Passage à l'échelle

Une méthodologie d'étude de l'influence du volume de collection sur les modèles de Recherche d'Information

Amélie IMAFOUO — Michel BEIGBEDER

Ecole Nationale Supérieure des Mines de Saint-Etienne
Centre Génie Industriel et Informatique (G2I)
158 Cours Fauriel
42023 Saint-Etienne Cedex 2, FRANCE
{imafouo, beigbeder}@emse.fr

RÉSUMÉ. *Peu de travaux en Recherche d'Information (RI) ont jusqu'alors abordé les questions d'efficience et d'efficacité des systèmes de RI dans le contexte du passage à l'échelle dans la taille des corpus. Nous proposons une démarche expérimentale reproductible (pour l'étude de l'influence du passage à l'échelle sur les modèles de RI) basée sur la construction d'une collection sur laquelle une caractéristique C donnée est la même quelle que soit la portion de collection selectionnée. Cette nouvelle collection dite "uniforme" peut être decoupée en sous-collections qui sont des " échantillons " de taille croissante de la collection entière et sur lesquelles des propriétés de modèles de RI sont étudiées. Nous appliquons notre démarche sur la collection WT10G de TREC9 avec comme caractéristique C la répartition des documents pertinents et comme propriétés les métriques d'évaluation de RI.*

ABSTRACT. *Few works in Information Retrieval (IR) field tackled the questions of IR Systems effectiveness and efficiency in the context of scalability in corpus size. We propose a general experimental methodology (which helps to study the scalability influence on IR models) based on the construction of a collection on which a given characteristic C is the same whatever be the portion of collection selected. This new collection called uniform can be split into sub-collection of growing size on which some given properties will be studied. We apply our methodology to WT10G (TREC9 collection), the characteristic C here is the distribution of relevant documents on a collection and properties are standards IR evaluation measures.*

MOTS-CLÉS : *passage à l'échelle, collection et sous-collections, évaluation de SRI*

KEYWORDS: *scalability, collection size, splitting collections, IR evaluation*

La quantité de documents disponibles sur le Web s'accroît au fil des années de façon quasi-exponentielle [LYM 03] tandis que ce support est de plus en plus utilisé, tant sur le plan professionnel que sur le plan personnel pour la recherche d'information (RI). Peu de travaux de RI ont jusqu'alors abordé les questions d'efficience et d'efficacité des SRI dans le contexte du passage à l'échelle dans la taille des corpus.

Nous proposons une démarche expérimentale reproductible pour l'étude de l'influence du passage à l'échelle sur différents facteurs de modèles de recherche d'information. Nous présentons tout d'abord des travaux de RI qui se sont interessés au problème du passage à l'échelle dans différentes phases du processus de recherche. Nous développons ensuite notre méthodologie et nous terminons par les expérimentations que nous avons mené sur la collection WT10G de TREC9.

1. Travaux antérieurs sur le passage à l'échelle

Le processus de recherche d'information consiste à fournir, en réponse à une demande de l'utilisateur (requête) les documents qui répondent au mieux à son besoin d'information (documents pertinents). Il se décline en plusieurs phases : construction de la collection, indexation, interrogation et évaluation. Nous passons en revue dans les sections suivantes les travaux associés à ces phases qui se sont interessés au passage à l'échelle.

1.1. *Construction de collection*

La première phase d'un processus de RI consiste en une collecte d'un ensemble de données (documents) au sein duquel la recherche va s'effectuer. Cette phase peut se faire par la mise en place d'une collection statique (Cranfield, TREC, CLEF) ou de facon dynamique (recherche sur le Web). Les collections statiques se composent d'un ensemble de documents (corpus), d'un ensemble de besoins d'informations (appelés *topics* dans le cadre de TREC) et d'un ensemble de jugements de pertinence sur les documents en utilisant les *topics*. L'utilisation des collections de documents de plus en plus larges depuis les années 90 (comme les collections TREC) fournit un environnement de test plus réaliste, mais limite l'étendue des jugements de pertinence. En effet, la technique de *pooling* actuellement utilisée pour les jugements de pertinence [VOO 97] rencontre de nombreuses limites avec la croissance du volume d'information. Cette technique consiste à interroger plusieurs SRI pour des besoins d'information donnés et à fusionner ensuite la liste des n premiers documents retournés par chaque SRI pour former un *pool* (dans TREC, n vaut 100). Ensuite un ensemble de jugements de pertinence sur les documents du *pool* est fourni par des juges humains. Pour des collections volumineuses, cela demanderait de nombreux juges et un temps considérable pour fournir des jugements de pertinence pour tous les documents. Les travaux de [VOO 00] montrent que les désaccords des juges humains [SOB 01] n'ont pas un impact significatif sur la fiabilité des résultats ; [ZOB 98] a également montré que malgré le biais introduit par l'incomplétude des jugements par le *pooling*, les ju-

gements de pertinence fournissent une base crédible pour l'évaluation de nouveaux SRI (SRI n'ayant pas participé à la campagne TREC). Des améliorations du *pooling* ont été proposées sans pour autant remédier à toutes les limites ; ainsi une variante des stratégies standards de *pooling* destinée à accroître le nombre de documents pertinents découverts a été proposée par [ZOB 98], la méthode ISJ, qui utilise un SRI interactif pour sélectionner les documents à juger et la méthode Move-To-front Pooling basée sur un nombre variable de documents pour chaque système participant au *pooling* suivant la performance de recherche de ce système est expliquée dans [COR 98], une méthode de pseudo-jugements de pertinence dans laquelle les juges humains sont remplacés par une technique aléatoire de sélection de documents pertinents est proposée dans [SOB 01]. Des méthodes pour permettre la comparaison de SRI sans utilisation du pooling émergent depuis peu [VOO 04].

