La prédiction efficace de la difficulté des requêtes : une tâche impossible?

Adrian-Gabriel Chifu, Léa Laporte, Josiane Mothe

La prédiction efficace de la difficulté des requêtes : une tâche impossible?

Adrian-Gabriel Chifu* — **Léa Laporte**** — **Josiane Mothe*****

* *IRIT UMR5505, CNRS, Université de Toulouse, Université Paul Sabatier (France)*
** *LIRIS UMR 5205, CNRS, Université de Lyon, INSA Lyon, France*
** *IRIT UMR5505, CNRS, Université de Toulouse, ESPE (France)*
adrian.chifu@irit.fr; josiane.mothe@irit.fr; lea.laporte@liris.cnrs.fr

RÉSUMÉ. *Les moteurs de recherche d'information (RI) retrouvent des réponses quelle que soit la requête, mais certaines requêtes sont difficiles (le système n'obtient pas de bonne performance en termes de mesure de RI). Pour les requêtes difficiles, des traitements adhoc doivent être appliqués. Prédire qu'une requête est difficile est donc crucial et différents prédicteurs ont été proposés. Dans cet articlenous étudions la variété de l'information captée par les prédicteurs existants et donc leur non redondance. Par ailleurs, nous montrons que les corrélationsentre les prédicteurs et les performance des systèmes donnent peu d'espoir sur la capacité de ces prédicteurs à être réellement efficaces. Enfin, nous étudions la capacité des prédicteurs à prédire les classes de difficulté des requêtes en nous appuyant sur une variété de méthodes exploratoires et d'apprentissage. Nous montrons que malgré les (faibles) corrélations observées avec les mesures de performance, les prédicteurs actuels conduisent à des performances de prédiction variables et sont donc difficilement utilisables dans une application concrète de RI.*

ABSTRACT. *Search engines found answers whatever the user query is, but some queries are more difficult than others for the system. For difficult queries, adhoc treatments must be applied. Predicting query difficulty is crucial and different predictors have been proposed. In this paper, we revisit these predictors. First we check the non statistical redundancy of predictors. Then, we show that the correlation between the values of predictors and system performance gives little hope on the ability of these predictors to be effective. Finally, we study the ability of predictors to predict the classes of difficulty by relying on a variety of exploratory and learning methods. We show that despite the (low) correlation with performance measures, current predictors are not robust enough to be used in practical IR applications.*

MOTS-CLÉS : *Recherche d'information, requête difficile, prédiction, analyse de données.*

KEYWORDS: *Information retrieval, query difficulty predictor, data mining, evaluation.*

1. Introduction

Si les moteurs de recherche d'information (RI) retrouvent des réponses quelle que soit la requête de l'utilisateur, certaines requêtes sont plus *difficiles* que d'autres ; c'est-à-dire que le système ne répond pas de façon satisfaisante et donc n'obtient pas de bonnes performances en termes de mesure de RI comme la précision ou le rappel. Intuitivement, répondre à un utilisateur qui s'intéresse à la biographie de François Hollande est plus simple que de trouver les documents pertinents pour la requête "orange".

Pour certaines requêtes, n'importe quelle technique de RI peut être utilisée, chacune pouvant répondre de façon satisfaisante à l'utilisateur. Pour d'autres requêtes en revanche, il a été montré une plus grande variabilité dans les résultats. Ainsi, certaines requêtes peuvent être qualifiées de faciles (celles pour lesquelles le système retrouve des documents pertinents) et d'autres difficiles (le système ne retrouve pas les documents pertinents ou peu de documents pertinents). Prédire qu'une requête est difficile est crucial. En effet, une requête difficile peut être améliorée par des traitements particuliers comme par exemple la désambiguisation des termes qui la composent .

Le challenge de la prédiction de la difficulté des requêtes a émergé au début des années 2000 (De Loupy et Bellot, 2000 ; Carpineto *et al.*, 2001 ; Cronen-Townsend *et al.*, 2002). Plusieurs prédicteurs de difficulté de requêtes ont ainsi été proposés. Ils sont de deux types : les prédicteurs pré-recherche se basent sur des caractéristiques intrinsèques aux requêtes et aux documents de la collection et peuvent être calculés indépendamment de toute recherche ; les prédicteurs post-recherche au contraire s'appuient sur des informations extraites une fois une première recherche effectuée (le score obtenu par les documents par exemple). Même si les prédicteurs pré-recherche sont plus facilement utilisables dans un contexte réel, les meilleurs prédicteurs actuels sont post-recherche, comme l'écart-type des scores des documents retrouvés (Shtok *et al.*, 2009).

Malgré les résultats ponctuels montrant l'efficacité de certains prédicteurs dans certains cas (Yom-Tov *et al.*, 2005), (Sarnikar *et al.*, 2014), ceux-ci restent insuffisants. En effet, prédire la difficulté n'est pas une fin en soit et différentes applications concrètes de la prédiction ont émergé. Par exemple, l'expansion sélective de requêtes s'appuie sur l'hypothèse que certaines requêtes doivent être traitées telles que soumises par l'utilisateur alors que d'autres seront plus efficaces si une expansion est réalisée au préalable (He et Ounis, 2007).

Dans cet article, nous montrons que les prédicteurs de la littérature ne sont pas redondants entre eux ; ce qui permet de penser que leur utilisation combinée pourrait améliorer la prédiction. Nous présentons également les résultats d'une analyse qui montre que chacun des prédicteurs n'est pas suffisamment corrélés au sens statistique (coefficients de Kendall et Spearman) avec les mesures de performance pour être considéré comme utilisable pour prédire la difficulté des requêtes. Finalement, nous étudions la combinaison des prédicteurs pour prédire des classes de difficultés et montrons que les résultats ne sont pas stables et dépendent des jeux de données.

