

THÈSE

Présentée devant

l'Université Paul Sabatier de Toulouse

en vue de l'obtention du

Doctorat de l'Université Paul Sabatier

Spécialité : **INFORMATIQUE**

Par

ASMA HEDIA BRINI

**Un Modèle de Recherche d'Information
basé sur les Réseaux Possibilistes**

Soutenue le 07 Décembre 2005, devant le jury composé de :

M. P. BOSC	Professeur à l'ENSSAT de Lannion Rapporteur
M. M. BOUGHANEM	Professeur à l'Université Paul Sabatier, Toulouse III Directeur de thèse
M. C. CHRISMENT	Professeur à l'Université Paul Sabatier, Toulouse III Examineur
M. D. DUBOIS	Directeur de Recherche CNRS à l'Université Paul Sabatier, Toulouse III Directeur de thèse
M. J.-M. PINON	Professeur à l'INSA de Lyon Rapporteur
M. G. ZURFLUH	Professeur à l'Université de Toulouse I Président

INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE

Centre National de la Recherche Scientifique - Institut National Polytechnique - Université Paul Sabatier
Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 04. Tel : 05.61.55.66.11

Résumé

Notre travail s'inscrit dans le domaine de la Recherche d'Information (RI). Nous nous sommes principalement intéressés à la modélisation de la pertinence et les pondérations des termes d'indexation. Plus précisément, nous donnons deux sens différents mais complémentaires à la notion de pertinence. Nous proposons une pertinence certaine et une pertinence plausible d'un document étant donnée une requête. Les documents sont restitués par ordre décroissant de leur nécessaire pertinence et ensuite, ou à défaut, par ordre décroissant de leur pertinence plausible ou possible. Ce modèle devrait être capable de répondre à des propositions du type :

- il est plausible à un certain degré que le document constitue une bonne réponse à la requête ;
- il est nécessaire, certain (dans le sens possibiliste), que le document réponde à la requête.

Le premier type de proposition vise à éliminer certains documents de la réponse ("weak plausibility"). La seconde réponse se focalise sur les documents qui seraient pertinents. Ces mesures duales de la pertinence sont basées sur les poids des termes. Le second problème que nous tentons de résoudre concerne la pondération des termes d'indexation. Nous proposons, comme pour la modélisation de la pertinence, de mesurer l'importance des termes d'indexation par deux mesures duales (i) la nécessité de représentativité et (ii) la possibilité de représentativité d'un terme d'un document. Un terme absent d'un document est non représentatif d'un document. Un terme certainement représentatif d'un document permet de pointer avec certitude vers ce document. Enfin, la troisième solution que nous apportons s'articule autour des termes considérés dans le calcul des scores de pertinence. Dans nos travaux, nous considérons que l'absence d'un terme de la requête dans la représentation des documents doit être une information à ne pas ignorer lors du calcul de la pertinence. A ce titre, l'approche que nous proposons permet de tenir compte explicitement de ce type de termes. Afin de prendre en compte les différentes réponses citées ci-dessus, nous proposons un modèle basé sur un réseau possibiliste. Les noeuds représentent les documents, termes d'indexation et la requête. Les relations de dépendance traduisant la représentativité d'un terme dans un document ou une requête sont quantifiées par des degrés de possibilité et de nécessité. Le dernier point auquel nous nous sommes intéressés concerne la distribution des termes dans la collection de documents. Plus précisément, un terme rare de la collection peut aider à pointer facilement vers certains documents. Le facteur le plus efficace proposé jusqu'ici a été la fréquence inverse du document dans la collection idf. Ce facteur donne la répartition d'un terme donné dans la collection. Cependant, selon notre conception, deux termes qui apparaissent

dans le même nombre de documents de la collection ne sont pas discriminants d'une manière équivalente. Nous proposons trois facteurs de discrimination et détaillons le comportement de chacun. Afin de valider nos propositions nous avons effectué plusieurs expérimentations sur une des collections standards de RI. Cette collection provient de la campagne d'évaluation CLEF (Cross Language Evaluation Forum). Nous avons pour cela comparé notre modèle à un des modèles les plus performants, en termes de rappel et précision, à savoir le modèle probabiliste OKAPI. Les résultats obtenus par notre modèle sur la collection utilisée sont meilleurs que ceux du système OKAPI sur la majorité des points de précision considérés.

Table des matières

Introduction générale	2
I La Recherche d'Information (RI)	12
1 La Recherche d'Information	13
1.1 Introduction	13
1.2 Les concepts fondamentaux de la recherche d'information	14
1.3 Pondération	18
1.3.1 L'apport de Luhn	18
1.3.2 Vue statistique	19
1.3.3 Vue probabiliste	20
1.4 Les modèles connus de la RI	22
1.4.1 Modèle vectoriel	23
1.4.1.1 Mesures de similarité	23
1.4.1.2 Pondération	25
1.4.1.3 Illustration	26
1.4.1.4 Conclusion	28
1.4.2 Stratégies de Recherche Probabilistes	29
1.4.2.1 Probabilité de pertinence	29
1.4.2.2 Modèle BIR	30
1.4.2.3 Pondération	32
1.4.2.4 Modèle de Poisson	33
1.4.2.5 Illustration	35
1.4.3 Autres modèles probabilistes	36
1.5 Techniques d'évaluation des SRIs	37
1.5.1 Rappel, précision et <i>fall out</i>	38
1.5.2 Interpolation	41
1.6 Conclusion	41

II	Réseaux Bayésiens et Recherche d'Information	44
2	Les Réseaux Bayésiens	45
2.1	Introduction	45
2.2	Définitions	46
2.3	Relations de dépendance	46
2.4	Calcul des probabilités	48
2.4.1	Axiomes de base	48
2.4.2	Probabilités conditionnelles	49
2.4.3	La règle de chaînage	50
3	Les modèles de Recherche d'Information basés sur les Réseaux Bayésiens	51
3.1	Introduction	51
3.2	Modélisations graphiques en RI	52
3.3	Le modèle inférentiel	53
3.3.1	Architecture générale	54
3.3.2	Calcul de la pertinence	56
3.3.3	Agrégation de la requête	57
3.3.4	Pondération des arcs $P(T_i D_j)$	58
3.3.5	Illustration	59
3.4	Le modèle de croyance	62
3.4.1	Architecture générale	62
3.4.2	Calcul de la pertinence	63
3.4.3	Probabilité des documents $P(D_j Par_{D_j})$	64
3.4.4	Probabilité de la requête $P(Q Par_Q)$	65
3.4.5	Généralisation des modèles classiques	66
3.5	Autres modèles basés sur les réseaux Bayésiens	66
3.5.1	Modèle d'Indrawan	67
3.5.2	Réseaux multi connectés pour la RI	69
3.5.3	Réseaux de croyance basés sur les expressions d'indexation	72
3.6	Conclusion et Discussions	73
III	Un Modèle de RI basé sur les Réseaux Possibilistes	76
4	Un modèle de Recherche d'Information basé sur les Réseaux Possibilistes	77
4.1	Introduction	77
4.2	Les Réseaux Possibilistes	78

4.2.1	La théorie des possibilités	78
4.2.1.1	Distribution de possibilité	78
4.2.1.2	Mesures de nécessité et de possibilité	79
4.2.1.3	Conditionnement possibiliste	80
4.2.2	Réseaux Possibilistes (RP)	80
4.2.2.1	Définitions	81
4.2.2.2	Réseaux possibilistes basés sur le minimum	81
4.2.2.3	Réseaux possibilistes basés sur le produit	82
4.2.2.4	Logique possibiliste	82
4.3	Un modèle de RI basé sur les réseaux possibilistes	83
4.3.1	Architecture générale du modèle	84
4.3.2	Evaluation de la requête	86
4.3.3	Agrégation des termes de la requête	89
4.3.3.1	Agrégations booléennes et quantifiée des termes de la requête	90
4.3.3.2	Noisy OR	92
4.3.4	Pondération des termes d'indexation	94
4.3.4.1	Arcs document-terme $\Pi(T_i D_j)$	94
4.3.4.2	Termes racines	102
4.3.5	Possibilité <i>a priori</i> des documents	102
4.4	Nouveaux facteurs de discrimination	103
4.4.1	Motivations	103
4.4.2	Pouvoir de discrimination	104
4.4.3	Discrimination par fréquence normalisée pondérée	108
4.4.4	Discrimination par entropie	110
4.4.5	Exemple comparatif	112
4.5	Illustration du modèle proposé	114
4.6	Conclusion	120
5	Expérimentations	122
5.1	Collection de tests	123
5.2	Protocole d'évaluation	123
5.3	Le modèle de base	126
5.4	Expérimentations et résultats	128
5.4.1	Impact des facteurs de discrimination « df »	128
5.4.2	Impact des techniques de pondération $\Pi(T_i D_j)$	130
5.4.3	Impact de la longueur des documents	133
5.4.4	Comparaison <i>idf</i> et <i>NDF</i>	135
5.4.5	Elimination des facteurs NDF_i du modele de base	137
5.5	Comparaison avec <i>OKAPI</i>	138
5.6	Conclusion	140

IV Conclusion générale	143
Conclusion Générale	144

Liste des tableaux

1.1	Poids des termes	27
1.2	Classement des documents par cosinus	27
1.3	Poids des termes obtenus par la méthode de pivot	27
1.4	Classement des documents selon la normalisation par pivot	28
1.5	Table de contingence des termes	33
1.6	Poids des termes	35
1.7	Classement des documents	36
1.8	Tableau de contingence de la pertinence	39
1.9	Exemple de calcul de rappel et précision pour les 5 premiers documents restitués	40
3.1	Probabilité conditionnelle du terme T_2 , $P(T_2 D_j)$	59
3.2	Probabilité conditionnelle du terme T_3 , $P(T_3 D_j)$	59
3.3	Probabilité conditionnelle du terme T_6 , $P(T_6 D_j)$	59
3.4	Probabilité conditionnelle de la requête $P(Q \theta)$	60
3.5	Classement des documents selon la propagation dans le modèle d'inférence	61
4.1	Agrégation quantifiée des termes de la requête $\Pi(Q \theta)$	92
4.2	Possibilités conditionnelles $\Pi(T_i D_j)$	99
4.3	Possibilités conditionnelles $\Pi(T_i D_j)$	101
4.4	Distributions des termes t_1 et t_2 dans les documents D_1 et D_2	106
4.5	Répartition des termes t_1 et t_2 dans la collection	107
4.6	Répartition des termes t_3 et t_4 dans la collection	107
4.7	Pouvoir discriminant des termes t_1, t_2, t_3, t_4	108
4.8	Fréquence maximale dans les documents	110
4.9	Répartition des termes dans les documents	110
4.10	Pouvoir discriminant des quatre termes	111
4.11	Facteurs de discrimination	113
4.12	Possibilités conditionnelles $\Pi(T_2 D_j)$	115
4.13	Possibilités conditionnelles $\Pi(T_3 D_j)$	115
4.14	Possibilités conditionnelles $\Pi(T_6 D_2)$	115
4.15	Possibilités <i>a priori</i> des documents $\Pi(D_j)$	116

4.16	Possibilités marginales des termes $\Pi(T_k)$	116
4.17	Possibilités conditionnelles des parents de Q	116
4.18	Possibilité de pertinence des documents	118
4.19	Possibilités conditionnelles des parents de Q , $\Pi(Q \theta)$	119
4.20	Classement des documents	119
5.1	Possibilités conditionnelles et marginales	126
5.2	Pourcentage de perte liée à l'utilisation de la pondération « positive »	132
5.3	Impact de la longueur sur les points de précisions	134
5.4	Discrimination positive vs discrimination négative	136
5.5	Pourcentage d'amélioration de notre approche comparée à l'approche probabiliste	139

Table des figures

1.1	Processus en U de la RI	14
1.2	Diagramme illustrant les mots significatifs	19
1.3	Courbes de précision-rappel pour deux requêtes Q_1 et Q_2	40
2.1	Connexion en série	47
2.2	Connexion divergente	47
2.3	Connexion convergente	48
3.1	Architecture générale	54
3.2	Architecture simplifiée	56
3.3	Architecture générale	63
3.4	Architecture globale	67
3.5	Architecture globale	70
4.1	Architecture générale du modèle	85
5.1	Points de précision	127
5.2	Normalisation des facteurs de discrimination	129
5.3	Pondération positive vs négative	130
5.4	Pondération positive vs négative	131
5.5	Pondération positive vs négative	131
5.6	Elimination de la longueur du calcul des scores de pertinence	133
5.7	Courbes de rappel des NDF avec et sans longueur	135
5.8	Impact des facteurs de discrimination sur la présence et l'absence des termes	136
5.9	Non prise en compte des termes de la requête absents	137
5.10	Comparatif des deux systèmes : Possibiliste et OKAPI	139
5.11	Comparatif des deux systèmes en terme de rappel précision : Possibiliste et OKAPI	140

Introduction générale

L'explosion sans précédent du volume d'informations disponibles sous des formats hétérogènes produites par des sources d'informations distribuées est un des résultats des développements de l'Internet et de l'informatique dans tous les secteurs d'activité. L'élaboration de systèmes automatisés pour gérer ces masses de données est devenue dans un tel contexte une nécessité.

Notre travail se situe dans le contexte de ces outils automatisés et plus précisément dans le domaine de la Recherche d'Information (RI). La RI est une branche en informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche des informations. Elle propose des outils, appelés systèmes de recherche d'information (SRI), dont l'objectif est de capitaliser un volume important d'information et d'offrir des moyens permettant de localiser les informations pertinentes relatives à un besoin en information d'un utilisateur exprimé à travers une requête. Nous nous intéressons dans le cadre de ce travail à l'information textuelle. Nous utiliserons indifféremment les termes information ou document, pour désigner la portion de texte renvoyée à l'utilisateur.

Contexte du travail

L'objectif principal d'un SRI est d'extraire à partir d'une collection de documents, ceux qui sont susceptibles de répondre à un besoin utilisateur. Dans ce contexte, pour évaluer la pertinence d'un document vis à vis d'un besoin exprimé par une requête, la majorité des SRI mesure un score de pertinence entre un document, représenté par ses mots clés et la requête. Cette mesure de pertinence, élément fondamental de tout SRI, est souvent formalisée à travers la notion de modèle de recherche d'information. Plusieurs modèles de RI ont été

proposés. D'une façon générale, les fondements mathématiques sur lesquels se basent ces modèles reposent sur l'utilisation de l'algèbre, la logique, la théorie des probabilités et les statistiques. Ces modèles peuvent être répertoriés en trois catégories selon leur définition de la pertinence [38] :

1. La pertinence est vue comme la similarité entre la requête et le document ([92] [109]) ;
2. La pertinence est modélisée par une variable aléatoire binaire et des modèles probabilistes sont utilisés pour calculer la valeur de la variable ;
3. L'incertitude sur la pertinence est produite par l'inférence de la requête à partir des documents ou inversement.

Quelle que soit la sémantique donnée à la représentation des objets (document ou requête) ou la définition de la pertinence, ces modèles ont un comportement général identique. La majorité d'entre eux représentent les documents et la requête par une liste de mots clés pondérés et le calcul de la pertinence est obtenu à partir de ces poids censés refléter l'importance des termes dans les documents. Plus précisément, pour la première catégorie, la pertinence est donnée par la similarité entre les représentations des documents et de la requête à un certain niveau [92] [109].

La seconde catégorie de modèles calcule la probabilité de pertinence des documents étant donnée une requête ou la probabilité de satisfaire la requête étant donné le document. Elle peut être subdivisée en deux sous catégories correspondant à la génération de documents [84] [82] [46] ou à la génération de requêtes [74] [85] [46]. Dans la première, [84], les poids des termes d'indexation dépendent de la distribution des termes dans les classes des documents pertinents et celle des documents non pertinents. Dans les modèles basés sur la génération de requêtes [74] [85] [46] les poids des termes dépendent de facteurs liés au jugement utilisateur. Dans les deux cas, les poids restent difficilement estimables.

Finalement, pour la dernière catégorie de modèles, la pertinence est modélisée par l'incertitude sur la déduction de la requête à partir du document. Ce type de modèles est basé sur des probabilités qui calculent l'incertitude liée aux inférences [114] [113]. Les représentations des objets (document ou requête) dans ces modèles ne dépend pas du cadre des inférences (les termes ne sont pas déduits à partir des objets comme pour la pertinence).

Problématique

Les problématiques auxquelles nous nous sommes intéressés concernent principalement la modélisation de la pertinence et les pondérations des termes d'indexation. Les modèles énoncés ci-dessus calculent la pertinence d'un document comme un score d'appariement entre le document et la requête ou une probabilité de pertinence d'un document vis à vis d'une requête. Ce score est une valeur unique, donnant l'événement (la pertinence) et son contraire notamment dans le modèle probabiliste. Quel que soit le modèle, le document est pertinent ou non pertinent à un certain degré. Le score de pertinence est calculé à partir des poids des termes de la requête et ceux des documents, ces poids sont considérés comme des données certaines. De plus, seuls les termes de la requête communs à ceux des documents sont considérés. Les termes de la requête absents des documents ne sont pas pris en compte.

Ces premiers constats nous amènent à poser intuitivement les questions suivantes :

1. Est-ce que la prise en compte d'une pertinence vue comme une variable binaire peut couvrir toute la sémantique de cette notion ?
2. Est-ce que les poids des termes d'indexation peuvent être considérés comme des données certaines pour représenter les documents *efficacement* ?
3. Et enfin, pourquoi les termes de la requête absents des représentations des documents ne sont pas pris en compte lors du calcul de la pertinence d'un document en réponse à une requête ?

La première question découle du fait que les modèles de RI, mesurent la **pertinence** d'un document en réponse à une requête utilisateur en se basant sur une unique valeur dite « score de pertinence ». De nombreux travaux de la littérature se sont penchés sur la notion de pertinence [23] [83] [101] [45] [52] [104] [102] [9] [65] [76] [12] [13]. Ces travaux s'accordent à dire qu'il n'existe pas une définition précise de la pertinence d'une part, et que cette notion est dynamique, multidimensionnelle et dépend de la perception de l'utilisateur. Un utilisateur peut juger les mêmes documents restitués en réponse à une requête donnée d'une manière différente à deux instants de temps différents. Ainsi, dans notre démarche, nous estimons qu'il n'est pas opportun d'affecter une unique valeur à la pertinence supposée englober la totalité de la sémantique de

cette notion. L'attribution d'une valeur unique de pertinence est une démarche *réductrice*, dans le sens où il est difficile de couvrir toute sa sémantique. La décision de restituer un document, donc d'évaluer sa pertinence, en se basant sur une seule valeur ne peut pas être fiable ni basée sur une certitude.

La seconde question découle de l'inconvénient d'affecter un poids unique à un terme. De manière générale, les approches de RI actuelles affectent un poids aux termes présents dans les documents. Ce poids est censé décrire le degré de représentativité du terme du contenu d'un document donné. Il est le produit de la combinaison de l'exhaustivité et de la spécificité d'un terme. Ces notions sont définies sur des données portant sur des échelles différentes. Les poids des termes sont généralement obtenus par la combinaison de $tf \times idf$. Le facteur tf donne le nombre d'apparitions d'un terme dans le document et idf le nombre d'apparitions d'un terme dans la collection. Le premier facteur permet de mesurer l'exhaustivité (par rapport au document) et le second la spécificité du terme dans la collection. A notre sens, ces facteurs mesurés sur des échelles différentes suggèrent implicitement des imprécisions et des incertitudes que les modèles existants ne traitent pas explicitement. Un seul poids est généralement obtenu (regroupant donc des informations différentes) pour quantifier à quel point un terme est « apte » à décrire ou représenter un document.

La troisième question concerne la non séparation entre la négation et l'absence. En effet, la pertinence d'un document est calculée à partir des poids des termes de la requête et ceux des documents. Les termes considérés sont ceux de la requête communs à ceux des documents. Cependant certains termes de la requête peuvent contribuer à pointer vers un sous ensemble particulier des documents de la collection. L'ignorance de ces termes lors du calcul des scores de pertinence des documents ne les contenant pas introduit de l'incertitude dans ce calcul. La vision qui assimile l'ignorance à la négation ne tient pas compte de toutes les informations disponibles, que nous considérons peu nombreuses, pour le calcul des scores de pertinence.

Contributions

Dans la littérature les modèles de RI sont catégorisés en fonction de la définition qu'ils donnent à la pertinence. Dans ce contexte, il s'agit de la pertinence d'un document vis à vis d'une requête. Les travaux que nous proposons s'inscrivent dans la définition d'un nouveau modèle de RI permettant notamment une nouvelle modélisation de la pertinence. Nous nous sommes particulièrement penchés dans nos travaux sur la résolution des trois problèmes soulevés dans la section précédente. Nous proposons :

- La redéfinition de la pertinence et sa quantification ;
- La redéfinition de la représentativité des documents et particulièrement la pondération affectée aux termes d'indexation ;
- La prise en compte et la quantification des termes de la requête absents des représentations des documents lors du calcul de la pertinence.

D'une manière générale, quel que soit le modèle de la littérature, et particulièrement ceux qui considèrent les poids des termes comme des probabilités de pertinence, l'incomplétude (ou imprécision) de l'information n'est pas considérée lors de la représentation d'un document ou de son évaluation étant donnée une requête.

A notre sens, une seule valeur de pertinence ne couvre pas l'imprécision et le vague intrinsèque à la notion de pertinence [13]. Nous cherchons à traduire la pertinence en essayant de modéliser différents aspects auxquels elle est reliée. Nous considérons une mesure de probabilité portant sur un événement et son contraire quelque peu restrictive. Nous proposons pour notre part de modéliser la pertinence dans un cadre possibiliste. La théorie des possibilités et la logique possibiliste proposent des mesures duales traitant l'incertitude liée à l'information d'une manière flexible et différente de la théorie des probabilités. Dans la littérature, les notions de possibilité ou de certitude reliées à un événement ou une information, sont laissées en marge, en RI, lors des calculs de la pertinence ou de la décision de la pertinence des documents.

Un des objectifs de ce travail est donc de proposer une approche moins restrictive pour la modélisation de la pertinence. En effet, nous donnons deux sens différents mais complémentaires à la notion de pertinence. Nous proposons une pertinence certaine et une pertinence plausible d'un document étant

donnée une requête. Les documents sont restitués par ordre décroissant de leur **nécessaire pertinence** et ensuite, ou à défaut, par ordre décroissant de leur **pertinence plausible ou possible**. Ces pertinences dépendent, comme dans les modèles actuels, des poids des termes de la requête et des documents. Tous les termes de la requête sont pris en compte qu'ils soient absents ou présents dans le document (pour lequel nous calculons la pertinence). Ce modèle devrait être capable de répondre à des propositions du type :

- il est plausible à un certain degré que le document constitue une bonne réponse à la requête ;
- il est nécessaire, certain (dans le sens possibiliste), que le document réponde à la requête.

Le premier type de proposition vise à éliminer certains documents de la réponse ("weak plausibility"). La seconde réponse se focalise sur les documents qui seraient pertinents. Dans le modèle proposé, un document contenant tous les termes de la requête constitue une réponse possiblement pertinente à la requête. Cette plausibilité doit être renforcée par une certitude provenant de la mesure de nécessité.

Ces mesures duales de la pertinence sont basées sur les poids des termes. Le second problème que nous tentons de résoudre concerne la pondération des termes d'indexation. Nous proposons, comme pour la modélisation de la pertinence, de mesurer l'importance des termes d'indexation par deux mesures duales (i) la **nécessité de représentativité** et (ii) la **possibilité de représentativité** d'un terme d'un document. Un terme absent d'un document est non représentatif d'un document. Un terme certainement représentatif d'un document permet de pointer avec certitude vers ce document. La possibilité de représentativité permet d'éliminer les termes non représentatifs du document et la nécessité de représentativité permet de focaliser vers certains documents les contenant.

Enfin, la troisième solution que nous apportons s'articule autour des termes considérés dans le calcul des scores de pertinence. Généralement, la pertinence d'un document donné est égale à la somme des produits des poids des termes de la requête communs à ceux du document. Dans nos travaux, nous considérons que l'absence d'un terme de la requête dans la représentation des documents

doit être une information à ne pas ignorer lors le calcul de la pertinence. A ce titre, l'approche que nous proposons permet de tenir compte explicitement de ce type de termes. En effet, la prise en compte de ces termes est une partie intégrante de notre modèle.

- L'utilisateur ignore la représentation des documents lors de la formulation de son besoin ;
- L'importance d'un terme dans la collection est liée à sa spécificité dans la collection.

Le premier point suggère qu'il y a une discordance entre le contenu *réel* des documents, le besoin *réel* de l'utilisateur, et leurs représentations respectives. Cette discordance est suscitée par les méthodes (statistiques) incertaines utilisées pour obtenir ces représentations. D'une part, dans le système de RI que nous proposons, la requête instancie le système et la propagation s'exécute sur tous les termes de la requête qu'ils soient absents ou présents dans les documents. Ainsi, nous considérons la requête comme l'information la plus *sûre* disponible pour le système.

Le second point suggère que l'absence d'un terme de la requête d'un document donné pénalise le score de pertinence de document. Cette pénalisation est fonction de l'importance du terme dans la collection. Un terme important dans une collection donnée est un terme spécifique de la collection. Il oriente le système vers un sous ensemble de documents particuliers, lorsque ce terme figure dans la requête. Pour éviter l'amalgame entre ces deux notions, la structuration de notre modèle permet de traiter les deux notions aussi bien en les séparant qu'en les combinant. Cette approche est possible au moyen des degrés de nécessité et de possibilité.

Afin de prendre en compte les différentes réponses citées ci-dessus, nous proposons un modèle basé sur un réseau possibiliste. Les noeuds représentent les documents, termes d'indexation et la requête. Les relations de dépendance traduisant la représentativité d'un terme dans un document ou une requête sont quantifiées par des degrés de possibilité et de nécessité. La pertinence d'un document étant donnée une requête est évaluée par une double mesure : possibilité et nécessité. La mesure de possibilité est utile pour filtrer les documents et la mesure de nécessité pour renforcer la pertinence des documents restants.

Le dernier point auquel nous nous sommes intéressés concerne la distribution

des termes dans la collection de documents. Plus précisément, un terme rare de la collection peut aider à pointer facilement vers certains documents. Le facteur le plus efficace proposé jusqu'ici a été la fréquence inverse du document dans la collection *idf*. Ce facteur donne la répartition d'un terme donné dans la collection. Cependant, selon notre conception, deux termes qui apparaissent dans le même nombre de documents de la collection ne sont pas discriminants d'une manière équivalente. En effet, un terme qui apparaît fréquemment à l'intérieur de peu de documents longs de la collection n'a pas la même importance qu'un terme qui apparaît rarement dans le même nombre de tels documents. Nous proposons trois facteurs de discrimination et détaillons le comportement de chacun. Nous verrons que leur effet implique des conséquences différentes en fonction de leur utilisation reliée aux mesures de possibilité et de nécessité.

Afin de valider nos propositions nous avons effectué plusieurs expérimentations sur des collections standards de RI. Ces collections proviennent des campagnes d'évaluation *TREC* (Text REtrieval Conference) et *CLEF* (Cross Language Evaluation Forum). Nous avons évalué l'impact d'une double mesure de pertinence et d'une double mesure de représentativité sur les performances de notre système. Nous avons pour cela comparé notre modèle à un des modèles les plus performants, en termes de rappel et précision, à savoir le modèle probabiliste *OKAPI*. Les résultats obtenus par notre modèle sur les collections utilisées sont meilleurs que ceux du système *OKAPI* sur la majorité des points de précision considérés.

Organisation du mémoire

Cette thèse est organisée en trois parties.

La première partie décrit le contexte général de notre travail. Ainsi nous commençons par définir les concepts fondamentaux de la Recherche d'Information (RI), dans le *Chapitre 1*. Nous étudions les modèles vectoriels et probabilistes. Nous nous intéressons particulièrement au sens de la pertinence donnée par ces modèles. Pour ce faire, nous décrivons les techniques utilisées pour la calculer ainsi que les techniques de pondération des termes d'indexation des objets manipulés, à savoir les documents et la requête.

La seconde partie est composée de deux chapitres. Dans le *Chapitre 2*, nous rappelons les définitions des Réseaux Bayésiens (RBs) et leur utilité. Dans le *Chapitre 3*, nous décrivons l'utilisation des RBs en RI. Nous commençons l'état de l'art par les modèles qui ont eu le plus de succès. Pour compléter l'état de l'art relatif à cette partie, nous survolons les majeures investigations effectuées en RI basées sur les RBS.

