Mise en oeuvre d'une base de données graphe pour l'analyse des logs de requêtes en recherche d'information

Josiane Mothe, Sagun Pai

Mise en oeuvre d'une base de données graphe pour l'analyse des logs de requêtes en recherche d'information

Josiane Mothe* — **Sagun Pai****

* *Institut de Recherche en Informatique de Toulouse, UMR5505 CNRS*
ESPE, Université Jean Jaurès, Université de Toulouse, Josiane.Mothe@irit.fr
** *Indian Institute of Technology Bombay, sagung.pai@gmail.com*

RÉSUMÉ. *Les travaux présentés dans cet article concernent la mise en oeuvre d'une base de données orientée graphe pour l'étude des reformulations de requêtes réalisées par les utilisateurs d'un moteur de recherche. Notre objectif est de rechercher des patrons de reformulation à des fins d'analyse linguistique. Nous nous sommes appuyés sur un log de connexion issu d'un moteur de recherche associé à la librairie digitale Revue.org. Après avoir extrait les sessions de recherche, nous avons défini plusieurs types de liens pouvant être extraits des reformulations et nous avons créé une base de données orientée graphe avec ces données. Nous nous sommes ensuite appuyés sur ces liens pour analyser la structure du graphe obtenu et extraire des éléments sur l'utilisation des termes. Nous avons proposé différents patrons de reformulation de requêtes qui pourraient être utilisés dans une reformulation automatique des requêtes.*

ABSTRACT. *The work presented in this paper makes use of graph-oriented database to study users' query reformulations using a search engine. Our goal is to extract query reformulation patterns for further linguistic analysis. The log of connections we use is an extract of the one from a search engine associated with a digital library (Revue.org). After having extracted search sessions, we have defined several types of links that can be extracted from reformulations and created a graph-oriented database with the data. We then analyze these relationships to learn from the structure of the obtained graph.*

MOTS-CLÉS : *Base de données orientée graphe, Recherche d'information, Reformulation de requêtes, Analyse de logs de connexion, Visualisation graphique.*

KEYWORDS: *Graph-based database, Information retrieval, Query reformulation, Query log analysis, Graphical visualization.*

1. Introduction

Dans cet article, nous nous intéressons à l'utilisation des bases de données graphes dans le domaine de la recherche d'information (RI) et plus particulièrement dans l'analyse de logs de connexion issus de moteurs de RI.

La majorité des services en ligne sont accessibles via des requêtes posées par les utilisateurs. Cela est par exemple le cas pour les services de RI tels que les moteurs du Web, mais l'accès par des requêtes est également présent dans la plupart des sites marchands. Pour ces services, il est primordial que les utilisateurs accèdent à l'information qu'ils souhaitent, qu'ils soient satisfaits des réponses obtenues pour qu'ils soient fidèles au service. Lorsque les utilisateurs interagissent avec un système en ligne, les trâces qu'ils laissent sont des sources très riches. Ces trâces ne sont pas gardées uniquement pour des raisons légales, mais également parce qu'elles peuvent permettre, après analyse, d'améliorer la qualité du service.

De nombreux travaux se sont ainsi intéressés à l'analyse des logs de connexion. Dans le domaine de la sécurité, l'analyse de logs peut permettre de détecter un comportement atypique, pouvant être un signe annonciateur d'une attaque (Oliner *et al.*, 2012). Un autre type d'exploitation des logs de connexion concerne l'analyse de l'activité des utilisateurs afin de mieux comprendre la façon dont ils interagissent avec le système. Dans ce domaine, de nombreuses études ont étudié les reformulations de requêtes réalisées dans les sessions de recherche (Rieh et Xie, 2006), (Huang et Efthimiadis, 2009), (Hassan *et al.*, 2013), (Eickhoff *et al.*, 2014) ou les caractéristiques de celles-ci (Mothe et Tanguy, 2007), (Kompaore *et al.*, 2008), (Mizzaro et Mothe, 2016). Pour certains, il s'agit d'identifier les stratégies de reformulation pour en produire une typologie (Border, 2002), (Anick, 2003), (Huang et Efthimiadis, 2009) ; pour d'autres, il s'agit d'étudier le lien entre les stratégies de reformulation et l'expertise des utilisateurs comme par exemple dans (Eickhoff *et al.*, 2014) ou même pour prédire le comportement d'un utilisateur dans le temps (Radinsky *et al.*, 2012). De nombreux travaux se sont ainsi intéressés à l'analyse de logs de connexion.