Le reste de l'article est organisé comme suit : la section 2 présente les travaux liés. Dans la section 3, nous décrivons les objectifs et points centraux de notre analyse en formulant les hypothèses que nous avons testées. Dans la section 4, nous présentons les données utilisées dans notre étude. Les sections suivantes détaillent les résultats obtenus : les corrélations entre prédicteurs (section 5), celles entre prédicteurs et mesures de performances (section 6) et la prédiction des classes de difficulté (section 7). La section 8 conclut cet article.

2. Etat de l'art

L'objectif des prédicteurs de difficulté de requêtes est de déterminer si le système est capable de bien répondre à la requête posée, c'est-à-dire s'il est capable de restituer les seuls documents pertinents qui répondent au besoin d'information sous-jacent. L'efficacité des prédicteurs est mesurée par la corrélation de la mesure de performance retenue (généralement une mesure de précision) avec les valeurs des prédicteurs. A la différence de (Grivolla, 2005) qui étudient les corrélations entre les performances des systèmes et les scores des documents retrouvés ; nous nous intéressons ici aux corrélations entre les performances des systèmes et les prédicteurs dédiés à la difficulté des requêtes.

Prédicteurs de la difficulté des requêtes. Ils sont généralement regroupés en prédicteurs pré-recherche et prédicteurs post-recherche.

Parmi les prédicteurs pré-recherche qui peuvent être calculés indépendamment de toute recherche, on peut citer l'*IDF* moyen des termes de la requête : il mesure le pouvoir discriminant des termes de la requête (Spärck Jones, 1972). Le *score de clarté* (clarity score) vise quant à lui à quantifier le niveau d'ambiguïté d'une requête ; il correspond à l'entropie relative entre le modèle de langue de la requête et le modèle de langue de la collection de documents (Cronen-Townsend *et al.*, 2002). On peut citer également le *nombre moyen de sens des termes de requête* et la *complexité de la requête* (Mothe et Tanguy, 2005). L'*étendue de la requête* (query scope) mesure le pourcentage de documents de la collection qui contiennent au moins un des termes de la requête (He et Ounis, 2004).

Les prédicteurs post-recherche nécessitent d'exécuter une recherche avec la requête dont on souhaite prédire la difficulté pour extraire des éléments issus des documents retrouvés. Le *gain pondéré d'information* (Zhou et Croft, 2007) mesure l'écart entre la moyenne des scores en haut de la liste des documents retrouvés et l'ensemble du corpus. L'accord entre les résultats de la requête complète et ceux obtenus avec les sous-requêtes est un autre prédicteur (Yom-Tov *et al.*, 2005). L'*engagement normalisé de la requête* (*NQC*) (Shtok *et al.*, 2009) est un prédicteur post-recherche qui mesure l'écart-type entre les scores des documents retrouvés. Le *retour sur la requête* (*QF*) (Zhou et Croft, 2007) calcule le chevauchement entre deux listes de documents retrouvés.

Les prédicteurs peuvent être utilisés indépendamment ou de façon combinée. Par exemple (Hauff, 2010) utilise les régressions linéaires. (Bashir, 2014) combine des prédicteurs pré-recherche en s'appuyant sur les algorithmes génétiques. Ces différentes études montrent que la combinaison de prédicteurs est plus efficace que l'usage indépendant de ces prédicteurs.

Evaluation de la qualité de la prédiction. Pour évaluer la qualité d'un prédicteur de difficulté, la plupart des travaux étudie la corrélation entre les résultats du prédicteur et la précision moyenne (AP). Généralement, les coefficients de Pearson, Spearman et Kendall sont utilisés (Hauff *et al.*, 2008). Dans la mesure où le caractère linéaire de la corrélation n'est pas garanti, il semblerait cependant plus opportun d'éviter l'utilisation du coefficient de Pearson.

Dans (Hauff *et al.*, 2008), 22 prédicteurs pré-recherche sont étudiés sur les collections *Robust*, *Gov2* et *WT10G* de TREC[1]. Les auteurs montrent que les corrélations diffèrent en fonction des collections et de la fonction de recherche (différents lissages Dirichlet sur un modèle de langue) mais que, globalement, *MaxIDF* est le prédicteur pré-recherche le plus commun aux différents paramétrages. (Hauff, 2010) presente une étude assez complète comparant les différents prédicteurs. L'auteur étudie à la fois les corrélations entre les prédicteurs de la même famille (spécificité, ambiguité, parenté des termes et sensitivité de l'ordonnancement) et les corrélations (Pearson et Kendall) entre les valeurs des prédicteurs et l'AP. A l'interieur d'une famille les corrélations varient beaucoup et peuvent atteindre une valeur approximative de $0,7$. Le prédicteur pré-recherche le plus corrélé avec la performance est le *MaxIDF* (coefficient de corrélation de Pearson égal à $0,532$ pour la collection TREC Robust), mais le plus robuste est le *MaxVAR* (la variance maximale de la pondération des termes), avec une corrélation de Pearson autour de $0,4$ quelle que soit la collection. Pour (Shtok *et al.*, 2010), *NQC* et *QF* sont les prédicteurs post-recherche les plus corrélés à la mesure AP .

Dans cet article, nous revisitons les prédicteurs de difficulté. Nous formulons différentes hypothèses que nous vérifions au travers de méthodes d'analyse adaptées.

3. Hypothèses et méthodologie

Hypothèse 1 : les prédicteurs ne sont pas corrélés entre eux. A notre connaissance aucune étude ne s'est intéressée à mesurer la corrélation entre les différents prédicteurs proposés dans la littérature. Or, s'il s'avère que les prédicteurs sont fortement corrélés, il serait intéressant de prospecter vers d'autres prédicteurs.

