HAL
open science

Analyse exploratoire des requêtes d'experts médicaux : cas des campagnes d'évaluation TREC et CLEF (regular paper)

Eya Znaidi, Lynda Tamine, Cécile Chouquet, Chiraz Latiri

To cite this version:

Eya Znaidi, Lynda Tamine, Cécile Chouquet, Chiraz Latiri. Analyse exploratoire des requêtes d'experts médicaux : cas des campagnes d'évaluation TREC et CLEF (regular paper). Symposium sur l'Ingénierie de l'Information Médicale, Lille, 01/07/2013-05/07/2013, Jul 2013, http://univ-lille1.fr, France. pp.(en ligne). hal-01010711

HAL Id: hal-01010711

https://hal.science/hal-01010711

Submitted on 20 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse exploratoire des requêtes d'experts médicaux : cas des campagnes d'évaluation TREC et CLEF

Eya Znaidi*, Lynda Tamine*
Cécile Chouquet**, Chiraz Latiri***

*IRIT - Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse
{znaidi,tamine}@irit.fr,
**IMT - Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse
cecile.chouquet@math.univ-toulouse.fr
***LIPAH - Faculté des Sciences de Tunis, 2092 El Manar Tunis
Chiraz.latiri@gnet.tn

Résumé. Dans ce papier, nous nous intéressons à l'analyse des besoins en information exprimés par des experts médicaux dans l'objectif de les caractériser puis mesurer l'impact de leur structure sur les résultats de recherche. À cet, effet, nous menons une étude exploratoire basée sur des analyses statistiques multidimensionnelles sur des collections de requêtes issues de campagnes d'évaluation internationales standards en l'occurrence TREC[1] et CLEF[2]. Notre étude révèle des variabilités significatives à la fois dans la morphologie des requêtes, que des besoins et des performances, que nous interprétons sur la base des objectifs et spécificités des tâches médicales associées. Les résultats de cette étude ont un impact sur la conception de systèmes de recherche d'information médicaux.

1 Introduction

La recherche d'information (RI) est un domaine qui étudie les modèles et les techniques permettant de sélectionner, à partir de corpus de documents, les informations dites pertinentes pour un utilisateur ayant exprimé un besoin en information à l'aide d'une requête Bayeza et Ribeiro-Neto (1999). Il est communément admis dans ce domaine, que l'expression des requêtes traduisant fidèlement les besoins en information est une tâche difficile aussi bien pour des novices que pour des experts Spink et Jansen (2004); White et Moris (2008). En conséquent, l'analyse et l'élicitation des besoins en information "cachés" derrière les requêtes, est devenu un réel défi dans le domaine. Plus spécifiquement, dans le domaine de la RI médicale, des études larges et exhaustives menées par *Pew Internet and American Life Project* révèlent que 80% des utilisateurs cherchent l'information de santé en ligne et que l'information retournée impacte leurs décisions quant à la prise en charge de leur propre santé ou celle de leurs proches Fox et Jones (2009). En outre, ces études montrent que la population d'utilisateurs est caractérisée par une variabilité significative à la fois sur le volet de l'âge que des niveaux d'expertise ; les besoins en information liés à la santé sont également très variés couvrant des besoins liés à la médication, santé et bien-être, traitements et pathologies.

[1] Text Retrieval Conference
[2] Cross-language Evaluation Forum

Cependant, force est de constater que malgré la diversité des Systèmes de RI (SRI), qu'ils soient généraux ou dédiés tels que PubMed[3], ainsi que la multiplicité de sources d'informations, les utilisateurs demeurent insatisfaits de la pertinence des résultats retournés par ces systèmes Zhang (2010). L'une des causes fondamentales à l'origine de ce constat, est la difficulté de formulation des requêtes de la part des utilisateurs, d'une part, et l'interprétation des besoins en informations induits du côté des SRI, d'autre part. Ceci a encouragé une branche de recherches liées à l'analyse et à l'élicitation des besoins en information médicale. Plus précisément, les investigations de recherche ont porté sur deux volets, le premier est lié à l'analyse du comportement des utilisateurs de SRI médicaux et l'autre, à l'analyse des requêtes exprimées par leurs utilisateurs. Concernant le comportement des utilisateurs, les travaux se sont globalement intéressés à la caractérisation des sessions de recherche, le principe de re-formulation de requêtes, les *clics* de souris, et l'influence du comportement sur la qualité des résultats Richesson et al. (2010). D'autres études se sont focalisées sur les pratiques de recherche comme le jugement de pertinence, les sites visités, les types d'information utilisés et ce, dans le but d'identifier les facteurs qui contribuent au succès ou à l'échec de la recherche Islamaj Dogan et al. (2009). La deuxième lignée de travaux de recherche a porté sur l'analyse de la formulation des requêtes. De nombreux travaux Cartright et al. (2011) ont montré l'impact d'attributs comme la présence de catégories médicales, d'acronymes, spécificité et longueur des requêtes sur les résultats des performances. Par ailleurs d'autres travaux ont montré l'impact du niveau d'expertise des utilisateurs sur la formulation des requêtes ainsi que sur les résultats de recherche White et Moris (2008); Lykke et al. (2012). Ces travaux ont également montré l'importance de l'environnement professionnel pour l'interprétation des besoins en information.