Pour traiter ce type de données, différents environnements informatiques se sont développés. Les bases de données orientées graphes sont particulièrement adaptées pour gérer des données "reliées" c'est à dire qui peuvent être représentées par des noeuds et des liens. En effet, les bases de données graphe reposent sur des concepts bien connus de la théorie des graphes, composés de noeuds reliés par des arcs, ces derniers étant éventuellement orientés. Dans les bases de données graphes, noeuds et liens peuvent être typés et pondérés.

Les travaux que nous présentons dans cet article visent également l'étude des reformulations de requêtes des utilisateurs et étudient l'adaptation des bases de données NoSQL relationnelles pour ce problème. Plus spécifiquement, nous nous sommes intéressés d'une part aux termes pivots des reformulations ainsi qu'aux patrons de reformulation dans une perspective linguistique. En effet, les taxonomies actuelles de typologie de reformulation de requêtes se focalisent surtout sur les aspects macro tels que l'ajout de termes, la suppression de termes, etc.(Ihadjadene et Chaudiron, 2008)

mais n'étudient pas le niveau plus fin des reformulations, à savoir quels termes sont utilisés pour les différentes perspectives de reformulation. Notre travail vise cet objectif. Ainsi, contrairement à d'autres travaux, nous nous focalisons sur les reformulations de requêtes sans considérer les autres actions des utilisateurs comme la consultation des documents. Dans ce travail, nous souhaitions également acquérir une expertise dans l'utilisation des bases de données graphe et plus spécifiquement de l'implantation Néo4j `https://neo4j.com/`.

A termes, les résultats de ce type d'analyse pourraient permettre d'aider l'utilisateur dans sa formulation du besoin, en fonction de la tâche qu'il souhaite réaliser ou en fonction de l'angle qu'il choisit pour sa recherche. Cet aspect n'est toutefois pas traité dans cet article et constitue une perspective à ce travail.

Le reste de cet article est organisé comme suit : dans la section 2, nous présentons les travaux reliés en nous focalisant sur les travaux basés sur l'analyse des reformulations de requêtes dans les logs. Dans la section 3, nous présentons les données que nous avons utilisées et la façon dont sont extraites les informations utiles à notre analyse. La section 4 présente la modélisation des requêtes que nous avons définie. Cette modélisation s'appuie sur une représentation sous forme de graphes. Différents types de liens entre termes ont ainsi été définis. La section 5 présente les résultats de l'analyse que nous avons conduite. Enfin, la section 6 conclut cet article et présente les perspectives à ce travail.

2. Travaux reliés

La représentation sous forme de graphes des informations est particulièrement adaptée à la modélisation des réseaux sociaux. Par exemple, (Ghosh *et al.*, 2012) analysent le développement des liens (*link farming*) afin d'accroitre son influence dans le réseau social Twitter, les noeuds représentant les utilisateurs. (Tchuente *et al.*, 2012) utilisent un réseau égocentrique pour définir un modèle social du profil des utilisateurs dans un réseau social.

La représentation sous forme de graphes d'informations sémantiques a également une longue histoire en particulier via les graphes conceptuels (Mugnier et Chein, 1996). En RI, ils ont été utilisés pour représenter des relations sémantiques entre termes (Dousset *et al.*, 2011) pouvant ensuite être utilisées pour spécialiser/ généraliser des termes des requêtes (Chevallet, 1994). D'autres formes de graphes ont également été utilisées pour permettre la reformulation de requêtes de façon automatique. Par exemple, (Zenz *et al.*, 2009) proposent un modèle dans lequel la connaissance qui est représentée sous forme de graphes est utilisée pour transformer une requête "sac de mots" en une requête sémantique. Pour cela, les auteurs s'appuient sur des patrons de requêtes basés sur la sémantique qui relie les concepts. (Bendersky et Croft, 2012) modèlisent les requêtes des utilisateurs sous la forme d'un hypergraphe (un arc pouvant lier plus de deux noeuds) dans lequel un noeud est un concept (terme, bigramme, entité nommée, etc.) ou un document et les liens des relations sémantiques

entre les noeuds (liens document/concept ou concepts entre eux). Cette structure est utilisée comme base à un modèle de RI.

