec prise en compte de google

Une part importante de Google a été remarquée lors de l'étude des "referers". C'est pourquoi, nous proposons en annexe une analyse des parcours utilisateurs incluant Google.

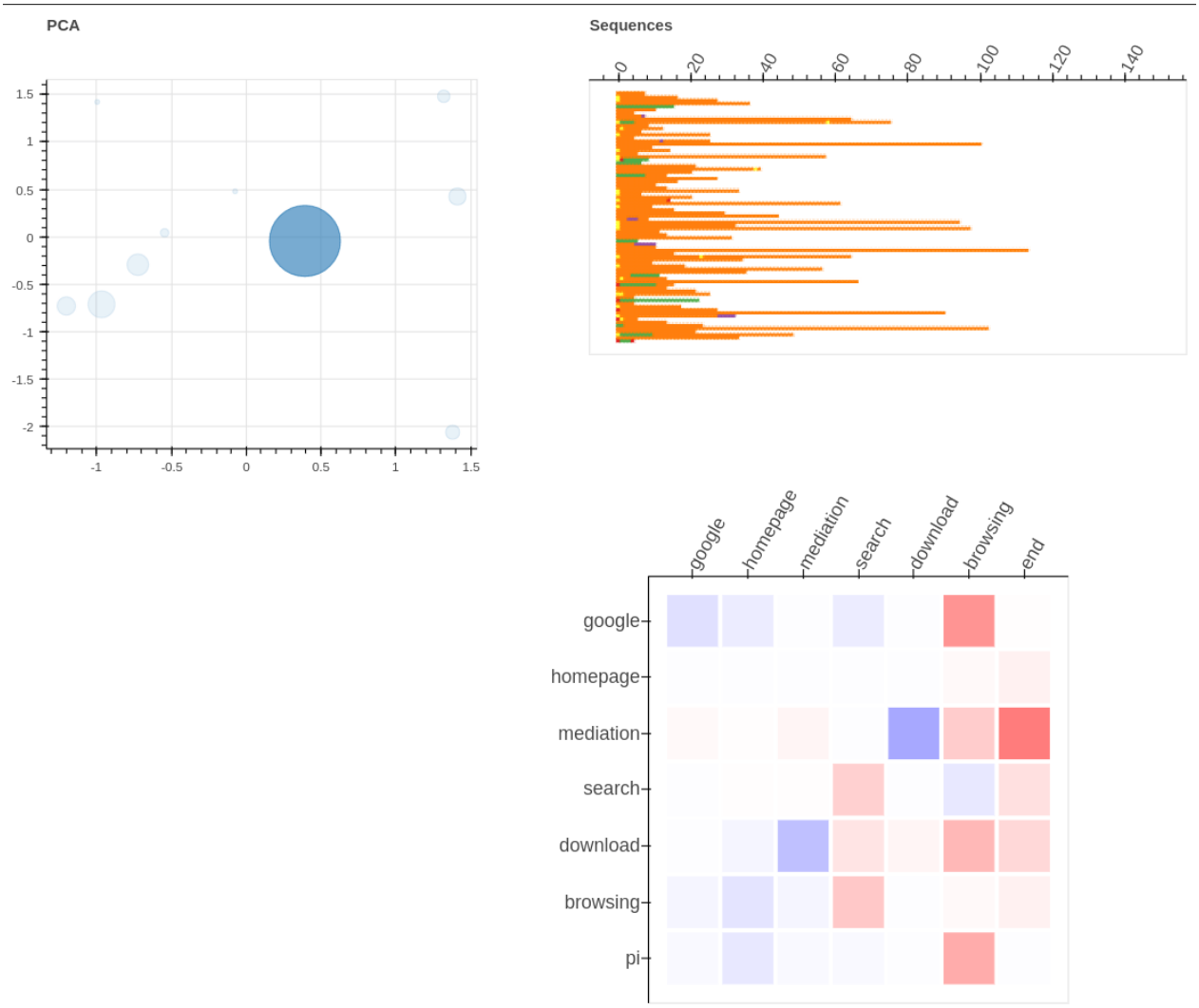

FIGURE C.1

Les figures C.1, C.2, C.3, C.4, C.5, C.6, C.7, C.8, C.9, C.10 sont des captures d'écran d'une visualisation interactive. Chacune des figures sont elles-mêmes composées de trois sous-figures.

Les sous-figures situées en-haut à gauche proviennent d'une analyse par composantes principales (ACP) de l'ensemble des matrices de transitions. L'ACP permet de compresser les paramètres des

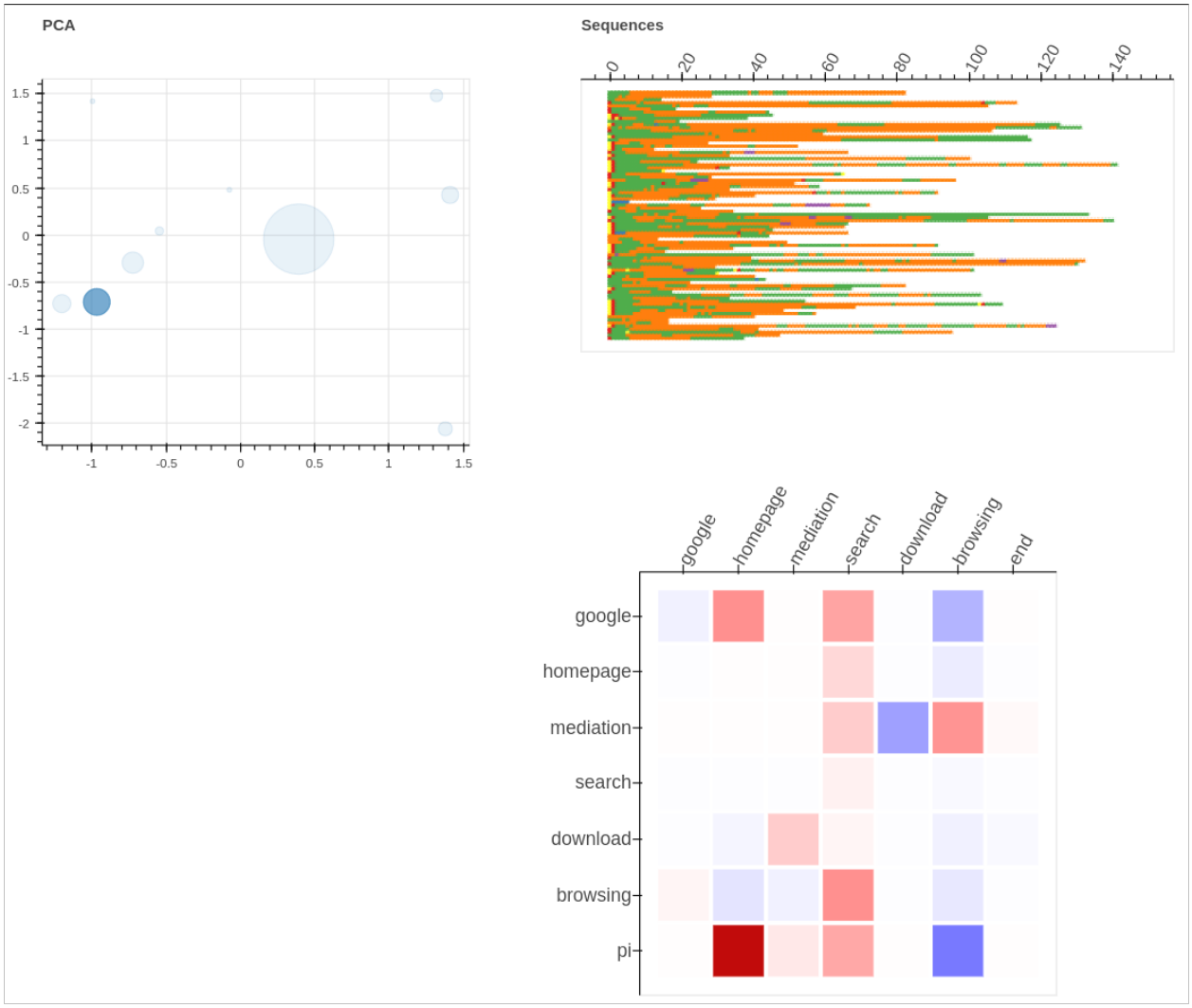

FIGURE C.2

matrices de transitions en 2 dimensions. Cette réduction de dimension nous permet de pouvoir représenter l'ensemble des clusters dans un unique graphique. Ainsi chaque bulle représente un cluster. La taille des bulles correspond également à la proportion d'usagers de Gallica appartenant à ce cluster. Les sous-figures situées en-haut à droite correspondent aux séquences issues de logs dont le code couleur est le suivant :

1. Navigation sur la page d'accueil :
2. Recherche Google :
3. Navigation sur les pages de médiations (collection + blog) :
4. Recherche Gallica :
5. Téléchargement d'un document :
6. Consultation "locale" dans l'interface Gallica :

Enfin les sous-figures en-bas à droite correspondent aux matrices de transitions propre à chaque cluster.

