

Université libanaise
Faculté des sciences

Université Paul Sabatier – I.R.I.T.

DEA d'Informatique

Coopération dans les sciences de traitement de l'information

Année universitaire 2005/2006

**Reformulation de requêtes dans un modèle de réseau possibiliste
pour la recherche d'information**

Préparé par Hassan CHOUAIB

Responsable(s) Mohand BOUGHANEM

Jury Dr. Bilal Chbaro

Dr. Ali Awada

Dr. Kabalan Barbar



REMERCIEMENTS

Je tiens à remercier très sincèrement Monsieur Mohand BOUGHANEM Professeur à l'université Paul Sabatier de Toulouse, pour m'avoir encadré, pour ses conseils judicieux et sa disponibilité. Qu'il soit assuré de ma profonde gratitude.

Je tiens à adresser mes remerciements à :

Monsieur Jean-Paul Bahsoun, Professeur à l'Université Paul Sabatier à Toulouse et directeur de l'UFR MIG de Toulouse.

Monsieur Bilal Chebaro, Docteur à l'Université Libanaise, pour son encadrement durant mon stage au Liban, sa patience, ses remarques et ses conseils.

Tous les membres de l'équipe SIG de l'IRIT, pour leur sympathie, leur gentillesse, leur compétence et leurs qualités humaines.

Enfin, j'adresse mes remerciements à mes parents pour leur soutien moral, leur encouragement et leur sacrifice.



Résumé

Mots clés :

système de recherche d'information, réinjection de pertinence(Relevance Feedback), théorie de possibilité, modèle possibiliste pour la recherche d'information, modèle de Rocchio.

Résumé :

Ce travail s'inscrit dans le cadre de la reformulation de requêtes pour les systèmes de recherche d'information. En effet nous avons ajouté un processus de reformulation de requête au modèle de réseaux possibilistes. Ce modèle modélise la pertinence en définissant deux aspects de la pertinence d'un document étant donnée une requête : la pertinence nécessaire et la pertinence possible. Les documents préférés et restitués en haut de liste sont les documents nécessairement pertinents. Alors nous avons profité de la double valeur de pertinence et des jugements de l'utilisateur pour trouver les meilleurs termes à ajouter dans la requête initiale. Des expérimentations sur la collection de test WT10g montrent l'intérêt de notre approche.



Abstract

Key Words:

Information Retrieval System, relevance feedback, theory of possibility, possibilistic model for information retrieval, Rocchio model

Abstract:

This work lies within the scope of the reformulation of requests for information retrieval system. We added a process of reformulation of request to the possibilistic network model. This model defines two aspects of the relevance of a document being given a request: relevance necessary and possible relevance. Then we benefited from the double value of relevance and the judgments of the user to find the best terms to be added in the initial request. Experiments on the collection of WT10g test show the interest of our approach.

TABLE DE MATIERES

INTRODUCTION GENERALE	1
------------------------------------	----------

CHAPITRE 1 PRINCIPE DE LA RECHERCHE D'INFORMATION

1.1- Introduction	4
1.2- Notions de base	4
1.3- Concepts clés de la recherche d'information (RI):.....	4
1.4 - Principaux modèles de recherche d'information :	5
1.4.1- Le modèle booléen.....	5
1.4.2-Modèle Booléen basé sur des ensembles flous.....	6
1.4.3-Le modèle vectoriel (Vector Space Model).....	8
1.4.4- Le modèle probabiliste :	9
1.4.4.1- Principe.....	10
1.4.4.2- Hypothèse d'indépendance et le modèle de recherche indépendant	11
1.4.5- Le modèle BNR (modèle RI basé sur les réseaux Bayésiens).....	13
1.4.5.1-Architecture générale du modèle.....	13
1.4.5.2- Estimation des distributions de probabilité	14
1.4.6- Le modèle possibiliste	15
1.4.6.1-Architecture générale du modèle.....	16
1.4.6.2- Evaluation de la requête	16
1.4.6.3 Agrégation des termes de la requête	18
1.4.6.4- Pondération des termes d'indexation.....	19
1.4.6.5- possibilité a priori des documents	20
1.5- Conclusion	21

CHAPITRE 2 LES TECHNIQUE DU "RELEVANCE FEEDBACK", ETAT DE L'ART

2.1- Introduction	23
2.1.1- Motivation de la technique du RF.....	23
2.2- Principes	24

2.3- Les principales techniques du RF	25
a- La technique du RF semi-automatique	25
b- La technique de RF automatique	25
2.4- RF et modèles de recherche	25
2.4.1- La procédure de RF	26
2.4.2- Les techniques de RF basées sur le modèle vectoriel.....	27
2.4.2.1- Le modèle Rocchio.....	27
2.4.2.2- Le modèle Ide	28
2.4.3- Les techniques de RF dans le modèle probabiliste :	29
2.4.4 - Les techniques de RF dans le modèle BNR	31
2.5- Conclusion	35

CHAPITRE 3 CONTRIBUTION ET REALISATION

3.1 Introduction	37
3.2- Motivation	37
3.3- Relevance feedback possibiliste	37
3.3.1- Définition de la fonction F.....	39
3.3.1.1- Formules basées sur la nécessité de termes ($N(T_i / D_j)$).....	39
3.3.1.2- Formules basées sur la possibilité de termes ($\Pi(T_i / D_j)$).....	40
3.3.1.3- Formule basée sur la possibilité et la nécessité	42
3.4- Exemple	42
3.4.1-Tableau de pertinences pour chaque terme.....	43
3.4.2 Choix des nouveaux termes	44
3.5- Conclusion	45

CHAPITRE 4 EXPERIMENTATIONS ET RESULTATS

4.1- Introduction	47
4.2- Collection de test	47
4.3-Evaluation des SRIs	47
4.3.1- Evaluation résiduelle	48

4.3.2-Protocole d'évaluation.....	49
4.4- Expérimentations et résultats.....	49
4.4.1- Apport des cinq formules pour $n=10$ $k=20$	50
4.4.2- Apport des cinq formules pour $n=15$ $k=20$	52
4.4.3- Apport des cinq formules pour $n=20$ $k=20$	53
4.4.4- Discussion de résultats.....	54
4.5 Reformulation Aveugle.....	54
4.6 Conclusion	55
CONCLUSION GENERALE ET PERSPECTIVES	56
REFERENCES.....	58



INTRODUCTION GENERALE

Une quantité toujours croissante d'information électronique (Internet, CD ROM...) est disponible et mise à la disposition du grand public. Le domaine de la recherche d'information (RI) aspire à fournir des méthodes rapides et efficaces de représentation, de gestion et de présentation de ces informations.

Dans ce contexte, les systèmes de recherche d'information (SRI) sont conçus pour chercher et restituer des documents pertinents à des besoins utilisateurs exprimés au moyen d'une requête. Cependant, il est difficile de choisir les termes adéquats à l'expression des besoins. La complexité de verbaliser un besoin en information peut augmenter lorsque le besoin est vague, lorsque la collection d'informations est peu familière à l'utilisateur ou lorsqu'il est inexpérimenté avec les SRIs. Par contre, il est plus facile pour un utilisateur de savoir si les documents restitués répondent à ses besoins c'est-à-dire qu'il peut évaluer la pertinence des documents.

Les techniques de reformulation de requête tiennent compte de ce dernier point. En effet, ces techniques tentent de construire automatiquement une représentation de requête en se basant sur les documents jugés pertinents par l'utilisateur. La reformulation de requête est un cycle d'activité : le système restitue, suite à une requête, des documents que l'utilisateur juge. Cette information (jugement) est utilisée par le système pour produire une nouvelle requête et restituer d'autres documents. Cette activité peut être itérative.

La reformulation de requête a fait ses preuves pour améliorer l'efficacité des SRIs pour certains types de recherche.

Notre but dans le cadre de ce mémoire est de proposer une nouvelle méthode de reformulation de requête dans le modèle des réseaux possibilistes pour la recherche d'information. Ce modèle proposé en 2005 par Asma BRINI [BRI 05c], donne une nouvelle approche pour modéliser la pertinence de documents étant donnée une requête. En effet la pertinence est mesurée selon deux dimensions : la pertinence possible et la pertinence nécessaire. Notre objectif est de

proposer un processus de reformulation de requête permettant de manipuler de manière efficace ces deux notions.

Nous avons structuré ce mémoire de la manière suivante : nous commencerons dans le premier chapitre par présenter la problématique de la recherche d'informations ainsi que les différents modèles proposés dans la littérature.

Le deuxième chapitre est consacré aux méthodes classiques de reformulation de requête dans les différents modèles proposées.

Le troisième chapitre présente la méthode de reformulation que nous avons proposée.

Enfin le quatrième chapitre présente les principaux résultats obtenus.

En conclusion, nous présenterons le bilan de notre travail.

CHAPITRE 1

1

PRINCIPE DE LA RECHERCHE D'INFORMATION

1.1- Introduction

Les systèmes de recherche d'information (**SRI**), servent d'interface entre une collection contenant des quantités considérables de documents et des utilisateurs cherchant des informations susceptibles de se trouver dans cette collection, en utilisant des requêtes. Ces systèmes intègrent un ensemble de techniques pouvant être résumées en quatre fonctions, qui sont :

- Stockage des informations
- Organisation de ces informations (processus d'indexation)
- Recherche d'informations : en réponse à des requêtes utilisateurs
- Restitution des informations pertinentes pour ces requêtes.

1.2- Notions de base

Un système de recherche d'information (**SRI**) est un système qui permet de retrouver dans une collection de documents ceux qui sont pertinents à une requête d'utilisateur.

Un SRI intègre un ensemble de modèles et de processus permettant de représenter, d'organiser, de questionner ce volume d'informations et de restituer l'information qui correspond le mieux au besoin de l'utilisateur, exprimé via une requête.

1.3- Concepts clés de la recherche d'information (RI):

Le processus de recherche d'information pertinente que le **SRI** est sensé restituer à un utilisateur, consiste en la mise en correspondance des représentations des informations contenues dans un fond documentaire et des besoins de cet utilisateur exprimés par une requête.

Dans la définition de **SRI**, il y a trois notions clés: documents, requête, pertinence.

- **Document:** Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur. Dans notre cas, nous traitons seulement des documents textuels.

- **Requête ou besoin:** Une requête exprime une interprétation du besoin d'information d'un utilisateur. Elle est en général de la forme suivante: "Trouvez les documents qui ...".

- **Pertinence:** Le but de la RI est de trouver seulement les documents pertinents. La notion de pertinence est très complexe. De façon générale, dans document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est sur cette notion de pertinence que le système doit juger si un document doit être donné à l'utilisateur comme réponse.

Cette notion de pertinence peut être appréhendée à deux niveaux :

- **Le niveau utilisateur :** A ce niveau, l'utilisateur a un besoin d'information dans sa tête, et il espère obtenir les documents pertinents pour répondre à ce besoin. La relation entre le besoin d'information et les documents attendus est la relation de pertinence (idéale, absolue, ...).
- **-Le niveau système :** A ce niveau, le système répond à la requête formulée par l'utilisateur par un ensemble de documents trouvés dans la base de documents qu'il possède.

Remarquez que la requête formulée par l'utilisateur n'est qu'une description partielle de son besoin d'information. Beaucoup d'études ont montré qu'il est très difficile, voire impossible, de formuler une requête qui décrit complètement et précisément un besoin d'information. Du côté de document, il y a aussi un changement entre les deux niveaux: les documents que l'on peut retrouver sont seulement les documents inclus dans la collection de documents. On ne peut souvent pas trouver des documents parfaitement pertinents à un besoin. Il arrive souvent qu'aucun document pertinent n'existe dans la collection.

1.4 - Principaux modèles de recherche d'information :

Différents modèles de recherche d'information ont été proposés. Le présent paragraphe a pour objectif d'en présenter les principaux.

1.4.1- Le modèle booléen

Pour ce type de modèle, un document est représenté comme un ensemble de termes, et une requête comme une expression logique de termes. La correspondance entre une requête et un document (notée par $RSV(D, q)$) est déterminée de la façon suivante:

$$RSV(D, t_i) = \begin{cases} 1 & \text{si } t_i \in D \\ 0 & \text{sinon} \end{cases}$$

$$RSV(D, q_1 \sqcap q_2) = \begin{cases} 1 & \text{si } RSV(D, q_1) = 1 \text{ et } RSV(D, q_2) = 1 \\ 0 & \text{sinon} \end{cases}$$

$$RSV(D, q_1 \sqcup q_2) = \begin{cases} 1 & \text{si } RSV(D, q_1) = 1 \text{ ou } RSV(D, q_2) = 1 \\ 0 & \text{sinon} \end{cases}$$

$$RSV(D, \neg q_1) = \begin{cases} 1 & \text{si } RSV(D, q_1) = 0 \\ 0 & \text{sinon} \end{cases}$$

Les avantages du modèle booléen :

- Le modèle est plus facile à implémenter et nécessite relativement peu de ressources [SALT90].
- Le langage de requête booléen est plus expressif que celui des autres modèles [CROF87].
- Ce modèle convient aux utilisateurs sachant exactement leurs besoins et en mesure de les formuler précisément avec le vocabulaire qu'ils maîtrisent parfaitement.

Les inconvénients du modèle booléen

- Il est difficile aux novices de formuler une requête combinant plusieurs opérateurs logiques, notamment pour les questions complexes. L'importance relative des mots clés ne peut pas être exprimée.
- Le classement des documents extraits par ordre de pertinence est difficile.
- La reformulation automatique des requêtes par la technique du RF est plus ardue.

1.4.2-Modèle Booléen basé sur des ensembles flous

Cette extension au modèle booléen standard vise à tenir compte de la pondération des termes dans les documents. Du côté requête, elle reste toujours une expression booléenne classique. Avec cette extension, un document est représenté comme un ensemble de termes pondérés comme suit:

$D_j = \{ \dots, (t_i, a_i), \dots \}$ où a_i est le degré d'appartenance du terme t_i au document D_j .

L'évaluation d'une requête peut prendre plusieurs formes. Une d'elles est la suivante:

$$RSV(D, t_i) = a_i$$

$$RSV(D, q_1 \square q_2) = \min(RSV(D, q_1), RSV(D, q_2)).$$

$$RSV(D, q_1 \sqcup q_2) = \max(RSV(D, q_1), RSV(D, q_2)).$$

$$RSV(D, \neg q_1) = 1 - RSV(D, q_1).$$

Dans cette évaluation, les opérateurs logiques \square et \sqcup sont évalués par min et max respectivement. C'est une des évaluations classiques proposées par L. Zadeh [Zadeh 65] dans le cadre des ensembles flous. Cependant, cette évaluation n'est pas parfaite. Par exemple, on n'a pas $RSV(D, q \square \neg q) \equiv 0$ et $RSV(D, q \sqcup \neg q) \equiv 1$. Du point de vue théorique, c'est gênant. Du point de vue pratique, quand on évalue une requête en forme de conjonction, on ne s'intéresse qu'à la partie la plus difficile, et quand on évalue une requête en forme de disjonction, c'est la partie la plus facile qui domine.

