# Étude des distributions de $tf$ et de $idf$ sur une collection de 5 millions de pages HTML

**Michel Beigbeder — Annabelle Mercier**

*École Nationale Supérieure des Mines de Saint-Etienne*
*158, cours Fauriel*
*F 42023 SAINT ETIENNE CEDEX 2*
*{michel.beigbeder, annabelle.mercier}@emse.fr*

RÉSUMÉ. *Nous étudions l'évolution des distributions des valeurs de la fréquence des termes et de la fréquence documentaire dans les vecteurs traditionnellement utilisés dans le modèle vectoriel de recherche d'informations en fonction du nombre de documents indexés. Nous construisons des collections dont la taille augmente d'un facteur 10 à chaque fois. Les documents utilisés sont extraits des pages HTML récoltées sur des sites de domaines géographiques francophones en décembre 2000.*

ABSTRACT. *We study the evolution of the distributions of term or document frequency in the vectors traditionally used in the vectorial information retrieval model according to the collection size. Around the same core of documents, we build collections whose size increases by a factor 10 each time. The documents used are extracted from HTML pages collected on sites of geographical fields French-speaking in December 2000.*

MOTS-CLÉS : *Toile mondiale, recherche d'informations, fréquence documentaire, fréquence des termes, informatique.*

KEYWORDS: *World Wide Web, information retrieval, document frequency, term frequency, computer science.*

## 1. Introduction

Nous sommes familiarisés depuis le milieu des années 90 avec les documents hypertextes par l'utilisation massive de la Toile[1]. Tout un chacun connaît par ailleurs le besoin d'outils de recherche d'informations performants sur ce médium, au vu justement de son succès qui a entrainé l'accessibilité potentielle à une quantité d'informations sans précédent concernant tous les domaines de la connaissance humaine. Pour tenir compte de cette quantité d'informations et garder une efficience[2] acceptable par les utilisateurs des outils de recherche de la Toile, il faut bien sûr utiliser des méthodes algorithmiques adaptées à la quantité de données manipulées. Ces méthodes passent par une étude et une adaptation aux données effectivement rencontrées, et se basent pour une grande part sur des méthodes de compression de ces données et des données dérivées (par exemple, l'*index* pour une application classique de recherche d'informations).

Du point de vue de leur efficacité les outils de recherche d'informations doivent prendre en compte les particularités de ce médium. Parmi celles-ci il y a bien évidement l'utilisation des liens qui caractérisent un hypertexte par rapport à un texte. Mais malgré l'utilisation de plus en plus fréquente de ces particularités, la base reste l'utilisation des contenus textuels des différentes pages récoltées par ces outils. Dans cette étude, nous voulons aborder le problème du passage à l'échelle en essayant d'évaluer l'effet de l'augmentation du nombre de documents sur les distributions des valeurs qui sont utilisées dans les représentants des pages par les méthodes traditionnelles dans le domaine de la recherche d'informations. Nous commencerons donc par rappeler brièvement quelles sont ces valeurs, et quelles transformations les méthodes classiques leur font subir. Puis après une présentation des données sur lesquelles nous avons travaillé, nous présenterons quelques statistiques extraites de notre collection avant de conclure.

## 2. Représentation des documents en recherche d'informations

Le modèle vectoriel dont le fondement remonte aux travaux de Salton [SAL 83] se base sur une représentation vectorielle des documents, ou encore sur une représentation matricielle de la collection de documents. Nous appelerons $D$ l'ensemble des documents, et $T$ l'ensemble des termes qui apparaissent dans ces documents. Par terme, nous désignons soit une suite de caractères alphanumériques trouvée dans au moins un document, soit cette chaîne de caractères après un traitement de normalisation (par exemple, développement des ligatures[3], transformation en minuscules, lemmatisation, etc.).

---

1. Le *Web*, ou encore le *World Wide Web*.
2. Nous distinguons *efficience* : qui est lié au rendement (rapidité et/ou quantité de ressources utilisées), de *efficacité* : qui est lié à la qualité du résultat.
3. Comme la ligature 'œ' par exemple.

| SMART[a] | fonction C[b] | | définition | image |
|---|---|---|---|---|
| `b**` | `tfwt_binary` | $tfb(d,t)$ | $=$ positif$(f(d,t))$ | $[0,1]$ |
| `n**` | | $tfn(d,t)$ | $=f(d,t)$ | $[0,+\infty]$ |
| `m**` | `tfwt_max`[c] | $tfm(d,t)$ | $=\frac{f(d,t)}{\max_{t'} f(d,t')}$ | $[0,1]$ |
| `a**` | `tfwt_aug`[c] | $tfa(d,t)$ | $=\frac{1}{2}+\frac{1}{2}\frac{f(d,t)}{\max_{t'} f_{d,t'}}$ | $[\frac{1}{2},1]$ |
| `s**` | `tfwt_square` | $tfs(d,t)$ | $=f(d,t)^2$ | $[0,+\infty]$ |
| `l**` | `tfwt_log` | $tfl(d,t)$ | $=1+\log f(d,t)$ | $[1,+\infty]$ |

[a] La lettre de cette colonne correspond à la configuration de SMART.
[b] La fonction C dans le source du programme SMART.
[c] La fonction de SMART ne divise pas par le $\max$, mais par $\max+0.00001$.

**Tableau 1.** *Les différentes fonctions $tf$*

La donnée de base, que nous noterons $f(d,t)$ pour construire les vecteurs, est la fréquence, autrement dit le nombre d'occurences, de chaque terme $t \in T$ dans chaque document $d \in D$. Une deuxième donnée $df(t)$[4] compte le nombre de documents ayant au moins une occurence du terme $t \in T$. La fonction $df$ peut se formuler en fonction de $f$ : $df(t) = \sum_{d \in D} \text{positif}(f(d,t))$ où $\text{positif} : \mathbb{N} \longrightarrow \mathbb{N}$ associe 0 à 0, et 1 à tout nombre entier strictement positif.