Le cluster représenté en figure C.1 regroupe le plus d'usagers. Il se caractérise par des sessions démarrant généralement par la consultation de documents et faisant intervenir de courtes phases de recherche.

Les clusters suivants (figures C.2, C.3, C.4, C.5) correspondent à des usagers accédant à Gallica par la page d'accueil depuis une recherche Google. Ces trois clusters diffèrent par les longeurs de

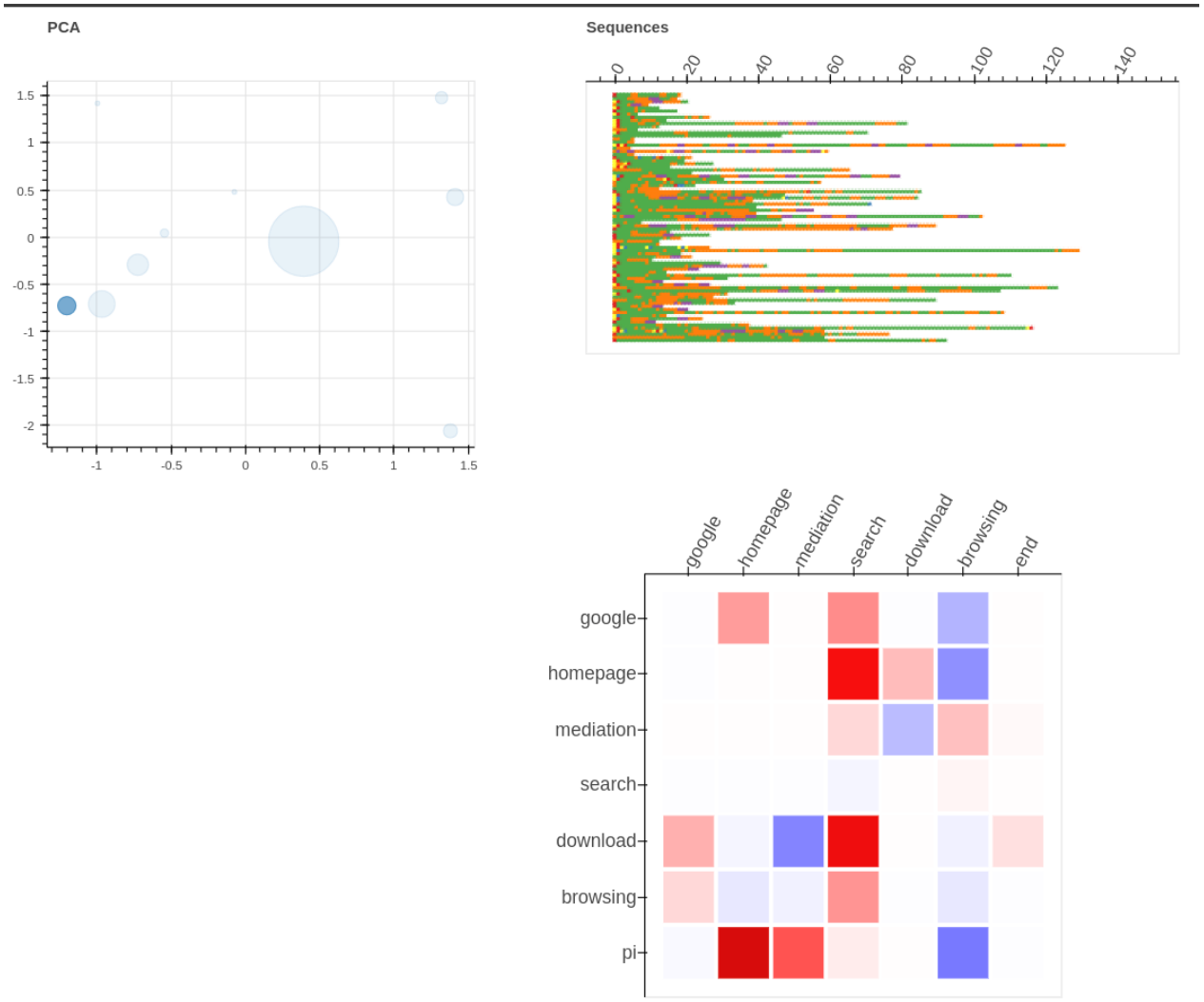

FIGURE C.3

leurs phrases de recherche. Les clusters C.2 et C.3 sont significativement plus enclin à faire appel au moteur de recherche de Gallica.

Le cluster C.7, regroupant une faible proportion d'utilisateurs, décrit les usagers qui interagissent avec les pages de médiations. Il est intéressant de remarquer que ces usagers accèdent souvent aux pages de médiation par le biais d'une recherche Google. Le cluster C.8 est similaire au cluster précédent C.7 mais se différencie par un accès direct (et non par Google) aux pages de médiations.

Les deux derniers clusters C.9 et C.10 sont les plus riches en enseignements. Ils décrivent des usages où Google est au centre des sessions. En effet on remarque une faible utilisation du moteur de recherche de Gallica et une importante présence de recherches Google. Ces usagers accèdent aux différents documents par le biais de multiples recherches Google.

Cette analyse permet de montrer que Google est un outil populaire chez les utilisateurs de Gallica. Pour une grande partie d'usagers, Google permet juste d'accéder directement à la page d'accueil, aux pages de médiations ou à un document particulier puis devient absent au cours du reste de leur session. Cependant pour une proportion non négligeable d'usagers (figures C.9 et C.10) Google est omniprésent au sein de leur session et se substitue même au moteur de recherche interne de Gallica.

