Prédire la difficulté des requêtes : la combinaison de mesures statistiques et sémantiques

Adrian-Gabriel Chifu

IRIT UMR5505, CNRS
Université de Toulouse, Université Paul Sabatier
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9, France
adrian.chifu@irit.fr

RÉSUMÉ. *La performance d'un Système de Recherche d'Information (SRI) est étroitement liée à la requête. Les requêtes pour lesquelles les SRI échouent sont appelées dans la littérature des « requêtes difficiles ». L'étude présentée dans cet article vise à analyser, adapater et combiner plusieurs prédicteurs de difficulté de requêtes. Nous avons considéré trois prédicteurs: un lié à l'ambiguïté des termes, un basé sur la fréquence des termes et une mesure de répartition des résultats. L'évaluation de la prédiction est basée sur la corrélation entre la difficulté prédite et la performance réelle des SRI. Nous montrons que la combinaison de ces prédicteurs donne de bons résultats. Le cadre d'évaluation est celui des collections TREC7 et TREC8 adhoc.*

ABSTRACT. *The performance of an Information Retrieval System (IRS) is closely related to the query. The queries that lead to retrieval failure are referenced in the literature as "difficult queries". This study aims at analysing, adapting and combining several difficulty predictors. The evaluation of the prediction is based on the correlation between the predicted difficulty and the IRS performance. As predictors, we have considered an ambiguity predictor, the IDF measure and a score distribution measure. We show that combining the proposed predictors, produce good results. The evaluation framework consists in the TREC7 and TREC8 ahdoc collections.*

MOTS-CLÉS : *Recherche d'Information, prédire la performance, difficulté des requêtes, ambiguïté des requêtes, combinaison des prédicteurs, corrélation des mesures*

KEYWORDS: *Information Retrieval, performance prediction, query difficulty, query ambiguity, combined predictors, measure correlation*

1. Introduction

Le succès d'une recherche sur un moteur dépend de nombreuses variables. Certaines requêtes conduisent à l'échec pour la plupart des SRI. Certains travaux ont étudié les caractéristiques de ces requêtes dites « difficiles » (Cronen-Townsend *et al.,* 2002, Harman *et al.,* 2009, Carmel *et al.,* 2010).

Cette difficulté pourrait être causée par l'ambiguïté, la formulation peu claire, le manque de contexte, la nature et la structure de la collection de documents, etc. (Carmel *et al.,* 2006, Mothe *et al.,* 2005). Ainsi, il parait intéressant de traiter les requêtes difficiles de façon spécifique. Pour cela, il est nécessaire de les prédire. Différents prédicteurs de nature hétérogène ont été suggérés dans la littérature.

Les mesures de prédiction de la difficulté peuvent être divisées en deux catégories : des mesures pré-recherche et des mesures post-recherche (Carmel *et al.,* 2010).

La prédiction pré-recherche estime la difficulté de la requête, avant que la recherche ne se déroule. De ce fait, seuls les termes de la requête, ainsi que quelques statistiques prédéfinies pour tous les termes dans des documents, peuvent être utilisés pour la prédiction. Par exemple, on peut considérer le prédicteur naïf de pré-recherche, basé sur la longueur de la requête. On peut penser que plus les requêtes sont longues, plus elles sont faciles, en raison de l'apport contextuel des mots. (He *et al.,* 2004) ont montré qu'il n'existe pas de corrélation entre la longueur des requêtes et la performance des systèmes sur ces requêtes. D'autres prédictieurs de pré-recherche sont basés sur des méthodes linguistiques. (Mothe *et al.,* 2005) ont extrait 16 différentes caractéristiques linguistiques des requêtes (caractéristiques morphologiques telles que le nombre moyen de morphèmes par mot de la requête, caractéristiques syntaxiques, ou semantiques comme le nombre moyen de synsets par mot dans WordNet). Les auteurs ont montré que la plupart des caractéristiques linguistiques ne sont pas fortement corrélées avec les performances des SRI. Hauff (Hauff, 2010a) montre que la *distance sémantique moyenne* des termes de la requête n'est pas non plus un bon prédicteur. Les prédicteurs statistiques de pré-recherche pour la difficulté des requêtes ont été largement étudiés. Deux statistiques de termes sont fréquemment utilisées : la « fréquence inverse » (*inverse document frequency, idf*) et la « fréquence inverse dans la collection » (*inverse collection term frequency, ictf*)(Hauff, 2010b). Le prédicteur « query scope » (*QS*), suggéré par He et Ounis (He *et al.,* 2004), mesure le pourcentage de documents contenant au moins l'un des termes de la requête dans la collection. Zhao (Zhao *et al.,* 2008) mesure la similarité basée sur un modèle d'espace vectoriel entre la requête et la collection. He et al. (He *et al.,* 2008) ont étudié des mesures basées sur la cohérence des requêtes. Les mesures de ce type nécessitent une analyse lourde pendant l'indexation, afin de pouvoir être exploitées au moment de la recherche.

