Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe

Simon Leva, Nicolas Faessel

Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe

Simon Leva[1] Nicolas Faessel[2]
(1) CLLE–ERSS : CNRS et Université de Toulouse (UMR 5263),
5 allées Antonio Machado, 31058 Toulouse Cedex 9
(2) IRIT : CNRS et Université de Toulouse (UMR 5505),
118 route de Narbonne, 31062 Toulouse Cedex 9
sleva@univ-tlse2.fr, nicolas.faessel@irit.fr

RÉSUMÉ

Les utilisateurs d'un système de recherche d'information mettent en œuvre des comportements de recherche complexes tels que la reformulation de requête et la recherche multitâche afin de satisfaire leurs besoins d'information. Ces comportements de recherche peuvent être observés à travers des journaux de requêtes, et constituent des indices permettant une meilleure compréhension des besoins des utilisateurs. Dans cette perspective, il est nécessaire de regrouper au sein d'une même session de recherche les requêtes reliées à un même besoin d'information. Nous proposons une méthode de détection automatique des sessions exploitant la collection de documents WIKIPÉDIA, basée sur la similarité des résultats renvoyés par l'interrogation de cette collection afin d'évaluer la similarité entre les requêtes. Cette méthode obtient de meilleures performances que les approches temporelle et lexicale traditionnellement employées pour la détection de sessions séquentielles, et peut être appliquée à la détection de sessions imbriquées. Ces expérimentations ont été réalisées sur des données provenant du portail *OpenEdition*.

ABSTRACT

Automatic search session detection exploiting results similarity from an external document collection

Search engines users apply complex search behaviours such as query reformulation and multitasking search to satisfy their information needs. These search behaviours may be observed through query logs, and constitute clues allowing a better understanding of users' needs. In this perspective, it is decisive to group queries related to the same information need into a unique search session. We propose an automatic session detection method exploiting the WIKIPEDIA documents collection, based on the similarity between the results returned for each query pair to estimate the similarity between queries. This method shows better performance than both temporal and lexical approaches traditionally used for successive session detection, and can be applied as well to multitasking search session detection. These experiments were conducted on a dataset originating from the *OpenEdition* Web portal.

MOTS-CLÉS : Recherche d'information, détection automatique de sessions de recherche, analyse de journal de requêtes.

KEYWORDS: Information retrieval, automatic search session detection, query log analysis.

1 Introduction

De plus en plus d'utilisateurs effectuent des recherches d'information sur le Web. Ils utilisent pour cela des moteurs de recherche, leur permettant d'exprimer leur besoin d'information sous la forme de requêtes constituées de mots-clés. Ces systèmes atteignent cependant leurs limites face à des requêtes comportant en moyenne deux ou trois mots-clés, n'exprimant pas un besoin d'information suffisamment explicite par rapport à l'ensemble des documents disponibles (Silverstein *et al.*, 1999). En particulier, les requêtes soumises par les utilisateurs tendent à être trop génériques ou trop spécifiques, nécessitant un certain nombre de reformulations avant d'obtenir un ensemble de documents pertinents (Downey *et al.*, 2007). Les requêtes d'un utilisateur sont donc rarement isolées, mais font essentiellement partie d'une session de recherche. Les sessions fournissent de nombreux indices sur l'objectif de l'utilisateur ou son expertise dans le domaine considéré, et constituent ainsi une unité qu'il peut être utile d'identifier en vue d'améliorer les performances d'un moteur de recherche.

Nous avons montré dans une précédente étude (Leva, 2013) que la segmentation d'un journal de requêtes en sessions de recherche n'est pas une tâche triviale pour des annotateurs humains, aboutissant à un taux d'accord modéré. Nous avons également observé que la réalisation de cette tâche a nécessité de la part des annotateurs une consultation de plusieurs ressources externes afin de pallier un manque de connaissances encyclopédiques, et ainsi permettre une prise de décision. Nous faisons donc l'hypothèse qu'il serait possible de développer une méthode de détection automatique des sessions basée sur la similarité entre les requêtes à partir de leur croisement avec une collection de documents.

Nous présentons dans cet article l'élaboration d'une méthode de détection de sessions exploitant les documents de la partie française de *Wikipédia* [1]. Cette méthode est évaluée sur une collection de référence construite à partir d'un journal de requêtes issu du portail *OpenEdition* [2], et les résultats obtenus sont comparés à deux méthodes de référence.

Dans une première section, nous présentons la notion de session à travers ses définitions et les différentes approches de détection automatique. Puis, nous présentons les données que nous avons utilisées en vue de nos expérimentations. Nous détaillons enfin les différentes méthodes de détection que nous avons mises en œuvre, avant d'en faire une évaluation.

