Analyse expérimentale sur la structure des index documentaires et leur impact sur l'efficacité de la recherche :

Cas de collections volumineuses

Soheila Karbasi — Lynda Lechani Tamine

IRIT – Equipe SIG
118 route de Narbonne
31062 Toulouse Cedex 4
{karbasi,lechani}@ irit.fr

RÉSUMÉ. *Cet article s'inscrit dans le cadre général de la problématique du passage à l'échelle dans la taille des corpus en l'abordant plus précisément sous l'angle des limites des représentations locales et globales des index documentaires. Une analyse globale de la structure de ces index est présentée en utilisant des collections de référence TREC. Cette analyse est suivie d'une évaluation expérimentale de leur impact sur l'efficacité de la recherche.*

ABSTRACT. *This article deals with the problem of scaling in information retrieval. More precisely, our goal is to analyse, in this context, the limits of index document indexes. An overall analysis on these index structures is presented by using the TREC collections. This analysis is followed by an experimental evaluation of their impact on the research retrieval effectiveness.*

MOTS-CLÉS : *passage à l'échelle, collections volumineuses, schéma de pondération TF IDF.*

KEYWORDS: *the passage on the scale, voluminous collections, TF IDF weighting schemes.*

1. Introduction

Avec l'avènement d'Internet, qui est sans doute l'un des réseaux les plus sollicités de nos jours, les facteurs volume de l'information et nombre d'utilisateurs se sont accrus de manière significative. En intégrant, de surcroît, les effets de la mise en œuvre d'intranets et d'extranets d'entreprises ainsi que les bibliothèques numériques, on prend conscience du volume grandissant d'information et du nombre exponentiel d'utilisateurs auxquels est confronté tout service d'information. Ceci traduit le phénomène du passage à l'échelle en recherche d'information. Dans ce contexte, où la problématique de la recherche d'information a pris une nouvelle dimension, on constate que les modèles et techniques traditionnelles connus dans ne conservent pas leur niveau de performances. En effet, compte tenu des exigences liées à l'efficacité et l'efficience, les outils de recherche d'information doivent prendre en compte les particularités liées au volume d'information. Les solutions apportées se basent pour une grande part, sur des méthodes de réduction de l'espace de représentation des informations à l'aide de nouveaux modèles formels et des méthodes de compression physique des données.

Dans cet article, nous abordons cette problématique générale sous l'angle particulier de l'impact de la structure des index documentaires sur l'efficacité de la recherche. Une analyse expérimentale préalable présente les caractéristiques des collections de test de référence TREC, et montre que l'accroissement des volumes des collections se traduit tant par l'accroissement de la taille des index que par leur hétérogénéité. Une étude sur l'impact de ces structures sur l'efficacité de la recherche, met en évidence le rôle dissocié de chacun des paramètres du schéma de pondération TF[1] IDF[2] sur les performances de recherche.

Cet article est organisé comme suit : la section 2 donne un large aperçu de la problématique du passage à l'échelle en recherche d'information. La section 3 présente un rappel sur la structure des index documentaires en focalisant sur les propriétés des schémas de pondération largement utilisés dans la littérature. Un intérêt particulier est porté au schéma de pondération de la forme TF IDF en ce sens, qu'on s'intéresse particulièrement à la déclinaison du volume sur ses spécificités. La section 4 présente notre étude expérimentale. On y développe notamment une étude préalable des caractéristiques des index documentaires des collections interrogées puis évaluons l'impact de la variation des paramètres de la fonction de pondération, sur l'efficacité de la recherche. Enfin une conclusion synthétise les résultats obtenus.

2. Problématique générale du passage à l'échelle

Le passage à l'échelle d'une technique et/ou algorithme désigne sa capacité à traiter des volumes considérables d'informations tout en conservant une complexité

1. Term Frequency
2. Inverse Document Frequency

du même ordre que celle induite par le traitement de volumes moins importants. En effet, depuis quelques années le volume d'informations ne se mesure plus en giga octets mais en téra-octets voire en péta-octets. Deux principaux facteurs sont à l'origine du passage à l'échelle : croissance du nombre d'utilisateurs et accroissement significatif des volumes d'information [Kobayashi et Takeda, 00]. Ceci a des incidences directes sur le processus de recherche d'information, notamment sur les phases de préparation des collections et évaluation des requêtes.