Pour vérifier l'hypothèse de la corrélation faible des prédicteurs actuels, nous avons utilisé les coefficients de corrélation de Kendall et de Spearman (adaptés à des corrélations non nécessairement linéaires). Plus spécifiquement, nous avons recalculé les valeurs de 31 prédicteurs de difficulté ou combinaisons de prédicteurs de difficulté, provenant de 4 familles de prédicteurs, sur différentes collections de test issues de la

1. TREC : Text REtrieval Conference `http://mitpress.mit.edu/books/trec`

campagne d'évaluation TREC. Nous avons ensuite calculé les corrélations entre les prédicteurs pris deux à deux. Les résultats confirment que les prédicteurs ne sont pas fortement corrélés. Le détail des résultats est présenté dans la section 5.

Hypothèse 2 : Les valeurs des prédicteurs sont corrélées aux valeurs à prédire. Dans la mesure où les prédicteurs ont été définis pour prédire la difficulté des requêtes, c'est-à-dire pour prédire les succès/échecs des systèmes de recherche, il est naturel de faire l'hypothèse que les valeurs des prédicteurs sont corrélés aux performances.

Actuellement, les mesures de performance majoritairement utilisées pour mesurer le succès d'une recherche sont la précision moyenne (AP) et la précision au rang 10 (P@10). Vérifier l'hypothèse 2 peut donc naturellement être réalisé en calculant la corrélation entre les valeurs des prédicteurs et les mesures de performances. La section 6 présente les résultats détaillés de cette analyse.

Bien que nous utilisions plus de prédicteurs, nos résultats sont conformes à ceux rapportés dans la littérature avec des corrélations généralement assez faibles : les meilleures corrélations significatives de Spearman sont de l'ordre de $0,5$.

Hypothèse 3 : Les prédicteurs ne sont pas robustes. La difficulté que nous avons rencontrée pour utiliser les prédicteurs de difficulté de façon efficace et stable dans des applications de RI nous a amenés à penser que les prédicteurs n'étaient pas suffisamment robustes. Notre hypothèse est donc que les prédicteurs ne permettent pas de partitionner efficacement les requêtes par rapport à leur difficulté.

Pour valider cette hypothèse, nous avons utilisé différentes méthodes d'exploration ou d'analyse de données afin d'étudier les classes de difficulté prédites et les classes observées.

Quelle que soit la méthode utilisée, nous avons montré qu'il n'était pas possible de prédire de façon efficace la classe de difficulté à laquelle appartient une requête. Les détails sont fournis en section 7.

4. Données analysées

4.1. *Prédicteurs de difficulté*

Dans notre étude, nous considérons différents prédicteurs de difficulté des requêtes calculés à partir du texte de la requête, de la collection de documents utilisée lors de la recherche et de la liste des documents retrouvés par la requête. Nous considérons d'une part la requête courte, issue du titre (notée T) et d'autre part la requête longue issue du titre et de la description (notée TD) des besoins d'information de TREC. Les prédicteurs et leurs variantes utilisées dans cette étude sont dérivés des classes de prédicteurs de difficulté de la littérature présentés dans la section 2. Nous les décrivons dans la suite de cette section.

L'ambiguïté des termes. Le *Nombre de sens de WordNet* (WNS) représente un prédicteur linguistique de pré-recherche, il correspond à une mesure de l'ambiguïté. Il

est calculé par le nombre moyen de sens dans WordNet[2] pour les termes de la requête (Mothe et Tanguy, 2005). Dans notre jeu de données, l'ambiguïté des termes est représentée par six caractéristiques dérivées de ce prédicteur : le maximum, la moyenne et la somme du nombre de sens des termes de la requête, calculés pour les requêtes courte T et étendue TD. Par exemple, *TD_wns_max* correspond au maximum de sens des mots de la requête TD.

La discrimination des termes. La *Fréquence Inverse* (*idf*) est un prédicteur statistique de pré-recherche mesurant si un terme est rare ou commun dans le corpus (Spärck Jones, 1972). Sa valeur pour une requête représente la moyenne des *idf* pour les termes de la requête. Dans la présente étude, nous utilisons quatre variantes : le minimum, le maximum, la moyenne et la somme des *idf*. Les *idf* sont calculés à la fois pour la requête T et la requête TD. Par exemple, le maximum des *idf* pour TD est noté *TD_max_idf*.

L'homogénéité des listes de documents : L'*écart-type* (*STD*) représente un prédicteur post-recherche statistique mesurant le degré de variation par rapport à la moyenne de la liste des scores attribués aux documents retrouvés pour la requête. Il s'agit d'une variante du *NQC* proposé par (Shtok *et al.*, 2009), sans normalisation. Dans notre jeu de données, l'homogénéité des listes de documents est caractérisée par 3 variables : la valeur des *STD* pour la liste des documents extraits de la requête T (notée *STD_T*), de la requête TD (notée *STD_TD*) et de leur différence (notée *STD_Diff_T-TD*).

La divergence des listes. Le *retour sur la requête* (QF) (Zhou et Croft, 2007) est un prédicteur post-recherche qui calcule le chevauchement entre deux listes de documents retrouvés. Nous considérons 4 listes de documents dans notre étude : issue du traitement de la requête courte avec et sans reformulation automatique (listes TRF et T respectivement) et du traitement de la requête étendue TD, avec et sans reformulation automatique (listes TDRF et TD respectivement). Pour obtenir ces listes, nous avons utilisé le moteur Indri qui implante le modèle de langue. Nous avons d'une part utilisé le modèle *Query Likelihood* et le modèle avec reformulation automatique de requêtes RM3 (les détails se trouvent en section 4.2) . A partir de ces paires de listes, le chevauchement représente le nombre de documents que ces listes ont en commun. Les listes peuvent être considérées partiellement, en ne s'intéressant qu'aux x premiers documents restitués. Dans notre jeu de données, nous considérons des combinaisons entre ces différentes listes et 4 valeurs pour x : (5, 10 , 50 , 100). Nous notons les prédicteurs *QF_L1_L2_x* où L1 est L2 sont les listes de documents retrouvés (T, TD, TRF, TDRF) et x le niveau de coupe.

Les combinaisons pré/post-recherche. Nous avons considéré 2 prédicteurs combinés : le produit entre le *WNS* de la requête T et la différence des *STD* ainsi que *IDF* multiplié par la différence des *STD*.

