Evaluer le passage à l'échelle dans des environnements à pertinence multivaluée

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. *La croissance continuelle et exponentielle des volumes d'information numérique affecte principalement des domaines comme celui de la Recherche d'Information (RI). Toutefois, peu de travaux en 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. Face à la masse grandissante d'information, il est préférable du point de vue de l'utilisateur moyen que les documents retournés soient classés par ordre de pertinence décroissante ; ce qui implique de prendre en compte de multiples niveaux de pertinence pour les documents. Nous établissons un lien entre ces deux pans de la RI (pertinence multivaluée et passage à l'échelle) et nous étudions des protocoles pouvant permettre d'évaluer l'abilité des systèmes de RI à retourner les documents de meilleur niveau de pertinence en tête des résultats quand la masse d'information croît.*

ABSTRACT. *Nowadays, many factors support a growing production of information. In modern large environments, for the user's point of view, it is desirable to have IRS that retrieve documents according to their relevance levels. Relevance levels have been studied in some previous IR works while some others (few) IR research works tackled the questions of IRS effectiveness and collections size. These latter works used standard IR measures on collections of increasing size to analyze IRS effectiveness scalability. In this work, we bring together these two issues in IR (multigraded relevance and scalability) by designing some new metrics for evaluating the ability of IRS to rank documents according to their relevance levels when collection size increases.*

MOTS-CLÉS : *passage à l'échelle, pertinence multivaluée, métriques, évaluation en RI*

KEYWORDS: *scalability, multigraded relevance level, metrics, IR evaluation*

1. Croissance d'information

De nombreux facteurs facilitent une production croissante d'information. Sur le web, la masse d'information correspondant aux pages statiques et aux pages calculées sans paramètres est estimée à 170 teraoctects. Une augmentation régulière de 30% par an pour ces pages a été constatée entre 1999 et 2002 [LYM 03]. Une étude estime la quantité d'information produite en 2002 à 5 exaoctects dont 92% serait stockée de façon numérique. On peut s'attendre à une continuité dans cette croissance car le nombre d'utilisateurs est sans cesse grandissant et leurs activités s'intensifient avec la mise à disposition de nombreux services (bibliothèques numériques, moteurs de recherche, annuaires, bases de données, ...). Plusieurs autres raisons justifient cette croissance d'information numérique : l'accès à des espaces de stockages de plus en plus vastes, les techniques de digitalisation sont de plus en plus performantes et il ya un besoin plus pressant d'échanger rapidement et efficacement de l'information. La question de la structuration et de l'accès aisé à cette masse d'information se pose dans des domaines comme la RI, les bibliothèques numériques, etc. Le repertoire de bibliothèques sur le web maintenu à l'Université de Californie (Libweb) liste couramment environ 7100 pages de bibliothèques dans près de 115 pays, de forme et de contenu très variables. De nombreux projets continuent de voir le jour [SAR 04], [RAI 00]. Le *web visible* (2,5 millions de documents et une augmentation quotidienne d'environ 7,3 millions de pages [BER 01],) et le *web invisible* (550 miliards de documents interconnectés ; environ 7.500 Teraoctects d'espace de stockage - soit 400 à 550 fois plus vaste que le *web visible* [BER 01]) constitue les deux zones du web.

Les estimations sur la quantité d'information véhiculée par messagerie électronique varient de 610 milliards à 1.100 milliards de messages envoyés(année 2001) et on peut estimer le nombre total de listes de diffusion à environ 36,5 milliards par an[LYM 03].

Face à cette croissante impressionnante d'information, de nombreux outils de recherche sont mis en oeuvre pour faciliter l'accès à l'information pertinente . comScore Media Metrix liste 166 moteurs de recherche en ligne en Mai 2005 ; 82,2% d'utilisateurs en ligne utilisent un moteur de recherche et la RI est la seconde activité la plus pratiquée en ligne d'après une étude de Pew Internet réalisé en 2005 [DOG 05].

