Les sessions de recherche comme contexte des requêtes

Simon Leva*

*CLLE-ERSS : CNRS et Université de Toulouse (UMR 5263),
5 allées Antonio Machado, 31058 Toulouse Cedex 9
sleva@univ-tlse2.fr
http://w3.erss.univ-tlse2.fr/textes/pagespersos/leva/

Résumé. La tâche d'identification des sessions des utilisateurs d'un moteur de recherche a suscité la construction de plusieurs collections de référence et l'élaboration de multiples méthodes de détection automatique. Cette tâche constitue en effet le point de départ de nombreuses études s'intéressant au contexte de la recherche et aux besoins d'information des utilisateurs. Nous détaillons dans cette étude la construction d'une collection de référence à partir d'un journal de requêtes issu du portail *OpenEdition*, et nous présentons une évaluation des annotations manuelles constituant cette collection. La référence obtenue contient 947 requêtes pour 406 sessions, avec un taux d'accord (Kappa de Cohen) entre les annotateurs allant de $0,47$ à $0,61$. Cette collection servira à l'évaluation de méthodes de détection automatique des sessions ainsi qu'à des études portant sur les reformulations de requêtes.

1 Introduction

Les utilisateurs d'un moteur de recherche sur le Web font appel à différentes stratégies afin de satisfaire leurs besoins d'information. En particulier, les requêtes soumises peuvent faire l'objet de plusieurs reformulations visant à préciser le besoin d'information initial. Les requêtes d'un utilisateur font ainsi souvent partie d'une session de recherche, et ne devraient pas être considérées de manière isolée. En effet, une session fournit de nombreux indices sur le contexte de la recherche, l'objectif de l'utilisateur ou son expertise dans le domaine considéré. Les reformulations de requêtes et les documents consultés sont des éléments particulièrement utiles pour une meilleure compréhension du besoin d'information, de son évolution et de sa satisfaction au fil d'une recherche. La notion de session est donc une notion clef en recherche d'information, et plus spécifiquement en recherche d'information contextuelle.

L'étude présentée dans cet article porte sur la construction d'une collection de référence manuellement annotée au niveau des sessions, ainsi que sur l'évaluation de la concordance des annotations. La collection constituée provient d'un journal de requêtes issu du portail de ressources électroniques dédiée aux sciences sociales *OpenEdition* [1]. La constitution d'une collection de référence dans cet environnement vise deux objectifs principaux : 1) servir de référence pour l'évaluation de méthodes de détection automatique des sessions ; 2) servir de

1. http ://www.openedition.org/

référence pour des études sur les types de reformulations effectuées par les utilisateurs de la plateforme, et à terme pour réaliser une typologie des requêtes.

Ce travail s'inscrit dans le cadre du projet ANR CAAS *(Contextual Analysis and Adaptive Search)* — programme Contint — coordonné par Josiane Mothe (IRIT), et faisant l'objet d'un partenariat entre l'Institut de Recherche en Informatique de Toulouse (IRIT), le Laboratoire Informatique d'Avignon (LIA) et l'Équipe de Recherche en Syntaxe et Sémantique du laboratoire Cognition, Langue, Langage, Ergonomie (CLLE-ERSS).

Dans cet article, nous nous intéressons tout d'abord à la notion de session telle que définie en recherche d'information, ainsi qu'aux collections de référence précédemment constituées (section 2). Puis, nous présentons la méthodologie de construction de la collection de référence issue du portail *OpenEdition* (section 3). Enfin, nous détaillons la collection construite et l'évaluation des annotations (section 4), avant d'envisager les perspectives de ce travail.

