Chapitre 3

L'importance des probabilités *a priori* pour la recherche des pages d'entrée[1]

3.1. Introduction

La recherche de page d'entrée soulève des caractéristiques qui la distingue de la recherche générale d'information, non seulement parce que les pages d'entrée sont des documents distincts des autres documents *web*, mais également parce que les buts et tâches sont différents. Dans les pistes *ad hoc* des campagnes d'évaluation présentes dans TREC, l'objectif consiste à dépister le maximum de documents pertinents. La tâche de page d'accueil (EP) concerne l'extraction de la page centrale d'une organisation servant de portail pour l'information sous-jacente dans le site. Notons toutefois que cette page n'est pas forcément la racine d'un site mais peut correspondre à la page d'entrée d'un sous-site (par exemple, celle de departement HMI à l'université de Twente). Compte tenu du fait que la recherche de page d'accueil a pour but d'obtenir un seul document, un système de recherche d'information devrait probablement être davantage optimisé vers une haute précision plutôt que de se préoccuper d'un taux de rappel élevé. Les utilisateurs de moteur de recherche préfèrent en général trouver une page d'accueil dans le premier écran de résultats. Du fait que les demandes de recherche sont habituellement très courtes, il s'avère très difficile de trouver une page d'accueil avec une grande précision. Cet article explore les diverses stratégies afin

1. Ce chapitre est basé sur l'article : « The Importance of Prior Probabilities for Entry Page Search », *Proceedings of the* 25^{th} *annual international ACM SIGIR Conference on Research and Cevelopment in Information Retrieval SIGIR'02*, © ACM, 2002, http://doi.acm.org/10.1145/564376.564383.

Chapitre rédigé par Wessel KRAAIJ, Thijs WESTERVELD et Djoerd HIEMSTRA.

d'améliorer les systèmes d'information conçus pour les tâches de recherche *ad hoc*, en tenant compte des caractéristiques des documents formant des pages d'entrée. Nos expériences pour les tâches de page d'accueil lors de TREC-2001 ont démontré que la structure des liens, les URL et les textes d'ancrage sont des sources intéressantes d'information pour la localisation de pages d'entrée [WES 01].

Les expériences d'autres groupes ont confirmé l'efficacité de ces caractéristiques [CRA 01a], [HAW 01b], [CRA 01b], [RA 01], [SAV 01a], [XI 01]. Dans cet article nous démontrons comment la connaissance des caractéristiques non liées au contenu d'une page *web* ainsi que sa probabilité d'être une page d'accueil, peuvent facilement être incorporées dans un modèle de RI basé sur les modèles statistiques de langage. Dans la section 3.2 nous discutons de travaux apparentés à l'ajustement de modèles de RI pour une tâche spécifique aussi que d'autres recherches entreprises concernant l'utilisation de sources d'information autres que le contenu du document. La section 3.3 décrit les principes de base en modélisation de la langue, ainsi que la manière dont les probabilités *a priori* peuvent être utilisées dans ce modèle. Dans la section 3.4 nous abordons diverses sources de connaissance *a priori* qui peuvent être utilisées pour une tâche de recherche de page d'accueil. Nous décrivons notre méthodologie d'évaluation dans la section 3.5 et présenterons nos expériences et nos résultats dans la section 3.6. Nous concluons avec une discussion de ces résultats et un résumé de nos principales conclusions.

3.2. Contexte et ouvrages reliés

Dans cette section, nous présentons les arguments en faveur d'une adaptation d'un système de RI à une tâche de recherche spécifique, ou même à une certaine collection de tests, afin d'optimiser la performance. Il est important cependant, d'employer des méthodes basées sur une théorie au lieu simplement d'adapter des solutions *ad hoc*.

3.2.1. *Optimisation des systèmes de recherche d'information*

L'utilisation des statistiques des termes s'avère primordiale pour la qualité des résultats sur la tâche de recherche *ad hoc*. Tous les modèles de RI proposant un classement de pertinence exploitent au moins deux statistiques : la fréquence d'un terme dans un document et la distribution d'un terme dans la collection de documents. Les modèles de RI ont évolué avec la plus grande disponibilité des collections tests. Les collections tests plus anciennes se composaient de résumés d'article et, dans ce cas, il était raisonnable de supposer que les documents possédaient des longueurs à peu près identiques. Puisque les collections tests sont basées sur des textes intégraux, cette hypothèse ne tient plus. Des modèles de RI ont donc été améliorés pour inclure une composante tenant compte de l'influence de la longueur du document. Les premières tentatives visant à combiner ces trois ingrédients principaux étaient plutôt *ad hoc* et

n'étaient supportées par aucune théorie. Par exemple, Salton et Buckley [SAL 88] ont évalué de nombreuses combinaisons de statistiques de fréquence de terme, de statistiques de fréquence de document ainsi que des mesures de normalisation tenant compte de la longueur du document, sans proposer un cadre formel explicite. La stratégie de combinaisons et de tests a été poursuivie par Zobel et Moffat [ZOB 98]. Ils ont examiné 720 stratégies basées sur différentes fonctions de ressemblance, stratégies de normalisation de longueur, pondération des termes, etc. et leur évaluation comprenait six collections de requêtes différentes. Ils ont conclu que leur stratégie d'essais multiples (4 semaines de temps de calcul) n'avait pas été capable de fournir un résultat définitif pouvant être maintenu à travers plusieurs collections tests. Au lieu de recourir à une démarche *ad hoc*, Robertson et Walker [ROB 94] ont expérimenté des approximations simples d'un modèle probabiliste. La formule OKAPI est une approximation d'un modèle théoriquement justifié, intégrant les trois composantes. Cette formule inclut des paramètres déterminants, comme par exemple, l'influence de la fréquence de terme ou l'influence de la normalisation basée sur la longueur du document. Les paramètres peuvent être ajustés pour s'accorder à la tâche particulière ou aux caractéristiques propres de la collection test.

Basé sur le succès de Robertson et Walker, Singhal *et al.* [SIN 96] ont montré que pour la tâche de recherche *ad hoc*, il y a une corrélation entre la longueur du document et la probabilité *a priori* de pertinence. La longueur du document s'avère un bon exemple d'information concernant un document et qui n'est pas directement liée à son contenu. Cependant, on peut corréler cette information à la pertinence du document. Ainsi, si nous établissons un système plutôt stupide retournant le plus long document quelque soit la requête, cette stratégie s'avère plus favorable que de retourner un document aléatoire.

3.2.2. *Caractéristiques non reliées au contenu des pages* web

Une source importante de connaissance *a priori* pour la RI sur le *web* est la structure des hyperliens. L'idée sous-jacente à la plupart des analyses de structure des liens consiste à admettre que s'il existe un lien du document A au document B, cela signifie probablement que les documents A et B traitent le même sujet et que l'auteur du document A recommande le document B. Deux des algorithmes les plus connus et utilisés pour l'analyse de structure des liens sont PageRank [BRI 98] et HITS [KLE 99]. Les deux méthodes basent leur calculs sur l'hypothèse qu'une page qui reçoit beaucoup de liens est fortement recommandée et donc s'avère probablement une source digne de foi. Cette valeur d'autorité d'un document est encore plus forte si les documents qui pointent possèdent une certaine autorité. L'efficacité de HITS et PageRank est présentée dans de nombreuses études [KLE 99], [BHA 98], [AME 00], [NG 01], [COH 00]. Cependant en mesurant la pertinence seulement, sans tenir compte de la qualité ou

de l'autorité, les méthodes basées sur HITS et PageRank n'ont pas réussi à améliorer l'efficacité de recherche *ad hoc* [HAW 00], [HAW 01a], [KRA 01b], [SAV 01b], [CRI 01].