La réparation de collections volumineuses pose des problèmes liés à l'accroissement de la taille des index, qui engendre un coût important de stockage [Witten et al., 99], difficulté des mises à jour de ces index [Berry et al., 99] et difficulté de maîtrise de l'évaluation des vocabulaires en fonction de la taille des collections [Williams et Zobel, 03]. Les solutions apportées à ces problèmes portent essentiellement sur la factorisation conceptuelle de l'information [Berry et al., 99] [Kokiopoulou et Saad, 04][Tang et al, 04] et compression physique des textes et index [Witten et al., 99]. Par ailleurs, l'évaluation des requêtes, dans le cas de collections volumineuses, pose le problème épineux de l'allongement des temps de réponse. En effet, la plupart des algorithmes de recherche d'information ont une complexité exponentielle. Cette complexité, conjuguée à la complexité des formulations d'ordonnancement des documents, dégrade la performance d'exécution de nombreux algorithmes en recherche d'information [Newby, 00]. Les résultats obtenus dans les campagnes TREC, plus précisément dans le cadre de la tâche *VLC*[3], confirment ce point [Hawking et al., 99]. Dans le but de pallier à ces problèmes, les propositions ont porté essentiellement sur la parallélisation des algorithmes de recherche d'information [Bailey et al., 96] [Jain et Goharian, 02], l'optimisation des accès aux fichiers inverses [Moffat et Zobel, 96] et réduction de la complexité des algorithmes proposés dans les modèles classiques de recherche d'information [Lee et Ren, 96]. En définitive, cet impact tangible du passage à l'échelle sur les phases du processus de la recherche d'information se décline dans la dégradation des performances tant du point de vue de l'efficacité que de l'efficience [Boughanem et al., 04]. Dans ce large contexte, notre objectif est d'analyser l'impact du volume sur la représentativité locale et globale des unités d'information, issues de l'étape de préparation des collections, puis en examiner les conséquences sur l'efficacité de la recherche. Plus précisément, nous nous intéressons à l'évaluation d'un schéma de pondération largement utilisé en recherche d'information, en l'occurrence le schéma TF-IDF, en analysant au préalable la structure de collections volumineuses afin d'en dégager les paramètres caractéristiques ayant un impact direct sur les performances de recherche dans de telles collections.

3. Structure d'un index documentaire

Malgré la diversité des modèles et techniques en recherche d'information, les bases de représentation des index documentaires demeurent quasi-équivalents. Le

3. Very Large Corpus

principe fondamental de leur organisation est généralement basé sur une structure composée de granules d'informations et de poids dont l'interprétation précise dépend du modèle théorique qui les supporte. Le granule d'information se traduit usuellement par les notions de terme, groupe de termes, concept etc. Le poids exprime le degré de représentativité locale du granule d'information, dont la portée est le document, ou sa représentativité globale, dont la portée est la collection de documents, ou alors une combinaison des deux types de représentativité. Aizawa [Aizawa, 03] identifie deux approches pour la classification des mesures conventionnelles pour la pertinence d'un terme relativement à l'index documentaire associé : approche basée sur l'utilisation et approche basée sur le fondement mathématique. La première approche aborde la classification des mesures sous l'angle de leur finalité. Dans ce sens on identifie deux principales classes de mesures : Mesures pour la sélection des termes et mesures pour la pondération des termes. Les Mesures pour la sélection des termes servent à choisir les termes importants dans une collection de documents ; elles sont particulièrement utilisées dans les processus de reformulation de requête par injection de pertinence (méthodes de sélection de termes pour l'expansion de requête) et méthodes de catégorisation de textes. Les mesures pour la pondération des termes sont utilisées pour estimer l'importance d'un terme dans la description du contenu d'un document. Cette classe de mesure est largement utilisée dans les schémas de pondération en recherche d'information. La seconde approche aborde la classification des mesures sous l'angle des heuristiques et principes formels servant de base à leur définition. C'est l'approche que nous développons dans le paragraphe suivant.

3.1. *Mesures mathématique pour la construction d'index*

Lors de la sélection et pondération des index, deux aspects ressortent : spécificité et exhaustivité. La spécificité traduit la représentativité des termes de l'index alors que l'exhaustivité traduit leur degré de couverture du contenu documentaire associé. Une méthode d'indexation performante doit faire un compromis entre ces deux aspects [Sparck Jones, 04]. On recense, en théorie, de nombreuses mesures : *mesure de la popularité*, basée sur la fréquence des termes ou la probabilité estimée de leurs occurrences [Luhn, 57], *mesure de la spécificité*, basée sur la quantité d'information ou l'entropie des termes, et qui quantifie les déviations dûes à l'aspect aléatoire des occurrences des termes [Sparck Jones, 72], m*esure de la discrimination* basée sur la contribution des termes à l'évaluation d'une fonction de discrimination spécifiée. Elle est souvent utilisée dans le modèle probabiliste et plus particulièrement lors de l'expansion de requête [Aizawa, 03], *mesure de la représentation* basée sur une combinaison de la fréquence de termes dans les documents et fréquence des termes dans la collection. Elle quantifie l'utilité des termes en spécifiant le document dans lequel ils apparaissent [Salton et Buckley, 88]. Cette mesure est largement utilisée dans le modèle vectoriel.

En résumé, la mesure de la popularité cible les termes fréquents dans les documents alors que la mesure de spécificité cible les termes rares dans la

collection. La difficulté de construire un index documentaire adéquat réside dans la réalisation d'un juste compromis entre ces deux types de mesures. Ce compromis fait, à juste titre, l'objet des mesures de représentation utilisées particulièrement dans les schémas de pondération de la forme TF-IDF.