2 État de l'art

Les journaux de requêtes *(query logs)* conservent une trace d'un certain nombre d'interactions entre des utilisateurs et un moteur de recherche (soumission et reformulation de requête, navigation sur les pages de résultats, consultation de documents...). Ces données permettent ainsi d'étudier les comportements de recherche des utilisateurs et fournissent des indices sur leurs besoins d'information. Dans cette perspective, la notion de *session de recherche* est centrale, entraînant le développement de diverses méthodes de détection automatique.

1. `http://fr.wikipedia.org`
2. `http://www.openedition.org/`

2.1 Définitions de la notion de session

Une session de recherche regroupe l'ensemble des requêtes soumises par un même utilisateur afin de satisfaire un même besoin d'information. Si cette idée de regrouper les différentes formulations d'un même besoin informationnel au sein d'une même unité est partagée dans les définitions de la notion de session, celles-ci font cependant l'objet de nombreuses variations. En effet, selon que leur structure soit envisagée de manière séquentielle ou imbriquée, les sessions vont comporter des caractéristiques différentes.

2.1.1 Structure séquentielle des sessions

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.*

Ainsi, les requêtes appartenant à une même session se succèderaient dans l'ordre chronologique de leur soumission, aboutissant à une organisation des requêtes successives en séquences. L'une des implications de cette conception est que les sessions sont marquées par une longueur, que celle-ci soit exprimée en termes de nombre de requêtes ou d'unité de temps. D'une part, une session peut contenir une seule ou plusieurs requêtes (Gayo Avello, 2009). Dans ce dernier cas, la requête initiale est suivie d'une ou plusieurs reformulations (He *et al.*, 2002). D'autre part, au niveau temporel, la durée d'une session peut varier de moins d'une minute (Spink *et al.*, 2006), quelques minutes (He et Göker, 2000), à quelques heures (Spink *et al.*, 2006). Dans ces différents cas, la durée d'une session reste courte et inférieure à une journée. En effet, l'identification des utilisateurs dans un journal de requête se basant sur l'adresse IP, et celle-ci pouvant changer toutes les 24 heures, il est difficile de retrouver un utilisateur unique au-delà de cette période (Gayo Avello, 2009). Certains auteurs fixent ainsi une fenêtre temporelle de 24 heures sur les requêtes provenant d'une même adresse IP (Jansen *et al.*, 2007; Gayo Avello, 2009), correspondant à la notion d'*épisode de recherche*. Un épisode peut donc comporter une ou plusieurs sessions de recherche.

2.1.2 Structure imbriquée des sessions

Si la vision des sessions en tant que séquences de requêtes successives coïncide avec les enregistrements temporellement ordonnés constitués par les journaux de requêtes, elle ne reflète pas la complexité des parcours de recherche des utilisateurs. En effet, ces derniers peuvent mener une recherche simultanément sur plusieurs thèmes (par exemple à travers l'utilisation de plusieurs onglets dans leur navigateur), ou interrompre momentanément leur recherche en cours pour s'intéresser à un nouveau besoin d'information. Ce comportement, correspondant à une recherche multitâche *multitasking search*, peut se traduire par une alternance entre des requêtes visant chacune un besoin d'information distinct, et donc par des sessions imbriquées entre elles. Comme le montre l'étude de (Spink *et al.*, 2006), les recherches multitâches peuvent être très fréquentes dans certains environnements : dans un journal de requêtes du moteur *AltaVista*, respectivement 81 % et 91 % des séquences de 2 et 3 requêtes portent sur plusieurs thèmes à la fois. (Jones et Klinkner, 2008) envisagent ainsi que les requêtes liées à un même besoin d'information ne

sont pas nécessairement contiguës, mais peuvent s'intercaler avec des requêtes liées à un autre besoin d'information, donnant lieu à une imbrication entre sessions. Malgré cette autre manière d'envisager la structure des sessions, la notion d'épisode et son implication temporelle reste applicable, car liée à la problématique d'identification des utilisateurs.

2.2 Méthodes de détection automatique des sessions

Selon la structure des sessions adoptée, les méthodes de détection automatique font appel à des caractéristiques des sessions et des ressources différentes. Ces méthodes peuvent ainsi exploiter la durée des sessions, le contenu lexical des requêtes, et des sources de connaissance externes.

