Vers une recherche d'information adaptative

Exploitation du contexte des requêtes et des documents dans l'environnement *Revues.org*

Clémentine Adam et Simon Leva
{clementine.adam,simon.leva}@univ-tlse2.fr

CLLE-ERSS & Université de Toulouse 2

21/10/2013

Plan

Plan

De la recherche d'information orientée système...

Principe

- L'utilisateur exprime son besoin d'information à travers la formulation d'une requête
- À l'aide d'un modèle, le système identifie dans la collection de documents les informations (documents) pertinentes en réponse à cette requête

- Approche initiale du domaine, apparue dans les années 1960
  - Méthodes de représentation des requêtes et des documents ?
  - Algorithmes d'appariement entre requêtes et documents ?
  - Méthodologies d'évaluation de l'efficacité de ces modèles ?
- Systèmes de recherche d'information conçus pour tout type d'utilisateur, de tâche de recherche et de document

. . . à la recherche d'information orientée contexte

Principe

- Influence du contexte et de la dimension cognitive de l'utilisateur au cours de l'interaction avec le système
- Des changements dans notre contexte de recherche peuvent influer sur notre besoin d'information
- Remise de l'utilisateur au centre du processus de recherche

- Généralisation du *World Wide Web* dans les années 1990
  - Méthodes de représentation des requêtes et des documents intégrant une vision cognitive de leur contenu ?
  - Modèles d'accès contextuel et personnalisé à l'information ?
  - Pertinence système *vs.* pertinence utilisateur ?
- Systèmes capables de reconnaître des contextes différents et de s'adapter à nos besoins et à nos comportements

Quels éléments de contexte considérer ?

1. **Contexte de l'utilisateur :** contexte personnel (âge, expertise, centres d'intérêts...) ou social (individu *vs.* communauté)
2. **Tâche de la recherche :** définie au début de la recherche et susceptible d'évoluer au cours de la recherche
3. **Contexte spatio-temporel :** lieu et heure de la recherche
4. **Contexte de l'information :** contexte direct de l'information (métadonnées) et source de l'information
5. **Mais aussi :** environnement de la recherche (privé *vs.* professionnel), moyen d'accès à l'information (ordinateur *vs.* téléphone portable), contexte psychologique de l'utilisateur, contexte de l'interaction avec le système...

Pour quels usages ?

Hypothèse 1

L'information contextuelle est tros riche et variée pour être traitée par la génération actuelle d'ordinateurs

- Le contexte pertinent n'est accessible qu'aux humains
- Les systèmes peuvent expliciter leur fonctionnement ou exploiter les interactions d'autres utilisateurs

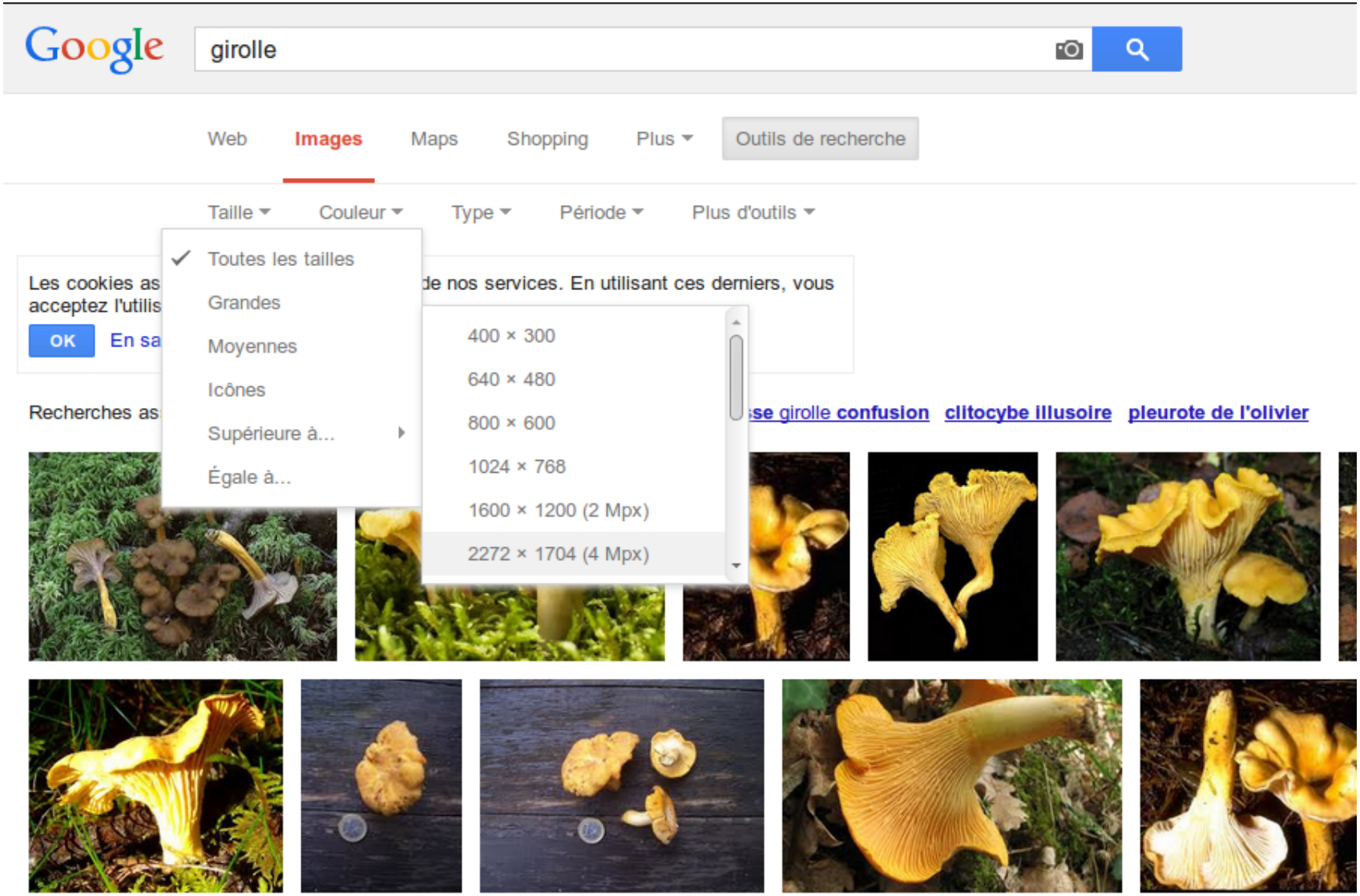
Hypothèse 2

Les systèmes contextuels peuvent directement exploiter des éléments de contexte dans leur fonctionnement