3.2. *Analyse du schéma TF- IDF*

Le schéma de pondération TF-IDF est basé sur une mesure de représentation. Cette mesure est basée sur une heuristique permettant de déterminer l'importance des termes dans un document et son efficacité a été justifiée depuis de longue date en recherche d'information. Ce schéma est souvent associé à la fréquence d'apparition du terme dans un document d'une part et dans la collection d'autre part. Par TF, on désigne une mesure qui traduit l'importance d'un terme dans un document, qualifiée de représentativité locale. Par IDF, on mesure si un terme, non uniformément distribué, est discriminant dans un document. La notion de discriminant se réfère à la qualité d'un terme qui distingue bien un document des autres documents de la collection. C'est-à-dire, un terme qui a une valeur de discrimination élevée doit apparaître seulement dans un petit nombre de documents. IDF a été proposée en 1972 par Spark Jones et traduit la représentativité globale d'un terme dans la collection. De nombreux travaux ont permis d'établir quelques bases théoriques pour cette fonction [Robertson, 04]. Intuitivement, on pense que l'augmentation de la taille de la collection aura un impact important sur les valeurs TF et IDF en raison de l'augmentation du nombre de documents et /ou du nombre total de termes dans la collection ou les deux. Une étude montre que l'influence des facteurs TF et IDF va en diminuant avec la taille des collections [Beigbeder et Mercier, 03]. Le passage à de grandes collections amplifie le problème de discrimination des termes puisque le nombre de termes fréquents n'augmente pas forcément avec la même proportion que le nombre de termes discriminants. Néanmoins, la valeur exacte de la fréquence des termes semble peu influencer sur les résultats de la recherche [Franz et McCarley, 02].

4. Evaluation expérimentale de l'impact du volume sur la pertinence des index documentaires

Dans le cadre de la présente étude, nous abordons le problème du passage à l'échelle en tentant d'évaluer l'impact du volume d'informations sur les représentativités locales et globales des termes d'indexation. Ces types de représentativité sont respectivement traduits à travers les poids calculés selon le schéma TF-IDF. Notre contribution consiste globalement à analyser la variation de la structure des index documentaires pour des collections de tailles différentes puis évaluer l'impact des paramètres associés à la formule de pondération des termes, sur l'efficacité de la recherche.

4.1. *Cadre expérimental*

L'étude expérimentale est réalisée en utilisant :

- Le SRI *Mercure* [Boughanem, 92] qui est un système basé sur un modèle connexioniste. Le schéma de pondération des index documentaires est de la forme TF IDF. Plus précisément, la formulation testée et retenue dans Mercure pour mesurer l'importance des termes dans les documents est la suivante :

$$d_{ij} = \frac{tf_{ij} * \left(h_1 + h_2 * \log\left(\frac{N}{n_i}\right)\right)}{h_3 + h_4 * \frac{Doclen_j}{AvgDoclen} + h_5 * tf_{ij}} \qquad [1]$$

Où :dij : poids du lien reliant le terme ti au document j , tfij : fréquence d'apparition du terme i dans le document j , ni : nombre de documents contenant le terme *i ,N* : nombre totale de documents dans la collection, *log (N/n) = idf* (le nombre de documents qui contiennent le terme i), *Docleni* = nombre de termes dans le document j (longueur de document dj), *AvgDoc* = nombre moyen de termes dans un document, *h1, h2, h3, h4, h5* : les paramètres constants. Des expérimentations préliminaires ont permis de poser : *h1= 0.1, h2= 0.8, h3= 0.2, h4= 0.7 et h5 = 1*

- Des collections de test fournies dans le cadre de la campagne d'évaluation TREC. Le tableau 1 présente les caractéristiques de ces collections après indexation par le système *Mercure*.
- Des mesures d'évaluation basées sur la précision moyenne et précision exacte.

4.2. *Analyse statistique sur la structure des collections de test*

Dans le but de caractériser l'impact des représentativités locales et globales des index documentaires issus des collections de test, nous avons en premier lieu procédé à la mesure détaillée de trois principaux paramètres : variation de la longueur des documents, variation de la fréquence documentaire et distribution des termes dans la collection. Cette analyse nous permettra d'orienter les expérimentations à mener pour l'évaluation de la recherche.

Caractéristique	***OSHUMED***	***GOV***	***WT10G***
Nombre de documents [4]	54,708	1,034,443	1,691,809
Nombre de termes	46,095	1,679,541	3,024,452
Nombre moyen de termes dans un document	129.1	250	267
Nombre moyen de termes distincts dans un document	77.8	143.5	156.104898
Volume de la collection à la base [5]	400 MO	18,1 GO	10 GO
Nombre de requêtes	63	50	50

Tableau 1. *Caractéristiques des collections de test*

4. Le nombre de documents qui sont indexés par le système Mercure.
5. Les valeurs présentées sont les volumes des collections compressées.

4.2.1. *Variation de la longueur des documents*

L'objectif de cette analyse est de mesurer l'impact différencié de l'augmentation de la taille des collections, en nombre de documents, sur la taille de l'index et longueur exacte des documents calculée sur la base du nombre de termes distincts contenus dans les descripteurs associés. Le tableau 2 présente les mesures obtenues sur les différentes collections de test. On note : T le nombre total de termes d'indexation dans la collection, $\overline{t(d)}$ le nombre moyen de termes distincts dans la collection, $\overline{t_c(d)}$ le taux d'indexation moyen dans la collection, $max_d\ (t(d))$ le nombre de termes distincts du plus long document et $max_c\ (t(d))$ le taux d'indexation maximal dans la collection. Ce tableau montre globalement un résultat prévisible qui est l'augmentation de la taille des index documentaires en fonction du nombre de documents contenus dans la collection. Cependant, les figures 1 et 2 montrent que les taux d'indexation sont d'autant plus importants dans les documents contenus dans les petites collections. Ceci infirme une supposition intuitive de l'accroissement proportionnel des longueurs des index documentaires en fonction de la taille de l'index global de la collection. L'analyse suivante complète les résultats obtenus lors de cette première analyse.