Davison [DAV 00] montre que l'hypothèse de localité thématique est admissible : quand des pages sont liées, elles sont susceptibles de partager un contenu sémantique relié. Ceci signifie que les documents qui sont liés aux documents pertinents sont potentiellement pertinents eux-mêmes. Cependant, l'exploitation de cette idée en attribuant des scores RI aux liens n'a pas encore démontré un clair succès dans une tâche de RI générale [KRA 01b], [GUR 01]. D'autre part, l'évaluation TREC-2001 a prouvé que l'analyse de structure des liens améliore les résultats d'un système RI basé seulement sur le contenu dans le cadre d'une tâche de recherche de page d'accueil [HAW 01b], [WES 01].

Une autre source de connaissance *a priori* dans RI sur le *web* est la présence des URL. Nous ne connaissons pas d'études (avant TREC-2001) dans laquelle la connaissance des URL a été employée pour une recherche *ad hoc*. Cependant, l'évaluation TREC-2001 a prouvé que le fait que les pages d'entrée tendent à avoir des URL plus courts que les autres documents peut être exploité avec succès par un algorithme de classement [HAW 01b], [WES 01], [RA 01], [SAV 01a], [XI 01].

3.2.3. *Probabilités* a priori *et les modèles de langue en recherche d'information*

Les connaissances *a priori* peuvent être utilisées d'une manière habituelle dans l'approche de modélisation de langue pour des systèmes de RI. L'approche de modélisation de langue emploie des modèles unigramme pour le classement des documents [PON 98]. Les premières expériences TREC par Hiemstra et Kraaij [HIE 98] montrent que la longueur des documents peut être utilisée comme probabilité *a priori* pour la recherche *ad hoc*, particulièrement pour des requêtes courtes. Miller *et al.* [MIL 98] ont combiné plusieurs sources d'information dans la probabilité *a priori* du document, y compris la longueur document, la source et la longueur moyenne des mots. Dans cet article nous adaptons l'approche de modélisation de langue standard pour la recherche des pages d'accueil en incluant les probabilités *a priori* dans le processus d'estimation. Nous montrons que pour cette tâche, l'utilisation appropriée des probabilités *a priori* donne des améliorations de plus de 100 % en comparaison avec la performance d'un système de base de modèle de langue. Bien que nous traitions principalement de la recherche de la page d'accueil dans ce chapitre, nous suivons une approche générique. Dans ce cadre, nous pouvons rapidement adapter notre système RI à d'autres tâches ou domaines, par exemple pour la génération de résumé automatique [KRA 01a].

3.3. Modèles de langue et probabilités *a priori*

Pour toutes nos expériences nous avons utilisé un système de recherche d'information basé sur un modèle de langue statistique [HIE 01]. Dans une telle modélisation, nous sommes intéressés par la probabilité qu'un document D soit pertinent étant donné une requête Q. Cette probabilité peut être estimée de la façon suivante en recourant au théorème de Bayes :

$$P(D|Q) \;=\; \frac{P(D)\,P(Q|D)}{P(Q)} \qquad [3.1]$$

La motivation principale pour l'application du théorème de Bayes est que les probabilités situées à droite peuvent être estimées avec plus de précision que les probabilités situées à gauche. Pour l'objectif de classement des documents, on peut simplement ignorer $P(Q)$ car cette constante ne modifie pas l'ordre du classement. La probabilité $P(Q|D)$ peut être estimée par un modèle de langue unigramme. L'approche standard de modélisation de langue représente la requête Q comme un séquence de n termes soit $T_1, \cdots, T_n$, et utilise une combinaison linéaire d'un modèle de document unigramme $P(T_i|D)$ et d'un modèle de la collection $P(T_i|C)$ permettant un lissage[2] selon l'équation suivante :

$$P(D|T_1, \cdots, T_n) \;\propto\; P(D) \prod_{i=1}^{n} (1-\lambda)P(T_i|C) + \lambda P(T_i|D) \qquad [3.2]$$

La probabilité $P(T_i|C)$ correspond à la probabilité de tirer un terme au hasard dans la collection, $P(T_i|D)$ définit la probabilité de tirer un terme au hasard dans le document et λ est le paramètre d'interpolation. Ce dernier s'estime habituellement comme une constante. Finalement, nous devrons estimer la probabilité *a priori* de pertinence d'un document $P(D)$[3].

3.3.1. *Estimation des probabilités* a priori

On distingue deux approches pour l'estimation de la probabilité *a priori* estimation directe en utilisant des données d'entraînement ou une approche plus heuristique.

Par exemple, on pourrait examiner l'hypothèse que la probabilité *a priori* est corrélée avec le longueur du document dans le contexte d'une tâche de RI *ad hoc*. La

2. Cette technique de lissage est aussi connu comme lissage *Jelinek-Mercer*.

3. La pertinence n'est pas modélisée explicitement dans la formule [3.2]. Voir [LAF 01] pour une formulation alternative du modèle incluant la pertinence comme une variable dans le modèle. Les deux modèles conduisent à une fonction équivalente de pondération des termes.

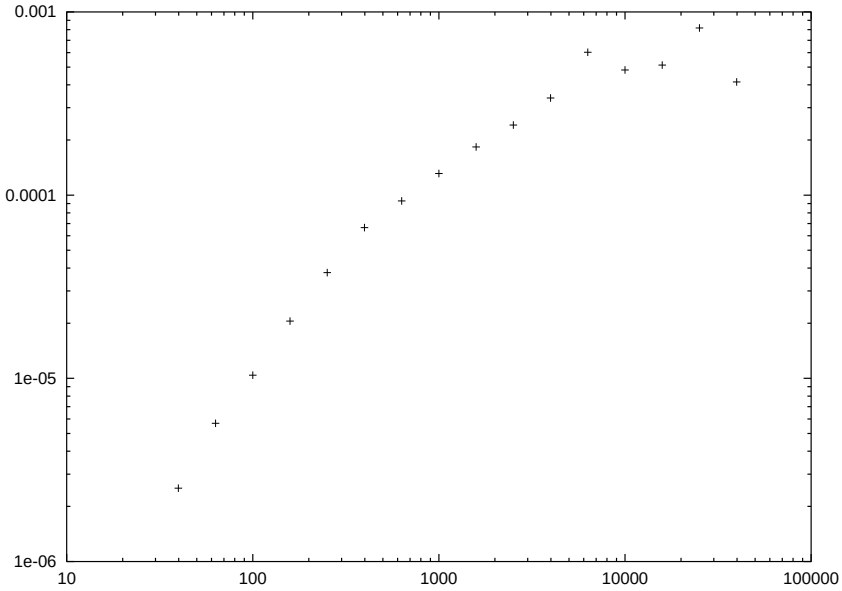

Figure 3.1. *Probabilité* a priori *du pertinence étant donnée la longueur du document (tâche* ad hoc search*)*

figure 3.1 montre le diagramme de la probabilité de la pertinence comparée à la longueur du document pour la tâche *ad hoc* de TREC-2001 *web track*. Pour le lissage des paramètres, nous avons construit 16 catégories sur un axe logarithme, pour représenter les différentes catégories de longueurs (variant entre deux et plus de 10 000 mots). Chaque point du diagramme marque la probabilité de pertinence des documents dans une de ces catégories. Ces dernières et les probabilités correspondantes définissent une mesure de probabilité discrète prenant une des 16 valeurs selon la catégorie concernée. Cette mesure de probabilité peut être utilisée directement dans l'équation [3.2]. Comme alternative, en regardant le diagramme, on pourrait rapprocher cette distribution avec une fonction linéaire de la longueur du document :

$$P_{doclen}(D) = P(R|D) = C \cdot doclen(D) \qquad [3.3]$$

dans laquelle R est une variable binaire aléatoire (*pertinent ou non*), $doclen(D)$ indique le nombre total de mots du document D et C est un constante que nous pouvons ignorer. En effet, seul le classement des documents nous intéresse et non la probabilité de pertinence. L'idée sous-jacente cette hypothèse est que les documents plus longs ont tendance à couvrir un éventail plus large de thèmes et, en conséquence, possèdent une probabilité de pertinence plus élevée. En effet, la relation linéaire entre la longueur

des documents et la probabilité de pertinence s'avère un modèle raisonnable pour diverses collections de RI *ad hoc* [KRA 01b]. L'inclusion de la probabilité *a priori* peut notamment améliorer la précision moyenne (MAP) jusqu'à une valeur absolue de 0,03 selon la longueur des requêtes.

