Répondre à des requêtes cliniques PICO[1]

Eya Znaidi* — **Lynda Tamine*** — **Chiraz Latiri****

* *IRIT, Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, France, 118 Route Narbonne, Toulouse, France*
{znaidi,tamine}@irit.fr
** *Laboratoire LIPAH, Faculté des Sciences de Tunis, Campus Universitaire Tunis El Manar, 1060 Tunis, Tunisie*
{chiraz.latiri}@gnet.tn

RÉSUMÉ. *Dans cet article, nous nous intéressons à l'évaluation de requêtes cliniques exprimées avec les facettes PICO (Population/Problem (P), Intervention (I), Comparaison (C) et Outcome (O)). Nous proposons l'application d'un opérateur d'agrégation prioritaire des scores qui permet : (1) d'agréger les scores de pertinence partiels issus de l'évaluation de représentations sémantiques associées aux sous-requêtes facettes et (2) contextualiser le score d'importance des facettes au document et requête en cours d'évaluation. Les expérimentations menées sur la collection standard CLIREC, comprenant* 423 *requêtes cliniques et plus de* 1.2 *millions de documents PubMed, mettent en évidence l'efficacité de notre approche comparativement aux autres modèles de l'état de l'art.*

ABSTRACT. *In this paper, we address the issue of answering PICO (Patient/Problem, Intervention, Comparison, Outcome) clinical queries. The contributions of this work include (1) a new document ranking model based on a prioritized aggregation operator that computes the global relevance score based on the relevance estimation of the semantic facet sub-queries and (2) leverages the importance of the facets according to the document and query under evaluation. The effectiveness of our clinical search approach is empirically evaluated using a clinical retrieval collection including* 423 *queries and more than* 1.2 *million of medical abstracts from PubMed. The experimental results show that our approach for PICO query answering significantly overpasses state-of-the-art document ranking models.*

MOTS-CLÉS : *RI Médicale, Requêtes Cliniques, Agrégation de scores, Représentation Sémantique*

KEYWORDS: *Medical IR, Clinical Queries, Semantic Query representation, Aggregation Scoring*

1. Population/Problem (P), Intervention (I), Comparaison (C) et Outcome (O)

1. Introduction

De nombreuses études ont clairement montré que la recherche d'information médicale est largement pratiquée aussi bien par des experts que par des novices (Fox et Duggan, 2013 ; Zhang, 2013). Plus spécifiquement, la recherche d'information médicale pratiquée par les experts, est un type de recherche verticale basé sur l'usage de ressources, comme les dossiers médicaux de patients ou la littérature scientifique médicale (eg. MEDLINE[1], Cochrane[2]), en vue de répondre à différents objectifs comme l'aide au diagnostic, l'aide à la prescription médicale ou encore la recherche clinique. Ce type de besoins en information induit en grande partie l'évaluation de requêtes cliniques ayant pour objectif de sélectionner, à partir de la littérature scientifique du domaine, des procédés cliniques et/ou des preuves d'études systématiques (Yang *et al.*, 2011). Plus globalement, les tâches de recherche d'information cliniques expertes, sont généralement conduites par les cliniciens dans le cadre de la médecine basée sur les faits connue sous l'acronyme EBM (*Evidence-Based Medicine*) (Sackett *et al.*, 1996).

Cette dernière consiste à utiliser de manière rigoureuse, explicite et judicieuse, les preuves scientifiques les plus récentes et plus pertinentes lors de la prise de décision concernant les soins à prodiguer à chaque patient. Sa pratique implique que l'on conjugue l'expertise clinique individuelle avec les meilleures preuves cliniques externes obtenues actuellement par la recherche systématique (Sackett *et al.*, 1996). La recherche de ces meilleures preuves d'études cliniques, à partir de la littérature scientifique, est à juste titre, l'objet de notre travail présenté dans cet article. Cela suppose d'abord la formulation d'un besoin en informations cliniques. Selon l'approche EBM, un moyen qui a été proposé aux experts en vue de clarifier leur besoin est de structurer leur requête selon la forme PICO, à savoir : Patient/Problem (P), Intervention (I), Comparison (C) et Outcome (O), appelés les éléments ou facettes PICO (Schardt *et al.*, 2007). A titre d'exemple, la formulation PICO de la question clinique Q : "*In people with recurrent aggression having any antiepileptic drug in any dosage, what is length of time of placebo for observer reported aggression ?"* est comme suit : $P \Rightarrow$ *"people with recurrent aggression"*, $I \Rightarrow$ *"any antiepileptic drug in any dosage"*, $C \Rightarrow$ *"length of time of placebo"*, $O \Rightarrow$ *"reported aggression"*.

En plus du verrou largement reconnu lié à l'ambiguité des expressions médicales et acronymes (Trieschnigg, 2010), l'évaluation de ce type de requêtes pose deux difficultés supplémentaires et non triviales à résoudre : (1) considérer la structure de la requête en facettes dans le processus d'appariement avec des documents qui ne sont pas ainsi structurés et (2) considérer leur importance relative dans l'estimation du score de pertinence. À notre connaissance, c'est un champ d'investigation peu exploré ; on recense en effet peu de travaux qui ont abordé spécifiquement le problème d'évaluation des questions cliniques PICO (Boudin *et al.*, 2010c ; Boudin *et al.*, 2010b ; Demner-Fushman et Lin, 2007). Dans l'ensemble des travaux précédents, une étape prélimi-