Une étude récente montre que seulement 3,2% des 1ères pages retournées sont identiques pour les trois meilleurs moteurs de recherche pour une requête donnée et les moteurs de recherche ont en commun un de leurs cinq premiers documents dans moins de 20% de cas [DOG 05]. En effet, la taille du web couplée à son dynamisme et à l'ambiguité (la subjectivité) des requêtes-utilisateurs font qu'il est difficile pour un moteur de recherche de fournir l'information la plus actualisée en temps réelle. Les études de Ding en 1996, de Lawrence et Giles[1] en 1998 et de A. Gulli et A. Signorini[2] pointent le peu de recouvrement entre les résultats des moteurs de recherche

1. Searching the www. Science, 280, 98-100
2. A comparative study of web search service performance, Proceedings of the ASIS'99, 136-142

interrogés avec les mêmes requêtes. Ainsi en menant une recherche uniquement sur Google, un utilisateur peut "rater" jusqu'à 70% des meilleures premières pages de résultats possibles. Une des solutions apportées à ce problème de peu de recouvrement entre différents moteurs de recherche est l'utilisation de métamoteurs de recherche (exemple : www.dogpile.com, www.seek.fr, ...etc). Des travaux de recherche se sont interessés au phénomène de croissance d'information et à son impact sur les différentes phases d'un processus de RI.

2. Evaluation dans des environnements larges : travaux antérieurs

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. Imafouo et al. [IMA 05] donne une vue d'ensemble des travaux de RI associés à ces différentes phases qui se sont interessés au passage à l'échelle. Pour ce qui concerne la phase d'évaluation, les travaux existants portent soit sur des métriques d'évaluation de SRI dans des collections volumineuses soit sur l'évaluation des performances des SRI quand la taille de collections augmente.

L'utilisation de collections volumineuses pour l'évaluation de SRI va 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. En effet, avec la croissance continue en taille des espaces de recherche, les techniques comme celle du *pooling* utilisé dans le cadre de TREC fournissent des collections de test avec des jugements de pertinence de moins en moins complets. Les métriques d'évaluation classiques ne sont pas robustes face à cette grande incomplétude. Les travaux de Voorhees [VOO 04] proposent une métrique d'évaluation basée sur les positions relatives des documents pertinents et des documents *jugés* non pertinents. Cette métrique (la *bpref*) est correlée aux métriques classiques quand des jugements de pertinence "complets" sont disponibles et elle est plus robuste à l'incomplétude des jugements de pertinence. Il est donc probable d'après cette étude que de grandes collections de test construites en utilisant le *pooling* resteront des bases d'expérimentations fiables malgré l'incomplétude des jugements de pertinence.

Concernant l'évaluation des performances de SRI face à la croissance de collections, les participants de la tâche Very large Collection (VLC) de TREC-6 ont noté une augmentation significative de la précision dans les premiers documents quand on passe d'un échantillon de la collection à la collection totale [HAW 99]. Sur cette base, Hawking et al. [HAW 03] ont établi différentes distributions de scores de documents pertinents et de documents non pertinents et les échantillons analytiques de collection construites à partir de ces distributions de score confirment les observations sus-cités ; puis de façon expérimentale, la haute précision est observée sur différents types d'échantillons de taille croissante : une amélioration de la précision sur les pre-

miers documents retournés quand la taille d'échantillons croît. Cette amélioration est justifiée par le nombre de documents pertinents plus élévé dans les échantillons plus grands, mais aussi dans la capacité du couple (requête, SRI) à classer les documents pertinents avant les documents non pertinents(puisque la précision diminue avec la taille de l'échantillon même dans le cas où elle n'est pas limitée par le nombre de documents pertinents présents dans l'échantillon).

Les travaux récents de Imafouo *et al.* [IMA 05] fournissent une méthodologie d'échantillonnage d'une grande collection en sous-collections ayant des caractéristiques communes. Ainsi, l'étude de l'influence de la taille sur les propriétés est affranchie du biais des caractéristiques (parfois très différentes entre sous-collections). Ces travaux ont montré que la rôle que joue la collection de documents dans le processus d'attribution de score à un document donné a un impact sur les performances du modèle quand on passe à l'échelle. Ces travaux étendent également les résultats de Hawking *et al* [HAW 03] sur l'amélioration de la haute précision avec la croissance en taille de collections à plusieurs modèles de RI.

Tous ces travaux ont utilisé les métriques classiques d'évaluation utilisées en RI et qui considèrent la pertinence comme une notion binaire. Toutefois, des travaux ont montré que la pertinence peut être mise sur une échelle à plusieurs valeurs.