3.3.2. *Combiner des probabilités* a priori

Si l'on dispose de diverses sources d'information pouvant influencer la probabilité *a priori* de pertinence, nous devons pouvoir les combiner. De telles sources sont, par exemple, la taille d'une page *web* ou son domaine. De manière générale, le problème peut être formalisé comme suit :

$$P(R|D) = P(R|\bar{\mathbf{x}}) = P(R|x_1, x_2, \ldots, x_n) \quad [3.4]$$

Dans la formule [3.4], R est un variable aléatoire binaire indiquant la pertinence, $\bar{\mathbf{x}}$ est une abréviation pour $(\bar{\mathbf{D}} = \bar{\mathbf{x}})$, où $\bar{\mathbf{D}}$ est une variable aléatoire *document* ayant comme valeur un vecteur de caractéristiques $\bar{\mathbf{x}} = (x_1, x_2, ..., x_n)$. Nous ne connaissons pas la distribution sous-jacente de $P(R|\bar{\mathbf{x}})$ ni même sa forme générale. De plus, nous disposons d'un ensemble très limité de données d'entraînement pour estimer les probabilités pour des valeurs différentes. Afin de pouvoir trouver une estimation de ces probabilités *a priori*, une approche possible consiste à appliquer l'hypothèse *bayésienne naïve*. Utilisant cette approche [MIT 97], [LEW 98] pour l'étiquetage automatique, nous avons la formulation générale suivante :

$$C' = argmax_{C_k} P(C_k|\bar{\mathbf{x}}) = \frac{P(\bar{\mathbf{x}}|C_k)P(C_k)}{P(\bar{\mathbf{x}})} \quad [3.5]$$

où C' représente la catégorie d'un objet qui est choisi parmi l'ensemble des classes disjointes mutuelles C_k. La règle de décision consiste à choisir le C_k maximisant le numérateur $P(\bar{\mathbf{x}}|C_k)P(C_k)$ étant donné un certain vecteur des attributs $\bar{\mathbf{x}}$. En particulier, en présence de nombreuses caractéristiques, il s'avère difficile d'estimer directement les probabilités sous-jacentes. Une stratégie courante est d'assumer que les caractéristiques sont conditionnellement indépendantes les unes des autres. Ainsi, on peut écrire : $P(\bar{\mathbf{x}}|C_k) = \prod_i P(x_i|C_k)$, ce qui correspond à l'hypothèse *bayésienne naïve*. Cette hypothèse réduit sensiblement la complexité car il reste moins de paramètres à estimer.

Le même problème d'estimation se rencontre pour le dénominateur. Dans ce cas cependant, ce terme peut être ignoré car pour les tâches d'étiquetage, ce terme est une constante. Dans notre cas, il n'y a que deux classes différentes c'est-à-dire que les

documents sont soit pertinents soit non pertinents (ou encore pages d'accueil ou non). Cependant, nous ne sommes pas seulement intéressés par la classe ayant la probabilité la plus élevée, mais nous voulons estimer cette probabilité pour un document donné ayant plusieurs attributs. Cette estimation va nous servir pour le classement des documents. Dans ce cas, on ne peut donc pas ignorer le dénominateur qui diffère pour chaque vecteur d'attributs. La formule [3.6] doit être utilisée pour l'estimation des probabilités *a priori* :

$$P(C_k|\bar{\mathbf{x}}) = \frac{\prod_{j=1}^{n} P(x_j|C_k)P(C_k)}{\sum_k(\prod_{j=1}^{n} P(x_j|C_k)P(C_k))} \quad [3.6]$$

Les paramètres de cette formule peuvent être estimés beaucoup plus facilement que dans le cas général indiqué dans l'équation [3.5]. A titre d'exemple, supposons que nous ayons trois attributs, chacun avec cinq valeurs. L'estimation directe de la probabilité de pertinence nécessiterait l'estimation de $5^3 = 125$ probabilités conditionnelles. En utilisant une approche bayésienne naïve, seulement 30 fréquences relatives devraient être calculées. Une approche alternative pourrait être le réduction du nombre de paramètres dans notre modèle, par exemple en considérant seulement trois classes (trois valeurs) par attribut, rapportant le nombre d'estimation à $3^3 = 27$ paramètres.

Dans nos expériences avec des probabilités *a priori* combinées, on a utilisé deux attributs (URL et #inlinks) chacun ayant quatre valeurs différentes d'attributs. Comme système de référence, nous avons supposée une indépendance conditionnelle des attributs étant donnée la pertinence et nous avons approximé la formule [3.6] en remplaçant le dénominateur par $\prod_j^n P(x_j)$, tout en évitant la nécessité de construire un ensemble d'entraînement pour pages *web* différentes de pages d'accueil. Donc, la formule peut être approximée par :

$$P(C_k|\bar{\mathbf{x}}) = \frac{\prod_{j=1}^{n} P(C_k|x_j)P(C_k)}{\prod_{j=1}^{n} P(C_k)} = K \prod_{j=1}^{n} P(C_k|x_j) \quad [3.7]$$

ce qui est équivalent au produit des probabilités *a priori* individuelles, à une constante près.

Nous avons comparé ce système avec un système où l'on avait estimé [3.4] directement, mais en utilisant un nombre de classes plus restreint. Nous avons défini sept classes disjointes de façon à ce que l'on rencontre au moins quatre exemples d'entraînement positifs pour chaque classe. Le paragraphe 3.6.2 décrit la procédure en détail.

3.3.3. *La combinaison des modèles de langue*

Tout comme nous pouvons avoir des sources d'information différentes pour estimer les probabilités *a priori*, nous pouvons aussi bien rencontrer des sources d'information différentes pour estimer les modèles du contenu des documents. A titre d'exemple, on peut considérer la possibilité d'estimer un modèle de langue d'un document en s'appuyant sur les textes d'ancrage de tous les hyperliens se dirigeant vers le document. Notez que le score du classement fourni par une requête sur des textes d'ancrage ne sont pas une probabilité *a priori*. Les textes d'ancrage et les textes des documents, proprement dit « contenu seulement », fournissent deux représentations textuelles très différentes des documents. Les systèmes basés sur des modèles de langue peuvent s'accommoder de plusieurs représentations de document d'une façon simple et directe, en définissant un modèle mixte. La manière privilégiée de combiner deux représentations serait modélisée par la formule suivante :

$$P(D|T_1,\cdots,T_n) \propto P(D)\prod_{i=1}^{n}\big((1-\lambda-\mu)P(T_t|C)+ \lambda P_{\text{content}}(T_i|D)+\mu P_{\text{anchor}}(T_i|D)\big) \quad [3.8]$$

Ainsi, la combinaison du modèle d'ancrage avec le modèle de contenu serait faite sur une base de « terme de requête par terme de requête » tandis que la combinaison avec la probabilité *a priori* s'opérait de manière séparée. Malheureusement, la version courante de notre SRI ne permet pas la combinaison de représentations de documents selon l'équation [3.8]. Afin de surmonter cette difficulté, les expériences avec textes d'ancrage décrites dans les prochaines sections ont été faites séparément des expériences basées sur le contenu. Les scores des deux approches seront combinés dans une deuxième phase. Cette approche a démontré son efficacité dans le domaine de la génération automatique du résumé [KRA 01a].

3.4. Les probabilités *a priori* pour la recherche de pages d'entrée

Dans une tâche de recherche de page d'accueil, un utilisateur souhaite la page principale d'une organisation ou d'un groupe spécifique dans une organisation. Une telle page *web* constitue le point d'entrée principal pour des informations de ce groupe. A partir de là, on peut habituellement se diriger vers les autres pages *web* du même groupe. Par exemple la page principale d'entrée de SIGIR 2007 est :
http://www.sigir2007.org.