Les travaux les plus en lien avec ceux présentés dans cet article concernent la représentation des flôts de requêtes sous forme de graphes. Dans (Boldi *et al.*, 2011), le graphe modélise les transitions dans les reformulations de requêtes. Chaque noeud du graphe représente une requête et un arc est créé dès lors que les deux requêtes se suivent dans une session. Les liens sont typés en fonction du type de reformulation observé (généralisation/spécialisation, correction de fautes, etc.). Le graphe est ensuite analysé pour extraire des modèles de reformulation et pour caractériser les patrons de reformulation de requêtes. L'étude des types de reformulation à d'ailleurs donné lieu à différentes taxonomies (Anick, 2003), (Jansen, 2007) ; une comparaison de différentes taxonomies est présentée dans (Huang et Efthimiadis, 2009). Ces taxonomies s'attachent à distinguer les stratégies de reformulation en ré-ordonnancement de termes, ajout de termes, suppression de termes, modification de la ponctuation, correction orthographique, substitution de termes, etc.. Ces études ne s'intéressent pas aux détails des reformulations comme par exemple quels sont les types d'ajout de termes ou de suppression de termes ; quels termes sont utilisés dans les substitutions, etc..

Notre approche est différente dans la mesure où les noeuds du graphe que nous construisons sont les termes des requêtes et les arcs entre les noeuds correspondent au type de phénomène observé entre les différentes reformulations de requêtes. Nous proposons ainsi trois types de liens entre les termes des requêtes : de co-occurrence (les termes liés sont utilisés pour (re-)formuler une requête), de reformulation (les termes de la requête sont liés aux termes de la requête suivante de la même session), de remplacement (les termes sont substitués entre deux formulations de requête d'une même session). Nous visons donc à l'observation de phénomènes au niveau des termes eux-mêmes (et non au niveau des requêtes comme cela est le cas dans les autres travaux du domaine).

La section suivante présente le log de connexion que nous avons utilisé dans notre étude.

3. Revue.org et log de connexion

3.1. *Moteur revue.org*

La plateforme revues.org donne accès à environ 400 revues scientifiques de Sciences Humaines et Sociales. Cette plateforme est constituée d'un moteur principal (OpenEdition) accessible via une barre de recherche et de nombreux moteurs verticaux. Ces moteurs verticaux permettent aux utilisateurs de faire des recherches dans une revue scientifique en particulier. Ce dispositif a pour effet de renvoyer une requête au moteur OpenEdition accompagnée d'un paramètre « restriction à la revue considérée ». Différents filtres de recherche documentaire sont également disponibles : sélection sur auteur, titre, année, type de publication, etc. Un exemple de présentation des

Figure 1. *Interface de Revue.org via OpenEdition.*

résultats d'une recherche soumise au moteur OpenEdition est visible dans la Figure 1. Les résultats sont proposés sous la forme suivante : titre, auteur et snippet (compilation de contextes d'apparition des mots de la requête dans le document). Il y a également une coloration des mots de la requête présents dans les résultats.

```
21:25:07 → REQ[le petit livre rouge] → Pub[Perspectives chinoises]¶
21:25:30 → ---->DOC[http://perspectiveschinoises.revues.org/4983]¶
21:25:38 → ---->DOC[http://perspectiveschinoises.revues.org/5576]¶
21:26:10 → REQ[révolution culturelle]→ Pub[Perspectives chinoises]¶
21:26:52 → ---->DOC[http://perspectiveschinoises.revues.org/181]¶
21:28:11 → REQ[littérature chine populaire] → Pub[Perspectives chinoises]¶
21:28:47 → REQ[littérature Mao]→Pub[Perspectives chinoises]¶
21:29:27 → ---->DOC[http://perspectiveschinoises.revues.org/727]¶
21:29:40 → ------>PDF[http://perspectiveschinoises.revues.org/pdf/727]¶
21:30:17 → REQ[littérature Mao]→Pub[Perspectives chinoises]¶
```

Figure 2. *Requêtes sur "le petit livre rouge".*

```
17:11:14 → REQ[linguistique textuelle]→Pub[Semen]¶
17:12:32 → REQ[linguistique textuelle]→TypDoc[Article] | Pub[Semen]¶
17:15:14 → REQ[définition de la linguistique textuelle ]TypDoc[Article] | Pub[Semen]¶
17:15:46 → REQ[concept de la linguistique textuelle]→ TypDoc[Article] | Pub[Semen]¶
17:16:09 → ---->DOC[http://semen.revues.org/1995]¶
17:17:32 → REQ[cohérence · et · cohésion · de · la · linguistique · textuelle]
   → TypDoc[Article] | Pub[Semen]¶
17:18:26 → REQ[théorie de la linguistique textuelle] → TypDoc[Article] | Pub[Semen]¶
18:31:28 → REQ[théorie de la linguistique textuelle] → TypDoc[Compte-rendu]¶
18:31:43 → ---->DOC[http://mots.revues.org/831]¶
```

Figure 3. *Requêtes sur la "linguistique textuelle".*

3.2. *Extraction des éléments utiles du log de connexion*

Le log de connexion qui nous a été gracieusement fourni comprend 1 521 912 requêtes soumises entre le 07/04/2010 et le 01/02/2012. Ce fichier de connexion a été filtré en fonction de la date des sessions ; les sessions retenues doivent contenir au moins une requête suivie de la consultation d'au moins un document. Par ailleurs, nous éliminons également les requêtes qui sont arrivées sur Revues.org à partir d'une requête initiale sur Google. Enfin, les requêtes répétées au cours d'un parcours de recherche ont été fusionnées.