Au contraire des méthodes de pré-recherche, les méthodes de prédiction post-recherche analysent les résultats de la recherche. La qualité de prédiction dépend fortement du processus de recherche, car différents systèmes fournissent différents documents pour une même requête. Les méthodes post-recherches peuvent être classées en trois principaux paradigmes : les méthodes basées sur « Clarity », les méthodes basées sur la robustesse et les méthodes basées sur les distributions des scores. L'approche « Clarity » pour prédire la performance d'un SRI est basée sur la mesure de *cohérence* (clarté) entre la liste des résultats et le corpus. La mesure « Clarity » appliquée par Cronen-Townsend (Cronen-Townsend *et al.,* 2002) est basée sur la divergence KL entre le modèle de langue de l'ensemble de résultats et le modèle de langue de la collection entière. Aslam et Pavlu (Aslam *et al.,* 2007) ont étudié la robustesse de la requête par rapport à l'utilisation de méthodes d'extraction différentes. Ils ont montré que le désaccord entre les listes de résultats en utilisant de multiples modèles de recherche est un indicateur de la difficulté de la requête. Enfin, l'analyse de la distribution des scores représente une manière plus rapide, car il n'est pas nécessaire de traiter les termes des documents. Un prédicteur proposé est le « Weighted Information Gain » (*WIG*) (Zhou *et al.,* 2007) qui mesure la divergence entre le score moyen des documents retrouvés en haut de la liste et de l'ensemble du corpus. Le prédicteur « Normalized Query Commitment » (*NQC*) (Shtok *et al.,* 2009) mesure l'écart type des scores des documents retrouvés, normalisé par le score de toute la collection.

Dans l'étude présentée dans ce papier, nous nous proposons de vérifier la corrélation entre certains prédicteurs de la difficulté et une mesure de difficulté de la requête (précision moyenne de la requête). Nous montrons que la corrélation pour chaque prédicteur considérés individuellement n'est pas élevée. Notre hypothèse est que comme les prédicteurs sont de nature différente, en les combinant, nous pourrions obtenir une mesure plus corrélée avec la difficulté.

L'article est organisé comme suit : dans la section 2, nous présentons l'état de l'art. Dans la section 3, nous décrivons les méthodes que nous proposons. La section 4 présente le cadre d'évaluation. Les résultats obtenus sont discutés dans la section 5. La section 6 conclut le papier.

2. Les mesures proposées

L'écart type (*STD*) est une mesure statistique qui indique le degré de variation d'une variable par rapport à la moyenne. Un STD faible suggère que toutes les données ont tendance à être très proches de la moyenne. Le STD utilise les listes des documents retrouvés par un moteur de recherche pour chaque requête et les scores de ces documents. Le prédicteur STD est une variante de NQC (Shtok *et al.,* 2009), sans normalisation. Pour une requête q et pour les N_q premiers documents retrouvés, STD est calculé par :

$$STD(q) = \left(\frac{1}{N_q} \sum_{i=1}^{N_q} \Big(score(D_q^i) - \frac{1}{N_q} \sum_{j=1}^{N_q} score(D_q^j) \Big)^2 \right)^{\frac{1}{2}}, \quad [1]$$

$score(D_q^i)$ représente le score de l'$i^{ème}$ document retrouvé pour la requête q.

