Mémoire de DEA

Évaluation et Prédiction des Difficultés de Requêtes dans la Recherche Documentaire pour l'Optimisation de Systèmes Interactifs

Jens Grivolla*

19 juin 2001

Dans le but de développer un système interactif de recherche documentaire, ce mémoire explore les possibilités de faire des estimations de performances de recherche en fonction de la requête posée.

Nous avons étudié différents paramètres caractérisants les requêtes et avons trouvé plusieurs scores calculables sur cette base qui permettent de faire de bonnes estimations qui pourrons être utilisées pour déterminer un dialogue à engager avec l'utilisateur.

Les meilleurs résultats ont été trouvés sur des requêtes courtes qui sont typiques dans un grand nombre d'applications, par exemple en utilisant des scores de spécificité permettant de discerner une classe de requêtes pour laquelle les performances des systèmes de recherche sont deux fois supérieures aux autres requêtes.

Néanmoins, nous avons également pu trouver certains paramètres applicables à d'autres types de requêtes, surtout en utilisant des classes de mots, ce qui permet de discerner environ 30% des requêtes pour lesquelles la qualité des documents trouvés est environ 20 à 30% supérieure aux reste des requêtes.

*jens.grivolla@lia.univ-avignon.fr

Table des matières

1 Introduction

1.1 La recherche documentaire

La recherche documentaire est un domaine d'importance croissante depuis déjà un certain nombre d'années. Avec l'omniprésence de bases de documents gigantesques, en particulier l'internet, il y a un besoin de systèmes performants pour rendre ces données utilisables.

Dans les dix dernières années, beaucoup de progrès ont été faits, et les systèmes de recherche dont nous disposons aujourd'hui sont bien supérieurs aux systèmes plus anciens en termes de précision et rappel des documents trouvés.

Cependant les résultats sont loins d'être parfaits et on note une certaine stagnation dans les performances depuis trois ou quatre ans, du moins pour les recherches classiques représentées par la partie *adhoc* des campagnes TREC (voir section 3.2.1). Pour pouvoir améliorer encore les performances on doit donc envisager d'autres approches, en particulier des approches interactives, permettant de préciser et améliorer les requêtes par un dialogue avec l'utilisateur.

1.2 Le but du projet

Il a été noté que la qualité des résultats varie fortement selon la requête, ainsi que d'un système à l'autre. Certains systèmes donnent des résultats supérieurs à d'autres systèmes sur une partie des requêtes, mais inférieures sur d'autres. On constate également que certaines requêtes donnent des mauvais résultats avec tous les systèmes de recherche, d'autres sont par contre faciles à traiter.

En faisant une analyse des requêtes par rapport à la qualité des résultats (sur une base connue, par exemple les campagnes TREC), on peut espérer pouvoir faire des prédictions sur les résultats qui pourront être obtenu, en se basant sur des caractéristiques de la requête traitée.

Cela peut permettre d'une part de choisir le système optimal pour la recherche, et d'autre part, dans le cas ou aucun système disponible ne donnera (probablement) de résultats satisfaisant, d'engager une interaction.

2 État de l'art

Bien que l'importance de l'analyse des requêtes ait été reconnue depuis quelques années, pratiquement aucun travail de recherche approfondi dans ce domaine ne semble avoir été publié.

Dans TREC-5, on note de fortes différences entre les requêtes courtes (titres seulement) et les requêtes entières qui donnent de bien meilleurs résultats [Voorhees and Harman, 1996]. Par rapport à TREC-4, les résultats obtenus sur les requêtes courtes ont déjà fortement évolué et certains systèmes optimisés pour cette tâche obtiennent de très bonnes performances.

Dans ce même article, une mesure de difficulté des requêtes est introduite[1], car une forte chute des performances est visible entre TREC-4 et TREC-5. Des premières investigations pour trouver des facteurs déterminant la difficulté ont montré qu'il n'y avait pas de corrélation significative avec la longueur des requêtes, ni le nombre de documents pertinents. L'article conclut que les requêtes sont plus complexes dans un sens général du terme et qu'il est difficile de prédire la difficulté d'une requête. Il est suggéré que plus de recherche devront être menées dans ce domaine.

Dans TREC-6, il est noté que des requêtes très courtes (titres uniquement) contenant des mots clés bien choisis peuvent donner de bons résultats, alors que l'absence de ces mots (requêtes sans les titres) détériore les résultats [Voorhees and Harman, 1997]. Il est suggéré que selon la longueur de la requête la stratégie de recherche devra être adaptée.

Dans le cadre de TREC-8, Karen Spärck Jones note une forte corrélation entre la performance des systèmes de recherche et la qualité d'information sur la requête ainsi que la "difficulté de la requête" (dans un sens général, non technique) [Spärck-Jones, 1999]. Par contre, d'après les résultats de TREC-7 et TREC-8, les résultats n'utilisant que le titre sont pratiquement aussi bons qu'avec les descriptions longues.

Également suite à TREC-8, Rorvig évalue la difficulté des ensembles de requêtes des différentes années [Rorvig, 1999]. Il réussit à établir des critères permettant des prédictions sur la difficulté de l'ensemble. Par contre, les mesures ne permettent pas d'estimations pour des requêtes individuelles. Rorvig confirme encore que les paramètres simples, comme la longueur de la requête, ne permettent pas de faire d'estimations :

> "Factors that cannot describe query difficulty are : (1) topic components (concepts, narratives, etc.), (2) topic length, (3) and topic construction (creating topics without regard to existing documents vs. the contrary practice). Document uniqueness is the only quantitative measure so far offered. Indeed, topic hardness appears to rest in that zone of phenomena that many can mutually observe, but cannot describe in terms that would eventually permit control."