4.2.2. *Répartition des documents en fonction de leur longueur*

Dans le but de cerner la distribution des termes d'indexation dans les documents contenus dans chaque collection, nous avons défini dix intervalles de taux d'indexation [0 0.1],]0.1 0.2], …,]0.9 1].

Caractéristique	***OSHUMED***	***GOV***	***WT10G***
T	46,095	1,679,541	3,024,452
$\overline{t(d)}$	77.8	143.5	156.1
$\overline{t_c(d)}$	0.1688	0.0085	0.0051
$max_d\ (t(d))$	332	7733	10000
$max_c\ (t(d))$	0.7203	0.4604	0.3306

Tableau 2. *Représentation des taux d'indexation dans les collections*

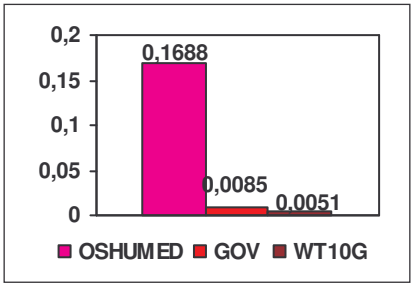

Figure 1.
Variation des taux d'indexation moyens en fonction des collections

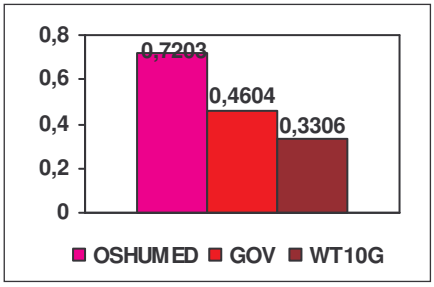

Figure 2.
Variation des taux d'index Max En fonction des collections

Pour chaque intervalle, nous avons ensuite calculé le nombre de documents dont les descripteurs ont des taux indexation inscrits dans l'intervalle correspondant. L'objectif de cette analyse est de mesurer l'impact différencié de l'augmentation de la taille des collections, en nombres de documents, sur la longueur des documents dans chaque collection. A cet effet nous nous sommes intéressés à deux types de longueur. La longueur exacte représentée par le nombre de termes d'indexation distincts contenus dans un document, et la longueur réelle représentée par le nombre d'occurrences des termes d'indexation contenues dans un document.

4.2.2.1. Répartition des documents en fonction de leur longueur exacte

Nous avons calculé les longueurs exactes des documents dans les différentes collections de test puis calculé pour chaque intervalle de taux d'indexation, le nombre et pourcentage de documents correspondants contenus dans chaque collection. Les résultats obtenus sont présentés sur les tableaux 3 et 4. On remarque globalement que la distribution des documents de la collection en fonction des taux d'indexation est plus étendue dans le cas de la plus petite collection *OSHUMED* que dans le cas des grandes collections *GOV* et *WT10G*. Plus précisément, on remarque que 90% des documents dans la collection *OSHUMED* ont des taux d'indexation dans l'intervalle [0.1 0.7]. En revanche, dans le cas des collections *GOV* et *WT10G*, plus de 90% des documents ont un taux d'indexation compris dans l'intervalle [0 0.1].

Taux d'indexation	*Nombre de documents (%)*		
	OSHUMED	*GOV*	*WT10G*
[0 0.1]	18132 (33.14 %)	1,016,116 (98.23 %)	1,674,520 (98.98 %)
] 0.1 0.2]	4578 (8.37 %)	16,936 (1.64 %)	14,271 (0.84 %)
] 0.2 0.3]	9040 (16.52 %)	1,236 (0.12 %)	1,779 (0.11 %)
] 0.3 0.4]	12052 (22.03 %)	119 (0.01 %)	497 (0.03 %)
] 0.4 0.5]	7656 (13.99 %)	31 (0.00 %)	189 (0.01 %)
] 0.5 0.6]	2505 (4.58 %)	4 (0.00 %)	94 (0.00 %)
] 0.6 0.7]	605 (1.11 %)	0 (0.00 %)	65 (0.00 %)
] 0.7 0.8]	115 (0.21 %)	0 (0.00 %)	55 (0.00 %)
] 0.8 0.9]	21 (0.04 %)	0 (0.00 %)	36 (0.00 %)
] 0.9 1]	4 (0.01 %)	1 (0.00 %)	303 (0.02 %)

Tableau 3. *Répartition des documents en fonction du taux d'indexation exact*

On remarque que la plupart des documents ont un taux d'indexation compris entre 1% et 3%. On montre ainsi globalement, que la longueur des documents est plus proche de la moyenne dans le cas de la petite collection alors que dans le cas des collections volumineuses, la plupart des documents ont une petite longueur. En combinant avec le résultat de la précédente étude, on peut supposer que l'augmentation du volume des collections a un impact plus important sur la diversité des termes contenus dans les documents que sur leur nombre. Ceci, nous amène à dire que l'accroissement du volume engendre d'avantage **l'hétérogénéité des index documentaires** plutôt qu'un accroissement de leur longueur exacte.