Dans ce papier, nous nous intéressons, à juste titre, à la caractérisation des requêtes des experts médicaux. À la différence des précédents travaux dans le domaine, notre objectif n'est pas d'évaluer l'impact du degré d'expertise sur les résultats de recherche en considérant une population large d'utilisateurs, mais de caractériser les requêtes expertes du point de vue de la formulation, en considérant la tâche médicale associée, qui constitue son contexte. Pour atteindre cet objectif, nous menons des analyses statistiques exploratoires sur des besoins en information d'experts du domaine biomédical, issus des campagnes d'évaluation TREC[4] et CLEF[5] et pour différentes tâches médicales. Nous nous focalisons sur deux principaux volets : le premier volet est d'identifier et d'analyser les corrélations entre attributs de requêtes qui sont la longueur, la spécificité, la clarté, selon le type clinique ou non clinique, en utilisant des mesures appropriées construites selon différentes sources d'évidence. Le deuxième volet porte sur l'étude de l'impact de ces attributs sur les performances de recherche des SRI, liées à ces requêtes.

La suite de cet article est organisée comme suit : la section 2 présente une synthèse des travaux de l'état de l'art sur la RI du point de vue du comportement des utilisateurs lors des sessions de recherche et de la forme de leurs requêtes, puis positionne notre contribution dans ce cadre et annonce les questions de recherche. La section 3 décrit les critères qui caractérisent le besoin en information des experts et présente les collections de données et tâches ayant fait

[3]http ://www.ncbi.nlm.nih.gov/pubmed
[4]Text Retrieval Conference
[5]Cross-language Evaluation Forum

l'objet de notre étude. La méthodologie de l'étude et les analyses sont présentées ainsi que les résultats discutés dans la section 4. Dans la section 5, nous concluons en mettant en évidence l'impact de ces analyses sur la conception des SRI médicaux.

2 Analyse des besoins en information biomédicale : synthèse des travaux et contribution

2.1 Synthèse des travaux

Les travaux de l'état de l'art portant sur l'analyse des besoins en information des utilisateurs de SRI biomédicaux ont été abordés sous l'angle de l'étude des stratégies de recherche d'une part et de l'analyse des requêtes d'autre part.
Concernant l'analyse des stratégies des utilisateurs, de nombreux travaux ont révélé des stratégies typiques des utilisateurs de SRI médicaux dépendantes cependant de leur niveau d'expertise Bhavnani et al. (2003). De manière générale, les travaux ont mis en évidence trois principales facettes du comportement : (1) *approche globale de recherche* Zhang (2010) : les études montrent que la recherche est basée sur un processus itératif essai-erreur caractérisé par des transitions entre recherche et navigation ; (2) *accès aux résultats* Tomes et Latter (2007) : de manière analogue aux autres utilisateurs de SRI, une préférence est clairement exprimée pour la haute précision ; (3) *intentions, buts et motivations* Oh (2012) : les résultats d'études empiriques montrent que la motivation est le principal facteur qui détermine l'échec ou le succès des sessions de recherche, plus particulièrement dans un cadre de recherche sociale. Plus spécifiquement, concernant les experts médicaux, les auteurs dans John et al. (2002), ont montré que les experts n'ont pas de stratégie optimale de recherche, qu'ils sont démotivés en se doutant de trouver des réponses crédibles à leurs requêtes et enfin, qu'ils ne choisissent pas des ressources fiables. En complément, les auteurs dans Lykke et al. (2012) ont analysé la différence entre les pratiques de recherche d'information initiée par des médecins en considérant le lieu d'émission : dans l'environnement professionnel ou en dehors. Ils ont conclu que les requêtes en milieu professionnel sont très ciblées visant une haute précision alors que les requêtes émises en dehors du milieu professionnel sont exprimées à l'aide d'un vocabulaire généraliste. Concernant les formes des requêtes, trois principales conclusions émergent d'études antérieures :

1. *Les requêtes médicales sont courtes* : plusieurs études Hong et al. (2002); Zeng et al. (2002); Natarajan et al. (2010) ont conclu que les requêtes sont généralement courtes, contenant moins de 3 termes avec un nombre moyen de termes égal à 2. Dans Zeng et al. (2002), les auteurs ont analysé les requêtes de MedlinePlus et les sessions de recherche d'information en santé dans les hôpitaux, et ont trouvé que le nombre de termes des requêtes est compris ente 1 et 3. Les mêmes résultats ont été trouvés dans Hong et al. (2002) qui ont analysé les requêtes de Healthlink sur la base de 377000 requêtes issues des fichiers de transactions (*log*).