Dans le cas de SRI utilisant plusieurs sources distribuées, [FRI 03] facilite le passage à l'échelle dans la taille des corpus en détectant et réduisant la duplication des documents identiques provenant de sources différentes. [SOB 01] s'intéresse à la réplication de la structure des hyperliens du Web sur une collection de taille réduite. Ceci concerne l'échantillonage des collections. Il s'agit de déterminer les propriétés d'une collection volumineuse, de construire une collection de taille réduite ayant les même propriétés et de réaliser des études sur la collection réduite (ce qui devrait être plus aisée que sur la collection entière) et surtout de pouvoir reporter sur la collection entière les résultats obtenus sur cette collection réduite.

1.2. *Indexation et interrogation*

Le temps d'indexation moyen de la collection augmente de manière très significative en fonction de la taille des collections [CHE 04]. La compression physique des informations reste la solution la plus souvent proposée pour cette limite dans le passage à l'échelle. L'utilisation des concepts agrégés pour représenter l'information (plutôt que des unités d'informations fines comme le terme ou les n-grams) est une piste qui permettra la réduction de l'espace de représentation.

Le temps moyen de traitement des requêtes augmente également de manière très significative en fonction de la taille des collections. Réduire la taille de l'ensemble sur lequel s'effectue la recherche en identifiant des sous-collections au sein de l'ensemble des données permettra de réduire ce temps. La difficulté est alors de déterminer les bases sur lesquelles la segmentation devra être effectuée (à base de questions auxquelles l'utilisateur répond ou d'un profil établi grâce à un historique des recherches [NEW 00], à base de métadonnées portant sur le besoin en information de l'utilisateur et/ou sur la nature ou l'usage des documents [CHE 04]).

1.3. *Evaluation*

Deux métriques permettent en général de réaliser l'évaluation de SRI : la précision mesure le taux de documents pertinents parmi les documents retournés ; le rappel mesure le taux de documents pertinents qui ont été effectivement retournés à l'utilisateur. La précision et le rappel mettent en exergue différents aspects de la performance d'un SRI. Les protocoles d'évaluation de la pertinence des résultats de recherche demeurent indépendants de la taille de la collection et de la diversité des corpus, et ceci peut engendrer des biais lors de la comparaison des performances entre SRI ; c'est l'un des enjeux de la tâche TeraByte introduite à TREC en 2004. Au sein des collections volumineuses, on remarque également une hétérogenéité plus importante et les descriptions statiques ne sont plus discriminantes. L'analyse de [BEI 03] montre que le passage à de grandes collections amplifie le problème de discrimination des termes puisque le nombre de termes fréquents n'augmente pas énormément et que la proportion de termes discriminants diminue. L'utilisation de ces collections pour l'évaluation de SRI va donc nécessiter l'apport de nouvelles métriques qui privilégient par exemple la précision sur les premiers documents sélectionnés (elle détermine la satisfaction de l'utilisateur dans des environnements comme le Web) ou qui prennent plus en compte les limites du rappel comme la *bpref* [VOO 04].

1.4. *Construction d'échantillons de tailles croissantes*

Un des objectifs de la collection Very large Collection (VLC) est de déterminer s'il devient plus aisé de trouver des documents pertinents quand la taille de la collection augmente. Suite aux observations des participants de la tâche Very large Collection (VLC) de TREC-6 (augmentation significative de la précision dans les premiers documents quand on passe d'un échantillon de la collection à la collection totale [HAW 99]), [HAW 03] applique la théorie de détection de signal à la RI en utilisant différentes formes de distribution de documents pertinents et de documents non pertinents et mène des expérimentations en constituant trois types d'échantillons de la collection entière :

– les échantillons uniformes : On crée n échantillons primaires de taille égale. Ainsi les sous-collections composées de taille $2/n, 3/n, \cdots, (n-1)/n$ sont constituées en composant les n échantillons comme le montre l'exemple suivant pour un échantillon de taille $3/7$: on contitue tout d'abord 7 échantillons primaires numérotés $0, 1, 2, 3, 4, 5, 6$. On crée ensuite 7 échantillons composites $(0, 1, 2), (1, 2, 3), (2, 3, 4), (3, 4, 5), (4, 5, 6), (5, 6, 0) et (6, 0, 1)$. Les tests se font sur chacune de ces sous-collections composées et le résultat reporté pour l'échantillon de taille $3/7$ correspond à une moyenne des résultats sur toutes ces sous-collections. Ainsi, les mesures moyennes reportées pour chaque taille de sous-collections prennent en compte toutes les données.

– Les échantillons répliqués : On prend les échantillons primaires de taille $1/10$ de la collection entière et on les replique un nombre de fois voulu.