2 La notion de session en recherche d'information

2.1 Définitions de la notion de session

L'une des premières définitions de la notion de session dans le cadre de l'étude d'un journal de requêtes a été proposée par Silverstein et al. (1999) :

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

Cette idée de regroupement des activités d'un utilisateur (soumission d'une requête, navigation sur la page de résultats, consultation d'un résultat) correspondant à un thème spécifique se retrouve chez Göker et He (2000), qui insistent également sur la proximité temporelle entre les activités au sein d'une session :

> *This paper focuses on the temporal ordering of activities clustered according to close proximity in time. [...] We group activities and refer to the resulting unit as a session. If we view a user with an interest in a specific topic as acting in a particular role, then [...] the activities in the same session are likely to correspond to one role.*

Spink et al. (2006) proposent une définition moins restrictive de la notion de session, qui est considérée comme une simple séquence de requêtes d'un utilisateur :

> *A session is the entire series of queries submitted by a user during one interaction with the Web search engine. Session length varied from less than a minute to a few hours.*

Cette définition est ensuite affinée par Jansen et al. (2007) et correspond alors à un épisode de recherche, notion distinguée de celle de session :

> *One can define a user episode on a Web search engine as a temporal series of interactions among a searcher, a Web system, and the content provided by that system within a specific period. [...] However, it is possible that one searching episode will be composed of one or more sessions. We define a session from a contextual viewpoint as a series of interactions by the user toward addressing a single information need.*

Tandis que les définitions précédentes envisagent les requêtes d'une session sous la forme de séquences, impliquant une certaine contiguïté entre les requêtes, Jones et Klinkner (2008) développent une vision hiérarchique de la notion de session. Ainsi, une session se compose d'une ou de plusieurs missions de recherche, qui se composent à leur tour d'un ou de plusieurs buts pouvant donner lieu à une ou plusieurs requêtes. De plus, les requêtes liées à un même but peuvent être imbriquées avec des requêtes visant un autre but :

> *A search session is all user activity within a fixed time window. [...] A search goal is an atomic information need, resulting in one or more queries. [...] The queries need not be contiguous, but may be interleaved with queries from other goals. [...] A search mission is a related set of information needs, resulting in one or more goals.*

Gayo Avello (2009) reprend la distinction entre épisode et session de recherche opérée par Jansen et al. (2007) tout en adoptant à nouveau une vision séquentielle de la notion de session :

> *[A searching episode] refers to the actions performed by a particular user within a search engine during, at most, one day. Such a searching episode can comprise one or more sessions where each of these includes one or more successive queries related to one single information need or goal.*

L'idée d'imbrication au sein d'une session entre des requêtes correspondant à des buts distincts est finalement reprise par Lucchese et al. (2011) :

> *Task-based sessions [are] sets of possibly non contiguous queries issued by the user of a Web Search Engine for carrying out a given task.*

Les définitions de la notion de session proposées dans la littérature se distinguent à plusieurs niveaux. Une session peut ainsi contenir des éléments de différente nature : des activités/interactions (Göker et He, 2000; Jansen et al., 2007) ou des requêtes (Silverstein et al., 1999; Spink et al., 2006; Jones et Klinkner, 2008; Gayo Avello, 2009; Lucchese et al., 2011). Cette différence est directement liée aux informations disponibles dans les journaux de requêtes, selon qu'ils fournissent uniquement les requêtes soumises par les utilisateurs, ou incluent également leurs actions. D'autre part, les sessions se distinguent par leur durée, plutôt courte (Silverstein et al., 1999; Göker et He, 2000) ou pouvant aller jusqu'à quelques heures (Spink et al., 2006), mais dans tous les cas inférieure à une journée (Gayo Avello, 2009). Cette contrainte temporelle s'explique notamment par le renouvellement de l'adresse IP des utilisateurs toutes les 24 heures, rendant impossible leur différenciation au-delà de cette période. Malgré ces points de divergence, la majorité des définitions s'accordent sur le fait qu'une session permet de regrouper des actions ou des requêtes liées à un même besoin d'information. Ces actions ou ces requêtes peuvent alors être envisagées de manière séquentielle (Silverstein et al., 1999; Göker et He, 2000; Jansen et al., 2007; Gayo Avello, 2009) ou imbriquée (Jones et Klinkner, 2008; Lucchese et al., 2011). Ce dernier cas est fréquemment rencontré lors d'une recherche multitâche *(multitasking search)* (Spink et al., 2006), l'utilisateur effectuant alors une recherche sur plusieurs thèmes simultanément.