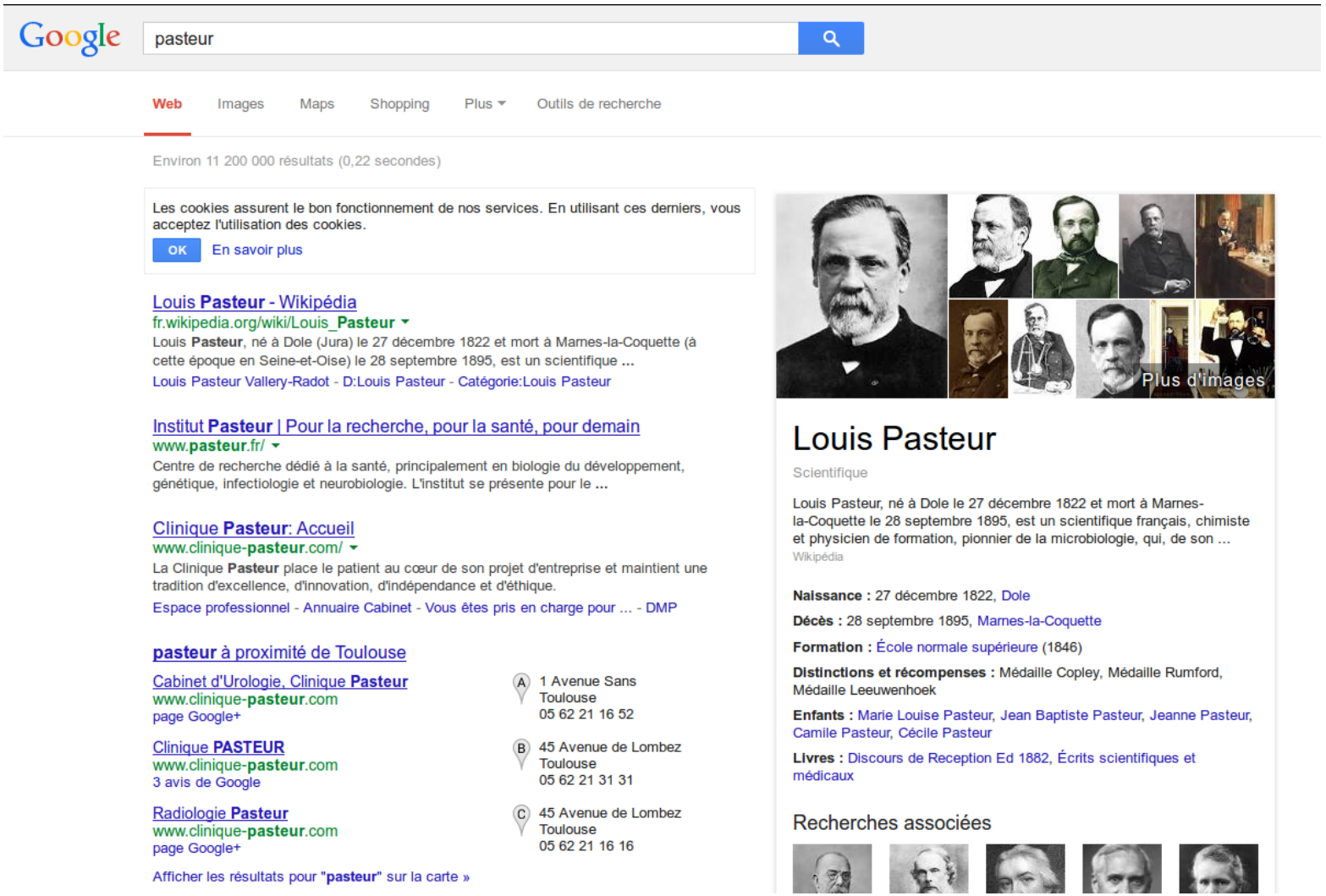
- Décisions du système incorporant des contraintes contextuelles
- Restriction du contexte à la tâche de la recherche

Exemple : exploitation des interactions précédentes

Exemple : exploitation des métadonnées



Exemple : tâche visant une information spécifique



Exploitation du contexte de l'interaction

Principe

Au cours du processus de recherche, l'utilisateur transmet de nombreuses informations au moteur de recherche à travers ses interactions avec le navigateur

- **Indicateurs explicites :** requêtes, clic sur des résultats...
- **Indicateurs implicites :** temps passé sur les pages de résultat, comportement de navigation...

- Inférence de l'objectif de l'utilisateur à partir de ses actions et de son comportement de recherche
  - Succès/difficulté/échec de la recherche ?
  - Type de relation entre une action et l'objectif de l'utilisateur ?
- Étude statistique des interactions entre utilisateur et système

Plan

CAAS : Contextual Analysis and Adaptive Search

Hypothèse

- Le recours au contexte peut améliorer les performances d'un système de recherche d'information et expliquer certains aspects du processus de recherche
- Les éléments de contexte considérés concernent les requêtes, les documents et le système de recherche d'information

Objectifs

- **Contrôle de la variété des contextes :** modèles représentant les divers aspects du contexte en RI
- **Reconnaissance et adaptation au contexte :** exploitation des modèles afin de choisir la meilleure stratégie de recherche en fonction du contexte détecté

Éléments de contexte

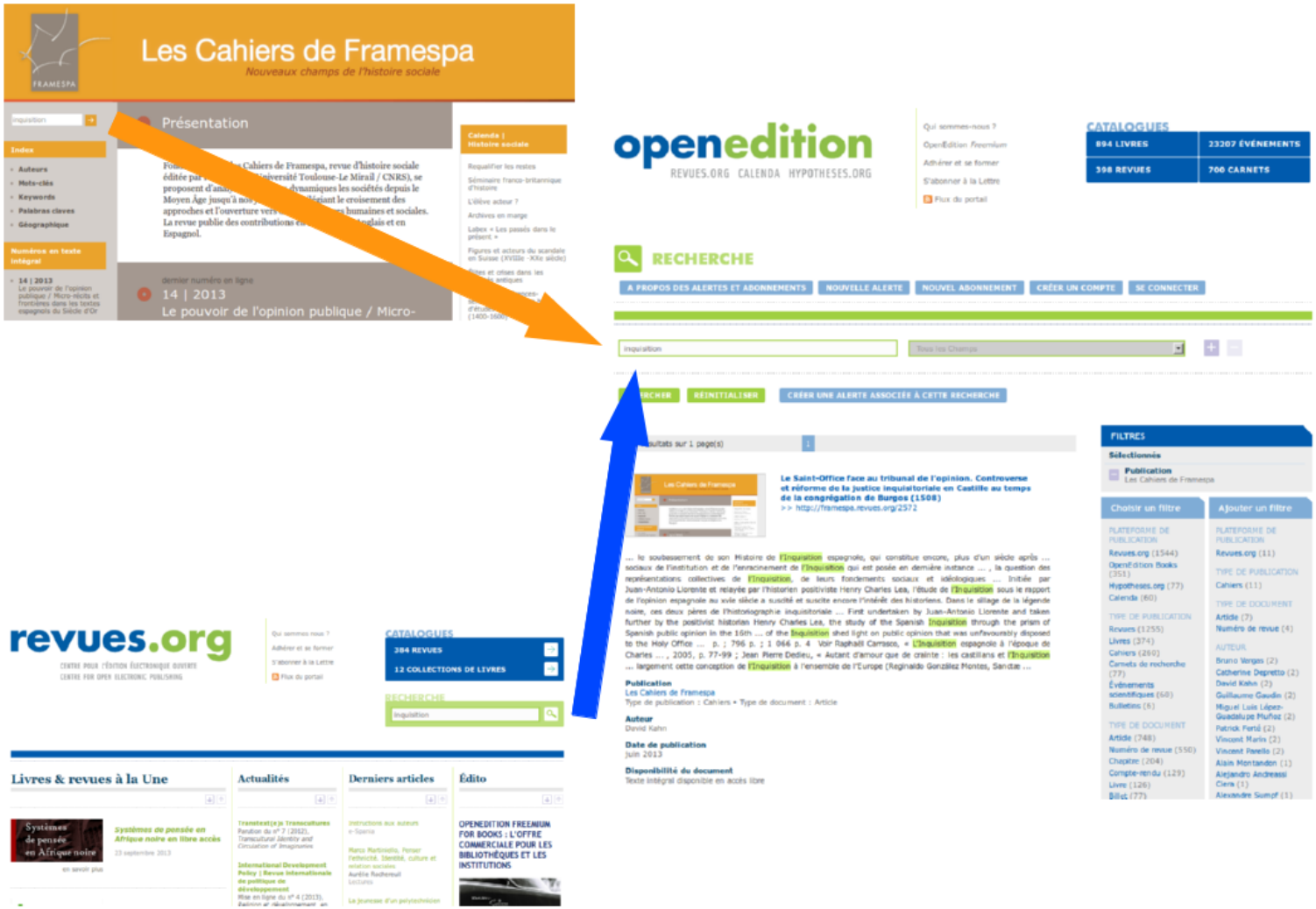
1. **Variété des requêtes**
   - Inférence des attentes de l'utilisateur à partir de ses requêtes ?
   - Corrélation entre des types de requêtes et des traits portant sur les requêtes afin de détecter des catégories
   - Corrélation entre des requêtes et les performances du système afin de prédire les requêtes difficiles
2. **Variété des documents**
   - Détection de traits portant sur les documents permettant de prédire leur pertinence par rapport à une requête soumise dans un contexte donné
3. **Variété des systèmes**
   - Évaluation des performances du système en fonction des traitements appliqués et du type de requête considéré