2. *Les termes de la requête ne sont pas fortement liés aux vocabulaires médicaux* : des études Yang et al. (2011); McCray et Tse (2003) ont identifié les sujets des requêtes en utilisant des critères linguistiques ; ces études montrent que les utilisateurs n'utilisent pas forcément les terminologies, en revanche, ils utilisent leur propre vocabulaires avec

des fautes typographiques et des abréviations. À titre d'exemple, Yang et al. (2011) ont étudié un moteur de recherche spécialisé dans les dossiers de patients appelé EMERSE. L'étude expérimentale a montré que que 18.9% des requêtes contiennent au moins un acronyme ; une autre étude développée dans McCray et Tse (2003) montre, suite à l'analyse de 4700 requêtes issues de ClinicalTrials.gov et MEDLINEplus, que l'échec des requêtes a été du principalement aux fautes typographiques et à l'utilisation d'abréviations.

3. *Les sujets des requêtes sont peu précis* : de nombreuses études Song et al. (2010); Liu et Huang (2011) ont montré que les sujets des requêtes sont vagues ; les auteurs ont alors proposé des systèmes de recommandation de requêtes en s'appuyant sur le résultat qui indique que les requêtes contenant des termes les mieux corrélés aux catégories médicales retournent plus de documents pertinents.

2.2 Aperçu de notre contribution

Les études présentées dans les travaux antérieurs concernent généralement des populations larges et/ou ciblées dans des scénarios de recherche, qui sont cependant non reproductibles. Dans ce papier, nous nous intéressons à l'analyse des requêtes d'experts médicaux, établies dans des scénarios de recherche reproductibles puisqu'elles sont issues de campagnes d'évaluation standards dans le domaine de la RI. De plus, notre analyse est exploratoire et adossée à des tâches biomédicales bien spécifiées dans le cadre de ces mêmes campagnes. À notre connaissance aucune étude de requêtes médicales capitalisées lors des différentes campagnes d'évaluation TREC et CLEF n'a été menée à ce jour. Lors de nos travaux antérieurs Znaidi et al. (2013), nous avons étudié les attributs quantitatifs (longueur, clarté et spécificité) pour caractériser les requêtes selon les tâches et selon la classification des requêtes cliniques et non cliniques. Nous avons effectué des analyses descriptives simples accompagnées de graphiques, des analyses de corrélations deux-à-deux par le coefficient de corrélation de Spearman. Nous avons conclu que la formulation des requêtes est différente selon la tâche de recherche en terme de nombre de termes et concepts utilisés. De plus, la spécificité de la requête des experts est influencé par la nature de la tâche indiquant d'une part l'utilisation d'acronymes et d'abréviation qui sont peu distribués dans la collection MEDLINE, d'une autre part, l'usage significatif par les experts de leurs connaissances du domaine. Par ailleurs, nos analyses ont montré que rechercher les informations sur les gènes et protéines favorise l'expression des requêtes clairement formulées alors que la description des cas de patients sont traduites de façon plus ambiguë par les experts ce qui laisse supposer que l'appariement de cas pathologiques est une tâche experte, qui ne se passe pas forcément par un fort partage entre la description du cas en cours et le cas sélectionné. Nous complétons cette précédente étude en adressant dans ce papier les questions de recherche suivantes :

- ***Quel est le degré de corrélation entre les attributs caractéristiques des requêtes en considérant les tâches médicales ?*** : pour répondre à cette question, nous formalisons tout d'abord un ensemble d'attributs de requête : longueur, spécificité, clarté et type (clinique *vs.* non clinique). Nous menons ensuite une analyse de corrélation multidimensionnelle entre attributs à l'aide d'une Analyse aux Composantes Principales (ACP), toutes tâches confondues.

- ***Quel est l'impact des attributs de requêtes sur les performances de recherche ?*** : pour répondre à cette question, nous nous appuyons sur les performances des requêtes en termes de mesures de précision, d'une part pour décrire chaque tâche médicale selon son niveau de performances et analyser les éventuelles corrélations entre les mesures de performance, et d'autre part pour expliquer les performances de recherche en fonction des attributs via une analyse de covariance multivariée (MANCOVA).