3. Pertinence Multivaluée

Pour les campagnes d'évaluation et de comparaison de SRI comme TREC, on utilise la technique du *pooling* pour former un ensemble de jugements de pertinence. Cette pertinence est généralement binaire. Cependant [REE 67], [SAR 75], [COO 71], [WIL 73] proposent des cadres de classifications des notions de pertinence, suggérant que la pertinence est un phénomène social et cognitif complexe et qu'il n'y a pas une seule pertinence. Différents degrés de pertinence ont été étudiés dans des travaux antérieurs. Tang *et al.* [TAN 99] montre qu'une échelle à sept points est optimale en terme de confiance dans les les jugements de pertinence. Spink *et al.* [SPI 98] utilisent trois niveaux de pertinence (*non pertinent, partiellement pertinent, pertinent*) dans les jugements de pertinence fournis par des utilisateurs réels. La collection du Web Track de TREC 2001 avait également adopté une échelle à trois niveaux pour la pertinence des documents [VOO 01], [BAI 01].

Pour Kekäläinen *et al.* [KEK 02], dans les environnements larges et modernes de RI, il est désirable d'avoir des SRI qui retournent des documents en fonction de leur niveau de pertinence. L'évaluation en RI doit donc pouvoir "récompenser" les SRI qui retournent les documents ayant le plus haut niveau de pertinence en tête des autres documents. Pour ce faire, il est necessaire de prendre en compte les différents niveaux de pertinence d'un document par rapport à un besoin d'information. Ces auteurs utilisent une collection avec une échelle à 4 niveaux de pertinence : (*hautement pertinent, suffisamment pertinent, marginalement pertinent et non pertinent*). Chacun de ces niveaux est quantifié par une valeur numérique et une des questions est le choix de ces

valeurs et le sens dont elles peuvent être porteuses : exemple : le niveau de pertinence "*hautement pertinent*" a la valeur 3 et le niveau "*marginalement pertinent*" a la valeur 1 ; doit-on interpreter cela comme suit : "un document hautement pertinent est 3 fois plus pertinent qu'un document marginalement pertinent" ? Ces auteurs proposent les métriques *generalised non-binary recall and precision*(précision et rappel non binaire notées gP et gR) qui sont des extensions des métriques classiques (précision et rappel) prenant en compte plusieurs niveaux de pertinence : soit R l'ensemble des n documents retournés et appartenant à l'ensemble des documents ayant un niveau de pertinence pour une requête donnée, $R \subseteq D$; chaque document d_i a le niveau de pertinence $r(d_i)$ qui est un nombre réel de l'intervalle $[0..1]$, on pose $gP = \sum_{d \in R} r(d)/n$ et $gR = \sum_{d \in R} r(d) / \sum_{d \in D} r(d)$. Comme les métriques de précision et de rappel classiques, ces métriques permettent des moyennes sur l'ensemble des requêtes, des moyennes de précision à des niveaux de rappel, des courbes de performances. Pour des environnements où les niveaux de pertinence ne sont pas compris dans l'intervalle $[0..1]$, ils peuvent être ramenés à cet intervalle par une normalisation (en faisant le rapport avec le plus grand niveau de pertinence par exemple).

Les métriques *Cumulative Gain* et *Discounted Cumulative Gain* sont également des métriques qui prennent en compte plusieurs niveaux de pertinence et proposées par Kekäläinen *et al.* [KEK 00]. Pour une collection donnée et un SRI, ces métriques calculent le gain cumulé d'information pertinente qu'on réalise au fur et à mesure qu'on parcourt la liste des résultats retournés par ce SRI sur cette collection ; dans le cas de la métrique *Discounted Cumulative Gain*, l'information pertinente à un rang donné est pondérée par une fonction décroissante du rang, avant d'être cumulée. Pour chacune de ces métriques, on obtient ainsi un vecteur qui donne pour chaque rang l'information pertinente cumulée du premier rang jusqu'à ce rang. Ces vecteurs peuvent être comparés à des vecteurs de gain d'information cumulée pour le cas d'un SRI *idéal* (SRI qui retourne les documents dans le meilleur ordre de niveau de pertinence, pour chaque requête)

Les métriques *Cumulative Gain* et *Discounted Cumulative Gain* ont été *adaptées* à l'évaluation des SRI qui travaillent sur des documents structurés (XML). Dans le cadre de la campagne *INEX*, la métrique XCG (Xml Cumulative Gain) est utilisée [KAZ 05].