Une requête dans une tâche de recherche de page d'accueil ne comprend habituellement que le nom de l'organisation ou du groupe dont on cherche la page d'accueil.

En général il n'y a que quelques réponses correctes, c'est-à-dire l'URL de la page d'accueil de l'organisation et, peut-être, quelques URL alternatifs dirigeant vers la même page et quelques sites miroirs.

La recherche de page d'accueil ressemble à la recherche de documents connus. Dans les deux types de recherches, un utilisateur souhaite une information spécifique. Cependant, dans la recherche de documents connus, l'utilisateur a déjà vu l'information requise tandis que dans la recherche de page d'accueil ce n'est pas forcément le cas.

Puisque la recherche de page d'accueil est une tâche différente de la recherche *ad hoc*, différentes probabilités *a priori* pourraient être nécessaires. Nous avons étudié quelles sont les sources d'information qui donnent les indications les plus utiles pour l'identification des pages d'entrée. De même, on s'est également interrogé pour savoir si les probabilités *a priori* sont en soit nécessaires pour une telle tâche.

Nous avons considéré trois sources différentes d'information : i) la longueur d'un document ; ii) le nombre de liens se dirigeant vers un document (*inlinks*) ; iii) la profondeur d'une page *web* dans un site, information donnée par l'URL. Les caractéristiques que nous n'avons pas étudiées, mais qui pourraient être utiles pour estimer si une page est une page d'entrée incluent : le nombre d'occurrences de certains mots-clé (par exemple « bienvenue »), le nombre de liens sortant dans un document et le nombre de fois où une page est visitée.

3.4.1. *Longueur du document*

Le paragraphe 3.3.1 a montré que les probabilités *a priori* basées sur la longueur du document sont utiles pour la recherche *ad hoc*. Ici nous étudions si la longueur d'un document s'avère également un indicateur utile de la probabilité qu'un document soit une page d'accueil. La figure 3.2 montre un graphique de la probabilité de la pertinence comparée à la longueur de la page, valeur calculée sur les données d'entraînement fournies pour la tâche de recherche de page d'accueil lors du *web track* de TREC-2001.

En effet la longueur du document permet d'estimer la probabilité qu'une page soit pertinente, puisque la distribution n'est pas uniforme. Les pages avec une longueur moyenne (60-1 000 mots) ont une probabilité plus élevée, avec un maximum autour de 100-200 mots. Cependant, les différences sont beaucoup moins marquées que pour la recherche *ad hoc*. Bien qu'il soit clair que la dépendance de la probabilité de la pertinence à l'égard de la longueur de document soit tout à fait différente de la recherche *ad hoc*, nous avons fait une expérience où nous avons employé notre meilleur modèle RI pour la recherche *ad hoc* (modèle de langue avec la probabilité *a priori* de la longueur telle que définie en [3.3]) sur la tâche EP. Le but de cette expérience consistait

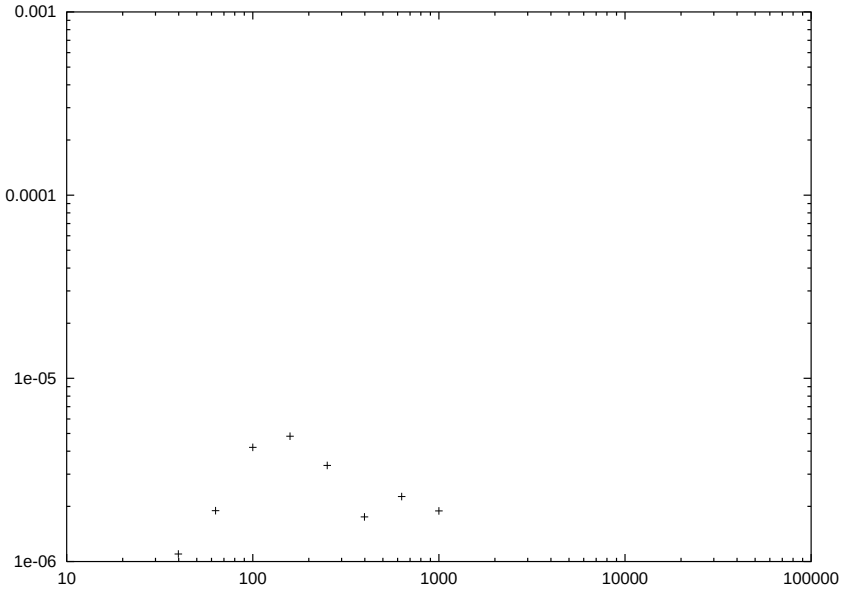

Figure 3.2. *Probabilité* a priori *de pertinence selon le longueur du document (*$P(entry\ page|doclent)$*), TREC-2001 collection d'apprentissage*

à démontrer que les modèles qui présentent une bonne performance pour la recherche *ad hoc* ne présentent pas nécessairement une performance similaire pour la recherche de pages d'entrée.

3.4.2. *Nombre des liens entrants*

Dans une collection d'hypertexte comme le *web*, les documents sont reliés par des hyperliens. Un hyperlien établit un lien entre un document source et un document cible. Notre probabilité *a priori* basée sur le nombre de liens entrants est basée sur l'observation que les pages d'entrée tendent à posséder un nombre plus élevé de liens entrants que d'autres documents (c'est-à-dire qu'elles sont mises en référence plus souvent). Le graphique de la probabilité d'être une page d'accueil *versus* le nombre de liens entrants dans la figure 3.3 montre cette corrélation pour les données d'entraînement pour la tâche de recherche de page d'accueil du *web track* de TREC-2001. Puisque le graphique suggère un rapport linéaire, nous définissons la probabilité *a priori* basée sur les liens entrants (*inlinks*) comme suit :

$$P_{inlink}(D) = P(R|D) = C \cdot inlinkCount(D) \qquad [3.9]$$

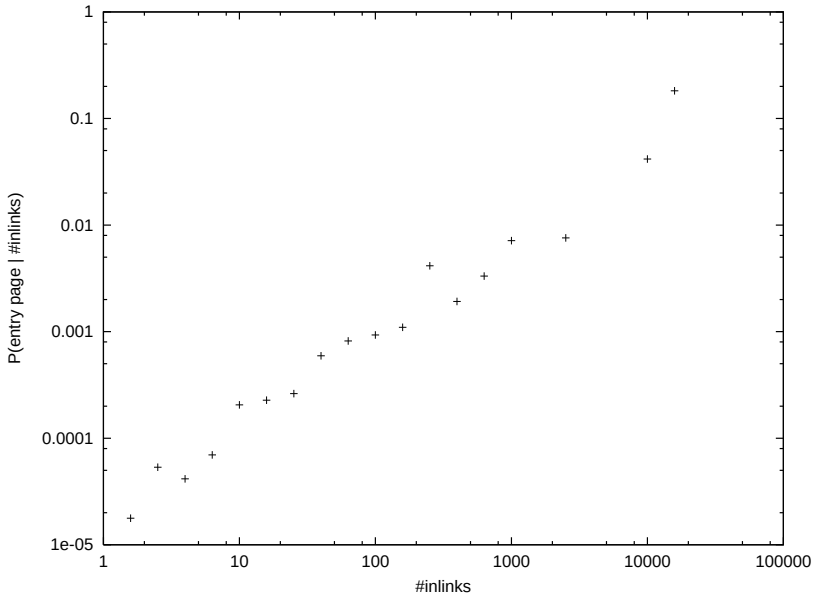

Figure 3.3. *Probabilité* a priori *de pertinence donnée le nombre des liens entrants ($P(entry\ page|inlinkCount)$), TREC-2001 collection d'apprentissage*

où R est une variable aléatoire binaire de « pertinence », $inlinkcount(D)$ est le nombre de liens entrants pour le document D et C une constante.