Intuitivement, on aimerait plutôt que les deux parties jouent toutes les deux un rôle dans l'évaluation. Ainsi, beaucoup d'autres formes d'évaluation ont été proposées. Une des formes est l'évaluation Lukaswicz [Loiseau 04] qui est la suivante:

$$RSV(D, t_i) = a_i$$

$$RSV(D, q_1 \square q_2) = RSV(D, q_1) * RSV(D, q_2).$$

$$RSV(D, q_1 \sqcup q_2) = RSV(D, q_1) + RSV(D, q_2) - RSV(D, q_1) * RSV(D, q_2).$$

$$RSV(D, \neg q_1) = 1 - RSV(D, q_1).$$

Dans cette évaluation, on voit que les deux parties d'une conjonction ou d'une disjonction contribuent en même temps, contrairement à celle de Zadeh. Cependant, elle a le même problème qui est $RSV(D, q \square \neg q) \neq 0$ et $RSV(D, q \sqcup \neg q) \neq 1$. En plus, $RSV(D, q \square q) \neq RSV(D, q)$ et $RSV(D, q \sqcup q) \neq RSV(D, q)$.

Si on compare ces extensions avec le modèle standard, il est assez facile de voir les avantages. Le plus important est qu'on peut mesurer le degré de correspondance entre un document et une requête dans $[0, 1]$. Ainsi, on peut ordonner les documents dans l'ordre décroissant de leur correspondance avec la requête. L'utilisateur peut parcourir cette liste ordonnée et décider où s'arrêter. Au niveau de la représentation, on a également une représentation plus raffinée. On peut exprimer dans quelle mesure un terme est important dans un document.

Ces évaluations ont été proposées à la fin des années 1970 et au début des années 1980. Maintenant, ces extensions sont devenues standard: la plupart des systèmes booléens utilisent un des ces modèles étendus.

1.4.3-Le modèle vectoriel (Vector Space Model)

Après le modèle booléen, le modèle qui a le plus influencé la recherche d'information est le modèle vectoriel qui a été créé au début des années 1970 par Gérard Salton et son équipe [SAL 71], [SAL 83] dans le système de recherche d'information SMART.

Dans ce type de modèle, les requêtes et les documents sont considérés de la même façon et représentés sous forme de vecteurs. Le processus de recherche utilise le calcul de distances entre ces vecteurs.

Formellement, dans un modèle vectoriel, on suppose que le poids $w_{d_{ij}}$ (resp. $w_{q_{ik}}$) associé au terme t_i dans le document D_j (resp. la requête Q_k) est positif. Les documents et requêtes sont des vecteurs dans un espace vectoriel de dimension N et définis comme suit :

$$D_j = (w_{d_{1j}}, w_{d_{2j}}, \dots, w_{d_{Nj}})$$

$$Q_k = (w_{q_{1k}}, w_{q_{2k}}, \dots, w_{q_{Nk}})$$

Le modèle vectoriel estime le degré de pertinence entre un document et la requête par un degré de corrélation entre leurs vecteurs associés. Cette corrélation peut être spécifiée par le calcul de similarité entre vecteurs, et qui peut être exprimée par le produit scalaire suivant :

$$Sim(D_j, Q_k) = \sum_{i=1}^N (w_{d_{ij}} * w_{q_{ik}})$$

Plusieurs fonctions de similarité ont été proposées dans la littérature. En voici quelques-unes des fonctions les plus répandues : les mesures de Cosinus, Jaccard et Dice.

Mesure de cosinus :

$$Sim(D_j, Q_k) = \frac{\sum_{i=1}^N (w_{d_{ij}} * w_{q_{ik}})}{\sqrt{\sum_{i=1}^N w_{d_{ij}}^2 * \sum_{i=1}^N w_{q_{ik}}^2}}$$

Mesure de Jaccard :

$$Sim(D_j, Q_k) = \frac{\sum_{i=1}^N (wd_{ij} * wq_{ik})}{\sum_{i=1}^N wd_{ij}^2 + \sum_{i=1}^N wq_{ik}^2 - \sum_{i=1}^N (wd_{ij} * wq_{ik})}$$

Mesure de Dice :

$$Sim(D_j, Q_k) = 2 * \frac{\sum_{i=1}^N (wd_{ij} * wq_{ik})}{\sum_{i=1}^N (wd_{ij}^2 + wq_{ik}^2)}$$

Les avantages du modèle vectoriel :

Il est possible d'assigner une pondération aux termes d'une requête.

- Le coefficient de similarité permet de :
 - de classer les documents par ordre de pertinence,
 - de déterminer le degré de similarité requête - document, document - document, phrase - phrase, etc..

Certains résultats de recherche tendent à prouver que les systèmes de recherche vectoriels sont plus performants que les systèmes de recherche booléens [TURT94].

Les inconvénients du modèle vectoriel :

Part des hypothèses suivantes :

- Les termes sont indépendants; ce n'est pas toujours le cas,
- Requêtes et documents sont essentiellement similaires alors que certains résultats produits par le calcul de similarité requête - document ne reflètent pas la réalité.

1.4.4- Le modèle probabiliste :

La théorie des probabilités est utilisée comme un moyen de modéliser le processus d'extraction de l'information. Dans les systèmes de recherche d'informations conventionnels, les documents sont extraits en réponse à une requête quand l'ensemble des termes clés d'un document

s'apparente dans une certaine mesure aux termes d'une requête. Dans de tels cas, les documents sont dits pertinents par rapport à cette requête.

1.4.4.1- Principe

Soit R et NR représentent respectivement la pertinence et la non-pertinence (ou de façon équivalente, l'ensemble de document pertinents et l'ensemble non-pertinent).

L'idée de base dans un modèle probabiliste est de tenter de déterminer les probabilités $P(R|D)$ et $P(NR|D)$ pour une requête donnée. Ces deux probabilités signifient respectivement : si on retrouve le document D, quelle est la probabilité qu'on obtient l'information pertinente et non-pertinente.

Dans un premier temps, travaillons dans le contexte suivant :

On ne considère que la présence et l'absence de termes dans les documents et dans les requêtes comme des caractéristiques observables. Autrement dit, les termes ne sont pas pondérés, mais prennent seulement les valeurs 0 (absent) ou 1 (présent).

On suppose qu'on a une requête fixe. On tente de déterminer les caractéristiques de R et NR pour cette requête donnée.

Donc, implicitement, $P(R|D)$ et $P(NR|D)$ correspondent plutôt à $P(R_Q|D)$ et $P(NR_Q|D)$ pour la requête Q, mais cet index peut être ignoré pour l'instant.

Si on peut calculer ces deux probabilités, alors on pourra classer les documents selon ces deux probabilités, ou selon la fonction (odd) suivant qui compare les deux probabilités :

$$O(D) = P(R|D) / P(NR|D)$$

Plus $O(D)$ est élevée pour un document, plus ce document doit être classé en haut.

Cependant, les deux probabilités nécessaires ne sont pas directement calculables. Ainsi, on utilise le théorème de Bayes:

$$P(R|D) = P(D|R) P(R) / P(D)$$

$$P(NR|D) = P(D|NR) P(NR) / P(D)$$

Où

$P(D|R)$ = la probabilité que D fait partie de l'ensemble pertinent,

$P(R)$ = la probabilité de pertinence, c'est-à-dire, si on choisit un document au hasard dans le corpus, la chance de tomber sur un document pertinent ;

$P(D)$ = la probabilité que le document soit choisi (si on prend au hasard un document dans le corpus, la chance de tomber sur D).

Appliquons dans $O(D)$, nous avons :

$$O(D) = P(R | D) / P(NR | D) = [P(D | R) P(R)] / [P(D | NR) P(NR)]$$

Comme pour la même requête, $P(R)$ et $P(NR)$ sont des constantes. Ainsi, nous pouvons ré-exprimer $O(D)$ comme suit :

$$O(D) \propto P(D|R) / P(D|NR)$$

($O(D)$ est proportionnelle à $P(D|R) / P(D|NR)$).

Étant donné que l'objectif de la RI est de déterminer le rang des documents, on peut très bien utiliser $P(D|R) / P(D|NR)$ à la place de $O(D)$ exacte. Donc, définissons $O(D)$ comme $P(D|R) / P(D|NR)$.

1.4.4.2- Hypothèse d'indépendance et le modèle de recherche indépendant

Comment estimer $P(D | R)$ et $P(D | NR)$? En général, on décompose le document en un ensemble d' "événements". Un événement dénote soit la présence ou l'absence d'un terme dans ce document, c'est-à-dire une série de $(t_i = x_i)$ où x_i est 0 ou 1 qui représentent l'absence et la présence du terme. Ainsi :

$$P(D | R) = P(t_1 = x_1, t_2 = x_2, t_3 = x_3, \dots | D)$$

$$P(D | NR) = P(t_1 = x_1, t_2 = x_2, t_3 = x_3, \dots | NR)$$

où $t_i = x_i$ correspond à la présence ou l'absence du terme t_i dans le document D .

Dans la théorie de probabilité, on sait que la probabilité de la combinaison de plusieurs événements ensemble doit être déterminée comme suit :

$$P(a, b, c, d \dots | R) = P(a | R) * P(b | a, R) * P(c | a, b, R) * P(d | a, b, c, R) * \dots$$

C'est-à-dire qu'il faut tenir compte des dépendances entre les événements, représentées dans cette formule par des probabilités conditionnelles. Il est vrai que dans le contexte de RI, les présences et les absences de termes sont dépendants. Par exemple, si le terme « informatique »

apparaît dans un document, il y a plus de chance que le terme « ordinateur » apparaisse aussi. Ainsi, nous avons :

$$P(\text{ordinateur} = 1 | \text{informatique} = 1) > P(\text{ordinateur} = 1)$$

Seulement, si on doit tenir compte de toutes les dépendances, le calcul de $P(D|R)$ et de $P(D|NR)$ sera très complexe, car il faut tenir compte des dépendances suivantes :

$$P(t_2 = x_2 | t_1 = x_1, R), P(t_3 = x_3 | t_1 = x_1, t_2 = x_2, R), \text{ etc.}$$

Si on veut estimer ces probabilités, on aura besoin d'un grand ensemble de documents jugés pertinents pour l'entraînement, ce qui n'est pas disponible. Ainsi, l'hypothèse d'indépendance est supposée pour simplifier le calcul :

Hypothèse d'indépendance: on suppose que les événements liés à différents termes sont indépendants.

Ainsi:

$$P(D|R) = \prod_{(t_i=x_i) \in D} P(t_i = x_i | R),$$

et

$$P(D|NR) = \prod_{(t_i=x_i) \in D} P(t_i = x_i | NR).$$

Avec

$$P(t_i | R) = \frac{r_i}{R}$$

(r_i : nombre de documents pertinents contenant t_i et R : nombre de documents pertinents total)

et

$$P(t_i | NR) = \frac{m_i - r_i}{M - R}$$

(m_i : nombre de documents non pertinents contenant t_i et M : nombre de documents de la collection)

Les avantages du modèle probabiliste :

Selon Savoy [SAV094], le modèle de recherche probabiliste est plus efficace que le modèle de recherche booléen, mais **moins** performant que le modèle de recherche vectoriel

Les inconvénients du modèle probabiliste :

Il n'existe pas de méthode d'estimation de la pertinence des termes avant toute extraction de document pertinent. Cette estimation se fait à posteriori.

1.4.5- Le modèle BNR (modèle RI basé sur les réseaux Bayésiens)

Les modèles probabilistes constituent un outil puissant pour les modèles de RI, car ils permettent de traiter d'une manière efficace l'incertitude intrinsèque au processus de RI. Plus récemment, des travaux ont essayé d'exploiter l'apport des Réseaux Bayésiens (RBs) pour définir des modèles de RI. L'avantage apporté par l'utilisation des réseaux a été principalement de pouvoir combiner des informations provenant de différentes sources pour restituer les documents qui seraient les plus pertinents étant donnée une requête.

1.4.5.1-Architecture générale du modèle

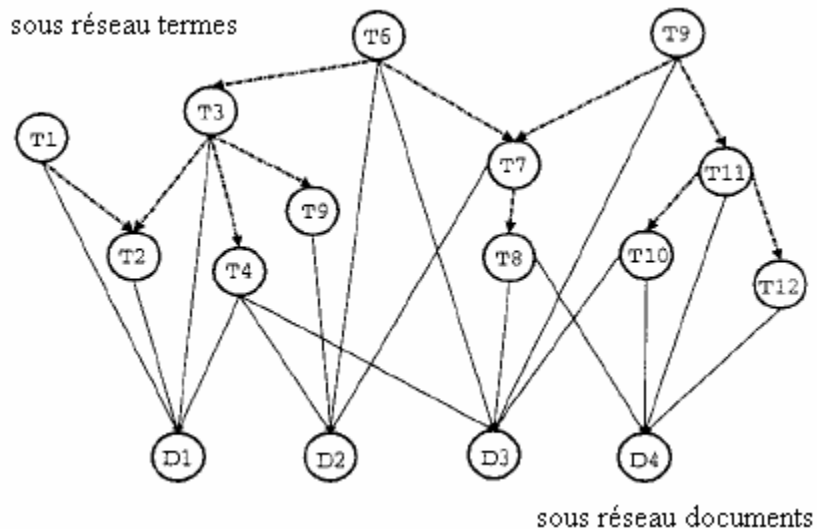


Figure 1.5.5.1.1- Structure du réseau dans un modèle BNR

Les nœuds du réseau dans un modèle BNR ont décomposés en deux ensembles de variables T et D :

L'ensemble des termes de la collection $T = (T_1, T_2, \dots, T_M)$, où M est le nombre de termes dans la collection

L'ensemble des documents de la collection $D = (D_1, D_2, \dots, D_N)$, où N est le nombre de documents dans la collection

Les domaines des nœuds sont binaires {vrai, faux} signifiant que le nœud est instancié ou non. T est l'ensemble des nœuds termes; une variable T_i associée à un terme prend ses valeurs dans le domaine $\text{dom}(T_i) = \{t_i, \bar{t}_i\}$, \bar{t}_i désigne que le terme T_i est non pertinent et t_i désigne qu'il est pertinent. Un terme est considéré pertinent si tous les documents qui contiennent sont jugés pertinent par l'utilisateur et non pertinent si non.