Nos conceptions rejoignent celles de Kekäläinen *et al.* [KEK 02] en ce qui concerne la prise en compte de multiples niveaux de pertinence. Face à la masse importante et grandissante d'information, un des défis majeurs pour les outils de recherche sera de pouvoir retourner en tête de liste de leurs résultats des documents du meilleur niveau de pertinence. Nous concevons des métriques permettant d'évaluer cette abilité dans les SRI quand la taille des collections augmente.

4. Evaluer le passage à l'échelle

4.1. *Importance d'un niveau de pertinence*

On suppose qu'on a attribué un niveau de pertinence à chaque document par rapport à chaque topic. Soit $\{niv_i\}$, $i = 1 \ldots, n$ l'ensemble des niveaux de pertinence de documents de la collection. Sur l'ensemble des documents on définit une relation d'équivalence R^t pour chaque topic t. Deux documents sont dans la même classe d'équivalence pour un topic donné s'ils ont le même niveau de pertinence au regard de ce topic.

Sur l'ensemble des niveaux de pertinence on définit un ordre total que nous noterons $\succ$ (tous les niveaux de pertinence sont comparables deux à deux). Ainsi nous aurons dans nos notations $niv_i \succ niv_j$ si $i > j$. Cette relation d'ordre total est suffisante pour donner la préférence qu'on souhaite dans la liste des documents retournés mais elle ne donne pas d'indication sur l'importance qu'on veut donner à un niveau de pertinence par rapport aux autres (l'importance relative entre les niveaux de pertinence). Or c'est l'importance qu'on attribue à un niveau de pertinence qui caractérise en fait la qualité/quantité d'information pertinente qu'on attend d'un document ayant ce niveau de pertinence. On peut vouloir crediter (resp penaliser) fortement des systèmes qui retournent les documents ayant le plus haut niveau de pertinence en tête de liste (resp pas en tête de liste) : dans ce cas, ce plus haut niveau de pertinence doit avoir une importance élevée (par rapport aux autres niveaux de pertinence) lors de l'évaluation des SRI. Par exemple pour des applications où on ne souhaite retenir que quelques documents mais de très hauts niveaux de pertinence. Il existe toutefois aussi des applications pour lesquelles on souhaite avoir de nombreux documents de bon niveau de pertinence. Pour ce type d'applications, on ne souhaite pas trop différencier un SRI qui retourne des documents d'un bon niveau de pertinence d'un SRI qui retourne des documents d'un très haut niveau de pertinence ; l'importance du "bon niveau" de pertinence et celle du "très haut niveau" de pertinence ne seront donc pas très éloignées.

Ainsi, une fonction I qui formalise l'importance des niveaux de pertinence dépendra de ce qu'on cherche à évaluer des SRI et du type d'applications sur lesquelles on voudrait pouvoir utiliser ces SRI. Toute fonction g caractérisant l'importance des niveaux de pertinence aura les propriétés suivantes (fonction positive et croissante) :

– $I(niv_i) > 0$ et $I(niv_i) > I(niv_j)$ si $niv_i \succ niv_j$ i.e. $i > j$

Des fonctions comme $I(niv_i) = a \times i + b$ ou $I(x) = a \times i^2 + b$ peuvent modéliser l'importance des niveaux de pertinence. Exemples : On a 5 niveaux de pertinence "TP" pour Très Pertinent, "P" pour Pertinent , "FP" pour Faiblement pertinent, "NJ" pour NonJugé et "NP" pour Non Pertinent, classés comme suit $TP \succ P \succ FP \succ NJ \succ NP$. On a les deux fonctions $I_1(niv_i) = i$ et $I_2(niv_i) = i^2$.

$I_1(TP) = 4$ et $I_1(FP) = 2$. $I_2(TP) = 16$ et $I_2(FP) = 4$. Pour la fonction I_1, TP est moins "éloigné" de FP que si on choisit la fonction I_2.