	OSHUMED	*GOV*	*WT10G*
$max_d(t(d))$	332	7,733	10,000
$\overline{t(d)}$	77.8	143.5	156.1
$(\overline{t(d)} / max_d(t(d)))$ %	23.43	1.86	1.56

Tableau 4. *Caractéristique des documents en fonction du taux d'indexation exact*

4.2.2.2. Répartition des documents en fonction de leur longueur réelle

Nous avons calculé les longueurs réelles des documents dans les différentes collections de test puis calculé pour chaque intervalle de taux d'indexation, le nombre et pourcentage de documents correspondants dans chaque collection. Les résultats obtenus sont présentés sur les tableaux 5 et 6.

Taux d'indexation	*Nombre de documents (%)*		
	OSHUMED	*GOV*	*WT10G*
[0 0.1]	18,780 (34.33 %)	971,759 (93.94 %)	1,691,245 (99.97 %)
] 0.1 0.2]	6,456 (11.80 %)	39,140 (3.78 %)	448 (0.03 %)
] 0.2 0.3]	8,793 (16.07 %)	11,722 (1.13 %)	85 (0.01 %)
] 0.3 0.4]	9,585 (17.52 %)	5,085 (0.49 %)	13 (0.00 %)
] 0.4 0.5]	6,610 (12.08 %)	3,373 (0.33 %)	4 (0.00 %)
] 0.5 0.6]	3,192 (5.83 %)	2,549 (0.25 %)	3 (0.00 %)
] 0.6 0.7]	854 (1.56 %)	719 (0.07 %)	0 (0.00 %)
] 0.7 0.8]	359 (0.66 %)	79 (0.01 %)	1 (0.00 %)
] 0.8 0.9]	76 (0.14 %)	13 (0.00 %)	2 (0.00 %)
] 0.9 1]	3 (0.06 %)	3 (0.00 %)	3 (0.00 %)

Tableau 5. *Répartition des documents en fonction du taux d'indexation réel*

Les tableaux 5 et 6 montrent que la répartition des documents des différentes collections en fonction de leurs longueurs réelle est comparable à la répartition en fonction de leurs longueurs exactes (paragraphe 4.2.2.1). En rapprochant ces résultats à la formule de pondération des index retenue dans le système Mercure, on peut supposer que dans le cas de grandes collections :

1. les valeurs de fréquence des termes contenus dans les collections volumineuse sont relativement proches,
2. le facteur $\left(\frac{Doclen}{AvgDoclen}\right)$ est relativement stable pour la plupart des documents.

Ceci, nous amène à dire que dans le cas de l'indexation des collections *GOV* et *WT10G*, ce deux facteurs *tf* et $\left(\frac{Doclen}{AvgDoclen}\right)$ sont peu discriminatoires.

	OSHUMED	*GOV*	*WT10G*
$max_d(t(d))$	570	12,271	276,861
$\overline{t(d)}$	129.1	250	267
$(\overline{t(d)} / max_d(t(d)))$ %	22.65	2.04	0.10

Tableau 6. *Caractéristique des documents en fonction du taux d'indexation réel*

4.2.3. *Analyse de la fréquence documentaire des termes*

L'objectif de cette analyse est de caractériser la fréquence des termes dans la collection. Il existe différentes fonctions de *idf* (la fréquence des termes dans la collection). La plupart de ces fonctions utilisent la valeur N (le nombre total de documents dans la collection) et n_i (le nombre de documents contenant le terme i). Pour caractériser la distribution des termes des différentes collections en fonction de leurs fréquences documentaires, nous nous somme basées sur le calcul de df (document frequency) qui égale à (n_i / N). A cet effet, nous avons défini dix intervalles de fréquence documentaire [0 0.1],] 0.1 0.2], …,] 0.9 1]. Ensuite pour chaque intervalle, nous avons calculé le nombre de termes d'indexation dont le *df* est inscrit dans l'intervalle correspondant. Les résultats obtenus sur les collections *OSHUMED*, *GOV* et *WT10G* sont présentés dans les tableaux 7 et 8. Ces tableaux montrent que, plus de 99% des termes ont une fréquence documentaire comprise dans l'intervalle [0 0.1] dans les trois collections. Ceci se traduit par le fait que la plupart des termes sont rares, c'est-à-dire peu fréquents dans l'ensemble des collections. Par conséquent, les valeurs associées du paramètre *idf* de ces termes sont rapprochées.

	OSHUMED	*GOV*	*WT10G*
DF_{min}	0.000,018	0.000,001	0.000,001
DF_{max}	0.655,218	0.192,905	0.118,177
DF_{avg}	0.001,690	0.000,077	0.000,040

Tableau 7. *Caractéristique des termes en fonction de la fréquence documentaire*

4.3. *Impact du schéma de pondération sur l'efficacité de la recherche*

La seconde étape de notre étude consiste à évaluer l'efficacité de la recherche sur les trois collections de test. Compte tenu des résultats obtenus lors de l'analyse statistique précédente, on s'est orientés vers l'ajustement de la formule de pondération des index documentaires utilisés dans Mercure et l'évaluation expérimentale de l'impact de cet ajustement sur la précision de la recherche. Plus précisément, on s'intéressera à l'évaluation de l'impact des paramètres fréquence documentaire inverse des termes et longueur des documents.