2.2 Collections de référence existantes

Plusieurs initiatives de constitution de collections de référence ont vu le jour dans le cadre de la détection automatique des sessions. Les collections ainsi constituées ont contribué au

développement de méthodes de détection automatique, notamment à travers des techniques d'apprentissage, mais ont également servi à l'évaluation de ces méthodes.

Göker et He (2000) ont mené une campagne d'annotation sur un journal de requêtes issu du réseau Intranet de l'agence de presse *Reuters* pour l'année 1999. Deux annotateurs experts en formulation de requête ont été chargés d'identifier les sessions de 1 440 utilisateurs pour un ensemble de 9 534 requêtes. La collection ainsi constituée a servi de référence pour l'évaluation d'une méthode de détection automatique basée sur la définition d'un seuil temporel entre les activités faisant partie d'une même session.

Jansen et al. (2007) ont procédé à l'annotation d'un échantillon de 2 000 requêtes soumises au méta-moteur de recherche *Dogpile* durant l'année 2005. Ces annotations ont été comparées à trois méthodes de détection automatique des sessions exploitant les informations d'adresse IP et de *cookies* soit seules, soit en les associant respectivement à un seuil temporel et à des patrons de reformulation de requêtes. L'exploitation de ces annotations a notamment permis d'identifier l'origine des erreurs générées par les méthodes automatiques évaluées.

Jones et Klinkner (2008) ont réalisé l'annotation d'un échantillon de 3 jours de requêtes soumises au moteur de recherche *Yahoo !* au cours de l'année 2007. La collection de référence constituée se compose de 2 922 sessions pour un ensemble de 8 226 requêtes soumises par 312 utilisateurs. La consigne donnée pour l'annotation était que les requêtes appartenant à une même session possèdent les mêmes critères de réussite en termes de satisfaction du besoin d'information de l'utilisateur. Cette collection leur a permis de tester l'opérabilité de leur définition de session, mission et but de recherche. L'annotation réalisée est de type ascendant, le groupe d'annotateurs ayant exploité les pages de résultats et les clics des utilisateurs afin de déterminer leurs buts, puis leurs missions et finalement leurs sessions. De plus, le cas des requêtes liées à un même but et imbriquées au sein d'une même session est envisagé.

Gayo Avello (2009) a mis en place une campagne d'annotation portant sur des échantillons de sept journaux de requêtes provenant de moteurs de recherche commerciaux (*AlltheWeb*, *Altavista*, *AOL* et *Excite*). La collection ainsi constituée comporte près de 95 000 requêtes soumises par près de 15 000 utilisateurs, pour un ensemble de près de 35 000 sessions manuellement identifiées. Il s'agit de l'initiative la plus ambitieuse de constitution d'une collection de référence. Un annotateur expert a été chargé d'évaluer si deux requêtes successives sont ou non thématiquement reliées. L'annotateur s'est basé uniquement sur le texte des requêtes, mais pouvait éventuellement faire appel à une ressource externe en cas de manque de connaissances. Cette collection de référence a servi de cadre commun pour l'évaluation et la comparaison de diverses méthodes de détection automatique.

Nous constatons que les différentes annotations manuelles des sessions recensées dans la littérature contiennent relativement peu de détails concernant le protocole d'annotation en lui-même. En particulier, aucune précision n'est indiquée quant aux difficultés rencontrées par les annotateurs. Il n'existe ainsi à notre connaissance aucune étude proposant d'évaluer l'accord inter-annotateur pour la tâche de détection des sessions. Par ailleurs, les études mentionnées n'indiquent pas de référence permettant d'accéder aux collections constituées.