Exemple : $(0), (0,0), ..., (0,0,0,0,0,0,0,0,0,0)$, $(1), ..., (9,9,9,9,9,9,9,9,9,9)$

– les échantillons " biaisés" : Les données de TREC-6 sont subdivisées en 5 sous-collections disjointes suivant leur origine. On fait de l'ensemble de documents de chacune des sources un échantillon, mais ces ensembles diffèrent en nombre de documents et en taille (de 235.4 Mo à 564.1 Mo) et leurs documents n'ont pas les mêmes probabilités de pertinence.

Les méthodes de [HAW 03] de construction de collection de taille croissante et la méthodologie que nous proposons se rejoignent dans leur objectif de fournir un cadre d'étude de l'influence du passage à l'échelle. Toutefois, la démarche que nous proposons se veut générale et peut être appliquée et adaptée selon les cas à de multiples caractéristiques en RI.

2. Notre démarche

2.1. *Hypothèses et méthodologie générale*

Notre objectif est d'étudier l'influence de la croissance de collections sur les modèles de RI. Nous devons pour ce faire mener des expérimentations sur des collections de taille croissante et analyser l'influence de la taille sur les propriétés de modèles de RI. La question est de savoir comment construire ces collections pour assurer une certaine fiablilité dans les résultats obtenus. Soit C une caractéristique de collection et P_i des propriétés qui dépendent de C. Le but de notre méthodologie est d'obtenir un ensemble de collections de taille croissante, similaires en ce qui concerne la caractéristique C. Pour ce faire, nous choisissons de construire une collection de départ sur laquelle la caractéristique C est la même quelle que soit la portion sélectionnée de la collection. Si nous disposons d'une telle collection, il est possible de la découper en portions (sous-collections) de taille croissante, d'étudier des propriétés P_i sur chaque portion et d'analyser l'influence de la taille de la portion sur ces propriétés. Le découpage de la collection initiale ne doit pas être contrainte. Ceci signifie qu'on pourra découper la collection initiale de différentes manières, la seule contrainte étant que la caractéristique C choisie soit la même sur toute portion choisie. Notre méthodologie comprend 4 étapes :

1) Cette étape suppose que nous avons une collection initiale. Dans ce cas nous étudions la caractéristique C sur cette collection dans le but de déterminer si elle satisfait déjà nos contraintes. Si nous n'avons pas de collection initiale, nous débutons à l'étape 2 qui suit. Si nous avons une collection initiale qui satisfait nos contraintes alors nous passons à l'étape 3 directement.

2) Construire une collection sur laquelle la caractéristique C soit la même sur toute portion choisie.

3) Echantilloner cette collection en portions de taille croissante.

4) Etudier les propriétés P_i sur chaque portion et analyser l'influence de la portion sur ces propriétés.

Nous appliquons cette méthodologie générale au cas particulier de l'évaluation en RI. Les valeurs des métriques d'évaluation en RI dépendent directement de la proportion des documents pertinents. Ainsi, l'évaluation est basée sur les documents pertinents connus et/ou retournés ; une étude de l'influence du passage à l'échelle sur l'évaluation de modèles de RI doit donc tenir compte de la proportion de documents pertinents dans chaque sous-collection utilisée. De plus,étant donné l'absence de contrainte sur la façon dont la collection pourra être découpée en sous-collections, et pour s'assurer que le nombre de documents pertinenets dans une portion donnée de la collection sera proportionnel à la taille de la portion, nous devons tenir compte de la distribution des documents pertinents. Ainsi, nous appliquons la méthodologie générale décrite précédemment avec comme caractéristique C la répartition de documents pertinents et comme propriétés les métriques d'évaluation de SRI.

2.2. *Etude de la caractéristique* C

Cette première étape consiste à étudier la propriété C sur la collection initiale. Dans notre cas d'étude, nous avons une collection initiale, il s'agit d'étudier la distribution de documents pertinents sur cette collection. Si les documents pertinents ne sont pas répartis de façon à ce que en sélectionnant n'importe quelle portion de collection, on obtienne la même distribution, nous passons à l'étape 2, sinon nous allons à l'étape 3 directement.

2.3. *Construction d'une collection uniforme suivant* C

Cette étape a pour but d'obtenir une collection qui peut être découpée de différentes façons, en respectant nos hypothèses sur la caratéristique C. Pour notre cas d'étude, la distribution des documents pertinents doit être la même quelle que soit la portion de collection choisie. Ainsi, le nombre de documents pertinents par *topic* et pour tous les *topics* est proportionnel à la taille de la portion. Pour obtenir une telle distribution, nous calculons la distance $E(t)$ que nous *souhaitons avoir* entre deux documents pertinents d'un *topic* t. Soit $P(t)$ l'ensemble des documents pertinents pour le *topic* t, T l'ensemble de tous les *topics* et D l'ensemble de tous les documents de la collection. Nous avons opté pour la distance :

$$E(t) = \frac{\mid D \mid - \mid \bigcup_{k \in T} P(k) \mid}{\mid P(t) \mid}$$

Ainsi au sein de la nouvelle collection, les documents pertinents pour le *topic* t seront séparés de $E(t)$ documents dits non pertinents (i.e. documents qui ne sont jugés pertinents pour aucun *topic*), et éventuellement de documents jugés pertinents pour d'autres *topics* différents de t. Pour des documents jugés pertinents pour plusieurs *topics*, ils sont insérés une seule fois dans la nouvelle collection, à la position définie par le premier des *topics* concernés traités. Ainsi, la distance réelle entre deux documents pertinents de t notée $E_r(t)$ est telle que $E_r(t) \leq E(t) + \mid (\bigcup_{k \in T} P(k)) - P(t) \mid$.