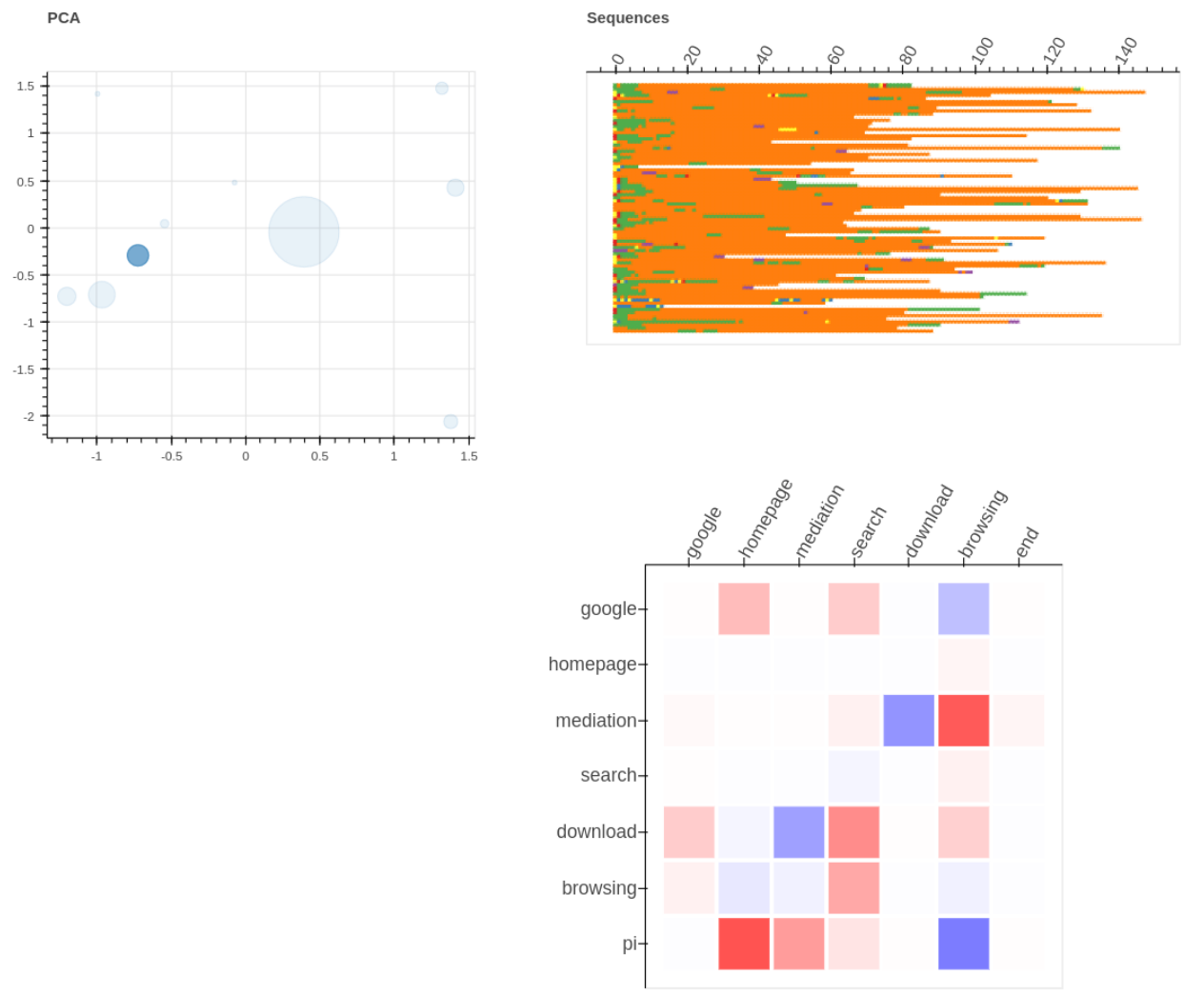
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.4

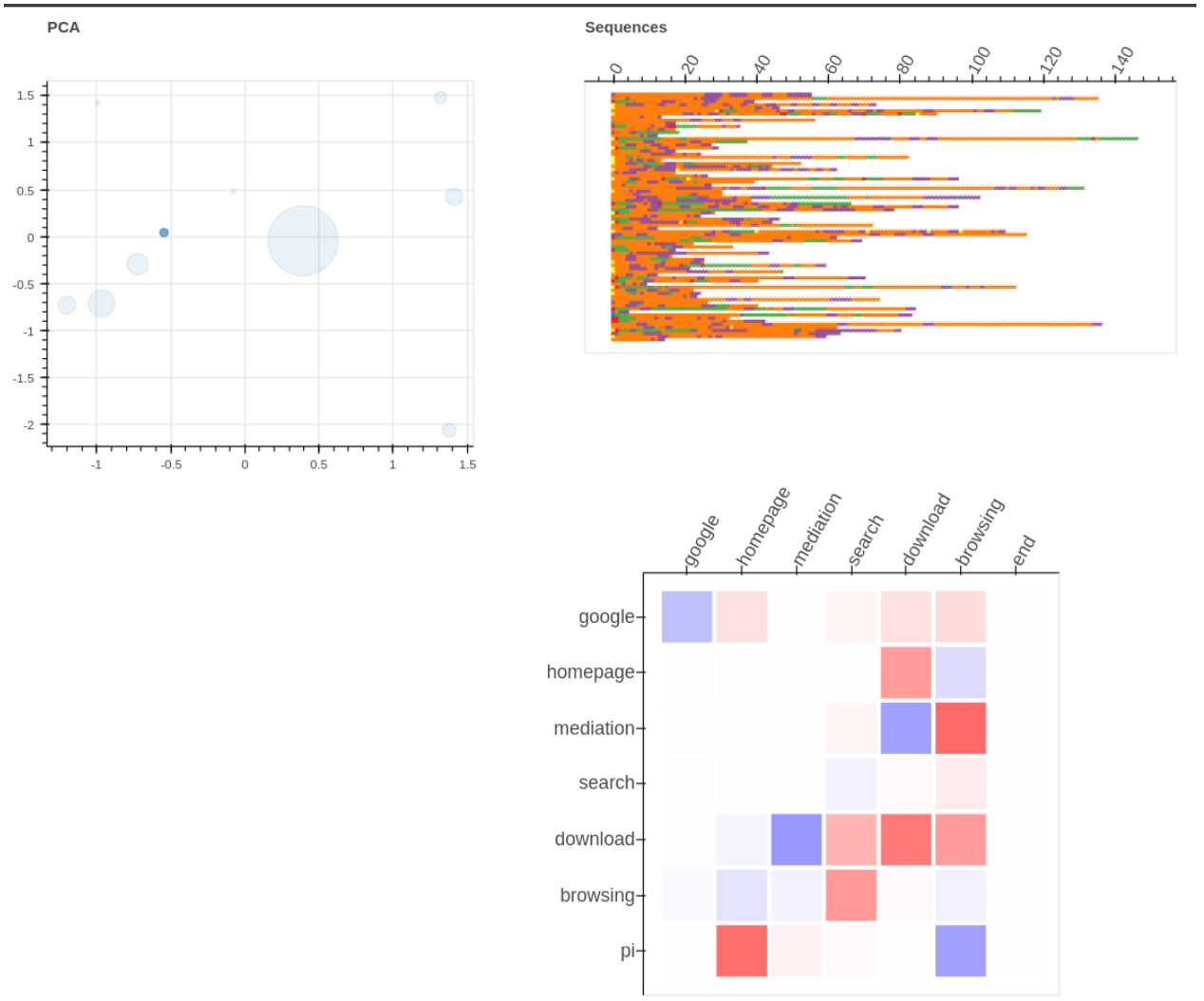
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.5