La « Fréquence Inverse » (*IDF*) mesure si un terme donné est rare ou commun par rapport à tous les documents. L'hypothèse est que plus les termes sont importants, plus la requête est facile. L'IDF pour une requête q ($IDF(q)$) peut être calculée par :

$$IDF(q) = \frac{1}{n_q} \sum_{t=1}^{n_q} \log_{10} \left(1 + \frac{N}{N_t} \right), \quad [2]$$

N représente le nombre total des documents dans la collection, N_t le nombre des documents contenant le terme t et n_q est le nombre de termes de la requête q.

Le « Nombre de Sens WordNet » (*WNS*) est un prédicteur linguistique pour la difficulté des requêtes, proposé par Mothe et Tanguy (Mothe *et al.*, 2005). Cette mesure d'ambiguïté de la requête est représentée par le nombre moyen de synsets par mot dans WordNet :

$$WNS(q) = \frac{1}{n_q} \sum_{t=1}^{n_q} senses_t, \quad [3]$$

n_q est le nombre de termes de la requête q et $senses_t$ est le nombre de synsets dans WordNet pour le terme t de la requête q.

Le STD, prédicteur post-recherche, utilise les scores des documents dans la liste retrouvée par le moteur de recherche Terrier[1] pour calculer l'écart type. Les deux autres prédicteurs (le IDF et le WNS) n'ont pas besoin d'effectuer effectivement la recherche pour pouvoir estimer la difficulté d'une requête.

Nous proposons **$\mathbf{COMB_1}$** et **$\mathbf{COMB_2}$**, deux combinaisons linéaires entre les prédicteurs mentionnés. La première prend en compte STD et WNS pour une requête q :

$$COMB_1(q) = \lambda \frac{STD(q)}{max(STD)} + (1 - \lambda) \frac{1}{WNS(q)}, \quad [4]$$

1. http ://terrier.org

$max(STD)$ représente la valeur maximale du STD pour toutes les requêtes de la collection. λ est un coefficient de pondération qui donne une importance différente à STD et à WNS dans le résultat final. On tient compte ici de l'inverse du WNS, parce que l'hypothèse est que plus la requête est ambigüe (WNS grand), moins elle est performante.

La deuxième combinaison de prédicteurs contient les mesures STD et IDF et suit la formule suivante :

$$COMB_2(q) = \lambda \frac{STD(q)}{max(STD)} + (1 - \lambda) \frac{IDF(q)}{max(IDF)}. \quad [5]$$

La combinaison des deux prédicteurs pré-recherche, IDF et WNS a été envisagée mais n'a pas donné de bons résultats.

3. Évaluation

Pour évaluer les prédicteurs proposés, nous avons utilisé les collections de données de la compétition TREC (Text REtrieval Conference). La tâche TREC *ad-hoc* analyse la performance des SRI qui recherchent dans un ensemble statique de documents à partir de besoins d'information (*topics*). Nous avons utilisé deux collections de la tâche *ad-hoc* : TREC7 et TREC8. Pour ces collections, la compétition fournit 528155 (environ 2 GO) documents (articles de journaux) et un ensemble de 50 besoins d'information par collection, en langue naturelle anglaise (topics 351–400 pour TREC7 et topics 401–450).

Pour le prédicteur STD, nous avons utilisé des exécutions construites par Terrier. L'ensemble de documents restitués correspond aux 1000 documents ayant les plus forts scores. Nous avons essayé plusieurs configurations pour les paramètres de Terrier. Pour chaque collection, nous avons choisi le paramétrage qui donne la meilleure précision moyenne (MAP). Ces valeurs de la MAP sont conformes avec la littérature (Zhong *et al.,* 2012).

La meilleure configuration en termes de MAP pour TREC7 est la suivante : une étape d'indexation dans deux pas, avec le BB2 (c = 1) comme modèle de pondération, le modèle d'expansion de requête KL sans paramètres (KLbfree), avec 3 documents pour l'expansion de requête, un terme doit apparaitre dans au moins deux documents afin d'être considéré comme pertinent lors de la reformulation de requête. Enfin, le nombre de termes à rajouter à la requête pour le processus d'expansion de requête est de 10. Les requêtes utilisent les parties description et narrative des topics. Pour TREC8 la configuration des paramètres utilise le modèle de pondération DFRee et seulement le titre comme requête.