D est l'ensemble des nœuds documents, une variable D_j prend ses valeurs dans le domaine le domaine $\text{dom}(D_j) = \{d_j, \bar{d}_j\}$, où \bar{d}_j signifie « le document D_j n'est pas pertinent » et d_j signifie « le document D_j est pertinent ». Un document est pertinent s'il répond au besoin utilisateur.

1.4.5.2- Estimation des distributions de probabilité

a) Estimation des distributions de probabilité

- **Nœud terme racine:** les probabilités a priori de ces nœuds sont données par

$$P(t_i) = \frac{1}{M} \text{ et } P(\bar{t}_i) = 1 - \frac{1}{M}$$

où M est le nombre de termes dans la collection

-**Nœud terme qui a des parents:** Pour les termes ayant des parents qui sont eux aussi des nœuds termes, la quantification des arcs les reliant est calculée par l'indice de Jaccard : la similarité entre deux ensembles de termes est donnée par le ratio entre le nombre d'éléments de l'intersection et l'union de ces deux ensembles.

-**Nœud Document :** le nombre de probabilité conditionnelle que nous avons besoin de calculer augmente exponentiellement par rapport aux nombre de parents ce qui pose un problème car si par exemple un document est indexé par n termes alors il faut calculer 2^n probabilités. Pour résoudre ce problème on considère θ_{D_j} l'ensemble des configurations possibles des parents de D_j .

Une configuration est non pertinente lorsque les instanciations des variables qu'elle contient ne sont pas conformes à la présence des termes dans le document.

Pour calculer la probabilité, on peut utiliser une fonction de probabilité basée sur la mesure de cosinus [Salton 83]:

Pour chacun de configuration θ_{D_j} , la probabilité qu'un document soit pertinent est donnée par:

$$P(D_j | \theta_{D_j}) = \alpha_j \sum_{\substack{T_i \in D_j \\ t_i \in \theta_{D_j}}} tf_{ij} idf_i^2$$

où tf_{ij} correspond au nombre d'occurrences du terme t_i dans le document D_j et idf_i est fréquence inverse dans le document. et α_j est donné par:

$$\alpha_j = \frac{1}{\alpha \sqrt{\sum_{T_i \in D_j} tf_{ij} idf_i^2}}$$

et α est une constante normalisée.

b) Calcul de la pertinence:

Les termes présents dans la requête propagent l'information à travers le réseau pour calculer la pertinence d'un document étant donnée la requête.

La probabilité $P(d_j/Q)$ de la pertinence d'un document étant donnée la requête est donné par:

$$P(d_j | Q) = \alpha_j \sum_{T_i \in D_j} tf_{ij} idf_i^2 P(t_i | Q)$$

on a $P(t_i | Q) = 1 \forall T_i \in Q$ alors l'équation devient:

$$P(d_j | Q) = \alpha_j \left(\sum_{T_i \in D_j \cap Q} tf_{ij} idf_i^2 + \sum_{T_i \in D_j \setminus Q} tf_{ij} idf_i^2 + P(t_i | Q) \right)$$

1.4.6- Le modèle possibiliste

Ce modèle est proposé dans [BRI 05], il est basé sur les réseaux possibilistes. La théorie des possibilités propose des mesures duales permettant le traitement de l'information incertaine.

La pertinence possible d'un document évalue le degré auquel un document peut être éliminé de la liste des réponses (documents restitués). La pertinence nécessaire mesure la certitude liée à la pertinence du document.

1.4.6.1-Architecture générale du modèle

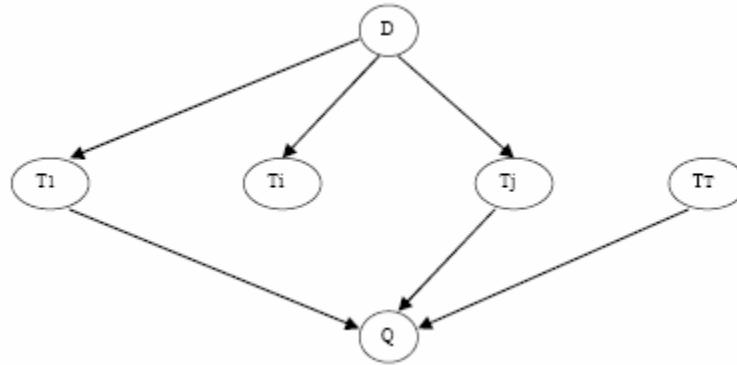


Figure 1.5.6.1.1- Architecture générale du modèle

La topologie du réseau est représentée dans la figure ci-dessus. D'un point de vue qualitatif, le graphe permet de représenter les nœuds documents, requête, termes d'indexation et permet d'exprimer les relations de dépendance existant entre ces nœuds. Un document (D_j) est instancié ou non, prenant ses valeurs dans le domaine $\text{dom}(D_j) = \{d_j, \bar{d}_j\}$. L'activation (ou instanciation) d'un nœud document $D_j = d_j$ (resp. \bar{d}_j) signifie que le document est pertinent ou non étant donnée une requête. Une requête, Q , prend ses valeurs dans le domaine $\text{dom}(Q) = \{q, \bar{q}\}$. Dans ce modèle est intéressé par l'instanciation de la requête, et a considéré que le cas $Q = q$, et on le note Q . Le domaine d'un nœud terme d'indexation, T_i , est $\text{dom}(T_i) = \{t_i, \bar{t}_i\}$. ($T_i = t_i$) signifie que le terme t_i est présent dans l'objet (document ou requête) et donc représentatif de l'objet. Un terme non-représentatif, \bar{t}_i , est un terme absent de la représentation de l'objet. Soit $T(D_j)$ (resp. $T(Q)$) l'ensemble des termes indexant le document D_j (resp. la requête Q).

1.4.6.2- Evaluation de la requête

L'évaluation de la requête dans ce modèle est effectuée par la propagation de l'information de la requête à travers le réseau dans ce modèle le processus de propagation est similaire à la propagation probabiliste bayésienne. Le processus d'évaluation consiste à propager l'information injectée par la requête.

Dans ce modèle, la pertinence est calculée par une double mesure:

- La **pertinence nécessaire**.
- La **pertinence possible**.

La **pertinence nécessaire** mesure à quel point un document doit faire partie de la liste des documents restitués sachant que la **pertinence possible** mesure à quel point un document constitue éventuellement une réponse à une requête donnée.

Ce modèle devrait être capable d'inférer des propositions de type :

- Il est plausible à un certain degré que le document est pertinent étant donnée la requête, $\Pi(dj / Q)$;
- Il est aussi certain (dans le sens possibiliste) que le document est pertinent à la requête, $N(dj/Q)$.

Le processus de propagation évalue les degrés de possibilité, $\Pi(dj / Q)$, et de nécessité, $N(dj/Q)$, par :

$$\Pi(dj / Q) = \frac{\Pi(Q \wedge dj)}{\Pi(Q)}$$

$$N(dj/Q) = 1 - \Pi(\bar{d}j / Q)$$

$$\text{avec } \Pi(\bar{d}j / Q) = \frac{\Pi(Q \wedge \bar{d}j)}{\Pi(Q)}$$

la possibilité de Q est :

$$\Pi(Q) = \max(\Pi(Q \wedge dj), \Pi(Q \wedge \bar{d}j))$$

Pour calculer $\Pi(dj / Q)$ il faut définir $\Pi(Q \wedge dj)$. Etant donnée la topologie du graphe, elle est de la forme :

$$\Pi(Q \wedge dj) = \max_{\forall \theta^l \in \theta} (\Pi(Q | \theta^l) \cdot \prod_{T_i \in T(D_j) \wedge T(Q)} \Pi(\theta_i^l | D_j) \cdot \Pi(D_j) \cdot \prod_{T_k \in T(Q) \setminus T(D_j)} \Pi(\theta_k^l))$$

avec :

θ : Les configurations possibles de l'ensemble des parents de Q,

θ_i^l : L'instanciation de T_i dans la configuration θ^l ;

θ^l : Une configuration possible de θ .

$T_i \in T(Q) \wedge T(D_j)$: Les termes de la requête qui indexent les documents, ces termes sont évalués dans le contexte de leurs parents par $\Pi(T_i | D_j)$.

$T_i \in T(Q) \setminus T(D_j)$: Les termes de la requête absents des documents pour ces termes une possibilité marginale est calculée, $\Pi(T_k)$

A l'issue du processus de propagation, Les documents sont restitués par ordre décroissant de pertinence nécessaire puis de pertinence possible.

Les documents qui ont une valeur de nécessité supérieure à 0 sont classés en premier et les documents possiblement pertinents sont classés après les documents nécessaires ou se retrouvent en haut de la liste lorsque le système ne trouve pas de documents nécessairement pertinents.

1.4.6.3 Agrégation des termes de la requête

La possibilité de la requête étant donnée les termes d'indexation, $\Pi(Q|\theta)$, dépend de l'interprétation de la requête. Plusieurs interprétations sont possibles. Les termes de la requête peuvent être connectés par une **conjonction**, une **disjonction**, ou par une **somme probabiliste**, ou encore une **somme probabiliste pondérée**.

- **conjonction** : Pour une requête booléenne, **ET**, le processus d'évaluation restitue les documents contenant tous les termes de la requête.

La possibilité de la requête Q étant donnée une configuration possible, θ^l , de θ de tous ses parents est donnée par :

$$\Pi(Q|\theta^l) = \begin{cases} 1 \text{ si } \forall T_i \in PAR_Q, \theta_i^l = \theta_i^Q \\ 0 \text{ si non} \end{cases}$$

-**Disjonction** : Pour une requête booléenne, **OU**, le document est plus ou moins pertinent s'il contient au moins un terme d'indexation de la requête.

La possibilité de la requête Q étant donnée une configuration possible, θ^l , de θ de tous ses parents est donnée par :

$$\Pi(Q|\theta^l) = \begin{cases} 1 \text{ si } \exists T_i \in PAR_Q, \theta_i^l = \theta_i^Q \\ 0 \text{ si non} \end{cases}$$

-Négation : La requête peut contenir la négation d'un terme, signifiant que l'utilisateur ne veut pas voir ce terme dans le document restitué. Lorsque le document contient ce terme alors la pertinence est nulle. La négation d'un terme est une opération unaire. Ainsi:

$$\Pi(Q | \theta^l) = \begin{cases} 1 & \text{si } \theta_i^l = \bar{t}_i \\ 0 & \text{si non} \end{cases}$$

1.4.6.4- Pondération des termes d'indexation

Pour évaluer la pertinence plausible et la pertinence nécessaire d'un document étant donnée une requête, il faut définir les arcs du réseau. Un arc reliant un nœud terme à un nœud document quantifie à quel point le terme est représentatif de ce document une absence d'arcs absence d'arc entre terme et document traduit l'absence du terme en question du document.

La pondération des arcs du réseau reliant les nœuds termes aux nœuds documents est donnée par les quatre formule suivantes :

$$\text{a) } \Pi(t_i | d_j) = ntf_{ij} \quad \text{Avec} \quad ntf_{ij} = \frac{tf_{ij}}{\max_{t_k \in d_j} (tf_{kj})}$$

$$\text{b) } \Pi(\bar{t}_i | d_j) = 1$$

$$\text{c) } \Pi(t_i | \bar{d}_j) :$$

Un terme discriminant dans la collection est un terme qui apparaît dans peu de documents de la collection. Dans ce modèle un terme discriminant est un terme qui est nécessairement représentatif d'un document et donc contribue certainement à le sélectionner parmi d'autres documents.

Un degré de nécessaire pertinence, ϕ_{ij} , d'un terme t_i pour représenter un document d_j comme un poids est définie par :

$$\phi_{ij} = \mu_1\left(\frac{N}{n_i}\right) * \mu_2(ntf_{ij})$$

où * : opérateur produit ;

μ_1, μ_2 : Fonctions de normalisation.

Par exemple si on prend μ_1 comme fonction logarithmique et μ_2 comme fonction d'identité alors:

$$\phi_{ij} = \frac{\text{Log}\left(\frac{N}{n_i}\right)}{\text{Log}(N)} * ntf_{ij}$$

Ce degré de nécessaire pertinence montre la nécessité qu'un terme implique un document et donc aide à restituer ce document :

$$N(t_i \rightarrow d_j) = \phi_{ij}$$

On a $\Pi(t_i | \bar{d}_j) = \frac{\Pi(\bar{d}_j \wedge t_i)}{\Pi(\bar{d}_j)}$ et puisque la possibilité a priori $\Pi(\bar{d}_j) = 1$ alors:

$$\Pi(t_i | \bar{d}_j) = \Pi(\bar{d}_j \wedge t_i) = 1 - N(t_i \rightarrow d_j) = 1 - \phi_{ij}$$

$$d) \Pi(\bar{t}_i | \bar{d}_j) = 1$$

Enfin le tableau ci-dessous résume les possibilités conditionnelles des termes d'indexation étant donnée l'instanciation du nœud document parent.

	d_j	\bar{d}_j
t_i	ntf_{ij}	$1 - \phi_{ij}$
\bar{t}_i	1	1

Tableau 1.5.6.4.1- Table de possibilités conditionnelles $\Pi(T_i | D_j)$

1.4.6.5- possibilité a priori des documents

En absence d'information sur les documents, la possibilité a priori d'un document est uniforme c.-à-d.: $\Pi(d_j) = \Pi(\bar{d}_j) = 1$.

Si on utilise des informations comme l'importance des termes et la longueur des documents etc.... alors la possibilité a priori d'un document peut calculer comme suit :

$$\Pi(d_j) = \frac{l_j}{\max_{k=1, \dots, N} l_k}$$

Avec:

l_j : est la longueur du document d_j en terme de fréquence et $l_j = \sum tf_{ij}$

Plus les documents sont courts plus leur pertinences diminuent.

De plus $\prod(\bar{d}_j) = 1$.

1.5- Conclusion

Nous avons commencé dans ce chapitre par positionner le contexte de la RI en donnant ses concepts fondamentaux ainsi que le fonctionnement global de tout SRI. Nous avons aussi détaillé les modèles les plus connus (booléen, vectoriel et les modèles probabiliste) de la RI et nous avons décrit plus en détail le modèle possibiliste car la reformulation que nous proposons est basée sur ce modèle.