Un article publié en 1998 par Bank, Over et Zhang traite le problème de classification des requêtes par rapport aux performances obtenues par différents systèmes de recherche utilisant six méthodes statistiques [Banks et al., 1998]. Les résultats ne sont pas satisfaisant, comme il est noté dans l'article : "None of the work we have done using the six approaches discussed here has provided the sort of insights we were seeking [...]".

Un article de Claude de Loupy et Patrice Bellot dans le cadre de LREC 2000 étudie quelques paramètres permettant des estimations de difficulté pour des requêtes individuelles [de Loupy and Bellot, 2000]. Une continuation de ce travail ce trouve dans la thèse de doctorat de Claude de Loupy [de Loupy, 2000]. Un résumé des résultats est présenté dans la section 4 de ce mémoire.

[1] réintroduite, pour être exact – elle avait été utilisée en TREC-2 à titre expérimental

3 Procédé

3.1 Approche générale

Afin de pouvoir déterminer des critères permettant une estimation de difficulté, il faut opposer des critères potentiels à une mesure de difficulté des requêtes. Ceci est fait sous forme d'une visualisation 2-dimensionelle avec sur un axe une mesure du critère caractérisant la requête et sur l'autre axe une mesure de difficulté basée sur les performances des systèmes de recherche.

Comme indicateur supplémentaire on peut calculer et dessiner une ligne correspondant à la difficulté moyenne par rapport au critère. Dans le cas de mesures discrètes pour le critère d'estimation et d'un nombre suffisant de requêtes pour chaque valeur de la mesure la moyenne est évidente, dans les autres cas il est nécessaire de regrouper les requêtes en classes selon le critère utilisé. Des expériences ont été faites d'une part avec des classes formées par découpage régulier (linéaire) entre les valeurs minimales et maximales de la mesure (avec un nombre de classes fixé de manière à réduire les effets aléatoires) et d'autre part avec des classes de taille (nombre de requêtes) fixe afin d'avoir un nombre suffisamment significatif de requêtes dans chaque classe. La deuxième méthode de classification s'est avérée bien supérieure à la première et a donc été utilisée comme indice pertinent qui permet de facilement reconnaître des corrélations entre le critère d'estimation et la difficulté des requêtes.

Alternativement on peut utiliser une approximation par des courbes béziers (une option inclue dans gnuplot) ce qui donne des résultats plus lisses. La meilleure visualisation est obtenue en calculant les courbes béziers sur la base d'une classification antérieure.

Dans quelques cas, le coefficient de corrélation linéaire[2] a été calculé comme indice supplémentaire, en particulier pour démontrer des différences entre différents critères ou entre différents ensembles de requêtes.

Pour évaluer la qualité d'un paramètre on s'intéresse d'une part au fait que la moyenne de difficulté montre une dépendance de ce critère (de manière linéaire ou sous une autre forme, par exemple gaussienne), une dépendance qui est reflété en partie par le coefficient de corrélation, mais aussi à la variance des performances en fonction du critère.

Même si en moyenne les performances ne dépendent pas fortement du paramètre, il peut y avoir un rapport entre le paramètre et la distribution de la difficulté des requête permettant de faire certaines prédictions. Par exemple, le seuil de performance minimale ou maximale peut varier selon le paramètre observé.

Une évaluation permettant de discerner de telles correspondances peut assez bien se faire de manière visuelle mais est difficilement formalisable par des critères mathématiques, surtout parce que toute régularité peut éventuellement être exploitée.

---

[2] $r(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y}$ où s_x et s_y sont les écarts-types de x et y : $s_\alpha^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i-\overline{\alpha})^2$

3.1.1 Définition de la difficulté

Selon l'utilisation prévue des estimations la difficulté d'une requête peut se définir de différentes manières. Il semble raisonnable d'appeler difficile des requête pour lesquelles la précision, c'est-à-dire le pourcentage de documents pertinents parmi les documents trouvés par le système de recherche, est faible.

Un score intéressant est la *précision moyenne*, qui tient compte aussi bien du rappel que de la précision. Pour chaque document pertinent dans le corpus (donc chaque document à trouver) on considère la précision correspondant à sa position dans la liste des documents rapportés, donc si le document se trouve en position n on considère la précision sur les n premiers documents. Si le document pertinent n'a pas été trouvé on prend 0. La précision moyenne est la moyenne arithmétique de ses valeurs.

Un autre score, qui diminue l'influence de l'ordre des réponses est la *précision relative (R-Prec)*, la précision sur les n document classés premiers, avec n le nombre de documents pertinents contenus dans le corpus.

Pour des analyses combinant les scores de plusieurs systèmes, la moyenne des précisions peut être utilisée. La plupart des graphes dans ce document montrent la moyenne (sur les systèmes) des précisions moyennes. Un aspect intéressant (mais qui n'a pas été traité dans ces expérimentations) serait également la variance des performances des différents systèmes (dans l'objectif d'une classification de difficulté plus nuancée ou pour l'analyse des différences entre les stratégies de recherche).