1. https ://www.nlm.nih.gov/bsd/pmresources.html accessible à l'aide PubMed
2. ://www.ncbi.nlm.nih.gov/pubmed, Cochrane http ://www.cochranelibrary.com/

naire à la recherche, est la détection des facettes PICO dans les documents. Au niveau de l'appariement requête-document, les auteurs dans (Demner-Fushman et Lin, 2007), se sont basés sur une approche d'appariement sémantique entre les types sémantiques extraits d'*UMLS* de chacune des facettes détectées dans les requêtes et les documents. Cependant, cette méthode ne prend pas en compte l'importance des facettes PICO dans le texte, lors du calcul des scores de pertinence. Boudin et al. (Boudin *et al.*, 2010c) ont proposé un modèle d'appariement qui considère l'importance de chaque facette PICO pour calculer le score de pertinence des documents. Toutefois, les scores d'importance sont calculés d'une manière statique, sur la base de la distribution des mots appartenant à chaque facette sur l'ensemble de la collection de documents. Dans cet article, nous proposons un modèle d'appariement requête PICO-document, qui à la différence des précédents travaux : (1) ne requiert pas l'identification préalable des facettes P, I, C et O dans les documents, (2) utilise un opérateur d'agrégation prioritaire (Pereira *et al.*, 2010) dans le calcul des scores d'appariement requête-document en personnalisant les poids de chaque facette selon la requête et le document en cours d'évaluation. Comme dans (Demner-Fushman et Lin, 2007), nous privilégions une représentation sémantique des requêtes et documents, qui est cependant basée sur la génération de graphes sémantiques par facette, développé dans notre précédente contribution (Znaidi *et al.*, 2015).

L'article est organisé comme suit. La section 2 présente une synthèse des travaux portant sur l'évaluation de requêtes PICO. La section 3 donne un large aperçu du processus d'évaluation de requête, annonce nos hypothèses de recherche puis détaille le modèle d'agrégation de scores de pertinence pour l'évaluation de requêtes PICO. Dans la section 4, nous présentons le cadre expérimental puis détaillons et analysons les résultats. Enfin, la section 5 conclut l'article et annonce les pistes de travaux futurs.

2. Evaluation de requêtes cliniques PICO

De nombreux précédents travaux ont montré que les requêtes médicales sont particulièrement complexes (Natarajan *et al.*, 2010 ; Suominen *et al.*, 2013). L'un des facteurs de complexité les plus abordés est incontestablement celui du fossé sémantique entre besoins en information (experts ou novices) et documents. Les solutions ont porté essentiellement sur l'utilisation de ressources sémantiques comme MeSH ou UMLS pour l'enrichissement des requêtes et/ou documents (Stokes *et al.*, 2009 ; Dinh et Tamine, 2012) ou alors dans le modèle d'appariement requête-document (Trieschnigg, 2010 ; Mao *et al.*, 2015). Parmi les requêtes médicales, les requêtes cliniques sont considérées, sous l'angle du système de recherche d'information, comme particulièrement difficiles car elles sont de nature exploratoire et l'évaluation de la pertinence des réponses candidates requiert des facettes d'informations contextuels difficiles à identifier (Francke *et al.*, 2008 ; Natarajan *et al.*, 2010). Sous l'angle de l'expert et du système de recherche d'information, une façon de préciser le besoin en information induit par une requête clinique est de la structurer en facettes PICO. Les travaux sur l'évaluation automatique des requêtes PICO sont peu abondants et se scindent

en deux volets. Dans la première catégorie de travaux (Boudin *et al.*, 2010a ; Zhao *et al.*, 2010), le problème principal adressé par les auteurs est la détection des facettes PICO, comme une étape en amont à la sélection de documents pertinents. La plupart des approches sont basées sur des techniques d'apprentissage supervisé afin d'identifier les éléments PICO à partir du texte. Par exemple, dans (Boudin *et al.*, 2010a), le processus d'identification des éléments PICO à partir du document a été conduit selon deux étapes : une première étape pour la segmentation du texte des documents en plusieurs phrases, puis dans une seconde étape, chaque phrase est transformée en un vecteur de propriétés utilisant les caractéristiques statistiques et linguistiques pour désigner les facettes P, IC et O. Les expérimentations sur un ensemble de 260000 résumés de *PubMed* ont montré que la combinaison linéaire de plusieurs classifieurs est l'approche la plus efficace pour la détection des éléments PICO.

Plus proche de notre contribution, la deuxième catégorie de travaux (Boudin *et al.*, 2010c ; Boudin *et al.*, 2010b ; Demner-Fushman et Lin, 2007) concerne la définition de modèles de recherche d'information qui exploitent les facettes PICO pour calculer les scores de pertinence des documents. Pour atteindre cet objectif, Boudin et al. (Boudin *et al.*, 2010c ; Boudin *et al.*, 2010b) ont proposé une extension de la version basique du modèle de langue (Song et Croft, 1999). Les auteurs ont modifié le modèle de pondération basé sur les termes des documents en tenant compte de la distribution des éléments PICO dans les différents passages de documents ainsi que la distribution des termes dans les différentes parties PICO. L'évaluation expérimentale conduite sur une collection de 1.5 millions de documents et 423 requêtes a montré que le modèle proposé a permis une amélioration de 28% de la *MAP*[3] sur l'ensemble des modèles de référence. Demner-Fushman et Lin (Demner-Fushman et Lin, 2007) ont également proposé un modèle unifié pour détecter et utiliser les éléments PICO dans une fonction de calcul de pertinence des documents S_{EBM}. Cette dernière est basée sur une combinaison linéaire des scores de pertinence partiels des documents, considérant trois éléments de l'EBM, à savoir, la structure PICO (S_{PICO}), la crédibilité de la preuve médicale (S_{SoE}) et le type de la tâche (S_{task}). Par exemple, le score S_{PICO} est basé sur une combinaison linéaire des scores des facettes P, I, C et O en prenant en compte l'appariement des mots entre le document et la facette de la question. Les expérimentations sur 24 questions cliniques ont montré que cette approche dépasse, en termes de performance, la recherche classique dans *PubMed*.