Ces deux données sont à la base des modèles communément appelés $tf \cdot idf$, c'est-à-dire des modèles où le score de pertinence d'un document à une requête est d'autant plus élevé qu'il contient de nombreuses fois les termes de la requête de l'utilisateur (facteur $tf$, *term frequency*), et que ces termes discriminent les documents entre eux (facteur $idf$, *inverse document frequency*). Pour une requête avec un seul terme $t$, il n'y a pas nécessairement proportionnalité comme le laisse sous-entendre l'expression $tf \cdot idf$ du score du document $d$ avec $f(d,t)$ mais seulement monotonie croissante. Ainsi, différentes fonctions ont été utilisées pour calculer le facteur $tf$. Le système SMART[5] [SAL 71] propose celles listées en table 1. Nous y avons indiqué l'intervalle des valeurs *(images)* prises par ces fonctions.

De même, pour ce qui concerne le facteur *idf*, il n'y a pas nécessairement proportionnalité inverse du score de $d$ avec $df(t)$ mais seulement monotonie décroissante. Plusieurs propositions classiques sont listées dans la table 2. Aux fonctions disponibles dans SMART, nous avons ajouté les fonctions $idfI$ et $idfP$ [WIT 99]. Notons que la fonction $idfp$ qui vient du modèle probabiliste [ROB 76] peut prendre des valeurs négatives.

4. Les initiales $df$ font référence à la « fréquence-document » ou fréquence documentaire (*document frequency* en anglais) du terme.
5. `ftp://ftp.cs.cornell.edu/pub/smart/smart.11.0.tar.Z`

| SMART[a] | fonction C[b] | | définition | image |
|---|---|---|---|---|
| `*n*` | | $idfn(t)$ | $= 1$ | $[1, 1]$ |
| `*i*` | `idfwt_idf` | $idfi(t)$ | $= \log \frac{\|D\|}{df(t)}$ | $[0, \log(\|D\|)]$ |
| MG[c] | | $idfI(t)$ | $= \log(1 + \frac{\|D\|}{df(t)})$ | $[\log(2), \log(\|D\| + 1)]$ |
| `*f*` | `idfwt_freq` | $idff(t)$ | $= \frac{1}{df(t)}$ | $[\frac{1}{\|D\|}, 1]$ |
| `*p*` | `idfwt_prob` | $idfp(t)$ | $= \log \frac{\|D\| - df(t)}{df(t)}$ | $[-\infty, \log(\frac{\|D\|-1}{1})]$ |
| MG[c,d] | | $idfP(t)$ | $= \log(1 + \frac{\max_{d,t'} f_{d,t'}}{df(t)})$ | $[\log(2), \log(1 + Cte)]$ |
| `*s*` | `idfwt_s_idf` | $idfs(t)$ | $= (\log \frac{\|D\|}{df(t)})^2$ | $[0, (\log(\|D\|))^2]$ |

[a] La lettre de cette colonne correspond à la configuration de SMART.
[b] La fonction C dans le source du programme SMART.
[c] Fonction non disponible dans SMART, mais citée dans [WIT 99].
[d] `*P*` dans SMART concerne la pondération des expressions (*phrases* en anglais).

**Tableau 2.** *Les différentes fonctions* $idf$

## 3. Les outils utilisés pour la construction des collections

Pour faire des tests de distributions à grande échelle obtenues avec ces différentes formules de pondération, nous avons eu besoin d'un ensemble de documents analogue à ce que l'on trouve sur la Toile. La meilleure source est donc de prendre un extrait représentatif de celle-ci. Nous avons accès à un ensemble de $5\,057\,642$ pages[6] collectées sur la Toile par le laboratoire CLIPS[7] de l'université de Grenoble en décembre 2000 grâce à un robot[8] qu'ils ont développé. Toutes les pages collectées sont dans des domaines d'origine géographique francophone — ce qui ne signifie pas que tous les documents sont en langue française, en effet de nombreux sites proposent plusieurs versions de leurs documents dans plusieurs langues.

Pour enlever les balises et pour remplacer les entités du langage HTML nous avons utilisé le programme LYNX[9] dans sa version 2.8.5 avec les options `-dump` et `-force_html`. Pour fabriquer les fréquences des termes $f(d, t)$ et les fréquences documentaires $df(t)$, nous avons utilisé l'outil MG dans sa version 1.2.1 [WIT 99]. Comme celui-ci ne gère pas les caractères hors du champ de l'ASCII, nous avons remplacé tous les caractères accentués par leur équivalent sans accent et les autres caractères (comme le symbole © par exemple) par un espace.

Pour étudier le passage à l'échelle, nous avons construit 6 corpus, WFR4.1 à WFR4.6, de tailles différentes en augmentant à chaque fois d'un facteur 10 le nombre de documents. Nous partons d'un noyau de 10 documents, le corpus WFR4.1, puis

6. `http://www-mrim.imag.fr/membres/mathias.gery/Robot/WebFr4_01_12_2000/domaines.html`
7. `http://www-clips.imag.fr/`
8. `http://www-mrim.imag.fr/membres/mathias.gery/CLIPS-Index/`
9. `http://lynx.browser.org/`.