Pour des utilisations particulières on peut s'intéresser à d'autres scores de difficulté, comme la précision sur les n premiers documents, avec n fixe choisi selon les besoins[3], et en particulier aux cas ou le ou les systèmes échouent complètement (ont donc une précision de 0). Pour combiner les résultats de plusieurs systèmes on peut calculer le pourcentage de tels échecs complets. Ce score est intéressant parce que en absence de documents pertinents on ne peut pas appliquer de mécanismes de *relevance feedback*[4] pour augmenter le nombre de documents pertinents retournés et on doit donc utiliser d'autres méthodes (interactives). Ce type de score ne s'est pas avéré idéal dans cette configuration expérimentale. En revanche, il pourra se réveler utile lors de futures expérimentations.

De manière générale, ces différents scores sont très fortement corrélés (comme on peut l'attendre), et le choix du score utilisé ne semble pas beaucoup influencer les résultats. Dans des applications particulières, on aura tout de même intéret à utiliser la mesure reflétant le plus directement les besoins dans le contexte d'utilisation.

3.2 Mode de travail et outils utilisés

L'extraction des données (requêtes et scores de performance) à partir des fichiers fournis par le NIST[5] à la suite des campagnes TREC, ainsi que les calculs de score des

[3] p.ex. $n = 5$ ou $n = 30$ selon l'attente de l'utilisateur qui dans la plupart des situations ne s'intéresse qu'aux premiers documents affichés

[4] interactif ou non

[5] `http ://www.nist.org`

différents paramètres sur les requêtes ont été réalisés par des programmes écrits en Perl[6]. Les résultats sous forme de fichiers pour gnuplot (avec et sans calcul de moyenne, etc.) ainsi que comme tableau complet comprenant tous les scores calculés ont ensuite été visualisés et ont servi au calcul de corrélation.

La recherche de paramètres potentiels s'est faite à l'aide d'une liste des requêtes classée par ordre de difficulté (créée à l'aide d'un programme Perl). Ceci permet de discerner les différences entre les requêtes faciles et difficiles et de développer des hypothèses sur l'influence de certaines caractéristiques des requêtes. L'efficacité de chaque paramètre trouvé a ensuite été vérifiée à l'aide du calcul d'un score représentant ce paramètre et la visualisation de la correspondance de ce score avec la difficulté des requêtes.

Cette approche diffère d'autres approches plus basées sur des analyses statistiques (comme celles appliquées (sans grand succès) dans [Banks et al., 1998]) par le fait d'inclure une phase plus intuitive, faisant confiance à nos capacités humaines de percevoir des régularités dans les requêtes, afin de déterminer des propriétés typiques des requêtes. Celles-ci serviron ensuite dans uns système de classification et de décision automatique (par exemple utilisant des arbres de décisions basés sur ces critères).

3.2.1 Les campagnes TREC

Toutes les expérimentations présentées dans ce rapport ont été effectuée sur la base des requêtes et des résumés des performances de recherche issus des campagnes TREC (*Text REtrieval Conference*).

Depuis 1992, le NIST organise annuellement ces campagnes d'évaluation qui permettent aux équipes travaillant dans le domaine de recherche documentaire de tester leurs systèmes sur un corpus de 500 000 documents et 50 requêtes chaque année, bénéficiant d'une évaluation manuelle des résultats.

TREC propose différentes disciplines, dont la plus classique (sur laquelle sont basées ces expérimentations) est le *adhoc*, la recherche non-interactive automatique à partir de requêtes écrites en langage naturel.

Les campagnes TREC ont permis de travailler sur une relativement grande quantité de données, mais présentent aussi quelques difficultés dues à l'évolution d'année en année des systèmes participants et de la définition de l'évaluation.

Pour disposer de suffisamment de données, les résultats des différentes années ont été mélangés dans une grande partie des expériences, alors que les systèmes participants ont changé avec chaque campagne.

De grands changements dans la construction des requêtes, en combinaison avec des petites pannes[7], ainsi qu'une documentation limitée pour certains aspects[8] n'ont pas toujours facilité le travail.

[6] `http ://www.perl.org`

[7] des requêtes mal formulées en TREC-6 (mots clés n'apparaissent que dans le titre, mais il y a des test sur la description seule)

[8] p.ex. concernant les parties des requêtes utilisées par les différents systèmes

3.2.2 WordNet

Pour les mesures de sens, synonymes, hyponymes, etc. la base de donnée lexicale WordNet[9] (version 1.6) a été utilisé. Elle contient des informations sur plus de 120 000 mots de la langue anglaise.

4 Les paramètres examinés

4.1 Constatations générales

En regardant les requêtes les plus faciles et les plus difficiles, on remarquera que dans les deux groupes les requêtes sont très variées. Comme le montrent aussi les analyses suivantes plus détaillées de certains paramètres, il n'y a pas de caractéristique qui à elle seule permette de faire des estimations sur la difficulté des requêtes.

On constatera également que la difficulté en termes de performances de recherche dépend surtout des attentes de l'utilisateur, du *besoin en information*, qui ne se reflètent pas toujours directement dans les requêtes mais sont à la base de l'évaluation des résultats livrés par le système de recherche. Dans le cas des campagnes TREC, cet aspect est visible dans le narratif fourni avec les requêtes et dans certains cas utilisé également pour la recherche.

Le narratif est difficile à exploiter de manière automatique car il ne fournit généralement pas de mots clés utile à la recherche. Il représente habituellement un approfondissement de la requête et permet surtout d'identifier les documents *non* souhaités par l'utilisateur.

Sur la base de TREC on peut essayer d'analyser le rapport entre la demande explicite (titre et description) et les attentes plutôt implicites (narratif) de l'utilisateur, ainsi que les performances des systèmes de recherche (évaluées en tenant compte des attentes implicites). Ceci doit être fait sur des sous-ensembles des requêtes (à partir de TREC 6), en regardant par exemple les performances obtenues en n'utilisant que les titres (ou alors titres et descriptions), et n'a pas encore été approfondi dans le cadre de ces études.