2.2.1 Méthode basée sur la durée des sessions

La première méthode de détection automatique des sessions à avoir été développée s'appuie sur la dimension temporelle des sessions, et envisage donc leur structure comme séquentielle. Cette méthode repose sur l'observation que plus la durée entre deux requêtes consécutives est longue, moins il est probable que ces requêtes renvoient à un même besoin d'information, et donc qu'elles appartiennent à une même session. Tout l'enjeu réside alors dans le choix d'un seuil temporel approprié fixant la durée maximale entre deux requêtes successives appartenant à la même session : 5 minutes (Silverstein *et al.*, 1999), 15 minutes (He et Göker, 2000), ou encore 30 minutes (Jansen *et al.*, 2007). Malgré sa forte utilisation due à sa simplicité de mise en œuvre, cette approche ne détecte ni les sessions très courtes résultant d'un changement soudain du besoin d'information, ni à l'inverse les sessions très longues au cours desquelles l'utilisateur peut effectuer des pauses importantes entre chaque requête. La prise en compte de ces cas nécessite en effet de s'appuyer sur d'autres indices de lien entre les requêtes.

2.2.2 Méthode basée sur le contenu lexical des requêtes

Afin de dépasser les limites de l'approche temporelle, une méthode de détection automatique exploitant le lien lexical entre les requêtes visant un même besoin d'information a été élaborée. L'hypothèse est alors que plus les requêtes ont un contenu lexical en commun, plus il est probable qu'elles appartiennent à une même session. La détection de ces liens lexicaux a principalement été envisagée à travers la tâche de détection des reformulations entre des requêtes successives. Plusieurs types de reformulation ont ainsi été définis (He *et al.*, 2002; Ozmutlu et Çavdur, 2005; Jansen *et al.*, 2007) : spécialisation (ajout d'un ou de plusieurs termes), généralisation (suppression d'un ou de plusieurs termes), reformulation (ajout et suppression d'un ou de plusieurs termes), etc. Si aucun de ces types de reformulation n'est identifié entre des requêtes successives, il est alors considéré que celles-ci ne sont pas lexicalement reliées, et appartiennent donc à des sessions différentes. Cette méthode a également été combinée avec la méthode temporelle, que ce soit à travers l'apprentissage automatique (He *et al.*, 2002; Ozmutlu et Çavdur, 2005) ou une interprétation géométrique (Gayo Avello, 2009). Néanmoins, l'approche lexicale possède deux principaux inconvénients : elle nécessite la présence d'au moins un mot commun entre les requêtes, et se heurte aux phénomènes de changement sémantique (synonymie, hyperonymie, hyponymie...). Le lien entre les requêtes visant un même besoin d'information

n'étant pas toujours lexicalement explicite mais pouvant être d'ordre sémantique, de nouvelles approches envisagent ainsi d'autres façons de détecter la similarité existant entre ces requêtes.

2.2.3 Méthode basée sur des sources de connaissance externes

Les approches temporelle et lexicale de détection automatique se basent sur une vision séquentielle des sessions, ne prenant une décision qu'à partir de la comparaison entre les requêtes successives d'un même utilisateur. De plus, ces approches ne permettent pas d'exploiter le lien de similarité souvent implicite entre les requêtes d'une même session. Plusieurs méthodes exploitent ainsi des sources de connaissance externes au journal de requête afin d'évaluer la similarité entre requêtes non plus de manière directe, mais à travers une représentation plus riche du contenu de ces requêtes. Ce niveau de représentation permet donc d'une part d'envisager la détection des sessions imbriquées, et d'autre part de prendre en compte toute la complexité de la tâche de détection des sessions.

(Jones et Klinkner, 2008) développent une approche basée sur un apprentissage supervisé exploitant des traits temporels, lexicaux, de cooccurrence des requêtes dans un journal plus étendu, et de similarité des requêtes avec les 50 premiers documents retournés en résultats. (Lucchese *et al.*, 2011) combinent deux mesures de similarité entre les requêtes : une similarité lexicale associant mesure de Jaccard et distance de Levenstein, ainsi qu'une similarité sémantique utilisant la mesure du cosinus sur une expansion des requêtes à l'aide des corpus WIKIPÉDIA et WIKTIONARY. Enfin, (Kramár et Bieliková, 2012) exploitent la similarité entre les métadonnées des documents pertinents cliqués pour chaque requête afin d'estimer la similarité entre chaque paire de requêtes. La pertinence des documents est ici estimée à l'aide d'un retour implicite, effectué par le système, des actions de l'utilisateur *(implicit feedback)*. Que ce soit au travers des documents constituant les résultats de la requête ou d'une collection de documents externe, ces approches présentent de meilleures performances que les approches lexicale et temporelle. En effet, elles permettent de prendre en considération la variété des comportements de recherche des utilisateurs, tant au niveau de la soumission de requêtes non contiguës qu'au niveau de la nature implicite du thème des requêtes. Cela pose également la question de l'évaluation des sessions imbriquées obtenues, présentant une perspective différente de celle des sessions séquentielles.