La dernière partie concerne nos contributions. Nous commençons le *Chapitre 4* par décrire sommairement les Réseaux Possibilistes. Nous définissons ensuite notre modèle, son architecture générale et le sens de ce modèle dans le domaine de la RI. Nous proposons trois indices de pouvoir de discrimination que nous affectons aux termes de la requête pour discriminer entre les documents de la collection. Finalement, le *Chapitre 5* décrit les expérimentations effectuées pour évaluer nos travaux.

En conclusion, nous dressons un bilan de nos travaux, en mettant en exergue nos propositions. Nous finissons par la proposition des perspectives possibles et nombreuses de nos travaux.

Première partie

La Recherche d'Information (RI)

Chapitre 1

La Recherche d'Information

1.1 Introduction

Un Système de Recherche d'Information (SRI) intègre un ensemble de modèles et de processus permettant de sélectionner des informations pertinentes en réponse au besoin en information d'un utilisateur représenté à l'aide d'une requête. Dans un contexte documentaire, un SRI permet de gérer une collection de documents stockés sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leur contenu sémantique. Un SRI peut être défini comme l'ensemble des procédures et des opérations permettant la gestion, la représentation, l'interrogation, la recherche, le stockage et la sélection des informations répondant aux besoins d'un utilisateur. L'interrogation du fonds documentaire à l'aide d'une requête nécessite la représentation de cette dernière sous une forme compatible avec celle des documents. Les fonctionnalités d'un SRI peuvent être déduites du processus global de la RI.

L'objectif de ce chapitre est de présenter les concepts de base de la RI. Dans la première section nous décrivons le processus global de la RI, communément connu sous le nom du processus en U . Nous donnons l'utilité et l'importance des opérations qui composent ce processus.

Nous décrivons dans la seconde section deux modèles connus de la RI, à savoir le modèle vectoriel et le modèle probabiliste. Nous rappelons essentiellement la modélisation de la pertinence dans ces modèles ainsi que la pondération des termes d'indexation.

Dans la dernière section, nous décrivons les techniques utilisées pour évaluer les performances des SRI.

1.2 Les concepts fondamentaux de la recherche d'information

La recherche des informations pertinentes qui répondent aux besoins d'un utilisateur, consiste à mettre en correspondance les représentations des informations contenues dans un fonds documentaire avec celles des besoins de l'utilisateur. Pour réaliser d'une façon efficace cette fonction, un SRI doit réaliser :

- l'indexation des documents de la collection et du besoin utilisateur ;
- l'appariement des représentations requête-documents pour le calcul de la pertinence des documents en réponse à un besoin utilisateur ;
- éventuellement, la reformulation de requêtes.

Pour résumer ces fonctions, nous pouvons représenter schématiquement certaines fonctionnalités d'un SRI par ce que l'on appelle communément le processus « en U », tel que dans la figure 1.1. Ce processus manipule des collections de documents et des besoins utilisateur pour rechercher, à partir d'une collection de documents, ceux qui répondent au mieux à un besoin utilisateur.

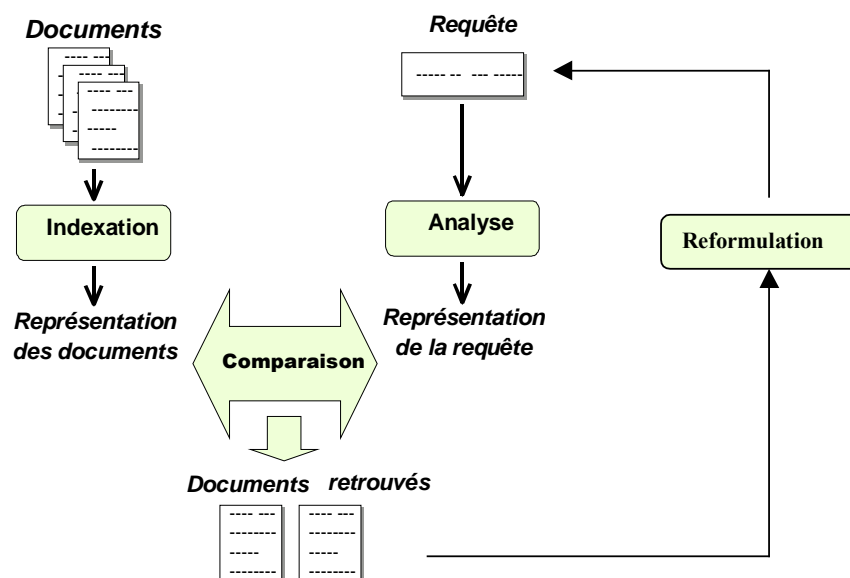


FIG. 1.1 – Processus en U de la RI

Collection de documents La collection de documents constitue l'ensemble des informations exploitables, compréhensibles et accessibles par l'utilisateur. Une collection comporte un ensemble de granules documentaires. Un granule de documents peut représenter tout ou une partie d'un document. Il représente l'unité sélectionnée en réponse à une requête de l'utilisateur. Nous nous limitons dans notre étude aux granules de documents textuels. Dans la suite de cette thèse, nous utilisons indifféremment les termes « document » ou « information » pour désigner un granule documentaire.

Besoin en information Un besoin en information est une représentation mentale de ce que l'utilisateur souhaite rechercher. Ce besoin est représenté sous forme d'une requête. La requête n'est donc qu'une représentation possible d'un besoin en information. La requête peut être exprimée de différentes manières [92] [89] [88] [69]. Les concepts de « requête » et « besoin » sont souvent confondus.

Processus d'indexation L'indexation des documents consiste à représenter les objets (document ou requête) par des descripteurs généralement représentés sous forme d'une liste de mots clés et de poids qui leur sont associés. Au moins deux découpages du sens de l'indexation existent. Ces découpages tentent de répondre à :

1. *Qui procède à l'indexation ?*
2. *Quel type d'analyse est opérée sur le texte ?*

Techniquement, l'indexation peut être manuelle, semi-automatique ou automatique [95], [94].

- indexation manuelle : chaque document est analysé par un spécialiste du domaine correspondant, ou par un documentaliste, qui détermine, selon ses connaissances, les unités syntaxiques ou mots-clés qui lui semblent les plus significatifs pour représenter le contenu du document. Les mots clés sont ensuite regroupés dans une liste plus ou moins structurée, utilisée pour les opérations d'indexation et d'interrogation.
- indexation automatique [75], [91], [70] : cette opération est réalisée à l'aide d'un processus entièrement informatisé, qui consiste à déterminer les unités syntaxiques ou mots-clés représentatifs des concepts et des thèmes exprimés dans le document. Les méthodes existantes vont d'une simple

extraction de mots simples (uniternes) à des analyses linguistiques et/ou statistiques permettant d'établir des relations sémantiques entre ces mots et de pondérer ces mots en fonction de leurs apparitions.

- indexation semi-automatique [73] [3] : cette méthode est une combinaison des deux précédentes. Le choix final reste au spécialiste du domaine ou documentaliste, qui intervient souvent dans l'établissement de relations sémantiques entre les mots-clés et le choix des multi-termes significatifs.

Un autre découpage du sens de l'indexation concerne l'analyse linguistique effectuée sur les mots ou concepts et leurs différents niveaux d'indexation. Nous avons répertorié au moins cinq grands niveaux d'analyse linguistique du texte intégral

- niveau morphologique : à ce niveau les accents sont supprimés et les mots extraits sont stockés sous la même casse (minuscules généralement). A cette étape les mots vides sont supprimés (prépositions, articles, etc).
- niveau lexical : les mots sont stockés sous une forme canonique (le radical, la racine, ou encore le lemme) [79] [41] [42]
- niveau syntaxique : la grammaire de la langue utilisée par les objets est utilisée pour extraire des groupes de mots ou des mots composés. Des syntagmes non utilisés dans l'objet peuvent être dérivés.
- niveau sémantique : les relations sémantiques sont déduites par la reconnaissance des concepts. Un thesaurus (dictionnaire de mots avec des relations entre eux de type synonymie, etc) est utilisé.
- niveau pragmatique : ce niveau est impossible à mettre en œuvre automatiquement car il s'agit de l'analyse du langage naturel par la connaissance du monde réel.

La pertinence La pertinence est une notion fondamentale et cruciale dans le domaine de la RI. Les travaux de recherche récents [45] [76] [9] s'accordent sur la difficulté de la définition de la pertinence. La pertinence d'un point de vue système [24] [22] est différente de celle d'un point de vue utilisateur. Nous nous intéressons particulièrement dans ce manuscrit à la pertinence utilisateur que nous désignerons par pertinence. La pertinence concerne d'une manière générale la restitution des documents pertinents en réponse à une requête utilisateur [23] [83]. La pertinence est liée à la perception de l'utilisateur, de plus elle est multidimensionnelle et évolue durant le temps d'une recherche [101]

[104] [52] [45] [76] [102] [9] [65] [12] [13].

Les modèles de RI définis dans la littérature (détaillés dans le second chapitre) mesurent cette pertinence comme un score, cherchant à évaluer la pertinence des documents vis à vis d'une requête. Cette pertinence est mesurée par une similarité de représentation document-requête (modèle vectoriel), une probabilité de pertinence des documents étant donnée une requête (modèle probabiliste).

Reformulation de requêtes La requête initiale est vue en RI comme un moyen permettant d'initialiser le processus de sélection d'informations pertinentes. La reformulation rentre dans un processus plus général d'optimisation de la fonction de pertinence qui a pour but de rapprocher la pertinence système de la pertinence utilisateur [11]. Ce processus permet de générer une requête plus adéquate que celle initialement formulée par l'utilisateur. Son principe est de modifier la requête de l'utilisateur par ajout de termes significatifs et/ou réestimation de leurs poids, selon deux approches :

- la première est basée sur les techniques d'association de termes (méthode directe) ;
- la seconde est basée sur les jugements utilisateurs (méthode indirecte), communément appelée Relevance feedback ou Retour de pertinence ; [96], [50].

Le principe de la méthode directe est d'ajouter à la requête initiale des termes sémantiquement proches. Cette proximité entre termes est obtenue selon différentes manières telles que des études sur le langage naturel, des mesures statistiques sur les documents de la collection, des calculs de corrélations entre termes [82], [72], [96], [80].

La seconde méthode (indirecte) permet à l'utilisateur de juger les documents restitués par le système pour repondérer les termes de la requête initiale ou ajouter-supprimer des termes qui se trouvent dans les documents jugés pertinents/non-pertinents [90] [54] [117], [89], [109], [13].

1.3 Pondération

La pondération est une composante importante dans le processus d'indexation de RI. En effet, le calcul de la pertinence d'un document en réponse à une requête utilisateur dépend des poids attribués aux termes indexant les documents communs à ceux de la requête. Il existe une étape importante avant la pondération des termes qui consiste à ne garder que les termes d'indexation intéressants. Par *intéressant* nous entendons les termes autres que les mots vides, par exemples les prépositions etc. Les méthodes proposées dans la littérature pour mesurer les termes « importants » se basent sur des outils sémantiques [94], statistiques [71] [60] [61] [99] [93] [100] [95] ou probabilistes [74] [6] [7] [51]. Un travail préliminaire à la pondération a été proposé par Luhn [70]. Les apports de Luhn concernent les statistiques sur le texte dans le but d'éliminer les termes « trop rares » ou « trop fréquents » dans la collection.

1.3.1 L'apport de Luhn

Luhn s'est inspiré de la loi de Zipf [119] pour déterminer les termes significatifs à l'intérieur d'un document. Le but est de retrouver le pouvoir de résolution d'un terme donné qui est sa capacité à identifier un document pertinent (appel) combinée à la capacité à l'isoler des documents non pertinents (précision). La loi de Zipf [119] stipule que le produit de la fréquence du terme par le rang du terme est approximativement constant. Cette loi a servi de point de départ pour les travaux de Luhn [70] [71], pour spécifier (arbitrairement) deux seuils (un seuil minimal et un seuil maximal) au delà desquels les termes ne sont pas intéressants pour représenter les documents. Le pouvoir de résolution des termes intéressants atteint un pic au milieu de l'intervalle délimité par les deux seuils. Le modèle de Luhn a été proposé pour sélectionner des termes significatifs à partir d'un document. La figure 1.2 illustre la méthode proposée par Luhn pour éliminer les mots inutiles lors d'une première étape de l'indexation des objets (documents, requêtes). Le seuil haut et bas dans la figure 1.2 correspondent respectivement aux seuils minimal et maximal.

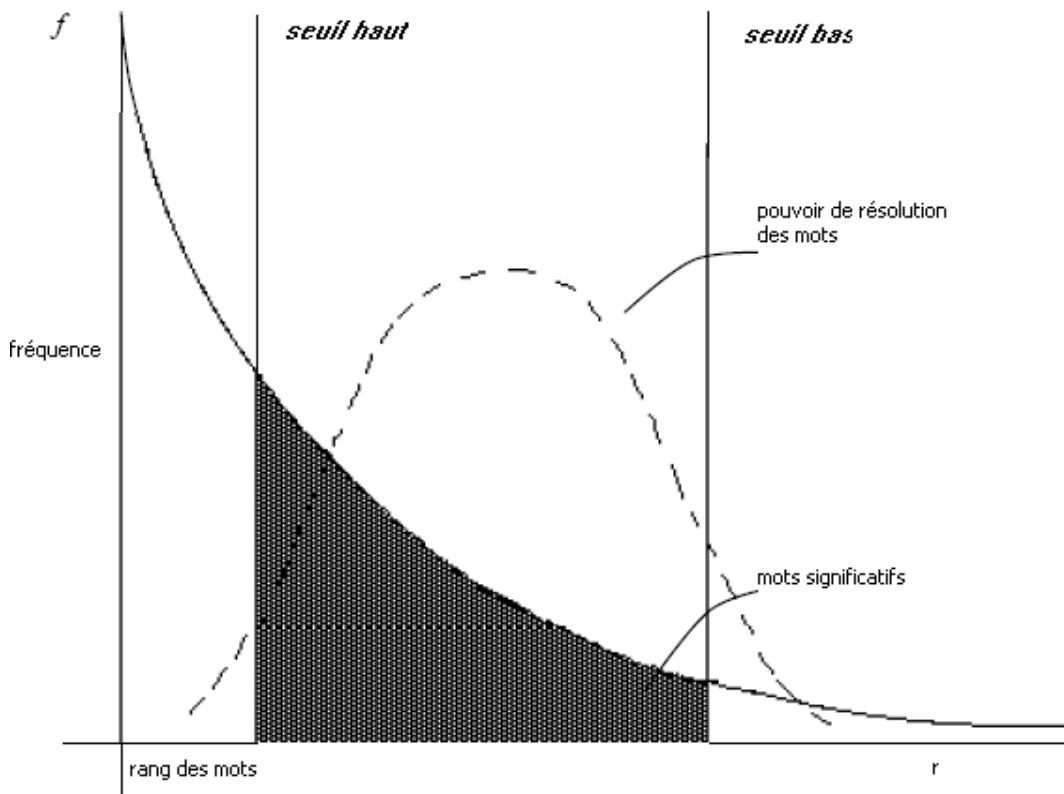


FIG. 1.2 – Diagramme illustrant les mots significatifs

1.3.2 Vue statistique

Selon les approches statistiques, l'efficacité de la pondération est guidée par deux facteurs : l'exhaustivité et la spécificité. Selon Keen et Digger [64], pour tout document, l'exhaustivité est donnée par le nombre de sujets différents indexant un objet. La spécificité est le pouvoir d'un langage d'indexation à décrire les sujets précisément. Ces deux notions sont largement corrélées à deux facteurs largement utilisés en RI : le rappel et la précision. Le rappel d'un Système de Recherche d'Information, SRI, est la proportion de documents pertinents réellement retournés en réponse à une requête utilisateur. La précision est la proportion de documents retournés réellement pertinents. Plusieurs travaux, particulièrement ceux de Lancaster [68] ont montré qu'un haut niveau d'exhaustivité (resp. spécificité) conduit à un haut (resp. bas) niveau de rappel et à un bas (resp. haut) niveau de précision. En contrepartie, un bas niveau d'exhaustivité (resp. spécificité) conduit à un bas (resp. haut) niveau de rappel et à un haut (resp. bas) niveau de précision.

Selon Sparck Jones et Salton [60] [61] [99] [93] [100] [62] il existe un compromis optimal entre spécificité et exhaustivité. Ces auteurs ont proposé des interprétations statistiques qui sont fonction de la distribution des termes dans les documents qu'ils indexent et dans la collection. L'exhaustivité est liée au nombre de termes d'indexation affectés à un document. La spécificité est liée au nombre de documents indexés par un terme donné. Dans les modèles les plus connus de la RI, le poids des termes est le résultat d'une vue fréquentiste, et est basé sur la combinaison de la fréquence des termes dans le document (tf) et de la fréquence inverse du terme dans les documents de la collection (idf). tf est le nombre d'apparitions du terme dans le document et idf est le ratio entre le nombre de documents de la collection et le nombre de documents contenant le terme. idf est vu comme un indicateur de la spécificité et est doté d'un pouvoir discriminant pour distinguer entre les documents de la collection.

Dans [92] [97], les auteurs ont proposé une approche itérative pour attribuer un pouvoir discriminant aux termes [99] dans le but de distinguer les « bons » termes des « mauvais ». Un « bon » terme est un terme qui permet de distinguer entre les documents alors qu'un « mauvais » terme est un terme qui rend les documents *similaires ou non dissociables*.

1.3.3 Vue probabiliste

Dans [6] [51], les auteurs ont posé des hypothèses statistiques concernant la distribution empirique des termes dans la collection. Les auteurs s'intéressent à la distribution des mots pour déterminer si un mot donné doit être retenu comme terme d'indexation. Le point de départ de leur travail sont les travaux de Rubinoff, Damerau et Dennis, qui ont montré que le comportement statistique des mots *spéciaux* est différent de celui des mots *de fonction*. Un mot *spécial* est un mot qui véhicule du sens, différent des propositions, mots vides etc. Ceux-ci constituent les mots de fonction et suivent une distribution de Poisson contrairement aux mots spéciaux. Soit la distribution d'un mot de fonction (exemple : et) sur un ensemble de textes, alors la probabilité, $f(n)$, que n occurrences d'un mot de fonction apparaissent dans un texte donné est :

$$f(n) = \frac{e^{-x} x^n}{n!}$$

Le nombre d'occurrences n varie d'un mot à l'autre, et pour un mot donné, il est proportionnel à la longueur du texte ;

x est la moyenne des nombres d'apparitions du mot dans l'ensemble des textes (documents).

Selon [6], les mots spéciaux sont orientés contenu alors que les mots de fonction ne le sont pas. Ainsi, un mot ayant une distribution aléatoire, telle que la distribution de Poisson, est non informatif sur le contenu du document qui le contient, et un mot ne suivant pas cette loi est informatif. Leur modèle considère que le document traite d'un mot à un certain degré, et qu'il est possible de subdiviser les documents en sous ensembles traitant chacun d'un mot au même degré. L'hypothèse posée à partir de cette subdivision est qu'un mot orienté contenu permet de distinguer entre plusieurs classes de documents selon le degré auquel le sujet référencé par le mot est traité dans chaque classe de documents. Ces termes seront des termes d'indexation intéressants. Par contre, un mot discriminant entre différents sous ensembles de documents ne l'est plus à l'intérieur d'un sous ensemble de documents (ou une classe donnée de documents).

Harter, a fourni deux hypothèses pour l'indexation automatique, basée sur les idées précédemment citées :

1. la probabilité qu'un document soit pertinent à un besoin portant sur un sujet donné, est fonction du degré auquel le document traite du sujet ;
2. le nombre d'*objets* dans le document est fonction du degré auquel le document traite le sujet référencé par le mot.

Ainsi, la distribution d'un mot, qui n'appartient qu'à deux sous ensembles, à différents degrés, peut être décrite par le mélange de deux distributions de Poisson, communément nommée une distribution « 2-Poisson » ((i) celle liée au fait qu'un mot orienté contenu permet de distinguer entre plusieurs classes de documents et ceci d'une manière dépendante du degré auquel le document traite du mot et (ii) la seconde distribution découle de la seconde hypothèse qui spécifie qu'à l'intérieur d'un sous ensemble la distribution d'un mot orienté « contenu » peut suivre la loi de Poisson). Ainsi, si nous gardons les mêmes notations que précédemment :

$$f(n) = \frac{P_1 e^{-x_1} x_1^n}{n!} + \frac{(1 - P_1) e^{-x_2} x_2^n}{n!}$$

avec

P_1 : la probabilité d'appartenance d'un document (pris aléatoirement) à l'une des deux classes ;

x_1, x_2 : les moyenne des nombres d'apparitions dans les deux classes.

1.4 Les modèles connus de la RI

Nous présentons dans cette section, certains modèles proposés dans la littérature pour la RI. Cette présentation n'inclut pas tous les modèles existants mais uniquement ceux dont nous nous sommes inspirés pour nos travaux. D'une façon générale, les fondements mathématiques sur lesquels se basent les modèles de la littérature reposent sur l'utilisation de l'algèbre, la logique, la théorie de la probabilité et les statistiques. Ces modèles peuvent être répertoriés en trois catégories selon leur définition de la pertinence d'un document vis à vis d'une requête utilisateur [38] :

1. La pertinence est vue comme la similarité entre la requête et le document ([92] [109]);
2. La pertinence est modélisée par une variable aléatoire binaire et des modèles probabilistes sont utilisés pour calculer la valeur de la variable;
3. La pertinence est dotée d'une incertitude produite sur l'inférence de la requête à partir des documents ou inversement.

Nous décrivons dans cette section, deux modèles de la première et seconde catégorie. Nous détaillons particulièrement les méthodes de pondération utilisées lors de l'indexation ainsi que les méthodes d'appariement document-requête. Nous donnons à la fin de la présentation de chaque modèle un exemple de collection comportant des documents et une requête que nous avons choisie exhaustive pour dérouler les algorithmes de pondération et de recherche. La requête est dite *exhaustive* parce qu'elle contient à la fois des termes rares et fréquents, et apparaissant aussi bien dans des documents longs que courts. Ce déroulement de l'algorithme est comparé dans la dernière partie de ce manuscrit au modèle que nous proposons et nous essaierons de synthétiser les similitudes et les différences entre les modèles grâce à l'exemple.

1.4.1 Modèle vectoriel

Le modèle vectoriel fait partie de la première catégorie de modèle. En 1957, Luhn [70] a proposé de représenter les textes (documents et information recherchée) sous forme de vecteurs pondérés et de procéder à un calcul statistique. Au début des années 70, Salton [92] [98] a développé cette idée, en proposant le modèle SMART (Salton's Magical Automatic Retriever of Text). Dans ce modèle, le sens d'un document est donné par les mots qui y figurent. Un coefficient de similarité est calculé pour restituer les documents dont le contenu (termes) correspond au contenu de la requête. Une mesure de distance calcule l'angle entre les deux représentations.

Un document de la collection est représenté sous forme d'un vecteur de termes pondérés. Lorsque la collection contient T termes, un document D_j est défini par : $(w_{D_{1j}}, \dots, w_{D_{Tj}})$. w_{ij} donne le poids du terme t_i dans le document D_j . Pareillement, une requête Q est représentée par : $Q(w_{Q_1}, \dots, w_{Q_T})$. Une telle représentation des termes de la requête et des documents implique qu'un nombre important des termes de l'univers du discours auront pour valeur zéro dans les représentations. Ce nombre sera d'autant plus grand que la collection est diversifiée et que le nombre de termes de l'univers du discours est grand.

Plusieurs mesures de similarité ont été utilisées et étudiées. Nous présentons dans ce qui suit, certaines d'entre elles ainsi que les pondérations utilisées dans ce modèle.

1.4.1.1 Mesures de similarité

La pertinence est vue comme une mesure de similarité. Nous discutons dans ce qui suit certaines mesures utilisées pour calculer la pertinence. Chaque mesure a en fait sa spécificité et sa manière d'interpréter la pertinence.

Un simple coefficient de similarité (CS) entre une requête Q et un document D_j est donné par le produit des deux vecteurs :

$$CS(Q, D_j) = \sum_{i=1, \dots, T} w_{Q_i} * w_{D_{ij}} \quad (1.1)$$

avec :

w_{Q_i} : le poids du terme i dans la requête Q ,

$w_{D_{ij}}$: le poids du terme i dans le document j .

Lorsque les poids appartiennent à l'intervalle $[0, 1]$, CS mesure en fait la cardinalité de l'ensemble $Q \wedge D_j$. Ce coefficient est une mesure d'intersection entre les termes d'indexation des documents et ceux de la requête si l'on voit les vecteurs comme des ensembles flous. Selon la formule donnée en 1.1, il existe une symétrie stricte dans le calcul dans la mesure où la requête et le document jouent le même rôle. De plus, l'absence d'un terme de la requête est interprétée comme la négation de ce terme. D'autres méthodes de comparaison de vecteurs ont été implémentées. La plus connue est celle utilisant le cosinus de l'angle entre la requête et le document :

$$CS_{cos}(Q, D_j) = \frac{\sum_{k=1}^t w_{Q_k} * w_{D_{kj}}}{\sqrt{\sum_{k=1}^t (w_{D_{kj}})^2 \sum_{k=1}^t (w_{Q_k}^2)}} \quad (1.2)$$

La similarité par cosinus, CS_{cos} , normalise la mesure 1.1 atténuant ainsi la longueur du document. La similarité par CS (cf. équation 1.1) favorise les documents longs par rapport aux documents courts. En effet, le nombre d'apparitions d'un terme dans un document long a plus de chance d'être plus grand que dans un document court. Pour ces mesures, seuls les termes d'indexation de la requête sont pris en compte. Quelle que soit la mesure utilisée dans le modèle vectoriel, les termes d'indexation n'apparaissant pas dans la requête (ayant des poids nuls) peuvent être ignorés et la négation est stricte dans ce modèle. D'autres mesures géométriques ont été expérimentées dans ce modèle, telles que la distance euclidienne, le coefficient de Jaccard, de Dice etc. Salton et Buckley ont expérimenté leurs mesures en utilisant des représentations de documents plus complexes, en tenant compte des termes reliés, phrases, termes de thesaurus. Cependant, ces méthodes n'ont pas montré une nette amélioration [95]. Singhal a proposé d'appliquer une transformation de normalisation par rapport à la longueur du document pour « augmenter » les scores des documents longs et diminuer les scores des documents courts. En effet, la probabilité de pertinence calculée par la campagne TREC montre une linéarité avec la longueur des documents. Cette fonction de transformation, dite normalisation de pivot, contient deux paramètres le pivot et le gradient (slope). Ces paramètres sont appris sur une collection ¹. Ainsi, le calcul de similarité est donné par la

¹La normalisation de cosinus originale a complètement été abandonnée dans l'état de l'art actuel

formule empirique :

$$CS_{Normalisee}(Q, D_j) = \sum_{i=1}^{UT_Q} \frac{w_{Q_i}}{(1-s) \times p + s \times UT_Q} \times \frac{\left(\frac{w_{D_j}}{1+\log(Avg_{D_j})}\right)}{(1-s) \times p + s \times UT_{D_j}} \quad (1.3)$$

avec :

UT_Q et UT_{D_j} le nombre de termes ayant un nombre d'apparitions égal à 1 dans Q et D_j respectivement ;

Avg_{D_j} : le nombre moyen d'apparitions de termes dans le document D_j ;

p : le nombre moyen de termes dans les documents, $\left(\frac{\sum_{j=1}^N |D_j|}{N}\right)$, $|D_j|$ la cardinalité du document D_j : le nombre de termes ;

s : constante trouvée expérimentalement, fixée à 0.2.

Cette utilisation de la normalisation par pivot a placé le système *SMART* parmi les meilleurs systèmes dans la campagne d'évaluation *TREC - 3* [108] [110]

1.4.1.2 Pondération

Le modèle de base a considéré les poids comme dans le modèle booléen : 0 et 1 pour désigner respectivement l'absence et la présence d'un terme dans le vecteur. Plus tard, les poids ont été calculés à partir d'une combinaison de $tf * idf$. Dans [97][95], des expérimentations ont été effectuées pour améliorer l'évaluation des poids des termes. La mesure qui a semblé donner les meilleures performances pour un terme i d'un document j , notée W_{ij} , est :

$$W_{ij} = \frac{w_{ij}}{\sqrt{\sum_{k=1}^t [w_{kj}^2]}} \quad \text{avec } w_{ij} = (1 + \log tf_{ij}) * nidf_i \quad (1.4)$$

et $nidf_i = \log \frac{N+1}{n_i}$ ou $\log \left(\frac{N}{n_i}\right)$; tf_{ij} : nombre d'apparitions du terme i dans le document j et t le nombre de termes dans le document j ;

n_i : le nombre de documents contenant le terme t_i et N le nombre de documents de la collection.