Attribuer à chaque niveau de pertinence une "valeur" (qui caractérise son importance) dans l'absolu ne signifie rien, mais dans le relatif par rapport aux autres niveaux de pertinence cette valeur prend du sens.

4.2. *Gain d'information entre deux niveaux de pertinence*

Face à deux documents ayant des niveaux de pertinence différents pour un topic, on n'attend pas la même quantité d'information pertinente de chacun d'eux. Il est interessant de pouvoir quantifier l'information pertinente gagnée (ou perdue) quand on passe d'un niveau de pertinence à un autre. Ce gain est une fonction des niveaux de pertinence : $Gain(niv_i, niv_j) = g(niv_i, niv_j)$. Quelles sont les caractéristiques de cette fonction f ?

– $g(niv_i, niv_j) > g(niv_i, niv_k)$ si $niv_j \succ niv_k$ i.e. si $j > k$ (1)

– $g(niv_i, niv_j) < g(niv_k, niv_j)$ si $niv_i \succ niv_k$ i.e. si $i > k$ (2)

– $g(niv_i, niv_i) = 0$ (3) : on ne gagne (resp. perd) pas d'information pertinente quand on reste sur le même niveau de pertinence (même si on change de document, car les documents d'un même niveau sont de la même classe d'équivalence).

Par déduction de (2) et (3), on a : $g(niv_i, niv_j) < 0$ si $niv_i \succ niv_j$ i.e. si $i > j$. En effet si on a $niv_i \succ niv_j$, alors cela signifie que la quantité d'information pertinente que contient tout document de niveau de pertinence niv_i est plus grande que la quantité d'information pertinente que contient tout document de niveau de pertinence niv_j. Donc, quand on passe d'un document de niveau de pertinence niv_i à un document de niveau de pertinence niv_j, on perd de l'information pertinente. De même, de (1) et (3) on déduit : $g(niv_i, niv_j) > 0$ si $niv_j \succ niv_i$ i.e. si $i < j$

Il est logique que la fonction de gain d'information pertinente entre deux niveaux de pertinence dépendent de l'importance accordée à chacun des niveaux, i.e. de la "valeur" attribuée à chaque niveau de pertinence. Ainsi $g(niv_i, niv_j) = h(I(niv_i), I(niv_j))$.

4.3. *La distance mathématique comme exemple de fonction gain*

De façon intuitive, le *gain* d'information pertinente entre deux niveaux de pertinence exprime une certaine de notion de *distance* entre les deux niveaux de pertinence. Cette distance a les propriétés standards qui sont :

– (symétrie) : $\forall i,\ j,\ d(niv_i, niv_j) = d(niv_j, niv_i)$

– (séparation) : $\forall i,\ d(niv_i, niv_i) = 0$

– (inégalité triangulaire) : $\forall i,\ j,\ k\, d(niv_i, niv_j) \leq d(niv_i, niv_k + d(niv_k, niv_j)$

Puisque chaque niveau de pertinence est associée à une valeur numérique qui caractérise son importance, on peut construire la *distance* entre les différents niveaux de pertinence en se basant sur les valeurs numériques d'importance qui leur sont associées et $d(niv_i, niv_j) = d(I(niv_i), I(niv_j))$.

On remarque que en prenant :

$$\begin{cases} g(niv_i, niv_j) = -d(I(niv_i), I(niv_j)) \, pour \, niv_i \succ niv_j \\ g(niv_i, niv_j) = d(I(niv_i), I(niv_j)) \, pour \, niv_j \succ niv_i \end{cases}$$

on respecte bien toutes les propriétés attendues de la fonction g qui modélise le *gain* entre deux niveaux de pertinence. Ainsi, la notion de *distance* mathématique peut être utilisée pour mettre en oeuvre notre fonction de *gain.*

Exemple : Nous reprenons l'exemple précédent et nous utilisons la distance de $d(x, y) =\mid x-y \mid$; on peut avoir $g(TP, FP) = -(4-2) = -2$ si on choisit I_1 comme fonction d'importance de niveau de pertinence et $g(TP, FP) = -(16 - 4) = -12$ si on choisit plutôt I_2.

