Comprendre les usages des corpus numérisés, approches fondées sur l'analyse des traces numériques

15 mars 2021 - Table ronde : regards croisés sur les défis et les enjeux de la mise à disposition de larges corpus numérisés

Cyrille Suire

Laboratoire L3i, La Rochelle Université

Enjeux

Pour la recherche :

- trouver et corriger les biais
- adapter l'algorithmique aux usages

Pour les fournisseurs de contenus :

- améliorer l'expérience utilisateur
- fournir de nouveaux services pertinents

Sources de biais

Production des contenus

- numérisation / OCR
- extraction de connaissance
- indexation / recherche
- ...

Comportement de recherche

- requêtes
- analyse des résultats
- filtrage
- ...

Sources de biais (exemple indexation)

ST	À	LA	RECHERCHE	DES	FEMMES	DANS	LA	PRESSE	ANCIENNE	NUMÉRISÉE
EF	À	LA	RECHERCHE	DES	FEMMES	DANS	LA	PRESSE	ANCIENNE	NUMÉRISÉE
SGF	À	LA	RECHERCHE	DES	FEMMES	DANS	LA	PRESSE	ANCIENNE	NUMÉRISÉE
FGF	À	LA	RECHERCHE	DES	FEMMES	DANS	LA	PRESSE	ANCIENNE	NUMÉRISÉE
LCF	à	la	recherche	des	femmes	dans	la	presse	ancienne	numérisée
SF			recherche		femmes			presse	ancienne	numérisée
FLSF			recherch		feme			pres	ancien	numeris

Exemple de prétraitement du texte pour l'indexation

Sources de biais (exemple indexation)

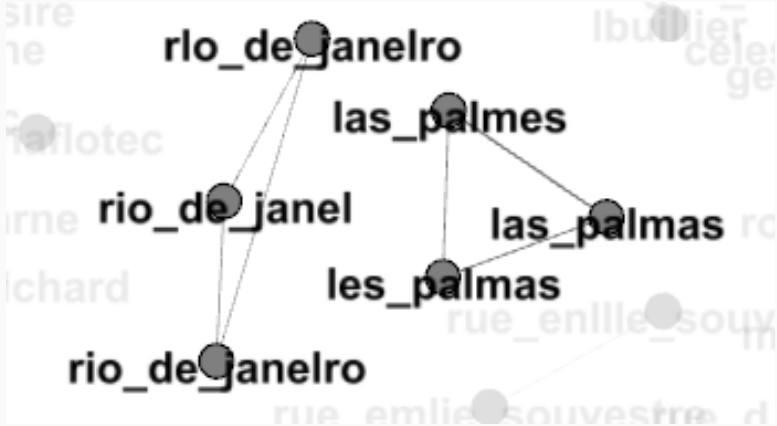
ST	What’s	Past	is	Prologue	The	NewsEye	International	Conference
EF	What’s	Past	is	Prologue	The	NewsEye	International	Conference
SGF	What’s	Past	is	Prologue	The	NewsEye	International	Conference
FGF	What’s	Past	is	Prologue	The	NewsEye	International	Conference
LCF	what’s	past	is	prologue	the	newseye	international	conference
SF	what’s	past	is	prologue	the	newseye	international	conference
FLSF	what’	past	is	prologu	the	newsey	international	conferenc

Exemple de prétraitement du texte pour l’indexation (texte en anglais, indexation en français)

Sources de biais (entités nommées)

Les stratégies et techniques de recherche d'information employées par les utilisateurs favorisent certains biais.