La mesure 1.4 a été justifiée par le fait qu'une fréquence élevée d'un terme peut fausser la comparaison entre la requête et le document. Une autre variante serait de considérer les poids des termes de la requête différents de celui du document. Ainsi, les auteurs ont proposé d'utiliser cette mesure pour les termes

d'indexation de la requête en négligeant le *nidf* pour les termes d'indexation des documents. Le dénominateur est une normalisation de type cosinus, permettant de donner une chance égale à tous les documents. Sans cette normalisation, les documents les plus longs, avec les fréquences les plus élevées génèrent une similarité avec la requête plus forte que celle des documents courts. D'autres pondérations ont été normalisées, par exemple $\log(tf)$ a été utilisé à la place du simple tf initial.

1.4.1.3 Illustration

La petite collection de documents que nous présentons dans cette section va servir d'exemple pour tous les modèles qui suivent. Soit la collection suivante :

$$\begin{aligned} D_1 &= \{4t_1, 6t_4\}; & D_2 &= \{20t_2, 10t_3, 15t_5, 5t_6\}; & D_3 &= \{t_2, t_3, t_5\}; \\ D_4 &= \{t_2, 15t_3, 10t_5\}; & D_5 &= \{15t_1, 15t_2, 15t_3\}; \\ Q &= \{t_2, t_3, t_6\} \end{aligned}$$

Dans cet exemple de collection, le document D_1 contient les termes t_1 et t_4 . De plus, le terme t_1 y apparaît 4 fois et le terme t_4 a un nombre d'apparitions égal à 6. Nous avons essayé de diversifier la requête, en prenant des termes qui apparaissent souvent dans la collection comme t_2 , et des termes rares tel que le terme t_6 . Nous avons aussi essayé d'avoir des documents courts (des sommes de fréquences d'apparition par document relativement petites) ainsi que des documents longs. Ces documents longs peuvent contenir, relativement un grand nombre de termes, comme D_2 , ou bien ils ont des fréquences d'apparitions de termes élevées avec un nombre de termes différents relativement petit (D_5).

Soit l'univers des termes de la collection : $UT = \{t_1, t_2, t_3, t_4, t_5, t_6\}$. Le tableau (tableau 1.1) présente un récapitulatif des poids des termes de la requête communs aux documents, obtenus par application de la formule (1.4). Le *nidf* est pris égal à $\log \frac{N}{n}$. Le *nidf* n'affecte pas les poids des termes des documents. Remarque : Lorsque les documents contiennent des poids égaux pour plusieurs termes il est difficile de donner des importances différentes aux termes.

La formule par cosinus restitue les documents tels que classés dans le tableau 1.2. Le premier document restitué est D_2 etc.

L'appariement par calcul d'angle par cosinus, utilise une normalisation sur la distance *euclidienne* des longueurs. Cette normalisation conduit à l'obtention

TAB. 1.1 – Poids des termes

	t_2	t_3	t_6
w_{i1}	0	0	0
w_{i2}	2.30	2.00	1.69
w_{i3}	1	1	0
w_{i4}	1	2.17	0
w_{i5}	2.17	2.17	0
w_{Qi}	0.09	0.09	0.69

TAB. 1.2 – Classement des documents par cosinus

D_2
$D_3; D_5$
D_4

du même score pour les documents D_3 et D_5 .

La méthode de normalisation par pivot nécessite le calcul des variables suivantes : $p = 3$, $s = 0.2$; $1 + \log(\text{Avg}_{tf_{D_1}}) = 1.69$, $1 + \log(\text{Avg}_{tf_{D_2}}) = 2.09$, $1 + \log(\text{Avg}_{tf_{D_3}}) = 1$, $1 + \log(\text{Avg}_{tf_{D_4}}) = 1.93$ $1 + \log(\text{Avg}_{tf_{D_5}}) = 2.17$.

La méthode qui utilise la normalisation par pivot (1.3), attribue les poids aux termes d'indexation tels que présentés dans le tableau (1.3). Pour cette

TAB. 1.3 – Poids des termes obtenus par la méthode de pivot

	t_2	t_3	t_6
w_{i1}	0	0	0
w_{i2}	3.97	1.98	0.99
w_{i3}	0.33	0.33	0
w_{i4}	0.19	2.97	0
w_{i5}	2.87	2.87	0
w_{Qi}	0.03	0.03	0.23

pondération, nous avons utilisé tf car cela a semblé donner de meilleurs résultats que l'utilisation de $1 + \log(tf)$ [110]. Les poids des termes attribués par cette méthode ne tiennent pas compte de la répartition des termes dans la collection en fonction de leur densité de distribution mais uniquement de leur présence-

absence à travers les documents de la collection.

Le classement par la méthode *pivot* restitue les documents selon l'ordre spécifié dans le tableau 1.4. Le rang des documents comparé à l'appariement par co-

TAB. 1.4 – Classement des documents selon la normalisation par pivot

D_2
D_5
D_4
D_3

sinus a changé en privilégiant les documents longs. La prise en compte de la longueur des documents a amélioré les performances du système SMART [110] et l'a placé en compétition avec OKAPI (décrit dans la section suivante) ou INQUERY (décrit dans la seconde partie de ce manuscrit).

1.4.1.4 Conclusion

La pertinence d'un document par rapport à une requête est supposée, dans ce modèle, reliée à la similarité entre un document et une requête à un niveau donné de représentation : plus la représentation du document est similaire à celle de la requête et plus le document est considéré pertinent. Bien qu'il ait été montré que ce modèle peut reproduire d'autres modèles, (par exemple probabiliste [51]), l'étude des représentations des objets est séparée de l'estimation de la pertinence. La séparation entre la mesure de la pertinence et la pondération des termes rend le modèle plus flexible mais il est difficile d'étudier l'interaction entre la pertinence et la représentation des objets (document ou requête). Un dernier point que nous relevons concerne les termes considérés dans le calcul de la pertinence. Ce modèle ainsi que les modèles existants de la littérature ne considèrent que les termes de la requête présents dans les documents lors du calcul de la pertinence. Ainsi, le poids d'un terme t_i de la requête absent du document multiplie par 0 le poids de ce même terme dans le vecteur représentant le document lors du calcul de la pertinence du document. Nous rappelons que dans ce modèle la similarité est obtenue par le produit des poids des termes document-requête.

1.4.2 Stratégies de Recherche Probabilistes

Différents types de modèles probabilistes ont été proposés [83] [46] [87] [89]. Ils diffèrent principalement sur leur façon d'estimer la probabilité de pertinence. Dans ce qui suit, nous allons décrire en détail le fonctionnement du modèle proposé dans [87]. Le système OKAPI basé sur ce modèle figure parmi les systèmes les plus efficaces en termes de rappel-précision [88]. Ce modèle [87] [89] [63] permet de tenir compte de différents types de fichiers d'items (documents), de différents types de besoins utilisateurs et d'une variété de requêtes utilisateur. D'un point de vue du modèle, la description des objets (documents, requêtes) fait partie de l'environnement du système, et le rôle du modèle est d'aboutir aux meilleures descriptions finales (représentations documents et requêtes) qui sont considérées pour la recherche. La relation de pertinence est le point crucial de ce modèle, et les statistiques des termes (distribution des termes dans les documents pertinents-non pertinents) sont utilisées pour l'estimation des paramètres du modèle. La restitution des documents en réponse à un besoin utilisateur est basée sur le « principe de classement probabiliste » (Probability Ranking Principle, *PRP*), stipulant une meilleure performance du système lorsque les documents sont restitués par ordre décroissant de leur pertinence. La reformulation de requête a été proposée pour ce type de modèle.

1.4.2.1 Probabilité de pertinence

Dans ce qui suit les représentations initiales des documents et des requêtes sont nommées respectivement par D et Q . De plus, ces descriptions sont supposées décomposables en unités ou composantes plus petites. La requête Q est en fait une instance du besoin en information, sa formulation initiale telle que formulée par l'utilisateur et son expression telle que soumise au système. La probabilité de pertinence tente de répondre dans ce système, pour chaque document et chaque requête, à la question : *Quelle est la probabilité que ce document soit pertinent pour cette requête ?* Ainsi, deux événements sont possibles ² :

- L , D est pertinent pour Q ,

²d'où l'appellation **BIR : Binary Independance Retrieval**.

- \bar{L} , D est non pertinent pour Q .

La probabilité qu'un document D soit pertinent pour Q , notée $P(L/D)$, quelle que soit sa représentation, et en utilisant la règle de Bayes est donnée par :

$$P(L/D) = \frac{P(D/L)P(L)}{P(D)} \quad (1.5)$$

Cette équation nécessiterait un développement de $P(D)$ non utile, une simplification par l'ajout du *log – produit* sans que le principe *PRP* ne soit altéré, donne :

$$\begin{aligned} \log \frac{P(L/D)}{P(\bar{L}/D)} &= \log \frac{P(D/L)P(L)}{P(D/\bar{L})P(\bar{L})} \\ &= \log \frac{P(D/L)}{P(D/\bar{L})} + \log \frac{P(L)}{P(\bar{L})} \end{aligned} \quad (1.6)$$

Le score d'appariement entre le document D^3 et la requête, noté $MS-PRIM(D)$ [87], est donné par :

$$MS - PRIM(D) = \log \frac{P(L/D)}{P(\bar{L}/D)} - \log \frac{P(L)}{P(\bar{L})} \quad (1.7)$$

Le dernier terme, $\frac{P(L)}{P(\bar{L})}$, est le même pour tous les documents de la collection, un classement de documents avec $MS - PRIM$ revient à un classement de $P(L/D)^4$.

Nous décrivons dans ce qui suit les méthodes utilisées pour estimer les différentes variables utilisées par les modèles probabilistes. Nous décrivons particulièrement le modèle « d'Indépendance Binaire », connu sous le modèle *BIR*, puis le modèle de Poisson.

1.4.2.2 Modèle BIR

Le modèle général a été défini en considérant les attributs définissant l'univers du discours des documents de la collection comme indépendants. En effet, dans chaque classe de documents (chaque classe de documents est définie par

³ D est pris dans sa totalité, ses termes d'indexation ne sont pas considérés pour l'instant

⁴ qui devrait être supérieure à $P(\bar{L}/D)$

L, \bar{L}) chaque attribut est indépendant de tous les autres attributs. Ici, nous ne parlons pas de termes t_i , car comme nous l'avons dit plus haut, un terme est plutôt considéré comme un attribut A_i de document. Un document est donc composé d'attributs qui sont binaires et prennent leurs valeurs dans le domaine $\{0, 1\}$, signifiant que le terme est présent ou absent de la représentation d'un document donné. Dans chaque classe de documents (pertinents, non-pertinents), chaque attribut est statistiquement indépendant de tous les autres attributs. Ainsi, les probabilités de pertinence (non pertinence) d'un document, notées $P(D/L)$ (respectivement $P(D/\bar{L})$), sont données par :

$$P(D/L) = \prod_i P(A_i = a_i/L)$$

$$P(D/\bar{L}) = \prod_i P(A_i = a_i/\bar{L})$$

A_i est le i ème attribut utilisé pour décrire le document D , et a_i est sa valeur. Ainsi :

$$MS - PRIM(D) = \sum_i \log \frac{P(A_i = a_i/L)}{P(A_i = a_i/\bar{L})} \quad (1.8)$$

Cette équation implique que le score de chaque document est calculé par la somme des parties, chacune relatant l'attribut qui décrit le document.

Les attributs manipulés dans le modèle caractérisent des termes. Soit $p(A_i = 1 | L)$, notée par p_i ; et $p(A_i = 1 | \bar{L})$, notée par \bar{p}_i . La fonction W donne le poids, w_i , pour la présence de l'attribut i :

$$w_i = \log \frac{p_i(1 - \bar{p}_i)}{\bar{p}_i(1 - p_i)} \quad (1.9)$$

Le score d'un document est égal à la somme des poids des termes (présents) dans le document. Les termes considérés pour le calcul des scores sont en réalité les termes de la requête. Ce modèle ne peut donc être appliqué qu'après l'estimation des paramètres p_i, \bar{p}_i , c'est-à-dire, pour chaque terme de la requête, il s'agit d'estimer la probabilité que le terme apparaisse dans un document pertinent, respectivement non pertinent.

Nous présentons dans ce qui suit, les méthodes utilisées pour calculer ces estimations. Ces méthodes peuvent utiliser des informations disponibles concernant la pertinence des documents. Dans d'autres cas, elles proviennent de statistiques sur les textes (documents).

1.4.2.3 Pondération

Interpréter le modèle probabiliste général signifie utiliser les informations disponibles sur les distributions des termes et des documents. Les termes de la requête constituent le point de départ de l'estimation de la pertinence des documents, cette pertinence est quantifiée par la présence/absence des termes d'indexation de la requête dans les documents. En 1976, Robertson et Sparck Jones [84], ont proposé quatre méthodes concurrentes pour estimer ces paramètres. Le postulat et hypothèse qui ont permis de trouver les meilleures pondérations sont :

- *Hypothèse*) Les distributions des termes dans les documents pertinents sont indépendantes et leurs distributions dans les documents non pertinents sont indépendantes ;
- *Postulat*) La pertinence probable est basée sur la présence et l'absence des termes de la recherche (requête) dans les documents.

La présence des termes de la requête dans un document contribue au calcul de la pertinence des documents à restituer. Cette contribution va dépendre du nombre d'apparitions du terme dans le document ainsi que du nombre total de documents qu'il indexe.

Une première estimation des variables mesure l'importance d'un terme dans la collection. Le poids affecté au terme est égal à la fréquence inverse normalisée, $nidf_i$ du terme, t_i :

$$w_i = nidf_i \quad (1.10)$$

Ici $nidf_i = \log \frac{N}{n_i}$; N est le nombre de documents dans la collection et n_i le nombre de documents contenant le terme t_i .

Plusieurs travaux dans la littérature ont montré l'amélioration de la pertinence apportée par ce $nidf_i$ [86] [89]. Une seconde pondération proposée dans la littérature, donnant le poids du terme i , notée par w'_i , est :

$$w'_i = \log \frac{N - n_i + 0.5}{n_i + 0.5} \quad (1.11)$$

Comparée à la pondération par $tf \cdot idf$, la pondération ainsi définie (formules 1.10, 1.11) du modèle BIR est pauvre dans la mesure où ni la fréquence du terme dans le document ni la longueur du document ne sont pris en compte.

Il est impossible avec cette mesure de donner des importances différentes aux documents contenant les mêmes termes de la requête. En 1979, Croft et Harper ont intégré la fréquence du terme dans un document dans le calcul du score.

Lorsque des informations concernant les termes, et particulièrement des informations indiquant la pertinence des documents indexés par les termes de la requête existent, une table de contingence des termes est présentée dans le tableau 1.5.

Avec R : le nombre de documents pertinents pour la requête traitée ; et r_i :

TAB. 1.5 – Table de contingence des termes

	Pertinent	Non pertinent	
Contenant le terme t_i	r_i	$n_i - r_i$	n_i
Ne contenant pas le terme t_i	$R - r_i$	$N - n_i - R + r_i$	$N - n_i$
	R	$N - R$	N

le nombre de documents pertinents contenant le terme ; pour simplifier, nous notons, $p(A_i = 1 | L) = p$ et $p(A_i = 1 | \bar{L}) = \bar{p}$, alors :

$$p = \frac{r_i}{R} \text{ et } \bar{p} = \frac{n_i - r_i}{N - R}.$$

Ainsi la formule 1.9, donne la fonction de pondération d'un terme t_i par :

$$w_i'' = \log \frac{r_i(N - n_i - R + r_i)}{(R - r_i)(n_i - r_i)} \quad (1.12)$$

Toute instanciation du modèle engendre des problèmes et des incertitudes liés aux informations sur lesquelles se basent ces estimations et aux techniques d'estimation. Pour pallier les problèmes d'incertitude, les auteurs proposent d'ajouter la valeur 0.5 aux valeurs centrales de la table de contingence :

$$w_i''' = \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)} \quad (1.13)$$

1.4.2.4 Modèle de Poisson

En 1994, Robertson et Walker [87] ont proposé le modèle probabiliste basé sur la distribution de Poisson pour pallier les inconvénients du modèle BIR. L'appariement entre un document et une requête est obtenu par le produit des poids des termes de la requête communs à ceux du document. La pondération

d'un terme est représentée dans ce modèle par p_{tf} (et $\overline{p_{tf}}$), signifiant la probabilité qu'un terme apparaisse avec la fréquence tf dans un document pertinent (respectivement non pertinent). $p_0, \overline{p_0}$, représentent l'absence du terme dans un document pertinent et non pertinent respectivement.

$$w = \log \frac{p_{tf} \overline{p_0}}{\overline{p_{tf}} p_0} \quad (1.14)$$

La fréquence des termes a été mesurée par la distribution de Poisson [51]. Harter, [51], a défini la propriété de document *élite* pour un terme, lorsque ce document traite du sujet auquel le terme fait référence. Dans [86], la fréquence du terme dépend de cette propriété. Le cheminement suivi pour aboutir à la pondération du terme qui découle de cette propriété n'est pas détaillé dans ce manuscrit, mais le poids final affecté à un terme t_i dans un document d_j , noté w_{ij} , est donné par :

$$w_{ij} = \frac{tf_{ij}(k_1 + 1)}{k_1 + tf_{ij}} \log \frac{N - n_i + 0.5}{n_i + 0.5} \quad (1.15)$$

où tf_{ij} est la fréquence du terme t_i dans le document d_j ;

k_1 détermine à quel point le poids attribué réagit par rapport à une augmentation de tf . Si $k_1 = 0$ le terme réduit l'effet de la présence-absence du terme dans le document. Si k_1 est élevé le poids est proportionnel à tf . Dans la campagne d'évaluation *TREC 4*, les meilleurs résultats ont été obtenus lorsque $k_1 = 1.2$ et 2 . Le poids ainsi trouvé est égal à 0 lorsque $tf = 0$; il croît linéairement avec tf ; une limite asymptotique est donnée. La seconde modification apportée à cette pondération tient compte de la longueur des documents. La longueur des documents est prise en compte car l'utilisation de la loi de Poisson suppose une longueur égale pour tous les documents. De plus, un document (d_1) ne doit pas être préféré à un document (d_2) uniquement parce que la fréquence d'apparition d'un terme est plus grande dans (d_1) et principalement parce que (d_1) est plus long que le document (d_2). La longueur des documents est « uniformisée » dans la collection, en la normalisant par rapport à la longueur moyenne des documents. Ainsi, le poids d'un terme t_i dans d_j , w'_{ij} , est défini par :

$$w'_{ij} = \frac{tf_{ij}(k_1 + 1)}{k_1 \times \left((1 - b) + b \frac{l_{d_j}}{avg - l_{d_j}} \right) + tf_{ij}} \log \frac{N - n_i + 0.5}{n_i + 0.5} \quad (1.16)$$

avec :

l_{d_j} : la longueur du document d_j ; $\sum_{i \in L} tf_{ij}$, les auteurs ont aussi proposé de

mesurer en octets les longueurs des documents ;

$avg - l_{d_j}$: la longueur moyenne du document d_j ; $\sum_{j \in N} \sum_{i \in T} \frac{tf_{ij}}{N}$; N le nombre de documents de la collection ; T le nombre de termes de la collection. Les expérimentations ont montré que $b = 0.75$ donne des résultats de recherche satisfaisants.

Les fonctions de pondération proposées ont été nombreuses, et celle qui a donné les meilleurs résultats est celle utilisée dans OKAPI BM25 (*BM pour Best Match*). Le poids du terme t_i de la requête, noté w_{Qi} , tient compte du nombre d'apparitions du terme t_i dans la requête, tf_{Qi} , et d'un paramètre k_2 , de valeur égale à 8 (trouvée expérimentalement). Ainsi :

$$w_{Qi} = \frac{tf_{Qi} \times (k_2 + 1)}{k_2 \times tf_{Qi}} \quad (1.17)$$

1.4.2.5 Illustration

L'utilisation du modèle de Poisson tient compte des poids tels que définis dans les formules 1.16 et 1.17. L'appariement entre un document et une requête est obtenu par le produit des poids des termes de la requête communs à ceux du document. Le tableau 1.6, donne les poids des termes de l'exemple présenté dans la section 1.4.1.3. Pour cet exemple, nous avons substitué $\log \frac{N-n+0.5}{n+0.5}$ par $\log \frac{N}{n}$, parce que la première attribue des poids négatifs. Dans cette application, nous

TAB. 1.6 – Poids des termes

	t_2	t_3	t_6
w_{i1}	0	0	0
w_{i2}	0.27	0.24	1.52
w_{i3}	0.18	0.18	0
w_{i4}	0.13	0.27	0
w_{i5}	0.26	0.26	0
w_{Qi}	1.125	1.125	1.125

avons utilisé les valeurs suivantes : $k_1 = 2$; $k_2 = 8$; $b = 0.75$; $avg - l_d = 26.8$. Par exemple le poids du terme t_2 dans le document d_2 , de longueur égale à 50 est égale à $w_{22} = \frac{20 \times 3}{2 \times 0.75 + 0.75 \times (\frac{50}{26.8}) + 20} * 0.09691001 \approx 0.2655 \approx 0.27$ ⁵.

⁵ Toutes les valeurs données dans le tableau 1.6 sont données à deux chiffres après la virgule et arrondies au supérieur

Le classement des documents dans le tableau 1.7 est identique à celui obtenu par la méthode pivot du modèle vectoriel. *Un terme a un poids positif si le*

TAB. 1.7 – Classement des documents

D_2
D_5
D_4
D_3

nombre de documents de la collection qu'il n'indexe pas est plus grand que le nombre de documents qu'il indexe. De plus, pour restituer un document, il faut que l'influence des termes rares de la collection soit plus importante dans le document en question que celle des termes fréquents dans la collection. Cette rareté est ici uniquement quantifiée par le nombre d'apparitions du terme dans la collection d'une manière inversement proportionnelle à sa non apparition.

1.4.3 Autres modèles probabilistes

Différents modèles probabilistes *binaires* ont été proposés dans la littérature. Ces modèles sont dits binaires parce qu'un terme appartient à une classe de pertinence et que deux classes de pertinence sont définies. Ainsi, il existe la classe des documents pertinents en réponse à une requête utilisateur et une classe de documents non pertinents à cette même requête. La quantification de la pertinence d'un document ne pouvant pas être connue avec exactitude, les modèles probabilistes tentent généralement de l'estimer. Tous les modèles probabilistes binaires existant dans la littérature conçoivent la pertinence de la même manière, seule son estimation diffère d'un modèle à l'autre. La probabilité de pertinence (L) sachant un document (D) et une requête (Q), notée $P(L | D, Q)$ peut être estimée directement par un modèle de régression polynomiale [47]. Dans ce modèle, la pertinence est supposée liée aux caractéristiques de D et de Q . L'avantage de cette régression est qu'elle permet l'apprentissage de la pertinence à partir des jugements passés donnés sur les requêtes ou les documents. La seconde méthode se base sur la génération de modèles de documents ou de requêtes qui tentent d'estimer $P(D, Q | L)$ [85], [46], [67].

Les modèles probabilistes les plus connus sont les modèles proposés par [84],

[84], [83], [46]. L'intégration de la mixture de 2-*Poisson* [87] a permis d'ajouter les fréquences d'apparition des termes dans la représentation des documents. En 1977, Rijsbergen [82] a étendu le modèle d'indépendance binaire en prenant en considération les relations de dépendance entre les termes.

Le modèle de langue a été introduit par Ponte [78] et développé par [53] [5] entre autres, utilisant des estimations probabilistes. Sa principale caractéristique concerne l'estimation d'un modèle de langue basé sur le document. La pertinence d'un document face à une requête est en rapport avec la probabilité que la requête puisse être générée par le modèle de langue du document. Ainsi, un document D incarne un sous langage, pour lequel un modèle de langue est construit. Le score du document face à une requête Q est déterminé par la probabilité que son modèle génère la requête. Un corpus contenant tous les mots de la collection est généré et pour que ce corpus soit aussi complet que possible des opérations de lissage ont été proposées. La procédure de lissage consiste à attribuer une probabilité non nulle aux séquences de mots non encore rencontrés dans les documents. Les apports de ce modèle résident en l'introduction d'une fonction efficace de classement probabiliste basée sur la génération de requêtes. Cette génération de requêtes se base sur des calculs statistiques.

1.5 Techniques d'évaluation des SRIs

L'évaluation d'un système de recherche d'information permet entre autres de mesurer sa viabilité, son utilité pour des utilisateurs (ces utilisateurs doivent être convaincus de la nécessité de ce SRI par exemple), sa capacité à satisfaire l'utilisateur. Selon Cleverdon [22], il existe 6 quantités principales mesurables :

- L'univers de discours de la collection : le degré auquel le système inclut l'information pertinente ;
- le temps de réponse : temps moyen entre la formulation de la requête et la réponse donnée par le système ;
- la présentation de la sortie ;
- l'effort demandé à l'utilisateur ;
- le rappel du système : la proportion de documents pertinents réellement retournés en réponse à une requête utilisateur ;
- la précision : proportion de documents retournés réellement pertinents.

L'efficacité de tout SRI est mesurée principalement par le rappel et la précision. La définition de ces mesures est centrée autour de la notion de la pertinence d'un document étant donnée une requête. Il est donc important pour pouvoir mesurer l'efficacité des SRIs sur des collections de documents, d'avoir des ensembles de requêtes pour lesquelles les pertinences sont connues. Pour répondre à la question qui consiste à savoir si le SRI restitue tous les documents pertinents de la collection, des collections de tests ont été mises en place. Nous citons à titre d'exemple *CACM*, *CRANFIELD*, *TREC*. Dans ce contexte, la campagne d'évaluation, probablement la plus connue dans le domaine de RI, est la campagne *TREC* (Text REtrieval Conference) qui a débuté dans les années 90 avec 25 participants. *TREC* a permis de réunir des recherches communes sur l'ensemble des méthodologies de RI et offre des moyens de comparaisons entre différents systèmes. Différents aspects sont évalués. A titre d'exemple, nous citons le temps de réponse des SRIs, les ressources mises en place par les SRIs, la capacité des SRIs à restituer des documents pertinents, etc. Lorsque les collections ne sont pas trop volumineuses, des experts jugent pour des requêtes données les pertinences des documents manuellement. Lorsqu'elles sont volumineuses, d'autres approches sont utilisées comme par exemple l'échantillonnage, le *pooling*, ou s'appuyant sur la recherche. Le *pooling* consiste à considérer l'union des sous ensembles des n premiers documents restitués par différents systèmes comme pertinents. L'échantillonnage consiste à estimer la taille des ensembles de documents réellement pertinents. L'approche basée sur la recherche consiste à opérer une recherche guidée qui prend fin lorsque l'expert juge que tous les documents pertinents ont été trouvés. Nous énumérons, dans ce chapitre, les techniques proposées dans la littérature pour mesurer l'efficacité des SRIs.

1.5.1 Rappel, précision et *fall out*

Ces mesures évaluent la capacité du système à retourner les documents pertinents sans restituer les documents non pertinents. Généralement les SRI retournent les documents classés par ordre décroissant de leur pertinence. Ce classement ou rang est un rangement des documents en fonction de leur pertinence étant donnée une requête. Plusieurs travaux se sont penchés sur cette notion de pertinence [65] [13], affirmant la subjectivité, la gradualité (etc) de cette notion. En effet, des documents restitués en réponse à une requête donnée

peuvent être jugés différemment par les utilisateurs. Des expérimentations ont permis de regrouper les jugements de pertinence donnés par des utilisateurs. La pertinence, pour ces mesures est évaluée par le niveau auquel un document est à *propos* ou *approprié pour* la requête. L'efficacité d'un système mesure sa capacité à satisfaire l'utilisateur en terme de pertinence des documents restitués vis à vis d'une requête. Dans [83], un tableau de contingence permet de mesurer cette pertinence, en fonction des documents restitués et non restitués (*Restitués*; *Non Restitués* dans le tableau 1.8). A et B sont des ensembles

TAB. 1.8 – Tableau de contingence de la pertinence

	<i>Pertinent</i>	<i>Non Pertinent</i>	
<i>Restitués</i>	$A \cap B$	$\bar{A} \cap B$	B
<i>Non Restitués</i>	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	\bar{B}
	A	\bar{A}	N

de documents. Selon le tableau de contingence 1.8, nous pouvons définir les mesures de rappel, précision et *fallout* par :

$$Précision = \frac{|A \cap B|}{|B|} \quad Rappel = \frac{|A \cap B|}{|A|} \quad Fallout = \frac{|\bar{A} \cap B|}{|\bar{A}|}$$

où $|\bullet|$ désigne la cardinalité.