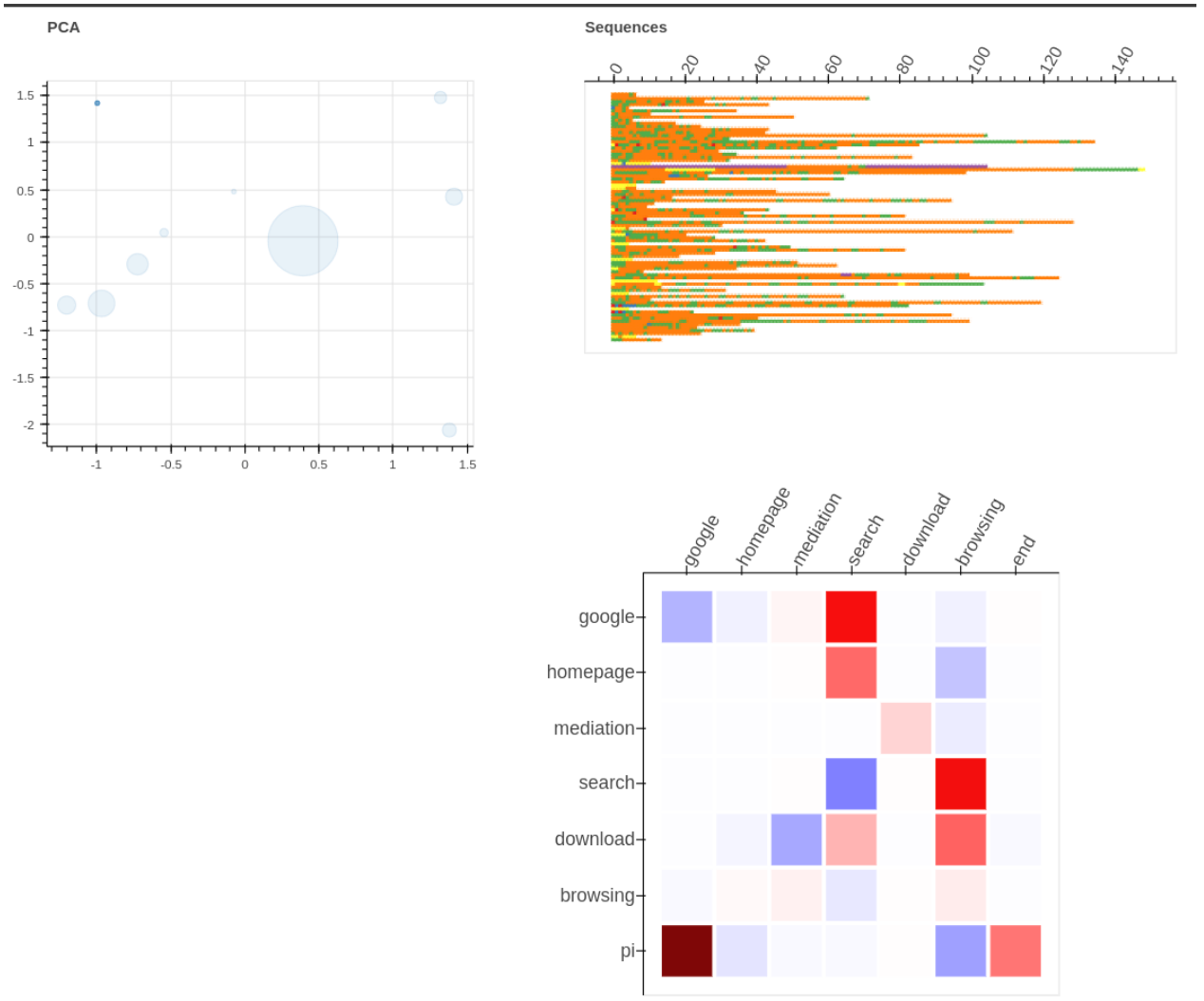
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.6

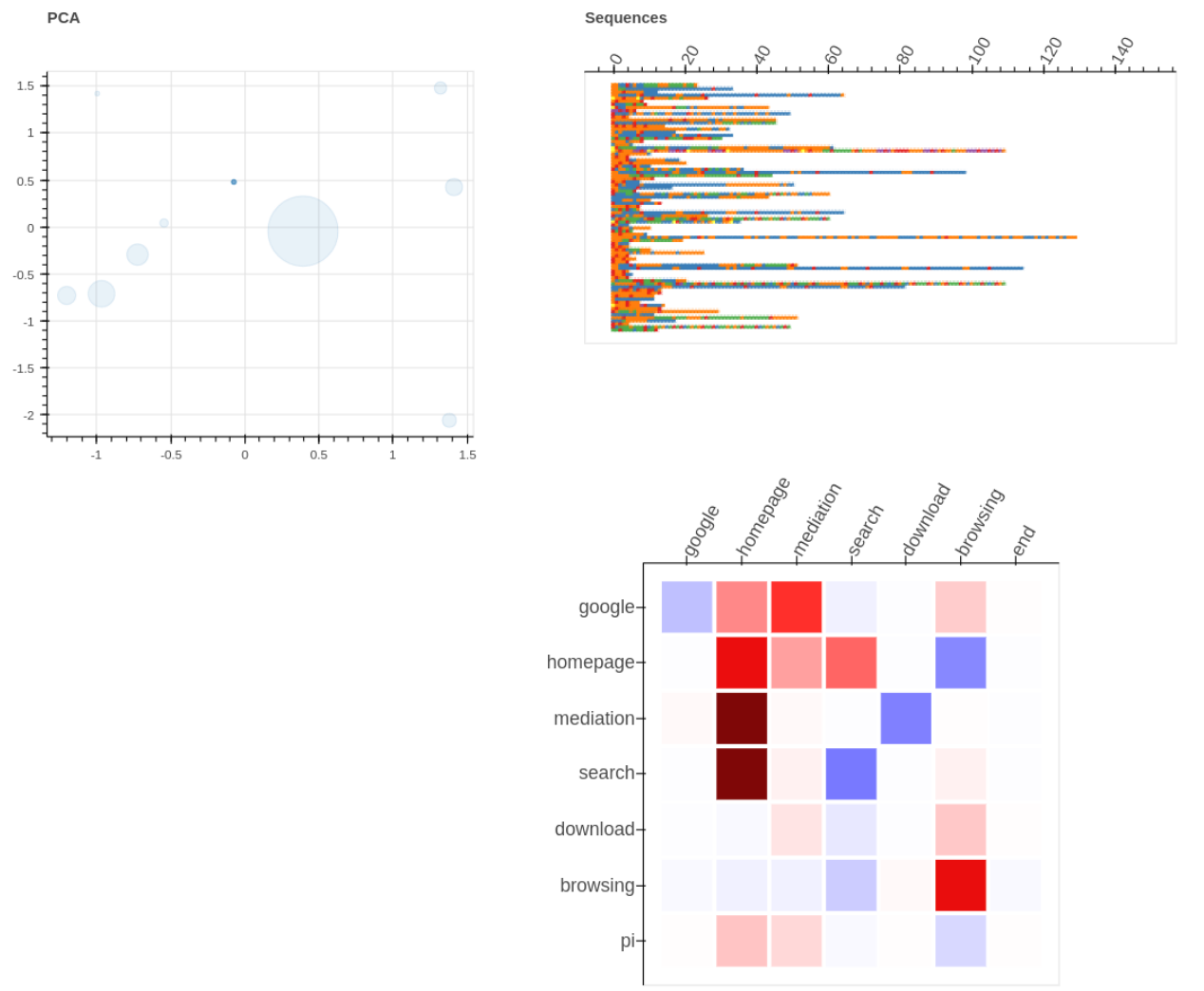
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.7

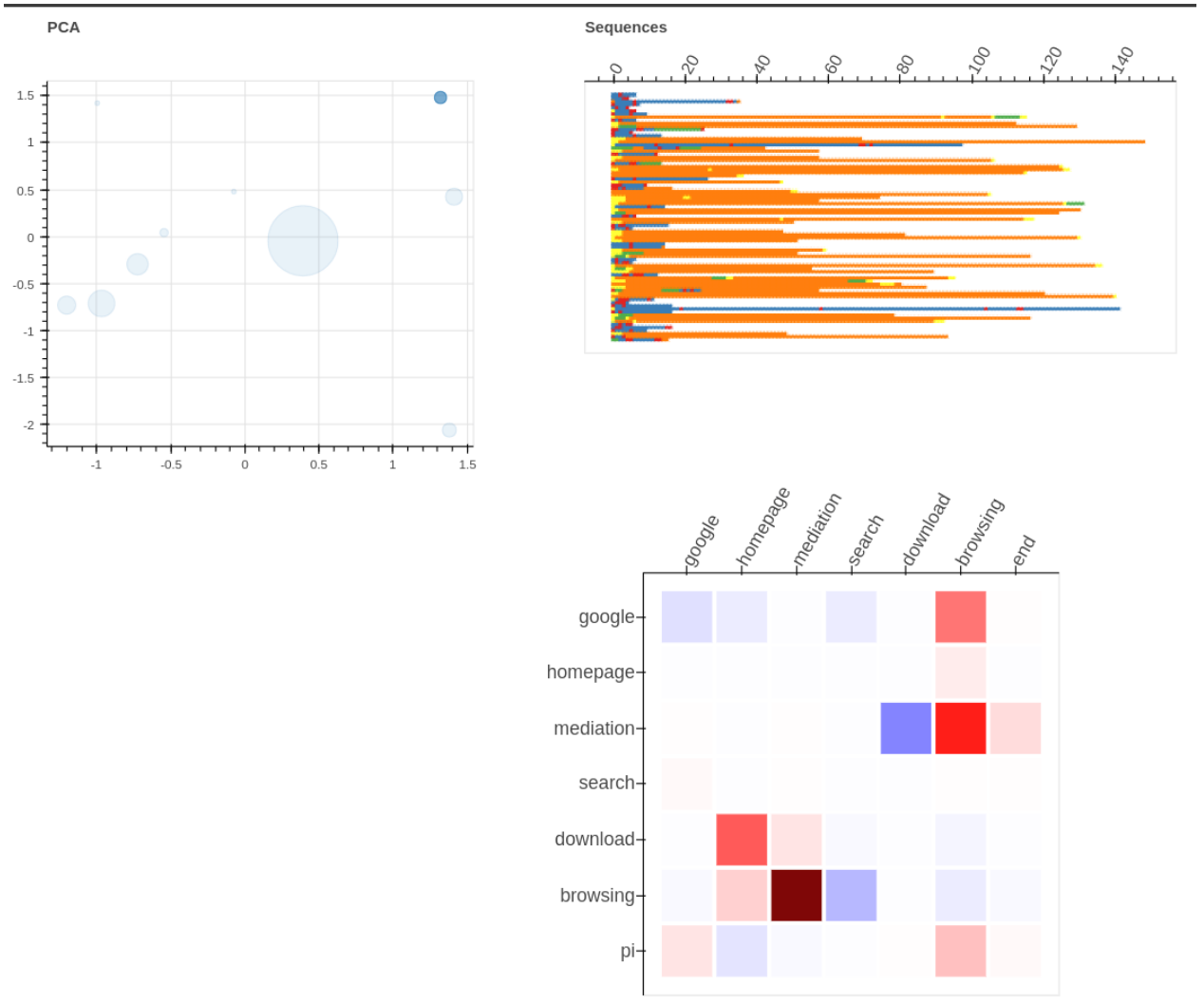
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.8