| Collection | $\|D\|$ | $\|T\|$ | $\#tf$ | Taille indexée | Taille de l'index | Temps |
|---|---|---|---|---|---|---|
| WFR4.1 | $10^1$ | 601 | 798 | 0.01 M | 0.02 M | 2 s |
| WFR4.2 | $10^2$ | 3 226 | 8 824 | 0.12 M | 0.12 M | 8 s |
| WFR4.3 | $10^3$ | 23 285 | 123 871 | 1.93 M | 1.21 M | 148 s |
| WFR4.4 | $10^4$ | 119 634 | 1 610 628 | 29.45 M | 13.01 M | 683 s |
| WFR4.5 | $10^5$ | 573 850 | 17 276 191 | 316.60 M | 128.49 M | 6343 s |
| WFR4.6 | $10^6$ | 3 049 927 | 187 560 895 | 3 398.28 M | 1 378.44 M | 66170 s |

**Tableau 3.** *Caractéristiques des collections* WFR4.1 *à* WFR4.6

nous ajoutons le nombre de documents nécessaires pour obtenir un corpus de $10^2$ documents que nous nommons WFR4.2 et ainsi de suite. La table 3 donne quelques caractéristiques essentielles sur ces 6 collections : nombre de documents, nombre de termes, nombre de couples $(d, t)$ pour lesquels $tf(d, t)$ est non nul (colonne intitulée $\#tf$), et enfin les tailles des données indexées et de l'index résultant et le temps d'indexation.

## 4. Expérimentations

Nous présentons ici deux types d'expérimentations pour les valeurs de $tf$ et $idf$. Le premier consiste à analyser pour chacune des collections l'impact du choix de la fonction. Pour cela, nous traçons sur le même graphique les différentes fonctions de $tf$ ou de $idf$ et nous comparons l'allure des courbes obtenues. Pour le deuxième type, nous choisissons une fonction particulière que nous traçons pour les collections WFR4.1 à WFR4.6 sur la même figure. Nous présentons ici la fonction $tfn$ qui montre le mieux la distribution des $tf(d, t)$, et $idfI$ qui est utilisée dans MG. Ceci nous permet donc d'étudier d'une part, l'impact des diverses fonctions pour une collection donnée et d'autre part, l'influence de la taille de la collection pour une fonction donnée.

### 4.1. *Construction des données*

L'utilitaire `mg_invf_dump` de MG permet d'obtenir pour chaque collection : $df(t)$ la fréquence documentaire, et $tf(d, t)$, la fréquence du terme $t$ dans le document $d$. Nous avons modifié ce programme pour qu'il calcule toutes les valeurs des fonctions $idf$. Nous avons aussi écrit un nouvel utilitaire `mg_vect_dump` pour calculer toutes les valeurs des fonctions $tf$.

Le nombre de données ainsi obtenues augmente très fortement en fonction de la taille de la collection, par exemple, pour WFR4.5, nous avons $573\,850$ points à tracer pour les $df$ et $17\,276\,191$ pour les $tf$. Comme il n'est pas possible d'envoyer autant de points à un programme de dessin, nous avons écrit un progamme `chunker` calculant un ensemble résumé de $N$ points (nous avons retenu pour les dessins $N = 5000$).

À partir du nombre de points en entrée, nous calculons une taille de paquet pour regrouper les points. Pour chaque paquet nous calculons un point moyen, la valeur minimale, la valeur maximale et l'écart type. Pour les dessins en échelle logarithmique en abscisse, la taille des paquets varie de façon à ce que les points à tracer soient régulièrement espacés.

Les données en entrée de ce programme (les valeurs prises par les différentes fonctions $idf$ et $tf$) sont triées numériquement par ordre décroissant. L'affichage se fait enfin avec GNUPLOT. Cependant pour les figures que nous présentons ici, nous ne nous sommes pas servi des valeurs minimale, maximale, ni de l'écart-type. En effet, pour pouvoir les utiliser de façon lisible il faut une sortie en couleur.

### 4.2. *Représentation des résultats*

Selon les cas, nous avons adopté une échelle linéaire ou logarithmique en ordonnée de façon à obtenir la meilleure séparation des courbes. Dans le cas logarithmique, les valeurs négatives de la fonction $idfp$ sont éliminées. Pour les abscisses nous avons tracé les courbes avec une échelle linéaire pour une vision globale mais aussi avec une échelle logarithmique pour mettre en évidence ce qui se passe au niveau des (grandes) valeurs de $tf$ ou de $df$ qui apparaissent en début de courbe. Cette échelle est en pourcentage du nombre de termes (resp. de couples $(d, t)$) de façon à ce que les courbes de la fonction $tfn$ (resp. $idfI$) tracées sur le même graphique pour les différentes collections soient comparables.

### 4.3. *Étude de la fréquence documentaire*

La première expérience consiste à étudier $df$ à travers les différentes fonctions $idf$. Nous rappelons que $df(t)$ est la fréquence documentaire, définie comme le nombre de documents dans lequel le terme $t$ apparaît.

#### 4.3.1. *Impact des fonctions $idf$ sur chaque collection*

Nous avons tracé les fonctions $idf$ décrites dans la section 2 pour chaque collection, nous ne montrons que celles pour WFR4.1 (fig. 1) et WFR4.6 (fig. 2) — pour les autres collections l'allure des courbes est semblable à celle de WFR4.6. Les quatre fonctions $idfP$, $idfI$, $idfi$, $idfp$ sont majorées par $idfs$ quasiment partout[10], et partout minorées par $idff$. Cette dernière se distingue par une amplitude de variation plus importante et devrait induire une meilleure discrimination pour les termes très fréquents. De plus nous avons noté que l'ordre des fonctions intermédiaires change selon la taille de la collection. Nous pouvons enfin remarquer que 50% des termes n'apparaissent que dans un seul document.

10. Plus précisement à partir de 7% (resp. 0.5%, 0.1%, 0.02%, 0.01%) pour WFR4.1 (resp. WFR4.2, WFR4.3, WFR4.4, WFR4.5 et WFR4.6).