3 Méthodologie de construction d'une référence annotée

3.1 Définition d'une session

Nous retenons dans le cadre de ce travail les définitions suivantes :
- un *épisode de recherche* correspond à l'ensemble des requêtes soumises à un moteur de recherche par un utilisateur donné durant au plus une journée ; cet épisode de recherche peut contenir une ou plusieurs sessions de recherche ;
- une session de recherche correspond à l'ensemble des requêtes reliées à un même besoin d'information ; ces requêtes peuvent être imbriquées au sein d'un même épisode de recherche dans le cas d'un épisode multitâche.

3.2 Annotation manuelle des sessions

La tâche d'annotation des sessions consiste à indiquer, pour un utilisateur et un épisode de recherche donné, à quelle session appartient chacune des requêtes incluses dans cet épisode. Autrement dit, il s'agit de segmenter les épisodes de recherche de chaque utilisateur en une ou plusieurs sessions. Nous avons élaboré un guide d'annotation pour familiariser les annotateurs avec les notions mobilisées tout en détaillant les aspects pratiques de la tâche.

Une session = un ensemble de requêtes liées Pour chaque épisode de recherche d'un utilisateur, les annotateurs doivent observer l'ensemble des requêtes soumises avant de prendre une décision et d'attribuer une session à chaque requête. Ce point est crucial, car le lien entre deux requêtes peut ne pas être évident de prime abord, mais apparaître dans une requête ultérieure. Par exemple, le lien n'est pas évident entre la requête `barriere de corail` suivie de la requête `nouvelle zélande`. Pourtant, ce lien s'éclaire avec la soumission d'une troisième requête `barriere de corail nouvelle zélande`. Il ne s'agit donc pas de considérer uniquement les couples de requêtes successives.

Des requêtes liées à un même besoin d'information Si les requêtes regroupées au sein d'une session visent un même besoin d'information, celui-ci n'apparaît pas toujours de manière explicite. Afin d'aiguiller les annotateurs, nous leur avons proposé une liste d'indices potentiels, pouvant apparaître de manière combinée. Ces indices relèvent de différents niveaux d'information :

1. *Indices textuels :* les requêtes possèdent des mots en commun, ou des parties de mot en commun. C'est par exemple le cas des requêtes `femmes moralistes`, `femmes moralistes 18 siecle` et `moralistes`.
2. *Indices sémantiques :* les requêtes possèdent des mots qui ne sont pas strictement identiques au niveau orthographique, mais qui sont néanmoins liés par une relation sémantique telle que la synonymie, l'hyperonymie ou l'hyponymie. Par exemple, les requêtes `foresterie` et `sylviculture` sont reliées par le fait que le terme *foresterie* est un synonyme en français québécois pour le terme *sylviculture*.
3. *Indices de proximité thématique :* les requêtes possèdent des mots qui ne sont ni identiques, ni liés par une relation sémantique classique, mais qui restent néanmoins lexicalement proches dans le cadre de thématiques ou d'objets spécifiques. Par exemple,

les requêtes `théâtre expérimental` et `grotowski` sont reliées par le fait que le polonais Jerzy Grotowski était un metteur en scène et théoricien du théâtre.

Utilisation de ressources externes En l'absence d'indices textuels, il peut parfois être nécessaire de faire appel à une ressource externe (dictionnaire, thésaurus, encyclopédie, etc.) afin de palier un manque de connaissances. Les annotateurs doivent alors indiquer la ressource utilisée et expliciter l'élément ayant permis de lever l'indécision. C'est par exemple le cas pour les requêtes `ifriqya` et `tunisie`, une réponse possible étant d'indiquer que l'encyclopédie *Wikipédia* mentionne *Le territoire de l'Ifriqiya correspond aujourd'hui à la Tunisie*.