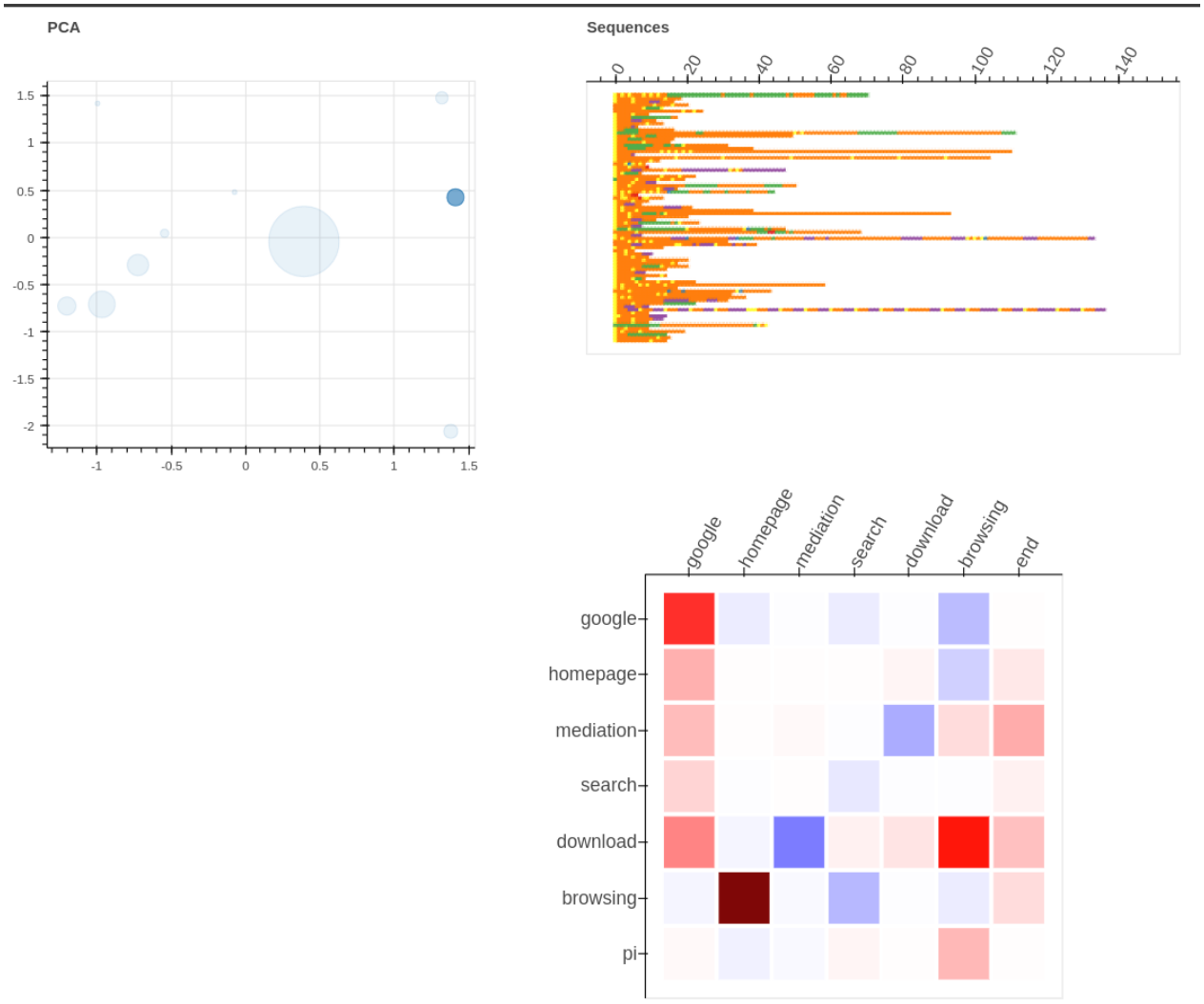
PCA
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
Sequences
0
20
40
60
80
100
120
140
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.9

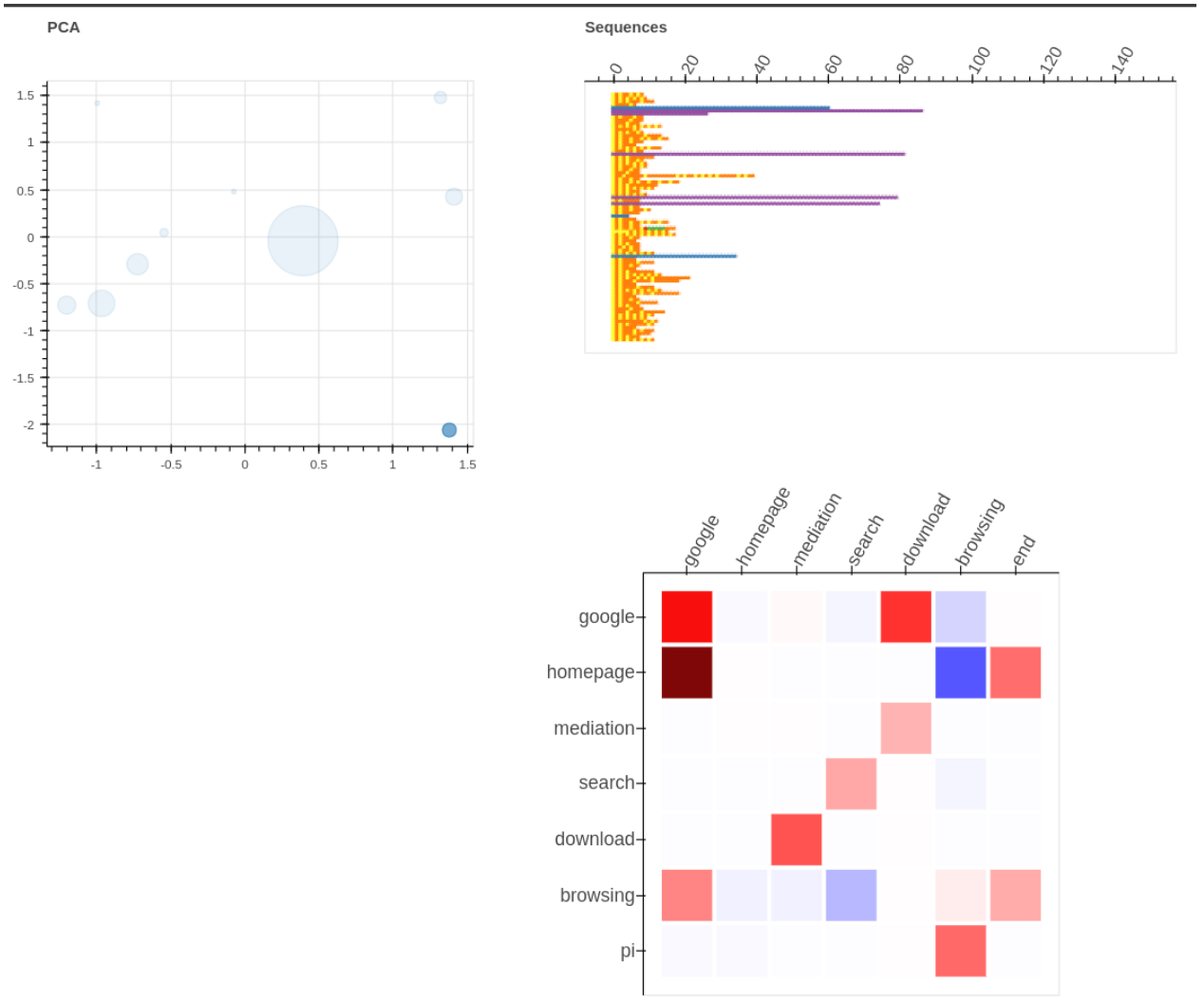
PCA
Sequences
0
20
40
60
80
100
120
140
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
-0.5
0
0.5
1
1.5
google
homepage
mediation
search
download
browsing
end
google
homepage
mediation
search
download
browsing
pi