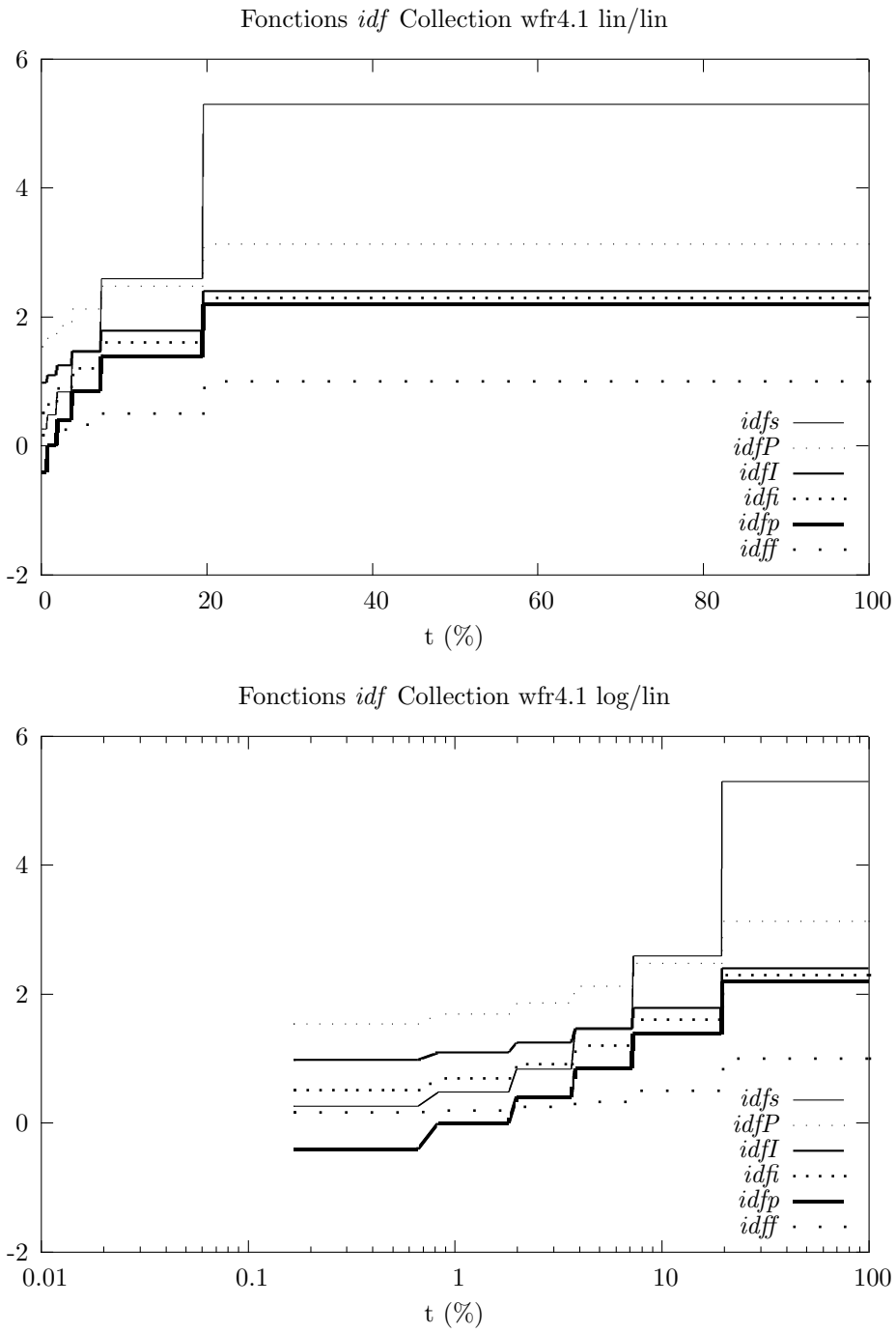


**Figure 1.** *Les différentes fonctions idf sur la collection* WFR4.1

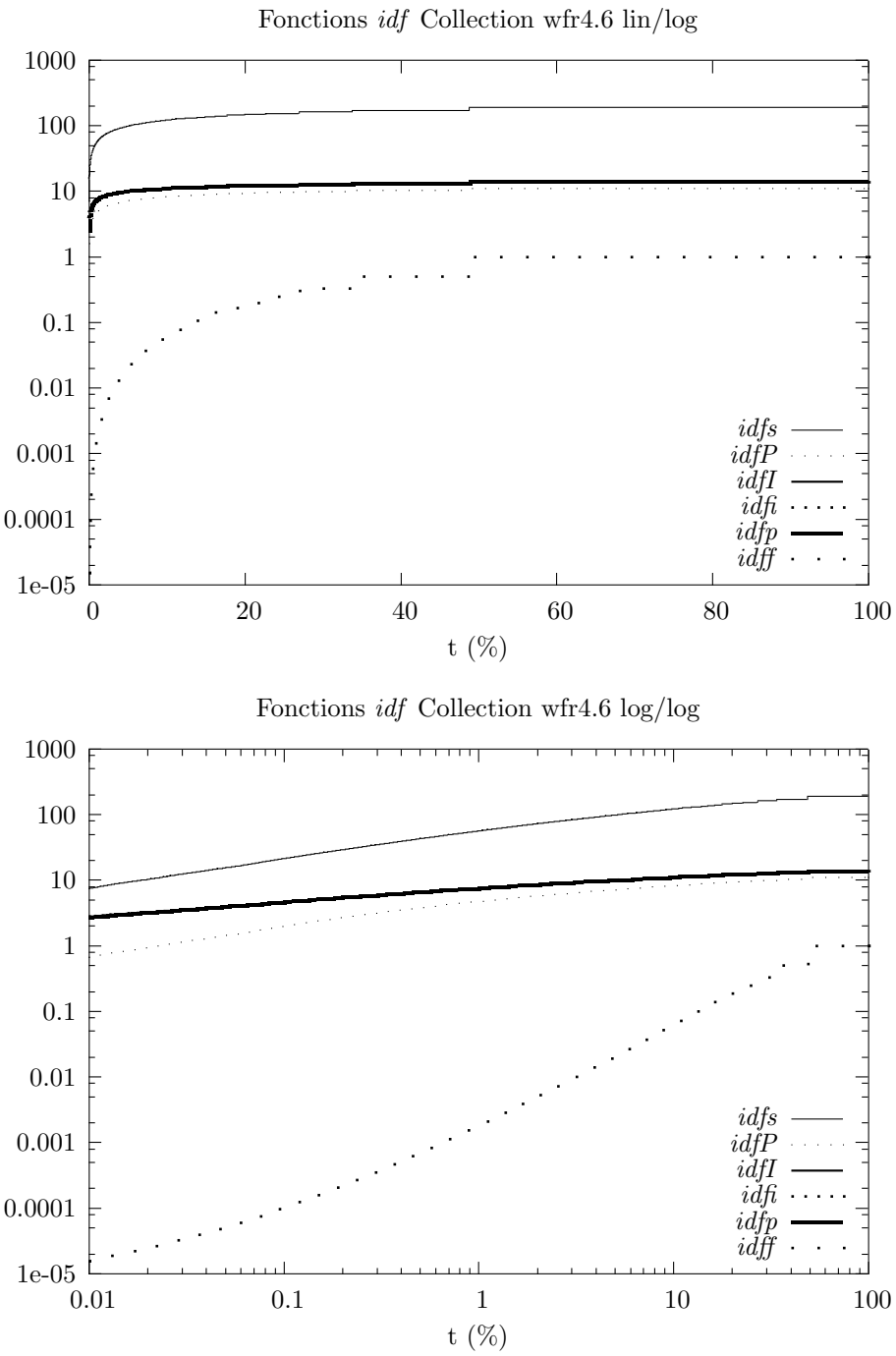


**Figure 2.** *Les différentes fonctions idf sur la collection* WFR4.6

4.3.2. *Passage à l'échelle pour la fonction $idfI$*

Nous avons choisi la fonction $idfI$ utilisée dans MG pour étudier l'influence du passage à l'échelle sur $df$. Les courbes pour les différentes collections sont comparables car nous avons normalisé[11] les valeurs en abscisse (cf. 4.2). À partir du tracé en lin/log de la figure 3, nous constatons que plus le nombre de documents est important plus les courbes se rapprochent d'une fonction constante. Le passage à l'échelle entre collection de taille de plus en plus grande montre qu'il est de plus en plus difficile de trouver des termes discriminants et que l'influence du facteur $idf$ va en diminuant.

## 4.4. *Étude de la fréquence des termes par document*

La seconde expérience nous permet d'étudier la distribution des valeurs de fréquences des termes dans les documents en fonction de la taille des collections. Nous rappelons que $tf(d,t)$ est la fréquence du terme $t$ dans le document $d$.

4.4.1. *Impact des fonctions $tf$ sur chaque collection*