Notez que nous n'avons pas différencié les liens du même serveur et les liens d'autres serveurs. Faire cette différence pourrait être intéressant. En effet ces deux types de liens jouent des rôles distincts. Les liens d'autres serveurs impliquent souvent une forme de recommandation, tandis que les liens du même serveur fonctionnent principalement pour un objectif de navigation. Cependant, ces deux types de liens pointent probablement plus souvent vers les pages d'entrée que vers d'autres pages. Les liens de recommandation dirigent habituellement au point d'entrée principal d'un site et les sites sont souvent organisés de telle manière que chaque page lie à la page d'accueil.

3.4.3. *Forme de l'URL*

L'URL d'une page constitue une troisième source d'information *a priori*. Un URL définit l'endroit unique d'un document sur le *web*. Il comprend le nom du serveur suivi d'un chemin de répertoire et du nom du fichier. Une première observation est que les documents dont l'URL correspond au nom du serveur uniquement ou suivi par un (ou

plusieurs) répertoire(s) sont souvent des pages d'accueil. De plus, au fur et à mesure que nous descendons plus profond dans l'arbre des répertoires, la quantité relative des pages d'entrée diminue. Ainsi la probabilité d'être une page d'accueil semble être inversement liée à la profondeur du chemin dans l'URL correspondant.

Nous définissons quatre types d'URL :

– *root* : un nom de domaine, suivi de façon optionnelle par « index.html » (par exemple http://www.sigir.org) ;

– *subroot* : un nom de domaine, suivi par un seul nom de répertoire, suivi de façon optionnelle par « index.html » (par exemple http://www.sigir.org/sigirlist/) ;

– *path* : un nom de domaine, suivi par un chemin arbitrairement profond, finissant avec « index.html » ou juste « / » (par exemple http://www.sigir.org/sigirlist/issues/) ;

– *file* : toutes les autres formes d'URL (par exemple http://www.sigir.org/resources.html).

L'analyse des données de WT10g, la collection utilisée pour le *web track* de TREC-2001 (c'est-à-dire les pages d'accueil correctes), démontre que les pages d'entrée ont une distribution très différente selon les quatre catégories d'URL (voir tableau 3.1).

URL type	Entry pages		WT10g	
root	79	(73,1%)	12258	(0,7%)
subroot	15	(13,9%)	37959	(2,2%)
path	8	(7,4%)	83734	(4,9%)
file	6	(5,6%)	1557719	(92,1%)

Tableau 3.1. *Distributions des pages d'entrée et WT10g selon les catégories URL*

Comme nous le supposions, les pages d'entrée semblent être principalement de type *root* tandis que la plupart des documents dans la collection sont de type *file*. Pour des liens entrants, nous avons été capables de définir un modèle pour la probabilité *a priori* de pertinence étant donné le nombre de liens entrants, cette estimation est une assez bonne approximation de la vraie distribution. Pour la probabilité *a priori* basée sur la catégorie de l'URL, il est moins facile de définir un modèle analytique. Par conséquent, nous avons estimé la probabilité *a priori* de l'URL directement sur la collection d'entraînement :

$$P_{URL}(D) = P(EP|URLtype(D) = t_i) = \frac{c(EP, t_i)}{c(t_i)} \quad [3.10]$$

avec $URLtype(D)$ correspondant à la catégorie d'URL du document D, $c(EP, t_i)$ est le nombre de documents qui sont à la fois une page d'entrée et de type t_i et $c(t_i)$ est

le nombre de documents de type t_i. Les probabilités résultantes pour les différentes catégories d'URL sont énumérées dans le tableau 3.2. Notez que les probabilités *a priori* divergent de plusieurs ordres de grandeur ; la probabilité d'une page de racine (*root*) d'être une page entrée est presque 2 000 fois plus élevée que pour n'importe quelle autre page.

$$\begin{array}{ll} P(Entry\ page \mid root) & 6,44 \cdot 10^{-03} \\ P(Entry\ page \mid subroot) & 3,95 \cdot 10^{-04} \\ P(Entry\ page \mid path) & 9,55 \cdot 10^{-05} \\ P(Entry\ page \mid file) & 3,85 \cdot 10^{-06} \end{array}$$

Tableau 3.2. *Probabilités* a priori *pour les catégories URL différentes estimées sur les données d'entraînement*

3.5. Evaluation

Nous avons comparé les différentes estimations des probabilités *a priori* présentées dans la section précédente en utilisant les collections et les jugements de pertinence de la tâche de recherche de page d'accueil (TREC-2001). TREC (TREC8, TREC9, TREC10), est une campagne d'évaluation annuelle visant l'application à grande échelle des technologies de RI. La tâche de recherche de page d'accueil a été conçue pour comparer et évaluer des méthodes pour dépister des pages d'entrée. Pour cette tâche, une requête se compose d'une expression simple énonçant le nom d'un groupe ou d'une organisation spécifique.

Worldnet Africa
Hunt Memorial Library
Haas Business School
University of Wisconsin Lidar Group

Tableau 3.3. *Exemple de requêtes pour la tâche de recherche de page d'entrée*

Le tableau 3.3 en présente quelques exemples. Un système devrait rechercher la page d'accueil pour le groupe ou l'organisation décrite dans le sujet (*topic*) (besoin d'information). Par exemple, lorsque le sujet est *Groupe Lidar de l'Université Wisconsin*, la page d'accueil pour ce groupe spécifique devrait être retournée à l'usager. L'extraction de la page d'accueil de l'Université Wisconsin ou pour un sous-groupe du groupe de Lidar serait compté comme incorrecte. Les sujets pour cette tâche ont été créés en choisissant aléatoirement un document (point de départ) à partir de la collection. Ensuite, en suivant des liens on parcourt le *web* jusqu'à ce que l'on rencontre un document que l'on peut considérer comme étant la page d'accueil (*homepage*) du site contenant le document qui nous a servi de point de départ. De cette façon, 145

requêtes ont été construites pour cette tâche. En outre un ensemble de 100 requêtes a été fourni pour l'entraînement. Nous avons employé ce dernier ensemble pour effectuer les graphiques des paragraphes 3.4.1 et 3.4.2 à partir desquels nous avons déduit les probabilités *a priori* concernant la longueur des documents et celle des liens entrants. Le même ensemble a été employé pour estimer les probabilités *a priori* basées sur la forme de l'URL comme nous l'avons décrit dans le paragraphe 3.4.3. Les expériences rapportées dans la prochaine section sont menées en utilisant l'ensemble des 145 requêtes d'évaluation.

La collection test pour cette tâche est la collection WT10g [BAI 03], un sous-ensemble (10GB) de la collection VLC2. Cette dernière correspond à une partie extraite du *web* fait par Internet Archive en 1997[4]. Le corpus WT10g est conçu pour avoir relativement une haute densité d'hyperliens inter-serveurs.

Pour chaque combinaison de probabilité *a priori* et de modèle de langue, nous avons retourné les 100 premières réponses. Les expériences ont été évaluées par la mesure rang moyen réciproque (MRR). Pour chaque requête la valeur réciproque ($1/n$) du rang (n) de la première page d'accueil correcte est calculée. Pour un ensemble de requêtes, nous procédons à une moyenne arithmétique de la performance de chaque requête. Si aucune bonne réponse n'est trouvée pour une requête spécifique, la réciproque du rang pour cette requête est définie comme étant 0.

3.6. Expérimentations

Dans nos expérimentations nous avons utilisé le modèle de RI décrit dans la section 3.3. La probabilité *a priori* dans notre modèle soit $P(D)$, est choisie parmi celles définies dans la section 3.4 (elle peut ne pas être prise en compte, dans ce cas, on considère des probabilités *a priori* uniformes). Les modèles de langue ($P(Q|D)$) ont été estimés sur la collection de documents (WT10g) ou sur la collection composée des textes d'ancrage. Cette dernière collection a été produite à partir des données originales de WT10g en recueillant tous les textes d'ancrage des liens se dirigeant vers le même document. Cet ensemble de texte forme un pseudo-document qui a alors été employé comme une description du document original[5]. Nous n'avons pas différencié dans cette évaluation les liens du même serveur et les liens d'autres serveurs. Les documents étaient traités comme suit :

4. http://www.archive.org d'Internet.