3 Analyse des requêtes médicales des experts issues des campagnes TREC et CLEF

3.1 Formalisation des attributs des requêtes

Dans notre étude, nous considérons un cadre de recherche d'information où un expert du domaine soumet une requête Q à une collection de documents cible C. Nous proposons quatre attributs qui caractérisent les requêtes : 1) la longueur, 2) la spécificité, 3) la clarté et 4) la catégorie de la requête. Nous proposons une formalisation de ces quatre attributs, et nous justifions leur utilisation.

1. *Longueur de la requête* : nous retenons deux facettes de la longueur : (1) longueur en nombre de mots significatifs, **LgT(Q)**, et (2) longueur en nombre de termes référant aux entrées préférées des concepts de la terminologie MeSH[6], **LgC(Q)**. Notre choix de la terminologie MeSH est justifié par sa large utilisation dans le domaine médical. Pour cela, nous exploitons notre technique d'extraction de concepts MeSH Dinh et Tamine (2011a,b).

2. *Spécificité de la requête* : la spécificité est considérée comme un critère important pour identifier les descripteurs Jones (1972). Dans notre étude, nous nous intéressons à deux facettes :

   - ***Spécificité terme-document DSpe(Q)*** : exprime la singularité des termes de la requête dans la collection. L'hypothèse de base est que plus les termes de la requête sont rares dans les documents, plus la requête est spécifique. Elle est calculée :

   $$DSpe(Q) = \frac{1}{LgT(Q)} \sum_{t_i \in termes(Q)} -\log(\frac{n_i}{N}) \quad (1)$$

   où $LgT(Q)$ est le nombre de termes de la requête, $termes(Q)$ est l'ensemble de termes de la requête, n_i est le nombre de documents contenant le terme t_i, N est le nombre total de documents de la collection C.

   - ***Spécificité hiérarchique HSpe(Q)*** : basée sur la profondeur du sens des termes de la terminologie MeSH. L'hypothèse de base est qu'un terme fils est plus spécifique que le terme parent. La spécificité hiérarchique est donnée par :

   $$HSpe(Q) = \frac{1}{LgC(Q)} \sum_{c_i \in Concepts(Q)} \frac{\textit{Niveau}(c_i) - 1}{\textit{Niveau max(MeSH)} - 1} \quad (2)$$

[6]MEdical Subject Headings

où $LgC(Q)$ est le nombre de concepts de la requête, $Concepts(Q)$ ensemble de concepts de la requête, $niveau(c_i)$ niveau du concept c_i dans MeSH , $Niveaumax(MeSH)$ est le niveau maximal de la hiérarchie MeSH.

3. *Clarté de la requête* : en général, une requête claire représente un sujet unique alors qu'une requête ambigüe peut évoquer différents sujets qui ne sont pas forcément reliés. Nous proposons deux facettes de clarté :

   - ***Score de clarté basé sur le sujet de la requête SClar(Q)*** : le score de clarté de la requête est calculé par le score de divergence de Kullback-Leiber entre le modèle de langue de la requête et le modèle de langue de la collection, donné par Cronen-Townsend et Croft (2002) :

$$SClar(Q) = \sum_{t \in V} P(t|Q) log_2 \frac{P(t|Q)}{P_{coll}(t)} \tag{3}$$

   où V désigne le vocabulaire de la collection, t un terme, $P_{coll(t)}$ est la fréquence relative du terme t et $P(t|Q)$ est estimée par : $P(t|Q) = \sum_{d \in R} P(w|D)P(D|Q)$. ou d est un document, R est l'ensemble de documents contenant au moins un terme de la requête.

   - ***Score de clarté basé sur la pertinence PClar(Q)*** : une requête est supposée être plus claire si elle contient des termes en commun avec les documents pertinents évalués par les experts. Cette hypothèse est la base des modèles de RI. Donc, PClar(Q) est calculée :

$$PClar(Q) = \frac{1}{|P(Q)|} \sum_{d \in P(Q)} \frac{|termes(Q) \cap |termes(d)|}{LgT(Q)} \tag{4}$$

   où $P(Q)$ ensemble de documents pertinents retournés pour une requête Q évaluée par les experts, $|termes(d)|$ (resp. $|termes(Q)|$) nombre de termes des documents (resp. termes de la requête).

4. *Catégorie de la requête* : nous nous intéressons à deux types de requêtes médicales : cliniques *vs.* non cliniques. Pour cela, nous utilisons le modèle PICO pour la classification des requêtes Boudin et al. (2010) : *P* correspond à la description des patients (sexe, morbidité, race, age etc.), *I* désigne une intervention, *C* correspond à une autre intervention permettant une comparaison ou un contrôle et *O* correspond aux résultats des expériences. Selon cette définition, nous effectuons l'annotation manuelle de toutes les requêtes de test en clinique (C) si la requête contient au moins 3 éléments PICO, non clinique (NC) sinon.