Les trois fonctions $tfs$, $tfn$, et $tfl$ (cf. figures 4 et 5) reproduisent les variations des trois fonctions $x \rightarrow x^2$, $x \rightarrow x$, et $x \rightarrow \log(x)$ sur la distribution des $tf$. Leur taux de variation et leur image[12] sont de ce fait nettement différents. Par contre elles ont le même comportement global en se terminant par des plateaux de plus en plus longs correspondant aux valeurs de $tf$(d,t) : 1, 2, 3, etc. Au contraire les deux autres fonctions commencent par un plateau (qui toutefois correspond à environ 1% des couples $(d,t)$) et se terminent avec un taux de variation croissant.

4.4.2. *Passage à l'échelle pour la fonction $tfn$*

Sur la figure 6 nous constatons une convergence de la distribution des $tf$. De plus, le mode d'affichage donne l'impression que la courbe de WFR4.4 majore celle de WFR4.5, ce qui n'est pas possible à cause du mode de construction des collections. En effet, tous les documents de WFR4.4 appartiennent aussi à WFR4.5. L'affichage que nous faisons utilise des résumés et en noir et blanc, nous ne pouvons pas visualiser les intervalles d'erreurs produit par ces résumés.

## 5. Conclusions et perspectives

Pour compléter cette étude, nous voulons aussi étudier les distributions des fonctions complètes de pondération, c'est-à-dire en prenant en compte les facteurs $tf \cdot idf$ que nous avons mentionnés, mais aussi le facteur de pondération sur chaque vecteur $\vec{d}$ représentant le document $d$ : $\vec{d} = (d_t)_{t \in T} = (tf(d,t) \cdot idf(t))_{t \in T}$.

---

11. Les points sont tracés pour un pourcentage du nombre maximum de termes pour la collection.
12. L'intervalle des valeurs prises.