3. Aperçu général de l'approche

3.1. *Hypothèses de recherche*

Soit une requête clinique Q avec les annotations PICO associées donnant lieu aux sous-requêtes Q_P, Q_{IC} et Q_O, manuellement ou automatiquement identifiées. Comme dans de précédents travaux (Boudin *et al.*, 2010c), nous considérons les fa-

3. Mean Average Precision

cettes I et C de façon regroupée comme elles sont associées au même type sémantique. Notre objectif est de sélectionner les documents d qui sont pertinents au besoin en information expert véhiculé derrière la requête Q et ce, en accord avec les facettes P, I/C et O. On s'appuie pour cela sur une représentation sémantique de la requête, issue d'un processus de génération de graphes sémantiques exploitant la ressource MeSH (Znaidi *et al.*, 2015). Nous proposons d'appliquer un opérateur d'agrégation prioritaire (Da Costa Pereira *et al.*, 2009) afin de calculer le score global de pertinence du document d par combinaison de ses scores partiels d'appariement avec les sous-requêtes facettes Q_P Q_{IC} et Q_O. Pour cela, nous nous appuyons sur les deux hypothèses suivantes :

– **H1.** Un document est d'autant plus pertinent qu'il s'apparie avec plus de facettes de la requête (Boudin *et al.*, 2010c ; Demner-Fushman et Lin, 2007).

– **H2.** En phase d'évaluation de la pertinence des résultats, l'expert médical n'accorde pas la même importance à l'adéquation des facettes. La facette I/C est plus importante que la facette P qui, à son tour, est plus importante que la facette O (Weifield et Finkelstein, 2005 ; Boudin *et al.*, 2010c).

3.2. *Processus de traitement de requêtes PICO*

Comme le montre la Figure 1, notre approche est basée sur quatre (4) étapes principales : (1) un appariement préliminaire document-requête basé sur les mots, qui retourne une liste initiale de documents pertinents candidats, (2) représentation sémantique des requêtes, qui retourne les graphes sémantiques conceptuels associés à chaque facette de la requête, (3) un appariement requête-document qui permet de sélectionner les concepts les mieux pondérés de chaque facette du graphe, et (4) calcul de pertinence des documents basé sur un opérateur d'agrégation prioritaire (Da Costa Pereira *et al.*, 2009). Une description détaillée des étapes (2) et (3) est donnée dans (Znaidi *et al.*, 2015). Le principe général de l'algorithme sous-jacent est le suivant :

– *Etape 2* : A partir des sous-requêtes facettes Q_P Q_{IC} et Q_O, un algorithme génère des graphes sémantiques de requêtes Q_P^c, Q_{IC}^c, Q_O^c, en effectuant une extraction de concepts $MeSH$ (Znaidi *et al.*, 2015) pour chacune des facettes en remontant de proche en proche des concepts extraits à partir des mots de la requête jusqu'à atteindre le plus proche concept commun de la hiérarchie de MeSH.

– *Etape 3* : Pour chaque arbre sous-requête associé à une facette Q_P^c, Q_{IC}^c, Q_O^c et chaque document d résultat de l'étape préliminaire (1), identifier les N_c meilleurs concepts associés en appliquant un algorithme de propagation de scores d'appariement concept-document par accumulation des scores des concepts associés au document d jusqu'aux concepts feuilles de MeSH.

Une illustration de la représentation sémantique issues des étapes (2) et (3), appliquée à la requête Q présentée en exemple "*In people with recurrent aggression having any antiepileptic drug in any dosage, what is length of time of placebo for observer reported aggression ?"* est donnée, respectivement, dans les Figure 2 et Figure 3.

Dans le présent article, notre principale contribution porte sur le calcul du score de pertinence des documents en réponse à une requête clinique, correspondant à l'étape (4).

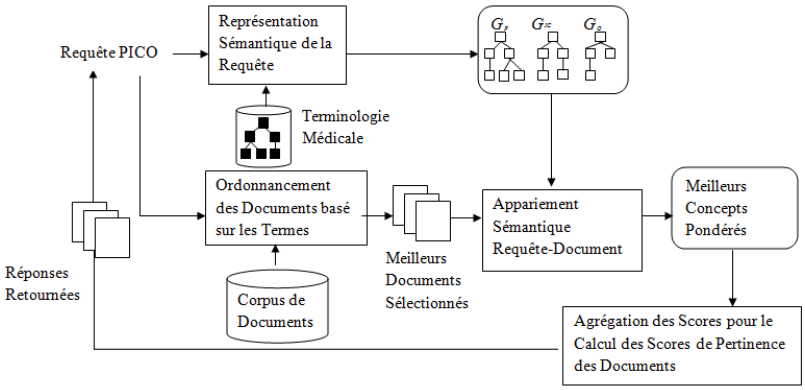

Figure 1 – Architecture du modèle pour le traitement des questions PICO.

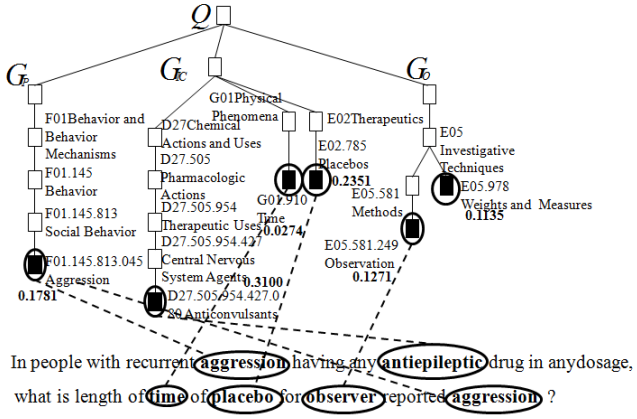

Figure 2 – Exemple de graphe sémantique d'une requête PICO

3.3. *Modèle d'agrégation pour l'estimation de la pertinence en réponse à des requêtes PICO*