Pour le IDF on a besoin de connaitre dans combien de documents un terme t d'une requête se retrouve (N_t, la formule 2). Terrier permet de calculer ces valeurs ; nous avons pris le lexique de termes créé après la radicalisation des termes.

Concernant le WNS, nous avons utilisé l'outil d'étiquetage grammatical de Stanford (POS Tagger). Nous recherchons ensuite dans WordNet le nombre de Synsets existant pour un terme donné, en fonction de son étiquetage grammatical.

Pour analyser la qualité de la prédiction nous avons calculé le coefficient de corrélation de Spearman pour deux variables : le prédicteur et la mesure de difficulté. Comme mesure de difficulté nous avons choisi la moyenne de la MAP de tous les systèmes qui ont participé à TREC7 et TREC8. La littérature du domaine a montré que cette mesure était robuste (Bigot *et al.,* 2011, Carmel *et al.,* 2006). (Carmel *et al.,* 2006) et (Aslam *et al.,* 2007) proposent aussi le système médian en termes de MAP comme mesure de difficulté. Dans la section 4.1 nous allons montrer qu'il est plus robuste de prendre la MAP moyenne, que la MAP d'un seul système. En effet, la variabilité des résultats est très importante en fonction du système considéré.

4. Résultats et discussion

4.1. *La variabilité des corrélations des prédicteurs*

Contrairement à plusieurs études qui utilisent les valeurs de AP d'un seul système pour évaluer leurs prédicteurs (Aslam *et al.,* 2007, Cronen-Townsend *et al.,* 2002), nous proposons d'utiliser la moyenne en termes de AP de plusieurs systèmes dans les calculs. Nous justifions cette approche par la variabilité des corrélations entre les prédicteurs et les MAP de différents SRI. Dans cette section, nous illustrons ce phénomène au travers de l'analyse de la corrélation du prédicteur IDF et de l'AP.

Pour les deux collections de données, TREC7 et TREC8, nous avons retenu 10 systèmes avec des performances qui couvrent la plage de MAP obtenue par l'ensemble des participants officiels. Parmi les 10 systèmes nous avons le meilleur système, le plus faible et aussi le système qui représente notre approche, c'est à dire le système simulé qui aurait obtenu la moyenne des AP par requête, pour tous les participants. Nous avons calculé, pour chaque système, le coefficient de corrélation Spearman et la p-value, entre IDF et la mesure de difficulté en termes de AP. Nous avons choisi le prédicteur IDF comme référence, car sa valeur est indépendante du système utilisé. Les corrélations obtenues se trouvent dans le tableau 1. Les résultats sont triés par ordre descendant, par rapport à la MAP. Le système qui représente la moyenne de tous les systèmes est nommé *AvgMAP*.

Les valeurs pour le coefficient de corrélation et pour la p-value varient beaucoup. Par exemple, pour des MAP entre 0.2609 et 0.3346 les corrélations varient entre 0.22 et 0.34 avec des p-values entre 0.0008 et 0.01857 (TREC8). De plus, pour TREC7 les meilleurs systèmes ont une p-value plus grande que 0.05 (0.10650 et 0.11830), les résultats ne sont donc pas statistiquement significatifs. En ce qui concerne les systèmes avec une MAP faible la confiance sur les corrélations obtenues n'est pas très intéressante, car ce sont des systèmes qui de toutes façons ne sont pas efficaces. Pour une MAP de 0.0287 nous avons une corrélation de 0.5083825 avec une p-value de 0.15610 et pour une MAP très proche (0.0273), nous avons une corrélation de 0.2158393 et une p-value de 0.00167.