La fonction de *Généralité* est une fonction de relation entre ces 3 mesures, et est notée dans ce qui suit par G . Elle mesure la densité de documents pertinents de la collection. Cette relation est donnée par :

$$Précision = \frac{Rappel \times G}{(Rappel \times G) + Fallout(1 - G)} \quad G = \frac{|A|}{N}$$

Pour chaque requête, une de ces mesures est calculée. Généralement le *rappel* et la *précision* sont les mesures les plus utilisées dans la littérature. Pour chaque requête soumise à un système de recherche, un tableau de précision-rappel peut être construit. Le tableau 1.9 illustre les calculs de précision et rappel pour les 5 premiers documents trouvés pour une requête donnée.

La collection utilisée contient 2 document pertinents à cette requête. Le premier document retrouvé est pertinent, et donc la précision et le rappel pour

TAB. 1.9 – Exemple de calcul de rappel et précision pour les 5 premiers documents restitués

<i>Rang du document</i>	<i>Pertinence</i>	<i>Rappel</i>	<i>Précision</i>
1	<i>P</i>	$\frac{1}{2}$	$\frac{1}{1}$
2	<i>NP</i>	$\frac{1}{2}$	$\frac{1}{2}$
3	<i>NP</i>	$\frac{1}{2}$	$\frac{1}{3}$
4	<i>P</i>	$\frac{2}{2}$	$\frac{2}{4}$
5	<i>NP</i>	$\frac{2}{2}$	$\frac{2}{5}$

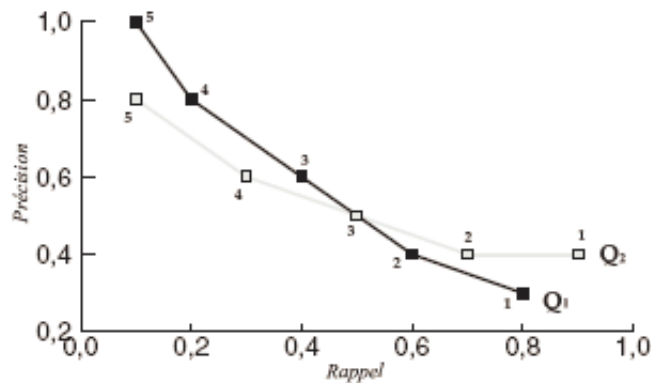


FIG. 1.3 – Courbes de précision-rappel pour deux requêtes Q_1 et Q_2

ce document valent $P = \frac{1}{1}$ et $R = \frac{1}{2}$.

Une façon d'évaluer un système est de tracer une courbe de précision-rappel. Ainsi, si le résultat de recherche dépend d'un certain paramètre, par exemple le rang d'un document restitué, alors pour chaque valeur du paramètre les valeurs de rappel et précision peuvent être calculées. Si λ est ce paramètre, alors P_λ exprime la précision, R_λ , le rappel, et la valeur de précision-rappel peut être représentée par le point (R_λ, P_λ) . Une liste ordonnée de paires précision-rappel peut être illustrée par une courbe, appelée courbe de précision-rappel. Le système parfait trouverait seulement les documents pertinents, avec une précision et un rappel de 100%. En pratique, ces deux taux varient en sens inverse, la précision diminue au fur et à mesure que le rappel augmente. La figure 1.3 est un exemple d'une courbe typique de précision-rappel de deux requêtes, où l'indice représente la valeur du paramètres λ .

La performance de chaque requête est généralement donnée par une courbe

de précision-rappel. L'ensemble des courbes, une pour chaque requête, sont combinées pour donner une courbe moyenne qui permet de mesurer les performances du système.

1.5.2 Interpolation

Une interpolation linéaire est opérée dans le but d'estimer la meilleure performance possible entre deux points adjacents observés. Soit (R_λ, P_λ) , l'ensemble des valeurs de rappel-précision obtenues en faisant varier le paramètre λ . Soit (R_θ, P_θ) un point observé tel que θ est une valeur de λ , pour laquelle une augmentation de rappel est produite. Ainsi, $G_s = (R_{\theta_s}, P_{\theta_s})$ est l'ensemble des points pour une requête. Pour une requête donnée, un point de rappel-précision est calculé pour chaque document pertinent. Pour interpoler deux points, on définit $P_s(R) = \{sup P : R' \geq R \text{ tq } (R', P) \in G_s\}$ tel que R est une valeur standard de rappel.

Pour chaque point de rappel standard défini, une précision interpolée est calculée. La valeur de la précision moyenne obtenue au point standard de rappel R , notée $P(R)$, est donnée par :

$$P(R) = \sum_{s \in S} \frac{P_s(R)}{|S|}$$

où $|S|$ est le nombre de requêtes et $P_s(R)$ est la précision de la requête s au niveau de rappel R . Comme les niveaux de rappel ne sont pas unifiés pour l'ensemble des requêtes, on retient dans la littérature, 11 points de rappel standards 0.00 à 1.00 à pas de 0.1.

L'ensemble des points observés est telle que la fonction d'interpolation est (monotone) décroissante (postulat trouvé expérimentalement) ce qui rend plus faciles les comparaisons entre courbes [83].

1.6 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 de la RI desquels nous nous sommes inspirés pour nos travaux. L'exemple que nous avons choisi dans les sections *Illustration* est utile pour pouvoir comparer, dans la dernière partie de ce manuscrit, les similitudes et les différences qui existent entre ces modèles et le modèle que nous proposons.

Les modèles proposés dans la littérature peuvent être répertoriés en 3 catégories [38] et ceci en fonction du sens de la pertinence définie :

1. la pertinence est en relation avec la similarité entre le document et la requête ;
2. la pertinence est modélisée par une variable aléatoire binaire et les modèles probabilistes tentent d'estimer la valeur de cette variable ;
3. l'incertitude intrinsèque à la pertinence est modélisée sur l'incertitude produite par l'inférence d'une requête à partir d'un document ou vice versa.

Nous avons présenté des modèles appartenant aux deux premières catégories dans la première partie de ce manuscrit à savoir le modèle vectoriel et les modèles probabilistes. Dans [114], il a été montré que le modèle booléen est un cas particulier de la troisième catégorie. Dans la deuxième partie de ce manuscrit nous discutons les modèles d'inférence probabilistes.

Deuxième partie

Réseaux Bayésiens et Recherche d'Information

Chapitre 2

Les Réseaux Bayésiens

2.1 Introduction

Les travaux présentés dans ce chapitre ont été utilisés dans divers domaines, notamment le diagnostic, la RI, l'intelligence artificielle, la théorie des probabilités et bien d'autres domaines. La modélisation graphique est un outil puissant pour la représentation et le traitement de la connaissance. La plupart des modèles graphiques, les Réseaux Bayésiens, les diagrammes d'influence, les chaînes de Markov, les arbres de décision, etc., s'appuient sur la théorie des probabilités.

Les Réseaux Bayésiens (RBs) sont utilisés dans divers domaines et constituent un outil puissant pour la représentation des connaissances et le traitement de l'incertitude. Un Réseau Bayésien (RB) est un graphe acyclique orienté dans lequel l'incertitude est mesurée par les probabilités marginales (noeud sans parent) ou conditionnelles. Un graphe est appréhendé selon un aspect qualitatif et un aspect quantitatif. L'aspect qualitatif est l'ensemble des noeuds du graphe représentant les variables du domaine traité ainsi que les relations d'indépendance entre ces variables. L'aspect quantitatif mesure les relations entre les variables au travers des arcs (liens orientés) les reliant.

Nous rappelons dans ce chapitre quelques notions de RBs. Nous donnons aussi les méthodes de calcul des probabilités conditionnelles dans ces réseaux.

2.2 Définitions

Soit $V = (A, B, ..)$ un ensemble fini de variables ;

$L = V \times V$;

$G = (V, L)$ est un graphe sur V et

L est l'ensemble des liens reliant une paire de variables de V .

Si le lien est orienté on parle alors d'arc et G est dit graphe orienté. Les variables représentent les événements ou propositions. Une variable peut avoir un nombre donné d'états. Par exemple, une variable peut décrire les couleurs possibles d'un objet, elle peut regrouper des maladies possibles $\{angine, grippe\}$. Les états peuvent être discrets ou continus. Les états d'une variable sont mutuellement exclusifs, la couleur d'un jeton ne peut pas être à la fois rouge et noire lorsque les états possibles sont $\{bleu, rouge, noir, blanc, jaune\}$. De plus,

- pour tout arc AB , A est l'origine ou le parent de B et B est le noeud final ou le fils de A ;
- un noeud racine est un noeud sans parent ;
- une feuille est un noeud sans fils ;
- un chemin est une séquence de noeuds reliés par des arcs ;
- une chaîne est une séquence de noeuds reliés par des liens ;
- un cycle est un chemin qui a le même noeud initial et final ;
- pour tout noeud $A \in V$ du graphe, $PARENTS_A$ est l'ensemble des parents de A .

2.3 Relations de dépendance

Les réseaux sont utiles pour calculer de façon locale l'impact de la modification d'une information d'une variable sur les états des autres variables. Le

changement d'un état d'une variable suite à la réception d'une information dans le réseau dépend de la topologie du graphe, et trois situations principales sont possibles.

Connexion en série Soit la situation de la figure 2.1. A a une influence sur B qui a une influence sur C . L'information peut circuler de A vers C ou de C vers A à travers B dans les deux cas. Par contre, si B est connue ou instanciée, la voie est bloquée et A et C deviennent indépendants. On dit dans ce cas, que A et C sont d-séparés étant donnée B , lorsque B est instanciée.

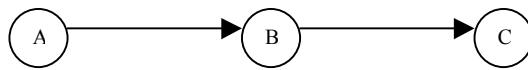


FIG. 2.1 – Connexion en série

Connexion divergente L'information peut passer entre les enfants de A lorsque la variable A est non instanciée. Dans la figure 2.2, les enfants B, C, D sont dits d-séparés par A .

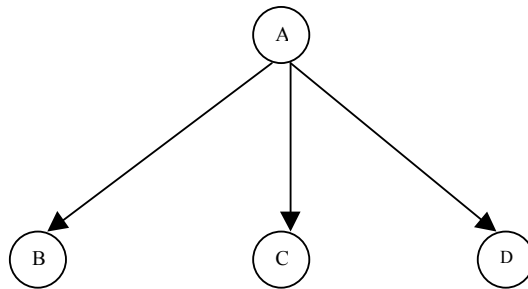


FIG. 2.2 – Connexion divergente

Connexion convergente Dans ce type de connexion telle que décrite dans la figure 2.3, lorsque aucune information n'est donnée sur le noeud fils mis à part l'information apportée par les parents, les parents sont dits dans ce cas indépendants. Par contre, si l'état du fils est connu alors la cause, c'est-à-dire un des états des parents va pouvoir donner de l'information sur les états des autres parents. L'information peut circuler dans une connexion convergente uniquement lorsque la variable de la connexion ou un de ses descendants a reçu de l'information.

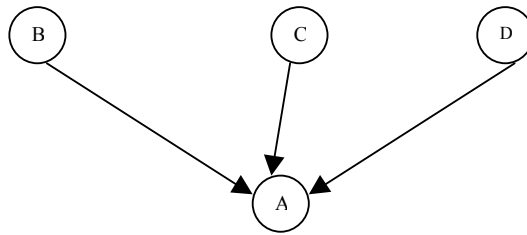


FIG. 2.3 – Connexion convergente

La d-séparation Les situations décrites ci-dessus recouvrent les manières possibles de transmettre l'information à travers un réseau. Deux variables distinctes A, B d'un réseau sont d-séparées, si pour tout chemin entre A et B , il existe une variable intermédiaire C , distincte de A et de B telle que :

- soit la connexion est en série ou divergente et C est instanciée ;
- ou la connexion est convergente et ni C , ni ses descendants ne sont instanciés.

Ainsi, si deux variables A et B sont d-séparées, alors tout changement d'état dans A n'aura pas d'impact sur l'état de B .

Un réseau Bayésien est doté d'une composante qualitative et d'une composante quantitative. Dans la section qui suit, nous donnons les techniques utilisées pour quantifier les liens existant entre toute paire de noeuds.

2.4 Calcul des probabilités

Les calculs de probabilité permettent de quantifier les liens reliant toute paire de noeuds du réseau. Plusieurs méthodes existent dans la littérature pour quantifier ces liens, nous nous restreignons dans cette partie au calcul Bayésien, à savoir les calculs classiques des probabilités.

2.4.1 Axiomes de base

La probabilité d'un événement A , notée $P(A)$ est un nombre de l'intervalle $[0, 1]$. Les probabilités obéissent aux axiomes suivants :

1. $P(A) = 1$ si A est certain ;
2. Si A et B sont mutuellement exclusifs, alors $P(A \cup B) = P(A) + P(B)$.

2.4.2 Probabilités conditionnelles

Les réseaux Bayésiens sont des modèles graphiques probabilistes permettant de représenter les influences entre des événements. Un réseau Bayésien est défini par un graphe acyclique orienté $G = (V, L)$. Dans ce graphe, V représente l'ensemble des noeuds du graphe et L l'ensemble des arcs reliant des paires de noeuds. Chaque noeud V_i représente une variable aléatoire associée à une distribution de probabilité, et chaque arc définit une influence du noeud de départ sur le noeud d'arrivée. La distribution de probabilité associée à une variable spécifie les probabilités de ses états conditionnellement aux états des variables qui l'influencent. On note $P(V_i | Parents(V_i))$ où $Parents(V_i)$ représentent l'ensemble des parents de la variable V_i . Ainsi

Définition Soit p une distribution de probabilité jointe sur un ensemble de variables V , et $G = (V, L)$ un graphe acyclique orienté. (G, p) est un réseau Bayésien si chaque variable $A \in V$ est conditionnellement indépendante de ses non descendants, noté $NONPARENTS_A$ étant donné l'ensemble de ses parents, $PARENTS_A$. Pour chaque variable A du graphe, les probabilités conditionnelles suivantes sont définies :

- Si $PARENTS_A = \emptyset$, ce qui signifie que le noeud A est un noeud racine, alors la probabilité *a priori* de A doit satisfaire

$$\sum_a P(a) = 1,$$

telle que a constitue l'ensemble des instances possibles de A

- Si $PARENTS_A \neq \emptyset$ alors la probabilité conditionnelle de A dans le contexte de ses parents est

$$\sum_a P(a | \theta_A) = 1,$$

où θ_A représente les instances possibles de l'ensemble des parents de A .

2.4.3 La règle de chaînage

Soit $V = (A_1, A_2, \dots, A_n)$ un ensemble de variables. La probabilité jointe $P(A_1, A_2, \dots, A_n)$ permet de calculer $P(A_i)$ et $P(A_i/c)$, telle que c est une information donnée. Le nombre et le temps de calculs effectués pour obtenir la probabilité $P(V)$ augmentent d'une manière exponentielle par rapport au nombre de variables contenues dans V . La règle de chaînage permet de calculer $P(V)$ d'une manière plus rapide lorsqu'il y a des dépendances entre les variables. Ainsi, la probabilité jointe est donnée par :

$$P(V) = \prod_{i=1}^n P(A_i \mid PARENTS_{A_i}) \quad (2.1)$$

où $PARENTS_{A_i}$ constitue l'ensemble des parents de A_i .

Chapitre 3

Les modèles de Recherche d'Information basés sur les Réseaux Bayésiens

3.1 Introduction

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. La composante qualitative du réseau permet de représenter les documents de la collection, les termes d'indexation ou concepts des documents ou de la requête, et du besoin utilisateur ou requête par des noeuds et de représenter les relations de dépendance (ou d'indépendance) existant entre ces variables par des arcs. L'aspect quantitatif du réseau permet d'évaluer les arcs reliant toute paire de noeuds au moyen de calcul de probabilités. La première utilisation des RBs en RI est apparue dans les années 80 [43], [44] mais s'est largement développée par les travaux de Turtle [112], [113]. D'autres travaux ont suivi [81], [107], [20], [35].

L'utilisation des RBs en RI a été un challenge à cause de deux principaux problèmes liés à leur utilisation :

1. le temps de calcul des distributions de probabilité et l'espace nécessaire à leur stockage augmentent d'une manière exponentielle avec le nombre de noeuds dans le réseau ;
2. la complexité de la propagation de l'information, c'est-à-dire les inférences nécessaires à propager l'information, dans un réseau est un problème NP-complet [25]¹.

Nous commençons en section 3.2 par survoler quelques modèles basés sur les réseaux de neurones pour montrer une modélisation graphique originale proposée pour la RI. De nombreux travaux utilisant une modélisation graphique de type RBs, ayant eu des succès différents, ont été définis. Nous décrivons les principaux modèles connus en sections 3.3 et 3.4. Nous nous intéressons particulièrement à la topologie de ces modèles ainsi qu'aux différents processus de propagation qui en découlent. Nous faisons une étude détaillée des pondérations lorsqu'elles sont différentes de celles des modèles présentés dans la première partie de ce manuscrit. Nous terminons ce chapitre par un bref état de l'art des modèles n'ayant pas eu un franc succès mais ayant certains aspects que nous jugeons pertinents que ce soit pour optimiser les calculs de propagation [56] [49] [1] [57], ou pour tenir compte des expressions pour indexer les documents [57] et enfin pour représenter les relations de dépendance entre les termes d'indexation des documents [34] [35] [36].

3.2 Modélisations graphiques en RI

Les réseaux Bayésiens n'ont pas été les seuls types de modélisation graphique proposés pour la RI. Nous citons à titre d'exemple les réseaux connexionnistes [66], [10], [29]. Ces réseaux sont similaires aux réseaux Bayésiens dans leur aspect graphique (noeuds et liens entre les noeuds), la quantification des liens ne se fait pas nécessairement dans un cadre probabiliste et la sémantique des liens ne traduit pas une relation de cause à effet. Ce type de modèle se base sur les fondements des réseaux de neurones biologiques, pour modéliser les documents et le besoin utilisateur et pour restituer les documents pertinents

¹Ceci parce que dans les réseaux *généraux*, il peut exister plusieurs chemins entre les paires de noeuds du graphe

par rapport à une requête donnée. Différentes granularités de l'abstraction d'un document peuvent être utilisées, les termes, les mots clés, les auteurs, etc. Le réseau de neurones est construit à partir des représentations initiales des documents et de la requête. L'évaluation des documents étant donnée la requête est fondée sur le principe de propagation à travers les connexions du réseau depuis les neurones représentant la requête vers les neurones documents. Les résultats sont présentés à l'utilisateur selon le niveau d'activation des neurones documents. Le modèle connexionniste est connu pour sa capacité d'apprentissage, ce qui permet aux SRIs d'être adaptatifs. Un aspect important de cette modélisation est la facilité de représenter les relations existantes entre termes (synonymie, co-occurrence, etc) et celles entre les documents (similitudes, référencement entre documents, etc). Une opération manuelle précédant la construction du modèle est nécessaire pour (i) définir les différentes couches du réseau (couches d'entrée, de sortie, intermédiaires etc) (ii) définir les neurones de chaque couche, (iii) définir la fonction d'entrée de chaque neurone, (iv) définir la fonction de sortie de chaque neurone, (v) définir les liens reliant les neurones ainsi que leur poids.

Les principaux modèles de RI sont représentés sous forme de couches. Les réseaux à couches [66] [10] sont constitués d'au moins deux couches (i) la première, appelée *couche d'entrée*, constituée des neurones descriptifs de la requête, (ii) la seconde, *couche de sortie*, constituée des neurones descriptifs des documents. D'autres couches, intermédiaires, ou *cachées* peuvent exister. Des poids sont attribués aux connexions qui sont orientées de la couche d'entrée vers la couche de sortie. L'évaluation des documents étant donnée une requête est fondée sur un processus d'activation ou propagation depuis les neurones de la couche d'entrée vers les neurones de la couche de sortie.

3.3 Le modèle inférentiel

La naissance du modèle d'inférence [111] [21] est le résultat de l'extension de deux idées : (i) la proposition d'utiliser des logiques non classiques pour déterminer le degré auquel un document implique ou correspond à une requête [115]; (ii) la notion d'inférence plausible et la possibilité de combiner plusieurs

sources pour inférer la probabilité de pertinence d'un document étant donnée une requête [31].

3.3.1 Architecture générale

Turtle [112], [111] a proposé un modèle de RI basé sur les réseaux d'inférence Bayésiens pour calculer la probabilité de pertinence d'un document étant donnée une requête. Ce modèle considère différentes représentations possibles des documents comme source d'information du contenu des documents et différentes représentations de la requête comme source d'information du besoin utilisateur. Deux composantes de réseau ont été définies : la première définit le réseau document ainsi que ses termes d'indexation et la seconde représente le besoin utilisateur et différentes représentations de la requête. La figure 3.1 décrit le modèle de base proposé. Le réseau document représente les noeuds documents

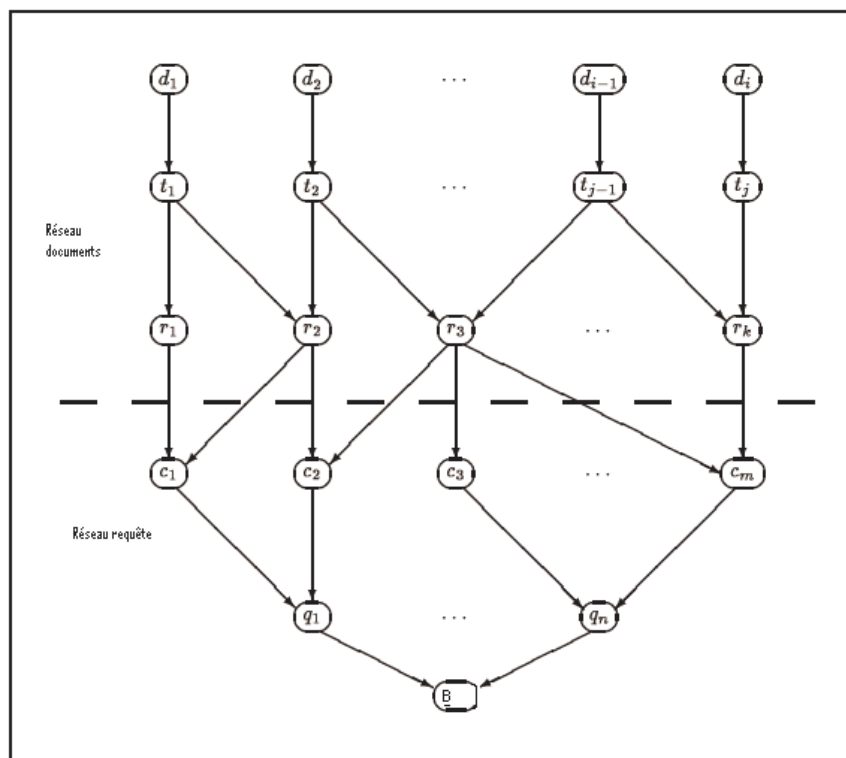


FIG. 3.1 – Architecture générale

(d_j) de la collection, les noeuds de représentation des textes (t_i), les noeuds de

représentation des concepts (r_k). Un noeud document correspond à l'événement qu'un document donné de la collection est observé. Un noeud de représentation de texte correspond au texte indexant le document et correspond à l'événement qu'un texte d'un document est rencontré. Les auteurs ont considéré dans leur approche de base les documents de type textes bruts mais suggèrent l'extension aux figures, images, aux documents multimédias etc. Il existe une correspondance 1 – 1 entre chaque représentation de texte et le document auquel il se réfère.

La dépendance entre un document et un texte est symbolisée par un arc entre les noeuds document et texte. Les noeuds de représentation de concepts correspondent à différentes techniques d'indexation utilisées pour obtenir les concepts d'indexation des documents comme par exemple une indexation automatique et une indexation manuelle. Un même concept peut ainsi être généré par les deux techniques d'indexation, et l'arc reliant dans ce contexte le concept au document aura deux sens différents.

Les domaines de tous les noeuds sont binaires $\{vrai, faux\}$ signifiant que le noeud est instancié ou non. Par exemple, un noeud représentant le texte aura pour instantiation *vrai* uniquement lorsque son noeud parent, document, est aussi instancié à *vrai*.

Le réseau requête est un graphe acyclique orienté représentant le besoin utilisateur (B) et des noeuds racines qui représentent les concepts de ce besoin (c_i). Plusieurs expressions de la requête peuvent être utilisées et représentées dans ce réseau (q_k). Les auteurs suggèrent de simplifier cette représentation en supprimant ces noeuds et de répercuter leur signification sur le noeud global B .

La valeur d'instanciation du noeud B est *vrai* lorsqu'elle désigne qu'un besoin utilisateur est rencontré dans un document. Le domaine des noeuds q_k est vrai pour désigner que la représentation de la requête est satisfaite. L'apport le plus important de ce modèle a été de pouvoir combiner l'information provenant de représentations différentes de documents ainsi que de combiner différentes formulations de la requête.

Une simplification du réseau a été proposée [111] et exposée dans la figure 3.2. Dans cette topologie, les noeuds documents sont des noeuds racine et il existe une relation entre les termes d'indexation et les documents ou la requête pour désigner que l'objet (document ou requête) contient le terme.

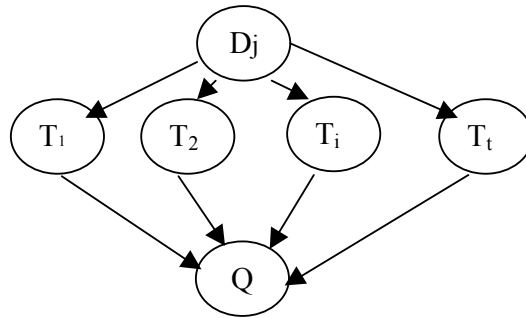


FIG. 3.2 – Architecture simplifiée

Soit D_j un document dont le domaine est $dom(D_j) = \{d_j, \bar{d}_j\}$.

Un terme d'indexation est référencé par T_i dans la figure, et le domaine des termes, noté $dom(T_i)$, est $dom(T_i) = \{t_i, \bar{t}_i\}$.

Le domaine de la requête est $dom(Q) = \{q, \bar{q}\}$. Nous ne sommes intéressés que par le cas où la requête introduit de l'information à travers le réseau, c'est-à-dire lorsque le noeud est instancié positivement, et nous noterons Q indifféremment lorsque cela ne prête pas à confusion.

3.3.2 Calcul de la pertinence

Le calcul de la pertinence revient dans ce modèle à instancier chaque document de la collection et à calculer la croyance de satisfaire la requête étant donné le document instancié. Le réseau pris dans sa globalité représente les dépendances qui existent entre une requête et les documents de la collection.

Un seul document est instancié positivement ($D_j = d_j$) à la fois. La propagation de l'information est déclenchée par cette instanciation.

La propagation dans ce modèle consiste à calculer pour chaque noeud la probabilité *a posteriori* étant données les probabilités *a priori* conditionnelles et marginales. La propagation tente de calculer la probabilité que l'information a été rencontrée étant donné un document instancié à $D_j = d_j$. Ce processus est réitéré pour tous les documents de la collection. Une liste des documents ordonnés par ordre décroissant de pertinence est restituée.

La probabilité conditionnelle d'un noeud est fonction de toutes les configurations possibles de ses noeuds parents. Soit θ l'ensemble des configurations

possibles des parents de Q , et θ_i^j une instance d'un noeud particulier T_i telle que dans la configuration de θ^j de θ . Par exemple, soit la requête Q composée des deux termes T_1 et T_2 , $Q = \{T_1, T_2\}$; alors l'ensemble des configurations possibles des parents de la requête, tel que leur domaine est binaire, est $\theta = \{\{t_1, t_2\}, \{t_1, \bar{t}_2\}, \{\bar{t}_1, t_2\}, \{\bar{t}_1, \bar{t}_2\}\}$. L'instance θ_1^1 du terme T_1 dans la première configuration de θ , $\theta^1 = \{t_1, t_2\}$, est $\theta_1^1 = t_1$.

