* 1. Introduction

Les systèmes de recherche d’information exploitent dans leur majorité deux classes de sources d’évidence pour trier les documents répondant à une requête. La première, la plus exploitée, est dépendante de la requête, elle concerne toutes les caractéristiques relatives à la distribution des termes de la requête dans le document et dans la collection (tf-idf). La seconde classe concerne des facteurs indépendants de la requête, elle mesure une sorte de qualité ou d’importance a priori du document. Parmi ces facteurs, on en distingue la longueur du document, le type d’URL du document, la présence d’URL dans le document, ses auteurs, etc.

Dans ce chapitre, nous nous organiserons comme suit : dans un premier lieu nous allons définir la pertinence à priori des documents , définir les caractéristiques qui permettent de classer les documents selon leur pertinence à priori, présenter le modèle de langue pour enfin , présenter les différentes méthodes de combinaison de scores à priori et dépendant de la requête.

* 1. Définition de la pertinence à priori de documents

La pertinence à priori des documents est la probabilité qu’un document soit plus pertinent qu’un autre car ils différent dans certaines propriétés indépendantes de la requête, ces propriétés peuvent être la taille de document, le nombre de liens entrants, etc. Si la probabilité a priori de pertinence d’un document n’est pas conditionnée par l’une de ces propriétés alors cette probabilité représente la probabilité de prélever un document de la collection. Par conséquent, tous les documents dans la collection ont la même probabilité d’être sélectionnés, et donc la probabilité a priori de pertinence de document peut être ignorée lors du classement des documents. Par contre, si la probabilité a priori est conditionnée par l’une de ces caractéristiques alors les documents de la collection n’ont pas la même probabilité a priori, et donc les documents ne sont pas équiprobables. Par exemple, si la caractéristique utilisée est le score de popularité de document alors un document populaire est plus probable d’être pertinent qu’un document moins populaire.

* 1. Les caractéristiques utilisées pour le classement à priori des documents

Plusieurs caractéristiques ont étés utilisées pour estimer la probabilité a priori d’un document, on peut citer : la longueur du document **[33][34**],la structures des liens**[33][36]**, le facteur temps **[37][38]**et le rapport information/bruit**[39].**

L’intuition derrière l’utilisation de ces caractéristiques est que : un document est plus probable d’être pertinent car : il est plus long, il est plus populaire, il est plus récent, ou contient plus d’informations que de bruit. Nous présentons ci-dessous quelques travaux utilisant ces caractéristiques.

* + 1. La structure de liens

L’idée derrière l’utilisation de la structure des liens est que les documents populaires ou les plus cités tendent à être plus pertinents. La méthode simple d’utilisation de la structure des liens est l’usage du nombre de liens entrants. La probabilité de pertinence a priori est alors exprimée comme suit :

P(d) = (II.1)

Où ) est le nombre de liens entrants dans le document.

C : est la collection de documents.

D’autres facteurs plus sophistiqués on été utilisés comme : le Hits **[40]** , le PageRank **[41]** qui est à l’origine du moteur de recherche Google. Le principe de cet algorithme consiste à ordonner les pages web selon leur popularité, en se basant sur l’hypothèse suivante : « une page est populaire (importante) quand elle est beaucoup citée ou citée par une page très populaire». L’estimation de cette popularité est formalisée comme suit :

*PR(p) = (1 – d) × + d ×* (II.2)

Où :

PR(p) : est le PageRank de la page p.

T : est le nombre total de pages sur le web (indexées) ;

d : est un paramètre fixé à 0.85 ;

C(pi) : est le nombre de liens sortant de la page pi , et k est le nombre de pages qui pointent la page p.

De nombreux travaux ont montré que l’incorporation du PageRank dans le classement des pages web, améliore les performances de recherche sur de grandes collections de type web **[42][43][44].**

* + 1. La taille de document

L’intuition de l’utilisation d’une telle caractéristique est qu’un document plus long tend à contenir plus d’informations et par conséquent il est plus probable d’être pertinent. Les résultats obtenus avec l’utilisation de cette caractéristique ont été mixtes et cela selon la collection utilisée **[33].**

La probabilité de pertinence a priori est proportionnelle à la taille du document, elle est exprimée ainsi:

P(d) = (II.3)

Où est la taille de document et est la taille de la collection.