3.2 Tâches médicales et collection de données

Pour notre étude, nous utilisons des données issues de deux campagnes d'évaluation en RI : Text REtrieval Conference (TREC)[7] et Conference and Labs of the Evaluation Forum

[7] http ://trec.nist.gov/

(CLEF)[8]. Nous exploitons les requêtes (notées Nb.Req), les documents (notés Nb.Doc) et les données composées de jugements de pertinence des médecins en respectant les différentes tâches de RI décrites ci-dessous :

- *Tâche TREC Medical (Nb.Req=35, Nb.D=95,701)* : La tâche de recherche consiste à identifier des cohortes pour une évaluation comparative efficace. Les requêtes représentent des ensembles de conditions symptômatiques de patients et les documents représentent des comptes rendus de visites médicales.

- *Tâches TREC Genomics* : c'est une des tâches les plus pérennes et les plus larges de TREC dans le domaine biomédical. Un des principaux objectifs de TRECGenomics est la recherche *adhoc*, où un chercheur en biomédecine qui recherche l'information sur les génomes soumet une requête à un moteur de recherche spécialisé dans la littérature scientifique biomédicale de la base MEDLINE[9]. Les tâches de RI *adhoc* ont évolué au fil des années : recherche des articles contenant les noms de gènes en 2003 *(Nb.Req=50, Nb.D=525,938)*, besoin en information de biologistes avec l'utilisation des acronymes en 2004 *(Nb.Req=50, Nb.D=4,591,008)* et les questions- réponses du domaine biomédical en 2006 *(Nb.Req=28, Nb.D=162,259)*.

- *Tâches de ImageCLEF (Nb.Req=10, Nb.D=55,634)* : les objectifs de la campagne CLEF ne cessent d'évoluer afin de couvrir différentes tâches de recherche d'images. ImageCLEF 2011 comprend trois tâches principales : classification de modalités, recherche d'images et recherche des cas de patients qui répondent potentiellement à un essai clinique. Ces dernières requêtes sont constituées d'une description textuelle de cas de patients, avec les données démographiques des patients, des symptômes, des descriptions d'imagerie médicale et des résultats de tests.

4 Résultats de l'analyse

Dans le travail présenté dans cet article, nous nous concentrons sur l'analyse des requêtes formulées par des experts du domaine médical. Notre analyse statistique se décompose en deux parties. La première a pour objectif l'analyse multidimensionnelle des corrélations entre les attributs. La seconde partie de notre analyse met en avant les mesures de performances et leurs éventuelles corrélations avec les attributs de requêtes. Pour ces deux parties, nous avons effectué des analyses en composantes principales (ACP). Les différences entre tâches ou entre requêtes cliniques et non-cliniques ont été testées par des analyses de variance ou des tests non-paramétriques de Kruskall-Wallis (adaptés aux petits échantillons). Dans un dernier temps, une analyse de covariance multidimensionnelle a permis de modéliser les mesures de performance en fonction des attributs quantitatifs et de la classification clinique ou non-clinique des requêtes.

Pour étudier les corrélations entre les six attributs des requêtes, nous avons réalisé une analyse en composantes principales (ACP) permettant de prendre en compte les corrélations entre attributs d'un point de vue multidimensionnel. L'étude des trois premiers axes principaux

[8] http ://www.clef-initiative.eu//
[9] http ://www.nlm.nih.gov/pubs/factsheets/medline.html

a permis de dégager les tendances principales de chaque collection. La Figure 1 représente la projection des 173 requêtes selon les deux premiers axes et permet de montrer que :