TREC7			
Nom	**MAP**	**Corr.**	**p-value**
CLARIT98COMB	0.3702	0.2329948	0.10650
t7miti1	0.3675	0.2505162	0.11830
uoftimgr	0.2755	0.2632893	0.02930
ok7as	0.2614	0.2039376	0.12190
FLab7atE	0.2020	0.2692437	0.04071
AvgMAP	*0.1992*	*0.2986315*	*0.01082*
Brkly24	0.1714	0.3014238	0.01299
APL985L	0.1576	0.2406300	0.03061
KD70000	0.0250	0.1737295	0.00688
dsir07a01	0.0117	0.2579675	0.46870

TREC8			
Nom	**MAP**	**Corr.**	**p-value**
perf-class	0.4726	0.3430094	0.00697
CL99SDopt2	0.3520	0.2228625	0.03048
8manexT3D1N0	0.3346	0.2201786	0.01857
ok8amxc	0.3169	0.2789503	0.00647
ibmg99b	0.2609	0.3445461	0.00080
AvgMAP	*0.2533*	*0.3175586*	*0.00165*
Dm8TFbn	0.1630	0.3768819	0.00010
AntHoc1	0.0287	0.5083825	0.15610
isa25t	0.0273	0.2158393	0.00167
isa25	0.0026	0.3642227	7.57E-06

Tableau 1. *Les corrélations entre le prédicteur IDF et la mesure de difficulté en termes de AP, par système et par collection. La MAP de chaque système est également indiquée*

Cette analyse montre donc que les corrélations obtenues varient de façon très importante en fonction des systèmes choisis comme référence à la corrélation, et cela même si l'on considère des systèmes globalement efficaces en termes de MAP. Cela nous a conduit à choisir plutôt un système moyen comme base au calcul. C'est cette référence qui est utilisée dans la suite.

4.2. *La corrélation basée sur la combinaison des prédicteurs*

Après avoir choisi la moyenne des AP comme mesure de difficulté, nous présentons dans le tableau 2 les corrélations entre cette mesure et les prédicteurs proposés dans cette étude, pour les collections TREC7 et TREC8. Pour les prédicteurs $COMB_1$ et $COMB_2$ présentés dans la section 2, nous avons considéré le paramètre λ qui donne les meilleurs coefficients de corrélation.

Parmi les trois prédicteurs proposés (STD, WNS et IDF) le plus corrélé avec la mesure de difficulté est STD et le plus faiblement corrélé est WNS. Les p-values pour WNS indiquent que les corrélations ne sont pas statistiquement significatives. Par contre, la combinaison entre STD et WNS ($COMB_1$) correspond à une corrélation forte, avec un degré de confiance élevé ($Corr$=0.5965426 et p-value=0.000231, pour TREC8).

Prédicteur	TREC7		TREC8	
	Corr.	p-value	Corr.	p-value
STD	0.5531333	0.03515	0.5446819	0.001353
WNS	−0.1796017	0.08747	−0.2814032	0.05086
IDF	0.2986315	0.01082	0.3175586	0.001649
$COMB_1$	0.4529652	0.01986	**0.5965426**	0.000231
$COMB_2$	**0.5601441**	**0.00567**	0.5849220	**0.000100**

Tableau 2. *Les corrélations entre les prédicteurs proposés et la mesure de difficulté, pour TREC7 et TREC8*

Le prédicteur IDF est plus fortement corrélé que le prédicteur WNS et c'est certainement pour cette raison que la combinaison entre STD et IDF ($COMB_2$) donne les meilleurs résultats.

En ce qui concerne le paramètre λ, la valeur qui produit les meilleurs résultats est 0.78 pour $COMB_1$ et 0.7 pour $COMB_2$. Le prédicteur avec la meilleure performance, STD, apporte plus de 70% dans le calcul du $COMB_2$. Pour $COMB_1$, WNS ayant une corrélation faible, STD prend une pondération plus importante (0.78).

Pour les deux collections (TREC7 et TREC8), le « Clarity score » (Cronen-Townsend *et al.,* 2002) a un coefficient moyen de corrélation de 0.535. Le prédicteur $COMB_2$ que nous proposons permet d'obtenir un coefficient moyen de 0.573, par rapport à AP d'un seul système (non spécifié par les auteurs). Le prédicteur « Clarity score » calcule l'entropie relative entre le modèle de langue des requêtes et le modèle de langue de la collection. Ce processus est lourd en temps de calcul puisqu'il faut calculer les scores de pertinence pour le modèle des requêtes (He *et al.,* 2004). A l'opposé, STD utilise directement les scores des documents retrouvés par le moteur de recherche et IDF utilise les fréquences des termes qui existent dans le lexique créé par le moteur d'indexation, donc le prédicteur $COMB_2$ est plus rapide à calculer.