2. http ://wordnet.princeton.edu/

4.2. *Collections*

Collections. Nous avons considéré trois collections de test de référence, issues de la tâche *ad hoc* de TREC : la collection *Robust* comprenant environ 2 Go de documents et 250 besoins d'information (301 à 450 et 601 à 700), *WT10G* comprenant environ 10 Go de documents Web et 100 besoins d'information (451 à 550) et *GOV2* comprenant environ 426 Go de documents et 150 besoins d'information (701 à 850).

Le calcul des caractéristiques post-recherche nécessite de réaliser une recherche. Pour chaque requête (T d'une part et TD d'autre part), nous avons réalisé deux recherches : une sans expansion et une avec expansion. Les collections ont d'abord été indexées : les mots vides ont été supprimés, les mots restants ont été racinisés par l'outil Krovetz (Krovetz, 1993). Pour la recherche sans expansion, nous avons utilisé le modèle de langue *query likelihood model* (QL) avec un lissage de Dirichlet ($\mu = 1000$). Cette valeur pour μ est considérée comme adaptée (Lavrenko, 2009). Pour l'expansion de requêtes, nous avons utilisé le modèle *Relevance Model 3* (RM3) (Lavrenko et Croft, 2001), qui correspond à une interpolation entre le modèle de Pseudo Relevance Feedback (RM1) et la requête initiale (QL). Le choix du paramètre pondérant l'importance relative des deux modèles reste un problème ouvert dans la littérature, nous l'avons fixé à $0,5$, donnant ainsi une importance identique à la requête initiale et à sa reformulation.

Modèles de recherche Plusieurs études ont montré que les systèmes obtenaient des résultats différents en fonction des requêtes. Lorsque nous calculons les corrélations entre les prédicteurs et les performances, nous nous appuyons sur les différents systèmes présentés : QL avec T, QL avec TD, RM3 avec T et avec TD. Les résultats obtenus pour chacune des collections et des modèles sont présentés dans le tableau 1 ; ils sont conformes aux résultats obtenus dans la littérature et constituent de bonnes références (Collins-Thompson, 2009).

Tableau 1. *Les exécutions retenues - MAP et P@10 pour chaque collection*

Collection	T		TD		TRF		TDRF	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
Robust	0,233	0,404	0,260	0,434	0,260	0,425	0,296	0,460
WT10G	0,194	0,300	0,215	0,357	0,220	0,317	0,244	0,378
GOV2	0,292	0,543	0,294	0,557	0,332	0,590	0,329	0,583

5. Corrélation entre les différents prédicteurs

Avant de vérifier la corrélation des différents prédicteurs deux à deux, nous présentons leurs valeurs initiales sous forme de boîtes à moustaches.

La figure 1 présente les valeurs de chaque prédicteur sur la collection Robust de TREC. Par exemple, le prédicteur *TD_max_idf* calculé sur l'ensemble des requêtes

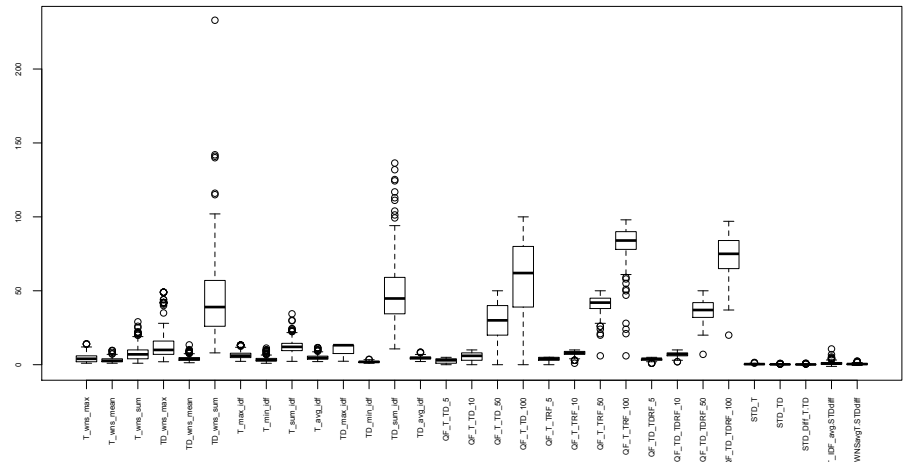

Figure 1. *Boîtes à moustaches - Valeurs initiales des caractéristiques des requêtes (Robust TREC)*

de cette collection prend ses valeurs entre 2,268 et 13,177. La moitié de ses valeurs sont inférieures à 6,119 (médiane matérialisée par le trait horizontal dans la boîte), au moins 25% de ses valeurs sont inférieures à 4.807 et au moins 75% de ses valeurs sont inférieures à 7,770 (valeurs matérialisées par la boîte). Les valeurs aberrantes sont représentées par des points en dessous ou au dessus des moustaches (moustaches matérialisées par des pointillés verticaux).

Les valeurs des prédicteurs sont très variables. Certains prédicteurs ont des valeurs très faibles et une faible variété (par exemple *STD_TD* prend ses valeurs entre 0,055 et 0,937) alors que d'autres ont un plus grand écart type (par exemple *QF_T_TD_100* prend ses valeurs entre 0 et 100). certaines analyses utilisent les valeurs centrées réduites afin d'homogéniser ces variables assez hétérogènes.

La figure 2 montre les corrélations de Kendall entre les différents prédicteurs retenus pour la collection Robust et considérés deux à deux. Les corrélations négatives sont moins nombreuses que les corrélations positives. Les seules corrélations négatives significatives supérieures à $-0,4$ sont entre les prédicteurs (*T_WNS_max* et *T_avg_diff*), (*T_WNS_mean* et *T_avg_diff*), (*T_WNS_sum* et *T_avg_diff*) et (*T_WNS_sum* et *T_min_diff*). Des corrélations fortes (supérieures à 0,6) sont observées pour 11 couples de caractéristiques présentés dans le Tableau 2.