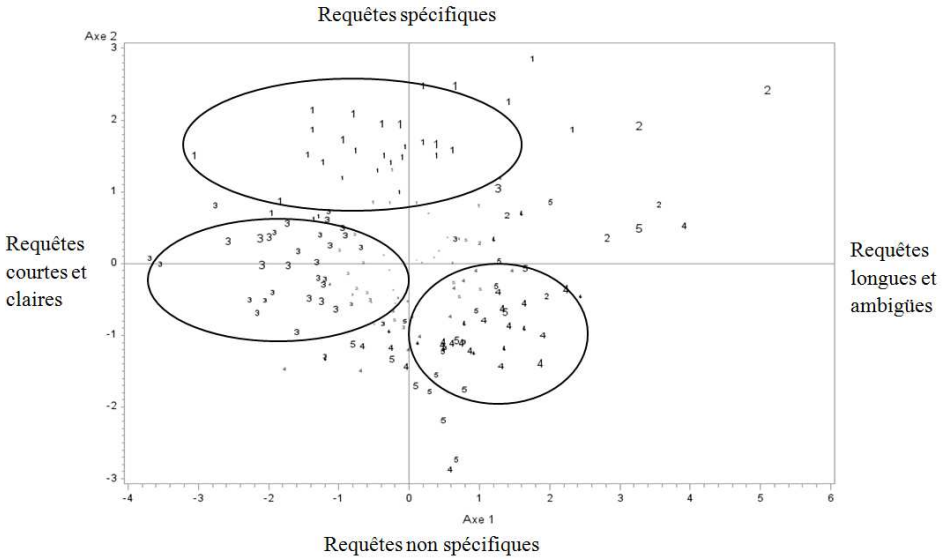

FIG. 1 – *Nuage de points résultats de l'ACP, 1 désigne la collection TRECMed ; 2 désigne la collection ImageCLEF ; 3 désigne la collection TRECGenom03 ; 4 désigne la collection TRECGenom04 et 5 désigne la collection TRECGenom06*

- la collection TRECGenomics2003, spécialisée dans le recherche sur les gènes et protéines, est caractérisée par des requêtes courtes en nombre de termes et de concepts, et claires (selon le score de clarté basée sur la pertinence).
- les requêtes de la collection TRECMed2011 regroupant des compte-rendus médicaux de patients sont spécifiques, claires et écrites avec un nombre assez important de concepts, mais avec un nombre réduit de termes non identifiés comme des concepts. Elles s'opposent aux requêtes des collections TRECGenomics2004 et TRECGenomics2006 dont la spécificité est plus faible.
- la collection ImageCLEF, principalement axée sur la recherche de cas de patients, se distingue des autres collections par un faible nombre de requêtes qui peuvent être longues, voire très longues.

En outre, l'interprétation du troisième axe de cette ACP (non présenté ici) révèle l'indépendance du score de clarté (basée sur le sujet de la requête) vis-à-vis des autres attributs : l'ambiguïté d'une requête ne dépend ni de sa longueur, ni de sa spécificité, et ceci dans les cinq (5) collections étudiées.

4.1 Impact des attributs des requêtes sur les performances de recherche

Cette partie des analyses est complémentaire à l'analyse des facettes des attributs présentée à la section précédente. Les performances de recherche sur l'ensemble des requêtes ont été me-

surées suite à une évaluation avec le système Terrier (http ://terrier.org/), et l'outil d'évaluation standard *trec eval*. Précisément, nous avons généré quatre (4) scores de performance calculés sur la base du modèle BM25 : MAP (*Mean Average Precision*) et les précisions aux rangs 10, 20 et 100 (notées respectivement P@10, P@20 et P@100). Dans un premier temps, une étude comparative des scores a été menée afin de caractériser les performances relatives à chaque tâche et de mettre en évidence d'éventuelles différences entre les tâches. Le Tableau 1 présente les moyennes *m* et écart-types (*sd*) des quatre scores calculés pour chaque tâche, et le résultat du test non-paramétrique de Kruskall-Wallis de comparaison des scores (*p-value*).
Cette analyse révèle que les performances des requêtes issues de la collection TRECGenomics2003 sont caractérisées par des scores moyens significativement plus faibles (inférieures à 0.07) que les quatre autres collections (dont les scores moyens sont compris entre 0.28 et 0.54). Ceci s'explique par la difficulté de la tâche lors de son année de lancement en 2003 : les documents résultats sont jugés effectivement pertinents seulement dans le cas où ils s'apparient avec des Gene Reference Into Function (GeneRIF), mais comme peu d'annotations textuelles GeneRIF étaient disponibles en 2003, les performances des SRI sont de fait sous-estimées Hersh et al. (2008). Soulignons que les scores moyens de ces quatre collections ne sont pas significativement différents, suggérant une performance homogène des requêtes des quatre collections. Par ailleurs, l'analyse des corrélations entre les quatre scores de performance permet de montrer que tous les scores sont très fortement corrélés positivement deux-à-deux ($p - value < 0.0001$), et ceci sur l'ensemble des collections : une requête évaluée comme performante par l'un des scores le sera également par les trois autres mesures.