Ceci introduit donc un biais sur l'uniformité qu'on souhaite avoir sur la collection. Etant donné que le nombre total de documents dans la collection est tel que $(\mid \bigcup_{k \in T} P(k) \mid) <<< \mid D \mid$, ce biais est peu significatif et n'influence pas l'uniformité de la distribution des documents pertinents que nous voulons établir sur la collection. Lors du découpage de la collection en portions (sous-collections), la taille de ces sous-collections reste suffisamment élevée pour que l'influence de ce biais soit négligée. De plus, ce biais est un compromis nécessaire pour obtenir une collection au sein de laquelle les documents pertinents sont à la fois répartis uniformément (à peu près) pour chaque *topic* et répartis uniformément (à peu près) si l'on considère l'ensemble des *topics*.

Voici l'algorithme général de notre démarche d'uniformaisation.

PertiEcrit=FAUX ;

$\forall t \in T$ {Initialiser C(t)=compteur(t) ; Calculer E(t) ;Constituer P(t) ; }

$NP = \{d \in D / \forall t \in T, d \notin P(t)\}$, $P = \bigcup_{t \in T} P(t)$

Tantque($NP \neq \emptyset$)et($\exists d \in P / NonMarquer(d, P)$){

$\forall t \in T$ {

Si (C(t)==E(t)) {d=document de P(t) ; Si (NonMarquer(d,P) {Ecrire(d) ;Marquer(d,P) ;}}

Re-Initialiser C(t) ;PertiEcrit=VRAI ;

}

Si (PertiEcrit==FAUX){d=document de NP ; Ecrire(d) ;$\forall t \in T$, C(t)=C(t)+1 ;}

}

Nous étayons ceci par un exemple. Supposons que la collection initiale soit composée de 30 documents $D = \{d_1, \cdots, d_{30}\}$ et $T = \{t_1, t_2\}$. Supposons $P(t_1) = \{d_1, d_7, d_{18}\}$ et $P(t_2) = \{d_7, d_{21}\}$. Le document d_7 est pertinent pour t_1 et pour t_2.

Nous calculons $E(t_1) = \frac{|D| - |\bigcup_{k \in T} P(k)|}{|P(t_1)|} = \frac{(30-4)}{3} \approx 8$ et $E(t_2) = \frac{(30-4)}{2} \approx 13$

Dans la collection uniforme, les documents sont ordonnés comme suit :

$$\{d_2, d_3, d_4, d_5, d_6, d_8, d_9, d_{10}, \underbrace{d_1}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, \overbrace{d_7}, d_{16}, d_{17}, d_{19},$$

$$d_{20}, d_{22}, d_{23}, d_{24}, d_{25}, \underbrace{d_{18}}, d_{26}, d_{27}, d_{28}, d_{29}, d_{30}, \overbrace{d_{21}}\}$$

Nous envisagions d'avoir $E(t_1)$ documents non pertinents entre deux documents pertinents de t_1. Dans le pire des cas pour notre exemple, deux documents pertinents de t_1 seront séparés de $E(t_1) + \mid (\bigcup_{k \in T} P(k)) - P(t_1) \mid = E(t_1) + 1$.

On remarque que le document d_7 est inséré une seule fois dans la collection. Parce que ce document est pertinent pour plusieurs *topics*, la distance réelle entre deux documents pertinents du *topic* t_1 est différente de la distance attendue. Le biais que cela

implique sur l'uniformité de la distribution de documents pertinents n'est pas significatif si la collection est volumineuse.

2.4. *Echantillonage de la collection*

Nous obtenons avec la démarche expliquée ci-dessus une collection " uniforme " et réutilisable pour différents types d'expérimentations. Cette collection peut être découpée en portions de tailles différentes et de différentes façons, puisque le nombre de documents pertinents sur une portion est proportionnel à sa taille.

Pour le cas de l'exemple précédent, pour obtenir des sous-collections de taille croissante, l'on peut découper la collection en $N = 3$ portions de taille $\frac{|D|}{N} = 10$

– Une première portion est $D_1 = \{d_2, d_3, d_4, d_5, d_6, d_8, d_9, d_{10}, d_1, d_{11}\}$

– Une seconde portion est $D_2 = \{d_{12}, d_{13}, d_{14}, d_{15}, d_7, d_{16}, d_{17}, d_{19}, d_{20}, d_{22}\}$

– Une troisième portion est $D_3 = \{d_{23}, d_{24}, d_{25}, d_{18}, d_{26}, d_{27}, d_{28}, d_{29}, d_{30}, d_{21}\}$

Ainsi, l'on peut construire des ensembles de sous-collections $\{S_1 = D_1, S_2 = D_1 \bigcup D_2, S_3 = D_1 \bigcup D_2 \bigcup D_3\}$ ou $\{S_1 = D_2, S_2 = D_1 \bigcup D_3, S_3 = D_1 \bigcup D_2 \bigcup D_3\}$. Dans les deux cas, nous obtenons trois sous-collections de taille croissante sur lesquelles la distribution de documents pertinents est la même.