5. Les expériences de cet article sont basées sur la collection d'ancrage de la CSIRO [CRA 01a, CRA 01b], tandis que les expériences rapportées dans [WES 01] sont basées sur une collection d'ancrage que nous avons construite nous-mêmes.

1) élimination des commentaires HTML et fragments Javascript ;

2) élimination des éléments de HTML ;

3) conversion des entités HTML en caractères ISO-Latin1 ;

4) élimination des mots vides ;

5) lemmatisation des termes en utilisant l'algorithme de Porter.

Le tableau 3.4 montre le rang réciproque moyen (MRR) pour des expériences avec des modèles évaluées à l'aide des pages *web* originales et la collection des textes d'ancrage. De plus nous avons considéré diverses probabilités *a priori*. Les dernières se basent sur le nombre de liens entrants ainsi que sur la forme de l'URL. Dans ces deux cas, l'efficacité de recherche est améliorée, en particulier la forme de l'URL apporte une amélioration très sensible. Nous poursuivrons cette discussion dans la section 3.7.

Méthode de classement	Contenu ($\lambda = 0,9$)	Ancrage ($\lambda = 0,9$)
$P(Q\|D)$	0,3375	0,4188
$P(Q\|D)P_{doclen}(D)$	0,2634	0,5600
$P(Q\|D)P_{URL}(D)$	**0,7705**	0,6301
$P(Q\|D)P_{inlink}(D)$	0,4974	0,5365

Tableau 3.4. *Résultats pour les différentes probabilités* a priori

3.6.1. *Combinaison des modèles avec les textes d'ancrage*

Nous avons combiné les résultats obtenus pour le contenu des pages *web* avec celui des textes d'ancrage par une interpolation linéaire : $RSV' = \alpha RSV_{page} + (1 - \alpha)RSV_{anchor}$.

α	MRR
0,5	0,3978
0,7	0,4703
0,8	**0,4920**
0,9	0,4797

Tableau 3.5. *Combinaison du texte des pages* web *et du texte d'ancrage*

Le tableau 3.5 démontre que bien que la stratégie de combinaison ne soit pas optimale (voir la section 3.3), la combinaison des deux représentations (contenu et texte d'ancrage) s'avère pertinente. Le meilleur résultat est réalisé pour $\alpha = 0,8$: MRR=0,4920 (si on se limite au texte d'ancrage on obtient : MRR = 0,4188). Nous

avons également étudié l'effet des probabilités *a priori* sur des systèmes combinés (voir tableau 3.6).

Méthode de classement	Contenu+Ancrages ($\alpha = 0,8$)
$P(Q\|D)$	0,4920
$P(Q\|D)P_{URL}(D)$	**0,7748**
$P(Q\|D)P_{inlink}(D)$	0,5963

Tableau 3.6. *Résultats pour différentes probabilités* a priori *différents (contenu+texte d'ancrage)*

3.6.2. *Combinaison des probabilités* a priori

Dans la section 3.4, nous avons vu que le nombre des liens entrants et la forme de l'URL peuvent fournir une information intéressante afin de déterminer si une page est une page d'accueil ou non. Les résultats du tableau 3.4 confirment cette hypothèse. Une combinaison de ces deux sources d'information pourrait également être utile. Dans le paragraphe 3.3.2 nous avons présenté de diverses approches pour combiner les probabilités *a priori*. Le tableau 3.8 montre les résultats pour les deux approches de combinaison. Toutes les expériences sont basées sur le corpus de page *web* et ne considèrent pas le texte d'ancrage. Dans ce cas, le problème principal demeure la combinaison des caractéristiques car il s'avère difficile d'estimer $P(EP|\bar{\mathbf{x}})$ directement. En effet, la collection d'entraînement est trop petite. Une approche simple consiste à multiplier les probabilités *a priori* individuelles, correspondant à une approximation d'un modèle fondé sur l'hypothèse indépendance conditionnelle. Cette expérience de combinaison permet d'obtenir une valeur MMR de 0,7504, ce qui reste inférieur à une expérience basée seulement sur la probabilité *a priori* de la forme de l'URL. Nous pensons que ce résultat décevant provient du fait qu'il existe une dépendance conditionnelle entre les caractéristiques ou que l'approximation du dénominateur n'est pas assez précise. Nous avons étudié une variante, sans admettre l'indépendance dans laquelle nous avons estimé directement $P(EP|\bar{\mathbf{x}})$, sur la base d'un modèle avec un nombre réduit de paramètres.

Pour cette approche nous ne pouvions pas travailler avec la formule [3.9]. Nous avons alors discrétisé l'espace des valeurs des liens entrants par le biais de quatre catégories sur un axe logarithme. Ensuite, nous avons défini sept catégories disjointes de documents. Le tableau 3.7 montre les diverses statistiques. Puisqu'il n'y a que quelques pages d'entrée dans l'ensemble d'entraînement de catégorie *non-root*, nous avons seulement subdivisé la classe *root* dans quatre catégories disjointes basés sur le nombre de liens entrants.

Catégorie de document t_i	pages d'entrée	WT10g
root avec 1-10 liens entrants	39 (36,1%)	8938 (0,5%)
root avec 11-100 liens entrants	25 (23,1%)	2905 (0,2%)
root avec 101-1 000 liens entrants	11 (10,2%)	377 (0,0%)
root avec 1 000+ liens entrants	4 (3,7%)	38 (0,0%)
subroot	15 (13,9%)	37959 (2,2%)
path	8 (7,4%)	83734 (4,9%)
file	6 (5,6%)	1557719 (92%)

Tableau 3.7. *Distribution selon les pages d'entrée et l'ensemble du corpus WT10g et selon les diverses catégories retenues*

A partir de ces statistiques nous avons calculé une probabilité *a priori* basée sur les deux caractéristiques liens entrants et forme de l'URL en utilisant :

$$P_{inlink-URL}(EP|dt(D) = t_i) = \frac{c(EP, t_i)}{c(t_i)} \quad [3.11]$$

dans laquelle $dt(D)$ indique la catégorie du document D et $c(.)$ est défini dans l'équation [3.10].

Le tableau 3.8 récapitule les résultats pour les expériences de combinaison. En plus des résultats avec les probabilités *a priori* simples (tableau 3.2), ce tableau indique l'évaluation dans laquelle la probabilité *a priori* basée sur les liens entrants a été estimée sur les données d'entraînement au lieu d'être estimée selon la formule [3.3]. Le MRR de cette approche (MRR=0,4752) s'avère inférieur à l'expérience basée sur la probabilité *a priori* représentée par [3.3], indiquant probablement que le nombre de catégories est trop faible.

Les expériences avec des combinaison de probabilités *a priori* montre une différence en performance : l'expérience basée sur une multiplication naïve des probabilités *a priori* a un résultat inférieur à celui de l'expérience basée sur la probabilité *a priori* de l'URL. En revanche, l'expérience estimant $P(EP|\bar{\mathbf{x}})$ directement et basée sur sept catégories disjointes améliore la MRR (dernière ligne du tableau 3.8). On en conclut alors qu'une combinaison des probabilités *a priori* donne une performance plus élevée que l'utilisation des probabilités *a priori* séparément.

3.7. Discussion

La tâche de recherche de pages d'entrée est sensiblement différente de la recherche *ad hoc*. La différence principale réside dans le fait que juste une page est pertinente.

Méthode	MRR sur contenu	description
$P(Q\|D)P_{URL}(D)$	0,7705	
$P(Q\|D)P_{inlink}(D)$	0,4974	probabilité *a priori* analytique [3.9]
$P(Q\|D)P_{inlink}(D)$	0,4752	4 catégories
$P(Q\|D)P_{URL+inlink}(D)$	0,7504	système de référence : produit des probabilités *a priori* (prob. *a priori* analytique (# liens entrants))
$P(Q\|D)P_{URL+inlink}(D)$	**0,7749**	estimation directe(7 catégories disjointes)

Tableau 3.8. *Combinaison des probabilités* a priori

Les techniques d'amélioration de rappel ne jouent pas un rôle, et risquent, au contraire, de détériorer la précision. Nous avons effectué des expériences avec des modèles s'appuyant uniquement sur le contenu des pages, ainsi que des approches combinant contenu et texte d'ancrage. Enfin nous avons tenu compte d'autres caractéristiques des pages comme le nombre de liens entrant et la forme URL.