Dans l'étape (4), représentée dans la Figure 1., nous disposons d'une représentation sémantique de chacune des requêtes facettes Q^c_P, Q^c_{IC}, Q^c_O à l'aide des N_c concepts pertinents issus de l'étape (2) et (3) (résumées dans la section précédente). Nous considérons que chaque facette P, I/C et O constitue un critère de pertinence et proposons une fonction d'agrégation F qui calcule, pour chaque document d, un score global de pertinence comme suit :

$$RSV_{PICO}(Q,d) = F(RSV_P(Q_P,d), RSV_{IC}(Q_{IC},d), RSV_O(Q_O,d)) \quad [1]$$

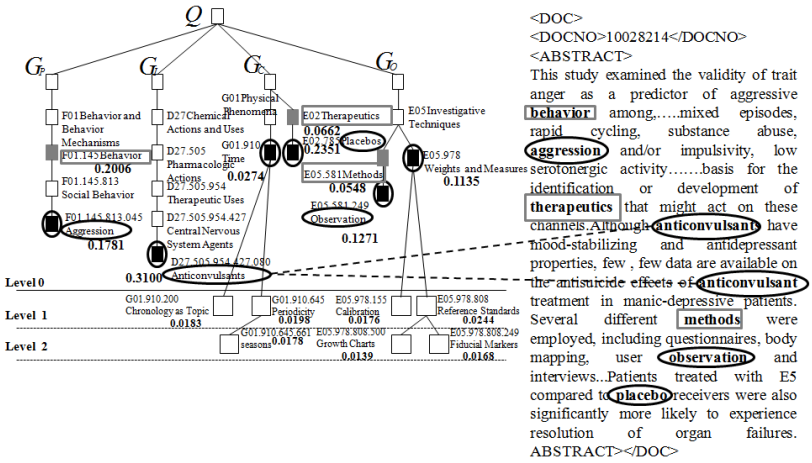

Figure 3 – Illustration de l'appariement requête-document

où F représente la fonction d'agrégation et P, IC et O sont les critères de pertinence PICO. Nous calculons pour chacun de ces critères un score d'importance qui reflète son poids ou contribution dans l'estimation de la pertinence du document.
Nous soutenons l'idée que les scores d'importance obéissent globalement à la hiérarchie posée par l'hypothèse H2 (Section 3.1) et confirmée par les travaux de l'état de l'art (Weifield et Finkelstein, 2005 ; Boudin *et al.*, 2010c) mais ne sont cependant pas fixes pour toutes les requêtes et documents. En effet, nous jugeons opportun de contextualiser les scores d'importance de ces facettes aux requêtes et documents en cours d'évaluation, à plus forte raison, si elles ne sont pas explicitement identifiées dans les documents. En clair, plus le document est supposé pertinent pour la facette la plus importante, moins il convient de considérer le score de son appariement avec la facette la moins importante. Pour répondre à cette intuition, nous proposons l'utilisation d'un opérateur prioritaire d'agrégation de scores (Da Costa Pereira *et al.*, 2009), qui permet de calculer un score global en tenant compte des scores partiels obtenus pour chaque critère ainsi que d'un ordonnancement dans l'importance de ces critères. Le poids de chaque critère est calculé en fonction des poids ainsi que le degré de satisfaction du critère le plus important. L'intuition sous-jacente est la suivante : plus le degré de satisfaction du critère le plus important est élevé, moins le degré de satisfaction du critère le moins important influencera le score global de pertinence.

En accord avec l' hypothèse H1 (Section 3.1), on calcule le score PICO comme suit :

$$RSV^4{}_{PICO}(Q,d) = \lambda_P * RSV(Q_P,d) + \lambda_{IC} * RSV(Q_{IC},d) + \lambda_O * RSV(Q_O,d) \quad [2]$$

4. Relevance Status Value

où $x \in \{P, IC, O\}$ et $RSV_x(Q, d)$ est calculé comme suit :

$$RSV_x(Q,d) = \frac{\sum_{C \in Q_x^C} SIM(C,d)}{\sum_{C \in \cup x Q_x^C} SIM(C,d)} \quad [3]$$

où :

– Q_x^C : est l'ensemble N_c de concepts pondérés, associés à chaque facette de la requête Q_x, résultat de l'étape 3.

– $SIM(C, d)$: est le degré de similarité entre les vecteurs TF/IDF du document d et concept C représenté par ses entrées préférées dans la terminologie MeSH.

Les poids d'importance des éléments PICO sont calculés selon le principe de l'opérateur d'agrégation des scores (Da Costa Pereira *et al.*, 2009), comme suit :

1) Hiérarchisation des poids d'importance des facettes en tenant compte de l'hypothèse H2 :

$$\lambda_P, \lambda_{IC}, \lambda_O \in [0..1], \quad where \quad \lambda_{IC} > \lambda_P > \lambda_O \quad and \quad \lambda_{IC} = 1 \quad [4]$$

2) Calcul de scores contextualisés de l'importance des facettes :

$$\begin{aligned} \lambda_P &= \lambda_{IC} * RSV(Q_{IC}^C, d) \\ \lambda_O &= \lambda_P * RSV(Q_P^C, d) \end{aligned} \quad [5]$$

Pour prendre en compte la pertinence du document en se basant sur les mots et sur les concepts comme recommandé dans (Stokes *et al.*, 2009), le score de pertinence du document d par rapport à la requête Q est la combinaison des scores de pertinence basés sur les concepts ($Score_{PICO}(Q, d)$) et le score de pertinence basé sur les mots ($Score_w(Q, d)$). Le score global de pertinence est calculé comme suit :

$$RSV(Q,d) = \alpha * RSV_{PICO}(Q,d) + (1-\alpha) * RSV_w(Q,d) \quad [6]$$

où $\alpha \in [0..1]$ est un paramètre fixé de façon empirique.