Le log de connexion ainsi obtenu comprend 130 141 requêtes. Nous avons ré-écrit le log de connexion pour nous centrer sur deux éléments : les requêtes formulées et les documents cliqués.

Les figures 2 et 3 donnent des exemples d'extraits de sessions de recherche.

Le log de connexion a été ensuite découpé en sessions.

Il existe différentes méthodes dans la littértaure pour le découpage en sessions. En 1999, (Silverstein, 1999) définissait une session comme une série de requêtes posées

par un même utilisateur dans une période courte de temps et qui regroupe une tentative d'un utilisateur pour répondre à un besoin d'information unique. (Jansen, 2007) considère la notion d'épisode de recherche (série temporelle d'actions réalisées dans une période de temps) qui peut regrouper plusieurs sessions, chacune visant un besoin d'information spécifique. Une définition similaire est retenue dans (Leva et Faessel, 2013) pour lesquels :

– 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 multi-tâche.

Concrètement, définir les frontières entre sessions dans un épisode peut être assez complexe. Aussi, plusieurs études se sont appuyées sur des frontières temporelles : une session est alors définie par rapport à une durée. (He et Harper, 2002) ont montré que cette méthode très simple à implanter donne un résultat de l'ordre de 73% de précision et de 62% de rappel.

Dans cette étude, nous avons découpé le log de connexion en considérant une fenêtre de temps de 24h.

Le tableau 1 donne quelques caractéristiques du log de connexion :

Nombre de requêtes	130 141
Nombre de sessions	45 654
Ecart type des longueurs de session	4,683
Longeur moyenne des requêtes	1,986
Ecart type des longeurs de requête	1,269

Tableau 1. *Caractéristiques du log de connexion utilisé.*

4. Encodage des liens entre termes

Dans cette étude, nous nous sommes focalisés sur les formulations de requêtes. Dans notre approche, une requête est composée d'un ensemble de mots. Les phénomènes de base de reformulation que nous avons retenus correspondent aux trois actions usuelles de base ci-dessous :

– suppression d'un ou plusieurs mots ;

– ajout d'un ou plusieurs mots ;

– reprise d'un ou plusieurs mots.

Requête	Mots ajoutés	Mots supprimés	Mots identiques
le petit livre rouge	le petit livre rouge		
révolution culturelle	révolution culturelle	le petit livre rouge	
littérature chine populaire	littérature chine populaire	révolution culturelle	
Littérature Mao	Mao	chine populaire	littérature

Figure 4. *Représentation des reformulations de la requête "le petit livre rouge" selon les trois actions de base.*

Requête	Mots ajoutés	Mots supprimés	Mots identiques
linguistique textuelle	linguistique textuelle		
définition linguistique textuelle	définition		linguistique textuelle
concept linguistique textuelle	concept	définition	linguistique textuelle

Figure 5. *Représentation des reformulations de la requête "linguistique textuelle" selon les trois actions de base.*

La reprise de mots consiste à réutiliser le même mot ou une variante proche (généralement issue de la correction typographique par l'utilisateur d'un mot mal orthographié). Ainsi, deux mots sont considérés comme identiques, soit s'ils sont composés exactement de la même série de lettres, soit si les deux séries de lettres diffèrent d'une distance d'édition de Levenshtein inférieure à 3. Notons que les séries numériques ne sont considérées identiques que si leur distance d'édition est de 0 ("*2011*" et "*2012*" sont des mots différents). De la même manière, les mots composés de 3 lettres ou moins ("*de*" et "*le*" sont bien considérés comme des mots différents).

Ainsi, la session de la figure 2 sera représentée comme indiqué dans le tableau 4. Celui de la figure 3 comme indiqué dans le tableau 5.

A partir de cette représentation, nous avons défini trois types de liens pouvant lier les mots d'une requête dans une formulation ou une reformulation. Par ailleurs, nous avons opté pour une représentation sous forme de graphes, dans lesquels les noeuds sont les mots des requêtes et les liens les différentes relations typées que nous avons définies et qui sont présentées et illustrées ci-dessous.