4.4. *Mesurer le passage à l'échelle*

Nous proposons des métriques pour analyser et évaluer le comportement des modèles de RI lorsqu'on utilise des ensembles de recherche de plus en plus grands. Soient deux collections $C1$ et $C2$ de taille croissante et un SRI S. On souhaite analyser / évaluer le comportement de S sur chacune des deux collections ; notamment, on voudrait déterminer si les performances de S s'améliorent, restent stables ou se détériorent lorsque la taille de la collection augmente. Pour ce faire, nous proposons deux groupes de métriques basées sur le gain d'information réalisée à chaque rang lorsqu'on parcourt les listes de résultats des différentes collections de taille croissante.

4.4.1. *Métriques 1*

Pour un topic donné, à chaque rang, on va calculer la quantité d'information pertinente gagnée à ce rang lorsqu'on change de collection. Pour un topic donné t, nous utilisons les notations suivantes :

– N est le point où l'on souhaite s'arrêter dans la liste des documents

– Tout document d retourné a un niveau de pertinence qui sera noté $NivPertinence(d)$

– Pour deux niveaux de pertinence niv_i et niv_j, on a le gain g qu'on réalise dans la qualité des résultats en passant du niveau de pertinence niv_i au niveau niv_j.

Nous définissons également

$$d^t(C_i) : \begin{pmatrix} \mathcal{N} \rightarrow C_i \\ k \mapsto d^t_k(C_i) \end{pmatrix}$$

Ainsi nous notons $Retrieved^t(C_i) = d^t(C_i)(\mathcal{N})$: la liste des documents retournés pour la collection C_i et $Retrieved^t_N(C_i) = d^t(C_i)([1, .., N])$: les N premiers

documents retournés pour la collection C_i. Pour un topic t, nous calculons le *passage* entre les listes de résultats de deux collections C_i et C_j à un rang k par

$$Passage_k^t(C_i, C_j) = g(NivPertinence(d_k^t(C_i)), NivPertinence(d_k^t(C_j)))$$

Ce passage est le *gain* d'information pertinente absolu entre le niveau de pertinence du document situé en position k de la liste de réponse de C_i pour le topic t et le niveau de pertinence du document situé en position k de la liste de réponse de C_j pour le topic t.

Pour avoir le gain d'information pertinente relatif (Gipr) à un rang, on utilise un coefficient de pondération (cp) qui dépend du rang. Ce coefficient de pondération (qui sera une fonction décroissante du rang) est necessaire pour reduire progressivement en fonction du rang l'impact du gain/perte d'information pertinente. Comme suggeré par Kekäläinen *et al.* [KEK 00], l'on peut faire cette réduction en pente rapide (avec une fonction comme l'inverse du rang $cp(k) = 1/k$) dans le cas où les documents en tête de liste sont ceux sur lesquels on veut focaliser l'évaluation ou en pente moins rapide (avec une fonction comme l'inverse du log du rang $cp(k) = 1/log_b(k)$).

Exemple : Pour $b = 2$, ${}^2log_2(2) = 1$ et $log_2(1024) = 10$. Ainsi pour le rang 1024, on garderait quand même encore $1/10$ème du gain/perte d'information pertinente alors que avec une reduction en pente rapide avec par exemple $p(k) = 1/k$, on ne garderait que $1/1024$ème du gain/perte d'information pertinent pour le rang 1024.

Ainsi, pour chaque topic t, nous obtenons un vecteur de passages

$$< Passage_1^t, \ldots, Passage_N^t >$$

. et un vecteur de passages pondérés

$$< cp(1) \times Passage_1^t, \ldots, cp(N) \times Passage_N^t >$$

Deux possibilités s'offrent à nous :

– soit on fait une somme des éléments du vecteur de passages pondérés d'un topic afin d'avoir une "valeur" unique pour chaque topic à une position donné N. Nous définissons donc la première métrique comme suit :

$$Metrique1_N^t(C_i, C_j) = \sum_{k=1}^{N} cp(k) \times Passage_k^t(C_i, C_j)$$

Ainsi pour chaque topic, nous obtenons une valeur unique à un point donné N. Cette métrique ne peut donc se calculer que sur l'ensembles de topics ayant des résultats sur chacune des deux collections C_i et C_j ; en d'autres termes sur des topics qui vérifient $\mid Retrieved^t(C_i) \mid \geq N$ et $\mid Retrieved^t(C_j) \mid \geq N$