4. Evaluation expérimentale

Les objectifs de l'évaluation expérimentale sont : (1) mesurer l'efficacité du modèle d'évaluation des requêtes PICO en analysant l'effet de chacun des éléments contributifs (représentation conceptuelle, principe de combinaison des facettes, principe de pondération des facettes) ; (2) analyser la robustesse du modèle en identifiant les raisons possibles d'échec vs. succès des requêtes.

Nombre de documents	1.212.040 résumés PubMed
Longueur moyenne de document	246 termes
Nombre de requêtes	423
Nombre moyen de termes de la requête	4.3 termes
Longueur moyenne de la requêtes (PICO)	18.7 termes
Nombre moyen de doc pertinents	19

Tableau 1 – Statistiques de la collection de test CLIREC

4.1. *Cadre d'évaluation*

4.1.1. *Collection de données*

Nous utilisons la collection de données CLIREC(*CLinical Information Retrieval Evaluation Collection*), construite dans le but d'évaluer la recherche d'information clinique (Boudin *et al.*, 2010c). Pour atteindre cet objectif, les auteurs ont construit la collection de test d'une manière semi-automatique à partir d'un ensemble de revues systématiques de la ressource *Cochrane*. Plus spécifiquement, les auteurs ont demandé à un groupe d'experts de générer les requêtes qui correspondent aux questions cliniques fournies par un sous-ensemble du répertoire *Cochrane*. Chaque question est manuellement annotée avec les parties P (Patient/Problème), I (Intervention), C (Comparaison) et O (Outcome, résultats). Pour constituer la vérité terrain, les auteurs ont tout d'abord extrait, pour chaque question clinique, les citations des documents associés à la section référence attachée à chaque revue ; cette section énonce en effet l'ensemble des études pertinentes qui traitent la question clinique considérée. Ensuite, une liste de documents pertinents (articles de journaux) associés à ces citations sont extraits de *PubMed* pour chaque question clinique.
Quelques caractéristiques et statistiques de la collection sont représentés dans le Tableau 1. Nous avons utilisé *MeSH*, qui est la terminologie la plus utilisée pour indexer les citations biomédicales (Stokes *et al.*, 2009). Chaque nœud de la terminologie représente un concept qui fait référence à une entrée préférée dans la terminologie.

4.1.2. *Protocole expérimental*

Nous avons adopté un protocole de validation croisée à base de 10 essais (*10-fold-cross validation*) pour le paramétrage de la fonction d'appariement puis le test de son efficacité. Nous avons calculé l'efficacité moyenne de la performance de recherche sur l'ensemble des 10 sous-collections d'essais construits en amont, après avoir identifié les valeurs optimales qui maximisent la MAP. Pour la mesure de l'efficacité de la recherche, nous utilisons les mesures de la précision moyenne (MAP), et $P@X$ ($X = 5, 10$) avec l'outil TREC-eval[5].

5. http//trec.nist.gov/trec eval

4.1.3. *Modèles de référence*

Nous avons comparé notre modèle de recherche d'information PICO, notée PSM, aux modèles de référence suivants, considérés avec leurs paramètres optimaux (optimisation de la MAP) :

– Deux modèles de recherche d'information de l'état de l'art : (1) le modèle probabiliste Okapi ($BM25$) avec es valeurs recommandées : $k1 = 1.2$, $k3 = 7$ et $b = 0.75$ (Robertson et Sparck Jones, 1988) et (2) le modèle de langue (LM) avec le lissage de Dirichlet avec le paramètre $\mu = 1000$ comme recommandé dans (Song et Croft, 1999).

– Une approche de reformulation sémantique de requêtes notée CQE, où nous avons étendu les requêtes initiales PICO Q avec les entrées préférées des meilleurs concepts MeSH retournés par l'étape 3 du processus d'extraction des concepts pertinents (Section 3.2). Nous avons utilisé le nombre de concepts qui optimise la MAP, soit $N_C = 4$ (Znaidi *et al.*, 2015).

– Un opérateur d'agrégation prioritaire, noté ($PSBM25$), appliqué aux ordonnancements issus de l'application du modèle $BM25$ pour l'évaluation de chacune des sous-requêtes Q_P, Q_{IC} et Q_O.

– Deux modèles de recherche d'information de l'état de l'art, spécifiques à l'évaluation de requêtes PICO : (1) un modèle d'agrégation de scores sans pondération, noté (AGM), proche du modèle présenté dans (Demner-Fushman et Lin, 2007), où le modèle $BM25$ est utilisé pour calculer la similarité entre les facettes des requêtes et les documents et, (2) *Positional Language Model* (PLM) présenté dans (Boudin *et al.*, 2010c), basé sur une extension du modèle de langue (LM). Les résultats présentés sont ceux rapportés dans (Boudin *et al.*, 2010c) sur la même collection de test et qui incluent uniquement les mesures de la MAP et la $P@5$.

5. Résultats

5.1. *Efficacité du modèle de pertinence basé sur l'agrégation des scores*

Dans un premier temps, nous avons identifié la valeur optimale du paramètre α utilisé dans l'équation 6 et ce en le variant dans l'intervalle $[0..1]$. Comme le montre la Figure 4, la valeur optimale du paramètre α est de 0.7 permettant d'atteindre un score de précision MAP égale à 0.170. Ces résultats montrent que le score basé sur les éléments PICO contribuent de façon significative au calcul du score global optimal. Nous retenons cette valeur pour le reste des expérimentations. Pour rappel, nous procédons à la variation du paramètre α dans le cadre de la validation croisée.