Nous observons que si les variantes d'une mesure (par exemple *T_WNS_sum*, *T_WNS_max* et *T_WNS_mean*) sont généralement très fortement corrélées, il peut arriver que certaines variantes explorent des espaces différents. Par exemple, *TD_min_idf* et *TD_sum_idf* sont faiblement voire non corrélées (coefficient de corré-

Tableau 2. *Les couples de caractéristiques avec une corrélation supérieure à* $0,6$

T_WNS_max ↔ *T_WNS_mean*	*TD_WNS_max* ↔ *TD_WNS_sum*	*QF_TD_TDRF_50* ↔ *QF_TD_TDRF_100*
T_WNS_max ↔ *T_WNS_sum*	*TD_WNS_mean* ↔ *TD_WNS_sum*	*STD_T* ↔*T_IDF_avg.STDdiff*
T_WNS_mean ↔ *T_WNS_sum*	*QF_T_TD_5* ↔ *QF_T_TD_10*	*STD_Diff_T.TD* ↔ *T_IDF_avg.STDdiff*
TD_WNS_max ↔ *TD_WNS_mean*	*QF_T_TD_100* ↔ *QF_T_TD_50*	

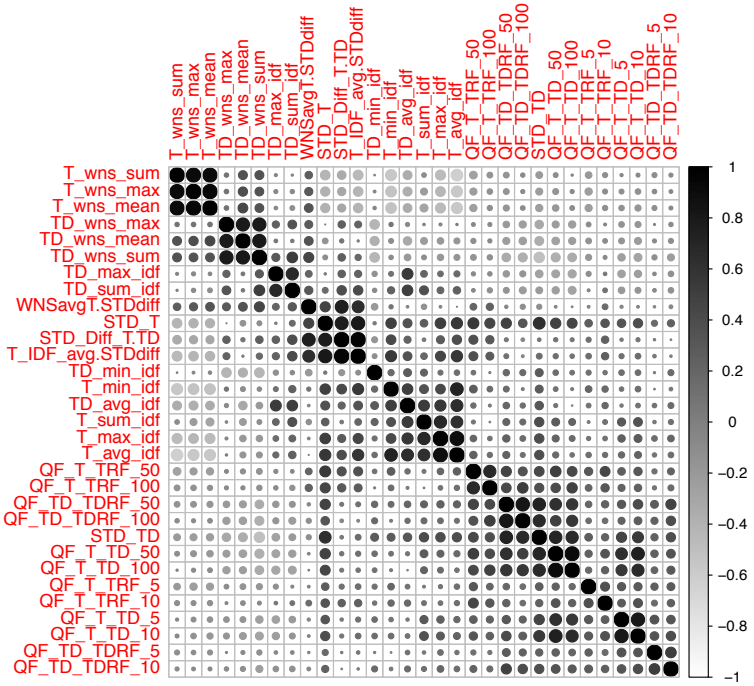

Figure 2. *Corrélation de Kendall entre les prédicteurs (requêtes Robust TREC).*

lation égal à -0,06). Ces résultats montrent qu'il peut être intéressant de s'appuyer sur différentes variantes des prédicteurs.

Sauf quelques exceptions entre des variantes d'un même prédicteur, il n'y a pas de corrélations fortes entre les prédicteurs pris deux à deux. Ce résultat indique que les prédicteurs ne sont pas redondants. Cela conduit à deux conclusions :

– la combinaison de prédicteurs peut potentiellement être plus efficace que l'utilisation indépendante des prédicteurs,

– la variété des prédicteurs de la littérature est bien réelle.

Les corrélations entre les différents prédicteurs peuvent également être visualisées sur les deux premiers axes principaux d'une analyse en composante principale (ACP).

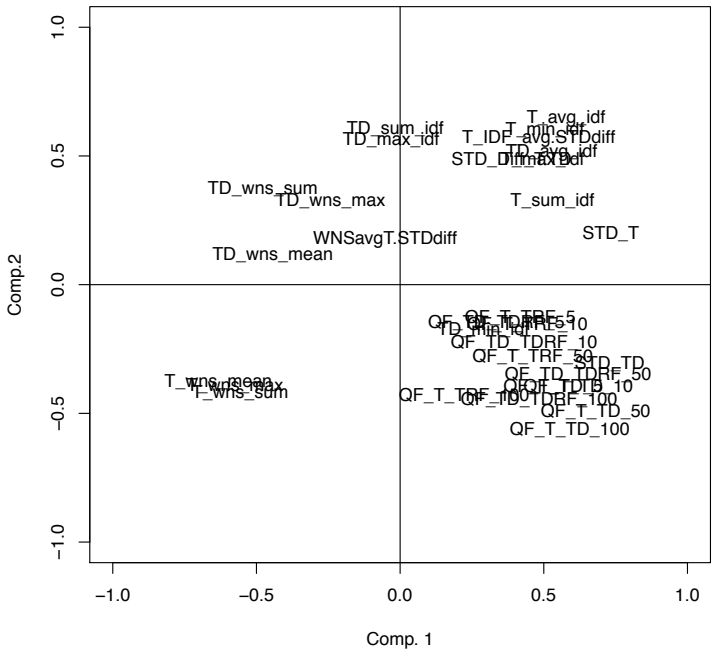

Figure 3. *Corrélation multiple (ACP) entre les prédicteurs (requêtes Robust TREC).*

La figure 3 identifie clairement différents paquets de variables corrélées (par exemple les prédicteurs à base de *QF* se trouvent tous en bas à droite). Cette figure montre également la complémentarité de ces prédicteurs . Elle est cohérente avec les résultats de la figure 2. Les résultats obtenus sur les autres collections amènent aux mêmes conclusions.