2.5. *Etude de l'influence de la taille de collection*

Dans cette étape, nous avons des sous-collections de taille croissante sur lesquelles la caractéristique C est la même (dans notre cas d'étude la distribution des documents pertinents). Nous étudions donc les propriétés P_i (dans notre cas d'étude les métriques d'évaluation de RI) sur chaque sous-collection et analysons le comportement de ces propriétés quand la taille de collection augmente.

3. Expérimentations avec WT10G

3.1. *Distribution des documents pertinents sur WT10G*

3.1.1. *Données*

Nous avons travaillé sur la collection de test de la conférence TREC denommée WT10G [BAI 01]. Les besoins d'information pour nos tests correspondent aux *topics* 451-500 pour lesquels un ensemble de jugements de pertinence des documents est fourni. Cette collection de test contient 1.692.096 documents dont 2.371 documents jugés pertinents.

$T = \{451, \cdots, 500\}$, D est WT10G et $\mid D \mid = 1.692.096$, $\sum \mid \bigcup_{t \in T} P(t) \mid = 2.371$

Tableau 1. *Statistiques sur nos requêtes : nombre de mots par requête*

Min	Max	Moyenne
2	8	4, 76

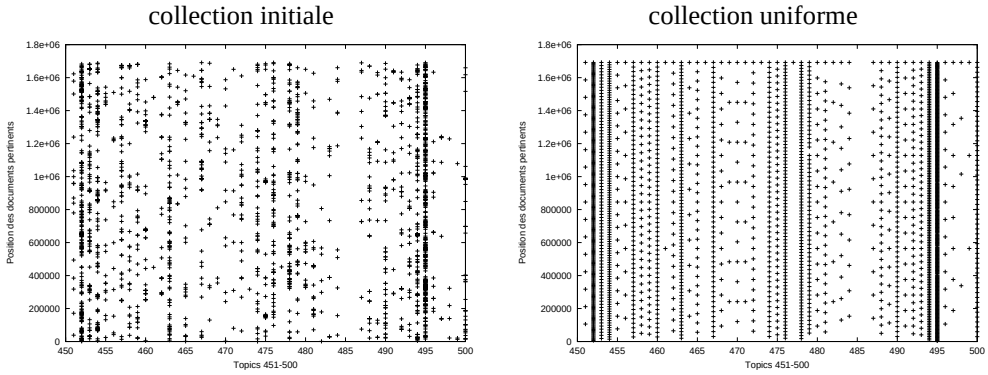
3.1.2. *Les requêtes utilisées*

A partir des *topics* de TREC9, nous avons construit manuellement un ensemble de requêtes pour l'interrogation des documents. Une requête est un ensemble de mots clés qui tiennent compte du titre du *topic* tel que fourni par TREC et du descriptif des documents pertinents attendus pour ce *topic* (partie DESC et partie NARR du *topic*). Le tableau 1 donne des statistiques sur le nombre de mots-clés par requêtes.

3.1.3. *Distribution des documents pertinents sur WT10G*

Nous avons étudié la répartition (position) des documents pertinents sur la collection WT10G par *topic* et nous obtenons le graphique de gauche de la figure 1.

Figure 1. *Documents pertinents par topic*



La répartition des documents pertinents est quelconque et elle varie selon les *topics*. Pour un *topic* donné, l'ensemble des documents jugés pertinents n'est pas réparti uniformément sur l'ensemble de la collection, le nombre de documents pertinents n'est pas une fonction linéaire de la taille de la collection de test. Etant donné que nous subdiviserons notre collection de test en sous-collections dont nous ferons croître la taille progressivement, il est important de tenir compte de la façon dont sont répartis les documents sur chaque sous-collection pour que les métriques de RI gardent tout leur intérêt quelle que soit la taille de la collection utilisée. Nous avons donc redistribué les documents pertinents au sein de la collection.

Tableau 2. *WT10G Uniforme : statistiques sur les longueurs de documents en nombre de caractères par collection*

Taille collection	Min	Max	doc vides	Moy	doc taille $\leq$Moy	doc taille $>$Moy
200.000	3	2.326.790	3	3.875,2	155.090	44.910
400.000	3	2.326.790	14	4.103,5	314.199	85.801
600.000	3	2.326.790	21	3.902	468.430	131.570
800.000	3	2.344.747	29	3.876,6	628.158	171.842
1.000.000	3	2.344.747	33	3.857,1	785.360	214.640
1.200.000	3	2.344.747	34	3.766,9	943.802	256.198
1.400.000	3	2.344.747	36	3.773,1	1.101.592	298.408
1.600.000	3	2.344.747	38	3.790,2	1.260.289	339.711

3.2. *Uniformisation de la collection*

La figure 1 montre la distribution de documents pertinents sur la collection uniforme. Dans cette nouvelle collection, le nombre de documents pertinents pour un *topic* donné est fonction linéaire de la taille de la collection et le nombre de documents pertinents (tous *topics* confondus) est également une fonction linéaire de la taille de la collection. Ainsi au sein de la nouvelle collection, les documents pertinents pour le *topic* t seront séparés de $E(t)$ documents dits non pertinents (i.e. documents qui ne sont jugés pertinents pour aucun *topic*), et éventuellement de documents jugés pertinents pour d'autres *topics* différents de t.