Plan

Le portail *OpenEdition*

- Développé par le Centre pour l'édition électronique ouverte
- Libre accès à un ensemble de ressources électroniques dans le domaine des sciences humaines et sociales
  - *Revues.org* : 384 revues, 12 collections de livres
  - *Calenda* : 23 306 évènements scientifiques
  - *Hypotheses* : 700 blogs et carnets de recherche
  - *OpenEdition Books :* 896 livres
- 30 millions de consultations en 2012
- Visiteurs provenant essentiellement du monde francophone, des États-Unis et d'Amérique du sud

Un environnement de recherche à double entrée



Recherche par champs et par filtres

Journal de requêtes (1)

1. **Log d'accès initial**
   - 46 millions de transactions avec le domaine *OpenEdition* du 07 avril 2010 au 1er février 2012
2. **Opérations de nettoyage et de filtrage**
   - Élimination des informations inexploitables ou erronées
   - Élimination des informations non pertinentes
3. **Informations disponibles et exploitables**
   - Identifiant de l'utilisateur
   - Date et heure de soumission de la requête
   - Texte de la requête
   - Origine de la requête
   - Paramétrage éventuel du moteur de recherche par l'utilisateur
4. **Log de requêtes final**
   - 1 057 471 requêtes / 227 302 utilisateurs
   - Requêtes en français, mais aussi en anglais ou espagnol

Journal de requêtes (2)

Utilisateur	Date et heure	Requête
130.104.10005	2012/01/25:16:02:41	temps
130.104.10005	2012/02/01:12:08:46	cahiers des sciences humaine
130.104.10005	2012/02/01:12:08:50	cahiers des sciences humaines
130.104.100926	2011/05/11:15:48:15	censure france
130.104.100926	2011/05/11:15:49:30	censure france presse
130.104.100926	2011/05/11:15:53:55	liberté de la presse france
130.104.10105	2011/12/07:00:36:12	absentéisme
130.104.10105	2011/12/07:00:42:34	absentéisme étudiant

Collection de documents

- Prise en compte uniquement des données de Revues.org

→ articles de sciences humaines et sociales

- Corpus de 253 millions de mots (62 204 documents)

Plan

Requêtes navigationnelles (1)

Définition

Requêtes visant à accéder à un document précis ou à une ressource particulière appartenant à la collection *Revues.org*

- **Références**
  - `Alexandre Lacassagne, Le médecin devant les cours d'assises, Paris, s.n., 1883`
  - `M. Tozy et M. Mahdi , « Aspects du droit communautaire dans l'Atlas marocain » , Droit et Société , 1990 , p. 219-227 .`
- **Titres**
  - `Produits locaux entre nature et culture : de la ferme voisine au terroir. Entretien avec Laurence Bérard`
  - `« L'Ovide moralisé et la tradition encyclopédique. Une approche générique comparative »`

Requêtes navigationnelles (2)

- **Auteurs**
  - corinne mélis
  - Gelinier O., Simon F.-X., Billard
- **Noms de revues**
  - Bulletin de la Société Préhistorique Française
  - ilcea
- **Extraits de documents**
  - Pour un débat sur la notion d'ethnie, voir le Fait ethnique en Iran et en Afghanistan, sous la direction de J. P Digard, éd. Du CNRS, 1988
  - "Maintenant que le théâtre est accessible sous forme de film, les théâtres vous paraissent-ils inutiles ?"

Requêtes informationnelles

Définition

Requêtes visant à accéder à une information contenue dans un ou plusieurs documents de la collection *Revues.org*

- **Thématiques générales**
  - `herméneutique`
  - `la femme à l'époque moderne`
- **Thématiques spécifiques**
  - `la loi n2000-516 du 15 juin 2000 renforçant la protection de la présomption d'innocence et des droits des victimes`
  - `Quelle citoyenneté pour les travailleurs migrants en République Populaire de Chine?`
- **Couplage thème/auteur**
  - `chiasme Merleau-Ponty`
  - `formes élémentaires vie religieuses durkheim`

Requêtes transactionnelles ?

Définition

Requêtes visant à obtenir une information sous une forme particulière

- `harris bibliographie`
- `biographie de maupassant`
- `quand les femmes ont pu voté pour la première fois?`
- `localisation des Araucans ou bien des Mapuches`
- `définition de la mobilité`
- `photo des enfants égyptiens allant travailler`

Plan

Problématique

- Les requêtes soumises par les utilisateurs d'un moteur de recherche comportent en moyenne 2 ou 3 mots-clés
  - Besoin d'information insuffisamment explicite par rapport à la collection de documents [Silverstein *et al.*, 1999]
  - Les requêtes tendent à être trop génériques ou trop spécifiques [Downey *et al.*, 2007]
  - Importance de la reformulation de requêtes
- Les requêtes soumises ne sont pas isolées, mais font partie d'une session de recherche
  - Une session fournit de nombreux indices sur le contexte de la recherche
- Construction d'une collection de référence manuellement annotée au niveau des sessions
  - Évaluation de méthodes de détection automatique des sessions

Définition des sessions : structure séquentielle

Définition [Silverstein *et al.*, 1999]

A session is a series of queries by a single user made whithin a small range of time. A session is meant to capture a single user's attempt to fill a single information need.

- Organisation des requêtes successives en séquences
- Caractérisation des sessions par leur longueur
- Notion d'épisode de recherche [Gayo Avello, 2009]

Utilisateur	Temps entre requêtes	Requête	Épisode	Session
158	0	masque	1	1
	339	double		2
	9107	exploitation agricole		3
	443	l'avenir familial de l'exploitation agricole		3

Définition des sessions : structure imbriquée

Problématique de la structure séquentielle des sessions

- Adaptée aux enregistrements des journaux de requêtes
- Ne reflète pas la complexité des parcours de recherche

Utilisateur	Temps entre requêtes	Requête	Épisode	Session
39	0	travail saisonnier	1	1
	630	industries de loisirs		2
	26	parc d'attraction		2
	207	fidélisation et emplois saisonniers		1
	387	travail saisonnier		1
	55	parc à thèmes		2
	31	parc astérix		2
	14	disneyland		2
	4	disneyland paris		2
	6	walibi		2
	7	compagnies des alpes		2

Collections de référence existantes

Auteur	Journal	Requêtes	Utilisateurs	Sessions
(Göker et He, 2000)	Reuters 1999	9 534	1 440	-
(Jansen *et al.*, 2007)	Dogpile 2005	2 000	-	-
(Jones et Klinkner, 2008)	Yahoo ! 2007	8 226	312	2 922
(Gayo Avello, 2009)	Excite 1997 Excite 1999 AlltheWeb 2001 Excite 2001 AltaVista 2002 AOL 2006	95 000	15 000	35 000

Définitions adoptées

Épisode de recherche

- Ensemble des requêtes soumises à un moteur de recherche par un utilisateur donné durant au plus une journée
- Un épisode peut contenir une ou plusieurs sessions de recherche

Session de recherche

- Ensemble des requêtes reliées à un même besoin d'information
- Les requêtes d'une même session peuvent être non séquentielles au sein d'un épisode de recherche