Le Tableau 2 présente les résultats de l'efficacité de recherche de notre modèle d'agrégation prioritaire sémantique, notée PSM, comparativement avec ceux des six modèles de référence cités ci-dessus, selon les mesures de précision (MAP et P@X) et du nombre de documents pertinents sélectionnés ($\#RR$). Les résultats montrent d'importantes améliorations significatives pour toutes les mesures d'évaluation et modèles, sauf pour le modèle PLM où l'amélioration est faible et non significative. Plus spécifiquement, pour la mesure de la MAP, les améliorations varient de $+4.60\%$ à $+52.36\%$. De plus, le nombre de documents pertinents retournés ($\#RR$) est plus im-

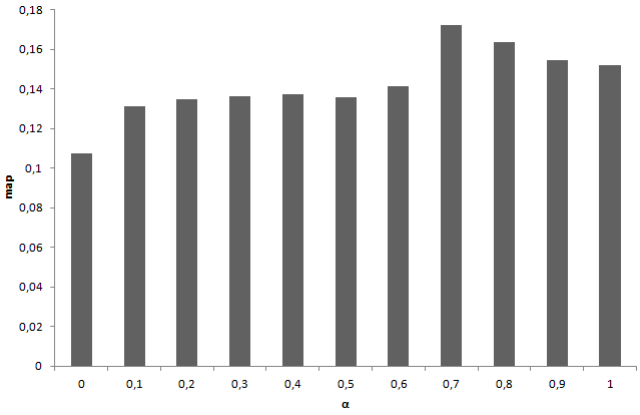

Figure 4 – Variation du paramètre de combinaison des scores basés mots et basés PICO (Formule 6)

portant en faveur du modèle PSM en comparaison avec l'ensemble des modèles de référence. D'une manière générale, ces résultats mettent en avant quatre (4) observations majeures :

1) Exploiter la sémantique cachée derrière les facettes des requêtes à travers les concepts, est efficace vu que le modèle PSM dépasse les modèles $BM25$ et LM mais aussi le modèle $PSBM25$ qui se base sur l'opérateur d'agrégation prioritaire.

2) L'importante amélioration de performance du modèle PSM par rapport au modèle CQE montre l'intérêt d'intégrer la structure de la requête en facettes P, I/C et O dans le calcul des scores de pertinence des documents.

3) Le modèle PSM donne de meilleurs résultats que le modèle AGM basé sur la représentation sémantique de la requête ainsi que l'agrégation additive des scores d'importance des facettes ; ces résultats montrent l'avantage d'assigner des scores partiels de pertinence, en adéquation avec l'importance des facettes PICO avec le document en cours d'évaluation.

4) Le modèle PSM est légèrement mais pas significativement plus pertinent que le modèle PLM ($+4.60\%$). Ce point sera particulièrement exploré lors de l'analyse de robustesse du modèle PSM présentée dans ce qui suit.

5.2. *Analyse de la robustesse du modèle d'agrégation de pertinence*

Ici, notre objectif est d'analyser la robustesse du modèle PSM. Pour rappel, un modèle de recherche d'information robuste doit impacter positivement la plupart des requêtes (Wang *et al.*, 2012). Pour cela, nous menons tout d'abord une analyse globale d'amélioration/baisse des performances sur l'ensemble des requêtes en comparaison avec les modèles de recherche d'information PICO de l'état de l'art en l'occurrence les modèles PLM et AGM. Ensuite, nous nous focalisons sur l'étude de cas de requêtes typiques du succès vs. échec du modèle PSM comparativement au modèle de

MODÈLE	Précision			% Change	# RR
	MAP	P@5	P@10		
$BM25$	0.112	0.1561	0.127	$+51.42\%^{**}$	4574
LM	0.111	0.156	0.130	$+52.36\%^{***}$	4491
CQE	0.14	0.163	0.142	$+47.71\%^{**}$	4625
$PSBM25$	0.123	0.151	0.139	$+37.94\%^{**}$	4904
AGM	0.121	0.148	0.135	$+40.09\%^{**}$	4835
PLM	0.163	0.240	–	$+4.60\%$	5770
PSM	**0.170**	**0.254**	**0.198**	–	**5894**

Tableau 2 – Evaluation comparative du modèle PSM. $\%Change$ indique le taux de changement du modèle PSM en termes de MAP. Les symboles *, ** and *** indiquent le degré de significativité du test de student : * : $0.01 < t \leq 0.05$; ** : $0.001 < t \leq 0.01$; *** : $t \leq 0.001$.

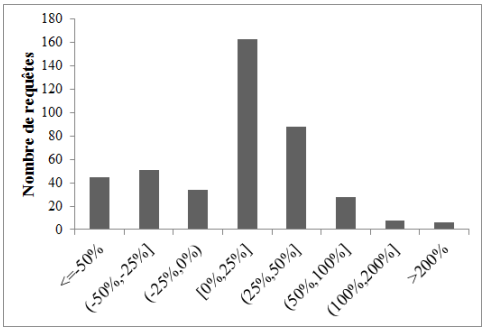

Figure 5 – Statistiques sur l'amélioration/dégradation en MAP comparé au modèle PLM

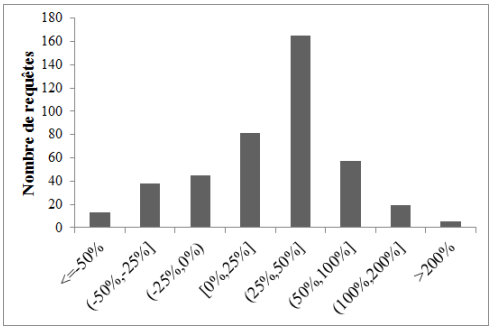

Figure 6 – Statistiques sur l'amélioration/dégradation en MAP comparé au modèle AGM

référence le plus performant pour lequel les améliorations observées sont non significatives (Section 5.1, Tableau 2).