Les paramètres trouvés sont en partie fortement corrélés. Une fois une base de paramètres établie, cet aspect devra être étudié afin d'éviter dans une application pratique des calculs coûteux à faible apport d'information. Des techniques comme l'analyse en composantes principales (ACP) pourront être utilisées pour définir le vecteur de paramètres d'estimation optimal.

4.2 Longueur de la requête

Confirmant les constatations mentionnées dans certains articles [Rorvig, 1999, Spärck-Jones, 1999, Voorhees and Harman, 1997], les expérimentations n'ont pas montré de corrélation entre la longueur des requêtes et les performances des systèmes de recherche.

Ceci contredit des expérimentations faites préalablement entre autres par Lu et Keefer[Lu and Keefer, 1994], ce qui suggère que selon le système utilisé, la longueur

[9] `http ://www.cogsci.princeton.edu/~wn/`

des requêtes peut avoir un effet sur les performances, mais que ce n'est pas le cas sur l'ensemble des systèmes participants à TREC.

Il est important de noter que ce n'est qu'à partir de TREC 5 que les recherches ont été faites sur différentes versions des requêtes (titre uniquement, description uniquement, et requête entière) et que donc tous les résultats analysés pour des requêtes très courtes proviennent de ces dernières années, alors que des stratégies d'expansion de requête étaient déjà très avancées. Un système moins performant à ce niveau pourrait réagir plus sensiblement à la longueur des requêtes. Malheureusement, les investigations n'ont pas pu être approfondies, ne disposant que des statistiques de TREC.

Le fait qu'aucune corrélation n'aie été visible non plus en n'analysant que les données TREC 2, 3 et 4[10] par rapport aux résultats obtenus dans d'autres expérimentations s'explique éventuellement par le fait que les systèmes utilisés dans TREC 4 étaient d'une manière générale plus performant que les systèmes plus anciens ou encore que la collection de requêtes était moins dure[11].

Les variations des longueurs des requêtes à l'intérieur d'un ensemble de requêtes (p.ex. entre 19 et 71 mots dans TREC 3) ne produit apparemment pas les mêmes effets que ceux obtenus dans les expérimentations par un raccourcissement artificiel des requêtes. Ceci doit s'expliquer par le fait que des requêtes "naturellement" courtes sont souvent des requêtes qui peuvent assez bien être exprimées par un nombre réduit de mots clés, ce qui compense l'aspect d'avoir éventuellement moins d'information par des mots différents dans la requête que dans des requêtes longues. De plus, les requêtes les plus courtes utilisées dans les premiers TREC sont toujours bien plus longues que celles construites dans les expérimentations.

4.3 Nombre de sens des mots de la requête

Ce paramètre a été utilisé avec beaucoup de succès pour l'estimation de difficulté de requêtes dans [de Loupy and Bellot, 2000, de Loupy, 2000].

Par contre, ayant initialement travaillé indépendemment de ces résultats, nous n'avions pas constaté une telle corrélation. En approfondissant les investigations, il s'est montré que les résultats obtenus par de Loupy et Bellot étaient très fortement liés d'une part à la base de requêtes traitée (ils avaient travaillé comme moi sur des données de TREC, mais en n'utilisant que les la version ultra-courte des requêtes de TREC 6) et d'autres part de leur manière particulière d'attribuer un score en fonction du nombre de sens des mots.

Faisant des tests en utilisant le nombre moyen, maximal et minimal de sens des mots contenus dans la requêtes, aucune corrélation utilisable pour faire des prédictions sur la dureté des requêtes n'était évidente.

La méthode proposée par Claude de Loupy et Patrice Bellot est la suivante :

> "Le score de la requête est incrémenté de 1 pour chaque terme ayant un seul sens (un mot inconnu de WordNet est considéré comme ayant un seul sens).

[10] TREC 4 contenant des requêtes plus courtes que TREC 2 et 3, bien que nettement plus longues que la version courte des requêtes dans les campagnes suivantes

[11] la difficulté d'ensembles de requêtes a été évaluée dans [Rorvig, 1999]

Si tous les termes ont un seul sens, le score est de 3. Si tous les termes ont 3 sens ou plus, le score est de -1 et si un mot n'est pas présent dans la base, le score est de -2."[12]

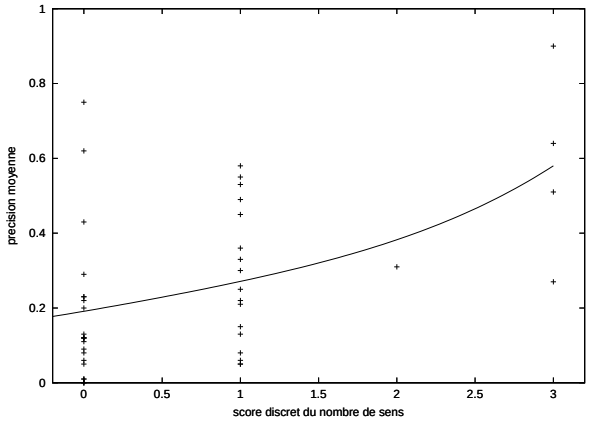

FIG. 1: score discret du nombre de sens – requêtes courtes de TREC 6

Il est évident qu'une telle manière d'attribuer un score est fortement liée à la condition que les requêtes soient très courtes (1 à 4 mots dans ce cas). Dans cette situation on peut supposer qu'ils s'agit de mots clés significatifs pour la recherche, ce qui n'est pas le cas pour des requêtes plus longues, contenant un narratif en langage naturel.