3 Données

Nous avons appliqué et évalué nos méthodes de détection automatique sur une collection de référence manuellement annotée en sessions provenant d'un journal de requêtes du portail *OpenEdition*. Nous présentons cet environnement de recherche ainsi que le journal de requêtes original avant de détailler la collection de référence utilisée.

3.1 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 363 revues et 16 collections de livres, *Calenda* recense plus de 21 000 évènements scientifiques en lettres et en sciences humaines et sociales, tandis qu'*Hypotheses.org* héberge 613 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.

3.2 Journal de requêtes initial

Nous avons exploité un journal de requêtes provenant du portail *OpenEdition* contenant une collection de 1 057 471 requêtes soumises par 227 302 utilisateurs durant la période du 07 avril 2010 au 1er février 2012. La langue principale des requêtes est le français, mais certaines sont également en anglais ou en espagnol. À la différence des requêtes soumises à un moteur de recherche généraliste, l'environnement d'*OpenEdition* se distingue par le fait que les requêtes proviennent essentiellement d'acteurs du monde académique et ciblent des revues, des évènements ou des blogs dans le domaine des sciences humaines et sociales.

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 *(access log)* original. En particulier, les données ont subi plusieurs opérations de nettoyage et de filtrage visant à éliminer les informations inexploitables (requêtes soumises par des robots d'indexation, suites de signes de ponctuation, etc.). 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.

3.3 Collection de référence

Dans une précédente étude (Leva, 2013), nous avons constitué une collection de référence à partir d'un échantillon du journal de requêtes *OpenEdition* comportant 947 requêtes soumises par 216 utilisateurs. Cette collection a été manuellement annotée en sessions afin de servir de référence à la fois pour l'évaluation de méthodes de détection automatique des sessions et pour des études sur les types de reformulations effectuées par les utilisateurs. L'ensemble des requêtes 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. La durée entre chaque requête successive est connue, mais cette information n'est pas présentée aux annotateurs. Trois annotateurs non spécialistes dans les domaines représentés par les documents ont été chargés de regrouper les requêtes de chaque épisode de recherche en une ou plusieurs sessions.

La tâche d'annotation des sessions a fait l'objet d'un guide d'annotation. Pour chaque épisode de recherche d'un utilisateur, les annotateurs devaient observer l'ensemble des requêtes soumises avant de prendre une décision et d'attribuer une session à chaque requête à travers un identifiant numérique unique. La collection de référence contient ainsi des sessions imbriquées. Afin d'identifier les requêtes visant un même besoin d'information, les annotateurs pouvaient

s'appuyer sur plusieurs indices, à la fois textuels, sémantiques, et de proximité thématique, ou s'appuyer à défaut sur des ressources externes permettant de pallier un manque de connaissances encyclopédiques. Un accord inter-annotateur a été évalué à l'aide du coefficient Kappa, le taux d'accord variant de modéré (0, 47 et 0, 57) à bon (0, 61). La collection de référence finale contient ainsi 406 sessions pour les 947 requêtes initiales, résultant des annotations faisant l'objet d'un accord entre au moins deux annotateurs.

4 Méthodes de détection automatique de sessions

Nous proposons d'utiliser différentes méthodes de détection automatiques de sessions : une méthode temporelle, une méthode lexicale se basant sur les indices lexicaux des requêtes pour identifier si une requête est une reformulation de la requête précédente, et enfin une méthode exploitant la collection de documents WIKIPÉDIA. Les deux premières méthodes, qui sont des méthodes provenant de la littérature, nous serviront de références pour la détection de ruptures de sessions dans le contexte d'identification de sessions de recherche séquentielles, que nous proposons dans la section 5.1.

4.1 Exploitation d'un seuil temporel

À partir d'un seuil fixant la durée maximale entre deux requêtes successives pouvant appartenir à la même session, la méthode temporelle détecte les ruptures (durée entre les requêtes supérieure au seuil) et les continuités (durée entre les requêtes inférieure au seuil) de session au sein de chaque épisode de recherche d'un même utilisateur. Chaque requête se voit ainsi attribuer un identifiant de session unique au sein d'un même épisode. Ce type d'approche n'envisageant les requêtes que d'un point de vue séquentiel, les sessions imbriquées ne sont pas repérées, et les identifiants de session sont constamment incrémentés pour chaque nouvelle session d'un épisode.