Parapar et al **[34]** ont proposé d’estimer cette probabilité *P(d)* en utilisant la taille compressée d’un document. La formule utilisée dans ce dernier travail est exprimée ainsi :

P(d) = (II.4)

Où *comp(d*) est la taille en octets du document ***d*** compressé (zippé) divisée sur la taille originale en octets du document. Ce nouveau facteur a été évalué et comparé au facteur taille originale de document en utilisant quatre collections TREC. Les résultats présentés montrent que la taille compressée d’un document obtient des améliorations de précision moyenne (MAP) allant de +0,4% à +3,1% par rapport à l’utilisation de la taille originale de document.

* + 1. La date de création de document

La caractéristique de la date de création de document est aussi utilisée sous l’intuition suivante : « Les documents récents tendent à être plus pertinents que les documents anciens », pour estimer la probabilité a priori d’un document, Li et Croft **[35]**ont proposé un modèle de langue qui permet d’intégrer la notion de « temps » dans l’évaluation de pertinence d’un document vis-à vis d’une requête, où ils assignent une plus grande probabilité de pertinence pour les documents ayant une date de création récente. Ainsi, ils expriment La probabilité de pertinence a priori d’un document sachant sa date de création, comme une distribution exponentielle, exprimée ainsi:

P() = λ (II.5)

Où TC est la date la plus récente dans toute la collection (exprimée en mois) et Td est la date de création du document.

Les évaluations réalisées ont montré que l’incorporation de la notion de temps en utilisant la distribution exponentielle est bénéfique pour la RI**.**

* + 1. Le rapport information/bruit

L’intuition de l’utilisation d’une telle caractéristique est qu’un document contenant le plus de tokens après le prétraitement est susceptible d’être plus pertinent qu’un document qui en a moins.

Il est défini comme le rapport entre la taille de document après prétraitement (élimination des mots vides et des balises HTML) et la taille de document sans prétraitement **[39].**

La probabilité de pertinence a priori est exprimée comme suit :

*P(d) =* (II.6)

Où *Itoken* est la taille de document après le prétraitement e*t Idocument* est la taille de document avant le prétraitement. Ainsi, un document avec moins de mots vides et peu de balises HTML produit un haut rapport information/ bruit, ce qui signifie que le document est de « bonne » qualité.

* + 1. Type d’URL de document

Kraaij et al **[33]** ont utilisé la forme (type) de l’URL pour estimer la probabilité qu’une page soit une page d’entrée. Elle est définie ainsi :

*P(d) = P(PE (d)) =* (II.7)

Où urltype est le type de l’URL de document d , c(PE, ti) est le nombre de documents du type d’URL « ti » qui sont des pages d'entrées « PE » pour un site web, il est obtenu à partir des évaluations de pertinence et c(ti) est le nombre de documents de type d’URL « ti ». Quatre types de catégories d’URL ont été définis : Racine, sous-racine, chemin (répertoire) et Fichier. Sur la base de ces quatre types d’URL, ils ont mené des expérimentations sur la collection web WT10g (collection utilisée dans TREC 2001) pour estimer la probabilité qu’une page soit une page d’entrée sachant son type d’URL. Ils ont constaté que cette source d’information est un bon indicateur pour prévoir la pertinence d’une page.

* 1. Le modèle de langue et la pertinence à priori de documents

Le modèle de langue permet de prendre en compte des fonctionnalités indépendantes des requêtes, c'est-à-dire simplement liées à un document sans avoir de requête. Ensuite, il suit une dérivation du modèle de récupération ML où la probabilité de pertinence p (r | Q, D), étant donné une requête et un document est estimée indirectement en invoquant la règle de Bayes.