Une session possède un identifiant numérique unique au sein d'un épisode Afin d'identifier chacune des sessions d'un utilisateur, il est demandé aux annotateurs d'attribuer à chaque requête d'un épisode le numéro de la session à laquelle elle appartient. La première session de chaque épisode est identifiée par 1, la seconde par 2, et ainsi de suite. Ce type de numérotation permet également d'identifier les requêtes imbriquées reliées à une même session.

4 Collection de référence

La collection de référence que nous avons constituée est issue de données provenant du portail *OpenEdition*. Nous présentons tout d'abord cet environnement de recherche avant de détailler le journal de requêtes exploité et les résultats de l'annotation de la collection.

4.1 Le portail *OpenEdition*

Le portail *OpenEdition* propose un libre accès à un ensemble de ressources électroniques dans le domaine des sciences humaines et sociales. Développé et dirigé par le Centre pour l'édition électronique ouverte (Cléo), il se compose de trois plateformes dont chacune est dédiée à une ressource électronique spécifique : *Revues.org* diffuse 353 revues et 22 collections de livres, *Calenda* recense plus de 20 000 évènements scientifiques en lettres et en sciences humaines et sociales, tandis qu'*Hypotheses.org* héberge 269 blogs et carnets de recherche.

Plusieurs points d'entrée permettent d'effectuer une recherche dans cet environnement varié. D'une part, un moteur de recherche principal est accessible sur la page d'accueil du portail *OpenEdition* et de la plateforme *Revues.org*. D'autre part, une recherche peut également se faire directement à partir du moteur de recherche situé sur le site d'une revue associée à *Revues.org*. Dans ces deux situations, les résultats sont présentés dans une interface commune, permettant de préciser le type d'information recherchée à l'aide de champs (titre, auteur, résumé, etc.) et de restreindre la recherche à l'aide de filtres (plateforme de publication et type de document visé, année de la publication, etc.).

4.2 Journal de requêtes exploité

Nos travaux s'appuient sur un journal de requêtes provenant du portail *OpenEdition*. Ce journal de requêtes contient une collection de 1 057 471 requêtes soumises par 227 302 utilisateurs durant la période du 07 avril 2010 au 1er février 2012. Les requêtes sont majoritairement formulées en français, mais certaines sont également en anglais ou en espagnol. À la différence

des requêtes soumises à un moteur de recherche généraliste, la particularité de l'environnement d'*OpenEdition* fait que les requêtes émanent principalement d'acteurs du monde académique, et ciblent des revues, des évènements ou des blogs.

La construction de ce journal de requêtes a nécessité la mise en œuvre de plusieurs traitements afin de ne conserver que les informations les plus fiables à partir du journal d'accès original. En particulier, les données ont été nettoyées et filtrées de manière à éliminer les informations inexploitables (requêtes soumises par des robots d'indexation, longues séquences de requêtes strictement identiques, suites de signes de ponctuation, etc.) et à contenir des requêtes provenant d'utilisateurs individuels. Les requêtes ont ensuite été regroupées par adresse IP et classées par ordre chronologique. Le journal de requêtes finalement obtenu comporte un identifiant pour chaque utilisateur — correspondant à l'adresse IP anonymisée —, la date et l'heure de soumission de chaque requête, ainsi que les requêtes soumises.

La tâche d'annotation des sessions a été menée sur un échantillon du journal de requêtes *OpenEdition* suffisamment large pour être représentatif des phénomènes en présence tout en restant aisé à traiter pour les annotateurs. Nous avons ainsi sélectionné aléatoirement une collection de 947 requêtes soumises par 216 utilisateurs. Cet échantillon a été automatiquement segmenté en 349 épisodes de recherche, correspondant pour chaque utilisateur à l'ensemble de ses requêtes soumises en une journée au plus [2]. Trois annotateurs — dont l'auteur de cette étude —, non spécialistes dans les domaines représentés par les documents, ont été chargés de segmenter les requêtes de chaque épisode de recherche en une ou plusieurs sessions.