4.3.1. *Evaluation de la recherche de base*

Le but de cette expérimentation est de montrer les performances du SRI sur les trois collections de test en utilisant la formule de pondération retenue dans le système *Mercure.*

Taux de Df	*Nombre de termes (%)*		
	OSHUMED	*GOV*	*WT10G*
[0 0.1]	45,986 (99.764 %)	1,679,316 (99.987 %)	3,024,251 (99.993 %)
] 0.1 0.2]	81 (0.176 %)	179 (0.011 %)	60 (0.0019 %)
] 0.2 0.3]	23 (0.050 %)	0 (0.000 %)	0 (0.000 %)
] 0.3 0.4]	3 (0.007 %)	0 (0.000 %)	0 (0.000 %)
] 0.4 0.5]	0 (0.000 %)	0 (0.000 %)	0 (0.000 %)
] 0.5 0.6]	1 (0.002 %)	0 (0.000 %)	0 (0.000 %)
] 0.6 0.7]	1 (0.002 %)	0 (0.000 %)	0 (0.000 %)
] 0.7 0.8]	0 (0.000 %)	0 (0.000 %)	0 (0.000 %)
] 0.8 0.9]	0 (0.000 %)	0 (0.000 %)	0 (0.000 %)
] 0.9 1]	0 (0.000 %)	0 (0.000 %)	0 (0.000 %)

Tableau 8. *Répartition des termes en fonction de la fréquence documentaire*

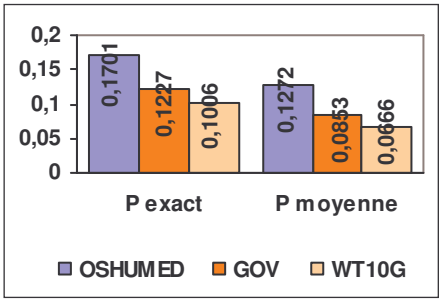

Figure 3. *Précision exacte et moyenne*

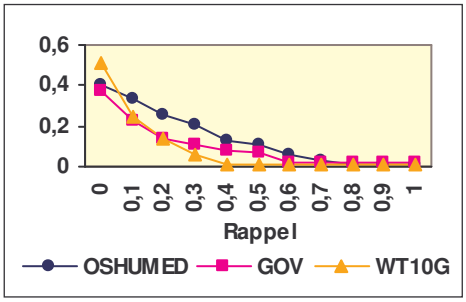

Figure 4. *Courbes des rappel- précision*

La figure 3 montre que les précisions exacte et moyenne sur la plus petite collection *OSHUMED* sont meilleures que celles obtenues sur les plus grandes collections *GOV* et *WT10G.* Les courbes des rappel-précision issues de l'évaluation

des trois collections, sont présentées dans la figure 4. Ceci montre globalement que le SRI présente des performances qui se dégradent en fonction de l'augmentation des volumes des collections interrogées.

4.3.2. *Evaluation de l'impact du paramètre IDF*

L'objectif de cette expérimentation est l'évaluer l'impact du paramètre *idf* sur la précision de la recherche dans le cas des différents collections. A cet effet, on a comparé les résultats obtenus, en utilisant deux différentes valeurs du paramètre h_2 de la formule d'indexation retenue dans le système Mercure. h_2 =0.8 est la valeur standard, h_2 =0 est la valeur expérimentée qui traduit l'absence du paramètre *idf* dans la formule d'indexation. Les tableaux 9, 10 et 11 montrent les résultats obtenus et les figures 5 ,6 et 7 illustrent respectivement ces résultats sur les collections de test. D'après les figures 5, 6 et 7 on remarque que, la présence du paramètre *idf* (h_2=0.8) améliore les précisions de la recherche pour trois collections. En outre, on remarque plus précisément que les taux d'accroissement moyenne de la précision croissent de manière proportionnelle à la taille des collections : (8.72 % pour la collection *OSHUMED*, 26.55 % pour la collection *GOV* et 58.19 % pour la collection *WT10G*). Ce résultat confirme la supposition posée dans le paragraphe 4.2.2 concernant les valeur des paramètres $\left(\frac{Doclen}{\text{particulier}}\right)$ et *tf*. En effet, ces paramètres étant peu discriminatoires dans le cas particulier des grandes collections, ils s'en suivent que l'annulation de paramètre *idf* dans la formule de pondération dégrade d'avantage la précision de la recherche.

ε_1 : Taux d'accroissement en % (entre les cas h_2=0 et h_2=0.8).

OSHUMED	h_2=0	h_2=0.8	$\varepsilon_{1\,(\%)}$
P5	0.1873	0.2159	15.27
P10	0.1571	0.1698	8.08
P15	0.1376	0.1386	0.73
P20	0.1230	0.1270	3.25
P30	0.0820	0.0847	3.29
P100	0.0246	0.0254	3.25
P200	0.0123	0.0127	3.25
P500	0.0049	0.0051	4.08
P1000	0.0025	0.0025	0.00
P Exact	0.1656	0.1701	2.72
P Moyenne	0.1170	0.1272	8.72

Tableau 9. *Taux d'accroissement de précision entre les cas (h_2=0 et h_2=0.8) (Collection OSHUMED)*

GOV	*h2=0*	*h2=0.8*	$\varepsilon_{I\,(\%)}$
P5	0,1878	0,2286	21.73
P10	0,1612	0,1755	8.87
P15	0,1497	0,1551	3.61
P20	0,1367	0,1490	9.73
P30	0,0912	0,0993	8.88
P100	0,0273	0,0298	9.16
P200	0.0137	0.0149	8.76
P500	0.0055	0.0060	9.09
P1000	0,0027	0,0030	11.11
P Exact	0,0948	0,1227	29.43
P Moyenne	0,0674	0,0853	26.55