CHAPITRE 2

2

LES TECHNIQUES DU "RELEVANCE FEEDBACK"

ETAT DE L'ART

2.1- Introduction

Le RF (Relevance Feedback) ou technique de modification des requêtes par analyse et incorporation des retours, est un processus de reformulation automatique de requêtes dont le but est de générer des requêtes optimales proches des besoins des utilisateurs. Cette reformulation qui se fait par interaction entre l'utilisateur et le système consiste en générale à modifier la pondération des termes de la requête initiale ou à leur substituer d'autres termes choisis pour leur caractère, notamment associatif, générique ou spécifique. Ces opérations de reformulation s'effectuent sur la base des indices fournis par l'utilisateur à travers, d'une part, la requête initiale et, d'autre part, les documents pertinents et non pertinents sélectionnés. Ce processus de recherche, de sélection de documents pertinents et non pertinents puis de génération automatique de requête se fait de façon itérative jusqu'à l'atteinte des objectifs à la satisfaction de l'utilisateur.

2.1.1- Motivation de la technique du RF

La technique du RF est utilisée en vue d'améliorer le repérage de documents, notamment: Pour les utilisateurs non avertis et peu familiers avec les techniques de formulation de requête en général et d'une interface de système de recherche en particulier. De tels utilisateurs auront à leur disposition un outil à même de les aider à reformuler efficacement leurs requêtes. Pour ceux-ci, la première requête soumise au système de recherche aura valeur de test.

quand l'utilisateur, même chevronné, a une connaissance approximative ou très limitée du domaine d'application et des caractéristiques de la collection qui fait l'objet de sa recherche. Dans de tels cas, l'utilisateur n'est pas toujours en mesure de fournir les termes appropriés lui permettant de spécifier de façon adéquate sa requête. La technique du RF supplée à sa connaissance limitée du domaine en enrichissant requête de termes susceptibles de l'améliorer substantiellement. Cela peut être utile dans les domaines en constante mutation où l'emploi de nouveaux termes argotiques est courant.

dans les environnements de recherche où le taux de rappel est critique tel que le domaine médical ou encore dans le cadre de l'exécution d'arrêts judiciaires où il est souvent nécessaire d'explorer tout indice potentiel.

L'emploi de RF procure des avantages réels [SALT90] :

Le RF libère l'utilisateur des contraintes liées à la reformulation de requête dans un environnement plus ou moins connu,

Permet de mener des opérations de recherche par étapes successives, favorisant ainsi une approche graduelle et rationnelle du champ d'intérêt.

Procure un outil approprié de recherche permettant de pondérer les termes suivant leur importance relative et d'adapter les recherches aux caractéristiques de La collection sur laquelle elle s'applique.

2.2- Principes

L'application de la technique du RF part du principe que tous les documents pertinents dépistés par un moteur de recherche suite à une requête ont des caractéristiques communes. Dans un modèle de recherche à base d'espace vectoriel, cela signifie que les vecteurs représentant les documents extraits et le vecteur de requête ont une similarité notable. Cela implique, par voie de conséquence, que les vecteurs de termes des documents non pertinents par rapport à la requête et les vecteurs de termes des documents pertinents sont dissemblables. Ces considérations ont conduit à la conception d'approches de reformulation automatique de requête. En effet, le premier prototype implémentant la technique du RF était basé sur le modèle de recherche vectorielle et consistait en un ensemble de termes, éventuellement pondérés, utilisés sans opérateurs booléens. Ainsi, les requêtes étaient exprimées sous forme de vecteurs :

$$Q_0 = (r_0, r_1, r_2, \dots, r_t)$$

où Q_0 est la requête initiale; r_i représente la pondération du terme i pouvant prendre les valeurs 0 ou 1 suivant l'absence ou la présence de ce terme dans la requête. Les termes de la requête initiale peuvent être des mots choisis dans un thesaurus, un dictionnaire de termes contrôlés, ou choisis librement parmi les mots du domaine.

À partir des documents pertinents produits par la requête Q_0 , le processus du RF génère automatiquement une nouvelle requête :

$$Q' = (r'_0, r'_1, r'_2, \dots, r'_t)$$

où r'_i représente la pondération du terme i de la première requête Q_0 modifiée dans la requête Q' .

2.3- Les principales techniques du RF.

Carol [CAR0971] distingue deux techniques principales du RF : la technique semi-automatique basée sur le modèle Rocchio et la technique automatique.

a- La technique du RF semi-automatique

La technique semi-automatique nécessite l'intervention de l'utilisateur qui doit identifier et sélectionner les documents pertinents et les documents non pertinents. Les travaux sur cette technique ont été menés par Rocchio à la fin des années 1970. Ces travaux ont été publiés en 1971 [ROCC 71] et ont été suivis de ceux de Ide [IDE 71]. Plus tard, les travaux sur le RF semi-automatique ont été enrichis par l'apport de la méthode probabiliste. Cette approche a été implémentée par Harper, Hamian, Croft, Spark Jones et Van Rijisbergen [CARO 97].

b- La technique de RF automatique

Selon Carol [CARO 97], dans les environnements où la technique du RF automatique est implémentée, un nombre prédéfini de documents extraits par la requête initiale sont réputés pertinents. Les procédures et formules utilisées dans l'approche du RF automatique sont des variantes des formules Rocchio et Ide qui permettent de faire abstraction des documents non pertinents.

2.4- RF et modèles de recherche

La technique du RF peut être appliquée aux principaux modèles de recherche que sont le modèle vectoriel, le modèle booléen et le modèle probabiliste. Nous étudions ici le modèle vectoriel qui est le plus connu et l'un des plus utilisés à ce jour

2.4.1- La procédure de RF

La procédure du RF peut se résumer dans la figure ci-dessous :

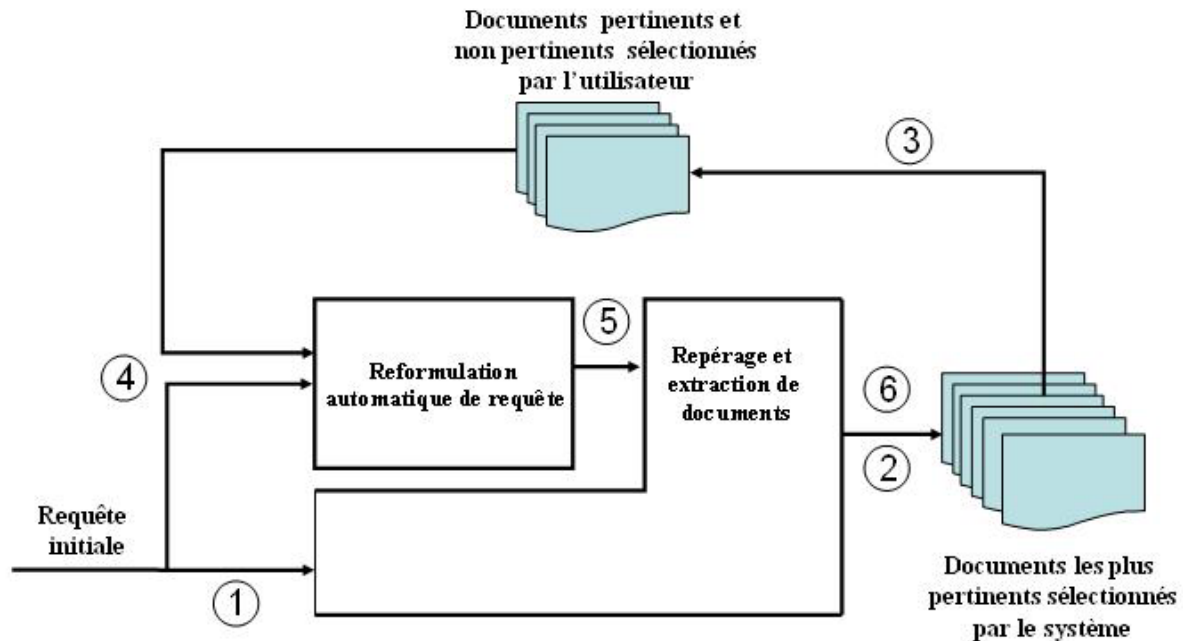


Figure 2.4.1 Processus de fonctionnement du RF

- 1 - Soumission d'une requête initiale formulée par l'utilisateur au système de repérage et d'extraction de documents,
- 2 - Traitement de la requête initiale par Le système de repérage,
- 3 - Sélection par l'utilisateur documents pertinents et non pertinents et cela, à partir de son point de vue,
- 4 - Composition automatique d'une nouvelle requête à partir des données issues de la requête initiale et des informations fournies par les documents pertinentes et non pertinentes sélectionnées,
- 5 - Traitement avec la nouvelle requête reformulée automatiquement et transmise au moteur de recherche,
- 6 - Génération des documents détectés avec la nouvelle requête reformulée.

2.4.2- Les techniques de RF basées sur le modèle vectoriel

Les techniques du RF appliquées au modèle de recherche vectoriel ont été dominées par les travaux de Rocchio [ROCC 71] puis de Ide [IDE 71]. Comme son nom l'indique, le RF dans le cadre de ce modèle part du principe que la requête initiale formulée par l'utilisateur sert au système à identifier une zone ou région de l'espace d'index de termes qui contient des documents pertinents. N'ayant pas d'autres informations sur les caractéristiques des documents enregistrés, la requête initiale constitue l'unique indice de départ. En introduisant dans le cycle la requête initiale et les documents pertinents et non pertinents courants sélectionnés, l'utilisateur, ce faisant, fournit des informations au système qui lui permettent de reformuler automatiquement le profil de la requête de sorte que les documents générés au fur et à mesure des itérations tendent de plus en plus à se rapprocher des besoins de l'utilisateur. Nous présenterons dans un premier temps le modèle Rocchio, puis le modèle Ide, une variante de modèle Rocchio et enfin le RF incrémental qui allie les deux premiers modèles.

2.4.2.1- Le modèle Rocchio

Le modèle de Rocchio se résume comme suit [ROCC 71] : Soit une série d'opérations de recherche et d'extraction d'informations initiées par une requête initiale Q_0 qui est par la suite modifiée en fonction des sorties produites par le système (utilisant Q_0 comme entrée). Soit Q' la requête modifiée obtenue et la plus proche de la requête optimale de l'utilisateur.

En tenant compte du fait que l'utilisateur a la possibilité de sélectionner parmi les documents extraits celles qui sont pertinentes et celles qui ne le sont pas, on peut assimiler cette action à un signal envoyé par l'utilisateur au système pour lui signifier un renforcement positif ou négatif sur la réponse obtenue. Sur la base de ces indices et de la requête initiale, il est possible de construire itérativement une requête modifiée de plus en plus optimale dont les résultats tendent à refléter les besoins de l'utilisateur.

L'efficacité de ce processus dépendra de la qualité de la requête initiale et du degré de convergence des itérations successives vers une requête optimale. Cette convergence est toujours exprimée par rapport au choix de l'usager.

La formule de base Rocchio est de la forme :

$$Q_1 = Q_0 + \frac{1}{n_1} \sum_{i=1}^{n_1} P_i - \frac{1}{n_2} \sum_{i=1}^{n_2} NP_i$$

avec :

Q_1 : est le vecteur de la nouvelle requête

Q_0 : Est le vecteur de la requête initiale

P_i : Vecteur de documents pertinent restitués et évalués

NP_i : Vecteur de documents non-pertinent restitués et évalués

n_1 : est le nombre de documents pertinent restitués et évalués

n_2 : est le nombre de documents non-pertinent restitués et évalués

Rocchio a par la suite développé ce modèle pour le formaliser par ce qu'il est convenu d'appeler le Standard Rocchio [ROCC7 1].

Standard Rocchio :

$$Q_1 = Q_0 + \beta \sum_{i=1}^{n_1} \frac{P_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{NP_i}{n_2}$$

Cette formule appelée "Standard Rocchio" introduit deux nouveaux paramètres, β et γ qui sont fixés arbitrairement et qui permettent de pondérer globalement la valeur moyenne des vecteurs des termes des documents pertinents et non pertinents.

À partir de cette formule, Rocchio a introduit des contraintes supplémentaires telles que la pondération des termes du vecteur original, ou une limite fixée au nombre de documents pertinents à considérer. Il a par la suite testé des variantes de sa formule de base sur des collections de tailles relativement réduites en ajoutant d'autres contraintes. L'une des contraintes étant de n'autoriser dans Q_1 que les termes préalablement présents dans Q_0 ou les termes présents au moins dans la moitié des documents pertinents, si ce nombre représentant cette moitié est supérieur au nombre de documents non pertinents contenant ces termes.

En positionnant les paramètres β et γ à 0.75 et 0.25 respectivement, la formule standard Rocchio a donné les meilleurs résultats en limitant l'effet produit par l'incorporation des documents non pertinents ou "feedback négatif" [SALT90]. Ainsi ces valeurs ci-dessus ont permis d'atteindre une amélioration de la précision de 19% à 156% suivant les collections avec une moyenne de 70%.

2.4.2.2- Le modèle Ide

Le modèle Ide de l'auteur du même nom est une variante du modèle de Rocchio [IDE 71]. Du modèle de Rocchio elle déduit la formule suivante qui lui sert de base à ses travaux :

$$Q_{i+1} = \pi Q_i + \omega Q_0 + \alpha \sum_1^{\min(na, n'p)} p_i + \mu \sum_1^{\min(nb, n's)} NP_i$$

où $n'_p + n'_s = N$ le nombre de document extraits et servant au processus du " feedback ". Les variables expérimentales étant : $\alpha, \omega, \mu, \pi, n_a, n_b$ et N .

Le paramètre α est positif et permet de pondérer tous les documents jugés pertinents par rapport à tous les éléments contribuant à la formation de ta requête (requête précédente, requête initiale R0 et documents non pertinents).

Le paramètre π permet d'augmenter la pondération de la requête précédente en fonction des documents du feedback. R0 est la requête initiale, Ri est la requête de la précédente itération, ω permet d'utiliser la requête initiale comme partie intégrante de la nouvelle requête, μ doit être théoriquement négatif pour tenir compte des documents non pertinents extraits. Les paramètres n_a, n_b permettent d'utiliser un nombre spécifique de documents pertinents et non pertinents dans la requête même quand les valeurs des paramètres n_a, n_b sont plus grands (utilisation de la fonction $\min()$).

La flexibilité de cette formule a permis à Ide non seulement de confirmer les résultats positifs obtenus par Rocchio, mais aussi d'étudier trois variantes de ce modèle [IDE71]:

Modèle basé sur l'utilisation exclusive de documents pertinents.

Modèle basé sur le nombre de documents N à extraire et à réintégrer dans le système à chaque itération du RF.

Modèle basé sur l'intégration d'un ou de deux documents non pertinents aux documents pertinents et à la requête initiale.