En effet, on a pu reproduire les résultats en utilisant (pratiquement) la même méthode sur la même base de requêtes (avec une corrélation linéaire de 0.522), mais il ne semble pas qu'il y ait de moyen direct d'appliquer ce score de manière utile à d'autres types de requêtes (corrélation linéaire de 0.122 sur l'ensemble des requêtes, inclus les requêtes longues).

Un résultat plus étonnant est que de légères variations dans la manière de calculer le score (p.ex. de soustraire des points pour des mots ayant beaucoups de sens) diminue dramatiquement la corrélation, ce qui montre une fois de plus la difficulté d'établir des paramètres utilisables.

On a également dû constater que les résultats sont beaucoup moins nets sur des requêtes de construction semblable de TREC 7 et 8 (la corrélation linéaire est de 0.185 sur TREC 8, 0.158 sur TREC 7). Il semble donc que le score utilisé est très spécifiquement optimisé sur un certain corpus de requêtes. Étant donné qu'on note tout de même une relation entre ce score et la difficulté des requêtes (d'ailleurs plus sur TREC 8 que sur TREC 7), il reste l'espoir qu'on puisse utiliser ce paramètre, surtout en combinaison avec d'autres critères.

On pourra essayer d'optimiser le score sur une base plus large (et donc éviter de se lier trop à un corpus limité), mais cela n'apportera sans doute pas de grands gains. Toutefois, en combinaison avec d'autres paramètres, ce score sous sa forme actuelle fournit une très bonne base.

[12] De plus, ils catégorisent les requêtes en *faciles*, *moyennes* et *difficiles*, mais cela n'a pas d'importance dans ce contexte.

Par contre, il sera nécessaire de traiter les cas pour lesquels le score n'est pas du tout utilisable actuellement, en particulier les requêtes longues. Le problème principal paraissant être le bruit introduit par des mots sans importance dans la requête, on devra essayer de sélectionner les quelques mots clés caractérisant le plus la requête pour ensuite calculer le score (éventuellement légèrement adapté) sur ces mots. Une première approche serait d'utiliser l'IDF pour estimer la contribution des mots de la requête, des analyses plus fines seront évidemment envisageables.

4.4 Nombre de synonymes des mots de la requête

La situation se présente de manière semblable au nombre de sens, les statistiques sur l'ensemble des systèmes et des années ne montrant aucune corrélation utilisable.

Alors qu'il est assez évident qu'un mot n'ayant qu'un seul sens aura plus de chance de permettre de bons résultats de recherche par rapport à des mots plus ambigus, l'effet que peuvent avoir les synonymes est moins clair. Un système ancien aura certainement des problèmes de rappel s'il ne trouve pas les documents pertinent contenant des synonymes du mot de la requête, mais les systèmes actuels utilisent souvent des méthodes d'expansion de requêtes comme le *blind relevance feedback* [Walker and de Vere, 1990], permettant de rajouter des mots sémantiquement proches des termes de la requête (synonymes, hyponymes et autres).

Il n'est donc pas clair si le nombre de synonymes des mots de la requête peut encore servir comme indicateur de difficulté. Les expérimentations ont montrées que ce paramètres n'est pas utilisable de manière brute, mais comme on peut voir dans le cas du nombre de sens, il peut tout de même y avoir un moyen d'utiliser ce critère.

Il n'est pas clair pour l'instant comment calculer un score utile à base des synonymes seuls. Par contre il sera intéressant d'analyser si par exemple la combinaison du nombre de synonymes et du nombre de sens peut donner de meilleur résultats que le nombre de sens seul.

Il est très probable que le nombre de synonymes ne soit utilisable (dans un premier temps) que sur des requêtes courtes. Les expérimentations ayant été faites en grande partie sur l'ensemble complet des requêtes (inclus les requêtes longues), il reste donc l'espoir d'un potentiel non encore exploité et qui mérite une attention plus détaillée.

4.5 Classes de mots

Les résultats les plus convainquants obtenus jusqu'à présent sur l'ensemble complet des requêtes dans ces expériences sont basés sur l'utilisation de dictionnaires spécifiques pour certaines classes de mots.

Les performances des systèmes de recherche sont très nettement supérieures sur des requêtes contenant des mots de certaines catégories par rapport au reste des requêtes, ce qui devrait permettre des prédictions assez bonnes sur un certain nombre de requêtes.

Pour les classes de mots analysées, seulement environ 10 à 15% des requêtes contiennent des mots d'une certaine classe, environ 30% contiennent des mots d'au moins une classe.

En utilisant de bons dictionnaires spécialisés et incluant d'autres classes on pourrait sans doute arriver à une couverture supérieure.

Il est clair que l'estimation de difficulté des requêtes devra se faire sur une combinaisons de plusieurs paramètres. Pour cela il n'est pas absolument essentiel que chaque critère individuel utilisé recouvre un très grand pourcentage des requêtes. Les classes de mots peuvent donc être extrêmement utiles comme indicateurs positifs pour les requêtes concernées.

4.5.1 Noms de pays

Un très bon indicateur qu'une requête permettra des bonnes performances et l'occurrence de *noms de pays* dans la requête. Sur le corpus analysé, 13% des requêtes contiennent de tels mots, et on constate des performances 22% supérieures au reste des requêtes (R-Prec = 0.55 par rapport à 0.45, précision moyenne = 0.27 par rapport à 0.22).