	Scores de performances : m (sd)			
Tâche/Score	*MAP*	*P@10*	*P@20*	*P@100*
TRECMed	0,32 (0,22)	0,45 (0,34)	0,43 (0,32)	0,28 (0,22)
ImageClef	0,30 (0,24)	0,54 (0,45)	0,48 (0,38)	0,31 (0,27)
TrecGenom03	0,067 (0,130)	0,038 (0,085)	0,045 (0,080)	0,028 (0,047)
TrecGenom04	0,36 (0,25)	0,53 (0,35)	0,47 (0,32)	0,34 (0,26)
TrecGenom06	0,34 (0,23)	0,45 (0,36)	0,11 (0,34)	0,31 (0,27)
p-value	$***$	$***$	$***$	$***$

TAB. 1 – *Scores moyens de performance par collection (et écart-type) avec la p-value du test de comparaison des scores entre collections ($***$: $p - value < 0,001$))*

L'objectif final de cette étude est d'évaluer l'impact des attributs sur les performances des requêtes. Les performances étant mesurées par quatre scores (fortement corrélés), elle constitue une variable multidimensionnelle. Nous avons donc mis en œuvre un modèle d'analyse de covariance multivariée permettant d'expliquer les performances en fonction des six attributs quantitatifs et du type de requêtes (cliniques ou non cliniques). Une démarche de sélection descendante a permis de mettre en évidence les attributs pouvant avoir un impact significatif sur les performances. Nous avons choisi d'illustrer ces résultats sur le score de performance P@100, présentés dans le Tableau 2. Nous avons pu mettre en évidence que, toutes choses égales par ailleurs :

Attributs	**Paramètres estimés (s.e.)**	**p-value**
Type de requête : C vs NC	-0.37 (0.12)	**
Longueur en concepts	0.04 (0.01)	**
Spécificité terme-document		
pour requête non-clinique	-0.41 (0.11)	***
pour requête clinique	0.74 (0.26)	**

TAB. 2 – *Résultats de la modélisation du score de performance $P@100$ en fonction des attributs des requêtes (par une MANCOVA) : estimations des paramètres associés aux attributs significatifs (et erreur standard, s.e.) et p-value (ns : p-value$> 0,05$; $*$: $0,01 <$p-value$< 0,05$; $**$: $0,001 <$p-value$< 0,01$; $***$: p-value$< 0,001$)*

- les requêtes sont d'autant plus performantes qu'elles contiennent un nombre important de concepts (le score P@100 augmente en moyenne de 0.04 point par concept supplémentaire) ;
- les requêtes identifiées comme cliniques sont moins performantes que les requêtes non-cliniques (de 0.4 point en moyenne sur le score P@100).
  Toutefois, cet effet du type de requête est à mettre en relation avec la spécificité terme-document de la requête. En effet, les requêtes cliniques sont d'autant plus performantes qu'elles sont spécifiques. En revanche, des requêtes non-cliniques auront tendance à être moins performantes quand leur spécificité augmente.

Ces résultats montrent globalement que, la recherche de cas pathologiques ou cohortes constitue une tâche plus difficile qu'une recherche *ad hoc* sur des gènes à titre d'exemple : plus précisément, (1) la longueur de la requête favorise la clarification de la requête et par conséquent sa "facilité" ; ceci est à la base même des techniques d'expansion de requêtes qui ont montré leur efficacité en RI ; (2) le type de la requête joue un rôle important conjointement à certains attributs caractéristiques du besoin en information des experts. Plus précisément, les requêtes cliniques, même si elles sont naturellement longues, demeurent difficiles si leur vocabulaire est général, non suffisamment caractéristique du cas pathologique considéré. En effet, les différences, même infimes, entre description du cas à travers la requête et cas retourné à partir de la base de cas, amèneront les experts à un jugement de pertinence négatif ; (3) en revanche, un besoin en information non clinique, ciblant des termes rares dans la collection, est plus difficile à satisfaire.

5 Conclusion

Dans ce travail, nous avons réalisé une étude statistique exploratoire sur les requêtes exprimées par les experts biomédicaux dans le cadre des campagnes d'évaluation TREC et CLEF. Les résultats de notre étude donnent un aperçu sur les spécificités des requêtes d'experts selon les différentes tâches. Trois attributs impactent les résultats de recherche, plus spécifiquement la longueur en termes, le score de clarté basé sur le sujet et la spécificité terme-document en fonction du type clinique ou non de la requête. Ces résultats suggèrent le besoin de contextualiser les modèles de RI médicale à la tâche. Plus précisément, un besoin de clarification et spécification par expansion/reformulation de requête serait appropriée pour des requêtes cliniques comparativement aux requêtes non cliniques. Au delà, en effectuant un croisement avec

les travaux de l'état de l'art, il en ressort un besoin de personnaliser la recherche, selon le niveau d'expertise des utilisateurs. Pour asseoir cette hypothèse, nous envisageons dans un futur proche, de mener une analyse exploratoire des besoins en information des experts *vs.* novices du domaine médical.

Références

Bayeza, R. et B. Ribeiro-Neto (1999). *Modern information retrieval*. Addisson Wesley.

Bhavnani, S. K., R. T. Jacob, J. Nardine, et F. A. Peck (2003). Exploring the distribution of online healthcare information. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pp. 816–817.