3.3. *Découpage en sous-collections*

Dans le cadre de nos expérimentations, nous avons constitué des portions de taille croissante par pas de 200.000 documents, en prenant les documents dans l'ordre d'apparition dans la collection. Nous obtenons ainsi 8 collections décrites dans la table 2. Nous avons travaillé sur 7 de ces sous-collections. Le choix du pas peut varier selon les besoins des expérimentations.

4. Expérimentations

4.1. *Modèles de RI utilisés*

– L'outil LUCY que nous avons utilisé implémente le modèle Okapi qui est une extension du modèle probabiliste.

– Nous avons également utilisé l'outil MG qui est basé sur le modèle vectoriel.

Les 3 autres modèles sont basés sur la proximité entre termes de la requête ; les implémentations que nous avons utilisées sont décrites en détail dans [MER 04].

– La méthode Cover Density Ranking de [CLA 00] permet de classer les documents pertinents du point de vue du système selon " la densité de couverture " des mots-clés de la requête dans les documents. Nous nommerons ce modèle *modèle de Clarke*.

– La méthode de [HAW 95] : Une requête est un ensemble de couples (u, a), composés d'une relation de proximité u et d'un coefficient d'importance a. Chaque élément de $I(u, a)$ (ensemble des intervalles d'un document qui satisfont la relation de proximité u) participe au score de pertinence de ce document. Nous nommerons ce modèle *modèle de Hawking*.

– La méthode de [RAS 03] attribue à chaque document, pour une requête donnée, un score calculé sur la base du modèle Okapi et en fonction de la proximité des termes de la requête dans le document. Nous nommerons ce modèle *modèle de Rasolofo*.

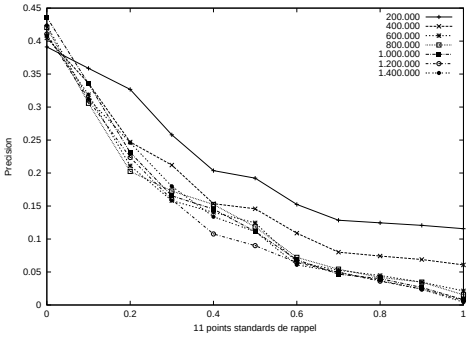
4.2. *Courbes rappel/précision*

Quand les documents pertinents ne sont pas répartis de façon uniforme sur la collection, il est délicat de mesurer l'impact qu'un accroissement de la taille aurait sur les métriques précision/rappel. Si les documents pertinents sont répartis de façon uniforme sur l'ensemble de la collection, alors nous pouvons observer l'impact de la taille sur le rapport rappel/précision.

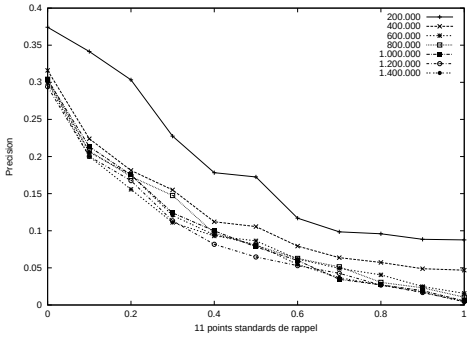
Les courbes de rappel/précision pour les 5 modèles de RI sont données sur les Fig. 2 (modèle de Clarke et modèle de Hawking), Fig. 3(modèle Okapi de l'outil Lucy et modèle de Rasolofo) et Fig. 4 (modèle vectoriel de l'outil MG).

Figure 2. *Rappel/précision pour les* 7 *sous-collections uniformes de WT10G*

modèle de Clarke



modèle de Hawking



Dans le cas des modèles de Hawking et de Clarke : on constate que pour les premiers niveaux de rappel, les courbes des grandes collections sont très proches, presque confondues. Pour les hauts niveaux de rappel, les courbes demeurent proches. Ceci signifie que ces deux modèles ont une certaine *stabilité* concernant la performance

Figure 3. *Rappel/précision pour les* 7 *sous-collections uniformes de WT10G*

modèle Okapi(Lucy) modèle de Rasolofo

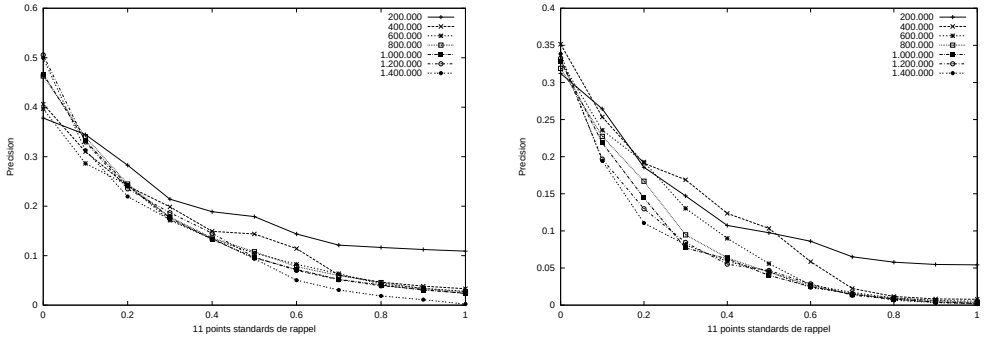
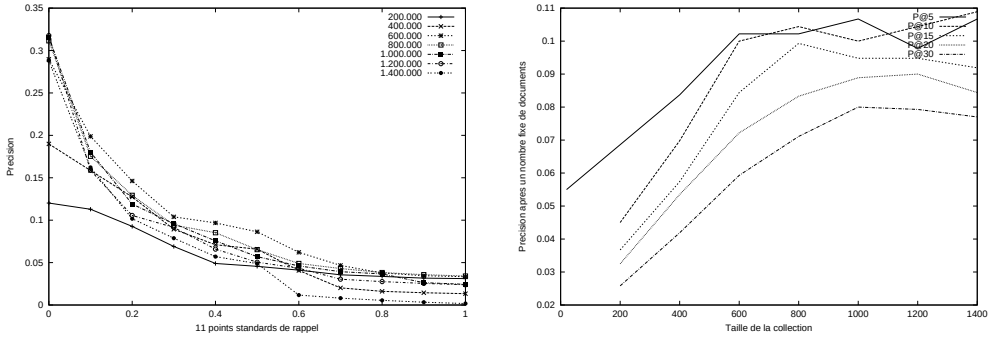