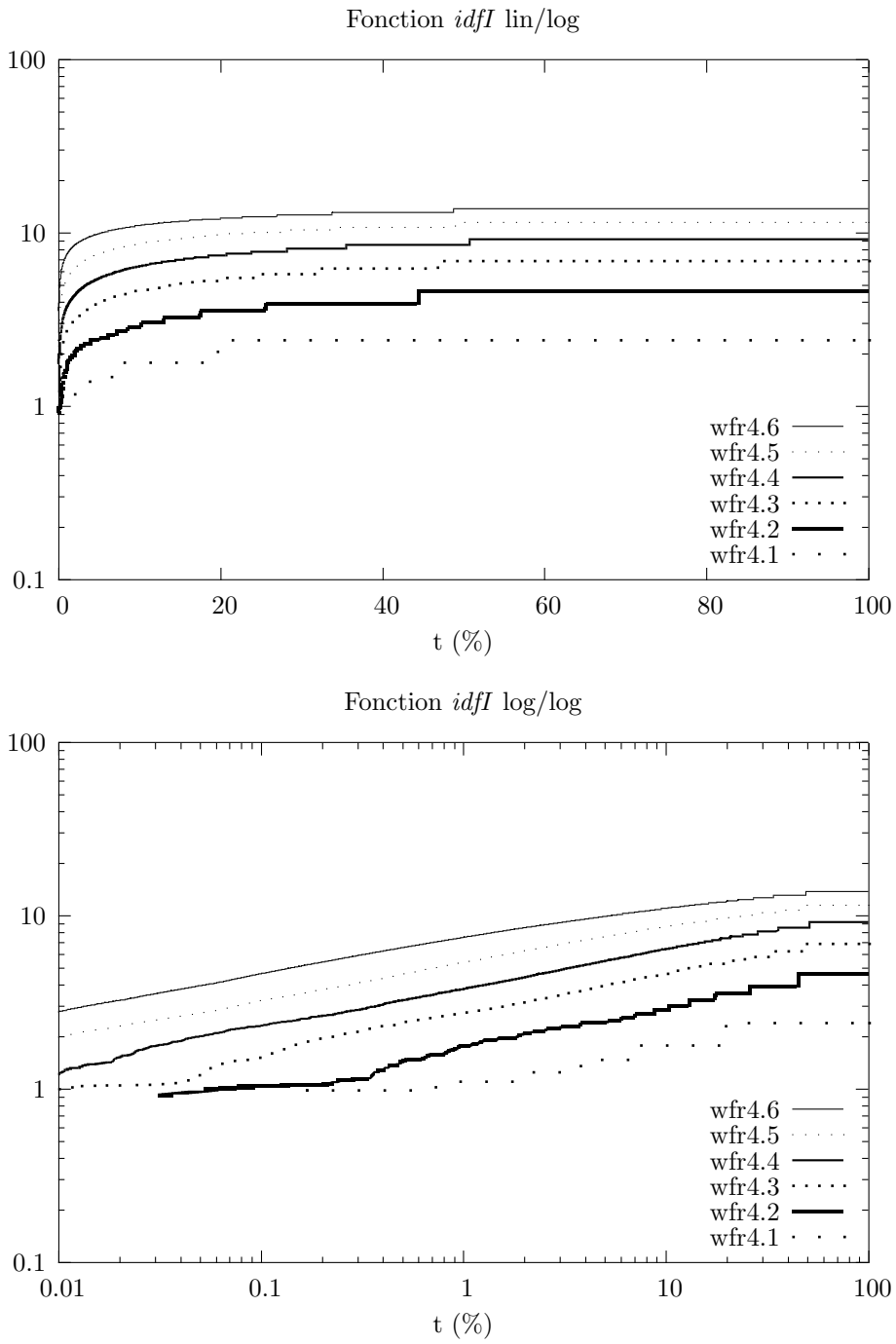


**Figure 3.** *Fonction idfI pour les collections* WFR4.1 *à* WFR4.6

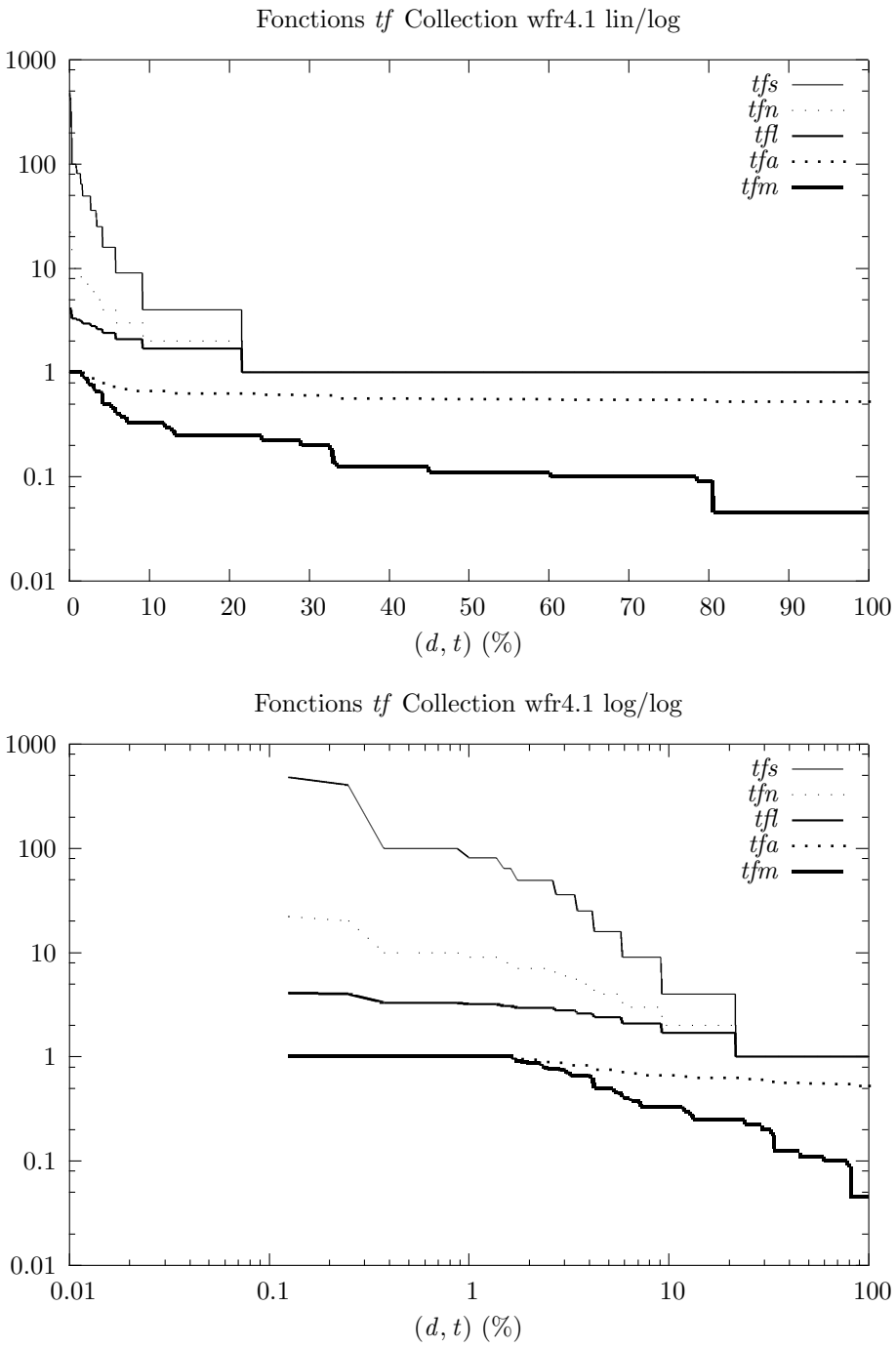


**Figure 4.** *Fonctions $tf$ pour la collection* WFR4.1

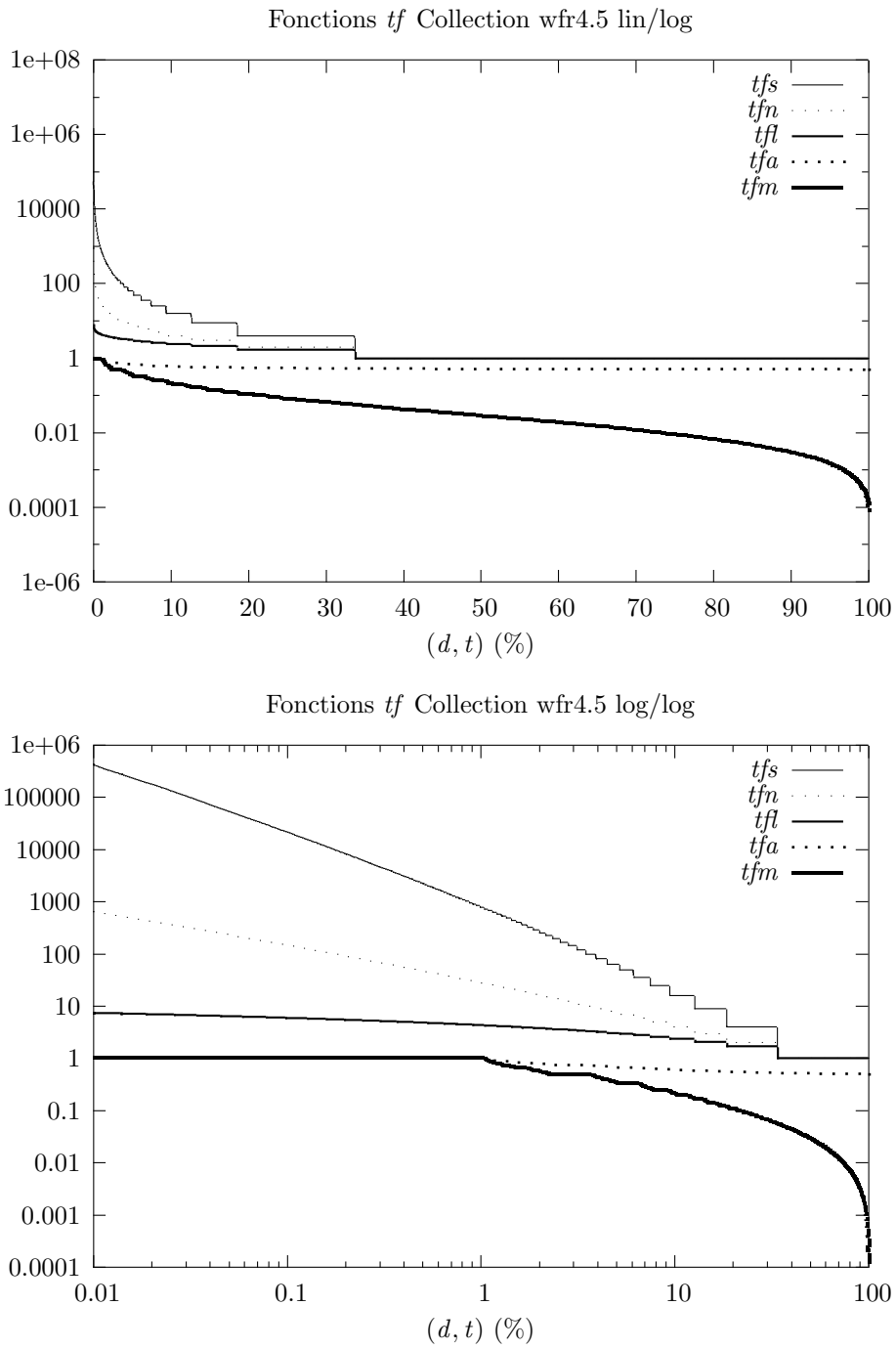


**Figure 5.** *Fonctions tf pour la collection* WFR4.5

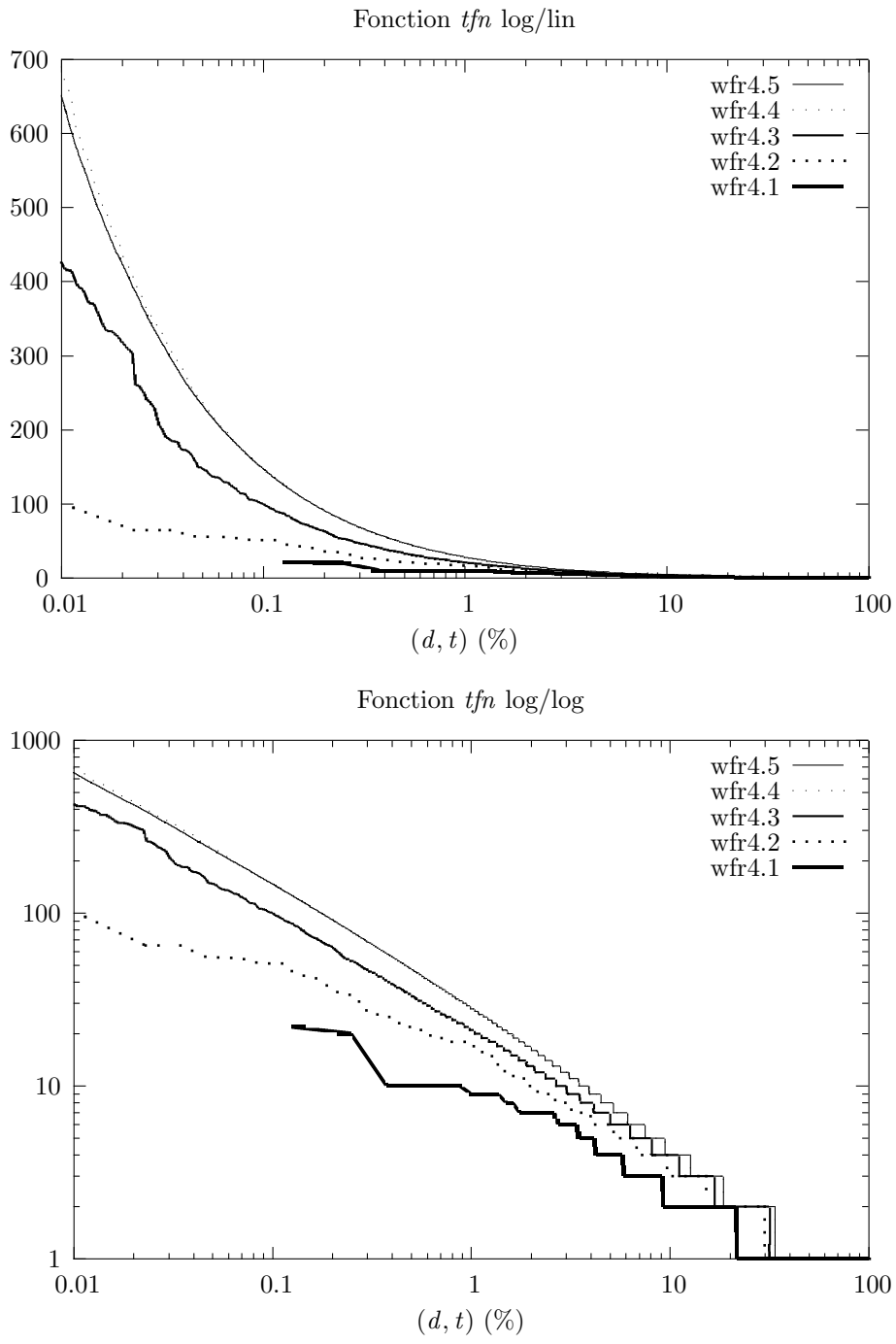


**Figure 6.** *Fonction $tfn$ pour les collections* WFR4.1 *à* WFR4.5

Nous avons défini les termes comme les suites de caractères alphanumériques trouvées dans les documents. De ce fait, les nombres sont aussi indexés et nous pensons que c'est à l'origine de l'accroissement du nombre de termes (facteur entre 4.79 et 7.21 d'une collection à la suivante). Il serait intéressant de regarder si en ne prenant que les suites de caractères alphabétiques