2.4.3- Les techniques de RF dans le modèle probabiliste :

Robertson et Sparck-Jones [ROBE 76] ont développé une formule de pondération des termes basée sur la distribution des termes de la requête dans les documents jugés pertinents et les documents jugés non-pertinents par l'utilisateur :

$$w_i = \log \frac{\frac{r_i}{R - r_i}}{\frac{n_i - r_i}{(N - n_i) - (R - r_i)}}$$

Avec:

w_i : poids d'un terme t_i de la requête;

r_i : nombre de documents pertinents contenant le terme t_i ;

R : nombre de documents pertinents pour la requête;

n_i : nombre de documents contenant le terme t_i ;

N : nombre de documents dans la collection;

Une variation de cette formule de base a été définie dans le but de calculer les nouveaux poids pour les termes de la nouvelle requête lors du processus de réinjection de pertinence [ROBE 76]:

$$w_i = \log \frac{p_i(1-u_i)}{u_i(1-p_i)}$$

p_i : probabilité ($d_i = 1/D$ est pertinent) = $(r_i + 0.5)/(R + 1)$;

u_i : probabilité ($d_i = 0/D$ est pertinent) = $(n_i - r_i + 0.5)/(N - R + 1)$;

$D = (d_1, d_2, \dots, d_n)$, $d_i = 1$ si le terme t_i indexe le document D , $d_i = 0$ sinon;

0,5 est un facteur d'ajustement

Harman [HARM 92] a montré que l'utilisation de la formule de 0,5 est un facteur d'ajustement permet une augmentation de la précision de l'ordre de 25% sur la base Cranfield.

Croft [CROFT 83] a défini une méthodologie de repondération en utilisant une version révisée de la formule de pondération de Sparck-Jones :

Recherche initiale: $w_{ijk} = (C + \text{idf}_i) \cdot f_{ik}$

$$\text{Feedback: } w_{ijk} = \left[C + \log \frac{p_{ij}(1 - q_{ij})}{q_{ij}(1 - p_{ij})} \right] \cdot f_{ik}$$

Avec :

w_{ijk} : le poids du terme t_i dans la requête j et le document k .

idf_i : fréquence absolue du terme t_i dans la collection,

p_{ij} : probabilité que le terme t_i soit assigné à un ensemble de documents pertinents pour une requête j ,

$$p_{ij} = \frac{r + 0,5}{r + 1} \text{ si } r > 0, p_{ij} = 0,01 \text{ si } r = 0,$$

q_{ij} : probabilité que le terme t_i apparaisse dans un ensemble de document non pertinent

$$q_{ij} = \frac{n - r + 0,5}{N - R + 1},$$

$$f_{ik} = K + (1 - K) \cdot \frac{\text{freq}_{ik}}{\max(\text{freq}_k)},$$

où :

freq_{ik} : la fréquence du terme t_i dans le document k ,

$\max(\text{freq}_k)$: la fréquence max d'un terme dans le document k ,

C, K : constantes.

2.4.4 - Les techniques de RF dans le modèle BNR

Soit b le nombre de documents jugé par l'utilisateur :

l'ensemble $\{D_{k_1} = d_{k_1}, \dots, D_{k_h} = d_{k_h}\}$ contient les documents pertinents et l'ensemble $\{D_{k_{h+1}} = \bar{d}_{k_{h+1}}, \dots, D_{k_b} = \bar{d}_{k_b}\}$ contient les documents non pertinents alors la nouvelle requête sera :

$$Q1 = Q \wedge [\{D_{k_1} = d_{k_1}, \dots, D_{k_h} = d_{k_h}, D_{k_{h+1}} = \bar{d}_{k_{h+1}}, \dots, D_{k_b} = \bar{d}_{k_b}\}].$$

Chaque nœud X non-instancié reçoit de tous ses nœuds parents un message sous forme de vecteur $\Pi_X(Z)$, il reçoit encore de tous ses nœuds fils Y un message sous forme vecteur $\lambda_Y(X)$.

Chaque nœud instancié X reçoit un message $\lambda_0(X)$ d'un nœud fils imaginaire avec "

$$\left\{ \begin{array}{l} \lambda_0(X) = (1,0) \text{ si } X = \bar{x} \\ \lambda_0(X) = (0,1) \text{ si } X = x \end{array} \right.$$

si l'évidence de X est partielle par rapport à une observation alors :

$$\lambda_0(X) = (P(\text{Obs} | \bar{x}), P(\text{Obs} | x))$$

Dans ce cas le plus important est le rapport $\frac{P(\text{Obs} | \bar{x})}{P(\text{Obs} | x)}$ et on peut dire que

$\lambda_0(X) = (P(\text{Obs} | \bar{x}), P(\text{Obs} | x))$ et $\lambda_0(X) = \left(\frac{P(\text{Obs} | \bar{x})}{P(\text{Obs} | x)}, 1\right)$ sont équivalentes et en plus, le rapport

Pour que tous les nœuds reçoivent λ_0 , on utilise le vecteur $\lambda_0(X) = (1,1)$ pour les nœuds non instancié.

Définition de quelques notations utilisées dans la suite:

R_Q : l'ensemble des documents restitués et évalués pour une requête Q

$|R_Q|$: cardinale de R_Q

n_r : Nombre de documents pertinents

$n_{\bar{r}}$: Nombre de documents non-pertinents

n_{t_i} : Nombre de documents restitués qui contiennent t_i

$n_{\bar{t}_i}$: Nombre de documents restitués qui ne contiennent pas t_i

n_{rt_i} : Nombre de documents pertinent qui contiennent t_i

$n_{\bar{r}\bar{t}_i}$: Nombre de documents non-pertinent qui contiennent t_i

$n_{r\bar{t}_i}$: Nombre de documents pertinent qui ne contiennent pas t_i

$n_{\bar{r}\bar{t}_i}$: Nombre de documents non-pertinent ne contiennent pas t_i

Les notions définies ci-dessus sont résumées dans le tableau ci-dessous suivant :

	$T_i = \bar{t}_i$	$T_i = t_i$	Total
Non-pertinent	$n_{\bar{r}\bar{t}_i}$	$n_{\bar{r}t_i}$	$n_{\bar{r}}$
Pertinent	$n_{r\bar{t}_i}$	n_{rt_i}	n_r
	$n_{\bar{t}_i}$	n_{t_i}	$ R_Q $

Tableau 2.4.4 Table de contingence des termes

On classe les termes indexant les documents restitués en trois catégories :

- Terme qui se trouve dans des documents pertinents seulement (termes positifs \mathfrak{S}^+).
- Terme qui se trouve dans des documents non-pertinents seulement (termes négatifs \mathfrak{S}^-).
- Terme qui se trouve dans des documents pertinents et non-pertinents (termes neutres \mathfrak{S}^\pm).

Il faut distinguer entre les termes indexant les documents trouvés et la requête et entre les autres qui indexent les documents trouvés et absents de la requête.

On note \mathcal{S}^q l'ensemble de termes indexant les documents trouvés et la requête et on va l'utiliser pour ré pondérer les termes de la requête initiale et on note \mathcal{S}^e l'ensemble des termes indexant les documents trouvés et absents de la requête et ce dernier ensemble est utilisé pour représenter les termes a ajouter (Expansion de la requête).

a) Repondération de termes de la requête initiale Q

Les termes de l'ensemble \mathcal{S}^q qui étaient instanciés comme pertinents, reçoivent un message $\lambda_0(T) = (0,1)$. Les termes de la requête initiale qui ocurrent seulement dans des documents non-pertinents ne sont pas considérés. Par conséquence, ils devraient être pénalisés en diminuant leur pertinence. Les autres termes ($\lambda_0(T) = (1,1)$) sont considérés comme des termes n'appartenant pas à la requête (nonquery term) et il est plus valable d'utiliser le vecteur $\lambda_0(T_i) = (\gamma_{t_i}, 1)$ à la place de $\lambda_0(T_i) = (1,1)$ avec $0 \leq \gamma_{t_i} \leq 1$.

la méthode proposée par [CAM 03] considère que γ_{t_i} est très sensible au nombre de documents non pertinents contenant T_i ($n_{\bar{r}t_i}$ définie ci dessus) et a montré que la meilleur valeur de γ_{t_i} est celle qui tend vers le vecteur $\lambda_0(T) = (1,1)$ et a proposé une fonction qui satisfait cette condition avec $0.5 \leq \gamma_{t_i} \leq 1$:

$$\lambda_0(T_i) = \left(1 - \frac{1}{n_{\bar{r}t_i} + 1}, 1\right)$$

D'autre part, les termes $\in \mathcal{S}^q \cap \mathcal{S}^+$ et ceux qui $\in \mathcal{S}^q \cap \mathcal{S}^\pm$ sont les plus important mais en principe on ne peut pas augmenter la pertinence de termes positifs ou neutres qui ocurrent dans la requête initiale car ils sont déjà complètement pertinents. Ainsi, la première approche simple qui s'appelle *tr-ins* et qui traite ce genre de termes propose que chacun de ces termes reçoive le message $\lambda_0(T) = (0,1)$.

Une autre approche est proposée pour augmenter la pertinence de ce genre des termes qui s'appelle *tr-rep*. Cette approche est basée sur la duplication de ces nœuds termes dans le réseau. Le nombre de duplication de chacun de ces nœuds est égal au nombre de document pertinents contenant ce terme (n_{rt_i}) et après il faut instancier les nœuds dupliqués comme pertinents.

Pour changer la structure du réseau, il suffit de connecter les nœuds termes dupliqués comme fils de T_i et connecter les nœuds documents fils de T_i comme fils des nœuds dupliqués.

La figure suivante montre une duplications de trois fois la terme T_i :

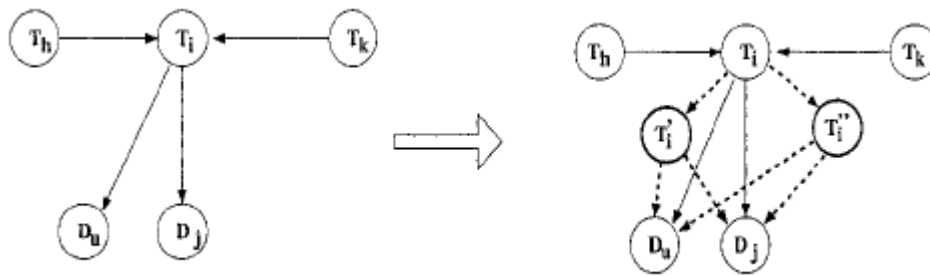


Figure 2.4.4.1 Duplication trois fois du terme T_i

Nous n'avons pas besoin de changé la structure du réseau. Le changement se fait virtuellement et il suffit de multiplier par n_{r_i} le facteur qui calcule le poids du terme dans la formule de probabilité générale et ceci pour chaque terme $\in \mathcal{S}^q \cup \mathcal{S}^+ \cup \mathcal{S}^-$. La pertinence de ces terme va augmenter n_{r_i} fois automatiquement.

b) Expansion de la requête (Ajouter des nouveaux termes sur la requête initiale):

Comme nous avons vu dans le paragraphe précédent il y a trois classe de termes (\mathcal{S}^+ , \mathcal{S}^- , \mathcal{S}^\pm). Puisque \mathcal{S}^e est l'ensemble de termes duquel nous pouvons choisir, les nouveaux ensembles des termes seront :

- Les termes négatifs qui $\in \mathcal{S}^e \cap \mathcal{S}^-$
- Les termes neutres qui $\in \mathcal{S}^e \cap \mathcal{S}^\pm$
- Les termes positifs qui $\in \mathcal{S}^e \cap \mathcal{S}^+$

Tous les termes négatifs sont instanciés comme non-pertinents et ont reçu le vecteur $\lambda_0(T_i) = (1,0)$. Les termes neutres ont reçu le vecteur $\lambda_0(T_i) = (1,1)$.

En générale on peut designer la probabilité qu'un terme soit pertinent ou non par $(p(r|t_i)$ respectivement $p(r|\bar{t}_i)$). Ces deux probabilité sont utilisés pour calculer la vecteur $\lambda_0(T_i)$ et dans ce cas:

$$\lambda_0(T_i) = (p(r | \bar{t}_i), p(r | t_i)) \text{ Où } \lambda_0(T_i) = \left(\frac{p(r | \bar{t}_i)}{p(r | t_i)}, 1 \right)$$

On a plusieurs méthodes pour calculer ces probabilités nous citons un ci-dessous :

-)Méthode qe-gmle:

$$P(r | t_i) = \frac{n_{rt_i}}{n_{t_i}} \text{ et } P(r | \bar{t}_i) = \frac{n_{r\bar{t}_i}}{n_{\bar{t}_i}}$$

2.5- Conclusion

L'application de la technique de RF, moyen efficace d'amélioration des performances du système de repérage d'information, permet d'augmenter substantiellement le niveau de précision par rapport à la requête initiale. Si plusieurs paramètres doivent être considérés pour une utilisation optimale des méthodes, il n'en demeure pas moins que les différentes variantes de la technique ont donné des résultats largement positifs. Le RF incrémental de par son interface utilisateur conviviale et sa formule unifiée et simplifiée devrait permettre de la vulgariser. Aujourd'hui, l'avantage procuré par cette technique est tel que plusieurs moteurs de recherche Web l'intègrent à leur mécanisme de recherche. L'impact direct est l'augmentation des requêtes soumises aux moteurs. Cette charge accrue sera d'autant plus limitée qu'il y aura convergence des résultats obtenus vers une satisfaction plus complète des usagers.

CHAPITRE 3

3

CONTRIBUTIONS ET REALISATIONS

3.1 Introduction

Nous présentons dans ce chapitre notre contribution au développement des systèmes de recherche d'information (SRI) dans le but de proposer un modèle de recherche d'information plus flexible. Notre contribution se traduit par la mise en œuvre d'un module permettant la reformulation de requête par réinjection de pertinence (Relevance Feedback) dans le modèle possibiliste. Ce modèle modélise la pertinence en définissant deux aspects de la pertinence d'un document étant donnée une requête : la pertinence nécessaire et la pertinence possible. Les documents préférés et restitués en haut de liste sont les documents nécessairement pertinents. Alors le but est de profiter de la double valeur de pertinence et des jugements (binaires ou graduels utilisateur) pour reformuler la requête.

3.2- Motivation

Les documents restitués par un SRI sont le résultat de comparaison de la représentation des documents à celle de la requête. Le problème est que cette correspondance ne donne pas toujours de « bons » résultats, les documents restitués ne satisfont pas toujours les besoins des utilisateurs.

La reformulation de requête par réinjection de pertinence (Relevance Feedback) qui est une combinaison de l'indexation des documents et du jugement utilisateur a donc été proposée pour pallier à ce problème

La problématique à laquelle nous nous intéressons concerne la reformulation de requêtes par réinjection de pertinence possibiliste. Particulièrement, nous voulons profiter des informations concernant les termes, qui sont fournies par le modèle possibiliste de point de vue pertinence (Possible et nécessaire), pour trouver les meilleurs termes d'indexés dans les documents jugés pertinents par l'utilisateur pour pouvoir reconstruire une nouvelle requête.