La propagation dans le réseau dont la topologie est donnée dans la figure 3.2 est :

$$P(Q | d_j) = \sum_{\forall \theta^k \in \theta} (P(Q | \theta^k) \cdot \prod_{T_i \in Q \wedge D_j} P(\theta_i^k | d_j) \cdot P(d_j)) \quad (3.1)$$

La quantification totale de la pertinence revient à quantifier chaque membre de la formule 3.1. Des probabilités *a priori* sont affectées aux documents de la collection, égales à $P(D_j = d_j) = \frac{1}{N}$, mais elles sont supprimées du calcul de la propagation globale parce que ce terme est considéré comme un coefficient uniforme appliqué à tous les documents de la collection.

La section suivante (3.3.3) décrit le traitement de la requête dans ce modèle. Il s'agit particulièrement des diverses possibilités utilisées pour connecter ses termes. Par la suite, nous donnons les pondérations attribuées aux termes d'indexation des documents (3.3.4).

3.3.3 Agrégation de la requête

Turtle a proposé cinq formes canoniques pouvant répondre à tout type de recherche. La requête peut être agrégée par les opérateurs booléens (*ET*, *OU*, et *NON*). D'autre part, l'utilisation des réseaux permet d'agréger la requête par la somme probabiliste ou une de ses variations la somme pondérée. Pour évaluer les probabilités conditionnelles $P(Q | \theta)$ d'un noeud Q ayant n parents, $\{\theta_1, \dots, \theta_n\}$, et, $P(\theta_1 = t_1) = p_1, \dots, P(\theta_n = t_n) = p_n$ les agrégations suivantes sont définies :

$$\begin{aligned} P_{Ou}(Q | \theta) &= 1 - (1 - p_1) - \dots - (1 - p_n) \\ P_{Et}(Q | \theta) &= p_1 \times \dots \times p_n \\ P_{Non}(Q | \theta_1) &= 1 - p_1 \\ P_{Somme}(Q | \theta) &= \frac{p_1 + \dots + p_n}{n} \\ P_{SommePonderee}(Q | \theta) &= \frac{(w_1 p_1 + \dots + w_n p_n) w_q}{w_1 + \dots + w_n} \end{aligned} \quad (3.2)$$

Nous remarquons que l'opérateur de négation *Non* est unaire. Lorsque la négation d'un terme est spécifiée dans la requête, la quantification de sa présence dans le document est obtenue par $1 - p_i$ telle que décrite par la formule 3.2. Ici, la négation du terme n'est pas son absence de la représentation du document lorsqu'il est spécifié dans la requête. Les termes de la requête absents des représentations des documents ne sont pas considérés dans le calcul de la pertinence d'un document en réponse à une requête. La somme probabiliste tient compte du nombre de parents instanciés positivement dans la configuration des parents ($|\theta_j = t_j|$) et la somme pondérée mesure la configuration positive en fonction du poids de chaque parent instancié positivement, ainsi que du poids de la requête w_q . Le poids utilisé peut être le facteur de discrimination *idf* ou une de ses variantes ou un poids attribué par l'utilisateur. Ces deux dernières techniques d'agrégation permettent un gain de temps lors des calculs de la pertinence puisque uniquement les termes présents dans une configuration (documents) sont considérés.

3.3.4 Pondération des arcs $P(T_i | D_j)$

Les arcs reliant les termes d'indexation aux documents sont pondérés par des variantes de *tf - idf*. Dans [111], une probabilité égale à 0.5 est affectée à un événement lorsque aucune information ne vient contredire ou renforcer sa véracité. La probabilité d'un événement égale à 0 donne la certitude que l'événement est faux, alors que si sa probabilité vaut 1 alors il est certain que l'événement est vrai. Ainsi :

$$P(t_i | d_j) = 0.5 + 0.5 * nt f_{ij} * nd f_i$$

$$nt f_{ij} = \frac{t f_{ij}}{\max_{r_k \in d_j} t f_{kj}} \quad nd f_i = \frac{\log(\frac{N}{n_i})}{\log(N)}$$

$$P(t_i | \overline{PARENTS_{T_i}}) = 0$$

$$P(\bar{t}_i | d_j) = 1 - P(t_i | d_j)$$

où $\overline{PARENTS_{T_i}}$ signifie que tous les parents de T_i sont instanciés à faux ; N est le nombre de documents de la collection, et n_i le nombre de documents contenant le terme t_i .

3.3.5 Illustration

Nous rappelons l'exemple donné en *Chapitre 1* :

$$D_1 = \{4t_1, 6t_4\}; \quad D_2 = \{20t_2, 10t_3, 15t_5, 5t_6\}; \quad D_3 = \{t_2, t_3, t_5\};$$

$$D_4 = \{t_2, 15t_3, 10t_5\}; \quad D_5 = \{15t_1, 15t_2, 15t_3\};$$

$$Q = \{t_2, t_3, t_6\}$$

Les tableaux 3.1, 3.2 et 3.3 illustrent les probabilités conditionnelles des termes de la requête, en fonction de l'instanciation du document qui les contient.

Nous rappelons qu'un seul document est instancié à la fois. Dans les tableaux

TAB. 3.1 – Probabilité conditionnelle du terme T_2 , $P(T_2 | D_j)$

	d_2	d_3	d_4	d_5
t_2	0.569	0.569	0.504	0.569
\bar{t}_2	0.431	0.431	0.496	0.431

des probabilités conditionnelles des termes, lorsque le document D_2 par exemple est instancié positivement, $D_2 = d_2$, alors tous les autres parents du terme sont instanciés à \bar{d}_j ; $j = 3, 4, 5$. Le terme T_6 n'indexe que le document D_2 . Pour

TAB. 3.2 – Probabilité conditionnelle du terme T_3 , $P(T_3 | D_j)$

	d_2	d_3	d_4	d_5
t_3	0.535	0.569	0.569	0.569
\bar{t}_3	0.465	0.431	0.431	0.431

TAB. 3.3 – Probabilité conditionnelle du terme T_6 , $P(T_6 | D_j)$

	d_2
t_6	0.625
\bar{t}_6	0.375

tout document ne le contenant pas, on a $P(t_6 | d_j) = 0$; et $P(\bar{t}_6 | d_j) = 1$.

Lorsque la requête est soumise au système, chaque document de la collection est instancié pour calculer la probabilité de satisfaction de la requête étant

donnée une instanciation d'un document. Un seul document est instancié à la fois.

Supposons comme dans [111], que plus le nombre de termes parents de la requête est grand, et plus la probabilité conditionnelle de la requête sachant la configuration de ses parents est grande. Rappelons que les parents de la requête sont dans ce cas les termes d'indexation présents dans la requête.

Le tableau 3.4 présente les probabilités conditionnelles de la requête étant données les 2^3 configurations possibles. Nous rappelons que 3 est le nombre de termes de la requête et que chaque terme a un domaine binaire. La croyance

TAB. 3.4 – Probabilité conditionnelle de la requête $P(Q | \theta)$

	q	\bar{q}
t_2, t_3, t_6	0.9	0.1
t_2, \bar{t}_3, t_6	0.7	0.3
\bar{t}_2, t_3, t_6	0.7	0.3
$\bar{t}_2, \bar{t}_3, t_6$	0.5	0.5
t_2, t_3, \bar{t}_6	0.5	0.5
$\bar{t}_2, t_3, \bar{t}_6$	0.3	0.7
$t_2, \bar{t}_3, \bar{t}_6$	0.3	0.7
$\bar{t}_2, \bar{t}_3, \bar{t}_6$	0.1	0.9

(de la satisfaction) de la requête Q lorsque le document D_2 est instancié est calculée par :

$$\begin{aligned}
 P(Q | d_2) = & (\\
 & P(Q | t_2 t_3 t_6) P(t_2 | d_2) P(t_3 | d_2) P(t_6 | d_2) + \\
 & P(Q | t_2 t_3 \bar{t}_6) P(t_2 | d_2) P(t_3 | d_2) P(\bar{t}_6 | d_2) + \\
 & P(Q | t_2 \bar{t}_3 t_6) P(t_2 | d_2) P(\bar{t}_3 | d_2) P(t_6 | d_2) + \\
 & P(Q | t_2 \bar{t}_3 \bar{t}_6) P(t_2 | d_2) P(\bar{t}_3 | d_2) P(\bar{t}_6 | d_2) + \\
 & P(Q | \bar{t}_2 t_3 t_6) P(\bar{t}_2 | d_2) P(t_3 | d_2) P(t_6 | d_2) + \\
 & P(Q | \bar{t}_2 t_3 \bar{t}_6) P(\bar{t}_2 | d_2) P(t_3 | d_2) P(\bar{t}_6 | d_2) + \\
 & P(Q | \bar{t}_2 \bar{t}_3 t_6) P(\bar{t}_2 | d_2) P(\bar{t}_3 | d_2) P(t_6 | d_2) + \\
 & P(Q | \bar{t}_2 \bar{t}_3 \bar{t}_6) P(\bar{t}_2 | d_2) P(\bar{t}_3 | d_2) P(\bar{t}_6 | d_2) \\
 &) \tag{3.3}
 \end{aligned}$$

ainsi en remplaçant par les valeurs, nous obtenons :

$$\begin{aligned}
 P(Q | d_2) = (& \\
 & 0.9 * 0.569 * 0.535 * 0.625 + \\
 & 0.7 * 0.569 * 0.465 * 0.625 + \\
 & 0.7 * 0.431 * 0.535 * 0.625 + \\
 & 0.5 * 0.431 * 0.465 * 0.625 + \\
 & 0.5 * 0.569 * 0.535 * 0.375 + \\
 & 0.3 * 0.431 * 0.535 * 0.375 + \\
 & 0.3 * 0.569 * 0.465 * 0.375 + \\
 & 0.1 * 0.431 * 0.465 * 0.375 \\
 &) \tag{3.4}
 \end{aligned}$$

Le classement des documents restitués est donné dans dans le tableau 3.5

Le document le plus pertinent est le document D_2 . Le résultat n'est pas

TAB. 3.5 – Classement des documents selon la propagation dans le modèle d'inférence

D_2
D_3, D_5
D_4

étonnant dans la mesure où ce document est le seul à contenir tous les termes de la requête et spécialement le terme T_6 qui est un terme de poids élevé. Ce poids élevé s'explique par la rareté du terme dans la collection donc la valeur de son facteur *idf*. Ainsi ce terme est intéressant pour discriminer entre les documents. Les documents D_3 et D_5 ont le même rang de classement. Ce résultat est principalement lié au fait que la fréquence est normalisée à l'intérieur de chaque document. De plus, la longueur des documents n'est pas prise en compte dans le calcul des pertinences. Ces deux documents peuvent être considérés comme *identiques* puisque les poids des termes considérés dans le calcul de la pertinence sont les mêmes. Le document D_4 est restitué en dernière position. Le poids le plus élevé est celui du terme T_3 . Cependant, les meilleures configurations des termes de la requête ne sont pas celles qui contiennent le terme T_3 . De plus, ce document ne contient pas le terme T_6 . La configuration t_2, t_3 ajoute un plus faible poids à la pertinence du document D_4 qu'à celle des documents

D_2, D_3, D_5 .

Des travaux plus récents, testés dans les campagnes d'évaluation *TREC3* ont tenu compte de la longueur des documents en reproduisant les caractéristiques du modèle probabiliste.

Ce modèle [113] [111] permet la généralisation des modèles booléens et probabilistes. De plus, il permet de répondre aussi bien à des requêtes « évoluées » qu'à la combinaison de phrases. Les auteurs ont proposé un processus de reformulation de requêtes.

3.4 Le modèle de croyance

La seconde principale utilisation des réseaux concerne le modèle basé sur les réseaux dits « de croyance » [81] [2]. Ce modèle est basé sur la définition préalable d'un espace d'échantillonnage qui permet de séparer clairement les portions de documents des portions de requêtes et donc de calculer d'une manière « efficace » [2] les degrés de croyance. Il permet de généraliser tous les modèles classiques de RI (booléens, probabilistes et vectoriels) ainsi que de générer les résultats trouvés par le modèle inférentiel [2]. L'avantage de la généralisation est qu'elle permet de combiner les caractéristiques des modèles classiques les considérant comme sources d'information.

Les relations de dépendance définies dans ce modèle diffèrent de celles de Turlte. Dans ce modèle le processus de recherche est déclenché par la réception de la requête.

3.4.1 Architecture générale

L'architecture générale du modèle de croyance est présentée dans la figure 3.3. L'univers de discours est donné par l'ensemble des termes d'indexation utilisés dans le système, noté U , et $U = \{T_1, \dots, T_T\}$ où T est le nombre de termes manipulés dans le système (pour représenter les documents ou la requête). θ est l'ensemble des configurations possibles sur U , donc 2^T configurations dans U sont possibles.

La définition des domaines d'un noeud terme donné est similaire à celle du

modèle de Turtle. Un terme appartient ou non à un concept. Un concept peut être une requête ou un document. Les termes d'indexation pointent vers les documents et la requête qu'ils indexent.

Un document D_j , est une variable aléatoire de domaine binaire, $dom(D_j) = \{d_j, \bar{d}_j\}$. Un document D_j de la collection est instancié à $D_j = d_j$ pour indiquer que le document couvre complètement U .

Une variable aléatoire binaire est associée à une requête Q de domaine $dom(Q) = \{q, \bar{q}\}$. $Q = q$ signifie que la requête couvre complètement l'espace des termes. La couverture de l'espace U par un concept (document ou requête) est la conformité du concept avec chaque élément de l'espace U .

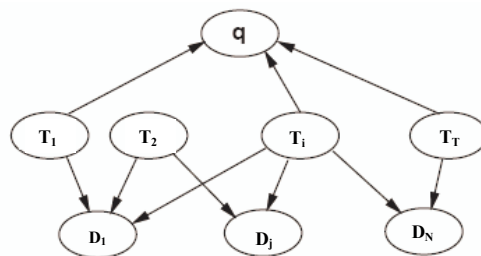


FIG. 3.3 – Architecture générale

A la réception d'un besoin utilisateur, la requête Q est instanciée et le processus de propagation est déclenché.

3.4.2 Calcul de la pertinence

Selon la topologie donnée par la figure 3.3, l'instanciation de la requête permet de calculer la probabilité de pertinence d'un document étant donnée une requête, $P(D_j | Q)$, donnée par la formule 3.5.

$$P(D_j | Q) = \frac{P(D_j \wedge Q)}{P(Q)} \quad (3.5)$$

La probabilité $P(Q)$ est calculée pour tous les documents de la collection, et est considérée comme une constante. Ainsi, une approximation possible de la

probabilité d'un document étant donnée une requête peut être

$$P(D_j | Q) \propto P(D_j \wedge Q) \quad (3.6)$$

D'après la topologie du réseau, l'instanciation des termes d'indexation (ici, cas de *d-séparation*) rend les noeuds documents et requête indépendants. Ainsi :

$$P(D_j | Q) \propto \sum_{\theta} P(D_j | \theta) \times P(Q | \theta) \times P(\theta) \quad (3.7)$$

θ représente l'ensemble des configurations possibles des termes de l'univers U . Ce modèle généralise les modèles classiques de la RI [2]. Nous donnons ici le traitement opéré sur le calcul de la propagation pour reproduire le modèle vectoriel. $\vec{d} = \{t_1, \dots, t_T\}$ et $\vec{q} = \{t'_1, \dots, t'_T\}$ désignent respectivement le vecteur document et le vecteur requête. Pour chaque document, une similarité par cosinus [94] entre un document et une requête est calculée. La probabilité d'un document étant donnée une configuration d'un concept est approximée par le produit entre les poids des termes du document et de la requête. Ainsi :

$$sim(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

où $|\cdot|$ désigne la cardinalité. Nous présentons dans ce qui suit, les mesures proposées pour calculer les probabilités conditionnelles d'un document donné (D_j) et de la requête (Q) étant donnés leurs parents respectifs, notés Par_{D_j} et Par_Q .

3.4.3 Probabilité des documents $P(D_j | Par_{D_j})$

L'univers de discours U définit l'ensemble des termes d'indexation du système. La probabilité $P(d_j)$ donne le degré auquel le document D_j couvre complètement l'espace des termes U . Cette couverture est calculée en contrastant chaque élément de U avec le document D_j , à travers $P(D_j | \theta)$ et en additionnant les contributions de chacun. Cette somme est pondérée par la probabilité $P(\theta)$ avec laquelle θ apparaît dans U . Cette probabilité répondrait à la croyance associée à la proposition *Est-il vrai que d_j couvre complètement U ?* et est donnée par :

$$P(d_j) = \sum_{\theta} P(d_j | \theta)P(\theta)$$

$$P(\theta) = \left(\frac{1}{2}\right)^T$$

De plus, toujours dans le contexte vectoriel décrit dans la section ci-dessus, nous avons :

$$P(d_j | \theta^Q) = \frac{\sum_{\theta_i^Q=1}^T w_{ij} \times w_{iq}}{\sqrt{\sum_{\theta_i^Q=1}^T w_{ij}^2} \times \sqrt{\sum_{\theta_i^Q=1}^T w_{iq}^2}}$$

$$P(\bar{d}_j | \theta^Q) = 1 - P(d_j | \theta^Q)$$

où θ_i^Q est la configuration des termes telle que donnée dans la requête Q , et w_{ij} , w_{iq} les poids du terme t_i dans le document d_j et la requête Q respectivement. Les poids w_{ij} sont des variantes de la pondération par *tf * idf* [95].

3.4.4 Probabilité de la requête $P(Q | Par_Q)$

La probabilité $P(Q)$ donne le degré auquel la requête couvre complètement l'espace des termes U . Cette probabilité répondrait à la croyance associée à la proposition *Est-il vrai que Q couvre complètement U ?*

$$P(Q) = \sum_{\theta} P(Q | \theta)P(\theta)$$

$$P(\theta) = \left(\frac{1}{2}\right)^T$$

Le calcul de cette équation nécessiterait 2^T calculs, où T est le nombre de termes manipulés par le système, mais en réalité uniquement les termes indexant la requête sont considérés.

La valeur 1 est attribuée aux arcs reliant les termes d'indexation à la requête lorsque tous les termes présents dans la requête sont instanciés positivement dans une configuration donnée des parents. Ainsi :

$$P(Q | \theta) = \left. \begin{aligned} &\{1 \text{ si } \forall T_i, \theta_i^Q = \theta_i \\ &= 0 \text{ sinon } \} \end{aligned} \right\}$$

$$P(\bar{Q} | \theta) = 1 - P(Q | \theta)$$

où θ_i^Q , θ_i l'instanciation du terme T_i dans la requête et dans θ respectivement.

3.4.5 Généralisation des modèles classiques

Un des apports majeurs de ce modèle a été de généraliser les modèles classiques de RI, à savoir les modèles booléens, probabilistes, vectoriels et de pouvoir reproduire le classement donné par le modèle inférentiel. Dans [81], un exemple de ces généralisations est donné.

3.5 Autres modèles basés sur les réseaux Bayésiens

Des extensions des deux modèles basés sur les réseaux Bayésiens ont été proposées [55] [26] [27] [28] [48] aussi bien pour résoudre des problèmes d'optimisation de calculs nécessités par les topologies des réseaux [49] [57] [1] [36] que pour les appliquer à des collections de documents de types hétérogènes [30] [107] [37], et contenant des liens hypertextes [43] [103] [33].

Nous décrivons brièvement dans ce qui suit les optimisations apportées aux modèles basés sur les RBs notamment le modèle d'Indrawan et celui de l'équipe de De Campos [1] [34]. Nous avons estimé ces modèles intéressants parce que dans le premier les index des documents sont basés sur des expressions (ensemble de termes reliés sémantiquement) [56] et dans le second des relations de dépendance entre termes sont ajoutées dans la topologie. Ces relations sont déduites de statistiques sur les documents [35] [1] [36] [34].

Les efforts de recherche ont été concentrés autour de l'optimisation dans la représentation des documents, dans la création de *super relations* entre paires (ou couples) de documents ou de termes. Ces relations supplémentaires permettent un gain de stockage et un gain de calcul lors de la propagation de l'information [49], [57], [34].

Une topologie plus flexible a été proposée dans [1] [35] prenant en considération aussi bien les relations entre termes d'indexation que les relations de dépendance entre les documents de la collection grâce à des moyens d'apprentissage dans le but de simplifier les calculs et d'améliorer les résultats de recherche [1]

3.5.1 Modèle d'Indrawan

Dans ce modèle [56] [57], le cycle de vie d'un système de recherche commence avec le développement des représentations des documents. Le réseau document constitue le noyau constant du réseau. Le second réseau est composé de la requête et est dynamique. Dans l'architecture générale (simplifiée) présentée

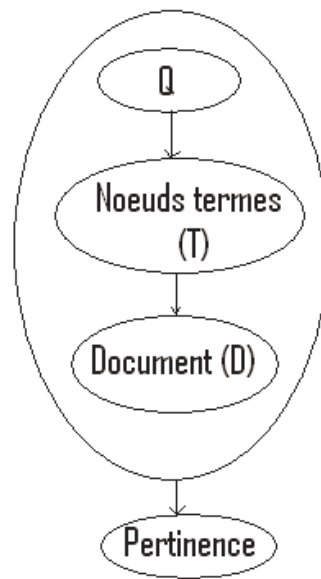


FIG. 3.4 – Architecture globale

dans la figure 3.4, le noeud requête est parent des noeuds termes qui sont aussi parents des noeuds documents. Les arcs sont orientés de Q vers les termes T et des noeuds termes vers les noeuds documents D . Ce modèle est le seul modèle à représenter un noeud « pertinence ». Ce noeud $Pert$ est explicité plus bas.

Représentation des documents Le réseau document est composé de deux couches de noeuds : une couche supérieure et une couche inférieure. La première (couche supérieure) est composée de l'univers de discours des mots clés de la collection. La couche inférieure contient les concepts, ou toute entité générée par la combinaison des mots clés. Une combinaison possible est le document. Un noeud terme implique l'existence des documents.

Représentation de la requête Le réseau requête explicite le besoin de l'utilisateur, (Q). Ce réseau contient un noeud racine qui symbolise le besoin

utilisateur, un ensemble de mots clés qui forment la description de la requête, relié à Q . Ce réseau de requête est temporaire.

Calcul de la pertinence Le processus de recherche est instancié par la réception de la requête. Le but est de calculer la probabilité de pertinence d'un document étant donnée la requête. Le processus de propagation est déclenché par la requête et l'information est propagée sur les noeuds documents. La probabilité de pertinence d'un document étant donnée une requête, selon la topologie du graphe (figure 3.4) est mesurée par 3.8.

$$\begin{aligned} P(Pert \mid D_j, Q) &= P(Pert \mid D_j, Q) \\ &= P(Pert \mid D_j, T_i, Q) \end{aligned} \quad (3.8)$$

Où $Pert$ est un noeud virtuel, noeud fils final du réseau tel que décrit dans la figure 3.4 et décrit l'événement « *la pertinence d'un document donnée en réponse à une requête* ».

D_j est un document de la collection et Q une requête.

Une simplification de la propagation suivant la règle de chaînage (définie au premier chapitre de cette partie) calcule la probabilité de pertinence d'un document, lorsque la requête devient connue, par $P(D_j \mid Q)$.

Une liste de documents ordonnés par probabilités de pertinence décroissantes est restituée. Un processus de réinjection de pertinence est proposé dans [58].

Comparaison avec le modèle inférentiel de Turtle La première différence entre le modèle d'inférence [111] [32] et le modèle d'Indrawan concerne la définition des relations de dépendance et a fortiori le sens des arcs entre les noeuds termes d'indexation et ceux des documents. Dans le modèle d'Indrawan [56], contrairement au modèle de Turtle, les auteurs considèrent que les termes impliquent le document. Les deux points de vue sont discutés et justifiés [56] [111].

Une seconde différence concerne la propagation et la représentation d'un noeud pertinence. Cependant, ce noeud n'est pas vraiment explicité puisque des simplifications sont opérées sur le calcul de la propagation. La requête instancie le système dans le modèle d'Indrawan alors que les documents sont instanciés dans le modèle de Turtle.

L'apport principal de ce modèle concerne l'optimisation proposée pour la propagation dans les réseaux. La complexité de calcul des probabilités conditionnelles croît de manière exponentielle avec le nombre de parents [25]. Pour pallier ce problème, des noeuds virtuels sont ajoutés pour catégoriser des ensembles de noeuds parents. Ainsi, le nombre de probabilités conditionnelles à calculer diminue. De plus, uniquement les termes de la requête sont pris en compte dans les calculs. Un second inconvénient concerne l'existence de boucle de la propagation. L'envoi de *message* [77] [59] des noeuds parents vers les noeuds fils et des noeuds fils vers les noeuds parents peut produire une boucle indirecte. Ce modèle permet de filtrer et de bloquer les *messages* qui circulent des noeuds fils vers les parents lorsque les valeurs de ces *messages* sont les mêmes que les *messages* originaux des parents.

3.5.2 Réseaux multi connectés pour la RI

Dans [34], [36], deux apports principaux sont fournis. Un réseau Bayésien, *BNR*, qui représente les relations de dépendance entre les termes d'indexation d'un document et des techniques qui réduisent le temps de calcul. Généralement les modèles de recherche existant supposent l'indépendance entre les termes pour faciliter les calculs, mais cette supposition entrave l'exactitude de ces modèles. Dans le modèle *BNR*, les relations entre les termes sont prises en compte. Ce modèle ne considère pas tous les termes de la collection : à partir du réseau initial un nouvel *arbre* (un graphe dans lequel il n'y a pas plus d'un chemin orienté reliant chaque paire de noeuds) est appris et contient un ensemble plus petit de termes et leurs relations de dépendance. Les noeuds restants sont complètement isolés. Cette nouvelle topologie représente non seulement les meilleures dépendances existantes entre les termes mais réduit également le temps de construction de l'arbre et le temps de propagation de l'information.

Topologie du modèle Le graphe acyclique orienté comprend deux ensembles de noeuds, comme montré dans la figure 3.5 :

1. L'ensemble des noeuds termes, T ; une variable T_i associée à un terme prend ses valeurs dans le domaine $\{t_i, \bar{t}_i\}$, où \bar{t}_i désigne le terme T_i comme n'étant pas pertinent et t_i comme pertinent. Un terme est pertinent si l'utilisateur considère qu'il va apparaître dans un document pertinent

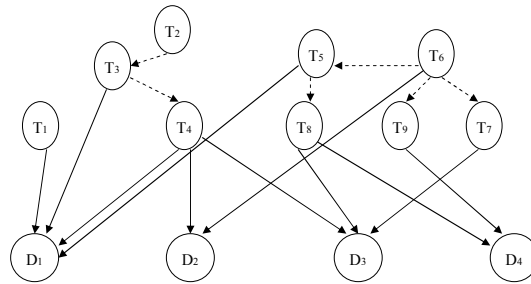


FIG. 3.5 – Architecture globale

(l'utilisateur utilise ce terme dans sa requête). Un terme est non pertinent si l'utilisateur pense que ce terme n'apparaît pas dans un document pertinent ;

2. L'ensemble des noeuds documents, D ; une variable D_j prend ses valeurs dans $\{d_j, \bar{d}_j\}$, 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.

La topologie proposée traduit une relation de dépendance d'un terme vers un document lorsque le document contient le terme. De plus, il n'y a pas de dépendance directe entre les documents. Enfin, étant donnée une requête, le degré de pertinence d'un document D_j peut être complètement déterminé par la connaissance de la pertinence de tous les termes d'indexation du document D_j . Lorsque cette information est absente, la connaissance de la pertinence pour la même requête d'un autre document D_k peut avoir une influence sur celle de D_j . Ce qui signifie que tout document D_j est conditionnellement indépendant de tout document D_k lorsque les valeurs de pertinence de tous les termes indexant le document D_j sont connues [36].

L'ensemble des parents d'un document est l'ensemble des termes indexant le document. Les dépendances entre les termes sont fonction des co-occurrences entre les termes.