FIGURE C.10

Annexe D

Illustration LDA

Les figures (D.2, D.3, D.4, D.5, D.6, D.7) illustrent 6 *topics* en rapportant les mots les plus représentatifs de chacun des *topics*. Les barres en bleues indiquent la fréquence des mots pour l'ensemble du catalogue de la BnF. Les barres rouges indiquent la fréquence des mots au sein du *topic* sélectionné.

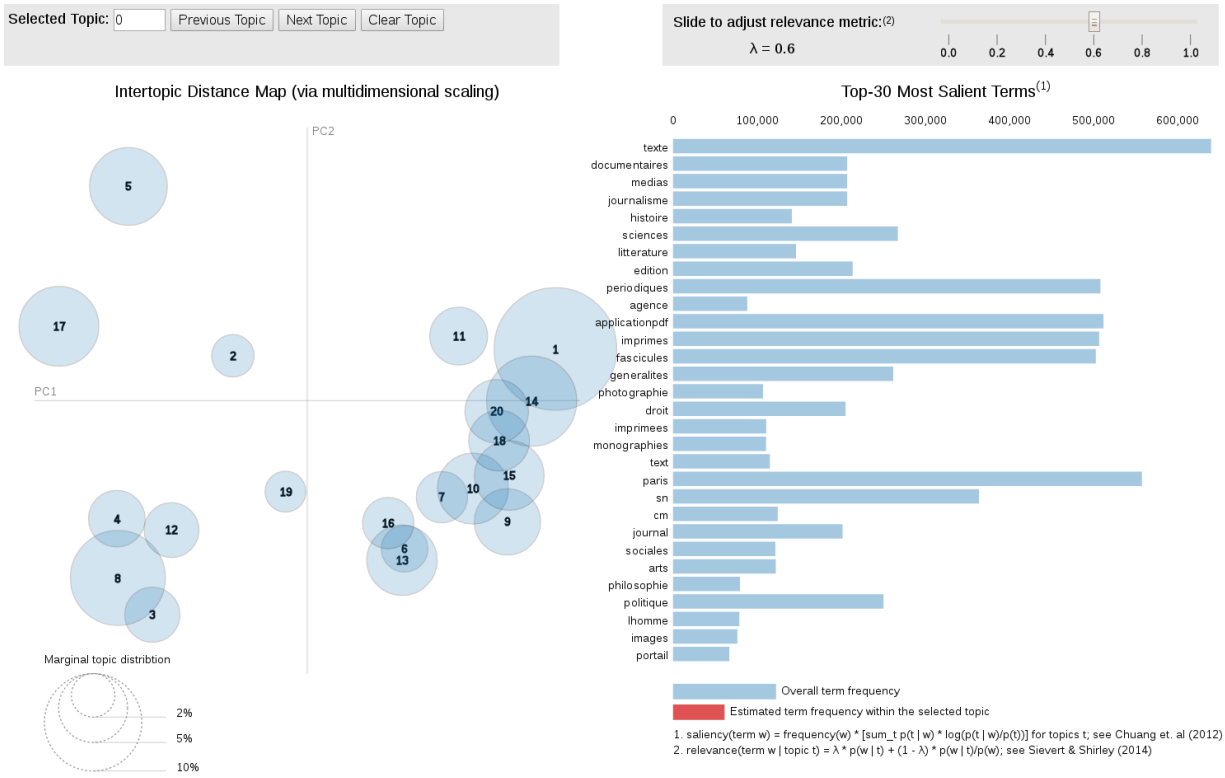

FIGURE D.1 – Segmentation des notices en 20 classes par la méthode LDA. Mots les plus discriminants pour la classification en "topics".