6. Corrélations entre prédicteurs et valeurs à prédire

Les valeurs à prédire (AP) obtenues sur les différentes collections varient de façon importante. Par exemple, pour les requêtes de la collection Robust de TREC avec les 4 exécutions retenues, l'AP varie de 0 à $0,9484$ sur l'ensemble des requêtes.

Nous nous sommes intéressés aux corrélations statistiquement significatives. Les corrélations obtenues avec le coefficient de Spearman (test du coefficient de Spearman) et celui de Kendall (test du coefficient de Kendall) sont cohérentes. Par exemple pour une exécution donnée, les deux coefficients déterminent les mêmes 5 prédicteurs les plus corrélés avec l'AP (à deux exceptions près). Quelle que soit l'exécution considérée (parmis les 4 étudiées) et quel que soit le coefficient de corrélation considéré, le prédicteur le plus corrélé avec l'AP est toujours *QF_T_TD_10* pour les

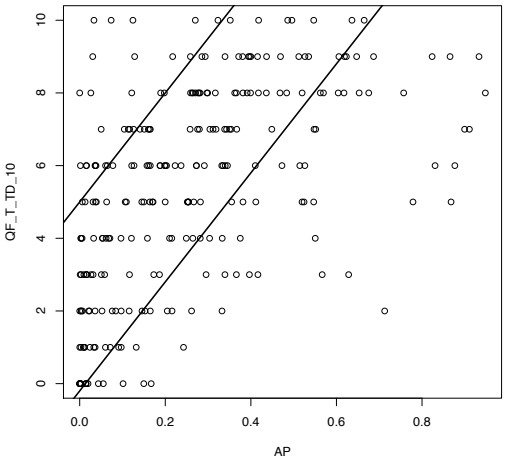

Figure 4. *AP pour l'exécution TRF et le prédicteur QF_T_TD_10 (Robust TREC).*

collections WT10G et Robust et *STD_TD* pour la collection Gov2. La corrélation de *QF_T_TD_10* avec l'AP sur Robust varie de $0,588$ à $0,469$ avec Spearman et de $0,436$ à $0,344$ avec Kendall, suivant les exécutions ; elle est un tout petit peu plus faible sur la collection WT10G. Le prédicteur *QF_T_TD_5* fait également et systématiquement partie des 5 prédicteurs les plus corrélés avec l'AP sur Robust et WT10G ; il est également corrélé de façon significative avec l'AP pour la collection Gov2.

Si nous considérons la $P@10$, les mêmes prédicteurs sont les plus corrélés : *QF_T_TD_10* pour les collections WT10G et Robust et *STD_TD* pour la collection Gov2.

La figure 4 représente les valeurs de l'AP pour l'exécution TRF et le prédicteur *QF_T_TD_10* sur l'ensemble des requêtes de la collection Robust de TREC. Ce prédicteur est le plus corrélé avec l'AP quelle que soit l'exécution. Malgré cette corrélation positive mais peu élevée numériquement $(0,588)$, la figure montre que le nuage de points est assez compact et que la corrélation existe bien.

Dans la majorité des cas, la corrélation est faible. Les corrélations les plus élevées sont proches de $0,6$, ce qui n'est généralement pas considéré comme une corrélation forte. Ainsi, pris indépendamment, les prédicteurs ne sont pas efficaces pour la prédiction de la mesure de performance du système.

7. Utilisation des prédicteurs pour prédire des classes de difficulté de requêtes

Dans cette section, nous avons étudié la capacité d'un modèle à apprendre la classification des requêtes selon des classes de difficulté.

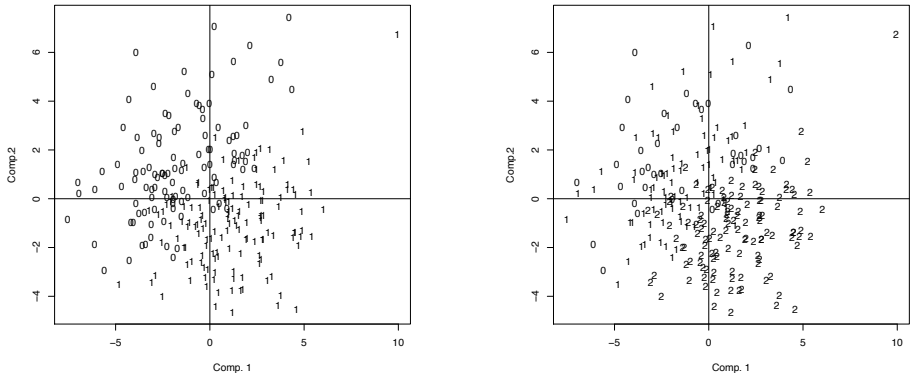

Figure 5. *ACP Prédicteurs - Requêtes selon leur classe de difficulté observée pour Robust TREC (à gauche avec 2 classes de difficulté, à droite avec 4).*

La littérature du domaine n'a pas proposé de valeur plancher de AP ou de $P@10$ en deça de laquelle la recherche serait considérée comme un échec ou la requête comme difficile. En effet, les valeurs dépendent des collections ; par exemple, la valeur moyenne de l'AP sur les 50 requêtes que comprend la collection TREC8 est d'environ $0,25$ alors qu'elle est plutôt de $0,20$ sur la collection *WT10G* et de $0,15$ sur *Clueweb 2009*. Ainsi, plutôt que de définir un seuil de valeur pour une mesure de performance, il est plus juste de définir des classes de valeurs comprenant un même effectif. Nous avons retenu deux classifications : à 2 et 4 classes. Dans le cas de 2 classes, l'attribution à l'une ou l'autre des classes est réalisée par rapport à l'AP obtenue par la requête : si l'AP est inférieure à la médiane, la requête est considérée comme difficile, sinon elle est considérée comme facile. Dans le cas de 4 classes, les quartiles définissent la classe d'appartenance : les requêtes sont partitionnées, suivant leur valeur de performance, en quatre classes comprenant chacune 25% des requêtes. Dans toute cette section, nous nous sommes appuyés sur une exécusion TRF, basée sur le titre des besoins d'information avec reformulation de requête automatique (cf. tableau 1). Les requêtes sont représentées via les valeurs de prédicteurs associées.