Estimation des distributions de probabilité Les noeuds racines étant les noeuds termes, les probabilités *a priori* de ces noeuds sont données par :

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

avec M le nombre de termes de la collection.

Pour les termes ayant des parents qui sont eux aussi des noeuds 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.

L'estimation de la probabilité des documents est plus problématique. Si un document est indexé par 30 termes, nous avons besoin de calculer 2^{30} probabilités. Un modèle canonique est alors utilisé pour représenter ces probabilités conditionnelles. Soit θ_{D_j} l'ensemble des configurations possibles des parents de D_j . Considérons le document D_j composé des deux termes $\{T_1, T_2\}$. Les configurations possibles sont $\{t_1, t_2\}$, $\{t_1, \bar{t}_2\}$, $\{\bar{t}_1, t_2\}$ et $\{\bar{t}_1, \bar{t}_2\}$. Etant donné l'ensemble des configurations des parents d'un document, certaines sont pertinentes, notées $R(\theta_{D_j})$, et d'autres pas. 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.

$$P(d_j | \theta_{D_j}) = \sum_{T_i \in R(\theta_{D_j})} w_{ij}$$

$$P(\bar{d}_j | \theta_{D_j}) = 1 - P(d_j | \theta_{D_j})$$

avec w_{ij} le poids du terme T_i dans le document D_j . Ce poids est compris entre 0 et 1 et est obtenu par des variations normalisées de *tf * idf* [92].

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, $P(d_j/Q)$. Les documents restitués sont classés par ordre décroissant de leur probabilité de pertinence. Les termes de la requête sont considérés comme l'information à propager. La probabilité de pertinence d'un document étant donnée une requête est égale à la somme des produits des poids des termes de la requête et des documents. Les termes considérés sont ceux de la requête présents dans le document. La somme des poids des termes présents dans le document et absents de la requête sont aussi considérés dans le calcul de la pertinence.

Réduction des relations de dépendance entre termes D'un point de vue temps de calcul, le modèle présenté peut contenir deux inconvénients majeurs en terme de temps liés : (i) au temps de construction de la structure de dépendance entre les termes, et (ii) au temps de calcul lors de la propagation de l'information.

Pour réduire ces coûts, le nombre de termes impliqués dans l'arbre est réduit entraînant une réduction de l'apprentissage du réseau ainsi que le temps de propagation. Cette réduction est possible grâce à la subdivision des termes en deux sous ensembles : termes « bons » et termes « mauvais ». Les « bons » termes sont ceux qui ont un pouvoir discriminant élevé calculé par la méthode de Salton [97] (exposé dans le premier chapitre de ce manuscrit). Les conclusions des expérimentations sur différentes collections de tests sont mitigées. En effet, la prise en compte des relations de dépendance entre les termes ne se sont pas toujours avérées efficaces (en termes de précision).

3.5.3 Réseaux de croyance basés sur les expressions d'indexation

Dans les travaux de Bruza [19] [55], les documents sont représentés par des expressions d'indexation. Une expression est composée d'une séquence de termes reliés par des connecteurs. Les relations entre les termes sont données explicitement par le contenu de l'information [40]. Toute expression est définie par une structure d'arbre. Un langage des index d'expression est défini à partir des expressions dérivées de la collection. Les relations sont catégorisées par type et regroupent les différents types de connecteurs trouvés dans la collection. Par exemple le connecteur « *de* » définit une relation de type *possession* dans l'expression *château de la reine*.

Le *pouvoir* des expressions d'indexation est l'ensemble de toutes ses sous expressions. Les distributions de probabilité jointes définies sur l'ensemble des indexes d'expression permettent de calculer la probabilité *a priori* d'une expression d'être associée à un document pertinent.

Des règles d'inférence plausibles associées au réseau permettent de faire des inférences. Par exemple, un document qui contient l'expression *pollution des rivières*, peut être considéré comme un document qui traite de la *pollution*.

La pertinence d'un document en réponse à une requête est obtenue à partir des règles d'inférence. Elle est calculée, pour un document donné, en déduisant la

requête des expressions utilisées pour représenter ce document.

3.6 Conclusion et Discussions

D'autres modèles basés sur les réseaux moins connus ont été proposés pour la RI que nous n'avons pas détaillés [28] [33]. Ces modèles sont basés sur des réseaux de requêtes. Un champ récent d'application dans le domaine de la RI dans lequel les réseaux Bayésiens ont une utilisation intéressante concerne notamment les collections hétérogènes (XML,..)

Les modèles basés sur les réseaux Bayésiens proposent un formalisme général pour la modélisation de la pertinence et la représentation des informations (documents, termes, requête). Plusieurs manières de considérer les relations de dépendance entre l'information d'observer les documents et de satisfaire le besoin utilisateur ont été définies.

Un premier constat que nous pouvons faire à travers cet état de l'art est qu'il existe deux principaux modèles basés sur les réseaux Bayésiens pour répondre aux besoins de la RI. Alors que le modèle inférentiel instancie le document à la réception d'une requête, le modèle de croyance instancie la requête. Une différence majeure dans la topologie de ces deux réseaux concerne le sens de la dépendance des termes d'indexation avec les documents. Alors que pour le modèle inférentiel cette dépendance, quantifiée par $P(t_i | d_j)$, va des documents vers ses termes d'indexation, dans le modèle de croyance la relation de dépendance est orientée des termes, qui constituent l'univers de discours, vers les documents et est quantifiable par $P(d_j | t_i)$.

Le modèle de croyance généralise les modèles traditionnels (probabiliste, vectoriel et booléen, d'inférence). Cette généralisation permet la combinaison et l'utilisation de leurs caractéristiques respectives dans un seul modèle et de reproduire leurs classements des documents restitués. Le modèle inférentiel ne peut pas reproduire le classement obtenu par le modèle vectoriel de par sa topologie.

Ces deux types de modèle ont permis de représenter graphiquement le processus global de RI sans pour autant accorder une grande importance ni donner un sens particulier à la quantification des liens. Les probabilités qui quantifient les arcs du réseau sont basées sur des heuristiques et s'inspirent des variations de $tf - idf$ [94] pour pondérer les arcs reliant les termes d'indexation aux documents.

La pertinence ne tient pas compte de l'imprécision engendrée par ces heuristiques. La notion de pertinence est définie d'une façon très générale qui permet la généralisation des modèles de base, mais est difficilement *raffinable*. De plus, l'évaluation des documents par rapport à une requête, ne prend en compte que les termes d'indexation présents à la fois dans les documents et la requête. En effet, l'absence des termes de la requête n'est pas traitée explicitement dans ces deux modèles, bien que dans le modèle de croyance les termes d'indexation de la requête constituent le point d'entrée du système (le processus de recherche est instancié par la réception de la requête).

Un dernier point concerne la définition ambiguë de la probabilité *a priori* d'un document dans le modèle inférentiel. Les documents de la collection sont représentés par des noeuds dans le réseau. Chaque noeud est de domaine binaire et la probabilité *a priori* d'un document devrait alors être égale à $\frac{1}{2}$ et non pas à $\frac{1}{N}$ comme défini dans [111]. Cette dernière définition ($P(d_j) = \frac{1}{N}$) signifierait que tous les documents sont représentés dans un seul noeud représentant tous les documents de la collection et que donc $dom(D_j) = \{d_1, \dots, d_N\}$. Cette définition de domaine des documents est justifiable puisqu'un seul document est instancié à la fois, excluant l'instanciation des autres documents de la collection.

A notre sens un cadre théorique intéressant permettant à la fois d'exprimer l'ignorance ainsi que tenir compte de l'imprécis et de l'incertain est possible par la théorie des possibilités. Nous présentons dans la troisième partie de ce manuscrit les motivations qui nous ont poussé à proposer un nouveau modèle de RI ainsi qu'à utiliser la théorie des possibilités dans le domaine de la RI.

Troisième partie

Un Modèle de RI basé sur les Réseaux Possibilistes

Chapitre 4

Un modèle de Recherche d'Information basé sur les Réseaux Possibilistes

4.1 Introduction

Les modèles actuels de la RI sont catégorisés en fonction de leur modélisation de la pertinence. Ces modèles calculent un score de pertinence traduisant un appariement entre la requête et le document [92] ou une probabilité de pertinence d'un document vis à vis d'une requête utilisateur [87]. Les documents sont restitués par ordre décroissant de leur pertinence.

Les données disponibles pour calculer la pertinence sont peu nombreuses, et il nous paraît difficile de traduire la pertinence par une unique valeur. En effet, les travaux récents sur la pertinence, s'accordent à dire qu'il est difficile de lui attribuer une définition précise. Intuitivement, il n'est pas « légitime » qu'une unique valeur puisse traduire correctement toute la sémantique de la pertinence.

La logique possibiliste permet une flexibilité dans le traitement de l'information disponible. Elle nous permet de modéliser et de quantifier la pertinence d'un document étant donnée une requête au travers de deux mesures : la nécessité et la possibilité. Les documents nécessairement pertinents sont ceux qui doivent figurer en haut de la liste des documents restitués et doivent permettre une cer-

taine efficacité du système. Les documents possiblement pertinents sont ceux qui répondraient éventuellement à la requête utilisateur. Ils figurent dans la liste des documents restitués classés à la suite des documents nécessairement pertinents ou à défaut (si le système n'en trouve pas) ils sont considérés comme une réponse plausible.

Ce chapitre est structuré de la manière suivante :

Nous décrivons tout d'abord le cadre théorique sur lequel repose notre approche, à savoir les réseaux possibilistes. Nous détaillons dans la seconde section le modèle que nous proposons. Ce modèle est basé sur un réseau possibiliste, défini par une composante qualitative et une composante quantitative :

- la composante qualitative représente les noeuds documents, termes d'indexation et la requête et les relations de dépendance existant entre eux ;
- la composante quantitative mesure les arcs entre les noeuds par les degrés de possibilité et de nécessité.

Les principaux résultats de ce Chapitre sont publiés dans [14] [16] [15] [18] [17]

4.2 Les Réseaux Possibilistes

4.2.1 La théorie des possibilités

La théorie des possibilités introduite par Zadeh [118] et développée par Dubois et Prade [39] traite l'incertitude sur l'intervalle $[0, 1]$, appelé échelle possibiliste, d'une manière qualitative ou quantitative. Nous nous restreignons, pour nos travaux, au cadre quantitatif.

4.2.1.1 Distribution de possibilité

La théorie des possibilités est basée sur les distributions de possibilité. Une distribution de possibilité, notée par π , est une application de Ω (l'univers de discours) vers l'échelle $[0, 1]$ traduisant une connaissance partielle sur le monde, noté ω . L'échelle possibiliste est définie de deux manières. Dans le cadre numérique les valeurs des possibilités traduisent souvent les bornes supérieures

des probabilités. Dans le cadre qualitatif, les valeurs de possibilité peuvent être considérées comme un ordre de classement des états possibles. La combinaison des distributions de possibilité, exprimée à l'aide des normes triangulaires (t-normes) dépend du cadre. Les opérateurs « produit » et « minimum » peuvent être utilisés pour combiner des distributions de possibilité indépendantes dans les cadres quantitatif et qualitatif respectivement.

Normalisation Une distribution de possibilité est dite α -normalisée, si son degré de normalisation, noté $a(\pi)$, est égal à α . Ainsi :

$$\alpha = a(\pi) = \max_{\omega} \pi(\omega)$$

Lorsque $\alpha = 1$, π est dite normalisée.

Marginalisation Soit une distribution de possibilité jointe, π sur Ω , une distribution marginale relative aux sous ensembles de variables peut être dérivée en utilisant l'opérateur *maximum*. Ainsi, $\forall X \subseteq V \forall x \in \text{dom}(X)$:

$$\pi(x) = \max_{\omega \in \Omega} \{ \pi(\omega) : \omega[X] = x \}$$

où V : ensemble de variables $\{A_1, A_2, \dots, A_N\}$;

X : sous ensemble de V ;

$\text{dom}(X)$: domaine de X , produit cartésien des domaines des variables de X ;

x : une instance de X , si $X = \{A_1, A_2, \dots, A_j\}$, alors $x = (a_1, a_2, \dots, a_j)$;

$\omega[X] = x$: configuration de X dans ω .

Une distribution de possibilité π sur Ω permet de qualifier les événements en terme de mesure de plausibilité et de certitude respectivement.

4.2.1.2 Mesures de nécessité et de possibilité

Dire qu'un événement est non possible n'implique pas seulement que son événement contraire est possible mais qu'il est certain. Deux mesures duales sont utilisées : la mesure de possibilité $\Pi(\phi)$, et la mesure de nécessité $N(\phi)$.

- La possibilité d'un événement A , notée $\Pi(A)$ est obtenue par $\Pi(A) = \max_{x \in A} \pi(x)$ et décrit la situation la plus normale dans laquelle A est vraie ;

- La nécessité $N(A) = \min_{x \notin A} 1 - \pi(x) = 1 - \Pi(\bar{A})$ d'un événement A reflète la situation la plus normale dans laquelle A est faux.

La distance entre $N(A)$ et $\Pi(A)$ évalue le niveau d'ignorance sur A . Rappelons que $N(A) > 0$ implique $\Pi(A) = 1$. Lorsque A est un ensemble flou, cette propriété n'est plus vérifiée et dans ce cas l'inégalité $N(A) \leq \Pi(A)$ est vérifiée.

4.2.1.3 Conditionnement possibiliste

En logique possibiliste, le conditionnement consiste à modifier la distribution de possibilité initiale π à l'arrivée d'une nouvelle information i . Soit ϕ , une sous classe de ω , $\phi = [i]$ l'ensemble des modèles de i . La distribution initiale π est remplacée par $\pi' = \pi(\bullet/\phi)$. Dans un cadre quantitatif, les éléments de ϕ sont proportionnellement modifiés :

$$\pi(\omega/_p\phi) = \left\{ \begin{array}{ll} \frac{\pi(\omega)}{\Pi(\phi)} & \text{si } \omega \in \phi \\ 0 & \text{sinon} \end{array} \right\} \quad (4.1)$$

avec : $/_p$: conditionnement basé sur le produit Dans un cadre qualitatif, le degré de possibilité maximal est affecté aux meilleurs éléments de ϕ :

$$\pi(\omega|_m\phi) = \left\{ \begin{array}{ll} 1 & \text{si } \pi(\omega) = \Pi(\omega) \text{ et } \omega \in \phi \\ \pi(\omega) & \text{si } \pi(\omega) < \Pi(\omega) \text{ et } \omega \in \phi \\ 0 & \text{sinon} \end{array} \right\} \quad (4.2)$$

$/_m$: conditionnement basé sur le minimum

4.2.2 Réseaux Possibilistes (RP)

Les travaux existant sur les réseaux possibilistes sont soit des adaptations directes de l'approche probabiliste [4], ou des méthodes d'apprentissage à partir de données imprécises [8]. La théorie des possibilités offre deux définitions du conditionnement, ce qui conduit à deux définitions des réseaux causaux possibilistes. Les réseaux possibilistes basés sur le produit sont très similaires aux réseaux probabilistes.

4.2.2.1 Définitions

Un graphe possibiliste orienté sur un ensemble de variables $V = \{A_1, A_2, \dots, A_N\}$ est caractérisé par une composante qualitative et une composante numérique. La première est un graphe acyclique orienté. La structure du graphe représente l'ensemble des variables ainsi que l'ensemble des relations d'indépendance. La seconde composante quantifie les liens du graphe en utilisant des distributions de possibilité conditionnelles de chaque noeud dans le contexte de ses parents. Ces distributions de possibilité doivent vérifier la contrainte de normalisation. Pour chaque variable A_i :

- Si A_i est un noeud racine et dom_{A_i} le domaine de A_i , la possibilité *a priori* de A_i doit satisfaire :

$$\max_{a_i} \Pi(a_i) = 1, \forall a_i \in dom_{A_i}$$

- Si A_i n'est pas un noeud racine, la distribution conditionnelle de A_i dans le contexte de ses parents doit satisfaire :

$$\max_{a_i} \Pi(a_i/\theta_{A_i}) = 1, \forall a_i \in dom_{A_i}$$

avec :

dom_{A_i} : le domaine de A_i ,

θ_{A_i} : l'ensemble des configurations possibles des parents de A_i .

4.2.2.2 Réseaux possibilistes basés sur le minimum

Un graphe possibiliste basé sur le minimum, noté par GP_M , est un graphe possibiliste où les possibilités conditionnelles sont obtenues par le conditionnement minimum (formule 4.2). La distribution de possibilité des réseaux possibilistes basée sur le minimum, notée par π_M , est obtenue par la règle de chaînage :

$$\pi_M(A_1, \dots, A_N) = \text{MIN}_{i=1..N} \Pi(A_i/\theta_{A_i})$$

avec :

MIN : l'opérateur minimum.

4.2.2.3 Réseaux possibilistes basés sur le produit

Un graphe possibiliste basé sur le produit, noté par GP_P , est un graphe possibiliste où les possibilités conditionnelles sont obtenues par le conditionnement produit (formule 4.1). La distribution de possibilité des réseaux possibilistes basés sur le produit, notée par π_P , est obtenue par la règle de chaînage :

$$\pi_P(A_1, \dots, A_N) = PROD_{i=1..N} \Pi(A_i / \theta_{A_i})$$

avec :

$PROD$: l'opérateur produit.

4.2.2.4 Logique possibiliste

Dans la logique possibiliste, les règles sont modélisées par des clauses logiques :

$$p \rightarrow q = \neg p \vee q$$

Des valeurs sont attachées aux bornes inférieures des degrés de nécessité et de possibilité de p et q qui sont considérées comme des propositions booléennes. Les axiomes de la théorie des possibilités permettent de modéliser p implique q avec un poids $\alpha > 0$ par l'inégalité

$$N(p \rightarrow q) \geq \alpha$$

ou d'une manière équivalente par

$$\Pi(p \wedge \neg q) \leq 1 - \alpha$$

pour signifier que $p \wedge \neg q$ est quelque peu impossible.

La distribution de possibilité exprimant cette information (connaissance) est π telle que :

$$\begin{aligned} \pi(x) &= 1 - \alpha \text{ si } p \wedge \neg q \text{ vraie à l'état } x \\ &= 1 \text{ sinon} \end{aligned}$$

La distribution de possibilité induite par plusieurs propositions, mesurée par des nécessités, est obtenue par une intersection floue (utilisant le minimum) des distributions de possibilité induites par chaque proposition.

4.3 Un modèle de RI basé sur les réseaux possibilistes

Les travaux que nous proposons s'inscrivent dans la définition d'un nouveau modèle de RI permettant notamment une nouvelle modélisation de la pertinence.

D'une manière générale, quel que soit le modèle de la littérature, et particulièrement ceux qui considèrent les poids des termes comme des probabilités de pertinence, l'incomplétude (ou imprécision) de l'information n'est pas considérée lors de la représentation d'un document ou de son évaluation étant donnée une requête.

Nous nous sommes particulièrement penchés dans nos travaux sur la résolution de trois points qui nous paraissent essentiels pour un modèle *efficace* et *viable* de la RI.

- Les modèles actuels de RI calculent généralement la pertinence d'un document vis à vis d'une requête par une somme de produits des poids des termes de la requête présents dans les documents. Nous estimons qu'il est difficile de traduire la notion de pertinence par une unique valeur. En effet, nous ne sommes pas « sûrs » que l'utilisation d'une unique valeur puisse exprimer toute la sémantique liée à cette notion. Nous suggérons de traduire différents aspects de la pertinence.

La théorie des probabilités permet de représenter un événement et son contraire. L'information majeure qui concerne la RI est la restitution des documents pertinents en réponse à une requête. Les modèles actuels considèrent généralement que la pertinence d'un document donné de la collection est indépendante de celle des autres documents de la collection. De ce fait, l'information disponible sur les documents non pertinents n'a pas d'impact sur la connaissance des documents pertinents. Par contre, une théorie qui permet de renforcer une information, notamment la pertinence d'un document donné, peut s'avérer intéressante dans le contexte de la RI.

Notre approche est basée sur les réseaux possibilistes. La théorie des possibilités propose des mesures duales permettant le traitement de l'information incertaine. Nous montrons dans notre approche que ces deux mesures sont complémentaires. 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 ;

- Le second point est étroitement lié au premier. La pertinence repose essentiellement sur les poids des termes. La pondération est à notre sens un élément fondamental de la RI. Les modèles actuels calculent le poids d'un terme dans un document donné en combinant son nombre d'apparitions dans ce document (tf) à son nombre d'apparitions dans tous les documents de la collection. Ce poids est censé mesurer l'importance d'un terme dans la collection. Dans notre vision, il est « risqué » de mesurer cette importance par une seule mesure ainsi obtenue. En effet, nous estimons qu'une telle approche induit une perte d'information. Nous suggérons de mesurer l'importance d'un terme dans un document par deux degrés. Le premier élimine les termes non intéressants ou non représentatifs. Le second renforce l'importance du terme ;
- Le dernier point découle de l'ignorance des termes de la requête absents dans les documents lors du calcul des scores de pertinence. Un terme important dans la collection est un terme qui permet de pointer vers un sous-ensemble de documents de la collection. Lorsque ce terme apparaît dans la requête, il permet au système de restituer les documents pertinents. Ce terme lorsqu'il est absent d'un document donné devrait pénaliser la pertinence du document.

Nous proposons donc un modèle de RI basé sur les réseaux possibilistes que nous détaillons dans ce qui suit.

4.3.1 Architecture générale du modèle

La topologie du réseau est représentée dans la figure (4.1). D'un point de vue qualitatif, le graphe permet de représenter les noeuds documents, requête, termes d'indexation et permet d'exprimer les relations de dépendance existant entre ces noeuds. Un document (D_j) est instancié ou non, prenant ses valeurs dans le domaine $\{d_j, \overline{d_j}\}$. L'activation (ou instanciation) d'un noeud document,

$D_j = d_j$ (resp. $\overline{d_j}$) signifie que le document est pertinent ou non étant donnée une requête. Une requête, Q , prend ses valeurs dans le domaine $\{q, \overline{q}\}$. Nous sommes intéressés par l'instanciation de la requête, nous ne considérons que le cas $Q = q$, et nous le notons Q . Le domaine d'un noeud terme d'indexation, T_i , est $\{t_i, \overline{t_i}\}$. ($T_i = t_i$) signifie que le terme t_i est présent dans l'objet (document ou requête) et donc *représentatif* de l'objet. Un terme *non-représentatif*, $\overline{t_i}$, est un terme absent de la représentation de l'objet. Soit $\mathcal{T}(D_j)$ (resp. $\mathcal{T}(Q)$) l'en-

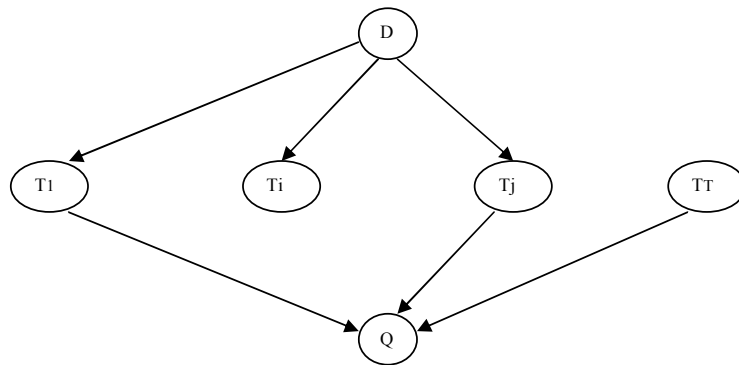


FIG. 4.1 – Architecture générale du modèle

semble des termes indexant le document D_j (resp. la requête Q). Considérons le sous-réseau document composé des noeuds documents et de leurs termes d'indexation. Les arcs sont orientés des noeuds documents vers les noeuds termes d'indexation exprimant les relations de dépendance existant entre les deux types de noeuds. Les termes de ce sous-réseau n'ont une existence que parce qu'ils apparaissent dans ces noeuds documents qui sont leurs parents. Considérons à présent le sous-réseau requête constitué du noeud requête et de ses termes d'indexation. La requête exprime une « demande » aux documents contenant certains termes mais en excluant d'autres. La requête propage l'information aux noeuds termes qui figurent dans la collection. Ces noeuds termes forment les noeuds parents de la requête. Un terme d'indexation de la requête n'apparaissant pas dans un document donné sera considéré comme un noeud terme racine, n'ayant pas de parents. Le système est instancié par la soumission de la requête. L'instanciation de la requête propage l'information à travers le réseau en activant les noeuds termes d'indexation, parents de la requête. Il existe une instanciación de l'ensemble des parents de la requête, PAR_Q , qui représente la requête dans sa forme la plus stricte (exactement telle que formulée par l'utilisateur). Soit θ^Q cette instanciación. L'ensemble des instances possibles des parents de la requête est noté θ . Nous montrerons plus tard dans

ce manuscrit, comment les valeurs sont affectées aux arcs.

4.3.2 Evaluation de la requête

L'évaluation de la requête est effectuée par la propagation de l'information apportée par la requête à travers le réseau. Dans ce modèle, le processus de propagation est similaire à la propagation probabiliste Bayésienne [4] [8]. Le processus d'évaluation consiste à propager l'information *injectée* par la requête. Les arcs reliés à la requête sont instanciés dans le but de calculer la pertinence des documents étant donnée cette requête.

Nous proposons donc de modéliser la pertinence par deux mesures différentes mais complémentaires traduisant des aspects liés à la perception que nous pouvons avoir de la pertinence. Nous ne prétendons pas dans nos travaux traduire toute la sémantique liée à cette notion. Nous estimons que *a priori* l'utilisation de plusieurs valeurs peut traduire différents aspects de la pertinence. Les décisions de restitution des documents en réponse à une requête reposent sur la valeur des scores de pertinence. Nous voulons montrer qu'intuitivement, avoir plus d'un score, en l'occurrence deux dans notre cas, apporterait une facilité quant à la décision de restitution des documents en réponse à une requête utilisateur. Nous sommes conscients, cependant, que les mesures sur lesquelles se basent les prises de décision ne doivent en aucun cas être contradictoires. Nous avons montré dans des travaux portant sur la réinjection de pertinence que la gradualité de la pertinence peut améliorer les performances d'un SRI [12] [13]. D'une manière générale, la théorie des possibilités permet de borner cinq domaines disjoints de propositions. Dans notre cas, la proposition est « la pertinence d'un document étant donnée une requête ». Nous notons cette proposition H pour simplifier. Alors elle est :

1. Impossible ; dans ce cas, $\Pi(H) = 0$ et $N(H) = 0$;
2. Pas tout à fait possible et non nécessaire. Dans ce cas $\Pi(H) > 0$ et $N(H) = 0$;
3. Complètement possible et non nécessaire : $\Pi(H) = 1$ et $N(H) = 0$;

4. Complètement possible et pas tout à fait nécessaire : $\Pi(H) = 1$ et $N(H) > 0$;
5. Certaine : $\Pi(H) = 1$ et $N(H) = 1$.

Nous modélisons donc ici, la pertinence par une double mesure. La **pertinence nécessaire** mesure à quel point un document doit faire partie de la liste des documents restitués. La **pertinence possible** mesure à quel point un document constitue éventuellement une réponse à une requête donnée. Le calcul de chacune de ces mesures repose sur des informations différentes. Notre 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, $\Pi(d_j | Q)$;
- Il est aussi certain (dans le sens possibiliste) que le document est pertinent à la requête, $N(d_j | Q)$.

Le premier type de proposition est censé éliminer les documents non pertinents. Le second se focalise sur le renforcement de la certitude de la pertinence. Etant donnée l'approche possibiliste choisie, nous cherchons à pouvoir restituer les documents nécessairement ou au moins possiblement pertinents étant donnée une requête. Ainsi, le processus de propagation évalue les degrés de possibilité, $\Pi(d_j | Q)$, et de nécessité, $N(d_j | Q)$, par :

$$\Pi(d_j | Q) = \frac{\Pi(Q \wedge d_j)}{\Pi(Q)}, \quad (4.3)$$

$$N(d_j | Q) = 1 - \Pi(\bar{d}_j | Q), \quad (4.4)$$

$$\text{avec, } \Pi(\bar{d}_j | Q) = \frac{\Pi(Q \wedge \bar{d}_j)}{\Pi(Q)}$$

La possibilité de Q est

$$\Pi(Q) = \max(\Pi(Q \wedge d_j), \Pi(Q \wedge \bar{d}_j)) \quad (4.5)$$

d'après [39],[4] nous avons :

$$\Pi(d_j | Q) = \min\left(1, \frac{\Pi(Q \wedge d_j)}{\Pi(Q \wedge \bar{d}_j)}\right) \quad (4.6)$$

Nous cherchons à définir $\Pi(Q \wedge D_j)$. Etant donnée la topologie du graphe, elle est de la forme :

$$\Pi(Q \wedge D_j) = \max_{\forall \theta^l \in \theta} (\Pi(Q | \theta^l) \cdot \prod_{T_i \in \mathcal{T}(Q) \wedge \mathcal{T}(D_j)} \Pi(\theta_i^l | D_j) \cdot \Pi(D_j) \cdot \prod_{T_k \in \mathcal{T}(Q) \setminus \mathcal{T}(D_j)} \Pi(\theta_k^l)) \quad (4.7)$$

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 .