4.1. *Relation de formulation*

Les liens de formulation (notés *QL* pour Query Link) : deux mots sont liés s'ils co-occurrent dans une requête. Par exemple, la figure 6 montre les mots et les liens QL des requêtes issues de la session de la figure 2. Ces liens sont bidirectionnels. Ils sont pondérés : le poids correspond soit à la fréquence de co-occurence dans les requêtes,

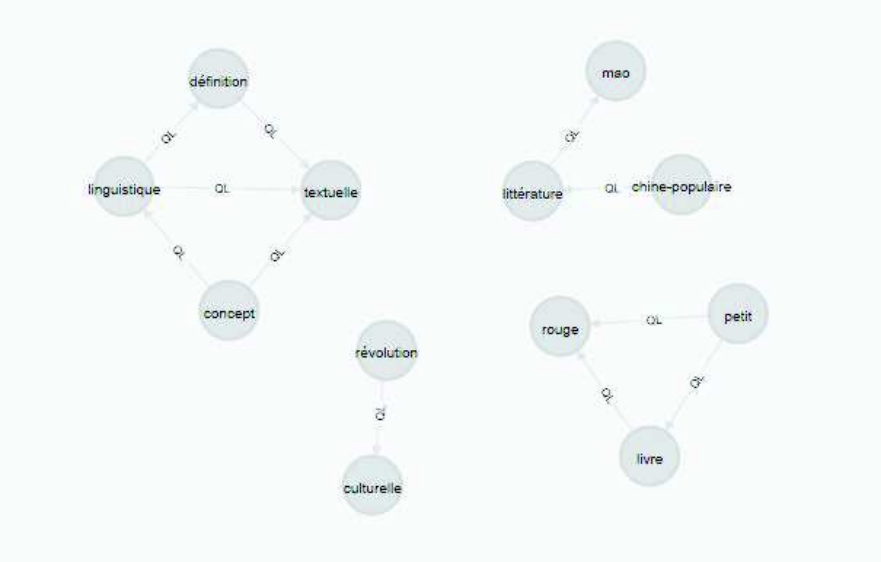

Figure 6. *Mots et relations QL pour la session des figures 2 et 3.*

soit à celle de la fréquence dans les sessions. Dans la figure présentée, il s'agit de la fréquence dans la session (comme une seule session est représentée, le poids est au maximum de 1 entre les mots).

4.2. *Relation de reformulation*

Les liens de reformulation (notés *QR pour Query Reformulation*) : chaque mot de la requête au temps t d'une session est lié à chacun des mots de la requête au temps $t + 1$ de la même session. Les liens sont orientés, de la requête du temps t vers la requête $t + 1$. Ils sont également pondérés, comme précédemment, le poids d'un lien peut être calculé en considérant chaque requête ou chaque session comme contribuant au poids. La figure 7 représente les liens QR pour les deux sessions précédentes.

4.3. *Relation de remplacement*

Les liens de remplacement (notés *RL pour Replacement Link*) : tout mot supprimé de la requête posée au temps t est lié à chacun des mots nouveaux de la requête au temps $t + 1$ de la même session. La figure 8 présente le graphe des relations RL entre les mots pour les deux sessions précédentes. Comme dans le cas des liens QR, les liens RL sont orientés et pondérés.