4.2 Exploitation des liens lexicaux pour la détection de reformulation

Dans leur étude, (Jansen *et al.*, 2007) proposent de détecter différents types de reformulation en utilisant les liens lexicaux entre deux requêtes. L'idée est de compter le nombre de mots communs à deux requêtes, et de déterminer ensuite, grâce à la longueur de chacune, si l'utilisateur a spécifié sa requête, ou bien l'a généralisée, etc. Dans le cas de sessions séquentielles, sans imbrications, on peut faire l'hypothèse que si une reformulation est détectée entre deux requêtes consécutives, c'est que celles-ci font partie d'une même session. Sinon, la dernière requête représente une nouvelle session : il y a eu une rupture de la session de recherche précédente. Dans le cadre de sessions imbriquées, la seule exploitation des liens sémantiques est plus délicate : la temporalité est implicitement utilisée dans la reformulation, car une requête ne peut être une reformulation que d'une requête antérieure. Ainsi, pour détecter les sessions imbriquées, il faut comparer toutes les requêtes entre elles en préservant leur ordre temporel.

4.3 Exploitation de la similarité des requêtes à l'aide de *Wikipédia*

Comme nous l'avons mentionné dans la section 2, l'utilisation de la temporalité et la détection des différents types de reformulation ne sont pas suffisantes pour la détection de sessions séquentielles, et encore moins dans le cadre des sessions imbriquées. Ainsi, nous proposons l'utilisation d'une source d'information externe qui est une version locale de *Wikipédia* en français, datant du 28 octobre 2012 [3]. Ce corpus a été indexé dans un moteur de recherche (*Terrier* [4]), permettant ainsi son interrogation avec les requêtes des épisodes de recherche.

Contrairement aux approches exploitant la sémantique des documents provenant d'une source externe (comme par exemple la *wikification* proposée par (Lucchese *et al.*, 2011; Kramár et Bieliková, 2012)), nous exploitons la liste des documents, ainsi que leur score, renvoyés par le système de recherche. Ainsi, les listes de résultats obtenus pour chaque requête d'un même épisode sont comparées, afin d'estimer si les requêtes forment une session de recherche.

Soit E un épisode de recherche. Soit Q_E l'ensemble des requêtes de l'épisode de recherche. Soit R_{q_a} l'ensemble des résultats de la requête $q_a \in Q_E$. R_{q_a} est constitué d'un ensemble de documents pondérés. Le poids des documents correspond aux scores obtenus par ces documents lors de l'interrogation dans un moteur de recherche. Le nombre de documents renvoyés par le moteur de recherche est fixé à un maximum de 1 000. Ce paramètre n'a pas été étudié dans le présent article, bien que le nombre de documents réellement pertinent dans *Wikipédia* pour une requête donnée est probablement plus faible. R_{q_a} peut être représenté comme un vecteur de dimension M, où M correspond à l'espace des documents du corpus, et dont les coordonnées sont données par le score des documents appartenant à M, obtenus pour la requête q_a. On peut représenter l'ensemble des documents renvoyés pour une requête comme un vecteur $\overrightarrow{v}_a = (s_{1,q_a}, s_{2,q_a}, \ldots, s_{i,q_a}, \ldots, s_{M,q_a})$, où s_{i,q_a} correspond au score du document i pour la requête q_a.

La similarité entre deux requêtes q_a et q_b est donnée par le cosinus de l'angle des vecteurs de documents répondant à ces deux requêtes :

$$sim(v_a, v_b) = \frac{\overrightarrow{v}_a \bullet \overrightarrow{v}_b}{|\overrightarrow{v}_a| \times |\overrightarrow{v}_b|} \tag{1}$$

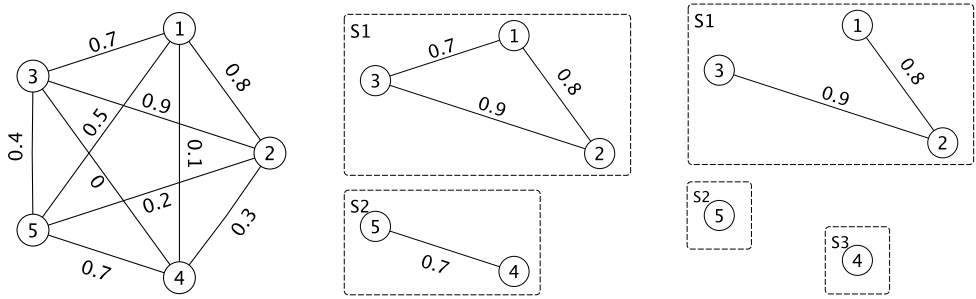
Un épisode peut être représenté comme un graphe valué non orienté complet $\mathscr{G}_{\mathscr{E}}(N, A)$, dont les nœuds N correspondent aux requêtes, et les arêtes A correspondent à la similarité entre deux requêtes (figure 1a). On peut définir le graphe de sessions $\mathscr{G}_{\mathscr{S}}(N, A')$, comme le sous graphe partiel potentiellement non connexe de $\mathscr{G}_{\mathscr{E}}$ dont l'ensemble des arêtes A' ont une valeur supérieure à un seuil t, dont chaque composante connexe forme une session de recherche.