Les configurations possibles des termes de la requête, Q , composée des termes $\{T_1, T_2\}$ sont $\theta = \{\{t_1, t_2\}, \{t_1, \bar{t}_2\}, \{\bar{t}_1, t_2\}, \{\bar{t}_1, \bar{t}_2\}\}$; L'instanciation θ_1^1 du terme T_1 dans la première configuration, $\theta^1 = \{t_1, t_2\}$, est $\theta_1^1 = t_1$.

Cette quantité (4.7) est calculée pour $D_j \in \{d_j, \bar{d}_j\}$. Nous remarquons que les termes $T_i \in \mathcal{T}(D_j) \setminus \mathcal{T}(Q)$, les termes présents dans le document absents de la requête, ne sont pas instanciés lors des calculs. Nous discuterons plus tard de ce point dans ce manuscrit. De plus, les termes de la requête qui indexent les documents, $T_i \in \mathcal{T}(Q) \cap \mathcal{T}(D_j)$, sont évalués dans le contexte de leurs parents par $\Pi(T_i | D_j)$, et séparés des termes de la requête absents des documents, pour lesquels une possibilité marginale est calculée, $\Pi(T_k)$.

A l'issue du processus de propagation, chaque document aura donc une valeur de nécessité et de possibilité de pertinence. Les documents répondant à la requête sont classés selon ces deux pertinences. Les documents sont restitués par ordre décroissant de pertinence nécessaire puis de pertinence possible. En effet, ceux classés en premier sont les documents qui ont une valeur de nécessité supérieure à 0. 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 (les documents ayant des degrés de nécessité de pertinence égale à 0).

Pour évaluer les documents étant donnée la requête, nous avons besoin de calculer chacun des facteurs utilisés dans 4.7. Nous décrivons dans ce qui suit, les différents traitements de la requête en fonction des configurations de ses termes ainsi que des connecteurs utilisés entre eux, $\Pi(Q | \theta)$. Les termes instanciés propagent l'information sur les documents qu'ils indexent, $\Pi(T_i | D_j)$. Nous définissons des postulats pour le calcul des poids des termes présents dans les documents et des termes racines, $\Pi(T_k)$. Les termes racines sont les termes présents dans la requête et absents des documents. Nous montrons par la suite, le calcul de la possibilité *a priori* des documents, $\Pi(D_j)$, en absence et en présence d'information sur les documents.

4.3.3 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*. Ces deux dernières agrégations ont déjà été proposées dans les travaux de Turtle [111].

L'idée majeure de l'agrégation de la requête est de mesurer la conformité d'une configuration possible, en l'occurrence celle trouvée dans un document donné, avec la configuration des termes de la requête. Pour ce faire, pour toute configuration, θ^l de θ , la possibilité conditionnelle $\Pi(Q | \theta^l)$ est spécifiée par des fonctions d'agrégation en fusionnant les fonctions de ressemblance élémentaires $\Pi(Q | \theta_i^l)$. Chaque $\Pi(Q | \theta_i^l)$ est le poids de la conformité entre l'instance θ_i^l du terme T_i avec celle de la requête (dans θ^Q). Une fonction de ressemblance élémentaire évalue donc à quel point une instance d'un terme dans une configuration donnée ressemble à l'instanciation de ce même terme dans la requête. Cette configuration est en fait la configuration telle que trouvée dans un document. Nous ne considérons pas les relations de dépendance entre couples de termes ici. Cependant ce type de relations pourrait être une information supplémentaire intéressante à exploiter. Ces relations sont exprimables aisément au moyen des réseaux. Nous discuterons de ce point plus loin dans le manuscrit.

Le stockage de toutes les configurations possibles des termes de la requête est coûteux en espace et le temps de calcul croît de manière exponentielle avec le nombre de termes parents de la requête. En effet, une requête, Q de domaine binaire, composée de 20 termes de domaines binaires aussi, nécessite 2×2^{20} calculs de configurations possibles. Dans notre cas, nous nous intéressons uniquement au cas $Q = q$, que nous notons Q pour simplifier. Une organisation possible serait de pondérer chaque terme de la requête et de calculer le poids de la jointure des termes de la requête. Lorsque l'utilisateur ne fournit aucune information sur les opérateurs d'agrégation de sa requête, l'unique connaissance disponible est l'importance du terme dans la collection. Cette connaissance est disponible pour chaque terme. Nous donnons dans ce qui suit les différentes techniques que nous proposons pour agréger les termes de la requête.

4.3.3.1 Agrégations booléennes et quantifiée des termes de la requête

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

$$\begin{aligned}\Pi(Q | \theta^l) &= 1 \text{ si } \theta_i^l = \theta_i^Q \\ &= 0 \text{ sinon}\end{aligned}\tag{4.8}$$

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

$$\begin{aligned}\Pi(Q | \theta^l) &= 1 \text{ si } \forall T_i \in PAR_Q, \theta_i^l = \theta_i^Q \\ &= 0 \text{ sinon}\end{aligned}\tag{4.9}$$

Dans 4.9, il faut que chaque terme T_i parent de la requête Q soit instancié dans θ comme dans la requête. Les documents pertinents pour ce type de requête sont les documents contenant simultanément tous ses termes. Lorsque les termes de la requête concernent un même sujet, des documents plausiblement ou nécessairement pertinents peuvent être restitués. Cependant, plus les termes de la requête sont nombreux et plus ils traitent de sujets différents, plus il est difficile de restituer des documents. Généralement, ce type de requête est trop strict.

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 pertinence finale d'un document augmente avec le nombre de termes de la requête présents. La conjonction pure est manipulée en remplaçant \forall par \exists dans la requête conjonctive 4.9.

$$\begin{aligned}\Pi(Q | \theta^l) &= 1 \text{ si } \exists T_i \in PAR_Q \text{ tel que } \theta_i^l = \theta_i^Q \\ &= 0 \text{ sinon}\end{aligned}\tag{4.10}$$

Cette interprétation est trop large pour discriminer entre les documents. Dans le cas de la disjonction, le système restitue les documents contenant au moins un terme de la requête. Les documents contenant tous les termes de la requête peuvent être restitués avec un score de pertinence plus faible qu'un

document ne contenant qu'un terme de la requête. Dans notre approche, le calcul de la pertinence d'un document vis à vis d'une requête dépend de la valeur *maximum* des instances des configurations des parents de la requête. Ce maximum atteint rapidement la valeur 1, il suffit pour cela qu'au moins un terme de la requête soit instancié telle que dans la configuration. Le score de pertinence finale d'un document donné dépend des poids des termes de la requête présents et absents dans une configuration donnée du document en question. Ainsi, soit une requête Q composée des deux termes T_1, T_2 . Il n'est pas impossible que le document D_1 contenant le terme T_1 se retrouve avec un score de pertinence plus élevé que celui du document D_2 contenant les deux termes de la requête.

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 :

$$\begin{aligned} \Pi(Q | \theta_i^l) &= 1 \text{ si } \theta_i^l = \bar{t}_i \\ &= 0 \text{ sinon} \end{aligned} \tag{4.11}$$

Le terme parent de la requête doit être instancié à *non représentatif* lorsque la requête contient la négation du terme.

Quantification Supposons qu'une requête est satisfaite par un document si elle contient au moins K termes communs avec le document. Nous considérons une fonction croissante, $f(\frac{K(\theta^l)}{n})$, tel que $K(\theta^l)$ est le nombre de termes de la requête instanciés dans une configuration donnée θ^l de PAR_Q , et que la requête contient n termes. Nous posons $f(0) = 0$ et $f(1) = 1$. f est un quantificateur flou [116]. Par exemple,

$$\begin{aligned} f(i/n) &= 1 \text{ si } i \geq \frac{K(\theta^l)}{n}, \\ &= 0 \text{ sinon} \end{aligned} \tag{4.12}$$

Pour l'agrégation donnée par 4.12 il faut qu'au moins K termes de la requête soient en conformité avec θ . D'une manière générale, f peut être une fonction

non booléenne.

L'approche quantifiée pour calculer la possibilité d'une requête Q étant donnée une configuration θ^l de tous ses parents, est donnée par :

$$\Pi(Q | \theta^l) = f\left(\frac{K(\theta^l)}{n}\right) \quad (4.13)$$

Le tableau 4.1 présente les résultats d'une quantification sur une requête, Q , contenant trois termes $\{T_1, T_2, T_3\}$. Pour cette quantification, la configuration est considérée « conforme » si au moins deux termes ont la même instanciation que dans la requête.

Le choix du nombre de termes *satisfaits* de la requête reste arbitraire. Dans ce

TAB. 4.1 – Agrégation quantifiée des termes de la requête $\Pi(Q | \theta)$

T_1	T_2	T_3	$\Pi(Q \theta)$
t_1	t_2	t_3	1
t_1	t_2	\bar{t}_3	1
t_1	\bar{t}_2	t_3	1
t_1	\bar{t}_2	\bar{t}_3	0
\bar{t}_1	t_2	t_3	1
\bar{t}_1	t_2	\bar{t}_3	0
\bar{t}_1	\bar{t}_2	t_3	0
\bar{t}_1	\bar{t}_2	\bar{t}_3	0

cas, cette attribution peut être une fonctionnalité du système, ou bien l'utilisateur peut spécifier dans sa requête le nombre de termes indexant le document à partir duquel il considère sa requête comme satisfaite. Par exemple, il peut introduire des quantificateurs du type « au moins deux termes » etc.

D'autre part, cette quantification, comme dans le cas d'une agrégation disjonctive de la requête, ne permet pas de discriminer entre les documents de la collection. En effet, seul le nombre de termes satisfaits est considéré. L'importance du terme satisfait (par exemple terme rare, terme fréquent dans la collection) n'est pas considérée.

4.3.3.2 Noisy OR

En général, les possibilités conditionnelles $\Pi(Q | \theta_i^l)$ ne sont pas agrégées par des booléens mais dépendent d'une évaluation appropriée des termes T_i .

La combinaison des termes de la requête peut être basée sur le « noisy-Or » [77]. Cet opérateur permet de quantifier les termes de la requête instanciés dans une configuration donnée comme dans la requête. Ces termes présents dans la configuration donnée conforme à la requête sont pondérés. Pour pouvoir discriminer entre les documents, plus ce nombre de terme croît, plus l'importance des termes instanciés avec la même valeur que dans la requête croît et plus la pertinence du document aura tendance à croître. Ce qui signifie que $\Pi(Q | \theta^l)$ est évaluée en termes de possibilités conditionnelles de la forme :

$$\Pi(Q | t_i \wedge_{k \neq i} \bar{t}_k) \quad (4.14)$$

et ce en utilisant une somme probabiliste. Soit

$$\Pi(Q | t_i \wedge_{k \neq i} \bar{t}_k) = 1 - q_i$$

Alors :

$$\begin{aligned} \Pi(Q | \theta^l) &= 0 \text{ si } \nexists T_i \in PAR_Q \text{ tel que } \theta_i^l = \theta_i^Q \\ &= \frac{1 - \prod_{i: t_i = \theta_i = \theta_i^Q} q_i}{1 - \prod_{T_k \in Par_Q} q_k} \text{ sinon} \end{aligned} \quad (4.15)$$

Uniquement, les termes instanciés positivement de la requête, $T_i = t_i$, apparaissent dans le numérateur. Le numérateur contient les termes de la configuration, dans le document en l'occurrence, ayant la même instanciation positive que dans la requête. La formule 4.15 permet de faire croître la pertinence finale d'un document donné. En effet, le score de pertinence d'un document donné croît de manière proportionnelle au nombre de termes qu'il contient ayant la même instanciation (positive) que dans la requête ¹.

Nous considérons que la présence ou l'absence d'un terme de la requête dans le document peut être quantifiée. Un terme fréquent dans toute la collection n'augmente pas forcément la pertinence du document étant donnée la requête. Par contre, un terme spécifique peut apporter une plus value à cette pertinence. Ainsi, plus un terme présent dans un document est spécifique, plus la pertinence du document en réponse à une requête qui contient ce terme augmente. La spécificité dans la littérature a été mesurée par la fréquence inverse

¹Nous supposons l'hypothèse du monde fermé ou Closed World Assumption (CWA) : $\Pi(Q | t_i) = \Pi(Q | t_i \wedge_{k \neq i} \bar{t}_k)$

du terme. Ainsi :

$$\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) = \frac{idf_i}{N} = nidf_i = 1 - q_i \quad (4.16)$$

Un avantage majeur de ce type d'agrégation est qu'il permet de résorber le problème de l'explosion combinatoire liée au calcul des possibilités conditionnelles. Nous rappelons qu'un des problèmes majeurs des réseaux Bayésiens est l'explosion combinatoire liée aux calculs des probabilités (ou possibilités dans notre cas) conditionnelles. Lorsque le nombre de parents ainsi que leurs domaines augmentent, le nombre de calculs des possibilités conditionnelles augmente d'une manière exponentielle.

4.3.4 Pondération des termes d'indexation

Nous présentons dans cette section les pondérations que nous avons proposées pour les termes d'indexation. Ces pondérations sont reliées aux relations de dépendance existantes entre un noeud terme et ses parents s'ils existent. En effet, lors du calcul de la pertinence d'un document étant donnée une requête, certains termes apparaissent dans le document et la requête et d'autres n'apparaissent que dans le document. Dans nos travaux actuels, les termes des documents absents des documents ne sont pas considérés lors des calculs de la pertinence. Cependant, ces termes peuvent être considérés. Un terme en relation sémantique ou statistique à un terme de la requête et présent dans un document peut apporter de l'information supplémentaire et peut constituer un élément intéressant à intégrer dans le calcul de la pertinence de ce document.

4.3.4.1 Arcs document-terme $\Pi(T_i \mid D_j)$

Pour évaluer la pertinence plausible et la pertinence certaine d'un document étant donnée une requête, nous avons besoin d'exprimer et de définir les arcs du réseau. Un arc reliant un noeud terme à un noeud document quantifie à quel point le terme est représentatif de ce document. Une absence d'arc entre un terme et un document traduit l'absence du terme en question du document. La représentativité des termes est selon notre approche considérée sous deux angles différents mais complémentaires. Nous rappelons que nous estimons que

la combinaison des facteurs $tf \times idf$ n'est pas l'unique approche permettant de donner un sens à la représentativité d'un terme d'un document donné. Ces deux facteurs sont définis sur des échelles différentes. Le premier est en rapport aux termes du document qu'il indexe. Le second facteur dépend des documents de la collection qu'il indexe. Les fréquences des termes d'un document donné sont intéressantes pour mesurer à quel point un document est exhaustif. La fréquence inverse permet de mesurer à quel point un terme est spécifique de la collection.

Nous voulons attribuer des poids aux termes sans induire de perte d'information. L'idéal serait d'avoir la connaissance de ces deux types d'information (spécificité et/ou exhaustivité). Dans la littérature, deux théories possibles sont connues par leur capacité d'interpréter sous deux angles une information ou une hypothèse. Ces deux théories sont la théorie de Dempster-Shafer et la théorie des possibilités. Dans notre approche, nous montrons que les résultats de l'une peuvent être retrouvés par l'autre et inversement.

Nous proposons donc dans ce qui suit deux méthodes pour quantifier les arcs reliant les documents aux termes qu'ils contiennent. Bien que ces deux méthodes paraissent similaires dans la définition de la représentativité d'un terme par rapport à un document qui le contient, elles produisent des performances différentes. Nous détaillons ce résultat dans le chapitre des expérimentations (*Chapitre 5*). Notre approche générale tente de distinguer les termes possiblement représentatifs des documents (ceux absents sont rejetés des représentations) de ceux qui sont nécessairement représentatifs. Ces derniers sont les termes qui suffisent à caractériser les documents.

Théorie de l'évidence pour pondérer les termes d'indexation La théorie de Dempster Shafer [105] (*DS*) est dérivée de l'approche bayésienne mais utilise deux mesures pour qualifier le degré de croyance que nous avons sur une hypothèse calculée à partir d'indices la confirmant ou l'infirmité. La théorie permet d'assigner une mesure de certitude aussi bien à des ensembles d'hypothèses qu'à des hypothèses seules. Cette approche permet de raisonner sur des ensembles d'hypothèses dans un premier temps et de se restreindre petit à petit aux hypothèses plausibles au fur et à mesure que de nouvelles évidences (informations) apparaissent. Contrairement aux mesures de probabilité, les fonctions de masse dans cette théorie sont définies sur tous les sous-ensembles du cadre de discernement et non pas seulement sur les singletons.

La théorie de DS a été introduite pour gérer l'incertitude et fusionner des informations provenant de sources multiples et hétérogènes.

Soit Ω l'ensemble des hypothèses exhaustives et mutuellement exclusives, et $P(\Omega)$ l'ensemble de toutes les disjonctions possibles sur Ω .

Supposons qu'il existe une source de données en relation d'une certaine manière avec Ω . On assigne aux éléments de $P(\Omega)$ une mesure de certitude grâce à l'affectation probabiliste m suivante :

1. $m(\emptyset) = 0$;
2. $\sum_{\forall E \in P(\Omega)} m(E) = 1$.

L'ensemble des hypothèses E vérifiant $m(E) > 0$ constitue l'*ensemble focal* de m . Les fonctions de croyance, Cr , et de plausibilité, Pl , sont définies par :

$$\forall E, F \in P(\Omega); Cr(F) = \sum_{E \subseteq F} m(E)$$

$$\forall F \in P(\Omega); Pl(F) = 1 - Cr(\overline{F})$$

La crédibilité, $Cr(F)$, est la somme de toutes les masses de croyance affectées aux ensembles qui font que F est certain. C'est donc une mesure de confiance totale accordée à chaque hypothèse ou sous-ensembles d'hypothèses.

La fonction de plausibilité, $Pl(F)$, est la somme des masses affectées aux ensembles qui ne sont pas inconsistants avec F , c'est-à-dire qui rendent F plausible.

Dans notre approche nous tentons de séparer les termes qui constitueraient de bons éléments pour représenter les documents de ceux qui permettent de pointer avec certitude vers certains documents. Ces termes lorsqu'ils figurent dans la requête doivent permettre d'éliminer les documents ne les contenant pas. Ils doivent aussi permettre de pointer vers un sous ensemble particulier de documents de la collection.

Une convention possible pour quantifier ces termes est basée sur les postulats donnés ci-dessous :

Postulat 1 : Un terme apparaissant dans un document est un terme représentatif de ce document ;

Postulat 2 : Un document est plus ou moins nécessairement sélectionné par un terme lorsque ce terme apparaît avec une fréquence élevée dans ce document et un nombre d'apparitions faible dans les autres documents de la collection.

Les arcs reliant les noeuds documents aux termes d'indexation sont évalués par des fonctions de masse proposées par la théorie de *DS* [105]. La théorie des probabilités imprécises suppose *l'existence* d'une mesure de probabilités P sur le référentiel Ω d'une étude donnée, mais celle-ci n'est pas parfaitement connue. Le modèle de Dempster est un cas particulier des probabilités imprécises. Les masses sont affectées à des propositions élémentaires et à leurs disjonctions. Le cadre de discernement, Θ_i , d'un terme T_i , est $\Theta_i = \{t_i, \bar{t}_i\}$. Les masses de probabilité sont affectées aux sous ensembles :

$$\{t_i\}, \{\bar{t}_i\}, \{t_i, \bar{t}_i\}, \emptyset$$

pour signifier respectivement que le terme est :

- sûrement représentatif;
- sûrement non représentatif;
- doté d'une représentativité encore non connue ou conflictuelle.

La masse affectée à t_i traduit la certitude de représentativité du document par le terme t_i lorsque ce terme est présent dans le document. La probabilité de base affectée (*bpa*), notée m , est une fonction telle que :

$$m(\emptyset) = 0, \quad m(t_i) + m(\bar{t}_i) + m(\Theta_i) = 1 \quad (4.17)$$

D'une manière générale, $m(A)$ mesure la croyance placée exactement sur A . nous définissons les fonctions de masse, dans le contexte de $D_j = d_j$, par :

$$m(\{t_i\} | d_j) = nt f_{ij}, \quad m(\{t_i, \bar{t}_i\} | d_j) = 1 - nt f_{ij},$$

où $nt f_{ij}$ la fréquence normalisée, $nt f_{ij} = \frac{t f_{ij}}{\max_{t_k \in d_j} (t f_{kj})}$.

La croyance placée exactement sur le terme t_i lorsqu'il est présent dans le document d_j est proportionnelle à son nombre d'apparitions dans ce document. Le terme t_i présent dans le document d_j est certainement représentatif de ce document au moins au degré $nt f_{ij}$ ². Cette certitude de représentativité dépend donc de l'importance du terme à l'intérieur du document qu'il indexe. La seconde masse, $1 - nt f_{ij}$, peut être attribuée librement à tout élément de Θ_i .

²Nous supposons à ce stade que nous avons éliminés les mots vides, etc.

Lorsqu'une fonction de masse m est établie sur une famille d'éléments focaux consonants (ou propositions emboîtées) sur Ω , la construction de la fonction de croyance correspondante, Cr , donne une mesure de nécessité, N , (et une mesure de possibilité duale Π), respectant les propriétés :

$$Cr = N \text{ et } Pl = \Pi$$

Dans notre contexte, nous voulons définir des degrés de possibilité de pertinence. Ils sont obtenus respectivement à partir des fonctions de masse. La possibilité conditionnelle, $\Pi(T_i \mid d_j)$ est donc définie comme la fonction de plausibilité de Shafer du fait qu'il y a emboîtement entre les ensembles focaux. Dans le contexte $D_j = d_j$, nous obtenons :

$$\begin{aligned} \Pi(t_i \mid d_j) &= 1; \\ \Pi(\bar{t}_i \mid d_j) &= 1 - ntf_{ij} \end{aligned} \quad (4.18)$$

Donc $1 - ntf_{ij}$ représente le degré de possibilité de non pertinence du terme t_i dans d_j .

Un terme important dans la collection est un terme qui apparaît avec une fréquence élevée dans peu de documents de la collection [61]. Nous supposons que la certitude de restituer un document pertinent au moyen d'un terme est reliée à l'importance du terme dans toute la collection. Nous traduisons l'importance d'un terme t_i pour restituer un document d_j par :

$$\phi_{ij} = \frac{\log \frac{N}{n_i}}{\log(N)} \cdot ntf_{ij} \quad (4.19)$$

avec N le nombre de documents de la collection ;

n_i le nombre de documents contenant le terme t_i .

Lorsque nous sommes dans le contexte de $D_j = \bar{d}_j$, une *bpa* est définie par :

$$m(\{\bar{t}_i\} \mid \bar{d}_j) = \phi_{ij}, \quad m(\{t_i, \bar{t}_i\} \mid \bar{d}_j) = 1 - \phi_{ij}$$

ϕ_{ij} est interprétée comme le degré auquel si d_j n'est pas pertinent, alors le terme t_i devrait être rejeté de la requête. En effet, dans ce cas, le terme t_i dirige la requête vers des documents non pertinents, et son utilisation dans la requête ne va pas permettre de restituer les documents pertinents. Si t_i est non discriminant pour d_j alors le fait que d_j ne soit pas un document pertinent nous laisse libre d'utiliser le terme t_i ou pas dans la requête. Comme précédemment, nous obtenons :

$$\begin{aligned} \Pi(\bar{t}_i \mid \bar{d}_j) &= 1 \\ \Pi(t_i \mid \bar{d}_j) &= 1 - \phi_{ij} \end{aligned} \quad (4.20)$$

Le tableau 4.2 montre les possibilités conditionnelles des termes dans le contexte de leur parent.

TAB. 4.2 – Possibilités conditionnelles $\Pi(T_i | D_j)$

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

Pondération des arcs basée sur la logique possibiliste La seconde approche que nous proposons pour pondérer les termes des documents est basée sur la logique possibiliste. Comme nous l'avons montré dans la section précédente, nous pouvons retrouver les degrés de possibilité et de nécessité à partir de la pondération par la théorie de Dempster Shafer. La différence repose sur le sens que nous donnons aux mesures. Nous décrivons plus loin dans cette section l'équivalent de la définition des poids dans le cadre de la théorie de Dempster.

Nous essayons dans notre approche encore une fois d'exprimer de manière plus complète, comparée aux modèles actuels, la pondération d'un terme. Une unification possible de la notion de représentativité serait : « la représentativité d'un terme par rapport à un document décrirait le point auquel un document est à propos du sujet concerné par le terme ».

De ce fait, dans notre cadre de travail, la théorie des possibilités, nous disposons deux degrés pour évaluer la possibilité et la nécessité attachées à des propositions.

Nous traduisons la nécessaire représentativité et la plausible représentativité d'un terme basé sur les deux postulats suivants :

Postulat 1 : Un terme est plus ou moins possiblement représentatif du document s'il apparaît fréquemment dans ce document ;

Postulat 2 : Un terme est plus ou moins nécessairement représentatif du document s'il apparaît fréquemment dans ce document et rarement dans les autres documents de la collection.

D'après le *Postulat 1*, $\Pi(t_i/d_j)$ peut être estimée à partir de la fréquence tf :

$$\Pi(t_i/d_j) = n f_{t_{ij}} \quad (4.21)$$

Un terme de poids 0 est un terme non compatible avec le document. Si son poids vaut 1, alors le terme est possiblement représentatif du document. Ici,

« représentatif » ne doit pas nécessairement être compris dans le sens général. Il signifie, dans ce contexte, « utilisé pour restituer ce document à partir de la collection ». Un terme représentatif dans le sens général, est un terme qui peut ne pas être utile, ni d'une grande aide pour restituer un document. Supposons un document de la collection qui traite de la *logique floue*. Le mot « floue » est très représentatif mais est potentiellement non utile s'il ne caractérise pas le document parmi d'autres documents ayant le même sujet (traitant du même domaine). Un terme n'apparaissant pas dans un document est un terme non compatible avec le document et s'il apparaît avec une fréquence maximale, alors le terme est un candidat possible pour le représenter ³.

Nous pouvons *basculer* facilement dans le cadre de la théorie de Dempster-Shafer. Pour ce faire, il aurait simplement fallu affecter les masses suivantes :

$$\begin{aligned} m(\{\bar{t}_i\} | d_j) &= 1 - ntf_{ij} \\ m(\Theta_i | d_j) &= ntf_{ij} \end{aligned}$$

Ici, la croyance qu'un terme est non représentatif d'un document est égale à $1 - ntf$. Un terme t_i non présent dans un document donné a une fréquence égale à 0. La croyance que ce terme est certainement non représentatif de ce document est donc égale à 1.

Dans le cadre de la théorie de *DS*, comme nous l'avons montré dans la section précédente, nous nous sommes penchés sur *l'information positive* alors que la théorie des possibilités utilise *l'information négative*. Dans le premier cadre (*DS*), nous avons mesuré la croyance que nous avons sur un terme à représenter un document qui le contient. Cette croyance est calculé par ntf . La certitude de représentativité d'un terme d'un document donné, dépend de sa distribution à l'intérieur de ce document. Dans la théorie des possibilités, la pondération telle que nous la suggérons, revient à mesurer le degré auquel un terme est non représentatif d'un document donné. Cependant, se plaçant du côté *possibiliste* la définition que nous avons faite nous a semblé *naturelle* et *intuitive*. Nous avons mesuré le degré auquel un terme est possiblement représentatif d'un document donné et nous l'avons défini comme fonction de sa fréquence. Il ne nous a pas semblé intéressant d'affirmer qu'un terme présent dans un document est complètement représentatif de ce document. En effet, il nous a paru impossible d'affirmer cette complétude de possibilité sachant qu'un terme t_1 pourrait

³A ce stade, nous laissons de côté les relations entre termes, telle que la synonymie par exemple

apparaître 1 fois dans un document d_1 contenant 1000 termes apparaissant en moyenne 30 fois dans ce document. Dans ce cas, t_1 est complètement négligeable dans ce document.