3.3- Relevance feedback possibiliste

Le but d'un SRI est de construire un système capable de chercher dans la collection, des documents en réponse à une requête utilisateur. Une recherche efficace dépend de facteurs

subjectifs : la capacité de l'utilisateur à exprimer son besoin d'information au moyen d'une requête et des caractéristiques du SRI. Nous avons proposé un processus de reformulation de requête par réinjection de pertinence (Relevance Feedback) pour le modèle possibiliste en intégrant la notion de possibilité et de nécessité en se basant sur la formule de Rocchio. Comme nous avons vu dans le deuxième chapitre, la formule de base de Rocchio est la suivante :

$$Q_1 = Q_0 + \frac{1}{n_1} \sum_{i=1}^{n_1} P_i - \frac{1}{n_2} \sum_{i=1}^{n_2} NP_i$$

En se basant sur cette formule, nous proposons d'y intégrer la possibilité et la nécessité de termes. Ceci nécessite un changement dans la formule.

Notre formule proposée est de la forme suivant :

$$Q_1 = \alpha Q_0 + \beta F(P) - \gamma F(NP)$$

Avec

Q_1 : est le vecteur de la nouvelle requête

Q_0 : Est le vecteur de la requête initiale

P : Liste de documents pertinents restitués et évalués

NP : Liste de documents non-pertinents restitués et évalués

F : Fonction qui combine les pondérations de chaque terme dans la liste des documents pertinents (respectivement Non-pertinents) pour trouver un poids final où à partir de ce poids nous pouvons choisir les meilleurs termes.

α : Paramètre positif permet de pondérer les termes de la requête initiale

β : Paramètre positif permet de pondérer les termes des documents jugés pertinents par rapport aux documents non pertinents

γ : Paramètre positif permet de pondérer les termes des documents jugés non-pertinent

3.3.1- Définition de la fonction F

Dans la liste de documents restitués (pertinent ou non-pertinent), un terme peut exister dans plusieurs documents, mais son poids possibiliste et nécessaire change d'un document à un autre. Alors, il faut trouver le moyen pour agréger tous les poids d'un même terme dans la liste des documents.

Nous avons proposé cinq formules, deux formules basées sur la nécessité, deux autres basées sur la possibilité et le cinquième est une combinaison des deux. Ces formules ont été définies dans le but de calculer les nouveaux poids pour les termes de la nouvelle requête lors du processus de réinjection de pertinence (Relevance Feedback)

La fonction F est alors une fonction qui applique l'une des cinq formules proposées sur l'une de deux listes de documents et qui trie le résultat final des poids des termes par ordre décroissant et renvoie les n premiers termes.

3.3.1.1- Formules basées sur la nécessité de termes ($N(T_i / D_j)$)

Les deux formules que nous avons proposées et qui sont basées sur la nécessité sont les suivantes :

- **Nécessité moyenne**

- **Nécessité * (r / R)**

a) Nécessité moyenne

Le poids final de chaque terme est donné par la formule :

$$poidsfinal(t_i) = \frac{1}{R} \sum N(t_i / D_j)$$

Avec

- $N(t_i / D_j)$: nécessité de t_i étant donnée D_j

- $D_j = \begin{cases} d_j & \text{Si nous sommes dans la liste de documents pertinents} \\ \bar{d}_j & \text{Si nous sommes dans la liste de documents non - pertinent} \end{cases}$

$$- R = \begin{cases} R_1 & \text{Si } D_j = d_j \\ R_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

R_1, R_2 sont respectivement le nombre de documents pertinents et le nombre de documents non- pertinents

b) Nécessité *(r / R)

Le poids final de chaque terme est donné par la formule :

$$poidsfinal(t_i) = \frac{r}{R} * \sum N(t_i / D_j)$$

Avec

- $N(t_i / D_j)$: nécessité de t_i étant donnée D_j

$$- D_j = \begin{cases} d_j & \text{Si nous sommes dans la liste de documents pertinents} \\ \bar{d}_j & \text{Si nous sommes dans la liste de documents non - pertinents} \end{cases}$$

$$- R = \begin{cases} R_1 & \text{Si } D_j = d_j \\ R_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

R_1, R_2 sont respectivement le nombre de documents pertinents et le nombre de documents non- pertinents

$$- r = \begin{cases} r_1 & \text{Si } D_j = d_j \\ r_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

r_1, r_2 sont respectivement le nombre de documents pertinents contenant le terme t_i et le nombre de documents non- pertinents contenant t_i

3.3.1.2- Formules basées sur la possibilité de termes ($\Pi(T_i / D_j)$)

Les deux formules que nous avons proposées et qui sont basées sur la possibilité sont les suivantes :

- **Possibilité moyenne**

- **possibilité *(r / R)**

a) Possibilité moyenne

Le poids final de chaque terme est donné par la formule :

$$poidsfinal(t_i) = \frac{1}{R} \sum \Pi(t_i / D_j)$$

Avec

- $\Pi(t_i / D_j)$: Possibilité de t_i étant donnée D_j

$$- D_j = \begin{cases} d_j & \text{Si nous sommes dans la liste de documents pertinents} \\ \bar{d}_j & \text{Si nous sommes dans la liste de documents non - pertinents} \end{cases}$$

$$- R = \begin{cases} R_1 & \text{Si } D_j = d_j \\ R_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

R_1, R_2 sont respectivement le nombre de documents pertinents et le nombre de documents non- pertinents

b) Possibilité *(r / R)

Le poids final de chaque terme est donné par la formule :

$$poidsfinal(t_i) = \frac{r}{R} * \sum \Pi(t_i / D_j)$$

Avec

- $\Pi(t_i / D_j)$: possibilité de t_i étant donnée D_j

$$- D_j = \begin{cases} d_j & \text{Si nous sommes dans la liste de documents pertinents} \\ \bar{d}_j & \text{Si nous sommes dans la liste de documents non - pertinents} \end{cases}$$

$$- R = \begin{cases} R_1 & \text{Si } D_j = d_j \\ R_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

R_1, R_2 sont respectivement le nombre de documents pertinent et le nombre de documents non-pertinents

$$- r = \begin{cases} r_1 & \text{Si } D_j = d_j \\ r_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

r_1, r_2 Sont respectivement le nombre de documents pertinents contenant le terme t_i et le nombre de documents non-pertinents contenant t_i .

3.3.1.3- Formule basée sur la possibilité et la nécessité

Cette formule mélange la possibilité et la nécessité d'un terme. La formule est donnée par :

$$poids_{final}(t_i) = \frac{1}{R} \sum \Pi(t_i / D_j) * N(t_i / D_j)$$

Avec

- $\Pi(t_i / D_j)$: Possibilité de t_i étant donnée D_j

- $N(t_i / D_j)$: nécessité de t_i étant donnée D_j

$$- D_j = \begin{cases} d_j & \text{Si nous sommes dans la liste de documents pertinents} \\ \bar{d}_j & \text{Si nous sommes dans la liste de documents non - pertinents} \end{cases}$$

$$- R = \begin{cases} R_1 & \text{Si } D_j = d_j \\ R_2 & \text{Si } D_j = \bar{d}_j \end{cases}$$

R_1, R_2 Sont respectivement le nombre de documents pertinent et le nombre de documents non-pertinents

3.4- Exemple

Soit une collection constituée de sept documents et la représentation de leurs termes d'indexes est :

$$D_1 = \{7t_1, 4t_2, 10t_3, 2t_{20}\} \quad , \quad D_2 = \{10t_2, 6t_4, 8t_5, 3t_6, t_7, 3t_{20}\}$$

$$D_3 = \{5t_6, t_7, 8t_{10}\} \quad , \quad D_4 = \{10t_3, t_6, 8t_8\}$$

$$D_5 = \{7t_4, 4t_5, 6t_9\} \quad , \quad D_6 = \{t_4, 2t_6, 7t_7\}$$

$$D_7 = \{16t_1, t_2, 2t_8\}$$

Considérons à présent que l'utilisateur a jugé D_1, D_2, D_5, D_6 pertinents et D_3, D_4, D_7 non pertinents.

Nous rappelons la méthode de calcul de pertinence d'un terme dans le modèle possibiliste qui est décrite dans le premier chapitre :

- pertinence possible :

$$\Pi(t_i / d_j) = ntf_{ij}$$

Avec $ntf_{ij} = \frac{tf_{ij}}{\max_{t_k \in d_j} (tf_{kj})}$ est la fréquence normalisée de terme t_i dans le document D_j

-pertinence nécessaire :

$$N(t_i / d_j) = \phi_{ij}$$

Avec $\phi_{ij} = \frac{\text{Log}(\frac{N}{n_i})}{\text{Log}(N)} * ntf_{ij}$ où N est le nombre de documents de la collection et n_i le nombre

de documents contenant t_i

Pour simplifier le calcul nous supposons que $\gamma = 0$ et nous travaillons sur les documents pertinents seulement.

3.4.1-Tableau de pertinences pour chaque terme

	t1	t2	t3	t4	t5	t6	t7	t9	t20
ni	2	3	2	3	2	4	3	1	2
nifd	0,6438	0,4354	0,6438	0,4354	0,64379	0,2876	0,4354	1	0,6438
Poss(ti/d1)	0,7	0,4	1						0,2
Nec(ti/d1)	0,4507	0,1742	0,6438						0,1288
Poss(ti/d2)		1		0,6	0,8	0,3	0,1		0,3
Nec(ti/d2)		0,4354		0,2613	0,51503	0,0863	0,0435		0,1931
Poss(ti/d5)				1	0,57143			0,8571	
Nec(ti/d5)				0,4354	0,36788			0,8571	
Poss(ti/d6)				0,1429		0,2857	1		
Nec(ti/d6)				0,0622		0,0822	0,4354		

Tableau 3.4.1.1 possibilité et nécessité des termes

Ce tableau contient les valeurs de possibilité et de nécessité de chaque terme étant donné un document.

Les cases vides signifient une valeur de pertinence nulle. Les termes présentés dans le tableau sont ceux présents dans des documents pertinents les autres non mentionnés appartiennent à des documents non-pertinents.

3.4.2 Choix des nouveaux termes

Nous avons appliqué les cinq formules sur les résultats du tableau 3.4.1.1 les résultats obtenus sont représentés dans le tableau suivant :

	t1	t2	t3	t4	t5	t6	t7	t9	t20
Nécessité* (r / R)	0,1127	0,3048	0,1609	0,5692	0,4414	0,0842	0,2395	0,2143	0,1609
Nécessité moyenne	0,1127	0,1524	0,1609	0,1897	0,2207	0,0421	0,1197	0,2143	0,0805
Nécessité*possibilité	0,0197	0,0533	0,0402	0,0827	0,0756	0,0062	0,0329	0,0459	0,0101
possibilité* (r / R)	0,175	0,7	0,25	1,3071	0,6857	0,2929	0,55	0,2143	0,25
possibilité moyenne	0,175	0,35	0,25	0,4357	0,3428	0,1464	0,275	0,2143	0,125

Tableau 3.4.2.1 poids final de termes

Si par exemple nous voulons prendre les cinq premiers termes :

- Pour la formule Nécessité * (r / R) nous choisissons les termes t4, t5, t2, t7, t9
- Pour la formule Nécessité moyenne nous choisissons les termes t5, t9, t4, t3, t2
- Pour la formule Nécessité * possibilité nous choisissons les termes t4, t5, t2, t9, t3
- Pour la formule Possibilité * (r / R) nous choisissons les termes t4, t2, t5, t7, t6
- Pour la formule Possibilité moyenne nous choisissons les termes t4, t2, t5, t7, t3

On constate à travers cet exemple que les formules proposées ne produisent pas les mêmes requêtes.

Nous avons effectué une série d'expérimentation sur une collection standard de corpus de TREC. Nous décrivons dans le chapitre suivant les résultats obtenues.

3.5- Conclusion

Nous avons décrit dans ce chapitre la méthode de reformulation de requête dans le modèle possibiliste. Nous avons proposé une nouvelle méthode possibiliste basée sur le modèle de reformulation de Rochio du modèle vectoriel. Cinq formules sont proposées pour calculer le poids final de termes d'indexation. Nous avons montré par un exemple que les cinq formules ne produisent pas les mêmes requêtes .

CHAPITRE 4

4

EXPEIRMENTATIONS ET RESULTATS

4.1- Introduction

Les expérimentations que nous décrivons dans ce chapitre ont été effectuées sur une collection standard de RI à savoir la collection de test WT10g.

L'objectif de ces expérimentations est de mesurer les performances et la viabilité de notre approche. Les expérimentations et les évaluations sont articulés autour de la comparaison de notre modèle et le modèle possibiliste de base en RI .

4.2- Collection de test

Pour mener nos expérimentations, nous avons utilisé la collection de pages Web constituée pour la neuvième conférence TREC, corpus nommé WT10g. Cet ensemble comprend 1 692 096 pages Web écrites en anglais pour un volume de 11033 mégaoctets. Nous disposons de 50 requêtes couvrant des domaines variés (par exemple "parkinson's disease", "hunger", "baltimore", "how e-mail benefits businesses", "mexican food culture"). Les requêtes correspondent à des exemples réels soumis au moteur de recherche Excite. Elles comportent en moyenne 2.4 mots (écart type de 0.6). De plus, les termes employés sont très souvent ambigus et très fréquents dans les pages Web disponibles. Afin de simuler un environnement distribué.

4.3-Evaluation des SRIs

L'évaluation des performances d'un SRI, peut être effectuée en mesurant différents paramètres et critères. Cleverdon a dressé une liste de paramètres et critères que l'on peut utiliser pour évaluer la qualité d'un SRI. On distingue parmi ces paramètres :

- Le **rappel** qui détermine la possibilité du SRI à retrouver tous les documents pertinents. Il est mesuré par le rapport entre le nombre de documents pertinents retrouvés et le nombre de documents pertinents contenus dans la base, pour une requête donnée.

- La **précision** détermine la capacité du SRI à ne présenter que les documents pertinents. Elle est égale au rapport entre le nombre de documents pertinents effectivement retrouvés et le nombre de documents sélectionnés en réponse à une requête donnée.

- La **précision exacte** est la précision au point où la précision vaut le rappel. Si la requête admet n documents pertinents, la précision exacte est la précision calculée à partir des n premiers documents de la liste ordonnée des documents restitués.

- La **précision moyenne** est la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée.