Supposons que les variables aléatoires D et Q désignent respectivement un document et une requête. Soit la variable aléatoire binaire R pour la pertinence r, p (r) = p (R = 1) et la non-pertinence , p () = p (R = 0).

*P(r* (II.8)

= *p(Q* (II.9)

= *p(Q* (II.10)

En supposant l'indépendance entre les requêtes et les documents p (D, Q) = p (D) p (Q), et étant donné que p (Q) n'affecte pas le classement (il est indépendant du document), l'équation devient

*P(r* (II.11)

Où p (Q | D, r) est la probabilité de la requête pour le document D et p (r | D) est la probabilité de pertinence à priori du document. Dans l'équation (II.11), nous supposons que la probabilité de pertinence à priori de document est beaucoup plus que la pertinence de la requête pour le même document pour obtenir une formule finale dépendante de p (r | D). La dérivation présentée précédemment a pris une hypothèse plus raisonnable, Q et D sont indépendants sous r, et à partir du rapport de cotes de pertinence, le score de pertinence final dépend de p (r | D) / (1 - p (r | D).

La requête est décomposée en ses termes Q = {q1, q2,. . . , qn} et supposons que, la pertinence r et le document D sont indépendants les uns des autres et sont générés par une distribution multinomiale.

P(Q (II.12)

Afin d'exclure les probabilités nulles pour les termes non vus dans un document, cette estimation doit être lissée.

La plupart des méthodes de lissage utilisent deux distributions, l'une pour les mots apparaissant dans le document (ps) et l'autre pour les mots invisibles (pu).

En introduisant le logarithme, on peut montrer que l'équation suivante (II.13) suffit pour fournir un classement de document en utilisant des sommes de logarithmes, équivalentes à celle que donnerait l'équation précédente (II.12).

log p(Q + n log (II.13)

où tf (qi, D) représente la fréquence du terme *qi* dans le document D, est un paramètre et p (qi | C) est le modèle de langue de la collection.

On peut citer deux techniques de lissage :

La technique de lissage la plus populaire et efficace est le lissage préalable de **Dirichlet**:

, αd = (II.14)

Où μ est un paramètre.

Mais il y’ a aussi le lissage de **Jelinek Mercer** : ce type de lissage consiste à combiner le modèle de langue considéré avec un ou plusieurs modèles de références estimés sur d’autres corpus d’apprentissages. Typiquement, dans le cas de collection de documents, on pourrait par exemple estimer le modèle de document en le combinant avec le modèle de la collection. Dans ce cas, le modèle de document est exprimé ainsi :

+λ (II.15)

Tel que m est le modèle de langue, d est le document. , C  est la collection et λ  est un paramètre.

Les modèles et sont estimés selon le maximum de vraisemblance.

* 1. Combinaison du score à priori et du score Document/Requête

Pour combiner le score à priori de document avec le score initial (document/ requête) p (Q | D, r) en utilisant la somme logarithmique standard présentée dans la formule II.16.

Si nous suivons une dérivation de somme logarithmique de l'équation (II.11) alors, la manière standard de combiner le document a priori avec la requête afin de produire un score de document serait:

*Score (D, Q) = log p(Q) + log p(r)* (II*.*16)

Tel que :

p(Q) : est la probabilité de pertinence de la requête Q pour chaque document D.

p(r) : est la probabilité de pertinence à priori du document D.

* 1. Conclusion

Dans ce chapitre nous avons défini la pertinence à priori de documents, ainsi que les caractéristiques utilisées pour effectuer ce classement .Ensuite, nous avons évoqué la pertinence à priori de documents dans le modèle de langue. Pour enfin présenter le modèle de combinaison de la probabilité à priori de document et de la requête en utilisant la somme logarithmique.

Dans le prochain chapitre nous allons mettre en œuvre une approche qui nous permettra d’étendre le modèle de recherche avec le score à priori de documents et constater une possible amélioration de notre système de recherche possible amélioration de notre système de recherche.