Analyse des logs de Gallica pour cartographier les profils des gallicanautes [a]

Philippe Chevallier (BNF), François Roueff (Télécom ParisTech)

a. D'après Nouvellet, Beaudouin, D'Alché-Buc, Prieur, and Roueff [2017]

Plan

Objectifs et données brutes

Mise en forme

Ce qu'il faut retenir

Objectifs et données brutes

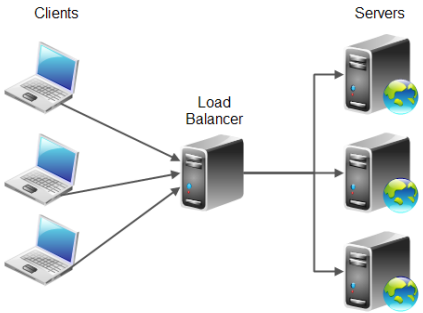
Mise en forme

Ce qu'il faut retenir

Objectifs

- Objectif général : améliorer le service de mise à disposition des ressources numériques aux *Gallicanautes*.
- Contexte : compléter des études et observations de terrain.
- Contrainte : exploration de toutes les données brutes disponibles (sur une durée donnée).
- Approche : modélisation des comportements d'utilisation des sites de la BnF.
- Méthodologie : classification non-supervisée des sessions vues comme séquences d'actions.
- Outils : Plateforme Teralab, programmation python, indexation par Elasticsearch.

Logs : architecture de Gallica/DataBnf



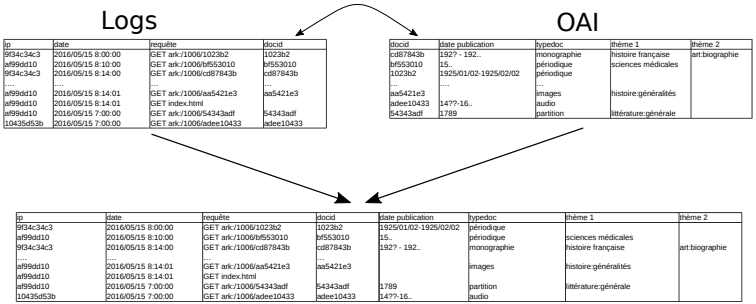
- Le Load Balancer redistribue les connexions entre les différents serveurs.
- Logs au niveau des 30 serveurs Apache.

Logs : format

Format des logs :

- IP : peut être partagée par des utilisateurs d'une même institution
- pays, ville
- date
- requête : ark/id unique, page static, moteur de recherche, ...
- referer : indique le site de provenance
- user-agent : information sur le navigateur de l'utilisateur

Enrichissement des données : OAI



ip	date	requête	docid
9f34c34c3	2016/05/15 8:00:00	GET ark:/1006/1023b2	1023b2
af99dd10	2016/05/15 8:10:00	GET ark:/1006/bf553010	bf553010
9f34c34c3	2016/05/15 8:14:00	GET ark:/1006/cd87843b	cd87843b
...	....	...	...
af99dd10	2016/05/15 8:14:01	GET ark:/1006/aa5421e3	aa5421e3
af99dd10	2016/05/15 8:14:01	GET index.html	
af99dd10	2016/05/15 7:00:00	GET ark:/1006/54343adf	54343adf
10435d53b	2016/05/15 7:00:00	GET ark:/1006/adee10433	adee10433

docid	date publication	typedoc	thème 1	thème 2
cd87843b	192? - 192..	monographie	histoire française	art:biographie
bf553010	15..	périodique	sciences médicales	
1023b2	1925/01/02-1925/02/02	périodique		
...	....	...		
aa5421e3		images	histoire:généralités	
adee10433	14??-16..	audio		
54343adf	1789	partition	littérature:générale	

ip	date	requête	docid	date publication	typedoc	thème 1	thème 2
9f34c34c3	2016/05/15 8:00:00	GET ark:/1006/1023b2	1023b2	1925/01/02-1925/02/02	périodique		
af99dd10	2016/05/15 8:10:00	GET ark:/1006/bf553010	bf553010	15..	périodique	sciences médicales	
9f34c34c3	2016/05/15 8:14:00	GET ark:/1006/cd87843b	cd87843b	192? - 192..	monographie	histoire française	art:biographie
...	....	...	...				
af99dd10	2016/05/15 8:14:01	GET ark:/1006/aa5421e3	aa5421e3		images	histoire:généralités	
af99dd10	2016/05/15 8:14:01	GET index.html					
af99dd10	2016/05/15 7:00:00	GET ark:/1006/54343adf	54343adf	1789	partition	littérature:générale	
10435d53b	2016/05/15 7:00:00	GET ark:/1006/adee10433	adee10433	14??-16..	audio		