4.5.2 Noms de personnes

Les noms de personnes ne sont pas tout à fait aussi puissant comme indicateur (performances environ 15% supérieures, avec une grande variance) et ne recouvrent qu'environ 5% des requêtes. On pourrait envisager (avec des pertes de couverture additionnelles) de se limiter aux personnes célèbres, les tests faits jusqu'à présent ne tiennent pas compte de cet aspect.

4.5.3 Abréviations

L'occurrence d'abréviations dans les requêtes est également fortement corrélée (positivement) avec les performances de recherche (R-Prec = 0.55 par rapport à 0.46, précision moyenne = 0.34 par rapport à 0.22) et distingue environ 9% des requêtes.

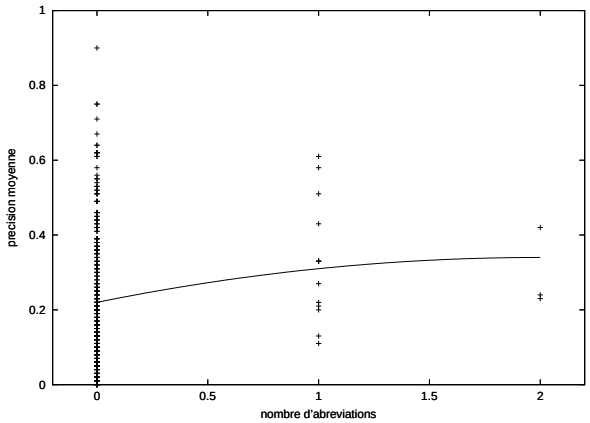

FIG. 2: occurrences d'abréviations – toutes requêtes, TREC 2 à 6

La grande variance des performances à l'intérieur des classes limite un peu la qualité des estimations sur cette base, mais l'occurrence d'abréviations est tout de même un indicateur très utile.

L'avantage de ce paramètre par rapport à d'autres classes de mots est qu'il est facile de reconnaître les abréviations automatiquement (sans dictionnaire) d'une manière assez sure.

4.5.4 Autres classes

À condition d'avoir des dictionnaires appropriés, d'autres classifications semblent prometteuses, en particulier les termes techniques, médicaux ou d'autres termes très spécialisés, ainsi que les événement historiques, du moins pour les recherches dans un corpus non spécialisé.

4.6 Spécificité

La spécificité des mots de la requête est un indicateur important pour l'estimation de difficulté. Plus les termes sont spécifiques, plus des documents contenant ces termes ont de chances de répondre aux besoins de l'utilisateur, alors que des termes très génériques ne pourront souvent pas suffisamment déterminer les documents pertinents.

4.6.1 Nombre d'hyponymes des mots de la requête

Une mesure de spécificité est le nombre d'hyponymes, c'est à dire de termes plus spécifique englobés par le mot donné. Donc plus un mot a d'hyponymes, moins il est spécifique. Ces propriétés ont d'ailleurs été prises en compte entre autre dans [Jourlin et al., 2000].

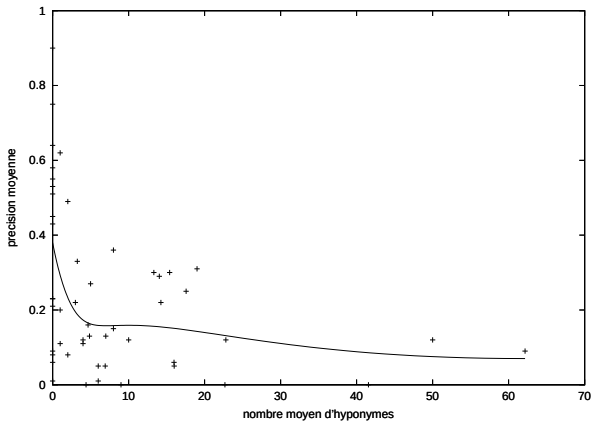

FIG. 3: nombre moyen d'hyponymes – requêtes courtes de TREC 6

Comme beaucoup d'autres mesures, le nombre d'hyponymes ne semble être utilisable que pour des requêtes très courtes, alors que sur l'ensemble de toutes les requêtes il n'y a pas de corrélation visible avec la dureté des requêtes. Il sera éventuellement possible d'élaborer un score plus sophistiqué sur la base de ce paramètre (semblable aux score

sur le nombre de sens) qui pourrait même être adaptable à des requêtes plus longues (en réduisant le bruit introduit par des mots sans importance).

4.6.2 Fréquence des termes (score discret)

Claude de Loupy et Patrice Bellot ont montré que la fréquence des termes peut être un bon indicateur de la difficulté d'une requête, du moins pour les requêtes courtes. Ils proposent de calculer les score suivant[13] :

> "Un mot apparaissant dans moins de n documents (nous avons choisi, de façon empirique, $n = 1000$) est considéré comme *un point de fixation* pour la réponse à la requête[...]
>
> Pour chaque mot de la requête apparaissant moins de n fois, le score est incrémenté de 1. Si tous les mots de la requête sont dans ce cas, le score de la requête est de 3 qui représente le score maximal que peut obtenir une requête. Si l'un des mots ne se trouve pas dans la base, la requête se voit affecter arbitrairement le score minimal de -2. Dans le cas où tous les mots apparaissent plus de 10000 fois dans la base, le score est égal à -1 (valeur arbitraire). Dans tous les autres cas, le score est de 0."