Le temps de réponse est égal à l'intervalle qui sépare l'instant auquel on envoie la demande et l'instant auquel le premier résultat s'affiche.

La présentation des résultats permet de mesurer l'ergonomie et la convivialité de l'interface utilisateur.

La mesure de ces paramètres se fait de plusieurs façons. Il faut déjà distinguer les paramètres calculés tels que les taux de rappel et de précision et ceux qui sont déterminés sur la base des observations ou par le biais d'autres techniques telles que les questionnaires et les interviews [SALT 83].

L'évaluation des performances du modèle que nous avons construit sera effectuée essentiellement sur la précision et le rappel.

4.3.1- Evaluation résiduelle

Pour évaluer l'apport de la reformulation de requête par rapport à la non reformulation nous avons utilisé la méthode résiduelle. Cette méthode est adaptée à l'évaluation d'un mécanisme de recherche d'information basé sur la relevance feedback (réinjection de pertinence). La mesure de taux de rappel et précision doit être effectuée avec précaution lorsqu'on évalue l'amélioration induite par l'intégration de la relevance feedback. En effet, le procédé de feedback est tel que tout document initialement restitué en réponse à une requête, paraît à nouveau avec un rang amélioré dans les itérations ultérieures du feedback.

La mesure doit plutôt estimer la capacité de la relevance feedback à restituer de nouveaux documents. L'une des solutions apportées est la méthode de collection résiduelle. Cette méthode préconise de ne pas considérer les documents préalablement jugés pour l'évaluation des résultats de l'itération feedback courante.

Dans ce méthode la mesure de pourcentage de nouveaux documents entre la réponse à la requête initiale et la réponse à la requête étendue, traduit la performance effective due à la relevance feedback.

4.3.2-Protocole d'évaluation

L'évaluation est effectuée selon le protocole TREC. Plus précisément, chaque requête est soumise au système de RI avec les paramètres fixés (facteurs de discrimination, type de pondération, etc.). Le système renvoie les 1000 premiers documents pour chaque requête.

Les valeurs de précision à P5, P10, ..., Pr. Ex, Pr.Moy sont calculées. Le ratio des documents pertinents parmi les 5 premiers documents restitués est la précision au point 5, P5. Certaines évaluations sont aussi mesurées aux points de rappel 0.1, 0.2, ..., 1.0. La définition ainsi que le calcul de ces notions de précision et de rappel ont été définies dans le premier chapitre.

Nous sommes parfois amenés à mesurer les pourcentages de perte ou de gain entre deux variations des variables du modèle. Ce pourcentage est obtenu d'une manière générale pour deux variables A et B mesurant le pourcentage de C par :

$$\%C = \frac{B - A}{A} * 100$$

Nous instancions les valeurs de A, B et C lors de leurs utilisations.

4.4- Expérimentations et résultats

Parmi les 1000 documents restitués par le système, l'utilisateur va juger m document pertinent et m document pertinent. Le jugement est utilisé pour extraire les n premier termes ayant le poids le plus élevé (de point de vue possibiliste) et qui sont introduits dans la requête lors de sa reconstruction. En effet, nous remplaçons la requête initiale avec n termes extraits de k premiers documents restitués et montrés à l'utilisateur parmi les m documents jugés. Le choix de ces variables (k et n) n'est pas fait aléatoirement mais après l'essai de plusieurs combinaisons, celle qui a donné les meilleurs résultats a pour n=10 et k=20. n a pris ses valeurs dans l'ensemble {5, 10, 15, 20} et k a pris ses valeurs dans l'ensemble {5,10,15,20,30}.

Les évaluations que nous avons faites ont essentiellement consisté à comparer la réinjection de pertinence (par les cinq formules proposées) et le modèle possibiliste de base.

Les trois paragraphes ci-dessous montrent l'apport de la reformulation de requêtes par réinjection de pertinence pour les cinq formules que nous avons proposées par rapport au modèle possibiliste de base (sans reformulation) avec les valeurs ($n=10,15,20$ et $k=20$) :

- **Nécessité*(r / R)**
- **Nécessité moyenne**
- **Nécessité * Possibilité**
- **Possibilité * (r / R)**
- **Possibilité moyenne**

En signalant qu'avec $k=20$ et pour tout n dans $\{5, 10, 15, 20\}$, la précision moyenne est amélioré plus que 53 %

4.4.1- Apport des cinq formules pour $n=10$ $k=20$

a- Tableau et figure de précision

Les valeurs de précision du modèle possibiliste de base et les cinq formules proposées pour $n=10$ et $k=20$ sont montrées dans le tableau et la figure ci-dessous

	P5	P10	P15	P20	P30	P100	P1000	Pr.Ex	Pr.Moy
Initial	0,184	0,154	0,1387	0,134	0,1253	0,0876	0,027	0,0932	0,0623
Nécessité* (r / R)	0,3526	0,3079	0,2789	0,2579	0,2219	0,1297	0,036	0,1904	0,1379
Nécessité moyenne	0,3389	0,2556	0,2259	0,2069	0,1769	0,1156	0,0338	0,1706	0,125
Nécessité*possibilité	0,3026	0,2462	0,202	0,1923	0,1718	0,1064	0,0291	0,1528	0,1143
Possibilité* (r / R)	0,3026	0,2513	0,212	0,1872	0,1684	0,099	0,0272	0,1567	0,1135
Possibilité moyenne	0,3026	0,2436	0,1983	0,1859	0,1692	0,0959	0,027	0,1473	0,113

Tableau 4.4.1.1 Valeurs de précision

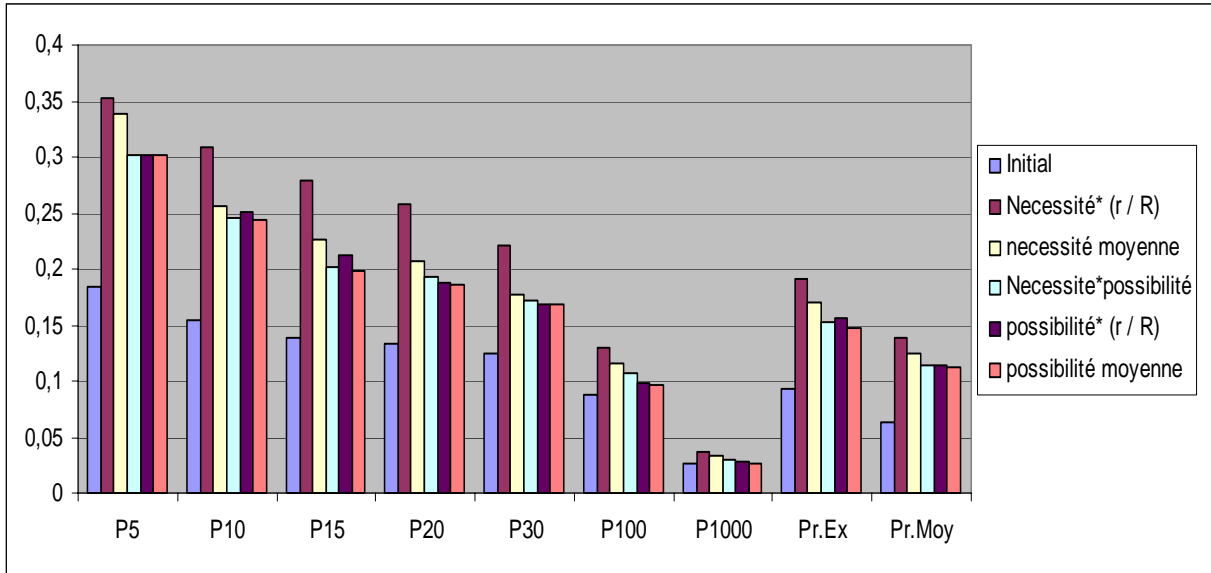


Figure 4.4.1.1 Apport de la réinjection de pertinence

b- Tableau de pourcentage d’amélioration pour P5, P10 et la précision moyenne

	P5	P10	Pr.Moy
Nécessité* (r / R)	91,6304	99,9351	121,348
Nécessité moyenne	84,1848	65,974	100,642
Nécessite*possibilité	64,4565	59,8701	83,4671
Possibilité* (r / R)	64,4565	63,1818	82,183
Possibilité moyenne	64,4565	58,1818	81,3804

Tableau 4.4.1.2 Pourcentage d’amélioration

L’amélioration apportée par la reformulation est nette sur les différents précisions 5, 10 , ... etc. comme l’indiquent le tableau 4.4.1.2 et la figure 4.4.1.1 ci-dessus et en plus la précision moyenne est améliorée plus que 81 % pour les cinq formules.

La formule **Nécessité* (r / R)** augmente la précision moyenne de la modèle possibiliste de base de 121 % ce qui montre l’importance des termes nécessaires (pertinence nécessaire) pour un document.

4.4.2- Apport des cinq formules pour n=15 k=20

a- Tableau et figure de précision

Les valeurs de précision du modèle possibiliste de base et les cinq formules proposées pour n=15 et k=20 sont montrés dans le tableau et la figure ci-dessous

	P5	P10	P15	P20	P30	P100	P1000	Pr.Ex	Pr.Moy
initial	0,184	0,154	0,1387	0,134	0,1253	0,0876	0,027	0,0932	0,0623
Nécessité* (r / R)	0,3947	0,3447	0,2807	0,2605	0,2193	0,1229	0,0322	0,1771	0,1327
Nécessité moyenne	0,35	0,2944	0,2389	0,2125	0,1898	0,1153	0,0348	0,1613	0,1231
Possibilité* (r / R)	0,3538	0,2667	0,2222	0,2090	0,1744	0,0992	0,0272	0,1660	0,1227
Nécessite*possibilité	0,3358	0,2718	0,2444	0,2192	0,1812	0,1123	0,0303	0,1626	0,1174
Possibilité moyenne	0,3282	0,2744	0,2171	0,2013	0,1735	0,0918	0,0246	0,1577	0,1166

Tableau 4.4.1.1 Valeurs de précision

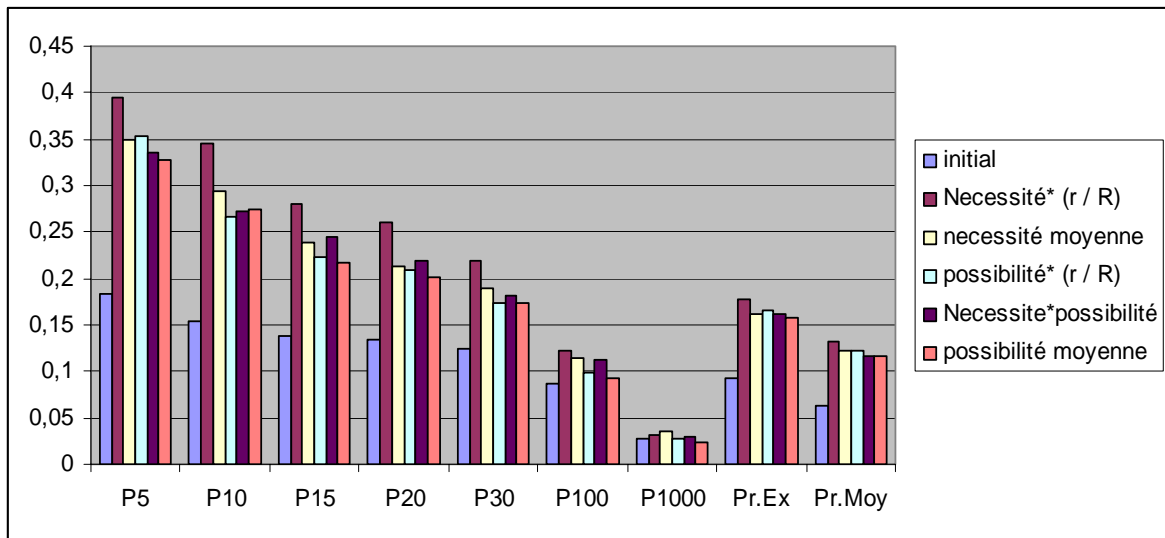


Figure 4.4.2.1 Apport de la réinjection de pertinence

b- Tableau de pourcentage d'amélioration pour P5,P10 et la précision moyenne

	P5	P10	Pr.Moy
Nécessité* (r / R)	114,511	123,831	113,002
Nécessité moyenne	90,2174	91,1688	97,5923
Possibilité* (r / R)	92,2826	73,1818	96,9502
Nécessite*possibilité	82,5	76,4935	88,443
Possibilité moyenne	78,3696	78,1818	87,1589

Tableau 4.4.2.2 Pourcentage d'amélioration

Pour $n=15$ et $k=20$, nous remarquons encore une amélioration nette sur les différents points de précision 5, 10, ... etc. comme l'indiquent le tableau 4.4.2.2 et la figure 4.4.2.1 ci-dessus et en plus la précision moyenne est améliorée plus que 87 % pour les cinq formules.

Nous remarquons en plus que l'amélioration des trois dernières formules est supérieur a celle du premier cas ($n=10, k=20$) mais elle est diminuée pour les deux premières formules.

Dans ce cas nous trouvons encore que la formule **Nécessité*** (r / R) est la meilleure parmi les autres et elle donne une amélioration 113%.