Moissonnage et nettoyage de l'entrepôt OAI :

- titre, auteur.
- date de publication.
- catégories

Tri des données de log

- Une session est une séquence de requêtes faite par un unique usager durant la visite d'un site particulier.
- La *sessionisation* est l'étape permettant de trier l'ensemble des requêtes en sessions.
  - Séquence de requêtes présentant la même IP
  - Nouvelle session lorsque l'IP devient inactive pendant 60min [1]
- Un robot internet est un logiciel qui réalise des tâches automatisées sur internet
  - "good bots" (robot d'indexation, droit d'auteur)
  - "bad bots" (robot moisonneur, botnet)

1. cf. enquête quali.

Objectifs et données brutes

Mise en forme

Ce qu'il faut retenir

Mise en forme/mise en “modèle”

- Mise en forme basique des données : collection de sessions, chacune vue comme une succession d’événements (plus ou moins riches) estampillés temporellement.
- Nécessité d’un modèle pour une classification non-supervisée permettant une interprétation [2]
  - Exemple 1 : suite ordonnée d’actions de l’utilisateur.
  - Exemple 1 enrichi : ajout d’une action basée sur le referer.
  - Exemple 2 : suite d’actions associées à des durées.

2. cf. enquête quali.

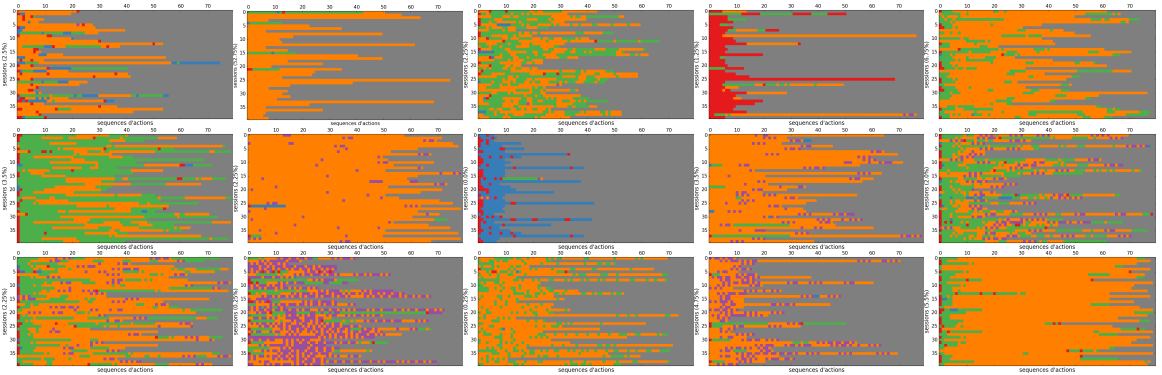
Exemple 1 : actions

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...) :

Une session de 10 actions

Représentation sous forme d'une "ligne" :

Exemple 1 : classification à 15 classes



Objectifs et données brutes

Mise en forme

Ce qu'il faut retenir

Importance de l'analyse des logs pour la connaissance des usages

- en complément (et non substitution) des autres méthodes
- prend en compte tous les usages, en particulier les plus difficiles à capter dans les enquêtes (usages "furtifs")
- permet une approche à la fois fine et quantifiée des parcours que le déclaratif rend imparfaitement
- ouvre des pistes d'amélioration concrètes en comparant le parcours idéal (pensé par le producteur) et le parcours réel (de l'usager)

Quelle place pour l'analyse de traces d'usage à la BnF ?

- limite des outils propriétaires (mesure d'audience, e.réputation, etc.) : pas d'accès aux données brutes ; requêtes prédéfinies
- choix fort d'en faire un objet de recherche avec Télécom ParisTech : transparence des traitements / modèles évolutifs / interprétation collective
- nécessite un dialogue entre :
  - les producteurs des données (DSI de la BnF)
  - les experts de la collection numérique (bibliothécaires de la BnF)
  - les chercheurs en science des données (département IDS de Télécom)
  - les chercheurs en sciences sociales (département SES de Télécom
- souhait de poursuivre ce type de traitements, en particulier pour mesurer l'impact des évolutions de l'interface sur les comportements

Bibliographie I

Adrien Nouvellet, Valérie Beaudouin, Florence D'Alché-Buc, Christophe Prieur, and François Roueff. Analyse des traces d'usage de gallica. Technical report, Télécom ParisTech, 2017.