nous aurions le même accroissement, et si cela donnerait des modifications dans l'allure des courbes que nous avons tracées.

D'ores et déjà, cette étude nous a permis d'analyser les valeurs de la fréquence documentaire et de la fréquence des termes dans les documents en fonction de la taille des collections, ceci pour les différentes formules classiques du domaine de la recherche documentaire. Il est important en effet d'avoir une idée des valeurs prises par ces différentes fonctions puisque d'une part elles servent directement dans les modèles de recherche d'informations, et que d'autre part elles sont souvent utilisées dans des fonctions de similarité dès que l'on veut comparer deux documents. Par conséquent connaître leur intervalle de variation peut permettre d'ajuster des coefficients de combinaison linéaire ou autre dans le cadre de l'utilisation de ces fonctions. Enfin 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 énormément et que la proportion de termes discriminants diminue.

## 6. Bibliographie

[ROB 76] ROBERTSON S., JONES K. S., « Relevance weighting of search terms », *Journal of the American Society for Information Science*, vol. 27, 1976, p. 129–146.

[SAL 71] SALTON G., *The SMART retrieval system : experiments in automatic document processing*, Automatic Computation, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[SAL 83] SALTON G., MCGILL. M. J., *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, 1983.

[WIT 99] WITTEN I. H., MOFFAT A., BELL T. C., *Managing Gigabytes : Compressing and Indexing Documents and Images*, Morgan Kaufmann, 1999.