Tableau 10. *Taux d'accroissement de précision entre les cas (h_2=0 et h_2=0.8) (Collection GOV)*

WT10G	*h2=0*	*h2=0.8*	$\varepsilon_{I\,(\%)}$
P5	0,2000	0,2960	48.00
P10	0,1900	0.3000	57.89
P15	0,1773	0.2653	49.63
P20	0,1750	0,2590	48.00
P30	0,1167	0,1727	47.99
P100	0,0350	0,0518	48.00
P200	0.0175	0.0259	48.00
P500	0.0070	0.0104	48.57
P1000	0,0035	0,0052	48.57
P Exact	0,0708	0,1006	42.09
P Moyenne	0,0421	0,0666	58.19

Tableau 11. *Taux d'accroissement de précision entre les cas (h_2=0 et h_2=0.8) (Collection WT10G)*

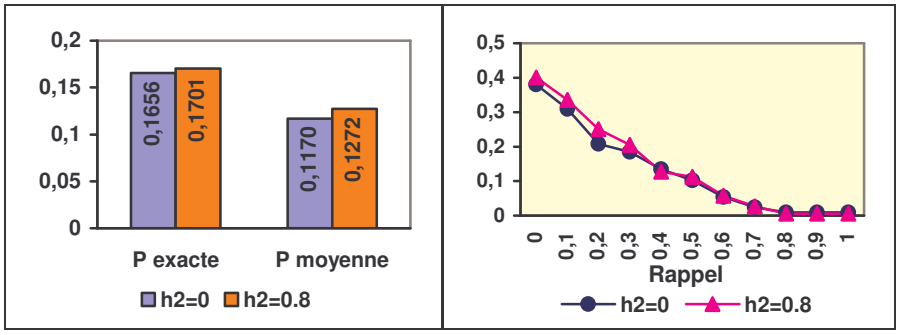

Figure 5. *Précision de la recherche (Collection OSHUMED)*

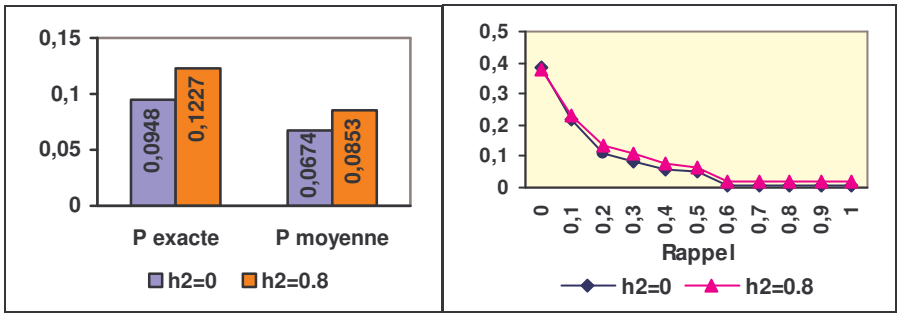

Figure 6. *Précision de la recherche (Collection GOV)*

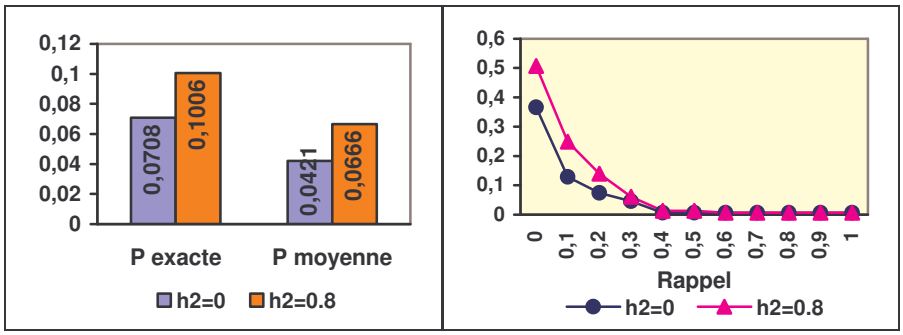

Figure 7. *Précision de la recherche (Collection WT10G)*

5. Conclusion

Inscrite dans le cadre de la problématique générale du passage à l'échelle, notre étude a pour objet d'analyser globalement l'impact du volume sur l'efficacité de la recherche et ce, sous l'angle du principe de construction des index documentaires. Plus précisément, on a focalisé sur l'analyse des facteurs longueur des documents et fréquence des termes, qui sont largement utilisés pour l'expression des schémas de pondération. Cette étude nous a permis de constater que le passage à de grandes collections amplifie le problème de discrimination des termes puisque le nombre de termes fréquents n'augmente pas de manière significative et que la proportion des termes discriminants diminue. Au terme de cette étude, nous pouvons tout d'abord dire que dans les grandes collections, la taille de la collection augmente en nombre de documents mais le nombre de termes distincts dans un document n'augmente pas dans la même proportion. Nous avons également montré que les longueurs des documents dans les grandes collections restent stables et que la plupart des documents ont une longueur proche de la longueur moyenne des documents dans le cas des grandes collections. En outre, l'étude a révélé que la plupart des termes dans les collections, particulièrement dans le cas des grandes collections, sont rares, ce

qui accroît d'avantage l'hétérogénéité des index que leur longueur. Ceci a de toute évidence un impact sur la valeur de *idf* mais également sur la proportion de termes discriminants.