Les figure 1b et 1c représentent deux graphes de sessions, déterminés respectivement pour un seuil de similarité $t = 0,7$ et $t = 0,8$.

5 Analyse des résultats

Nous proposons une analyse des résultats de notre méthode de détection automatique (cf. section 4.3) en prenant en compte à la fois une perspective séquentielle et imbriquée sur les

3. `http://dumps.wikimedia.org/frwiki/20121028/frwiki-20121028-pages-articles.xml.bz2`
4. `http://terrier.org`



(a) Graphe de l'épisode $\mathscr{G}_{\mathscr{E}}$ (b) Graphe de sessions pour le seuil 0.7 (c) Graphe de sessions pour le seuil 0.8

FIGURE 1 – Représentation en graphe des épisodes et sessions de recherche

sessions détectées. Au niveau de l'évaluation des sessions séquentielles, nous effectuons une comparaison entre notre méthode et celles basées sur une approche temporelle et lexicale.

5.1 Évaluation des sessions séquentielles

Nous avons exploité la collection de référence OPENEDITION pour l'évaluation des trois méthodes de détection automatique des sessions ainsi que pour le réglage des paramètres internes de chacune de ces méthodes. Dans le cadre de sessions séquentielles, cette évaluation peut être réalisée à l'aide des mesures de précision, rappel et F-mesure, appliquées au nombre de ruptures et de continuations de sessions détectées par chaque système. Une rupture peut être détectée entre deux requêtes consécutives si, dans la méthode temporelle, le temps entre les deux requête est supérieur à un délai donné, et dans le cas de la méthode lexicale, aucun cas de reformulation n'est identifié. Ce type d'évaluation ne permet donc pas de refléter entièrement les performances d'un système dans le cadre de la détection de sessions imbriquées.

5.1.1 Mesures d'évaluation

Les équations suivantes présentent les mesures d'évaluation proposées par (Gayo Avello, 2009) dans le cadre d'une comparaison avec une référence manuelle. La précision P correspond au nombre de ruptures de session présentant un accord entre le système et la référence par rapport au nombre de ruptures de session détectées par le système. Le rappel R correspond au nombre de ruptures de session présentant un accord entre le système et la référence par rapport au nombre de ruptures de session de la référence. La F-mesure F permet de pondérer rappel et précision. Nous avons automatiquement annoté notre collection de référence en ruptures et continuations de session, les cas où deux requêtes successives d'un même épisode possèdent un même identifiant de session correspondant à une continuation, les cas contraires à une rupture. Nous avons également supprimé les cas triviaux pour lesquels les systèmes n'ont aucune décision à prendre (la première requête de chaque épisode et un épisode ne contenant qu'une seule requête constituent toujours une rupture de session) afin d'évaluer leurs performances réelles,

réduisant à 598 le nombre de requêtes de la collection de référence initiale.

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

5.1.2 Méthode temporelle

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

(a) Mesures d'évaluation

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

(b) Matrice de confusion

TABLE 1 – Résultats de l'évaluation de la méthode temporelle.

Le tableau 1a présente les performances de la méthode temporelle pour une durée maximale des sessions de 640 secondes. Après avoir testé cette méthode avec des valeurs de seuil allant de 10 à 5 120 secondes, ce seuil offre en effet les meilleurs résultats sur notre collection.

Le tableau 1b présente la matrice de confusion des résultats de la méthode temporelle. Cette méthode possède une efficacité de 84 %, correspondant à la proportion des vrais positifs (22) et des vrais négatifs (479) par rapport à l'ensemble des cas traités (598). Nous observons que la méthode temporelle entraîne autant de faux positifs (tableau 2, utilisateur 18) que de faux négatifs (tableau 2, utilisateur 32). Cela est directement lié à l'incapacité de cette méthode de s'adapter à la durée propre à chaque session. Ainsi, il s'est écoulé 4 436 secondes entre les requêtes de l'utilisateur 18 qui sont explicitement liées au niveau lexical, et 315 secondes entre les requêtes de l'utilisateur 32 qui portent sur des thèmes distincts.