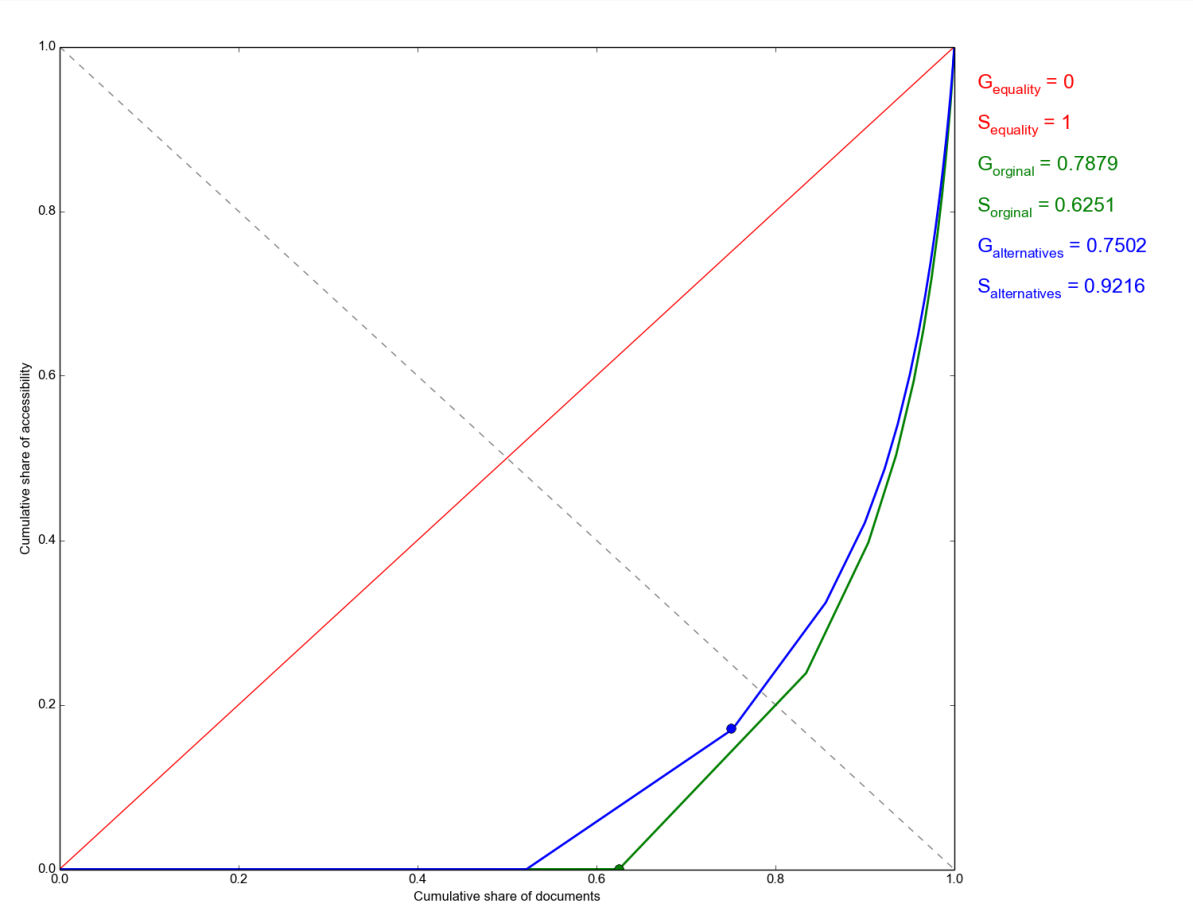
- Recherche par entités nommées



Exemple d'entités nommées mal reconnues

Sources de biais (exploitation des résultats)