Il s'avère intéressant de noter qu'un système basé uniquement sur le texte d'ancrage surpasse un système basé uniquement sur le texte de la page. Les textes d'ancrage contiennent souvent le nom de la page d'entrée de l'organisation. La combinaison de ces deux représentations textuelles par interpolation simple s'avère efficace : MRR = 0,4920 (+17%).

Afin de discriminer entre les pages d'accueil et les autres pages d'un site, le contenu seul ne s'avère pas efficace. En effet, les mots de la requête peuvent se trouver dans les deux types de pages. Par conséquent, d'autres caractéristiques doivent être employées. Nous avons examiné trois caractéristiques différentes : la longueur du document, le nombre de liens entrants et la forme de l'URL. Une probabilité *a priori* basée sur une fonction linéaire de la longueur du document a diminué de manière significative le MRR pour les expériences par rapport à une évaluation basée sur les pages *web* (contenu). Ceci s'explique puisqu'une telle fonction ne correspond pas du tout aux données d'entraînement. Néanmoins il est intéressant de voir qu'une probabilité *a priori* mal choisie peut détériorer les résultats. La probabilité *a priori* estimée comme fonction linéaire de longueur du document semble améliorer les résultats par rapport au texte d'ancrage, mais c'est une illusion. En effet, la longueur d'un document composé des textes d'ancrage est une concaténation des textes d'ancrage des liens se dirigeant vers une page *web* particulière. Ainsi, la longueur de ce pseudo document est linéairement liée au nombre de liens entrants. Les probabilités *a priori* basées sur les liens entrants ou sur la longueur de ce pseudo document ont des gains de performance

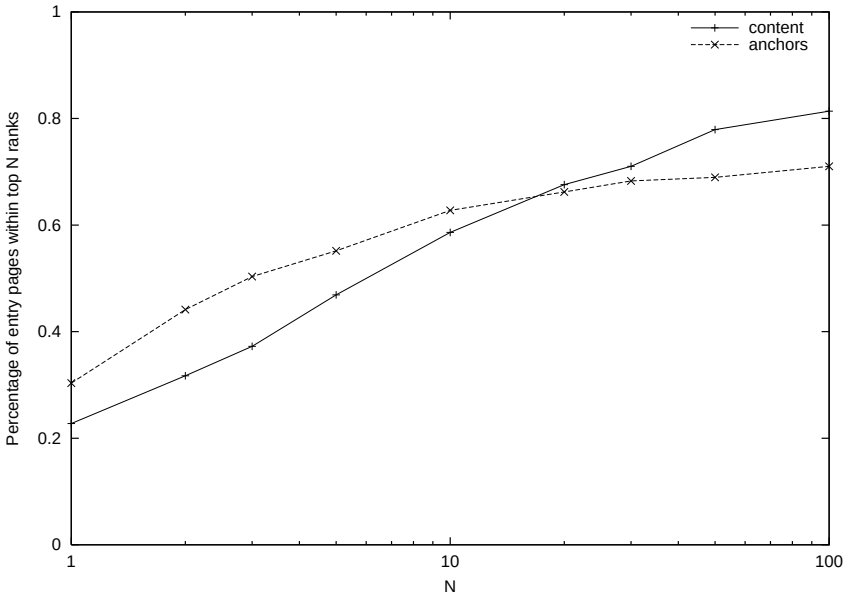

Figure 3.4. *Succès à N pour les expériences basée sur le contenu des pages* web *et les texte d'ancrage*

similaires. Il semble cependant y avoir une amélioration additionnelle avec la probabilité *a priori* basée sur la longueur du document ; il semble que des textes d'ancrage plus longs sont corrélés avec des pages d'entrée. Les meilleurs résultats sont réalisés par les expériences avec la probabilité *a priori* basée sur la forme de l'URL. Étonnamment, le meilleur résultat a été obtenu avec une expérience basée sur des textes de pages *web* originales. Ceci peut être expliqué par le fait que ce système possède un plus grand rappel que le système basé sur des textes d'ancrage. La probabilité *a priori* basée sur l'URL est un dispositif puissant pour améliorer la précision et qui se combine bien avec le système utilisant le contenu des pages. La figure 3.4 illustre ceci : tandis que le système d'ancrage possède un taux initial élevé de succès ($30,3\%$ des pages d'entrée ont été trouvées au grade 1 ; $22,8\%$ pour le contenu), le système basé sur le contenu des pages trouve en fin de compte plus de pages d'entrée ($81,4\%$ trouvé parmi les 100 requêtes ; $71,0\%$ pour des ancrages). Nous avons également étudié si nous pourrions combiner les caractéristiques URL et le nombre de liens entrants. Une simple multiplication des probabilités *a priori* (en assumant l'indépendance conditionnelle et statistique) a détérioré les résultats. Une autre stratégie, fondée sur un nombre réduit de classes, mais sans aucune supposition d'indépendance a permis un petit gain en comparaison avec la probabilité *a priori* de la catégorie d'URL. Nous pensons que nous pourrions améliorer ces résultats. Avec plus de données d'entraînement

nous pourrions estimer les probabilités conditionnelles communes des deux caractéristiques d'une manière plus précise. Le fait que la probabilité *a priori* des liens entrants a obtenu une plus mauvaise performance que la basée sur une fonction analytique [3.9] indique que le nombre des catégories est probablement trop faible. Cette faible valeur pourrait également avoir contribué aux résultats décevants des expériences basées sur une combinaison avec la probabilité *a priori* estimée sur les données d'entraînement. Nous avons évalué la dernière hypothèse, par un raffinage plus poussé de la catégorie *root* par le biais de 9 catégories (basé sur des classes de #inlinks). Ceci a eu comme conséquence une autre amélioration (MRR = 0,7832 ; 72% des réponses avec la bonne réponse en première position ; 89% dans les 10 premiers rangs). Ce résultat confirme notre hypothèse. Nous avons également refait l'expérience avec la probabilité *a priori* basée sur des catégories de nombre des liens entrants, avec 9 catégories au lieu de 4 catégories : le MRR obtenu s'élevait à 0,5064, ce qui est encore meilleur que la probabilité *a priori* des liens entrants analytique. On peut ainsi conclure que le nombre de classes est critique pour la performance d'un système utilisant des probabilités *a priori* estimées sur des donnée d'entraînement.

3.8. Conclusion

Plusieurs caractéristiques des pages *web* peuvent aider à améliorer l'efficacité d'un système de recherche de page d'accueil. Les principales sont le nombre des liens entrants et la forme de l'URL. Nous avons démontré qu'un système de RI basé sur des modèles de langue unigramme peut s'adapter à ces types d'information très facilement, en estimant la probabilité *a posteriori* d'une page donnée selon ses caractéristiques et en employant cette probabilité comme probabilité *a priori* dans le modèle. La caractéristique de la forme de l'URL s'est avérée avoir la puissance prédictive la plus forte, récupérant plus de 70% de pages d'entrée en premier position, et jusqu'à 89% dans les 10 premières places. Les expériences initiales combinant les deux caractéristiques ont montré une autre amélioration. La performance du modèle unigramme combiné avec la probabilité *a priori* est extrêmement sensible au nombre de paramètres. Toutefois, au vu de la relative faible taille de la collection d'entraînement nous ne pouvons pas modéliser l'ensemble des probabilités et des relations entre les caractéristiques de manière stable.

Remerciements

Cette recherche a été financée par le projet DRUID de l'institut de télématique, aux Pays-Bas. Nous voudrions remercier David Hawking et Nick Craswell de CSIRO pour avoir rendu disponible leur collection de textes d'ancrage. Nous remercions Lyne Blanchette et Jacques Savoy d'avoir corrigé le manuscrit français.