4.4.3- Apport des cinq formules pour $n=20$ $k=20$

a- Tableau et figure de précision

Les valeurs de précision du modèle possibiliste de base et les cinq formules proposées pour $n=20$ et $k=20$ sont montrés dans le tableau et la figure ci-dessous

	P5	P10	P15	P20	P30	P100	P1000	Pr.Ex	Pr.Moy
initial	0,184	0,154	0,1387	0,134	0,1253	0,0876	0,027	0,0932	0,0623
Nécessité* (r / R)	0,4105	0,3211	0,2737	0,2553	0,2096	0,1171	0,0288	0,1747	0,1259
possibilité* (r / R)	0,3795	0,2718	0,2376	0,2090	0,1761	0,0962	0,0245	0,1525	0,1093
possibilité moyenne	0,359	0,2769	0,2222	0,2038	0,1769	0,0956	0,0248	0,1450	0,1062
nécessité moyenne	0,3278	0,2528	0,2111	0,1861	0,1583	0,0956	0,0296	0,1387	0,1028
Nécessité*possibilité	0,3179	0,2308	0,1863	0,1782	0,1521	0,0882	0,0262	0,1318	0,0954

Tableau 4.4.3.1 Valeurs des précisions

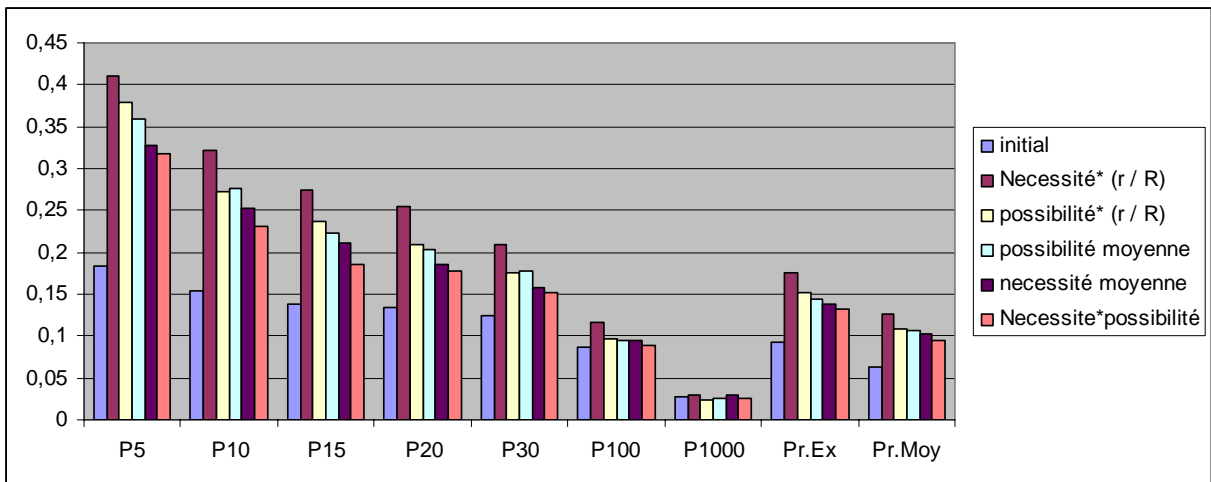


Figure 4.43.1 Apport de la réinjection de pertinence

b- Tableau de pourcentage d'amélioration pour P5, P10 et la précision moyenne

	P5	P10	Pr.Moy
Nécessité* (r / R)	123,098	108,506	102,087
Possibilité* (r / R)	106,25	76,4935	75,4414
Possibilité moyenne	95,1087	79,8052	70,4655
Nécessité moyenne	78,1522	64,1558	65,008
Nécessite*possibilité	72,7717	49,8701	53,13

Tableau 4.4.3.2 Pourcentage d'amélioration

Nous constatons que même avec $n=20$ $k=20$ l'amélioration est nette mais elle commence à diminuer pour les cinq formules mais il reste plus que 53 %.

4.4.4- Discussion de résultats

On constate d'une manière générale que les cinq formules proposées améliorent les résultats par rapport au modèle possibiliste de base sans reformulation de requête. En effet, sur tous les graphes, nous pouvons constater une amélioration de précision aux différents points P5, P10, P15, P20, P30, P100, P1000 représentant le nombre de documents pertinents parmi les 5, 10, 15, 20, 30, 100, 1000 premiers documents. L'amélioration apportée par la reformulation est nette sur les précisions à 5 à 10 et à 30. Ceci est intéressant car il signifie que les documents qui répondent à la requête utilisateur sont positionnés parmi les premiers documents restitués et la précision moyenne est améliorée de plus que 53% pour le cinq formules, elle atteint 121% pour la formule ($Nécessité * \frac{r}{R}$) (avec $n=10, k=20$) par rapport à la précision moyenne du modèle possibiliste de base (sans reformulation de requête) nous pouvons conclure que cette formule est la plus stable parmi les autres et celle qui a donné les meilleurs résultats pour toutes les configuration de n, α, β, γ .

4.5 Reformulation Aveugle

Nous avons effectué les mêmes expérimentations en considérant une reformulation aveugle. Dans ce type de reformulation aucun jugement de pertinence n'est effectué sur les documents

restitués, mais on suppose d'une manière aveugle que les k premiers documents sélectionnés par la requête comme étant pertinents puis on construit la nouvelle requête en considérant le même principe que précédemment.

Dans ce cas l'évaluation est une évaluation habituelle par contre elle est une évaluation résiduelle dans le cas précédent.

Le tableau suivant donne un récapitulatif des résultats avec $k=10$:

Initial	0,208	0,192	0,1827	0,166	0,147	0,098	0,0286	0,123	0,0862
Possibilité moyenne	0,192	0,142	0,1253	0,117	0,099	0,062	0,0183	0,0727	0,0575
Nécessité moyenne	0,2041	0,1633	0,1361	0,1204	0,116	0,072	0,023	0,0911	0,0742
Nécessité*possibilité	0,184	0,138	0,1253	0,122	0,109	0,064	0,0223	0,0918	0,0671
possibilité* (r / R)	0,16	0,156	0,1293	0,112	0,093	0,055	0,0175	0,0856	0,0633
Nécessité* (r / R)	0,212	0,166	0,1493	0,145	0,129	0,08	0,0231	0,1037	0,0755

Tableau 4.5.1 Valeurs de précision

Nous remarquons une diminution de précision pour les cinq formules car parmi les dix premiers documents restitués par le moteur possibiliste nous pouvons trouver quatre ou cinq documents, sachant que dans le cas de jugement de l'utilisateur les dix documents sont pertinents

4.6 Conclusion

Nous avons décrit dans ce chapitre les expérimentations effectuées pour mesurer la viabilité de notre approche de reformulation. D'une manière générale, ces expérimentations s'articulent autour de l'impact des cinq formules proposées sur les performances du moteur possibiliste de base.

Nous avons comparé la performance de notre système à celle obtenue dans le système possibiliste de base (sans reformulation de requête). La précision moyenne que nous avons obtenu est supérieure de 53% à celle obtenue par le système possibiliste de base pour $k=20$ et n dans $\{5, 10, 15, 20\}$ dans les cas des cinq formules proposées et atteint 121 % pour la formule qui calcule les nécessités normalisées des termes avec $n=10$.



CONCLUSION GENERALE ET PERSPECTIVES

Les travaux présentés dans ce rapport s'inscrivent dans le cadre des études sur les systèmes de recherche d'information. Ils concernent la reformulation de requête par réinjection de pertinence dans un modèle possibiliste pour la recherche d'information.

La reformulation de requête est une technique utilisée en recherche interactive d'information pour permettre à l'utilisateur de fournir des informations supplémentaires dans le but d'aider le système à restituer les documents les plus pertinents.

Nous avons proposé un processus de reformulation de requête par réinjection de pertinence basé sur le jugement de l'utilisateur sur les documents restitués en intégrant la possibilité et la nécessité d'un terme. L'intégration de ces deux degrés de pertinence nous aide à préciser les termes à ajouter dans la nouvelle requête. Suivant les formules que nous avons proposées, nous avons choisi les n premiers termes par ordre décroissant de leur pertinence finale (possible et nécessaire). Les résultats de ce processus ont effectivement amélioré les performances du moteur possibiliste de base dans la restitution de documents en réponse aux besoins d'utilisateurs. La précision moyenne a augmentée de plus de 53% pour les cinq formules proposées et elle atteint 121% pour la formule de nécessité normalisée et pour $n=10$. Ces résultats montrent que l'introduction de la possibilité et de la nécessité est intéressante et viable pour la reformulation par réinjection de pertinence.

Comme perspectives du travail, nous envisageons d'appliquer la reformulation de requête par la méthode de Relevance Feedback en se basant sur le réseau bayésien en faisant la propagation inverse des informations dans le réseau possibiliste bayésien.

Cette propagation consiste à calculer la pertinence (possible et nécessaire) d'une nouvelle requête étant donné la liste des documents jugés pertinents par l'utilisateur pour trouver la

meilleure configuration des termes ($\Pi(\theta/D_1, D_2, D_3, Q)$ et $N(\theta/D_1, D_2, D_3, Q)$ avec D_1, D_2, D_3 sont les documents jugés pertinents et θ est la nouvelle configuration des termes) pour l'utiliser dans la nouvelle requête.

REFERENCES

[BUEL 85] D.A. Buell, D.H. Kraft. *A Model for a Weighted Retrieval System*. Journal of the American Society for Information Science . 1985. 32(3) : 211-216

[BOUG 95] M.Boughanem. *Un système de Recherche d'Informations Orienté Objet Basé sur l'Approche Connexionniste*, Rapport de Recherche, Toulouse (France), décembre 1992.
Ce Rapport récapitule un ensemble d'études sur les systèmes de recherche d'information avec le modèle connexionniste.

[BOUG 00] M.Boughanem. *Contribution à la Formalisation et à la Spécification des Systèmes de Recherche et de Filtrage d'Information*, Habilitation à Diriger les Recherches, Université Paul Sabatier de Toulouse, 23 Novembre 2000.
Cette Habilitation récapitule un ensemble d'études sur les systèmes de recherche d'information avec le modèle connexionniste Mercure, système de filtrage d'information, les algorithmes génétiques pour la RI et le croisement de langues. Décrit également le projet TREC ...

[BOUG 98] M.Boughanem, C. Chrisment, C. SOULE-DUPUY. *Query modification based on Relevance Back-propagation in ad-hoc environment. IPM : Information Process and Mangement*.
Ce Rapport récapitule un ensemble d'études sur les systèmes de recherche d'information avec le modèle connexionniste.

[BOUG 97] M.Boughanem, C. SOULE-DUPUY. *Query modification based on Relevance Back-propagation.*, 5^{ème} Conférence internationale RIAO Recherche d'Information Assistée par Ordinateur. 1997.
L'article présente une méthode de reformulation automatique de requêtes basée sur la rétropropagation de la pertinence dans un modèle connexionniste.

[BRI 04c] BRINI A., BOUGHANEM M., DUBOIS D., « Vers une approche possibiliste pour la Recherche d'Information », *Veille Stratégique Scientifique et Technologique, (VSST 2004)*, 2004, p. 55-65.

[BRI 05a] BRINI A., BOUGHANEM M., DUBOIS D., « A Model for Information Retrieval based on Possibilistic Networks », *Proc. of the symposium on String Processing and Information REtrieval (SPIRE 2005), LNCS, Springer, 2005, p. 271-282.*

[**BRI 05b**] BRINI A., CAMPOS L., DUBOIS D., BOUGHANEM M., « Query Propagation in Possibilistic Information Retrieval Networks », *Proc. of the Conference of the European Society for Fuzzy Logic and Technology, (EUSFLAT 2005)*, 2005.

[**BRI 05c**] BRINI A. H., « un modèle de recherche d'information basé sur les réseaux possibilistes », Thèse de doctorat, Université de Toulouse III, Université Paul Sabatier (UPS), 2005.

[**BUCK 94**] C. Buckley, G. Salton, J. Allan. *The Effect of Adding Information in a Relevance Feedback Environment*, (SIGIR). 1994.

Application de la technique de réinjection de pertinence dans la base TREC

[**CAM 02**] L.M.de Campos, J.M. Fernández, and J. Huete. Document Instantiation for Relevance Feedback in the Bayesian Network Retrieval Model . In. *Proc. of the Intelligent Data Analysis Confer.* 2001.

[**CAM 03**] DE CAMPOS L., FERNANDEZ-LUNA J., HUETE J., « Implementing relevance feedback in the Bayesian network retrieval model », *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 54, no 4, 2003, p. 302-313.

[**CARO 97**] Budi Yuwono, Savio L.Y. Lam, Jerry H. Ying, Dik L. Lee, A World Wide Web Resource Discovery System <http://www.w3.org/conferences/www4/papers/66>

[**DUBO 88**] D. Dubois, H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press : New York, 1988.

[**HAIN 93**] D. Haines, W.B. Croft. *Relevance Feedback and Inference Networks*. Conference on Research and Development in Information Retrieval (SIGIR). 1993

Les auteurs proposent d'attribuer des poids moins importants aux termes à ajouter lors de la reconstruction de la requête.

[**HAR 92**] HARMAN D., « Relevance feedback and other query modification techniques », *Information Retrieval : Data Structures and Algorithms*, William B. Frakes and Ricardo Baeza-Yates, editors, Prentice Hall, Englewood, Cliffs, NJ, 1992, p. 241-263.

[**IDE 71**] Ide, E, (1971). *New experiments in relevance feedback*. In G. Salton (Ed.) *The Smart System-Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Les citations de documents sont montrés à l'utilisateur pour qu'il puisse juger leur pertinence. Nouvelle manière de réinjection de pertinence.

[**IAN 01**] "Abduction, explanation and relevance feedback " Department of Computing Science Submitted for the Degree of Doctor at the University of Glasgow 2001

[LUND 97] C. Lundquist, D.Grossman, O. Frieder. *Improving Relevance Feedback in the Vector Space Model.*, In The Proceedings of the 6th. ACM Annual Conference on Information Knowledge Management (CIKM 97).

[ROCC 71] J.J.Rocchio. *Relevance Feedback in Information Retrieavl*, in The Smart System Experiments in Automatic Document Processing in Automatic Document Processing, edi Prentice-Gall. 1971. 313-323.

Le but de l'article est de proposer une méthode de reformulation automatique de requêtes dans le cas du modèle vectoriel. La requête apprise est dans ce cas exprimée en fonction de la requête originale, du centroïde des documents pertinents et des documents non pertinents

[SALT, 71] G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Inc, NJ.1971.

Le modèle vectoriel est peut-être le modèle le plus populaire parmi les autres modèles de recherche d'information. La majeure partie des recherches tourne autour du modèle SMART développé à l'Université de Cornell, cet article décrit brièvement le système SMART, représentation, indexation, recherche, etc...."

[SALT 83] G. Salton, M.J. MacGill. *Extended Boolean Information Retrieval*, Communications of the ACM 1983. Vol. 26, N°12, 1022-1036, 1983.

[SALT 89] G.Salton.. *Automatic Text Processing*, The Transformation Analysis and Retrieval of Information by Computer. 1989. Addison Wesley.

[SALT 90] G. Salton & C. Buckley. *Improving Retrieval Performance By Relevance Feedback*, Journal of The American Society for Information Science. 1990. 41(4) : 288-297.

Dans cet article est définie la pondération des poids des termes lors de la reformulation. On attribue des poids plus importants aux termes des documents pertinents.

[SAVO 91] J. Savoy, D. Dubois : *Bayesian Inference Networks in Hypertext*, Intelligent Multimedia Information Systems and Management (RIAO), 662-681, 1991.

[TURT 91] H.R. TURTLE, W.B. CROFT. *Efficient Probabilistic Inference for Text Retrieval*. Intelligent Multimedia Information Systems and Management (RIAO). 1991.

[Loiseau 04] *Recherche flexible d'information par filtrage flou qualitatif*. Thèse de doctorat, Université Paul Sabatier, Toulouse, décembre 2004.

[Zadeh 65] Zadeh, L. (1965). Fuzzy sets. Information and control 8, pages 338–353.