Consignes d'annotation manuelle des sessions

- Une session correspond à un ensemble de requêtes liées
  - Observation de l'ensemble des requêtes soumises avant d'attribuer une session à chaque requête : `barrière de corail` / `nouvelle zélande` / `barrière de corail nouvelle zélande`
- Des requêtes liées à un même besoin d'information implicite au sein d'une session
  - **Indices textuels :** `femmes moralistes` / `femmes moralistes 18 siecle` / `moralistes`
  - **Indices sémantiques :** `foresterie` / `sylviculture`
  - **Proximité thématique :** `théâtre expérimental` / `grotowski`
- Utilisation de sources de connaissance externes
- Une session possède un identifiant unique au sein d'un épisode

Collection constituée pour l'annotation

- Sélection aléatoire de 947 requêtes/216 utilisateurs
- Segmentation automatique en 349 épisodes de recherche

Utilisateur	Requête	Épisode	Session
39	`travail saisonnier`	1	1
	`industries de loisirs`		2
	`parc d'attraction`		2
	`fidélisation et emplois saisonniers`		1
	`travail saisonnier`		1
	`parc à thèmes`		2
	`parc astérix`		2
	`disneyland`		2
	`disneyland paris`		2
	`walibi`		2
	`compagnies des alpes`		2

Comparaison des annotations (1)

- Comparaison basée sur l'identification d'une nouvelle session (NS) ou d'une continuation de la session précédente (CS) par les annotateurs
  - Et non sur les identifiants de session

Utilisateur	Requête	Épisode	Ann. 1		Ann. 3	
142	`Flandre Wallonie`	2	1	NS	1	NS
	`BHV`		1	CS	2	NS
	`conflit périphérie`		1	CS	1	NS
	`conflit périphérie Bruxelles`		1	CS	1	CS
	`citoyen définition`		**2**	**NS**	**3**	**NS**
	`définition citoyen`		2	CS	3	CS

- Pas de prise en compte des cas triviaux
  - Première requête d'un épisode de recherche
  - Un épisode de recherche ne contient qu'une seule requête et constitue donc une seule session

Comparaison des annotations (2)

		Ann. 2			Ann. 3		
		CS	NS	Total	CS	NS	Total
Ann. 1	CS	**490**	57	547	**514**	33	547
	NS	12	**39**	51	11	**40**	51
	Total	502	96	**598**	525	73	**598**

		Ann. 3		
		CS	NS	Total
Ann. 2	CS	**482**	20	502
	NS	43	**53**	96
	Total	525	73	**598**

Comparaison des annotations (3)

- Évaluation du degré de concordance entre chaque paire d'annotateurs

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

Paires d'annotateurs	κ
Annotateurs 1 et 2	$0,47$
Annotateurs 1 et 3	$0,61$
Annotateurs 2 et 3	$0,57$

- Sélection pour chaque requête des annotations faisant l'objet d'un accord entre au moins deux annotateurs
- 406 sessions identifiées pour 947 requêtes
- $2,33$ requêtes par session en moyenne

Exemple de désaccord entre les annotateurs

Utilisateur	Requête	Épisode	Session ann. 1	Session ann. 2
7	`bank`	1	1	1
	`bank crisis`		1	1
	`China party`		2	1
	`china financial`		2	1
	`bank`		1	1
72	`philosophie`	1	1	1
	`jean lasrière`		1	2
	`jean ladrière`		1	2
	`"jean ladrière"`		1	2

- Difficulté de choisir le lien pertinent entre les requêtes parmi plusieurs liens possibles
- Difficulté d'exploitation systématique d'une ressource externe pour les liens sémantiques faibles

Plan

Détection automatique : méthode temporelle

Principe [Silverstein *et al.*, 1999 ; He et Göker, 2000]

- Plus la durée entre deux requêtes consécutives est longue, moins il est probable que ces requêtes appartiennent à la même session
- Choix d'un seuil temporel approprié

- Simplicité de mise en œuvre
- Ne détecte pas les sessions très courtes résultant d'un changement soudain du besoin d'information
- Ne détecte pas les sessions très longues comportant des pauses importantes entre chaque requête

Détection automatique : méthode lexicale

Principe [He *et al.*, 2002 ; Ozmutlu et Çavdur, 2005]

- Plus les requêtes ont un contenu lexical en commun, plus il est probable que ces requêtes appartiennent à la même session
- Détection des reformulations entre requêtes successives

- Combinaison possible avec la méthode temporelle
- Nécessite la présence d'au moins un mot commun entre les requêtes
- Ne détecte pas les requêtes sémantiquement reliées mais ne comportant pas de lien lexical explicite

Détection automatique : méthodes basées sur des sources de connaissance externes

Principe

- Exploitation de sources de connaissances externes au journal de requêtes
- Évaluation de la similarité entre les requêtes à partir d'une représentation plus riche de leur contenu

- Divers types de sources de connaissances
  - autre journal de requêtes, documents retournés [Jones et Klinkner, 2008]
  - *Wikipédia* et *Wiktionary* [Lucchese *et al.*, 2011]
  - métadonnées des documents pertinents cliqués [Kramár et Bieliková, 2012]
- Approche présentant les meilleures performances

Baseline : méthode temporelle

Implémentation [Jansen *et al.*, 2007]

- Seuil temporel fixant la durée maximale entre deux requêtes successives appartenant à la même session
- Rupture : durée entre les requêtes supérieure au seuil
- Continuité : durée entre les requêtes inférieure au seuil

Utilisateur	Temps entre requêtes	Requête	Épisode	Session
54	0	alexandrie arabe	1	1
	12	alexandrie sources arbaes		2
	5	alexandrie sources arabes		2
	18	alexandrie cadiz sources arabes		3
	5	alexandrie cadiz sources		3