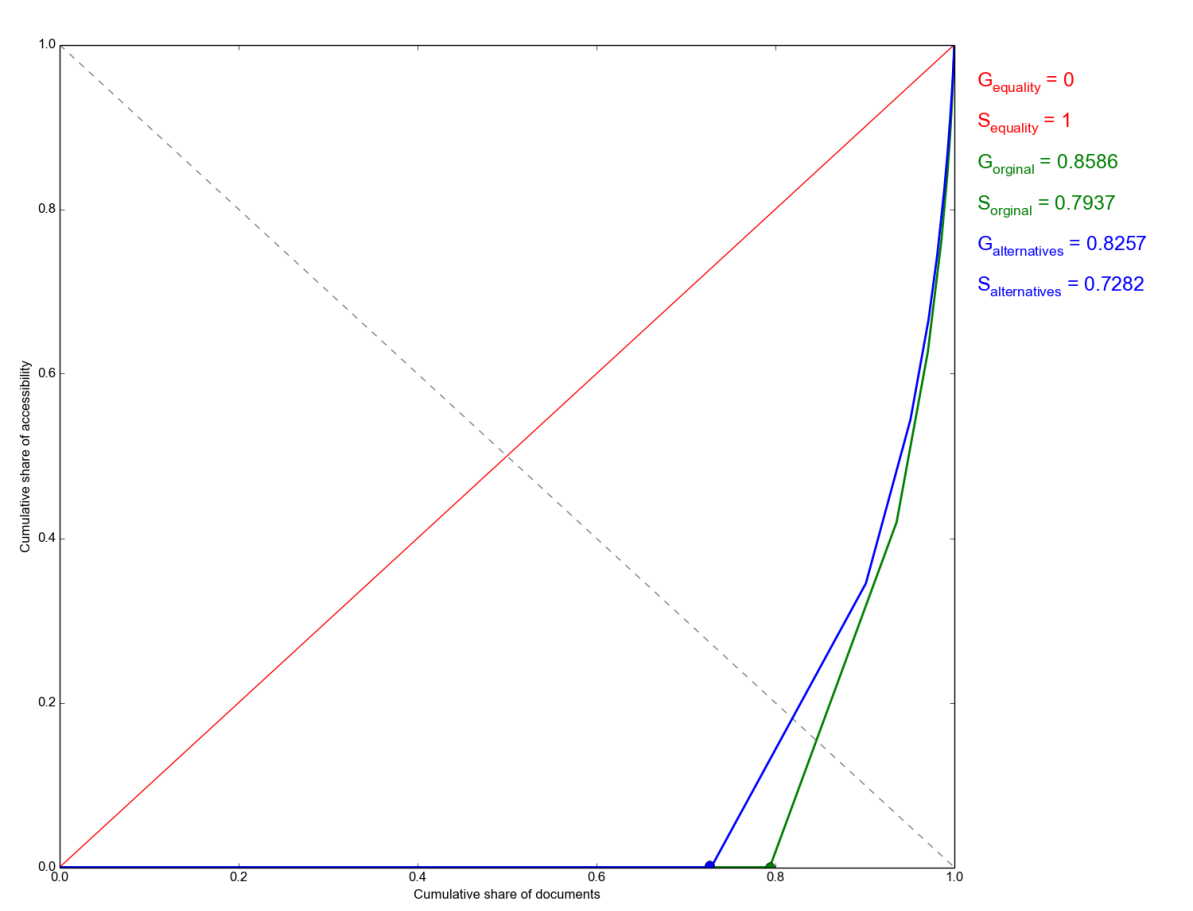
- Analyse des résultats, illusion de l'exhaustivité



Accessibilité d'un corpus de documents (n=50)

Sources de biais (comportement de recherche d'information)

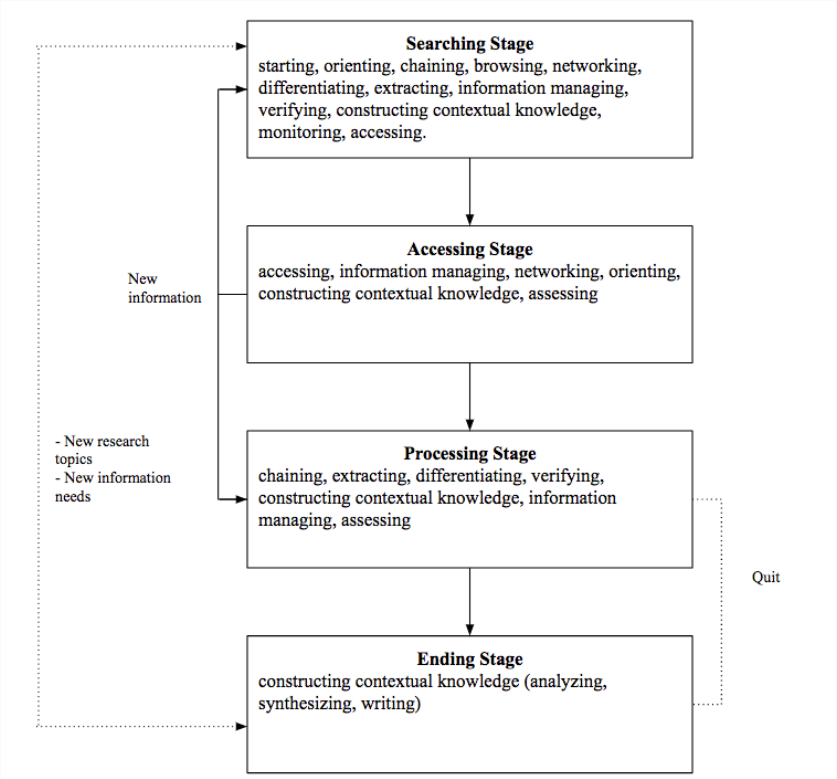
- Analyse des résultats, illusion de l'exhaustivité