7.1. *Corrélations multiples par une analyse à postériori*

La figure 5 montre l'ACP lorsque les prédicteurs sont les variables et les requêtes les individus. Chaque requête est représentée par un chiffre qui indique la classe à laquelle elle appartient ; la partie gauche correspond au cas de 2 classes et la partie droite au cas de 4 classes.

Dans les deux cas, on note que les classes se sont pas clairement séparées par les deux premiers axes de l'ACP. Dans le cas de 2 classes, cela pourrait être dû à la construction des classes puisque des requêtes qui obtiennent une valeur de l'AP

proche de la médiane peuvent être associées à deux classes différentes même si leurs valeurs sont proches. En revanche, dans le cas de 4 classes, le fait que les classes extrêmes (notées 0 et 3) ne soient pas clairement distinguées par l'ACP montre que les combinaisons de prédicteurs ne permettent pas de séparer les requêtes selon leur difficulté.

7.2. *Classification k-moyenne*

L'ACP ne permet pas de visualiser les classes de difficulté des requêtes ; nous avons voulu vérifier que d'autres méthodes ne le permettaient pas. Nous avons ainsi utilisé la classification k-moyenne où k correspond au nombre de classes souhaitées (2 ou 4). Les requêtes sont regroupées par rapport à la ressemblance de leur représentation (ici correspondant aux valeurs de chaque prédicteur). Nous avons utilisé la distance Euclidienne. Plutôt que de prendre les initiateurs de groupe au hasard, nous avons *aidé* l'algorithme en lui fournissant les barycentres des classes attendues.

Le tableau 3 presente la matrice de confusion entre les classes observées et les classes obtenues par les k-moyennes. Le taux d'erreur (nombre de requêtes mal classées / nombre de requêtes) est de 36% dans le cas de deux classes et de 62% dans le cas de 4 classes. Dans ce dernier cas pour les classes extrêmes (très difficiles ou très faciles), la classification k-moyenne a un taux de réussite au mieux de 55% (34 requêtes sur 62 sont bien classées).

Tableau 3. *Matrice de confusion-Erreur entre la classification k-moyenne et la classe observée - Robust TREC (à gauche avec 2 classes de difficulté, à droite avec 4.*

k-moy. \ Obs.	C0	C1
C0	71	35
C1	54	89

k-moy. \ Obs.	C0	C1	C2	C3
C0	34	15	11	10
C1	12	7	4	8
C2	8	21	23	13
C3	9	19	24	31

7.3. *Classification SVM et arbre de décision*

Dans cette section, nous avons utilisé une méthode d'apprentissage supervisé, les Séparateurs à Vaste Marge (SVM), qui ont fait leurs preuves dans de nombreux domaines. Nous avons considéré d'une part un noyau linéaire et d'autre part un noyau gaussien. Nous avons également utilisé une autre méthode d'apprentissage supervisé, les arbres de décision binaires (algorithme CART). Dans une première phase, nous avons souhaité vérifier que la classification pouvait être réalisée par ces méthodes en utilisant l'ensemble des requêtes.

Le tableau 4 présente les erreurs de classification lorsque 2 classes de requêtes sont définies.

Tableau 4. *Taux d'erreur des méthodes supervisées SVM et CART*

Méthode	GOV2	Robust	WT10G
SVM linéaires	0,27	0,27	0,22
SVM gaussiens	0,26	0,14	0
Arbre de décision	0,15	0,15	0,17

Apprendre sur l'ensemble des données n'a pas d'utilité en soit, mais cela permet de connaître la capacité de l'algorithme à potentiellement classer les données. On note que quelle que soit la méthode il reste de 15 à 20% d'erreur de classification. Ceci peut paraître correct et donc nous a conduit à étudier la véritable capacité d'apprentissage.

Nous avons donc ensuite utilisé ces mêmes méthodes dans un usage plus classique de données d'apprentissage / données de test. Nous avons appris sur 80% des données et testé sur les 20% restant en réalisant 5 partitions. Cette procédure a été appliquée sur les trois collections ; les partitions de requêtes sont différentes pour chaque collection.

Tableau 5. *Erreur de classification des SVM linéaires sur les différentes collections*

Collection	GOV2		ROBUST		WT10G	
	Appr.	Test	Appr.	Test	Appr.	Test
Partition 1	0,24	0,34	0,22	0,29	0,29	0,42
Partition 2	0,23	0,34	0,21	0,28	0,09	0,32
Partition 3	0,25	0,33	0,26	0,18	0,35	0,16
Partition 4	0,13	0,50	0,23	0,38	0,05	0,50
Partition 5	0,23	0,40	0,25	0,34	0,17	0,40
Moyenne	0,21	0,38	0,23	0,29	0,19	0,36

Le tableau 5 présente les taux d'erreur dans le cas des SVM linéaires. Le taux d'erreur moyen sur le test est de l'ordre de 40%. On note également une grande disparité en fonction des tirages puisque les partitions 4 de Gov2 et WT10G ne font pas mieux que le hasard (50% d'erreur). Les taux d'erreurs sont bien plus importants sur les jeux de test que sur les jeux d'apprentissage. On note en effet des cas de sur-apprentissage potentiel puisque les taux d'erreur moyens sur l'apprentissage sont de moins de 20% alors que ceux sur les tests sont de 30 à 40%.

Dans le cas des SVM gaussiens (cf. tableau 6), le sur-apprentissage est encore plus important (en particulier sur la collection WT10G). Le taux d'erreur moyen sur le test est de l'ordre de 30% quelle que soit la collection.