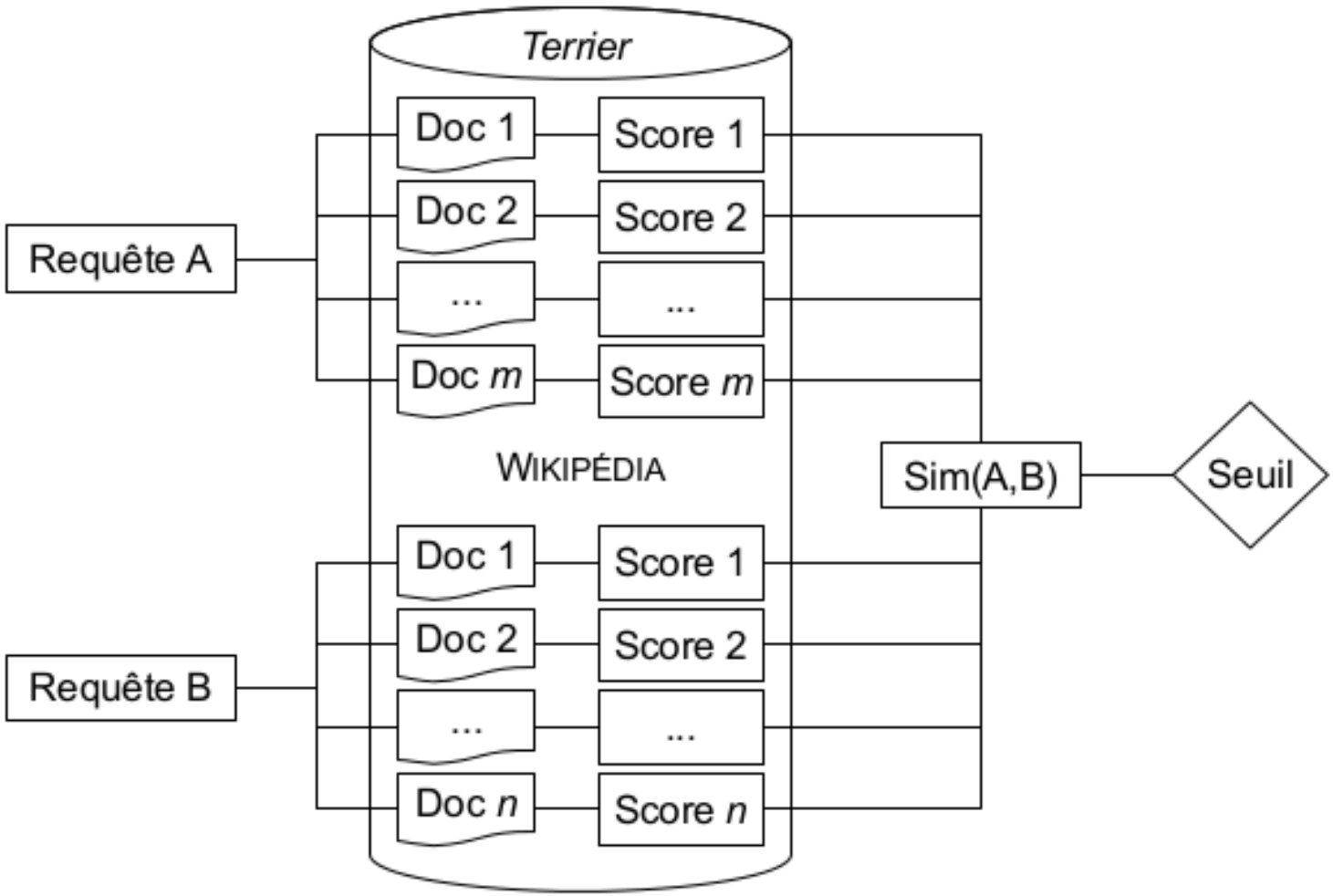
Baseline : méthode lexicale

Implémentation [Jansen *et al.*, 2007]

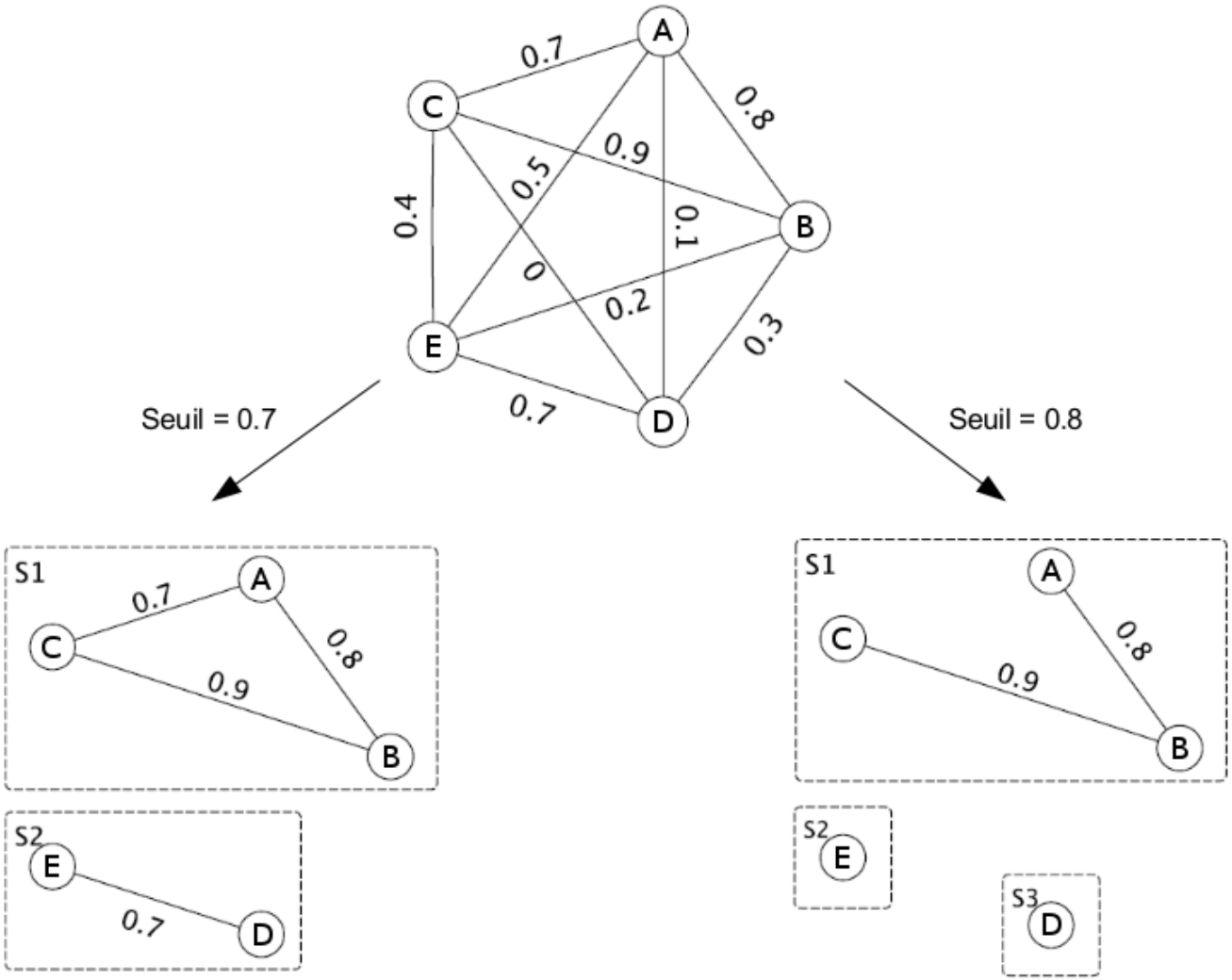
- Calcul du nombre de mots communs et de la différence de longueur entre deux requêtes successives afin de détecter des types de reformulation
- Rupture : aucun type de reformulation détecté
- Continuité : reformulation détectée entre les requêtes

Utilisateur	Temps entre requêtes	Requête	Épisode	Session
54	0	`alexandrie arabe`	1	1
	12	`alexandrie sources arbaes`		1
	5	`alexandrie sources arabes`		1
	18	`alexandrie cadiz sources arabes`		1
	5	`alexandrie cadiz sources`		1

Méthode développée : similarité à l'aide de *Wikipédia* (1)



Méthode développée : similarité à l'aide de *Wikipédia* (2)



Évaluation des sessions séquentielles : mesures d'évaluation

Principe [Gayo Avello, 2009]

- Mesures de précision P, rappel R et F-mesure F
- Comparaison entre le nombre de ruptures de sessions détectées par le système et le nombre de ruptures de sessions de la référence

$$P = \frac{N_{RuptureCorrecte}}{N_{RuptureSysteme}} \qquad R = \frac{N_{RuptureCorrecte}}{N_{RuptureReference}} \qquad F = \frac{2PR}{P+R}$$

- Annotation automatique de la collection de référence en ruptures et continuations de sessions
- Suppression des cas triviaux n'entraînant aucune décision de la part du système
  - Collection de référence réduite à 598 requêtes

Évaluation des sessions séquentielles : méthode temporelle

Précision	0,31
Rappel	0,31
F-mesure	0,31

		Rupt.	Cont.	
Réf.	Rupt.	22	48	70
	Cont.	49	479	528
		71	527	598

- Seuil temporel variant de 10 à 5 120 secondes
- Meilleure performance pour un seuil de 640 secondes
- Efficacité de 84 %

Util.	Requête	Réf.	Syst.
18	`loup`	1	1
	`Le monde agricole confronté au loup`	1	2
32	`Après la catastrophe`	1	1
	`Recherches sociologiques et anthropologiques`	2	1

Évaluation des sessions séquentielles : méthode lexicale

Précision	0, 24
Rappel	1
F-mesure	0, 38

		Rupt.	Cont.	
Réf.	Rupt.	70	0	70
	Cont.	225	303	528
		295	303	598

- Efficacité de 62 %
- Aucun faux négatif

Util.	Requête	Réf.	Syst.
35	`éclaircissants`	1	1
	`peau claire`	1	2
	`tshoko`	1	3
	`maquillage afrique`	1	4
	`dépigmentation`	1	5