– soit on fait une somme des élements des vecteurs (absolus ou pondérés) de tous les topics rang par rang et on a un seul vecteur final de N éléments pour les topics :

$$< cp(1) \times \sum_t (Passage_1^t(C_i, C_j)), \ldots, cp(N) \times \sum_t (Passage_N^t(C_i, C_j)) >$$

Ce vecteur somme a la même taille que les vecteurs de passages pondérés de chaque topic ; son $i$ème élément est la somme des $i$ème éléments des vecteurs 'absolus ou pondérés) de chaque topic. On peut donc visualiser ces vecteurs sous forme de courbes gain/perte (sur tous le topics) par rang.

4.4.2. *Metrique2*

Elles sont basées sur le même principe que les Métrique1. Pour une collection donnée le SRI S retourne une liste de résultats $Retrieved(C_0)$ pour un topic donné. Nous construisons la liste de résultats ***idéale*** pour ce topic comme suit :

– soit $Documents^t(niv_i)$ l'ensemble des documents niveaux de pertinence niv_i pour le topic t et présents dans la collection C_0

– N est le point où l'on souhaite s'arrêter dans la liste des documents

On crée une liste de documents $Retrieved\_ideal(C)$ qui est la liste "idéale" qui aurait dûe être retournée par le SRI pour ce topic. On introduit d'abord les "meilleurs" documents, ensuite les bons,etc...Pour chaque niveau de pertinence niv_i, en partant du meilleur vers le pire, on introduit

$$nb = \left(\begin{array}{l} N - \sum_{j=i+1}^{n} \mid Documents(niv_j) \mid \text{ si } N - \sum_{j=i}^{n} \mid Documents(niv_j) \mid < 0 \\ sinon \mid Documents(niv_i) \mid \end{array} \right)$$

documents "fictifs" ayant ce niveau de pertinence. Pour deux documents d_i et d_j pris dans cette liste, si d_i est placé avant d_j alors d_i est "plus pertinent" que d_j pour le topic, i.e. implication suivante est toujours vérifiée :

$$i < j \Longrightarrow NivPertinence(d_i) > NivPertinence(d_j)$$

Exemple : On suppose qu'on a 3 niveaux de pertinence niv_1 , niv_2 , niv_3. Soit le topic t tel que $\mid Documents^t(niv_3) \mid = 7$, $\mid Documents^t(niv_2) \mid = 10$, $\mid Documents^t(niv_1) \mid = 25$ et on choisit $N = 30$.

Pour chaque niveau de pertinence, on calcule le nombre de documents ayant ce niveau de pertinence à insérer dans la liste idéale de résultats de taille $N = 30$.

La liste "idéale" formée sera constituée comme suit :

$$\underbrace{niv_3, \ldots, niv_3}_{7fois} , \underbrace{niv_2, \ldots, niv_2}_{10fois} , \underbrace{niv_1, \ldots, niv_1}_{13fois}$$

Comme dans le cas précedent, deux possibilités s'offrent à nous :

– on calcule à un point donné N et pour un topic,

$$Metrique2_N(C) = Metrique1_N(Retrived(C), Retrieved\_ideal(C))$$

Cette métrique donne le gain moyen sur la qualité qu'on réalise en passant de la liste de résultats obtenues sur la collection C par rapport à une liste de résultats "idéale".

– on fait une somme des élements des vecteurs pondérés de tous les topics rang par rang et on a un seul vecteur final de N éléments pour les topics :

$$< cp(1) \times \sum_t (Passage_1^t(C, C\_ideal)), \ldots, cp(N) \times \sum_t (Passage_N(t)(C, C\_ideal)) >$$

Comme précedemment, ce vecteur somme a la même taille que les vecteurs de passages pondérés de chaque topic ; son $i$ème élément est la somme des $i$ème éléments des vecteurs pondérés de chaque topic. On peut donc visualiser ces vecteurs sous forme de courbes gain/perte (sur tous le topics) par rang.