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

TABLE 2 – Exemple de faux positif et de faux négatif induits par la méthode temporelle.

5.1.3 Méthode lexicale

Le tableau 3a présente les performances de la méthode lexicale sur notre collection de référence.

Le tableau 3b présente la matrice de confusion des résultats de la méthode lexicale, montrant une efficacité de 62 %. La méthode lexicale n'entraîne aucun faux négatif. En effet, de par son fonctionnement, cette méthode considère par défaut qu'il existe une rupture de session entre les requêtes si aucun type de reformulation n'est détecté, ce qui explique le taux de rappel de 1

Précision	0,24
Rappel	1
F-mesure	0,38

(a) Mesures d'évaluation

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

(b) Matrice de confusion

TABLE 3 – Résultats de l'évaluation de la méthode lexicale.

observé. Cette méthode est donc sensible au contenu lexical des requêtes, et les faux positifs proviennent de l'absence de mots communs entre les requêtes, des fautes de frappe non palliées par la distance d'édition, ainsi que de la non détection des liens sémantiques entre les requêtes. C'est le cas dans le tableau 4 pour l'utilisateur 35, dont l'ensemble des requêtes renvoie au thème des produits éclaircissants pour la peau.

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

TABLE 4 – Exemple de faux positif induit par la méthode lexicale.

5.1.4 Méthode basée sur *Wikipédia*

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

(a) Mesures d'évaluation

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

(b) Matrice de confusion

TABLE 5 – Résultats de l'évaluation de la méthode basée sur *Wikipédia*.

Le tableau 5a présente les performances de la méthode basée sur *Wikipédia* pour un seuil de similarité entre requêtes fixé à 0,005. Après avoir testé cette méthode avec des valeurs de seuil allant de $1 \cdot 10^{-5}$ à 0,5, ce seuil offre en effet les meilleurs résultats sur notre collection.

Le tableau 5b présente la matrice de confusion des résultats de la méthode basée sur *Wikipédia*, montrant une efficacité de 77 %. Cette méthode génère presque neuf fois plus de faux positifs que de faux négatifs. Le cas de l'utilisateur 7 dans le tableau 6 est un exemple de faux négatif. Ces cas correspondent à des requêtes contenant peu de termes ou des termes génériques, facilitant la découverte de liens thématiques entre elles, et constituent également des cas ambigus pour les annotateurs. Il aurait ainsi été possible de regrouper l'ensemble des requêtes de l'utilisateur 7 au sein d'une même session ayant pour thème la finance. Les perspectives d'améliorations sont

donc faibles pour ce type d'erreurs. En revanche, les faux positifs sont dûs à des fautes de frappe et pour l'essentiel aux limites de la collection de documents de *Wikipédia*, qui contiennent peu ou pas d'information concernant certaines requêtes très caractéristiques de l'environnement de recherche *OpenEdition*. C'est le cas de l'utilisateur 71, qui effectue une recherche portant sur un auteur de revue spécifique de la plateforme *Revues.org*. Une solution serait alors d'exploiter la collection de documents d'*OpenEdition* conjointement à ceux de *Wikipédia* de manière à avoir une couverture à la fois spécifique et générique sur les sujets introduis par les requêtes.

Nous pouvons néanmoins observer qu'au niveau de la détection des ruptures de sessions, la méthode basée sur *Wikipédia* donne une précision égale à celle de la méthode temporelle (0,31) tout en améliorant le rappel (0,8). Les performances globales de cette approche, représentées par une F-mesure de 0,45, sont meilleures que celles des approches temporelle et lexicale implémentées.

Util.	Requête	Réf.	Syst.
71	`philippe dasseto`	2	2
	`dasseto`	2	3
	`Dasseto`	2	4
7	`bank`	1	1
	`bank crisis`	1	1
	`China party`	2	1
	`china financial`	2	1
	`bank`	1	1

TABLE 6 – Exemple de faux positif et de faux négatif induits par la méthode basée sur *Wikipédia*.

5.2 Évaluation des sessions imbriquées

Les données sur lesquelles nous avons effectué nos expérimentations contiennent des sessions de recherche imbriquées. Cette expérimentation permet de valider, non plus la détection des points de rupture, mais bien les sessions renvoyées par notre système par rapport aux sessions de référence annotées manuellement.