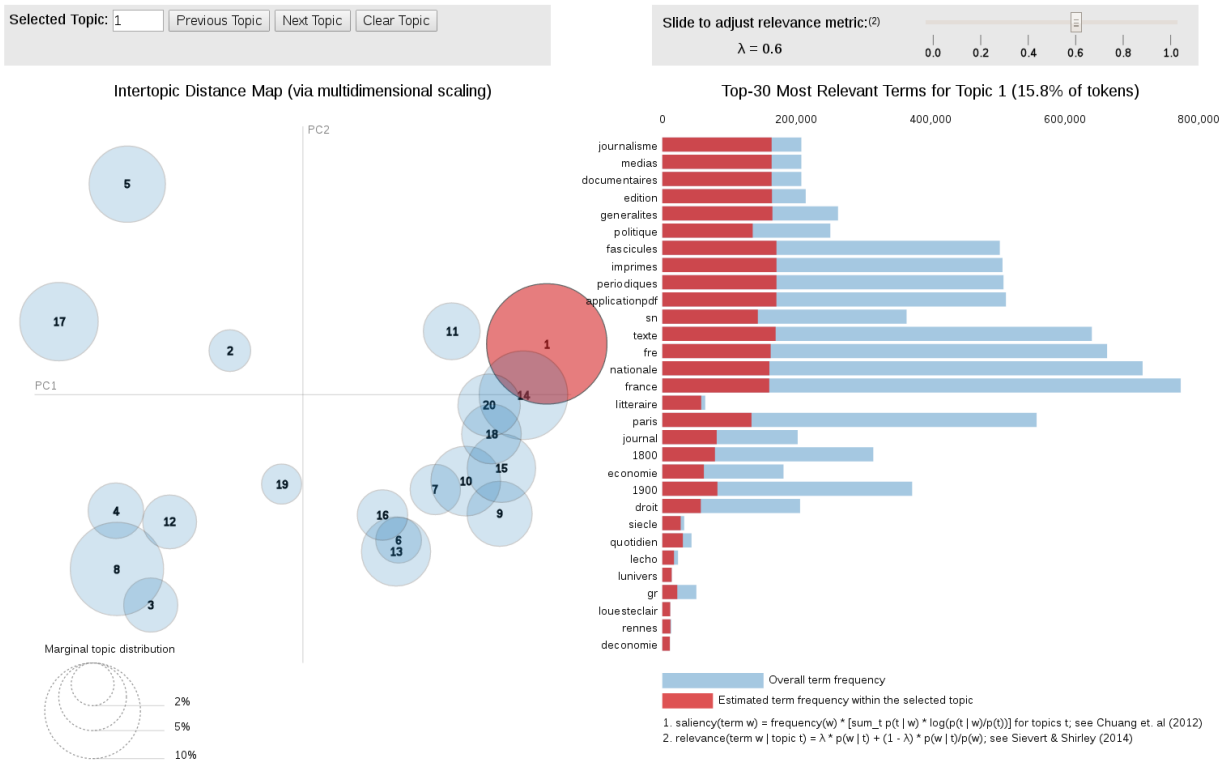

FIGURE D.2 – Distribution des mots pour le cluster 1

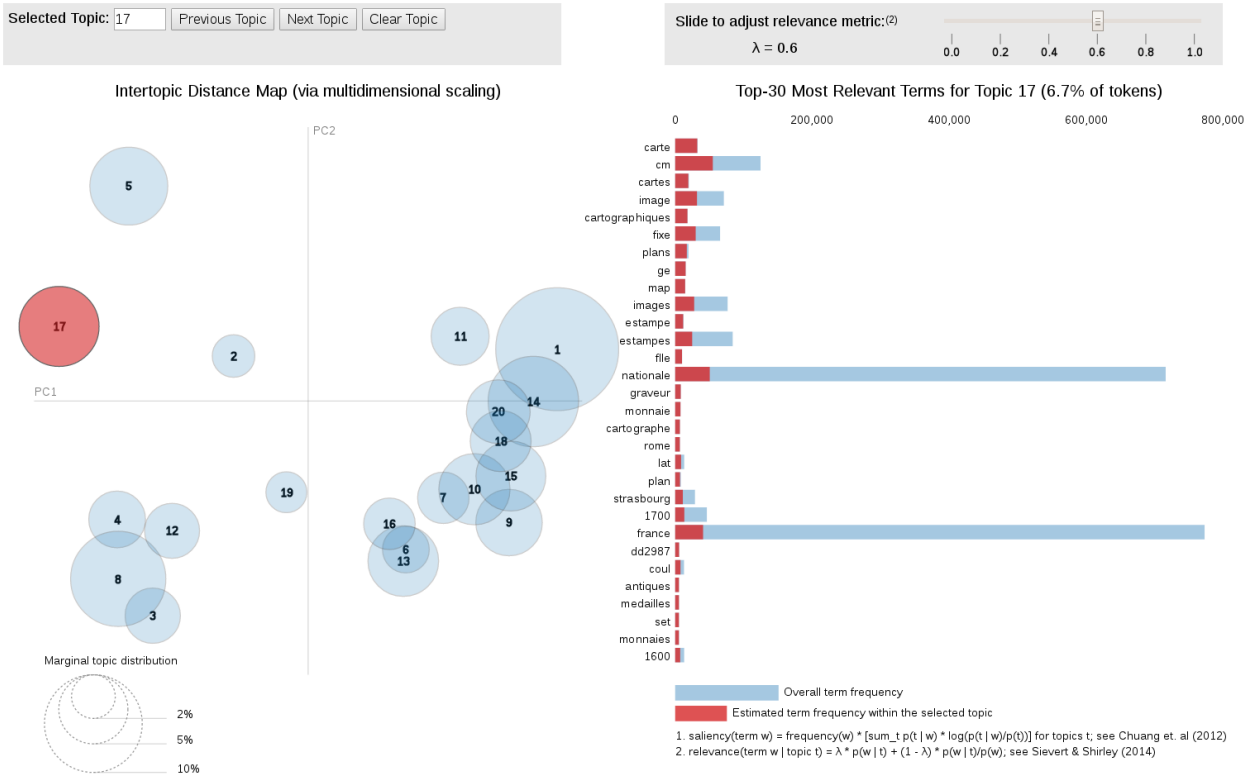

FIGURE D.3 – Distribution des mots pour le cluster 17

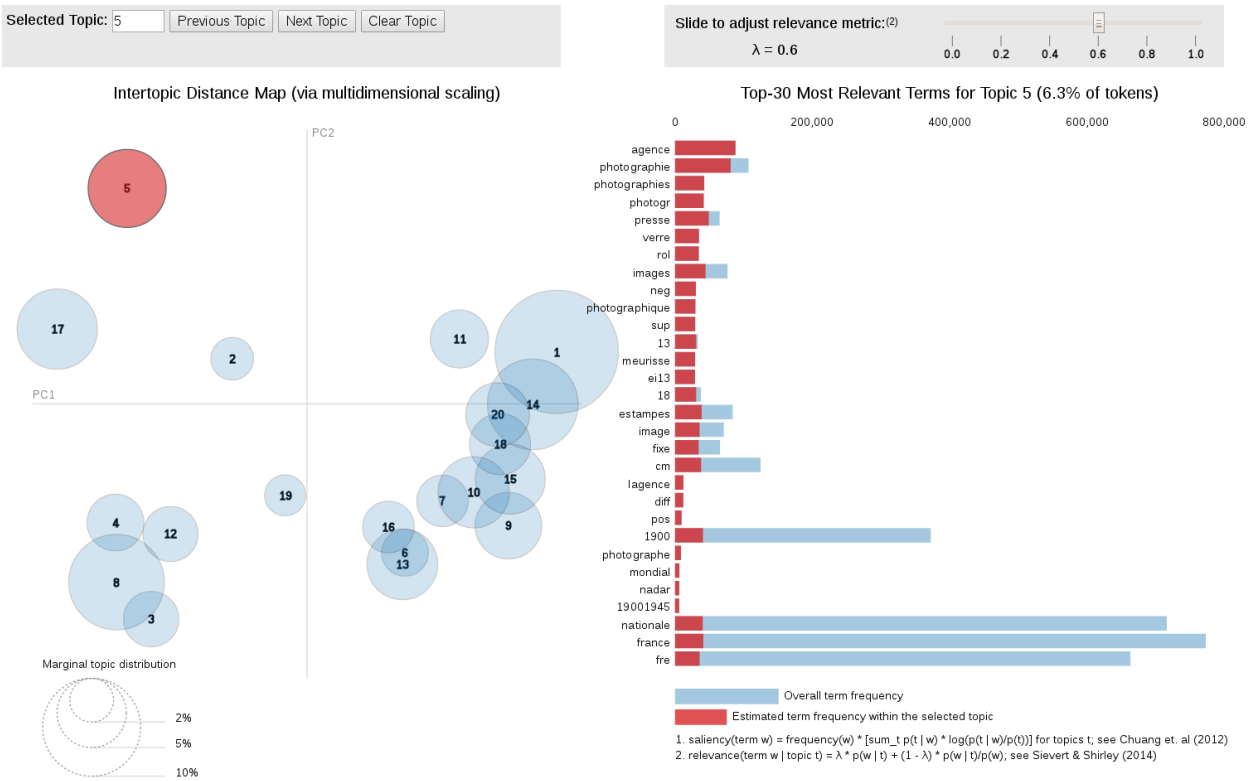

FIGURE D.4 – Distribution des mots pour le cluster 5