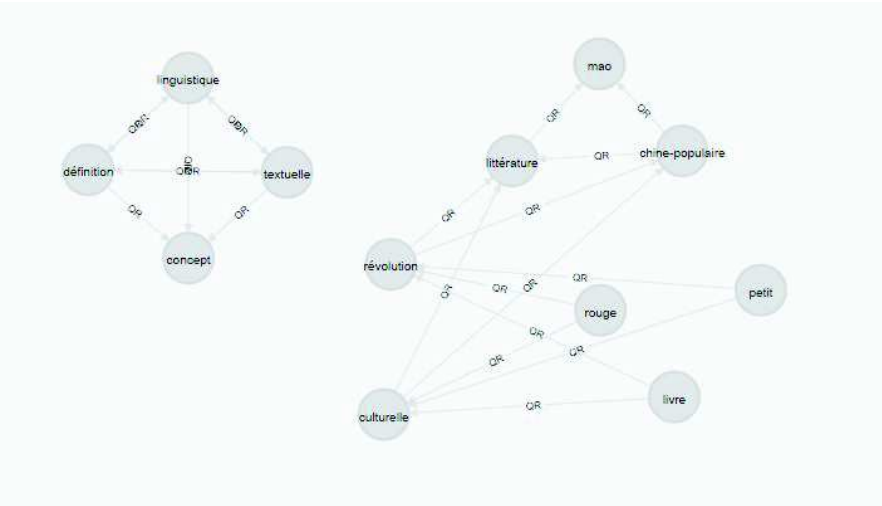

Figure 7. *Mots et relations QR pour les sessions des figure 2 et 3.*

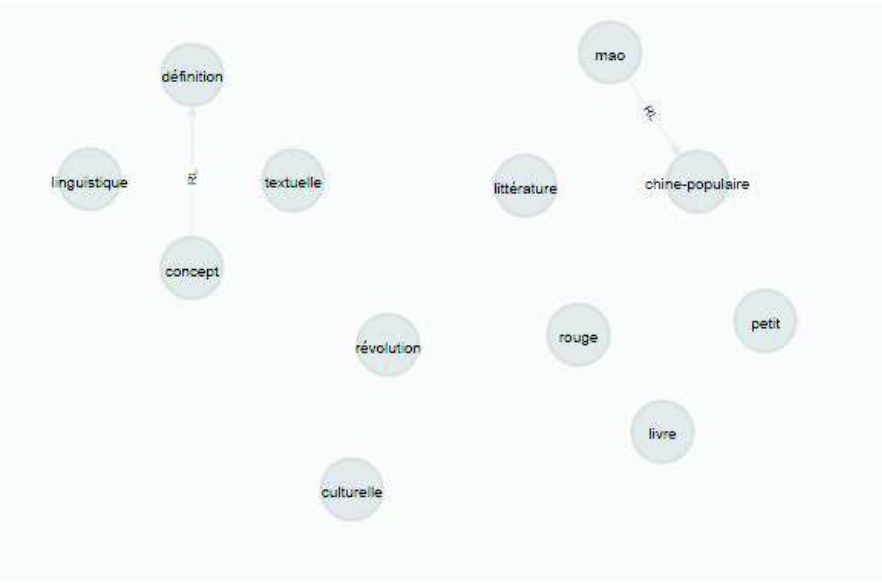

Figure 8. *Mots et relations RL pour la session des figures 2 et 3.*

5. Analyse des formulations et des reformulations de requêtes

Nous nous sommes intéressés à plusieurs aspects concernant l'utilisation des termes dans les requêtes. Nous avons d'abord extrait les termes dans l'objectif de savoir si les mêmes termes étaient liés avec tous les types de liens. Nous nous sommes ensuite intéressés à quelques schémas particuliers.

5.1. *Fréquences des termes*

Les termes les plus fréquents (hors mots outils) sont présentés dans le tableau 2. Sur la partie droite, il s'agit des termes les plus fréquents toutes requêtes confondues alors que la partie gauche compte les sessions différentes dans lesquelles le terme apparait.

Fréquence dans les requêtes	Fréquence dans les sessions
recherche (1637)	recherche (1051)
france (1482)	histoire (602)
politique (1407)	politique (545)
histoire (1396)	france (518)
travail (1255)	tourisme (444)
tourisme (1255)	revue (428)
communication (1137)	travail (402)
développement (1063)	communication (386)
sociologie (972)	développement (376)
afrique (945)	sociologie (370)

Tableau 2. *Mots les plus fréquents dans les requêtes et dans les sessions.*

Naturellement, le nombre de sessions dans lesquelles un terme apparaît est bien inférieur à celui de requêtes (divisé par 2 ou 3). A deux exceptions prêt, *afrique* du côté des requêtes et *revue* du côté des sessions, les termes les plus fréquents sont les mêmes que l'on considère les requêtes ou les sessions.

Le tableau 3 présente les termes qui sont les plus liés avec d'autres en fonction du type de lien (QR et RL). En utilisant le language $Cypher$ associé à la base de données graphe $No4j$ que nous avons utilisée pour l'implantation, pour les liens de type QR, la requête est : "match (n)-[r :QR]-(x) return n,count(n) order by count(n) desc limit 5".