Évaluation des sessions séquentielles : méthode *Wikipédia*

Précision	0, 31
Rappel	0, 8
F-mesure	0, 45

		Rupt.	Cont.	
Réf.	Rupt.	56	14	70
	Cont.	124	404	528
		180	418	598

- Seuil de similarité variant de $1 \cdot 10^{-5}$ à $5 \cdot 10^{-1}$
- Meilleure performance pour un seuil de $5 \cdot 10^{-3}$
- Efficacité de 77 %

Util.	Requête	Réf.	Syst.
71	`philippe dasseto`	1	1
	`dasseto`	1	2
	`Dasseto`	1	3
106	`divorce`	1	1
	`adoption`	2	1

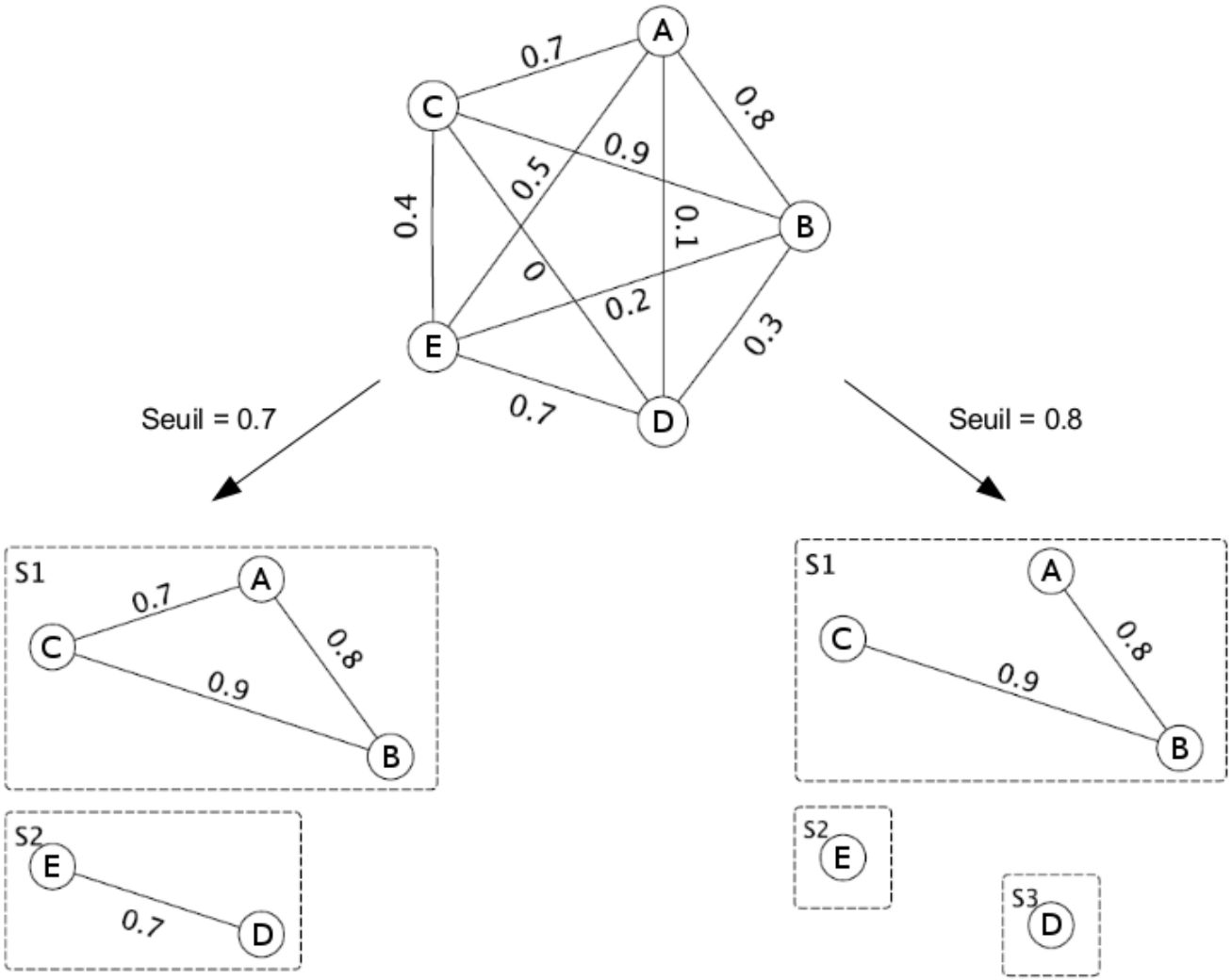
Évaluation des sessions imbriquées : mesures d'évaluation

Principe [Lucchese *et al.*, 2011]

- Index de Rand R [Rand, 1971]
- Index de Jaccard J [Jaccard, 1901]
- Vérification de la cohérence de répartition des paires de requêtes entre les sessions détectées par le système et les sessions de référence
- f_{ij} paire de requêtes apparaissant dans une session système identique ($i = 0$) ou différente ($i = 1$) et dans une session de référence identique ($j = 0$) ou différente ($j = 1$)

$$R = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \qquad J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Évaluation des sessions imbriquées : illustration



Évaluation des sessions imbriquées : méthode de similarité

Seuil	Rand	Jaccard
$1 \cdot 10^{-5}$	0,786	0,726
$5 \cdot 10^{-5}$	0,786	0,726
$1 \cdot 10^{-4}$	0,786	0,726
$5 \cdot 10^{-4}$	0,788	0,728
$1 \cdot 10^{-3}$	0,786	0,722
$5 \cdot 10^{-3}$	0,758	0,668
$1 \cdot 10^{-2}$	0,728	0,626

- Meilleur résultat pour un seuil de similarité de $5 \cdot 10^{-4}$
  - Liste des résultats obtenus pour chaque requête sensible aux types de reformulation
  - Seuil optimal pour la détection des sessions dépendant de notre jeu de données

Plan

Les données de départ

Données brutes

- Métadonnées des articles de Revues.org récupérées sur le serveur OAI-PMH (au 24/01/2013)
  - → auteur(s), date de publication, date de mise en ligne/modification, titre, langue, url
  - Articles correspondants au format HTML moissonnés sur les serveurs

Filtrage

- Identification de la langue avec *Textcat Language Guesser* → 85 % d'articles en français
- Identification des articles en texte intégral → 82 % des documents

Le corpus de travail

Corpus constitué de documents divers (des articles de revues, mais aussi des compte-rendus de lecture, des entretiens, des annonces d'événements scientifiques...) en texte intégral et en français :

Nombre de mots	253 millions
Nombre d'articles	62 204
Nombre de revues	279

TABLE : Caractéristiques du corpus mobilisé

Traitements

Les traitements effectués :

- Normalisation du HTML
- Repérage et analyse des références bibliographique avec Bilbo (`http://bilbo.hypotheses.org/`)
- Analyse syntaxique avec Talismane
- Analyse distributionnelle

Analyse distributionnelle du corpus

Principe de l'analyse distributionnelle

Mesure de la similarité sémantique entre les mots/termes du corpus en fonction des contextes syntaxiques qu'ils partagent

Exemple : les voisins de *linguistique*

voisin	contextes partagés
sémiotique	génératif, structural, comparé, textuel
philologie	roman, sémitique, comparé, historique
sémantique	génératif, structural, cognitif, formel
vs dans Wikipédia :	
grammaire	génératif, comparé, formel, historique
anthropologie	s'intéresser_suj, étudier_suj, discipline_de
géologie	professeur_de, utiliser_en, enseigner_obj

Analyse distributionnelle du corpus

Les voisins de Revues.org sont exploitables pour :

- analyser les requêtes (typage des reformulations)

Analyse distributionnelle du corpus

Les voisins de Revues.org sont exploitables pour :