Accessibilité d'un corpus de documents (n=10)

Etudier le comportement des utilisateurs, théorie

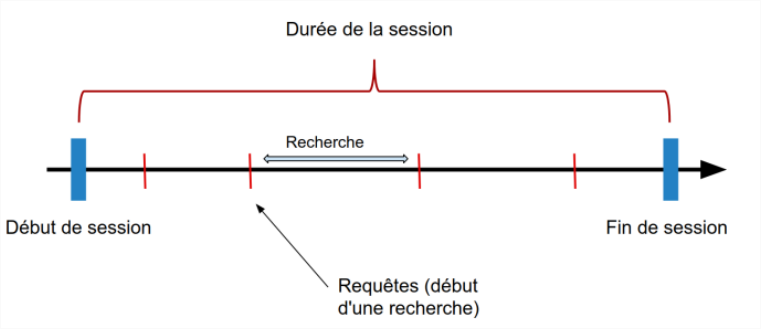
Pour étudier le comportement de recherche d'information des utilisateurs, nous disposons de modèles théoriques.



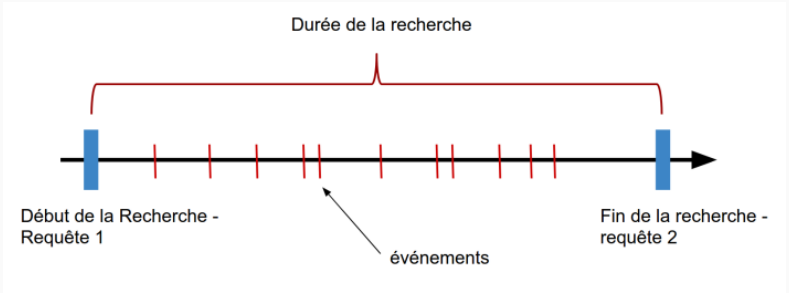
Exemple d'un modèle théorique (modèle de Rhee)

Etudier le comportement des utilisateurs, approche statistique

Objectif : trouver et valider des indicateurs permettant de discriminer des types d'activités de recherche d'information.



Session de recherche



Détail d'une recherche

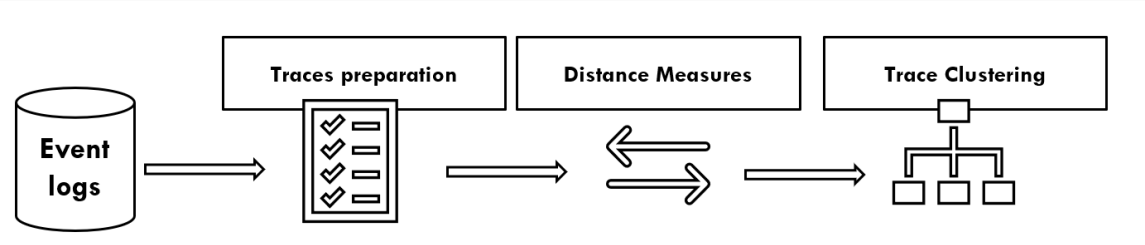
Etudier le comportement des utilisateurs, approche statistique

	intégralité de la tâche (f)		
	T4		
	T1	T2	T3
longueur des requêtes (f_1)			
p	*	***	***
Z	-2.56	-4.05	-4.08
durée de la session (f_2)			
p	***	***	***
Z	-4.33	-4.55	-4.53
docs. visibles (f_3)			
p	**	0.18	*
Z	-2.64	-1.34	-2.09
docs. sélectionnés (f_4)			
p	**	***	***
Z	-3.21	-3.42	-4.2
position des docs. sélectionnés (f_5)			
p	***	***	***
Z	-3.39	-3.92	-3.51
docs. consultés(f_6)			
p	**	**	***
Z	-3.63	-3.21	-4.04
durée d'exploitation (f_7)			
p	***	0.28	***
Z	-3.39	-1.07	-4.6

*Pertinence statistique d'indicateurs entre différentes tâches de recherche d'information * pour p < 0.05, ** pour p < 0.01, *** pour p < 0.001*

Etudier le comportement des utilisateurs, modélisation de processus

Objectif : trouver des modèles de comportements de recherche d'information et les représenter.