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
26	génération	1	1
	les jeunes		1
	ruralité		2

TAB. 1 – *Épisodes annotés en sessions extraits de la collection OpenEdition.*

Le tableau 1 présente un extrait de la collection *OpenEdition* après annotation. Pour chaque utilisateur est précisé l'ensemble de ses requêtes ainsi que les épisodes correspondants. Nous indiquons ici l'annotation de chaque épisode en sessions réalisée par l'annotateur 1.

2. Dans le cas d'une séquence de requêtes soumises avant et après minuit, chevauchant donc deux journées, la mise en place d'un seuil temporel évite la séparaison en deux épisodes distincts.

Le cas de l'utilisateur 39 correspond à un épisode multitâche, illustrant l'imbrication entre des requêtes relevant de thématiques de recherche différentes. Nous pouvons effectivement identifier d'une part un besoin d'information lié au thème du travail saisonnier et d'autre part un besoin d'information lié au thème des parcs de loisirs. Ces besoins d'information ont donné lieu à deux sessions distinctes, respectivement marquées par les identifiants 1 et 2, et ce malgré une discontinuité entre les requêtes.

Le cas de l'utilisateur 26 constitue un exemple de difficulté d'annotation. Si le lien est clair entre les deux premières requêtes, la troisième pose problème : s'agit-il d'une précision de la thématique précédente, auquel cas la requête porterait sur la ruralité selon les générations, ou s'agit-il d'une nouvelle thématique ? Étant donné qu'il n'y a ici aucune indication explicite de lien entre les requêtes, par exemple sous la forme d'une quatrième requête `ruralité chez les jeunes`, l'annotateur 1 a considéré qu'il s'agit de deux thématiques et donc de deux sessions distinctes.

4.3 Résultats de l'annotation

	Annotateur 1	Annotateur 2	Annotateur 3
Nombre de sessions	393	428	410

TAB. 2 – *Nombre de sessions identifiées dans la collection par chaque annotateur.*

Le tableau 2 présente le nombre de sessions identifiées dans la collection *OpenEdition* par chaque annotateur. Si ce nombre varie, il reste cependant proche de 410 sessions identifiées en moyenne. Afin de comprendre les différences au niveau du nombre de sessions identifiées et d'estimer le taux de concordance entre les annotateurs, nous confrontons chaque paire d'annotations obtenues dans une matrice de confusion. Ce type de représentation permet de visualiser, pour un nombre de catégories identique entre deux annotateurs, le nombre d'annotations communes et d'annotations différentes entre ces annotateurs.

		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**

TAB. 3 – *Matrices de confusion des annotations effectuées par l'annotateur 1 et les annotateurs 2 et 3.*

Les tableaux 3 et 4 représentent les matrices de confusion des annotations obtenues pour chaque paire d'annotateurs. Nous n'exploitons pas directement les identifiants de sessions attribués à chaque requête : dans les cas de détection de plus de deux sessions et d'épisode multitâche, une nouvelle session peut en effet recevoir un identifiant différent entre les annotateurs

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

TAB. 4 – *Matrice de confusion des annotations effectuées par les annotateurs 2 et 3.*

même si ces derniers s'accordent sur le fait qu'une requête marque le début d'une nouvelle session et non la continuation de la session précédente. Une telle situation est représentée dans le tableau 5, où la requête `citoyen définition` fait partie de la session 2 pour l'annotateur 1 et de la session 3 pour l'annotateur 3, tout en constituant le début d'une nouvelle session pour les deux annotateurs. Afin d'éviter ce biais, nous nous basons donc sur l'identification par les annotateurs d'une nouvelle session (noté NS dans les matrices de confusion et le tableau 5) ou d'une continuation de la session précédente (noté CS dans les matrices de confusion et le tableau 5) au sein d'un même épisode. De plus, nous ne prenons pas en compte lors de la constitution des matrices de confusion les cas triviaux — un épisode de recherche ne contient qu'une seule requête et constitue donc une seule session — pour lesquels aucune alternative ne s'offre aux annotateurs, aboutissant à une annotation identique. Cela explique le fait que seules 598 requêtes sont comptabilisées au total au lieu des 947 initialement annotées.