Enfin, les expérimentations nous ont montré que le fait de « bien » discriminer les paramètres de la formule d'appariement, en l'occurrence, *tf*, *idf* et *longueur des documents*, permettrait d'améliorer les précisions des résultats des systèmes de recherche d'information dans le cas de collections volumineuses. Dans ce sens, des investigations sont envisagées dans de futurs travaux dans le but de mieux caractériser les distributions des termes dans les collections et ce, afin de mettre en évidence les corrélations entre ces facteurs et proposer des schémas de pondération dépendant de la taille des collections interrogées.

6. Bibliographie

[Aizawa, 03] Akiko Aizawa , *An Information-Theoretic Perspective of tf-idf Measures*, Information Processing & Management, Vol.39, No.1, pp 45-65 , 2003.

[Bailey et al., 96] P. Bailey, and D. Hawking, *A Parallel Architecture for Query processing over A Terabyte of Text*, CS Tech Report, The Australian National University, June 1996.

[Beigbeder et Mercier, 03] Michel Beigbeder and Annabelle Mercier. *Etude des distributions de tf et de idf sur une collection de 5 millions de pages html.* In Atelier de recherche d'informations sur le passage à l'échelle Congrès INFORSID 2003, Nancy, France, juin 2003.

[Berry et al., 99] M.W Berry, Z. Darmac, E.R jessup, Martrices, *vector spaces and information retrieval*, SIAM Review, Vol 41(2),pp 335-362,1999.

[Boughanem, 92] Boughanem M., *Les systèmes de Recherche d'Information : d'un modèle classique à un modèle connexionniste*, Thèses de Doctorat de l'Université Paul Sabatier, Toulouse, 1992.

[Boughanem et al., 04] Boughanem M., Tamine L., Chevallet J.P., Martinez J., Calabretto S., Rapport final de l'AS-91 du RTP-9 *"Passage à l'échelle dans la taille des corpus"*, 2004, site Web: http://www.irit.fr/ASVolume/.

[Franz et McCarley, 02] Martin Franz and J. Scott McCarley. *How many bits are needed to store term frequencies?* In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp 377-378. ACM Press, 2002.

[Hawking et al., 99] David Hawking, Nick Craswell, and Peter Bailey. *Overview of the TREC-8 Web Track*. In Proceedings of the 8th Text Retrieval Conference, pp 131-150. NIST Special Publication, 1999.

[Jain et Goharian, 02] A. Jain, N. Goharian, *On Parallel Implementation of Sparse Matrix Information Retrieval Engine*, The 2002 International Multi-conferences in Computer Science: on Information and Knowledge Engineering (IKE), 2002.

[Kobayashi et Takeda, 00] M. Kaboyashi, K. Takeda, *Information retrieval on the web*, ACM Computing Surveys,32(2), pp 144 -173, 2000.

[Kokiapoulou et Saad, 04] E. Kokiapoulou, Y. Saad. Polynomial filtering in latent semantic indexing for information retrieval. In Proceedings of the 27th annual international ACM SIGIR Conference on research and development in information retrieval, pp 104-111, Sheffield 25-29 July 2004.

[Lee et Ren, 96] Dik Lun Lee, Liming Ren. *Document Ranking on Weight-Partitioned Signature Files*. ACM Transactions, 14(2), pp 109-137 (1996).

[Luhn, 57] Luhn, H. P. (1957). *A statistical approach to mechanized encoding searching of literary information*. IBM Journal of Research and Development, pages 309 – 317.

[Moffat et Zobel, 96] Alistair Moffat, Justin Zobel: *Self-Indexing Inverted Files for Fast Text Retrieval*. ACM Trans. Inf. Syst. 14(4): 349-379 (1996).

[Newby, 00] Gregory B.Newby. *The science of Large-Scale Information Retrieval*. Internet Archive Symposium 2000.

[Robertson, 04] Robertson, S.E. (2004), "*Understanding inverse document frequency: on theoretical arguments for IDF*", Journal of Documentation, Vol. 60 No. 5, pp 503-520, 2004.

[Salton et Buckley, 88] G. Salton, C. Buckley, *Weighting approaches in automatic text retrieval*. Information processing and Management, 24(5), 513-523.

[Sparck Jones, 72] Sparck Jones, K. (1972), *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, 28(1), pp. 11-21, 1972.

[Sparck Jones, 04] Sparck Jones, K. (2004), *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, Vol. 60, No. 5, pp. 493-502, 2004.

[Tang et al, 2004] C. Tang, S. Dwarkadas, Z. Xu. On scaling latent semantic indexing for large peer to peer systems. In Proceedings of the 27th annual international ACM SIGIR Conference on research and development in information retrieval, pp 112-121, Sheffield 25-29 July 2004.

[Williams et Zobel, 03] H.E Williams , J. Zobel, *Searchable words on the web*. International journal of digital libraries, 2003.

[Witten et al., 99] T.B.I Witten, A. Moffat. *Managing Gigabytes, compressing and indexing documents and images.* Morgan Kaufmann publishers, second edition, Butterworths, 1999.