5.2.1. *Analyse globale*

Les Figures 5 et 6 résument les statistiques concernant le nombre de requêtes dont la performance du modèle PSM a été plus élevée vs. dégradée en comparaison avec les modèles PLM et AGM. L'axe des abscisses représente le taux l'amélioration (+) vs. dégradation (-) de la performance en termes de la mesure MAP. L'axe des ordonnées représente le nombre de requêtes concernées par cette amélioration vs. dégradation. Les barres à gauche de $[0\,25\%]$ représentent les requêtes dont les performances sont plus élevées en considérant les modèles de comparaison AGM et PLM. Les barres à droite (incluant $[0\,25\%]$) représentent les requêtes dont la performance est plus élevée pour le modèle PSM. On peut observer d'après ces figures que la proportion des requêtes avec une amélioration de la performance plus élevée en faveur du modèle PSM, est plus importante. A partir de la Figure 5, nous constatons plus particulièrement un grand nombre de requêtes avec une amélioration de performance dans l'intervalle $[0\%, 25\%]$ suivi de $(25\%, 50\%]$ par rapport au modèle PLM. Dans la Figure 6, nous observons un plus grand nombre de requêtes avec une amélioration de performance dans $(25\%\,50\%]$ suivi de l'intervalle $[0\%\,25\%]$ par rapport au modèle AGM. Nous remarquons également que l'amélioration est plus importante par rapport au modèle AGM avec un plus grand nombre de requêtes dans $[25\%\,50\%]$. Cela montre l'utilité et l'efficacité d'appliquer l'opérateur d'agrégation de scores avec des poids d'importances variables d'une requête à l'autre. La performance de quelques requêtes dépasse les 100%. Pour résumer, comparé aux modèles PLM et AGM, les résultats montrent que le modèle PSM est robuste.

5.2.2. *Analyse au niveau requête*

Dans le but de comprendre les raisons de l'amélioration vs. dégradation de la performance des requêtes du modèle PSM comparativement au modèle de référence PSM, nous avons analysé les deux meilleures requêtes $R+$ et les deux plus faibles $R-$ en termes de performance MAP. Le Tableau 3 présente les ensembles $R+$ et $R-$ en y mentionnant l'identifiant de la requête (Id), sa description (Des) ainsi que le taux de changement (augmentation vs. dégradation) par rapport au modèle PLM ($\%Change$), la longueur en nombre de mots ($\#T$), le nombre de concepts extraits ($\#C$) et le score de clarté ($\#Cla$) de la requête. Ce score traduit le degré d'appariement de la requête Q avec la collection. Il est calculé sur la base d'une mesure de divergence entre le modèle de langue de la requête et celui de la collection (Steve et Croft, 2002) :

$$Cla(Q) = \sum_{t \in V} P(t|Q) log_2 \frac{P(t|Q)}{P_{coll}(t)} \quad [7]$$

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

Q	Id	Desc	$\%Change$	$\#T$	$\#C$	$\#Cla$
R^+	$M6.2$	***(P)** In obese patients diabetes(**P**)(**IC**) orlistat Placebo(**IC**) (**O**)Weight loss(**O**).*	$+78.08\%$	7	4	0.080
	$Q37.2$	***(P)** Adults 18 years or more migraine(**P**)(**IC**)aspirin plus an antiemetic placebo (**IC**) **O**)pain free(**O**).*	$+74.60\%$	8	5	0.071
R^-	$C21.1$	***(P)** Adults 14 years and older GORD (**P**)(**IC**)Medical management : proton pump inhibitors/histamine receptor antagonists Laparoscopic fundoplication surgery(**IC**)(**O**) Health-related quality of life (**O**).*	-45.07%	18	7	0.062
	$Q48.3$	***(P)** Adults 18 years or older rheumatoid arthritis (**P**)(**IC**) methotrexate combined with other non-biologic disease modifying anti-rheumatic drugs (DMARDs) methotrexate alone(**IC**)(**O**)ACR response of non-MTX DMARDS inadequate response(**O**).*	-83.21%	26	6	0.020

Tableau 3 – Analyse comparative de cas types de requêtes types PSM vs. PLM

Les résultats présentés dans le Tableau 3 montrent que le modèle PSM présente une meilleure performance pour les requêtes relativement plus courtes ($M6.2 et Q37.2$) ; inversement, le modèle PLM est plus performant pour les requêtes longues ($C21.1$ et $Q48.3$). L'amélioration de la MAP atteint 78.08% et 74.60% respectivement pour les requêtes $M6.2$ et $Q37.2$. Cela peut être expliqué par le fait qu'en comparaison avec le modèle PLM, le risque d'ambiguïté étant plus élevé pour les requêtes plus courtes, la représentation conceptuelle des facettes des requêtes proposée dans le modèle PSM permet de réduire l'effet négatif du défaut de correspondance entre les représentations document-requête sur les mots comme proposé dans le modèle PLM. À l'opposé, pour les requêtes longues, la dégradation de la MAP atteint respectivement -45.07% et -83.21% pour les requêtes $C21.1$ et $Q48.3$. On remarque que les requêtes sont bien plus longues en mots (7, 8 vs.18, 26) mais pas autant plus longues en concepts avec le même ordre de grandeur (4, 5 vs. 7, 6). Même si on constate que ces requêtes sont moins claires (0.062 et 0.020) que les requêtes plus longues (0.080 et 0.071), en défaveur du modèle PLM, ce dernier s'avère plus performant. Ceci peut être expliqué par le fait que plus la requête est longue, plus le risque d'appariement des mots avec les documents candidats est grand ; cependant la non prise en compte du contexte de la facette, comme cela est fait dans le modèle PSM, conduit au calcul d'un score de pertinence selon la formule 3 qui pourrait être erroné. A titre d'exemple, pour la requête $Q48.3$, des mots comme *older*, *rheumatoid* et *arthritis methotrexate* présents dans un document candidat peuvent s'apparier indifféremment avec les facettes IC et P dans le cas du modèle PSM. En revanche, dans

le modèle PLM, ce défaut d'appariement lié à la facette ne peut survenir puisque les documents sont préalablement annotés et l'appariement est effectué facette-facette.