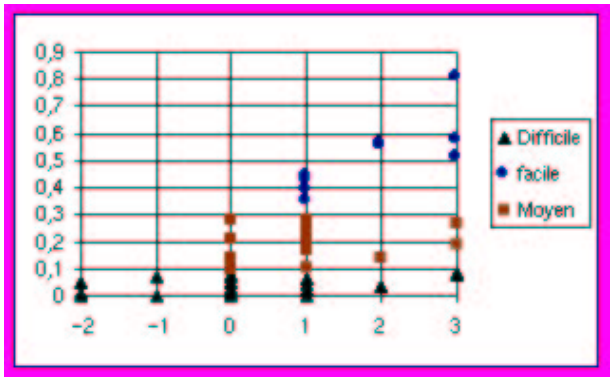

FIG. 4: score discret utilisant la fréquence des termes – requêtes courtes de TREC 6

Comme le score discret pour le nombre de sens, ce score ne peut pas être appliqué directement à des requêtes plus longues, mais il semble être plus facilement adaptable. Il n'a pas encore été testé sur d'autres ensembles de requêtes que les requêtes courtes (titre uniquement) de TREC 6.

4.6.3 IDF

Une mesure très utilisée dans la recherche documentaire pour estimer l'apport d'information d'un mot est l'IDF (*inverse document frequency* :

$$IDF(\lambda) = -\log\left(\frac{documents\ contenant\ x}{nombre\ total\ de\ documents}\right)$$

[13] citation raccourcie

En combinaison avec d'autres mesures, l'IDF peut servir à développer de bons scores pour la prédiction de difficulté.

4.6.4 Coefficient de Bookstein

Une autre mesure de la spécificité d'un terme, le coefficient de Bookstein $B(\lambda)$, est défini ainsi :

$$B\left(\lambda\right) = \frac{N\left(\lambda\right)}{E\left(\lambda\right)}$$

où $N(\lambda)$ est le nombre de documents contenant λ et $E(\lambda)$ le nombre théorique qui devraient contenir λ si la répartition était de type aléatoire.

$$E\left(\lambda\right) = D \cdot \left(1 - \left(1 - \frac{1}{D}\right)^{T(\lambda)}\right)$$

où D est le nombre total de documents dans la base et $T(\lambda)$ le nombre total d'occurrences de λ dans la base (compté autant de fois qu'il apparaît dans un document pour tous les documents).

Le coefficient de Bookstein n'a pas été étudié individuellement, mais est utilisé dans le score 4.7.

4.7 Score de de Loupy & Bellot

Claude de Loupy et Patrice Bellot ont proposé des scores combinant plusieurs paramètres :

$$score1(Q) = \max_{x \in Q}\left(\frac{IDF(x)}{S(x)}\right)$$

où $S(x)$ et le score sur le nombre de sens présenté dans 4.3,

$$score2(Q) = \max_{x \in Q}\left(\frac{IDF(x)}{B(x) \cdot S(x)}\right)$$

où $B(x)$ est le coefficient de Bookstein.

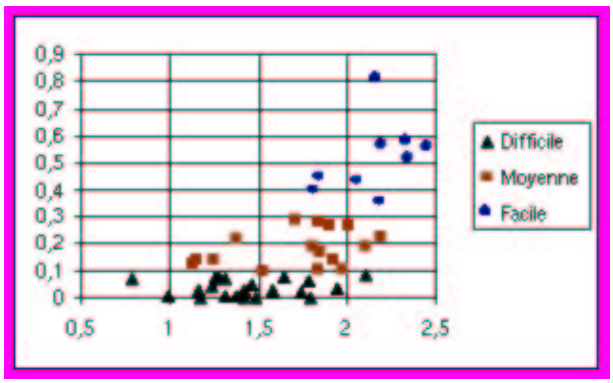

FIG. 5: Score basé sur IDF, nombre de sens et coefficient de Bookstein ($score2(Q)$)– requêtes courtes de TREC 6

Ces scores (surtout la deuxième version) donne de très bons résultats pour les requêtes courtes issue de TREC 6, il devra être modifié pour être appliquable à d'autres types de requêtes. Aussi n'a t'il pas encore été validé sur d'autres ensembles de requêtes courtes.

4.8 Négations dans le narratif

Dans les campagnes TREC, les attentes de l'utilisateur sont le mieux exprimé dans le narratif, qui concrétise la demande formulée dans titre et description de la requête. Les expériences ont montré que l'apport du narratif à la recherche automatique est très limité, pourtant il permet à l'évaluateur humain de mieux déterminer la pertinence des documents trouvés.

De brèves expérimentations ont été faites pour essayer de trouver des critères formels qui décrivent l'influence que les attentes exprimées dans le narratif ont sur les performances, en particulier en observant l'occurrence de négations (*'not relevant', ...*) dans le narratif qui indiquerait que certains documents correspondant à la requête au niveau de titre et description ne sont, en fait, pas pertinent. Les résultats n'en sont pas encore convainquants, mais cette voie paraît prometteuse (plus pour une meilleure compréhension du problème que pour l'application pratique).

5 Conclusions et Perspectives

Les expérimentations ont montré qu'on peut obtenir des estimations utilisables valables pour tous les systèmes de recherche pour certaines requêtes. Par contre, il ne semble pas être possible de trouver de paramètres caractérisant suffisamment bien la difficulté des requêtes qui soient appliquables de manière générale.