- analyser les requêtes (typage des reformulations)

Croisement des voisins de Revues.org avec les substitutions de mots dans les reformulations de requêtes

`relations entre espace pouvoir et` **`société`**
`relations entre espace pouvoir et` **`identité`**

`performance` `système d'information`
`évaluer` `système d'information`

`musique et` **`apprentissage`**
`musique` `et` **`mémorisation`**
`rythmique` `et mémorisation`

→ environ 50 % des paires de mots impliquées dans des substitutions sont des voisins

Analyse distributionnelle du corpus

Les voisins de Revues.org sont exploitables pour :

- analyser les requêtes (typage des reformulations)
- mesurer certaines caractéristiques des documents

Plan

Caractérisation des documents

Calcul d'un ensemble de caractéristiques :

- Pour chaque document du corpus
- Pour chaque paire <requête,document renvoyé>

Caractérisation des documents

Caractéristiques des documents :

- Caractéristiques structurelles de base : longueur du texte, présence et quantité de certains éléments (résumé, tableaux, figures, références bibliographiques, notes), nombre et niveaux des titres de section, etc.
- Complexité syntaxique (sur la base de l'analyse de TALISMANE) : profondeur de l'arbre, distance moyenne des dépendances syntaxiques
  Exemple :
  - Forte complexité : http ://1895.revues.org/1632
  - Faible complexité : http ://cem.revues.org/11450

Revue du livre : Mary Ann Doane, The Emergence of Cinematic Time : Modernity, Contigency, the Archive (1895)

S'inscrivant, d'un point de vue méthodologique, dans la perspective d'une archéologie foucaldienne des savoirs (ici sur le temps et sa représentation), ce livre construit une partie de ce que l'on pourrait appeler une « épistémologie du cinéma » déjà balisée par les études de Friedrich Kittler (1986) ou Jonathan Crary (1990 et 1999) – citées à plusieurs reprises par Doane –, mais aussi par des ouvrages collectifs comme ceux édités par Leo Charney et Vanessa R. Schwartz (1995) ou plus récemment par François Albera, Marta Braun et André Gaudreault (2002). Tous ces travaux manifestent, à des degrés divers, une volonté de mettre à jour l'*épistémé* qui voit l'émergence du cinématographe, c'est-à-dire à déchiffrer la grille symbolique à travers laquelle cette période appréhende divers savoirs, discours et pratiques, et, plus précisément, à rendre compte du rôle joué par le cinéma à l'intérieur des mutations épistémiques engendrées par la modernité.

AVALLON - Église collégiale Saint-Lazare (Bulletin du centre d'études médiévales)

Le site

La tradition mentionne la fondation d'une collégiale au IXe siècle. L'église actuelle a été reconstruite au XIIe siècle. Au XIXe siècle fut découverte, sous le chevet et antérieur à celui-ci, une petite crypte. Transformée plus tard en espace technique pour le calorifère, elle attend aujourd'hui une restauration.

La fouille

Les recherches se sont concentrées sur l'analyse du bâti et des enduits de la crypte. Les relevés et les observations ont souligné des problèmes d'accès et de datation du site. Elles ont révélé au moins trois états successifs de construction antérieurs à l'édifice actuel (XIIe siècle).

Caractérisation des documents

Caractéristiques des paires <requête,document>

- fréquence et couverture des mots de la requête dans le document
- présence des mots de la requête dans des éléments spécifiques : titre du document, auteur, résumé, titres de section...
- caractéristiques des chaines lexicales construites à partir des mots de la requête en utilisant les Voisins de Revues.org : densité, couverture

Exemple : `assr_23491.html`

- caractéristiques des snippets générés pour chaque paire <requête,document> : structure (nombre de « morceaux »), fréquence des mots de la requête, cohésion lexicale...

Exemple :
`http://search.openedition.org/?q=caricature+religion`

Jean-Claudes Gardes, Guillaume Doizy (textes réunis par), Ridiculosa 15, Caricature et Religion(s)

Brest, Université de Bretagne occidentale, décembre 2008, 574 p.
>> http://assr.revues.org/23491

... à caricaturer la religion, les mots « détournement », « déplacement », « métonymie », « décalage ... de religions, les caricatures du pape et de la papauté y ajoutaient un piment de vulgarité scatologique devenue ...) de l'université de Bretagne occidentale, rend compte d'une série de manifestations intitulées « Caricature ...

Publication
Archives de sciences sociales des religions
Type de publication : Revues • Type de document : Compte-rendu

Auteur
Jean-Louis Schlegel

Date de publication
décembre 2011

Disponibilité du document
Texte intégral disponible en accès restreint sur Cairn

Grands Hommes vus d'en bas

L'iconographie officielle et ses usages populaires

>> http://gradhiva.revues.org/1632

... » (Duprat 2002 : 10). Ces caricatures inversent l'esthétique officielle de la grandeur en peignant ... à la caricature politique contemporaine, comme en témoigne l'émission télévisée du Bébête Show (Collovald 1992 ... par la caricature, qui pousse à l'excès cette fétichisation grotesque pour en dévoiler la vacuité (Mbembe 1996 ... de la panthéonisation, cf. Ben-Amos 1990. 5 Sur la caricature du Bébête Show et des Guignols de l'info, cf. également ... ou demi-dieu ? La mise en place d'une religion napoléonienne », Romantisme 28(100) : 131-141. Bourdieu ... -Pascal 1996 « Les ambivalences dans la caricature des dirigeants politiques. Illustrations africaines ... de papier : la caricature de Henri III à Louis XVI. Paris, Belin. Durkheim, Émile 1975 (1883) « Le rôle ...

Publication

Gradhiva

Type de publication : Revues • Type de document : Editorial

Auteurs

Julien Bonhomme, Nicolas Jaoul

Date de publication

mai 2010

Disponibilité du document

Texte intégral disponible en accès restreint sur Cairn

Plan

Quels éléments des logs peut-on croiser avec les caractéristiques des documents ?

Liens entre les caractéristiques du document et :

- sa « popularité » globale : le nombre de clics sur celui-ci dans l'ensemble des logs
- pour une requête donnée :
  - le fait qu'il ait été consulté ou non (document cliqué *vs* collection de documents renvoyés par le moteur de recherche)
  - le temps de consultation

Vers un point de vue plus global... :

- Typologie des requêtes selon leurs interactions avec les documents renvoyés par le moteur de recherche
- Typologie des documents selon l'ensemble des requêtes par lesquelles ils sont abordés

Popularité et caractéristiques des documents

Méthode

- Pour chaque document, recherche de l'ensemble des accès dans les logs disponibles (en évitant les doublons d'utilisateur+requête)
- Croisement avec l'ensemble des caractéristiques de documents et recherche de corrélations

Résultats

Les documents plus souvent cliqués :

- sont plus longs (effet du moteur de recherche)
- sont plus structurés (plus de titres et plus de niveaux de titres)
- ont plus tendance à contenir des références bibliographiques

Temps passé et caractéristiques des documents

Méthode

- Extraction de toutes les séquences requête - accès à un document - requête (pour un même utilisateur)
  - → La différence entre l'heure de la seconde requête et l'heure d'accès au document donne la durée de consultation
- Croisement avec l'ensemble des caractéristiques (du document, de la paire document-requête) et recherche de corrélations

Temps passé et caractéristiques des documents

Résultats

- Le temps passé sur un document est légèrement corrélé avec sa longueur.
- Mais pas avec :
  - la fréquence / couverture des mots de la requête dans le document
  - la cohésion lexicale avec les mots de la requête dans le document
  - la complexité syntaxique du document

Choix d'un document suite à une requête

Méthode

- Pour chaque requête ayant donné lieu à au moins un clic, prise en compte de tous les documents affichés par le moteur de recherche sur la même page que le(s) document(s) cliqué(s)
- Recherche de corrélations entre les caractéristiques des documents et leur choix au sein de la collection de documents renvoyés