Figure 4. *Rappel/précision et P@n pour les* 7 *sous-collections uniformes de WT10G - modèle vectoriel (Mg)*

Rappel/Précision P@n



rappel/précision quand la taille de collection augmente et respectent ainsi nos hypothèses. On remarque également que le score attribué à un document pour ces deux modèles ne dépend que du document et de la requête, mais non de la collection.

Pour le modèle OKAPI, le rapport rappel/précision est bien meilleur pour les grandes collections que pour les petites sur les premiers niveaux de rappel. Quand le niveau de rappel augmente (jusqu'à 10%), la tendance change et les courbes deviennent plus proches. On sait que pour le modèle OKAPI, le score attribué à un document dépend des autres documents de la collection ; ce score n'est pas indépendant de la taille de collection. La *stabilité* est moins bonne que celle des deux modèles précédents.

Pour le modèle de Rasolofo, les grandes collections sont moins bien placées que les petites collections pour les tous premiers niveaux de rappel. Pour les niveaux de rappel de 10% à 30%, les courbes sont ordonnées de la plus petite collection à la plus

grande et les courbes ne sont pas proches. Pour les grandes collections, les courbes commencent à être proches à partir du niveau de rappel 30%. Ce modèle combine la proximité entre les termes de la requête et un score Okapi pour attribuer un score final à tout document, autrement dit, ce modèle combine une fonction qui ne dépend pas de la collection et une autre qui en dépend. Sa *stabilité* commence à des niveaux de rappel hauts.

Pour le modèle vectoriel, les courbes des grandes collections sont proches pour les premiers niveaux de rappel (jusqu'à 10%). Mais l'ordre des courbes varie en fonction du niveau de rappel et les courbes se chevauchent plus ou moins. Ceci montre une certaine *instabilité* concernant le rapport rappel/précision quand la taille de collection croît. Pour ce modèle, le score attribué à un document dépend directement des autres documents de la collection (au travers du *idf* qui dépend de la collection).

Le rôle de la collection dans l'attribution de scores aux documents affecte le passage à l'échelle du modèle (en ce qui concerne le rapport rappel/précision). Les modèles de RI pour lesquels le score attribué à un document dépend uniquement de la requête et du document passent mieux à l'échelle (ils sont plus stables) pour la performance rappel/précision que ceux pour lesquels le score dépend de la collection.

4.3. *Haute précision*

La précision après un nombre fixe de documents influe directement sur la satisfaction de l'utilisateur dans des environnements comme le Web et elle est une métrique d'évaluation de SRI facile à interpréter. Les travaux de [HAW 03] ont utilisé un unique modèle de RI (Okapi BM25 au travers du système PADRE) pour méner des expérimentations sur la précision après un nombre fixe de documents. Pour les 5 modèles de RI utilisés, nos résultats rejoignent ceux que [HAW 03] a obtenu pour les $P@20$. Nous étendons ces résultats aux précisions $P@5$, $P@10$, $P@15$, $P@30$. La figure 4,graphique de droite (modèle vectoriel) montre l'évolution de ces précisions quand la taille de la collection croît. Pour les 4 autres modèles utilisés, nous obtenons des courbes similaires. Ces résultats montrent que la précision sur les premiers documents retournés augmente avec la taille de l'échantillon considéré de la collection.

4.4. *Positions des premiers documents pertinents*

Notre étude sur les positions des 1ers documents pertinents retourné montre que la position moyenne du premier document pertinent retourné baisse sensiblement (donc se rapproche des premiers documents retournés) quand la taille de collection croît. De plus, le pourcentage de *topics* pour lesquels le premier document pertinent retourné est parmi les 10 premiers documents augmente rapidement avec la taille de la collection.Nous avons retenu uniquement les *topics* pour lesquels le premier document pertinent apparaît parmi les 1000 premiers documents retournés pour toutes les sous-collections (31 *topics* sur 50).