3.9. Bibliographie

[AME 00] AMENTO B., TERVEEN L., HILL W., « Does "authority" mean quality ? Predicting expert quality ratings of Web documents », dans [CRO 00], p. 296-303, 2000.

[BAI 03] BAILEY P., CRASWELL N., HAWKING D., « Engineering a multi-purpose test collection for web retrieval experiments », *Information Processing and Management*, vol. 39, n° 6, p. 853-871, 2003.

[BHA 98] BHARAT K., HENZINGER M. R., « Improved algorithms for topic distillation in a hyperlinked environment », dans [CRO 98], p. 104–111, 1998.

[BRI 98] BRIN S., PAGE L., « The anatomy of a large-scale hypertextual Web search engine », *Computer Networks and ISDN Systems*, vol. 30, n°1–7, p. 107–117, 1998.

[COH 00] COHN D., CHANG H., « Learning to Probabilistically Identify Authoritative Documents », *Proceedings of the* 17^{th} *International Conference on Machine Learning*, 2000.

[CRA 01a] CRASWELL N., HAWKING D., ROBERTSON S., « Effective site finding using link anchor information », dans [CRO 00], p. 250-7, 2001.

[CRA 01b] CRASWELL N., HAWKING D., WILKINSON R., WU M., « TREC10 Web and Interactive Tracks at CSIRO », dans [VOO 01b], p. 261–268, 2001.

[CRI 01] CRIVELLARI F., MELUCCI M., « Web Document retrieval using Passage Retrieval, Connectivity Information, and Automatic Link Weighting - TREC-9 Report », dans [VOO 01a], p. 611–620, 2001.

[CRO 98] CROFT W. B., MOFFAT A., VAN RIJSBERGEN C., WILKINSON R., ZOBEL J. (Dir.), *Proceedings of the* 21^{st} *Annual International Conference on Research and Development in Information Retrieval*, 1998.

[CRO 00] CROFT W. B., HARPER D. J., KRAFT D. H., ZOBEL J. (Dir.), *Proceedings of the* 23^{rd} *Annual International Conference on Research and Development in Information Retrieval*, 2000.

[DAV 00] DAVISON B. D., « Topical locality in the Web », dans [CRO 00], p. 272-9, 2000.

[GUR 01] GURRIN C., SMEATON A. F., « Dublin City University Experiments in Connectivity Analysis for TREC-9 », dans [VOO 01a], p. 179–188, 2001.

[HAW 00] HAWKING D., VOORHEES E., CRASWELL N., BAILEY P., « Overview of the TREC-8 Web Track », vol. 8, National Institute of Standards and Technology, NIST, p. 131–148, 2000.

[HAW 01a] HAWKING D., « Overview of the TREC-9 Web Track », dans [VOO 01a], p. 87–102, 2001.

[HAW 01b] HAWKING D., CRASWELL N., « Overview of the TREC-2001 Web Track », dans [VOO 01b], p. 25–31, 2001.

[HIE 98] HIEMSTRA D., KRAAIJ W., « Twenty-One at TREC-7 : Ad hoc and Cross-language track », dans [VOO 99], p. 227-238, 1998.

[HIE 01] HIEMSTRA D., Using language models for information retrieval, PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[KLE 99] KLEINBERG J. M., « Authoritative sources in a hyperlinked environment », *Journal of the ACM*, vol. 46, n° 5, p. 604–632, 1999.

[KRA 01a] KRAAIJ W., SPITTERS M., VAN DER HEIJDEN M., « Combining a mixture language model and Naive Bayes for multi-document summarisation », *Working notes of the DUC2001 workshop (SIGIR2001), New Orleans*, 2001.

[KRA 01b] KRAAIJ W., WESTERVELD T., « TNO/UT at TREC-9 : How different are Web Documents ? », dans [VOO 01a], p. 665–671, 2001.

[LAF 01] LAFFERTY J., ZHAI C., « Probabilistic IR Models Based on Document and Query Generation », dans CALLAN J., CROFT B., LAFFERTY J. (dir.), *Proceedings of the workshop on Language Modeling and Informati on Retrieval*, 2001.

[LEW 98] LEWIS D. D., « Naive (Bayes) at forty : The independence assumption in information retrieval. », dans NÉDELLEC C., ROUVEIROL C. (dir.), *Proceedings of ECML-98*, 10^{th} *European Conference on Machine Learning*, n° 1398, Lecture Notes in Computer Science, Chemnitz, DE, Springer Verlag, Heidelberg, DE, p. 4–15, 1998.

[MIL 98] MILLER D. R. H., LEEK T., SCHWARTZ R. M., « BBN at TREC-7 : using Hidden Markov Models for Information Retrieval », dans [VOO 99], p. 133-142, 1998.

[MIT 97] MITCHELL T., *Machine Learning*, McGraw Hill, 1997.

[NG 01] NG A. Y., ZHENG A. X., JORDAN M. I., « Stable algorithms for link analysis », dans [CRO 00], p. 258-266, 2001.

[PON 98] PONTE J. M., CROFT W. B., « A Language Modeling Approach to Information Retrieval », dans [CRO 98], p. 275-281, 1998.

[RA 01] RA D.-Y., PARK E.-K., JANG J.-S., « Yonsei/ETRI at TREC-10 : Utilizing Web Document Properties », dans [VOO 01b], p. 643–650, 2001.

[ROB 94] ROBERTSON S. E., WALKER S., « Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval », *Proceedings of the* 17^{th} *ACM Conference on Research and Development in Information Retrieval (SIGIR'94)*, p. 232-241, 1994.

[SAL 88] SALTON G., BUCKLEY C., « Term-Weighting Approaches in Automatic Text Retrieval », *Information Processing & Management*, vol. 24, n° 5, p. 513-523, 1988.

[SAV 01a] SAVOY J., RASOLOFO Y., « Report on the TREC-10 Experiment : Distributed Collections and Entrypage Searching », dans [VOO 01b], p. 578–590, 2001.

[SAV 01b] SAVOY J., RASOLOFO Y., « Report on the TREC-9 Experiment : Link-Based Retrieval and Distributed Collections », dans [VOO 01a], p. 579–588, 2001.

[SIN 96] SINGHAL A., BUCKLEY C., MITRA M., « Pivoted Document Length Normalization », *Proceedings of the* 19^{th} *ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*, p. 21-29, 1996.

[VOO 99] VOORHEES E. M., HARMAN D. K. (Dir.), *The Seventh Text Retrieval Conference (TREC7)*, vol. 7, National Institute of Standards and Technology, NIST, 1999.

[VOO 01a] VOORHEES E. M., HARMAN D. K. (Dir.), *The Ninth Text Retrieval Conference (TREC9)*, vol. 9, National Institute of Standards and Technology, NIST, 2001.

[VOO 01b] VOORHEES E. M., HARMAN D. K. (Dir.), *The Tenth Text Retrieval Conference (TREC-2001)*, vol. 10, National Institute of Standards and Technology, NIST, 2001.

[WES 01] WESTERVELD T., HIEMSTRA D., KRAAIJ W., « Retrieving Web Pages Using Content, Links, URL's and Anchors », dans [VOO 01b], p. 52–61, 2001.

[XI 01] XI W., FOX E. A., « Machine Learning Approaches for Homepage Finding Tasks at TREC-10 », dans [VOO 01b], p. 633–642, 2001.

[ZOB 98] ZOBEL J., MOFFAT A., « Exploring the Similarity Space », *SIGIR FORUM*, vol. 32, n° 1, p. 18-34, 1998.

Index

web, 75–78, 81–86, 88, 91, 93–95
estimation, 82, 87
forme de l'URL, 89, 91
hyperlien, 77, 83, 85, 89
modèle de langue, 78, 79, 83, 84
page d'entrée, 75, 93
probabilité *a priori*, 76–79, 81–85, 87–95
texte d'ancrage, 90, 91, 93, 94