Boudin, F., J. Nie, J. C. Bartlett, R. Grad, P. Pluye, et M. Dawes (2010). Combining classifiers for robust pico element detection. *BMC Medical Informatics and Decision Making*, 1–6.

Cartright, M.-A., R. W. White, et E. Horvitz (2011). Intentions and attention in exploratory health search. In *SIGIR '11*, New York, NY, USA, pp. 65–74. ACM.

Cronen-Townsend, S. et W. B. Croft (2002). Quantifying query ambiguity. HLT '02, pp. 104–109.

Dinh, D. et L. Tamine (2011a). Biomedical concept extraction based on combining the content-based and word order similarities. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, New York, NY, USA, pp. 1159–1163. ACM.

Dinh, D. et L. Tamine (2011b). Combining global and local semantic contexts for improving biomedical information retrieval. In *ECIR*, pp. 375–386.

Fox, S. et S. Jones (2009). The social life of health information. Technical report, Pew Internet & American Life Project.

Hersh, W., , et B. Croft (2008). Trec genomics special issue overview. *Information retrieval journal*.

Hong, Y., N. Cruz, G. Marnas, E. Early, et R. Gillis (2002). A query analysis of consumer health information retrieval. In *AMIA*, pp. 791–792.

Islamaj Dogan, R., G. C. Murray, A. Névéol, et Z. Lu (2009). Understanding pubmed user search behavior through log analysis. *Database 2009*.

John, E., J. Osheroff, M. Ebell, M. Chambliss, D. Vinson, . J.J. Stevermer, et E. Pifer (2002). Obstacles to answering doctors' questions about patient care with evidence : Qualitative study. *BMJ 324*(7339), 710.

Jones, S. (1972). A statistical interpretation of term specificity and its application to retrieval. *Journal of documentation 28*(1), 11–20.

Liu, R.-L. et Y.-C. Huang (2011). Medical query generation by term-category correlation. *Inf. Process. Manage. 47*(1), 68–79.

Lykke, M., S. Price, et L. M. L. Delcambre (2012). How doctors search : A study of query behaviour and the impact on search results. *Inf. Process. Manage. 48*(6), 1151–1170.

McCray, A. T. et T. Tse (2003). Understanding search failures in consumer health information systems. *AMIA*, 430–434.

Natarajan, K., D. Stein, S. Jain, et N. Elhadad (2010). An analysis of clinical queries in an electronic health record search utility. *International journal of medical informatics 79*(7), 515–522.

Oh, S. (2012). The characteristics and motivations of health answerers for sharing information, knowledge, and experiences in online environments. *JASIST 63*(3), 543–557.

Richesson, R. L., D. Shereff, C. Spisla, N. Albarracin, D. Konicek, et J. E. Andrews (2010). The use of snomed ct to support retrieval and re-use of question and answer sets for patient registries. *I. J. Functional Informatics and Personalised Medicine 3*(4), 342–365.

Song, M., H. Spallek, D. Polk, T. Schleyer, et T. Wali (2010). How information systems should support the information needs of general dentists in clinical settings : suggestions from a qualitative study. *BMC medical informatics and decision making 10*(7).

Spink, A. et B. Jansen (2004). *Web Search : Public Searching of the Web*. Kluwer Academic Publishers.

Tomes, E. et C. Latter (2007). How consumers search for health information. *Health informatics journal 13*(3), 223–235.

White, R. et D. Moris (2008). How medical expertise influences seb search behaviour. In *SIGIR '08*, pp. 791–792.

Yang, L., Q. Mei, K. Zheng, et D. A. Hanauer (2011). Query log analysis of an electronic health record search engine. *AMIA 2011*, 915–924.

Zeng, Q., S. Kogan, N. Ash, R. A. Greenes, et A. A. Boxwala (2002). Characteristics of consumer terminology for health information retrieval. *Methods of information in medicine 41*(4), 289–298.

Zhang, Y. (2010). Contextualizing consumer health information searching : an analysis of questions in a social q&#38 ;a community. IHI '10, New York, NY, USA, pp. 210–219. ACM.

Znaidi, E., L. Tamine, C. Chouquet, et C. Latiri (2013). Characterizing health-related information needs of domain experts. In *Artificial Intelligence in Medicine*, Volume 7885 of *Lecture Notes in Computer Science*, pp. 48–57. Springer Berlin Heidelberg.

Summary

In this paper, we are interested in the analysis of medical expert queries within the goal of characterizing them and then measure the impact of their structure on the retrieval perfermormances. To achieve this goal, we conduct an exploratory study by means of multidimensional statistical analysis using query collections issued from TREC and CLEF international evaluation compaigns. The results of our study reveal significant variations of the query according to both structure and performances due to the objectives and specificities of the underlying medical tasks. Our study arises implications on the design of medical information retrieval systems.