Nous utilisons ici certaines des métriques définies par (Lucchese *et al.*, 2011). Ces mesures sont l'index de Rand (Rand, 1971) et l'index de Jaccard (Jaccard, 1901), qui considèrent des paires de requêtes et permettent de vérifier la cohérence de répartition de ces paires entre les sessions système et les sessions de référence d'un même épisode de recherche.

On considère f_{11} le nombre de paires qui sont dans une même session calculée et dans une même session de référence, f_{00} le nombre de paires reparties dans des sessions calculées différentes, ainsi que dans des sessions de référence, f_{01} le nombre de paires qui sont dans une même session calculée mais dans des sessions de références différentes, f_{10} le nombre de paires qui sont des sessions calculées différentes, mais dans une même session de référence.

L'index de Rand est défini comme suit :

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

et l'index de Jaccard :

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

Seuil	0,00001	0,00005	0,0001	0,0005	0,001	0,005	0,01
Rand	0,78556	0,78556	0,78556	**0,78796**	0,78646	0,75840	0,72841
Jaccard	0,72588	0,72588	0,72588	**0,72829**	0,72228	0,66834	0,62587

TABLE 7 – Similarité des sessions système par rapport aux sessions de référence.

Les résultats du tableau 7 montrent que le meilleur seuil de similarité pour la génération automatique des sessions est égal à 0,0005. Ce seuil, très bas, s'explique par le fait que la liste des résultats obtenus pour chaque requête est très sensible aux différents types de reformulation. En effet, l'ajout ou la suppression d'un mot dans la requête peut changer totalement les résultats renvoyés par le moteur de recherche utilisé. Ainsi, bien que l'utilisation de la similarité entre les résultats des requêtes puisse aider à détecter des sessions imbriquées, il semble évident que le seuil de détection optimal de ces sessions est dépendant de notre jeu de données.

6 Conclusion

Nous avons proposé une méthode de détection des sessions de recherche basée sur l'utilisation d'une ressource externe. Cette méthode, exploitant les résultats d'un moteur de recherche sur la collection de documents de *Wikipédia*, a été validée dans le cas de sessions de recherche séquentielles. Nous avons également proposé une extension dans le cadre de la recherche de sessions imbriquées. Une perspective à court terme est de valider la détection de sessions imbriquées par rapport aux autres approches de la littérature utilisées dans ce cadre.

Les résultats préliminaires obtenus sont encourageants, et nous envisageons d'étudier la combinaison des trois approches, à savoir temporelle, lexicale, et basée sur l'utilisation de *Wikipédia*. En effet, il nous semble que ces approches sont complémentaires : l'approche utilisant la similarité des requêtes calculée au moyen d'une ressource externe est très sensible aux reformulations identifiées dans l'approche lexicale. Une idée serait que lorsque la reformulation est identifiée de manière triviale, celle-ci prédomine dans la détection de session. Si ce n'est pas le cas, le système peut utiliser la similarité pour détecter les sessions. Une autre piste que nous envisageons concerne l'exploitation de la collection de documents de la plateforme *OpenEdition*, sans doute mieux adaptée au contenu des requêtes de notre journal.

Remerciements

Ce travail s'inscrit dans le projet ANR CAAS (*Contextual Analysis and Adaptive Search*, programme Contint) coordonné par Josiane Mothe (IRIT), 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). Nous adressons également nos plus vifs remerciements à Marin Dacos

et l'équipe du Centre pour l'édition électronique ouverte (Cléo) pour leur collaboration et la mise à disposition des données du portail *OpenEdition*.

Références

Downey, D., Dumais, S. et Horvitz, E. (2007). Models of searching and browsing : Languages, studies, and applications. *In Proc. IJCAI*, pages 2740–2747.

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.

He, D. et Göker, A. (2000). Detecting Session Boundaries from Web User Logs. *In Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pages 57–66.

He, D., Göker, A. et Harper, D. J. (2002). Combining Evidence for Automatic Web Session Identification. *Information Processing and Management*, 38(5):727–742.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Jansen, B. J., Spink, A., Blakely, C. et Koshman, S. (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 Klinkner, K. L. (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*, pages 699–708.

Kramár, T. et Bieliková, M. (2012). Detecting Search Sessions Using Document Metadata and Implicit Feedback. *In Proceedings of the WSCD 2012 Workshop on Web Search Click Data*.

Leva, S. (2013). Les sessions de recherche comme contexte des requêtes. *In Actes de l'atelier Contextualisation de Messages Courts – 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*, pages 1–12.

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

Ozmutlu, H. C. et Çavdur, F. (2005). Application of Automatic Topic Identification on Excite Web Search Engine Data Logs. *Information Processing and Management*, 41(5):1243–1262.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850.

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

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