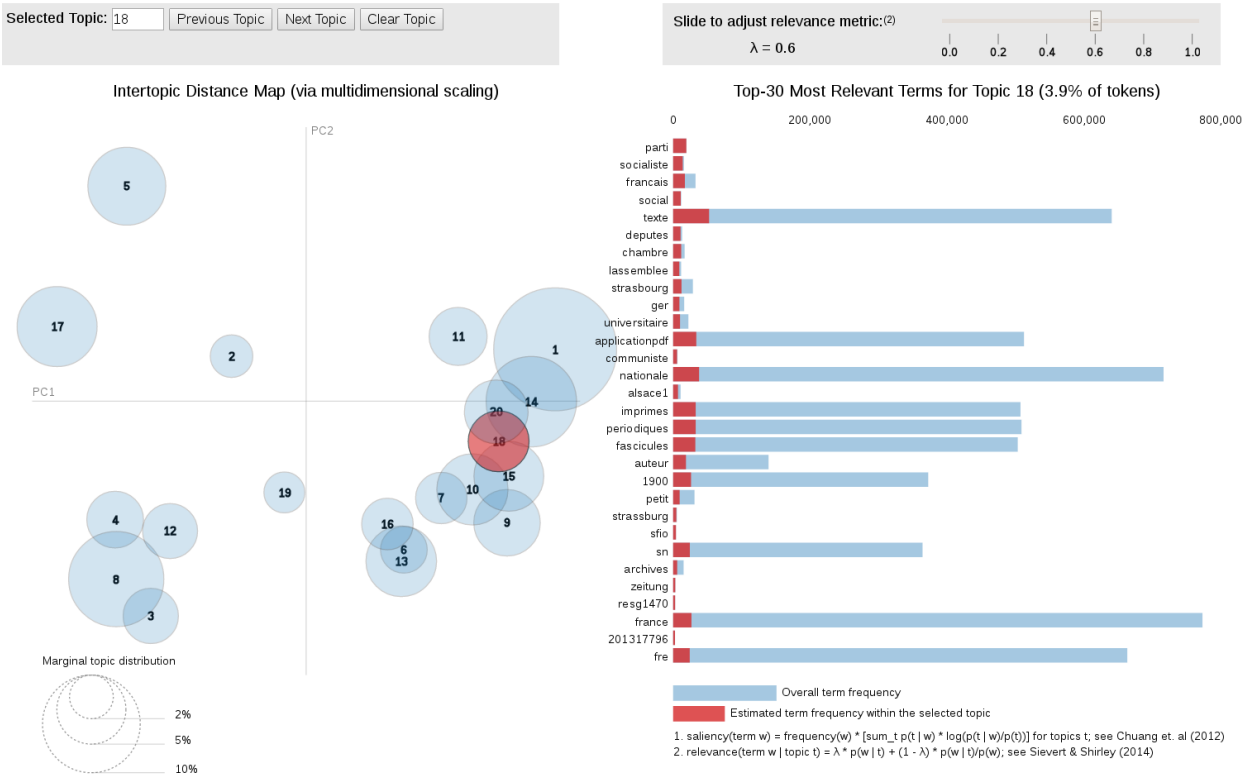

FIGURE D.5 – Distribution des mots pour le cluster 18

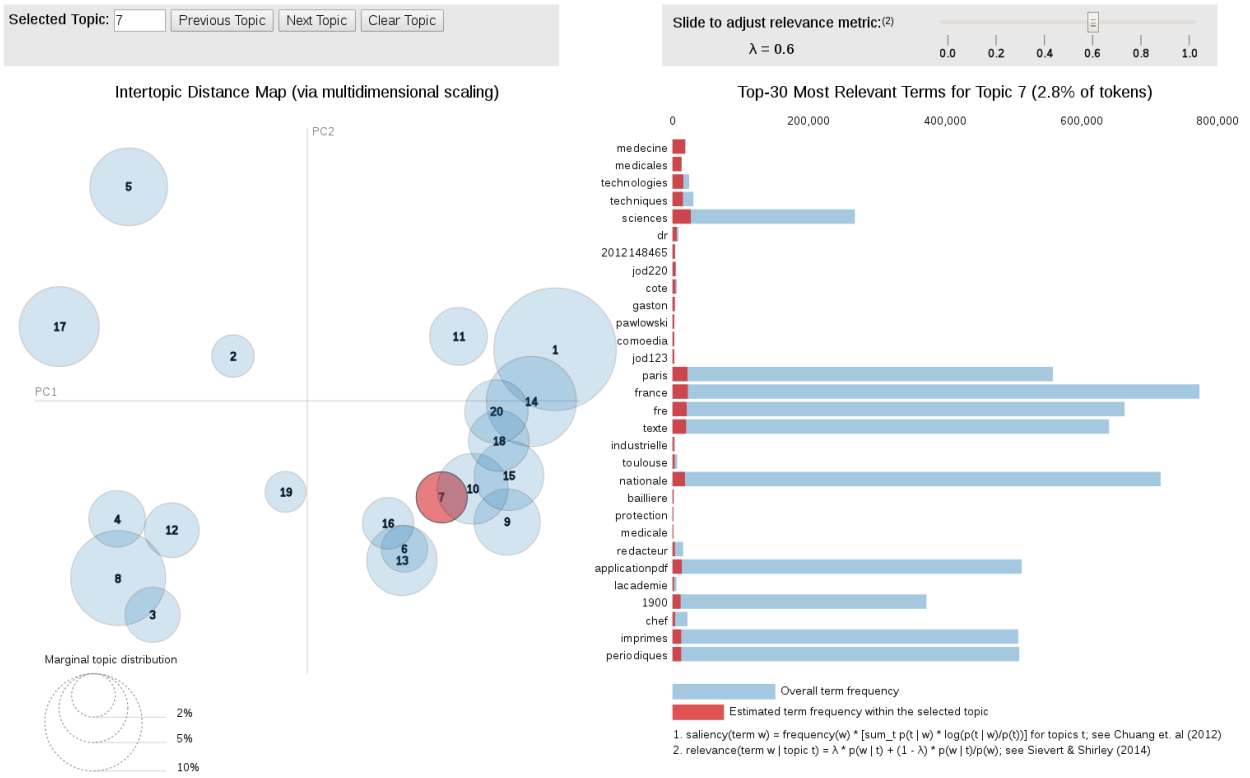

FIGURE D.6 – Distribution des mots pour le cluster 7

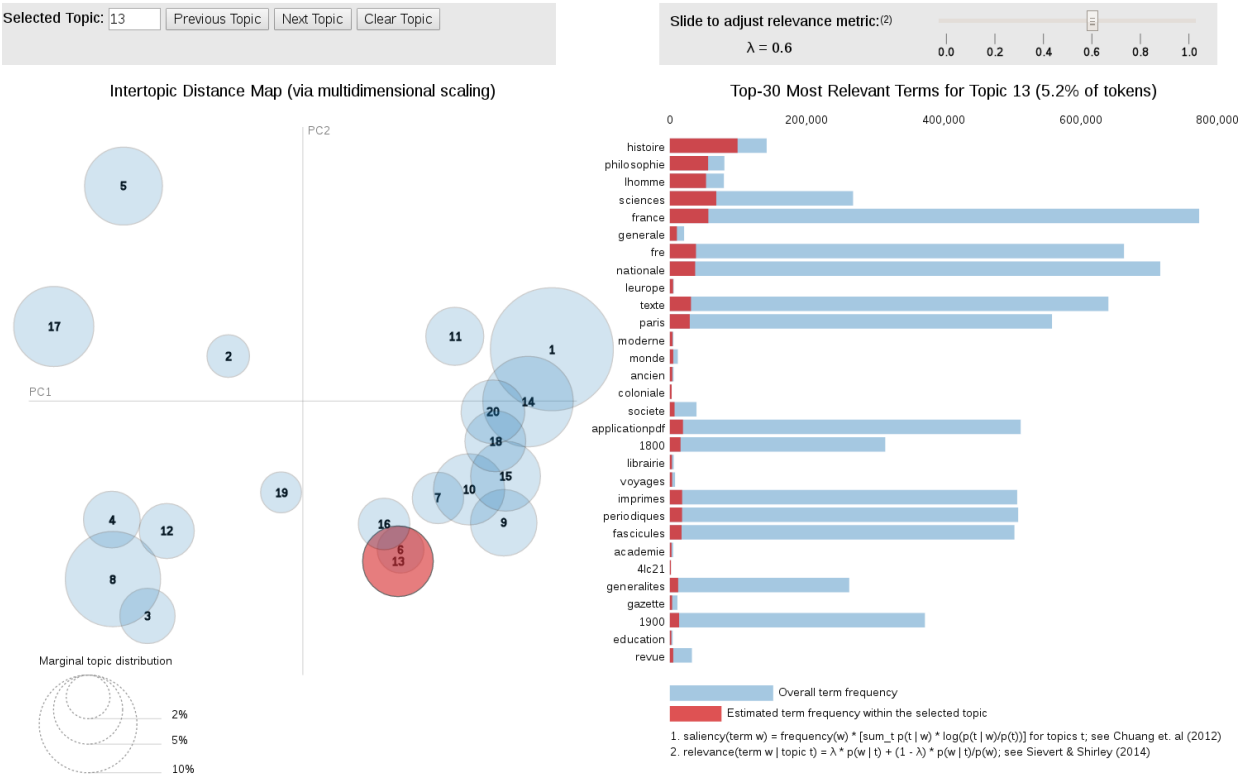

FIGURE D.7 – Distribution des mots pour le cluster 13