Résultats

- Résultats triviaux : corrélation avec la fréquence des termes de la requête dans le snippet/document, la couverture...
- Pas d'apport des indices plus fins : par exemple de la cohésion lexicale de la requête avec le document ou avec le snippet

Vers une typologie du lien requête-document

→ Approche plus qualitative

- Pour un document, extraction de l'ensemble des requêtes qui y ont abouti dans les parcours de recherche
- Visualisation des requêtes dans le document
- Typologie des requêtes en fonction de leurs liens avec les documents qu'elles renvoient
- Typologie des documents en fonction des points de vue à travers lesquels ils sont abordés

Une typologie des requêtes en lien avec les documents qu'elles visent

Exemples :

- La présentation de soi. Ethos et identité verbale
- Colloque « Le muet a la parole »

Partie du document ciblée

- élément unique / zone restreinte du document
- ensemble du document
- élément particulier (saillant) : titre, auteur

Exemples

Requêtes titre ou auteur

```
Jean-Paul Fourmentraux
monique selim
gunthert
alain rabatel
yves defrance
a) Pierre DELEAGE, 2010. « Mythe et chant rituel chez
les Sharanahua »
Les jeunes sur Internet. Se construire un autre
chez-soi
```

Exemples

Requêtes ciblant l'ensemble des documents

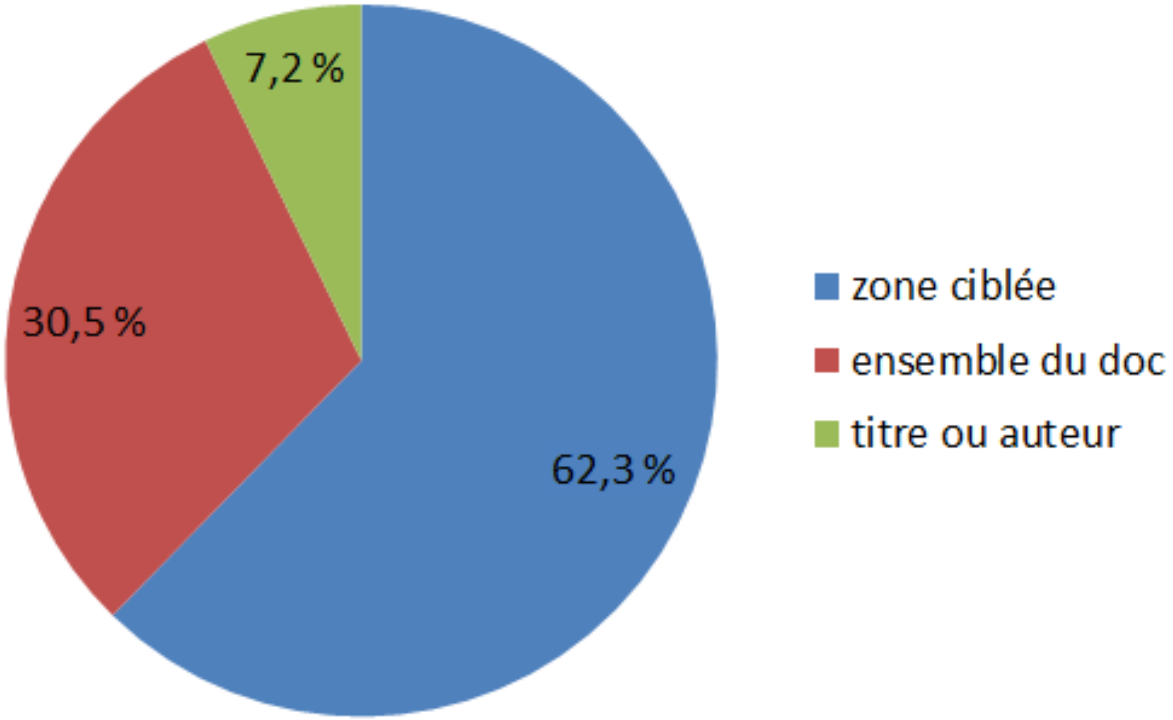
articles influence sociale processus de changement et enjeux de pouvoir au travail ; formation professionnelle ; tourisme solidaire ; projet urbain ; communication interne en entreprise ; les femmes immigrantes ; travail et emploi des femmes ; formation continue ; politique culturelle ; engagement parents apprentissage enfants ;

Requête ciblant une zone restreinte des documents

diphonie ; OGM ; intertextualité polyphonie ; amish ; la démocratie participative ; ostéopathie ; nudisme ; téléspectateur ; sans-abris ; médecine tibétaine ; Heteronormativite ; kerbrat ; l'amour ; GERARD MENDEL ;

Typologie des requêtes

Représentation des différents types de requêtes



Une typologie des documents selon la façon dont ils sont abordés par les requêtes

- Pour chaque document, analyse de l'ensemble des requêtes ayant conduit un utilisateur à y accéder

→ Différents profils de documents selon la façon dont ils sont abordés par les requêtes :

  - Proportions de requêtes titre/auteur, portant sur l'ensemble du document ou ciblées
    Exemples :
    - Un document abordé presque uniquement par l'anecdote : Une cuiller à pot pour demander la pluie
    - Un document toujours entièrement balayé par les requêtes : Les risques du journalisme dans les conflits armés
  - Homogénéité des ensembles de requêtes liées aux documents

Homogénéité des ensembles de requêtes : exemples

Article : « Le vécu de l'accident nucléaire de Fukushima, Japon : les paroles des enfants » (Akiko IDA)
Requêtes : `FUKUSHIMA ; accident nucléaire 2011 ; articles de presse accident nucleaire japon ; japon nucleaire ; japon ; Accident nucléaire Japon ; accident nucléaire fukushima 2011 ; Fukushima ; accident nucléaire au Japon ; fukushima daiichi ; fukushima ; centrale nucleaire de fukushima`

Article : Stylistique des fantômes (Philippe-Alain Michaud)
Requêtes : `gladiator ; stylistique ; feu d'artifice ; Philippe-Alain Michaud ; platon`

$\rightarrow$ Calculer un score d'homogénéité (comment est mesurée la similarité entre deux requêtes : distance d'édition ? similarité sémantique ?)

Conclusion et perspectives

Des données riches et pléthoriques

→ beaucoup d'énergie consacrée à l'acquisition, au nettoyage, au tri, au recoupement des données...

→ de nombreux types d'informations à croiser :

- sur les requêtes : caractéristiques des requêtes, inscription dans une session de recherche, types de reformulations...
- caractéristiques des documents, informations tirées du corpus, interactions document-requête...
- comportement de l'utilisateur à travers différents indicateurs : clics, informations temporelles...

Conclusion et perspectives

Création/extraction de nouvelles données

- une base distributionnelle : les Voisins de Revues.org
- une collection de référence comprenant des épisodes de recherche annotés en sessions de recherche
- extraction de données : par exemple, listes de paires de requêtes présentant des liens sémantiques

Conclusion et perspectives

Quelques résultats

- Méthodes de détection de sessions de recherche
- Analyse des mouvements sémantiques dans les reformulations de requête
- Quelques corrélations entre accès aux documents et caractéristiques

MAIS des indicateurs du comportement / de la satisfaction de l'utilisateur assez frustes, un corpus très hétérogène...

→ difficultés à faire émerger des résultats qui valorisent l'approche linguistique / TAL

Conclusion et perspectives

Perspectives

- Avancer dans la typologie (des requêtes, des reformulations, des documents...) :
  - Faire une typologie *a posteriori* des requêtes, exploitant des informations contextuelles issues des sessions (position de la requête dans la session, consultation d'un document, etc.)
- Vers des études :
  - plus qualitatives ?
  - plus circonscrites / contrôlées
    - se limiter à un type de requête ? $\rightarrow$ étude des requêtes auteur