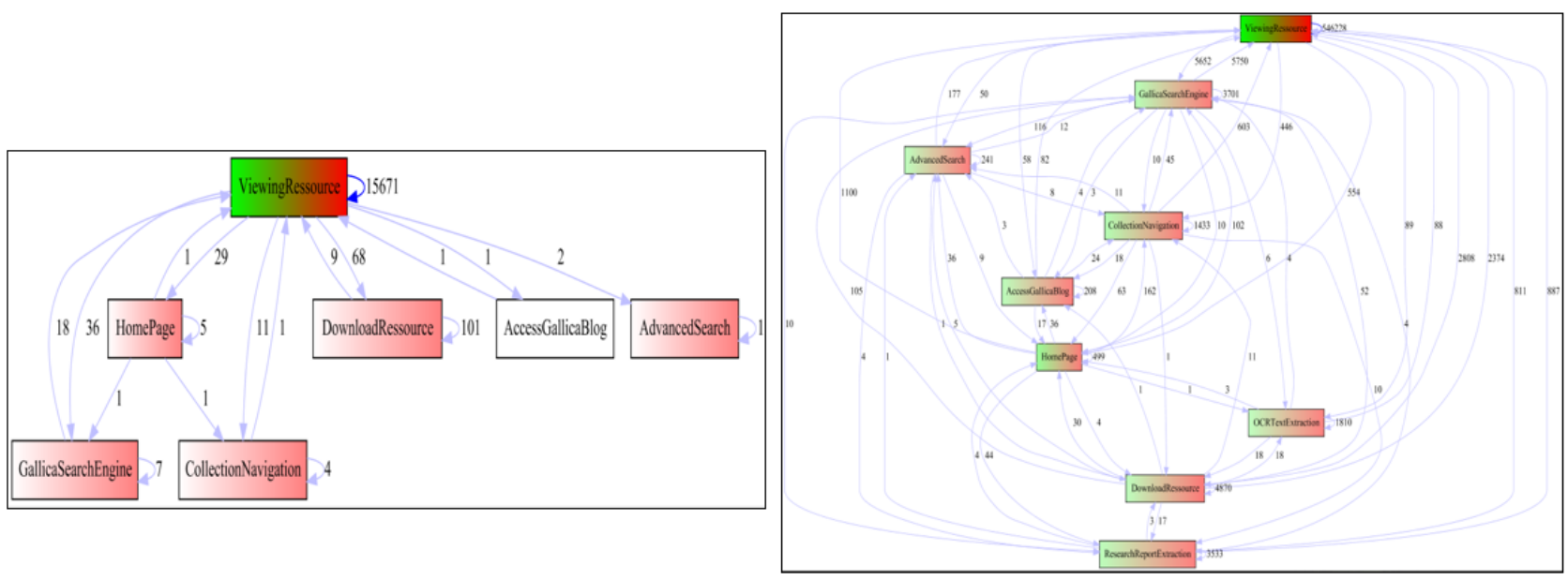


Principe de fonctionnement

```
##1a47161ad98134bf072fe5ea3573fca6##Japan##Tokyo##- - [10/Apr/2017:07:52:14
+0200] "GET /accueil/?mode=desktop HTTP/1.1" 200 11311 "-" "Mozilla/5.0
(Macintosh; Intel Mac OS X 10_12_4) AppleWebKit/603.1.30 (KHTML, like Gecko)
Version/10.1 Safari/603.1.30" "JSESSIONID=AD97; xtidc=1605; xtan18798=-;
xtant18798=1; rxVisitor=1479; xtvrn=$18798$" 50886
```

Exemple de log

Etudier le comportement des utilisateurs, modélisation de processus



Exemple de sortie de la méthode