Il n'est pas étonnant de voir que les termes les plus fréquents (tableau 2) se retrouvent ici. Par ailleurs, il n'apparait pas de distinction claire entre les types de lien (hormis une fréquence bien moindre pour les liens de remplacement). Les mêmes termes semblent être indifféremment utilisés dans la requête initiale, dans ses reformulations ou dans les substitutions.

avec un lien QR	avec un lien RL
france (1393)	france (307)
politique (1272)	politique (278)
recherche (1164)	histoire (210)
histoire (1138)	travail (201)
travail (980)	développement (189)

Tableau 3. *Mots les plus utilisés dans les reformulations de requêtes.*

5.2. *Patrons de reformulation*

Nos travaux futurs vont s'orienter vers la recherche de patrons de formulation ou de reformulation. Plusieurs patrons vont retenir notre attention :

– Les paires les plus fréquentes de termes (t1, t2) telles que t2 remplace t1. Ce patron pourrait servir à la reformulation automatique. En effet, si l'on constate que souvent les utilisateurs remplacent le terme t1 par t2, il pourrait être judicieux d'étendre automatiquement et de façon transparente pour l'utilisateur la requête initiale contenant le terme t1 avec le terme t2,

– Les termes qui sont présents dans les requêtes finales. L'idée est de retenir la formulation finale de la requête après plusieurs reformulations. Ce patron est particulièrement intéressant dans le cas où les requêtes de la session n'ont donné lieu à aucune consultation de documents alors que la requête finale a, elle, donné lieu à la consultation d'un ou plusieurs documents. Une analyse linguistique plus fine de ces requêtes pourra également peut être apporter des éléments sur les éléments qui ont conduit au succès de la requête,

– Les termes qui sont inutilement ajoutés : dans une session, il s'agit des termes qui sont ajoutés à une requête puis supprimés dans la formulation suivante. Ces termes sont particulièrement utiles à connaître si la requête associée n'a donné lieu à aucune consultation de document. Il peut s'agit potentiellement de termes ne permettant pas de préciser de façon satisfaisante une requête, voire peut être même des termes qui amènent du bruit documentaire. Une fois identifiés, si ces termes se trouvent dans une requête, ils pourront être automatiquement supprimés par le moteur (ou le moteur pourra leur donner une importance moindre).

6. Conclusions et perspectives

Dans cet article, nous avons présenté la mise en place d'une base de données graphe à partir d'un log de connexion contenant les requêtes des utilisateurs ainsi que les sessions dans lesquelles elles apparaissent. Nous avons retenu trois types de relations entre les noeuds qui permettent de représenter les liens explicites ou impli-

cites entre les termes que les utilisateurs ont fait au travers des requêtes formulées sur le moteur de recherche.

Une première analyse des fréquences des termes a pu être réalisée. Nous avons également pensé à différents patrons qu'il pourrait être intéressant d'extraire à la partir de la base de données graphe ; ces patrons pourront nous renseigner sur les schémas utilisés par les utilisateurs dans leurs formulations et reformulations de requêtes.

Dans un travail futur, la base de données sera étendue en intégrant des noeuds documents permettant de savoir si les termes ont conduit à la sélection d'un ou plusieurs documents. Cette information complémentaire pourra nous permettre d'extraire des patrons plus riches sur l'importance des termes dans les requêtes ou leurs reformulations.

Un autre prolongement de ce travail concerne l'évaluation de l'impact des patrons de reformulation sur la pertinences des résultats de recherche. Nous pourrions ainsi étendre nos travaux basés sur la structure des textes pour la reformulation de requêtes (Ermakova *et al.*, 2016) et la desambiguisation de termes dans les requêtes (Chifu *et al.*, 2015).

Remerciements

Ces travaux ont été rendus possible grâce au soutien de l'Ambassade de France en Inde pour la visite de Sagun Pai à l'Institut de Recherche en Informatique de Toulouse en 2014 dans le cadre des bourses Charpak. Ils ont également bénéficié du soutien de l'Agence Nationale pour la Recherche au projet au projet CAAS Contextual Analysis and Adaptive Search, ANR-10-CONT-001.

7. Bibliographie

Anick P., « Using Terminological Feedback for Web Search Refinement : A Log-based Study », *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, ACM, p. 88-95, 2003.

Bendersky M., Croft W. B., « Modeling higher-order term dependencies in information retrieval using query hypergraphs », *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 941-950, 2012.

Boldi P., Bonchi F., Castillo C., Vigna S., « Query reformulation mining : models, patterns, and applications », *Information retrieval*, vol. 14, n° 3, p. 257-289, 2011.

Border A., « A taxonomy of web search », *SIGIR Forum*, vol. 36, n° 2, p. 3 - 10, 2002.

Chevallet J.-P., « Utilisation des graphes conceptuels pour des systèmes de recherche d'informations orientés vers la précision des réponses », *actes des journées' graphes conceptuels' du PRC-GDR IA,*, p. 35-53, 1994.