5. Conclusion

Pour conclure, dans cette étude nous avons montré qu'il est plus robuste de considérer la moyenne de plusieurs systèmes comme mesure de difficulté pour comparer les corrélations avec différents prédicteurs, à cause de la variabilité générée par un seul système. Nous avons en effet montré que pour des systèmes avec une MAP similaire, les coefficients de corrélation avec le même prédicteur de difficulté peuvent être très différents. De plus, pour TREC7, la corrélation entre les AP et le prédicteur de difficulté lorsque le meilleur système est choisi n'est pas significative (p-value).

D'une autre coté, nous avons prouvé que même si les prédicteurs sont de nature hétérogène (linguistique, statistique, de pré-recherche, de post-recherche, etc.), leur combinaison peut produire une corrélation plus forte avec la me-

sure de difficulté. Dans cette étude nous avons considéré deux combinaisons linéaires.

Une première perspective à ce travail concerne l'étude de différentes formes de combinaison des prédicteurs. Nous pensons également agrandir la plage de prédicteurs et leurs combinaisons. Enfin, une autre piste est basée sur le choix du système le plus adapté pour servir de base aux indicateurs post-recherche.

6. Bibliographie

Aslam J. A., Pavlu V., « Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions », *Proc. of ECIR 2007*, vol. 4425 of *Lecture Notes in Computer Science*, Springer, p. 198-209, 2007.

Bigot A., Chrisment C., Dkaki T., Hubert G., Mothe J., « Fusing different information retrieval systems according to query-topics : a study based on correlation in information retrieval systems and TREC topics », *Information Retrieval Journal*, vol. 14, n° 6, p. 617-648, 2011.

Carmel D., Yom-Tov E., « Estimating the Query Difficulty for Information Retrieval », *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 2, n° 1, p. 1-89, January, 2010.

Carmel D., Yom-Tov E., Darlow A., Pelleg D., « What makes a query difficult ? », *Proc. of SIGIR 2006*, ACM Press, p. 390 - 397, 2006.

Cronen-Townsend S., Croft W. B., « Quantifying query ambiguity », *Proc. of HLT 2002*, p. 104-109, 2002.

Harman D., Buckley C., « Overview of the Reliable Information Access Workshop », *Information Retrieval*, vol. 12, n° 6, p. 615-641, July, 2009.

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

Hauff C., Predicting the Effectiveness of Queries and Retrieval Systems, PhD thesis, Centre for Telematics and Information Technology University of Twente, January, 2010b.

He B., Ounis I., « Inferring query performance using pre-retrieval predictors », *Proc. of SPIRE 2004*, Springer Berlin Heidelberg, Padova, p. 43 - 54, 2004.

He J., Larson M., de Rijke M., « Using Coherence-Based Measures to Predict Query Difficulty », *Proc. of ECIR 2008*, vol. 4956 of *Lecture Notes in Computer Science*, Springer, p. 689-694, 2008.

Mothe J., Tanguy L., « Linguistic features to predict query difficulty », *Proc. of SIGIR, Predicting query difficulty - methods and applications workshop*, ACM, p. 7-10, 2005.

Shtok A., Kurland O., Carmel D., « Predicting Query Performance by Query-Drift Estimation », *Proc. of ICTIR 2009*, vol. 5766 of *Lecture Notes in Computer Science*, Springer, p. 305-312, 2009.

Zhao Y., Scholer F., Tsegay Y., « Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence », *Proc. of ECIR 2008*, vol. 4956 of *Lecture Notes in Computer Science*, Springer, p. 52-64, 2008.

Zhong Z., Ng H. T., « Word Sense Disambiguation Improves Information Retrieval », *ACL (1)*, The Association for Computer Linguistics, p. 273-282, 2012.

Zhou Y., Croft W. B., « Query performance prediction in web search environments », *Proc. of SIGIR 2007*, ACM, p. 543-550, 2007.