5. Conclusions

Nous proposons une démarche reproductible qui permet d'étudier l'influence du passage à l'échelle sur les modèles de RI. Cette démarche passe par l'uniformisation d'une collection volumineuse par rapport à une caractéristique C (dans notre cas d'étude, la répartition des documents pertinents sur la collection) et son échantillonage en sous-collections de taille croissante. Ces sous-collections nous permettent ensuite d'étudier l'impact de l'augmentation en volume du corpus sur des propriétés dépendantes de la caractéristique C (dans notre cas d'étude les métriques d'évaluation de SRI). Nos résultats montrent que les modèles de RI pour lesquels l'attribution de score à un document ne dépend pas de la collection (mais uniquement du document et de la requête) sont plus stables pour les performances rappel/ précision quand la taille de collection croît. Nous travaillons à mieux cerner l'impact que le rôle de la collection dans l'attribution de score a sur les modèles de RI. Nos résultats pour la haute précision rejoignent (et étendent à plusieurs niveaux de coupure et à 5 modèles de RI) ceux de [HAW 03], à savoir que la précision après un nombre fixe de documents retournés augmente avec la taille de la collection. La croissance en taille de la collection améliore également la position du premier document pertinent retourné.

La collection uniforme WT10G est en cours d'utilisation pour l'étude de l'impact de la taille de collection sur d'autres métriques d'évaluation de RI (nouvelle métrique comme la *bpref* de [VOO 04] qui semble plus robuste à l'incomplétude des jugements de pertinence).

La croissance du volume d'information est continue et il devient incontournable de s'intéresser à la façon dont les modèles de recherche vont se comporter face à des espaces de recherche de plus en plus larges. Notre méthodologie vise à permettre de telles études. Elle peut être utilisée pour bâtir des collections "uniformes" par rapport à d'autres caractéristiques (répartition des termes de requêtes) pour étudier d'autres propriétés de RI.

6. Bibliographie

[BAI 01] BAILEY P., CRASWELL N., HAWKING D., « Engineering a multipurpose test collection for Web retrieval experiments DRAFT », *Proceedings of the 24th annual international ACM SIGIR conference*, 2001.

[BEI 03] BEIGBEDER M., MERCIER A., « Etude des distributions de tf et de idf sur une collection de 5 millions de pages html », *Atelier de recherche d'informations sur le passage à l'échelle Congrès INFORSID 2003*, Nancy, France, 2003.

[CHE 04] CHEVALLET J. P., MARTINEZ J., BOUGHANEM M., LECHANI-TAMINE L., CALABRETTO S., « Rapport final de l'AS-91 du RTP-9 'Passage à l'échelle dans la taille des corpus' », 2004.

[CLA 00] CLARKE C. L. A., CORMACK G., TUDHOPE E., « Relevance ranking for one to three term queries », *Information Processing and Management*, vol. 26, n° 2, 2000, p. 291-311.

[COR 98] CORMACK G., PALMER C., CLARKE C. L. A., « Efficient construction of large test collections », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, p. 282-289.

[FRI 03] FRIEDER O., GROSSMAN D., « On scalable Information retrieval Systems », *Invited Paper, 2nd IEEE International Symposium on Network Computing and Applications.*, Massachusett, Cambridge, 2003.

[HAW 95] HAWKING D., THISTLEWAITE P., « Proximity operators - so near and yet so far », *Proceedings of the Fourth Text Retrieval Conference TREC-4*, 1995, p. 131-143.

[HAW 99] HAWKING D., THISTLEWAITE P., « Scaling up the TREC collection », *Information retrieval*, vol. 1, n° 1, 1999, p. 115-137.

[HAW 03] HAWKING D., ROBERTSON S., « On collection size and retrieval effectiveness », *Information retrieval*, vol. 6, n° 1, 2003, p. 99-105.

[LYM 03] LYMAN P., VARIAN H. R., SWEARINGEN K., CHARLES P., GOOD N., JORDAN L. L., PAL J., « How much informations 2003 », http ://www.sims.berkeley.edu/research/projects/how-much-info-2003/, October 2003.

[MER 04] MERCIER A., « Etude comparative de trois approches utilisant la proximité entre les termes de la requête pour le calcul des scores des documents », *INFORSID 2004 - 22ème congrès informatique des organisations et des systèmes d'information et de décision*, 2004.

[NEW 00] NEWBY G. B., « The science of large scale information retrieval », Internet archives, 2000.

[RAS 03] RASOLOFO Y., SAVOY J., « Term proximity scoring for keyword-based retrieval systems », *Proceedings of European Conference on Information Retrieval Research*, 2003, p. 207-218.

[SOB 01] SOBOROFF I., NICHOLAS C., CAHAN P., « Ranking retrieval systems without relevance judgments », *Proceedings of the 24th annual international ACM Conference on research and Development in Information Retrieval*, 2001, p. 66-73.

[VOO 97] VOORHEES E., HARMAN D., « Overview of the sixth text retrieval conference (TREC-6) », *NIST Special Plublication 500-420-The Sixth text retrieval Conference*, 1997.

[VOO 00] VOORHEES E., « Variations in relevance judgments and the measurement of retrieval effectiveness », *Information and Processing Management*, , n° 36, 2000, p. 697-716.

[VOO 04] VOORHEES E., BUCKLEY C., « Retrieval evaluation with incomplete information », *Proceedings of the 27th annual international conference on Research and development in information retrieval*, 2004, p. 25-32.

[ZOB 98] ZOBEL J., « How reliable are the results of large scale information retrieval experiments », *Proceedings of the 21th ACM SIGIR Conference on research and development in information retrieval*, 1998, p. 307-314.