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

TAB. 5 – *Exemple de désaccord au niveau des identifiants de session et d'accord au niveau de la détection d'une nouvelle session entre les annotateurs 1 et 3.*

Par exemple, nous voyons dans le tableau 3 qu'il y a 39 requêtes pour lesquelles l'annotateur 1 et l'annotateur 2 ont considéré qu'elles marquent le début d'une nouvelle session au sein d'un même épisode, constituant donc des annotations identiques, tandis qu'il y a 12 requêtes pour lesquelles l'annotateur 1 a considéré qu'elles marquent le début d'une nouvelle session tandis que l'annotateur 2 les a annotées comme une continuation de la session précédente.

Nous utilisons le coefficient Kappa (Cohen, 1960) afin d'évaluer le degré de concordance entre chaque paire d'annotateurs au niveau de l'identification d'une requête comme débutant une nouvelle session ou poursuivant la session précédente. Ce coefficient se base sur une différence relative entre l'accord réel observé et un accord aléatoire, rendant ainsi possible une

comparaison. Il se définit par le rapport :

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

où P_a est la proportion d'accord observée entre deux annotateurs, et P_e la proportion d'accord aléatoire. Le coefficient Kappa correspond donc à une estimation de l'accord optimal entre les annotateurs après avoir retranché la part de cet accord dû au hasard — un accord parfait correspondant à un κ égal à 1. Le tableau 6 présente le coefficient Kappa obtenu pour chaque paire d'annotateurs. Le taux d'accord varie de modéré ($0,47$ et $0,57$) à bon ($0,61$).

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

TAB. 6 – *Accord inter-annotateur estimé avec le coefficient Kappa.*

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

TAB. 7 – *Exemples de désaccord entre les annotateurs 1 et 2.*

Le tableau 7 présente des exemples de désaccord au niveau de l'identification des sessions entre les annotateurs 1 et 2. Pour l'utilisateur 7, l'annotateur 1 a considéré que les requêtes `China party` et `china financial` portent sur une thématique à part entière, tandis que l'annotateur 2 les a reliées avec la thématique des autres requêtes de cet épisode, relative à la banque. Il s'agit d'un cas difficile, car s'il est possible de trouver un lien entre ces requêtes à l'aide des termes *financial* et *bank*, il est également possible de séparer les requêtes contenant le terme *bank* de celles contenant le terme *china*. Pour l'utilisateur 72, l'annotateur 1 a identifié un lien que l'annotateur 2 n'a pas trouvé : Jean Ladrière était un philosophe et logicien belge. Nous mettons donc en avant deux types d'erreurs susceptibles d'intervenir lors de l'identification des sessions, provenant d'une part de la difficulté de choisir le lien le plus pertinent parmi plusieurs liens possibles, et d'autre part de la difficulté d'exploitation systématique d'une ressource externe pour les liens sémantiques faibles.

4.4 Collection de référence constituée

Étant donnée la présence de points de désaccord entre les annotateurs, nous avons constitué notre collection de référence en sélectionnant pour chaque requête les annotations faisant l'objet d'un accord entre au moins deux annotateurs, ou, lorsque les identifiants de session diffèrent entre les trois annotateurs, en prenant en compte la détection ou non d'une nouvelle session. Pour les 947 requêtes constituant la collection de référence, 406 sessions ont ainsi été identifiées, chaque session contenant en moyenne $2,33$ requêtes. Ce résultat est comparable au cas d'une collection provenant d'un moteur de recherche généraliste sur le Web, pour lequel chaque session contient entre $2,33$ et $2,96$ requêtes (Gayo Avello, 2009).