Dans le cas d'un arbre de décision CART, les taux d'erreur sur le test varient de 35 à 51% en fonction des collections alors que le taux d'erreur sur l'ensemble d'apprentissage est entre 12 et 20%. Ceci montre un sur-apprentissage et donc une instabilité des résultats. Par ailleurs, nous notons que CART identifie sur les 5 partitions

Tableau 6. *Erreur de classification des SVM gaussiens sur les différentes collections*

Collection	GOV2		ROBUST		WT10G	
	Appr.	Test	Appr.	Test	Appr.	Test
Partition 1	0	0,27	0,18	0,33	0	0,47
Partition 2	0,16	0,31	0,26	0,22	0,17	0,32
Partition 3	0,24	0,37	0,27	0,18	0	0,37
Partition 4	0	0,40	0,23	0,40	0	0,25
Partition 5	0,24	0,37	0,17	0,30	0	0,25
Moyenne	0,13	0,34	0,21	0,29	0,03	0,33

$QF\_T\_TD\_10$ comme étant le paramètre le plus discriminant pour la collection Robust et STD_TD pour la collection Gov2. Ce résultat est cohérent avec les résultats obtenus avec les corrélations présentées dans la section 5. Concernant la collection WT10G, le paramètre le plus discriminant dépend des tirages.

8. Conclusion et perspectives

Cet article présente une analyse portant sur les prédicteurs de difficulté de requêtes. Nous nous sommes intéressés à 4 familles de prédicteurs et leurs différentes variantes. Nous avons montré sur 3 collections de référence issues de TREC que les prédicteurs de la littérature étaient assez faiblement corrélés entre eux, ce qui montre la potentialité de les combiner lors de la prédiction. Nous avons également montré que ces prédicteurs étaient corrélés (mais pas fortement corrélés) avec les mesures de performance qu'ils sont supposés prédire. En revanche, et malgré l'utilisation de différentes méthodes d'analyse et d'apprentissage, nous avons montré que les prédicteurs ne parvenaient pas à prédire correctement des classes de difficulté.

Ces résultats peuvent expliquer certains résultats ponctuels sur les applications des prédicteurs de difficulté montrés dans la littérature mais expliquent également les nombreux échecs dans d'autres tentatives, en particulier lorsqu'il s'agit de généraliser ces résultats. Dans nos prochains travaux, nous souhaitons analyser différentes mesures de performance ainsi que différents types de systèmes.

Ces travaux s'inscrivent dans le cadre du projet CAAS Contextual Analysis and Adaptive Search, ANR-10-CONT-001, financé par l'ANR dans l'appel Contint 2010.

9. Bibliographie

Bashir S., « Combining pre-retrieval query quality predictors using genetic programming », *Applied Intelligence*, vol. 40, n° 3, p. 525-535, 2014.

Carpineto C., de Mori R., Romano G., Bigi B., « An information-theoretic approach to automatic query expansion », *ACM Trans. Inf. Syst.*, vol. 19, n° 1, p. 1-27, 2001.

Collins-Thompson K., « Reducing the risk of query expansion via robust constrained optimization », *CIKM*, p. 837-846, 2009.

Cronen-Townsend S., Zhou Y., Croft W. B., « Predicting query performance », *International ACM Conference on Research and Development in Information Retrieval, SIGIR*, p. 299 - 306, 2002.

De Loupy C., Bellot P., « Evaluation of document retrieval systems and query difficulty », *Workshop Using Evaluation within HLT Programs : Results and trends*, p. 31-38, 2000.

Grivolla J., « Une méthode pour l'évaluation automatiquede la "difficulté" d'une requête », *Proceedings of CORIA, Grenoble, France*, p. 39-49, 2005.

Hauff C., « Predicting the effectiveness of queries and retrieval systems », *SIGIR Forum*, vol. 44, n° 1, p. 88, 2010.

Hauff C., Hiemstra D., de Jong F., « A survey of pre-retrieval query performance predictors », *ACM Conference on Information and Knowledge Management, CIKM*, p. 1419-1420, 2008.

He B., Ounis I., « Inferring query performance using pre-retrieval predictors », *International Conference, SPIRE*, p. 43 - 54, 2004.

He B., Ounis I., « Combining fields for query expansion and adaptive query expansion », *Information Processing & Management*, vol. 43, n° 5, p. 1294-1307, 2007.

Krovetz R., « Viewing Morphology As an Inference Process », *International Conference on Research and Development in Information Retrieval SIGIR*, p. 191-202, 1993.

Lavrenko V., *A generatice theory of relevance*, vol. 26 of *The Information Retrieval Series*, 2009.

Lavrenko V., Croft W. B., « Relevance based language models », *International Conference on Research and Development in Information Retrieval, SIGIR*, p. 120-127, 2001.

Mothe J., Tanguy L., « Linguistic features to predict query difficulty », *Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty - methods and applications workshop*, p. 7-10, 2005.

Sarnikar S., Zhang Z., Zhao J. L., « Query-performance prediction for effective query routing in domain-specific repositories », *JASIST*, vol. 65, n° 8, p. 1597-1614, 2014.

Shtok A., Kurland O., Carmel D., « Predicting Query Performance by Query-Drift Estimation », *Advances in Information Retrieval Theory, Second International Conference on the Theory of Information Retrieval, ICTIR*, vol. 5766 of *LNCS*, p. 305-312, 2009.

Shtok A., Kurland O., Carmel D., « Using Statistical Decision Theory and Relevance Models for Query-performance Prediction », *International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 259-266, 2010.

Spärck Jones K., « A statistical interpretation of term specificity and its application in retrieval », *Journal of Documentation*, vol. 28, n° 1, p. 11-21, 1972.

Yom-Tov E., Fine S., Carmel D., Darlow A., « Learning to Estimate Query Difficulty : Including Applications to Missing Content Detection and Distributed Information Retrieval », *International Conference on Research and Development in Information Retrieval*, SIGIR, p. 512-519, 2005.

Zhou Y., Croft W. B., « Query performance prediction in web search environments », *International Conference on Research and Development in Information Retrieval, SIGIR*, p. 543-550, 2007.