Dans la suite, les recherches devront donc être plus concentrées sur des sous-ensembles spécifiques des requêtes (par exemple les requêtes courtes) selon l'application prévue. Les requêtes des campagnes TREC sont très variées, alors que dans la pratique on sera souvent confronté à des requêtes d'un type particulier. Pour la plupart des applications il n'est donc pas nécessaire de trouver des critères aussi généraux que ceux qu'on peut espérer découvrir en se basant sur TREC.

D'autre part, il sera bon de se concentrer sur un (ou un nombre très limité) de systèmes de recherche. Les études faites jusqu'ici utilisaient généralement des moyennes de scores de différent systèmes ce qui introduit un certain 'bruit'. Les résultats existants des campagnes TREC ne permettent pas bien d'analyser les faiblesses et points forts de systèmes individuels car les systèmes participants changent d'une année sur l'autre.

Pour avoir suffisamment de données (une base suffisamment grande de requêtes traitées) il sera nécessaire d'appliquer les systèmes à un grand corpus de requêtes (plusieurs années de TREC ou des requêtes plus spécifiques). Ceci sera possible grâce à deux systèmes de type différent disponibles au LIA. Sur cette base on pourra également faire des comparaisons détaillées des performances des systèmes pour éventuellement non seulement faire des estimations de performances afin d'engager en un dialogue avec l'utilisateur, mais aussi choisir la stratégie de recherche adéquate pour chaque requête

et ainsi améliorer les performances. Comme a été dit dans la section 3.2, certaines techniques de classification automatique des requêtes présentées dans [Banks et al., 1998] pourront être utilisées pour faciliter une analyse approfondie (non automatique).

Il sera également intéressant d'utiliser les documents trouvés par le système de recherche dans une première passe pour faire les estimations. En particulier, on pourrait essayer de classer les documents d'après des critères de ressemblance ou même des analyses sémantiques plus fines afin de détecter d'éventuelles ambiguïtés dans les requêtes et de permettre de concrétiser la demande sur la base des sujets traités dans les différentes classes de documents.

Nous avons dans ces travaux exploré diverses voies pour déterminer les paramètres les importants pour le dialogue. Nous espérons par la suite developper un système de catégorisation automatique de requêtes et d'interaction basé sur les arbres de décisions [Kuhn and de Mori, 1995].

En plus de l'importance immédiate en particulier pour le développement de systèmes interactifs, un grand intérêt de ce sujet de recherche est dans la diversité des approches envisageables. Il est ainsi possible de regrouper autour d'un objectif principal différents domaines de recherche dont les résultats seront certainement utilisables également dans d'autres contextes d'application en recherche documentaire et traitement de langue naturelle.

Références

[Banks et al., 1998] Banks, D., Over, P., and Zhang, N.-F. (1998). Blind men and elephants : Six approaches to trec data. *Information Retrieval*.

[Crestani et al., 1997] Crestani, F., Sanderson, M., Theophylactou, M., and Lalmas, M. (1997). Short queries, natural language and spoken document retrieval : Experiments at glasgow university. In *TREC-6 Proceedings*. http ://trec.nist.gov/pubs/trec6/t6_proceedings.html.

[de Loupy, 2000] de Loupy, C. (2000). *Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire*. PhD thesis, Laboratoire Informatique d'Avignon.

[de Loupy and Bellot, 2000] de Loupy, C. and Bellot, P. (2000). Evaluation of document retrieval systems and query difficulty. In *Acte de "Using Evaluation within HLT Programs : Results and Trends"*, pages 31–38, Athènes, Grèce. http ://www.limsi.fr/TLP/CLASS/ClassD43.pdf.

[Jourlin et al., 2000] Jourlin, P., Johnson, S. E., Spärck-Jones, K., and Woodland, P. C. (2000). Spoken document representations for probabilistic retrieval. *Speech Communication*.

[Kuhn and de Mori, 1995] Kuhn, R. and de Mori, R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

[Lu and Keefer, 1994] Lu, X. A. and Keefer, R. B. (1994). Query expansion/reduction and its impact on retrieval effectiveness. In *TREC-3 Proceedings*. http ://trec.nist.gov/pubs/trec3/t3_proceedings.html.

[Rorvig, 1999] Rorvig, M. (1999). Retrieval performance and visual dispersion of query sets. In *TREC-8 Proceedings*. http ://trec.nist.gov/pubs/trec8/t8_proceedings.html.

[Spärck-Jones, 1999] Spärck-Jones, K. (1999). Summary performance comparisons trec-2 through trec-8. In *TREC-8 Proceedings, Appendix B*. http ://trec.nist.gov/pubs/trec8/t8_proceedings.html.

[Voorhees and Harman, 1996] Voorhees, E. M. and Harman, D. (1996). Overview of the fifth text retrieval conference. In *TREC-5 Proceedings*. http ://trec.nist.gov/pubs/trec5/t5_proceedings.html.

[Voorhees and Harman, 1997] Voorhees, E. M. and Harman, D. (1997). Overview of the sixth text retrieval conference. In *TREC-6 Proceedings*. http ://trec.nist.gov/pubs/trec6/t6_proceedings.html.

[Walker and de Vere, 1990] Walker, S. and de Vere, R. (1990). Improving subject retrieval in online catalogues : 2. relevance feedback and query expansion. *British Library Research Paper 72*.

[Wilkinson et al., 1995] Wilkinson, R., Zobel, J., and Sacks-Davis, R. (1995). Similarity measures for short queries. In *TREC-4 Proceedings*. http ://trec.nist.gov/pubs/trec4/t4_proceedings.html.