5 Conclusion

Dans cet article, nous avons détaillé les différentes étapes de construction d'une collection de référence annotée au niveau des sessions. La collection ainsi constituée se compose de 947 requêtes soumises par 216 utilisateurs sur le portail *OpenEdition*, correspondant à 406 sessions de recherche. Nous avons également proposé une mesure de l'accord entre les annotations obtenues en utilisant le coefficient Kappa de Cohen. Nous avons ainsi montré que la tâche d'annotation manuelle des sessions de la collection possède un taux de concordance globalement modéré entre les annotateurs, avec un Kappa allant de $0,47$ à $0,61$.

Les cas de désaccord entre les annotateurs confirment qu'il s'agit d'une tâche non triviale, et soulèvent plusieurs difficultés essentiellement liées à l'identification du besoin d'information de l'utilisateur. Il s'avère notamment difficile de trancher lorsque les requêtes contiennent peu de mots et que ces derniers possèdent un sens suffisamment large pour donner lieu à une multiplicité de liens possibles entre les requêtes.

La collection de référence constituée servira à l'évaluation de diverses méthodes de détection automatique des sessions, et notamment de méthodes traitant le cas des épisodes multitâches. Nous pourrons alors procéder à la segmentation en sessions du journal de requêtes *OpenEdition* dans son ensemble à l'aide de la méthode la plus appropriée. Cette étape constitue le point de départ d'une étude plus globale sur les reformulations de requêtes, se basant sur le fait que chaque session est marquée par un contexte de recherche caractéristique observable à travers l'ensemble des requêtes soumises. Ces requêtes pourront alors faire l'objet d'une typologie en fonction de la similarité entre les contextes qui les contiennent.

6 Remerciements

Nous adressons nos plus vifs remerciements à Marin Dacos et l'équipe du Cléo pour leur collaboration et l'accès aux données du portail *OpenEdition*. Nous tenons également à remercier nos annotateurs Clémentine et Nicolas pour leur disponibilité et l'intérêt porté à ce travail.

Références

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*(1), 37–46.

Gayo Avello, D. (2009). A Survey on Session Detection Methods in Query Logs and a Proposal for Future Evaluation. *Information Sciences 179*(12), 1822–1843.

Göker, A. et D. He (2000). Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning. In P. Brusilovsky, O. Stock, et C. Strapparava (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems*, Volume 1892 of *Lecture Notes in Computer Science*, pp. 319–322. Springer-Verlag Berlin Heidelberg.

Jansen, B. J., A. Spink, C. Blakely, et S. Koshman (2007). Defining a Session on Web Search Engines. *Journal of the American Society for Information Science and Technology 58*(6), 862–871.

Jones, R. et K. L. Klinkner (2008). Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 699–708.

Lucchese, C., S. Orlando, R. Perego, F. Silvestri, et G. Tolomei (2011). Identifying Task-Based Sessions in Search Engine Query Logs. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 277–286.

Silverstein, C., H. Marais, M. Henzinger, et M. Moricz (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum 33*(1), 6–12.

Spink, A., M. Park, B. J. Jansen, et J. Pedersen (2006). Multitasking During Web Search Sessions. *Information Processing and Management 42*(1), 264–275.

Summary

Identifying the sessions of a search engine's users aroused the creation of several benchmark collections and the elaboration of different automatic detection methods. This task actually represents the starting point of numerous studies dealing with the research context and the users' information needs. We detail within this study the creation of a benchmark collection based on a query log from the *OpenEdition* portal. We present as well an evaluation of the manual annotations which constitute this collection. The resulting benchmark contains 947 queries corresponding to 406 sessions, with an inter-annotator agreement (Cohen's Kappa) varying from 0.47 to 0.61. This collection will be exploited for both the evaluation of sessions automatic detection methods and studies of query reformulations.