Chifu A., Hristea F., Mothe J., Popescu M., « Word Sense Discrimination in Information Retrieval : A Spectral Clustering-based Approach », *Information Processing & Management*, vol. 51, p. 16-31, mars, 2015.

Dousset B., El Haddadi A., Mothe J., « Knowledge discovery from people and semantic networks - Analyzing texts in languages we do not understand », *E-Journal on Digital Enterprise*, vol. 30, p. (en ligne), juin, 2011.

Eickhoff C., Teevan J., White R., Dumais S., « Lessons from the Journey : A Query Log Analysis of Within-session Learning », *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, ACM, p. 223-232, 2014.

Ermakova L., Mothe J., Nikitina E., « Proximity Relevance Model for Query Expansion », *ACM Symposium on Applied Computing (SAC), Pisa, Italy, 04/04/2016-08/04/2016*, ACM, http ://www.acm.org/, p. 1054-1059, april, 2016.

Ghosh S., Viswanath B., Kooti F., Sharma N. K., Korlam G., Benevenuto F., Ganguly N., Gummadi K. P., « Understanding and combating link farming in the twitter social network », *Proceedings of the 21st international conference on World Wide Web*, ACM, p. 61-70, 2012.

Hassan A., Shi X., Craswell N., Ramsey B., « Beyond clicks : query reformulation as a predictor of search satisfaction », *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM, p. 2019-2028, 2013.

He D. G.-A., Harper D., « Combining evidence for automatic web session identification », 2002.

Huang J., Efthimiadis E. N., « Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs », *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, ACM, p. 77-86, 2009.

Ihadjadene M., Chaudiron S., « Quelles analyses de l'usage des moteurs de recherche. Questions méthodologiques », *Questions de communication*, n° 14, p. 17-32, 2008.

Jansen B. J. A. S. C. B. S. K., « Defining a Session on Web Search Engines », *Journal of the American Society for Information Science and Technology*, vol. 58, n° 6, p. 862-871, 2007.

Kompaore N. D., Mothe J., Tanguy L., « Combining indexing methods and query sizes in information retrieval in French », *International Conference on Enterprise Information Systems (ICEIS), barcelone, 12/06/2008-16/06/2008*, ICM, http ://matwbn.icm.edu.pl/, p. 149-154, 2008.

Leva S., Faessel N., « Détection automatique des sessions de recherche par similarité des résultats provenant d'une collection de documents externe (regular paper) », *in* and (ed.), *Rencontres des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Sables D'Olonne, France, 17/06/2013-21/06/2013*, CNRS, http ://www.cnrs.fr, p. 217-230, juin, 2013.

Mizzaro S., Mothe J., « Why do you Think this Query is Difficult ? A User Study on Human Query Prediction (short paper) », *ACM SIGIR Special Interest Group on Information Retrieval (SIGIR), Pisa, Italy, 17/07/2016-21/07/2016*, ACM, http ://www.acm.org/, p. 1073-1076, juillet, 2016.

Mothe J., Tanguy L., « Linguistic Analysis of Users' Queries : towards an adaptive Information Retrieval System », *Signal-Image Technologies and Internet-Based System, 2007. SITIS'07. Third International IEEE Conference on*, IEEE, p. 77-84, 2007.

Mugnier M.-L., Chein M., « Représenter des connaissances et raisonner avec des graphes », *Revue d'intelligence artificielle*, vol. 10, n° 1, p. 7-56, 1996.

Oliner A., Ganapathi A., Xu W., « Advances and Challenges in Log Analysis », *Communication of the ACM*, vol. 55, n° 2, p. 55-61, February, 2012.

Radinsky K., Svore K., Dumais S., Teevan J., Bocharov A., Horvitz E., « Modeling and Predicting Behavioral Dynamics on the Web », *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, ACM, p. 599-608, 2012.

Rieh S. Y., Xie H. I., « Analysis of multiple query reformulations on the web : The interactive information retrieval context », *Information Processing & Management*, vol. 42, n° 3, p. 751 - 768, 2006.

Silverstein C. H. M. M. H. M. M., « Analysis of a Very Large Web Search Engine Query Log », *SIGIR Forum*, vol. 33, n° 1, p. 6-12, 1999.

Tchuente D., Canut C. M.-F., Baptiste-Jessel N., Péninou A., Sèdes F., « Modèle et techniques de dérivation de profils utilisateurs à partir de réseaux sociaux égocentrés. », *INFORSID*, p. 207-222, 2012.

Zenz G., Zhou X., Minack E., Siberski W., Nejdl W., « From keywords to semantic queries—Incremental query construction on the Semantic Web », *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 7, n° 3, p. 166-176, 2009.