6. Conclusion et perspectives

Dans cet article, nous avons proposé l'application d'un opérateur d'agrégation prioritaire pour l'évaluation de requêtes cliniques PICO. L'opérateur ne requiert pas une annotation préalable des facettes PICO dans les documents et permet d'adapter le score d'importance des facettes aux documents et requêtes en cours d'évaluation. Les expérimentations conduites sur la collection *CLIREC* ont montré que l'opérateur proposé est significativement plus performant que la majorité des modèles de référence basés sur l'appariement mot-mot, la reformulation sémantique des requêtes et des modèles d'agrégation classiques des ordonnancements issus de l'évaluation de chaque sous-requête associée à une facette. Bien que robuste, l'analyse des performances de recherche au niveau requête montre que l'opérateur PSM présente des limites. En effet, le modèle d'ordonnancement des documents ne prend pas en compte le lien entre le contexte des mots dans les documents et leur contexte dans la requête, représenté par les facettes auxquelles les mots font référence. Nous planifions de vérifier le bien fondé de ces limites en menant une analyse statistique sur l'ensemble de la collection qui permettrait de déterminer les facteurs d'échec des requêtes. Suivra, une réflexion quant à l'intégration de ces facteurs comme éléments contextuels dans l'opérateur d'agrégation prioritaire des scores.

7. Bibliographie

Boudin F., Nie J.-Y., Bartlett J., Grad R., Pluye P., Dawes M., « Combining classifiers for robust PICO element detection », *BMC Medical Informatics and Decision Making*, vol. 10, n° 1, p. 29, 2010a.

Boudin F., Nie J. Y., Dawes M., « Clinical information retrieval using document and PICO structure », *NAACL HLT*, p. 822-830, 2010b.

Boudin F., Nie J.-Y., Dawes M., « Positional language models for clinical information retrieval », *EMNLP*, p. 108-115, 2010c.

Da Costa Pereira C., Dragoni M., Pasi G., « Multidimensional Relevance : A New Aggregation Criterion », *in* M. Boughanem, C. Berrut, J. Mothe, C. Soule-Dupuy (eds), *Advances in Information Retrieval*, vol. 5478 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, chapter 25, p. 264-275, 2009.

Demner-Fushman D., Lin J., « Answering Clinical Questions with Knowledge-Based and Statistical Techniques », *Comput. Linguist.*, vol. 33, n° 1, p. 63-103, 2007.

Dinh D., Tamine L., « Towards a context sensitive approach to searching information based on domain specific knowledge sources », *Web Semantics : Science, Services and Agents on the World Wide Web*, 2012.

Fox K., Duggan M., « Health online 2013 », 2013.

Francke A., Smit M., de Veer A., « Factors influencing the implementation of clinical guidelines for health care professionals : a systematic meta-review », *BMC Medical Information Decision Making*, vol. 8, n° 38, p. 1-11, 2008.

Mao J., Lu K., Mu X., Li G., « Mining Document, Concept, and Term Associations for Effective Biomedical Retrieval : Introducing MeSH-enhanced Retrieval Models », *Inf. Retr.*, vol. 18, nº 5, p. 413-444, 2015.

Natarajan K., Stein D., Jain S., Elhadad N., « An analysis of clinical queries in an electronic health record search utility », *International journal of medical information*, vol. 79, nº 7, p. 515-522, 2010.

Pereira C., Dragoni D., Pasi G., « Multidimensionnal relevance : a new aggregation criterion », *ECIR 2010*, p. 264-275, 2010.

Robertson S. E., Sparck Jones K., « Document Retrieval Systems », Taylor Graham Publishing, London, UK, UK, chapter Relevance Weighting of Search Terms, p. 143-160, 1988.

Sackett D. L., Rosenberg W. M. C., Gray J. A. M., Haynes R. B., Richardson W. S., « Evidence based medicine : what it is and what it isn't », *BMJ*, vol. 312, nº 7023, p. 71-72, 1996.

Schardt C., Adams M., Owens T., Keitz S., Fontelo P., « Utilization of the PICO framework to improve searching PubMed for clinical questions », *BMC Medical Informatics and Decision Making*, vol. 7, nº 1, p. 16+, 2007.

Song F., Croft W. B., « A General Language Model for Information Retrieval », CIKM '99, p. 316-321, 1999.

Steve C. R., Croft W., « Quantifying query ambiguity », *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, p. 104-109, 2002.

Stokes N., Cavedon Y., Zobel J., « Exploring criteria for succesful query expansion in the genomic domain », *Information retrieval*, vol. 12, p. 17-50, 2009.

Suominen H., Salanter S., Velupillai S., Chapman W., Savova G., Elhadad N., Pradhan S., South B., Mowery D., Jones G., Leveling J., Kelly L., Goeuriot L., Martinez D., Zuccon G.*Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, vol. 8138 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 212-231, 2013.

Trieschnigg D., Proof of concept : concept-based biomedical information retrieval, PhD thesis, University of Twente, 2010.

Wang L., Bennett P. N., Collins-Thompson K., « Robust Ranking Models via Risk-sensitive Optimization », SIGIR '12, p. 761-770, 2012.

Weifield J., Finkelstein K., « How to answer your clinical questions more efficiently », *Family practice management*, vol. 12, nº 7, p. 37, 2005.

Yang L., Mei Q., K.Zheng, Hanauer D. A., « Query log analysis of an electronic health record search engine », *Proceedings of the annual symposium AMIA*, AMIA '11, p. 915-924, 2011.

Zhang Y., « Searching for specific health-related information in MedlinePlus : behavioral patterns and user experience », *Journal of the American Society for Information Science and Technology (JASIST)*, 2013.

Zhao J., yen Kan M., Procter P. M., Zubaidah S., Yip W. K., Li G. M., « Improving Search for Evidence-based Practice using Information Extraction », *BMC Medical Informatics and Decision Making*, 2010.

Znaidi E., Tamine L., Latiri C., « Answering PICO Clinical Questions : a Semantic Graph-Based Approach (short paper) », *Conference on Artificial Intelligence in Medicine (AIME)*, 2015.