4.5. *Conclusions*

Dans ce travail, nous proposons des métriques pour évaluer la manière dont les SRI passent à l'échelle. Ces métriques s'appuient sur la notion de pertinence multivaluée. Leur but est de fournir des informations sur la cohérence entre l'ordre des documents retournés par un SRI et les niveaux de pertinence de ces documents. Les métriques standards de RI sont basées sur une notion de pertinence plutôt binaire. Les métriques comme la *Discounted Cumulative Gain* ou la *Cumulative Gain* se basent également sur la pertinence multivaluée ; mais pour une collection donnée et un SRI, ces métriques calculent le gain d'information pertinente qu'on réalise au fur et à mesure qu'on parcourt la liste des résultats retournés par ce SRI sur cette collection. Les métriques que nous proposons calculent plutôt le gain d'information pertinente réalisée lorsqu'un même SRI est utilisé sur différentes collections. Ainsi, en appliquant nos métriques sur des sous-collections de taille croissante d'une très grande collection, l'on peut évaluer l'abilité du SRI à classer les documents en fonctions de leur degré de pertinence lorsque l'on passe à l'échelle dans la taille de collection. Nous travaillons sur ce point de façon expérimentale avec différents SRI. Nous avons ébauché une formalisation de l'attribution des valeurs numériques d'importance aux niveaux de pertinence et nous travaillons à son amélioration.Nous travaillons également à la mise en relation entre nos métriques et les métriques standards ou des métriques prenant en compte plusieurs niveaux de pertinence.

5. 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.

[BER 01] BERGMAN M., « The Deep Web : surfacing hidden value », *The Journal of Electronic Publishing*, vol. 7, n° 1, 2001.

[COO 71] COOPER W. S., « A definition of relevance for information retrieval », *Information Storage and Retrieval*, , 1971.

[DOG 05] DOGPILE, « Different engines, different results », http ://compareseachengines.dogpile.com/OverlapAnalysis.pdf, 2005.

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

[IMA 05] IMAFOUO A., BEIGBEDER M., « Scalability influence on retrieval models : An experimental methodology », *27th European Conference on Information Retrieval*, 2005.

[KAZ 05] KAZAI G., LALMAS M., « Notes on what to measure in INEX », *INEX Workshop on Element Retrieval Methodology*, 2005.

[KEK 00] KEKÄLÄINEN J., JÄRVELIN K., « IR evaluation methods for retrieving highly relevant documents », *Proceedings of the 23th annual international ACM SIGIR Conference*, 2000, p. 41-48.

[KEK 02] KEKÄLÄINEN J., JÄRVELIN K., « Using graded relevance assessments in IR evaluation », *Journal of the American Society for Information Science and Technology*, vol. 53, n° 13, 2002, p. 1120 - 1129.

[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.

[RAI 00] RAITT D., « Digital libraries initiatives across Europe », *Computers in libraries*, vol. 20, n° 10, 2000, p. 26-35.

[REE 67] REES A. M., SCHULZ D. G., « A field experimental approach to the study of relevance assessments in relation to document searching. 2 vols. », rapport n° NSF Contract No. C-423, 1967, Center for Documentation and Communication Research, School of Library Science.

[SAR 75] SARACEVIC T., « Relevance : A review of and a framework for the thinking on the notion in information science », *Journal of the American Society for Information Science*, vol. 26, 1975, p. 321-343.

[SAR 04] SARACEVIC T., « Evaluation of digital libraries :an overview », *Workshop on the Evaluation of Digital Libraries*, 2004.

[SPI 98] SPINK A., GREISDORF H., BATEMAN J., « From highly relevant to not relevant : examining different regions of relevance », *Information Processing and Management : an International Journal*, vol. 34, n° 5, 1998, p. 599-621.

[TAN 99] TANG R., WILLIAM M. SHAW J., VEVEA J. L., « Towards the identification of the optimal number of relevance categories », *Journal of the American Society for Information Science*, vol. 50, n° 3, 1999, p. 254-264.

[VOO 01] VOORHEES E. M., « Evaluation by highly relevant documents », *Proceedings of the 24th annual international ACM SIGIR Conference*, 2001, p. 74-82.

[VOO 04] VOORHEES E., BUCKLEY C., « Retrieval evaluation with incomplete information », *Proceedings of the 27th annual international Conference*, 2004, p. 25-32.

[WIL 73] WILSON P., « Situational relevance », *Information Storage and Retrieval*, vol. 9, n° 8, 1973, p. 457-471.