Un terme discriminant dans la collection est un terme qui apparaît (souvent) dans peu de documents de la collection. Nous supposons qu'un terme discriminant est un terme qui est nécessairement représentatif d'un document et donc contribue certainement à le sélectionner parmi d'autres documents. Nous définissons un degré de nécessaire pertinence, ϕ_{ij} , d'un terme t_i pour représenter un document d_j comme un poids de la forme :

$$\phi_{ij} = \mu_1 \left(\frac{N}{n_i} \right) * \mu_2 (nf_{t_{ij}}) \quad (4.22)$$

où $*$: opérateur produit ;

μ_1, μ_2 : fonctions de normalisation. Par exemple, μ_1 fonction logarithmique, μ_2 fonction identité. Alors :

$$\phi_{ij} = \frac{\log \frac{N}{n_i}}{\log(N)} \cdot nf_{t_{ij}} \quad (4.23)$$

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

$$N(t_i \rightarrow d_j) = \phi_{ij} \quad (4.24)$$

puisque

$$\Pi(\overline{d_j}) = 1 \text{ a priori,} \quad (4.25)$$

$$\text{alors } \Pi(t_i | \overline{d_j}) = \Pi(t_i \wedge \overline{d_j}) = 1 - N(t_i \rightarrow d_j) = 1 - \phi_{ij} \quad (4.26)$$

$$\text{et } \Pi(\overline{t_i} | \overline{d_j}) = 1 \quad (4.27)$$

Dans le tableau 4.3, nous résumons les possibilités conditionnelles des termes d'indexation étant donnée l'instanciation du noeud document parent.

TAB. 4.3 – Possibilités conditionnelles $\Pi(T_i | D_j)$

	d_j	$\overline{d_j}$
t_i	$nf_{t_{ij}}$	$1 - \phi_{ij}$
$\overline{t_i}$	1	1

4.3.4.2 Termes racines

Dans notre approche, un terme discriminant absent du document, diminue la pertinence de ce document. Un terme non discriminant a moins d'influence sur la pertinence qu'un terme discriminant.

Dans la littérature, un facteur connu pour mesurer le pouvoir discriminant d'un terme t_i dans la collection est la fréquence inverse du terme dans la collection, idf_i , ou $nidf_i$. Une valeur du facteur idf élevée signifie que le terme est discriminant et inversement. L'impact de l'absence d'un terme de la requête du document est mesurée dans notre cas par :

$$\begin{aligned} \forall T_i \notin \mathcal{T}(D_j), \quad \Pi(\theta_i) &= 1 \text{ si } \theta_i^Q = \bar{t}_i \\ &= 1 - nidf_i \text{ sinon} \end{aligned} \quad (4.28)$$

avec $nidf_i = \frac{\log \frac{N}{n_i}}{\log(N)}$. Un terme non discriminant a une faible valeur de $nidf$ donc une possibilité marginale élevée qui ne décroîtra pas de manière significative la pertinence du document. Un terme discriminant absent du document peut décroître, voire annuler, la pertinence du document.

Une des caractéristiques de notre modèle est de pouvoir rester flexible. Ainsi, lorsque un terme donné a une valeur de $nidf = 0$, nous ajoutons une constante pour ne pas éliminer le document ne le contenant pas de la liste des documents restitués. Nous proposons à la suite de la définition de notre modèle trois nouveaux facteurs permettant de calculer le pouvoir discriminant d'un terme dans la collection.

4.3.5 Possibilité *a priori* des documents

En absence d'information, la possibilité *a priori* d'un noeud document est uniforme

$$\Pi(d_j) = \Pi(\bar{d}_j) = 1$$

Nous pouvons obtenir des connaissances sur les documents étant donnée l'importance de leurs termes, leur longueur etc. Cette connaissance peut être donnée par un utilisateur, le profil utilisateur etc. Si nous sommes intéressés par les documents longs, la possibilité *a priori* d'un document instancié à $D_j = d_j$

devient :

$$\Pi(d_j) = \frac{l_j}{\max_{k=1,\dots,N} l_k} = nl_{d_j} \quad (4.29)$$

avec l_j la longueur du document d_j en terme de fréquence ; $l_j = \sum_i tf_{ij}$. Plus le document est court, moins il est pertinent. De plus, $\Pi(\overline{d_j}) = 1$.

4.4 Nouveaux facteurs de discrimination

Nous exposons dans cette section les motivations qui nous ont conduit à proposer de nouveaux facteurs de discrimination pour les termes de la collection. Nous proposons trois facteurs, que nous détaillons et illustrons d'un exemple. Nous montrons aussi un comparatif de ces facteurs avec les facteurs de discrimination connus de la littérature.

4.4.1 Motivations

Le processus d'indexation produit des représentations des documents et de la requête sous forme de liste de termes pondérés. Les poids affectés aux termes sont le plus souvent le résultat d'une vue fréquentiste des probabilités parce qu'il n'y a pas d'autres informations ou connaissances disponibles. Les informations utilisées pour la pondération des termes d'indexation sont, comme nous l'avons mentionné, obtenues par la combinaison des facteurs $tf \times idf$. Les modèles actuels ne tiennent généralement compte que des termes de la requête présents dans les documents pour le calcul de la pertinence. Il nous paraît plus judicieux selon notre approche de considérer tous les termes de la requête, présents ou absents dans les documents. En effet, lors du calcul de la pertinence, nous considérons que peu de données sont disponibles. A notre sens, l'ignorance des termes de la requête absents des documents réduit d'autant plus ces données. Cette information (absence des termes) découle naturellement de la topologie du réseau présentée dans la figure 4.1.

L'utilisateur ignore la représentation des documents lors de la formulation de son besoin. Pourtant, les termes qu'il utilise dans sa requête sont généralement

d'une grande importance pour le système. Dans notre vision, ces termes sont la donnée la plus *sûre* sur laquelle peut se baser le calcul de la pertinence d'un document. Le processus de propagation de notre système déclenché par la réception d'une requête donnée propage l'information à travers ces termes *actifs*. Dans ce qui suit, nous proposons trois facteurs, df_k ; $k \in \{1..3\}$, qui donnent aux termes un pouvoir discriminant. Ce pouvoir peut être utilisé de différentes façons.

- Il peut aider à discriminer entre les documents de la collection. Dans ce cas, il est affecté aux termes des documents communs à ceux de la requête. Le facteur *idf* est généralement utilisé dans ce sens dans les modèles actuels ;
- Il peut éliminer des documents de la liste des documents restitués en réponse à une requête. Dans ce cas, il est affecté aux termes de la requête absents des documents.

Nous notons tout de même que le facteur *idf* et les facteurs que nous proposons peuvent être utilisés indifféremment de la première ou la seconde façon. Cependant, dans nos expérimentations sur des collections de tests, l'utilisation de ces facteurs n'a pas le même effet sur le résultat de la recherche en fonction de leur utilisation (renforcer la pertinence vs décroître la pertinence).

Les facteurs que nous proposons s'appuient sur les connaissances de la répartition des termes dans les documents en fonction de leur densité dans les documents. Chacun de ces facteurs gère l'incertitude engendrée par le processus d'indexation d'une manière spécifique. Ces facteurs sont exposés et argumentés dans ce qui suit.

4.4.2 Pouvoir de discrimination

La densité d'un terme $\frac{tf_{ij}}{l_j}$ ($l_j = \sum_{t_i \in d_j} tf_{ij}$), indique à quel point un terme est présent dans le document. Un terme, t_1 , qui apparaît 10 fois dans un document de longueur 100, n'a pas la même importance dans le document qu'un terme, t_2 , qui apparaît 50 fois dans ce même document. Dans ce cas, le terme t_2 est plus dense que le terme t_1 dans ce document. La situation idéale serait que tous les termes de la requête utilisateur soient denses mais dans peu de

documents de la collection. Un terme dense dans peu de documents de la collection permet de pointer vers un sous-ensemble particulier de documents de la collection. Dans ce cas, ce terme est dit doté d'un **pouvoir discriminant**. Nous définissons un facteur de discrimination, df_{1_i} , du terme t_i qui mesure le pouvoir discriminant de ce terme. Il est inversement proportionnel à la distribution de la densité du terme t_i dans la collection.

Ce facteur est maximisé pour un terme donné, lorsque ce terme apparaît peu de fois dans les documents et dans peu de documents de la collection. Ce terme lorsqu'il apparaît dans la requête pourrait à notre sens aider à pointer vers ce petit sous ensemble de documents. La distribution de la densité du terme t_i dans la collection est égale à la somme de ses densités dans les documents qu'il indexe. Plusieurs normalisations de la densité sont possibles et nous en proposons deux dans les mesures 4.30 et 4.31.

$$df_{1_i} = \frac{N}{\sum_{j: t_i \ni d_j} \alpha_{ij}}, \quad \alpha_{ij} = \frac{\frac{tf_{ij}}{l_j}}{\max_{t_l \in T, d_k \in N} \frac{tf_{lk}}{l_k}} \quad (4.30)$$

avec :

l_j : longueur du document D_j ;

N : ensemble des documents de la collection ;

T : ensemble des termes de la collection.

Le facteur α_{ij} est normalisé par rapport au maximum obtenu sur la densité de tous les termes de la collection (4.30) et le facteur $\alpha_{1_{ij}}$ par rapport aux documents indexés par le terme t_i (4.31).

$$df_{1_i} = \frac{N}{\sum_{j: t_i \ni d_j} \alpha_{1_{ij}}}, \quad \alpha_{1_{ij}} = \frac{\frac{tf_{ij}}{l_j}}{\max_{k: t_i \ni d_k} \frac{tf_{ik}}{l_k}} \quad (4.31)$$

La normalisation est un point crucial dans ces mesures. Les conclusions sur le comportement du facteur df_{1_i} d'un terme t_i dépendent de la normalisation choisie. Soient deux termes t_1 , t_2 répartis dans les documents D_1 , D_2 tels que présentés dans le tableau 4.4.

Il est clair dans cet exemple, que la valeur du dénominateur dans la mesure 4.31 pour les deux termes est presque identique. Les sommes obtenues pour les termes t_1 et t_2 valent respectivement $\sum_j \alpha_{1_{1j}} = 1.96$ et $\sum_j \alpha_{1_{2j}} = 1.545$. Une normalisation de la densité d'un terme t_i par rapport à la distribution de la densité de ce terme dans les documents qu'il indexe ne permet pas de conclure sur l'importance du terme dans la collection puisque dans ce cas il est

TAB. 4.4 – Distributions des termes t_1 et t_2 dans les documents D_1 et D_2

	D_1	D_2
t_1	10	20
t_2	2	2
	D_1	D_2
l_j	12	22

uniformisé par rapport à sa distribution dans la collection. Une normalisation par rapport à la distribution de tous les termes de la collection serait plus porteuse de sens. Pour ces mêmes deux termes, nous obtenons $\sum_j \alpha_{1j} = 1.916$ et $\sum_j \alpha_{2j} = 0.283$.

Un inconvénient du facteur df_1 proposé concerne son dénominateur. Une valeur élevée de ce dénominateur n'impliquerait pas forcément que le terme a un pouvoir discriminant (comme nous l'entendons). En effet, il faut connaître les raisons de cette valeur élevée et pouvoir affirmer si elle provient de faibles densités mais dans un nombre élevé de documents ou s'il s'agit de densités élevées dans peu de documents de la collection. Dans le premier cas, le terme ne discrimine pas entre les documents de la collection.

Un terme qui apparaît dans un document court peut s'avérer plus discriminant qu'un terme qui apparaît dans un document long.

Nous présentons dans ce qui suit, une illustration du comportement de ce facteur sur quatre termes de la collection.

Illustration Soit une collection contenant 4 documents, et les termes t_1 , t_2 qui apparaissent dans les documents D_1 , D_2 . Le tableau 4.5 donne la répartition des termes t_1 et t_2 dans les documents qu'ils indexent. La mesure utilisée dans cet exemple est celle donnée par 4.31. α_{1j} , α_{2j} correspondent aux valeurs de t_1 et t_2 respectivement ;

l_j désigne la longueur du document d_j , tf le nombre d'apparitions du terme dans le document.

Considérons deux autres termes t_3 et t_4 qui sont répartis dans les documents D_3 et D_4 tels que présentés dans le tableau 4.6.

Le tableau 4.7 présente la valeur df_{1i} obtenue pour chaque terme t_i . Les termes

TAB. 4.5 – Répartition des termes t_1 et t_2 dans la collection

	D_1	D_2
tf_{t_1}	29	49
tf_{t_2}	2	5
l_j	31	54
α_{1j}	1	0.969
α_{2j}	0.068	0.098

TAB. 4.6 – Répartition des termes t_3 et t_4 dans la collection

	D_3	D_4
tf_{t_3}	1	1
tf_{t_4}	9	4
l_j	10	5
α_{3j}	0.106	0.213
α_{4j}	0.962	0.855

t_3 et t_2 ne sont pas denses dans les documents qu'ils indexent. Ils apparaissent relativement dans peu de documents de la collection. Ces termes maximisent le facteur df_1 . Ce comportement vérifie bien ce que nous avons prédit. Les documents contenant le terme t_2 sont plus longs que ceux contenant le terme t_3 . La longueur des documents n'est pas éliminée des calculs puisque le poids du facteur df_{1_2} du terme t_2 est plus élevé que celui du terme t_3 .

Le terme t_1 minimise le facteur df . Cependant, nous aurions peut être préféré qu'il le maximise puisque ce terme est plus dense dans les documents dans lesquels il apparaît que les termes t_2 et t_3 . Le terme t_4 est encore moins préféré et ceci est lié à son apparition dans des documents relativement courts de la collection.

Dans cet exemple nous aurions envie de dire que les termes les plus discriminants minimisent le facteur df_1 . Cependant, il n'est pas garanti que la valeur élevée d'un dénominateur donné dans les mesures 4.31 et 4.30 traduit une densité élevée dans peu de documents de la collection. Elle pourrait traduire une faible densité dans un grand nombre de documents de la collection. Le second facteur de discrimination que nous proposons tente de tenir compte davantage de l'importance d'un terme donné à l'intérieur des documents qu'il indexe.

TAB. 4.7 – Pouvoir discriminant des termes t_1, t_2, t_3, t_4

	t_1	t_2	t_3	t_4
df_{1_i}	2.03	23.81	12.47	2.20

4.4.3 Discrimination par fréquence normalisée pondérée

Le second facteur proposé, df_{2_i} , considère la distribution des termes dans la collection étant données leur fréquence et la longueur des documents qu'ils indexent. Généralement les documents longs ont des nombres d'apparition des termes plus grands que les documents courts. Si nous prenons en exemple ce manuscrit, le mot *information* risque d'avoir un nombre d'apparitions plus grand que dans un article de 12 pages traitant de la recherche d'information. D'autre part, les documents longs peuvent contenir des termes d'indexation plus variés que les documents courts. Nous définissons un facteur, α'_{ij} , qui quantifie l'importance du terme t_i dans le document D_j en relation avec la longueur du document, l_j , par :

$$\alpha'_{ij} = ntf_{ij} \times l_j \quad (4.32)$$

avec :

$$ntf_{ij} = \frac{tf_{ij}}{\max_k tf_{kj}}$$

L'importance d'un terme dans un document est quantifiée par sa fréquence normalisée. Une fréquence normalisée élevée pour un terme dans un document donné, indique que ce terme serait possiblement représentatif de ce document (ou apporte un sens au sujet décrit par le document). Une valeur élevée de la mesure 4.32 indique que le terme est *important* dans un document long.

Le facteur α' *somme* mesure l'importance d'un terme dans les documents. Un terme qui apparaît « fortement » dans les documents longs aura une valeur α' *somme* élevée. Une valeur élevée de la fréquence normalisée pondérée α' *somme* _{i} du terme t_i indiquerait son apparition élevée dans des documents longs. Inversement, une faible valeur de la mesure 4.33 suggère les points :

- (1) le terme apparaît peu fréquemment dans les documents ;
- (2) le terme apparaît dans des documents courts ;

– ou (1) et (2) à la fois.

L'importance de la distribution du terme t_i dans la collection est quantifiée par :

$$\alpha'_{somme_i} = \sum_{k: t_i \ni D_k} n t f_{ik} \times l_k \quad (4.33)$$

Intuitivement, nous considérons que plus un terme apparaît fortement dans un grand nombre documents de la collection et moins ce terme est spécifique à la collection et inversement. Ainsi, le pouvoir discriminant, df_2 , ou de spécificité est alors inversement proportionnel au facteur α'_{somme} et est mesuré par 4.34. Les termes les plus spécifiques devraient maximiser le facteur df_2 .

$$df_{2_i} = \frac{\sum_{j=1}^N l_j}{\alpha'_{somme_i'}} \quad (4.34)$$

avec N : nombre de documents dans la collection.

D'une manière générale, ce facteur semble plus adéquat pour retrouver les termes qui apparaissent à la fois faiblement et dans des documents courts. Dans ce cas, la valeur α'_{somme} est faible et ces termes maximisent le facteur df_2 . Lorsque α'_{somme_i} du terme t_i est élevée il est délicat d'affirmer que ce terme apparaît fortement dans des documents longs de la collection et dans un nombre de documents faible.

Pour conclure, le facteur df_2 est intéressant pour éliminer les termes non intéressants de la collection. Par exemple, lorsque un terme de la requête a une faible valeur df_2 , cela signifie que ce terme ne permet pas de discriminer entre les documents de la collection. La seule contribution de ce terme serait d'aider à restituer des documents courts dans lesquels il apparaît faiblement. Nous pouvons *à la limite* ne pas le considérer lors du calcul du score de pertinence d'un document vis à vis de la requête (la requête qui contient ce terme).

Illustration Le tableau 4.8 donne le nombre maximal d'apparitions, max_{t_f} , dans chaque document. Supposons la répartition des termes t_1, t_2, t_3, t_4 telle que donnée dans le tableau 4.9. Le tableau 4.10 donne le df_{2_i} obtenu pour chacun des termes. Les termes les plus rares dans la collection maximisent df_2 . La spécificité ici mesure à quel point un terme est rare dans les documents et

TAB. 4.8 – Fréquence maximale dans les documents

	D_1	D_2	D_3	D_4
max_{tf}	29	49	9	4

TAB. 4.9 – Répartition des termes dans les documents

	D_1	D_2	D_3	D_4
tf_{t_1}	29	49	0	0
tf_{t_2}	2	5	0	0
tf_{t_3}	0	0	1	1
tf_{t_4}	0	0	9	4
l_j	30	50	10	5
α'_{1j}	31	54	0	0
α'_{2j}	2.137	5.510	0	0
α'_{3j}	0	0	1.11	1.25
α'_{4j}	0	0	10	5

dans la collection d'une manière générale. Nous classons les termes par ordre décroissant de leur spécificité comme (t_3, t_2, t_4, t_1) . Le terme t_1 est le moins préféré au sens du facteur df_2 . La longueur des documents est éliminée par ce facteur. *A priori* les documents courts sont les préférés de cette mesure. Le terme t_3 apparaît faiblement dans des documents relativement courts de la collection. Ce terme est préféré au terme t_2 qui apparaît faiblement dans des documents relativement longs. Les termes t_1 et t_4 minimisent le facteur. Ils apparaissent fortement dans des documents courts (le cas du terme t_4) et longs (cas du terme t_1). Ces illustrations ne sont pas exhaustives, la collection de documents contient très peu de documents et très peu de termes. Dans l'exemple comparatif avec les facteurs connus de discrimination ces résultats sont mieux explicités.

4.4.4 Discrimination par entropie

La formule de Shannon, l'entropie, a été tout d'abord élaborée en vue de modéliser la transmission de signaux électriques [106]. Alors que Shannon parle de transmission de quantité d'information, nous proposons une formule repo-

TAB. 4.10 – Pouvoir discriminant des quatre termes

	t_1	t_2	t_3	t_4
df_{2i}	1.176	13.075	42.356	6.666

sant sur la théorie d'entropie pour mesurer la production d'information apportée par les termes de la collection. L'information produite par la densité d'un terme dans un document donné, peut être agrégée pour mesurer l'entropie d'un terme dans tous les documents de la collection. Nous essayons d'évaluer la concentration (ou la dispersion) d'un terme dans la collection par le facteur de discrimination, df_{3i} .

$$df_{3i} = - \sum_j p_{ij} \log p_{ij}; \quad (4.35)$$

$$p_{ij} = \frac{\frac{tf_{ij}}{l_j}}{\sum_{l,k} \frac{tf_{lk}}{l_k}}$$

p_{ij} est la probabilité de densité du terme t_i dans le document D_j . Cette probabilité peut être obtenue par rapport à tous les documents contenant le terme t_i . Dans ce cas, dans 4.35, $l = i$ et k est tel que $t_i \ni D_k$. La seconde normalisation est en rapport à tous les termes de la collection. Dans ce cas, l est tel que $t_l \in T$; T est le nombre de termes de la collection; $l = 1, \dots, T$ et $k = 1, \dots, N$, avec N le nombre de documents de la collection. Nous avons choisi le second cas pour une meilleure conformité avec le cadre probabiliste.

Plus la valeur df_{3i} est faible, moins la distribution est uniforme, et plus le terme serait intéressant pour pointer vers certains documents particuliers et inversement. Pour mesurer la production de non information lorsqu'un terme de la requête est absent du document, nous pouvons affecter le facteur df_3 aux termes de la requête absents des documents. Plus le terme est doté d'un pouvoir discriminant, moins le document ne le contenant pas est pertinent pour la requête. Ce facteur de discrimination généralise le facteur connu de spécificité idf . Il est clair que deux documents ayant la même valeur de idf peuvent avoir des valeurs de df_3 différentes. idf est moins discriminant que df_3 .

4.4.5 Exemple comparatif

Le but de cette section est de comparer les différents facteurs de discrimination proposés avec deux facteurs de discrimination connus dans la littérature. La collection est composée de 5 documents :

$$D_1 = \{6t_4, 4t_9\}, D_2 = \{3t_2, 10t_3, 15t_5, 6t_6, 10t_7, 12t_8\},$$

$$D_3 = \{t_1, t_2, t_5\}, D_4 = \{t_1, 15t_3, t_4\}, D_5 = \{15t_1, 15t_2, 15t_3\}$$

La notation ainsi définie, désigne que les termes t_4 et t_9 apparaissent respectivement 6 et 4 fois dans le document D_1 .

Nous avons tenté d'avoir une collection exhaustive, les longueurs des documents varient. Le terme t_3 a la même importance dans deux documents distincts, mais la distribution des termes dans ces deux documents est différente. Certains termes apparaissent peu fréquemment dans la collection, et peuvent apparaître aussi bien dans des documents longs que courts. Différents facteurs de discrimination ont été testés, pour trouver une corrélation entre les termes d'indexation et les documents. Ces facteurs tentent de discriminer entre les documents de la collection et sont comparés à des discriminants connus tel que *idf* ou $\log\left(\frac{N-n_i+0.5}{n_i+0.5}\right)$ utilisés respectivement dans les systèmes *SMART* et *OKAPI*.

Les facteurs appliqués aux termes de la collection sont présentés dans le tableau 4.11. Les facteurs ndf_i ; $i = 1, \dots, 3$ sont normalisés par rapport à tous les termes de la collection pour obtenir une valeur comprise entre 0 et 1. Ainsi :

$$ndf_{1_i} = \frac{df_{1_i}}{\max_{t \in T} df_{1_t}}, \quad ndf_{2_i} = \frac{df_{2_i}}{\max_{t \in T} df_{2_t}}, \quad ndf_{3_i} = \frac{df_{3_i}}{\max_{t \in T} df_{3_t}}$$

T est le nombre de termes de la collection.

Le facteur ndf_3 dans le tableau 4.11 minimise l'entropie lorsque les termes apparaissent rarement dans la collection et dans les documents longs. Ce facteur préfère les termes t_6 et t_7 qui apparaissent dans un seul document de la collection, D_2 . Ce document est le plus long de la collection. Ce facteur semble préférer les termes qui apparaissent rarement dans les documents. Le terme t_6 semble plus discriminant que le terme t_7 bien que celui-ci (t_7) soit plus fréquent dans le document D_2 . Cette dernière constatation est vraie pour les couples de termes (t_6, t_7) , (t_7, t_8) et (t_6, t_8) . Le terme t_6 est le plus spécifique alors que son apparition dans ce document est la plus faible (parmi t_7 et t_8). Le terme t_9

TAB. 4.11 – Facteurs de discrimination

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
$nidf_i$	0.222	0.222	0.222	0.398	0.398	0.699	0.699	0.699	0.699
$\log\left(\frac{N-n_i+0.5}{n_i+0.5}\right)$	-0.146	-0.146	-0.146	0.146	0.146	0.477	0.477	0.477	0.477
ndf_{1_i}	0.147	0.148	0.076	0.162	0.178	1	0.6	0.5	0.267
ndf_{2_i}	0.135	0.112	0.067	0.598	0.112	0.297	0.178	0.18	1
ndf_{3_i}	0.682	0.676	1	0.506	0.556	0.135	0.196	0.222	0.333

est aussi relativement doté d'un pouvoir discriminant sur les documents de la collection. Ce terme apparaît relativement peu fréquemment dans un document relativement court.

Le facteur ndf_2 préfère les termes rares dans les documents courts. Le terme t_9 maximise ce facteur. Ce terme apparaît dans un seul document relativement court de la collection. Le terme t_4 figure aussi parmi les termes qui maximisent ce facteur. En effet, ce terme n'apparaît que dans deux documents relativement courts. Le classement de ce terme s'explique donc aussi bien par le fait que le terme n'apparaît pas fréquemment dans les documents (D_4) que par le fait que les documents qu'il indexe sont relativement courts (D_1 et D_4). Le terme t_3 est le moins spécifique par ndf_2 . Il apparaît avec un nombre d'apparitions relativement élevé et dans des documents relativement longs.

Le facteur ndf_1 retrouve à peu près les résultats du facteur ndf_3 . Les termes les plus rares de la collection maximisent ce facteur.

Le facteur $nidf$ et les facteurs $nidf_1$ et $nidf_3$ retrouvent d'une manière équivalente les classements des termes. Alors que le facteur $nidf$ calcule la discrimination d'un terme donné en fonction de sa simple absence-présence dans les documents, les facteurs que nous proposons tentent de donner un sens plus *précis* à la discrimination. Nous nous intéressons davantage à la présence-absence qu'à l'importance des termes à l'intérieur des documents de la collection. Cette « importance » est mesurée par la densité du terme dans le document ($\frac{tf}{l}$) ou par la « fréquence normalisée pondérée » d'un terme dans le document.

4.5 Illustration du modèle proposé

Nous reprenons dans cette section l'exemple de la collection donnée dans le premier chapitre de ce manuscrit. Il s'agit de répondre à la requête Q contenant une fois chacun des termes t_2, t_3 , et t_6 . Nous comparons à la fin de cette section notre classement aux classements des modèles vectoriel, probabiliste et d'inférence.

$$\begin{aligned} D_1 &= \{4t_1, 6t_4\}; & D_2 &= \{20t_2, 10t_3, 15t_5, 5t_6\}; & D_3 &= \{t_2, t_3, t_5\}; \\ D_4 &= \{t_2, 15t_3, 10t_5\}; & D_5 &= \{15t_1, 15t_2, 15t_3\}; \\ Q &= \{t_2, t_3, t_6\} \end{aligned}$$

La réception de la requête instancie le système. Le processus de propagation de l'information apportée par la requête entraîne le calcul des possibilités conditionnelles de chaque document étant donnée la requête selon la topologie du graphe. Pour calculer les possibilités de pertinence d'un document donné, nous avons besoin de calculer la possibilité jointe $\Pi(Q \wedge D_j)$ donnée par :

$$\Pi(Q \wedge D_j) = \max_{\theta^l \in \theta} (\Pi(Q \mid \theta^l) \cdot \prod_{T_i \in \mathcal{T}(Q) \wedge \mathcal{T}(D_j)} \Pi(\theta_i^l \mid D_j) \cdot \Pi(D_j) \cdot \prod_{T_k \in \mathcal{T}(Q) \setminus \mathcal{T}(D_j)} \Pi(\theta_k^l)) \quad (4.36)$$

où D_j prend ses instances dans $\{d_j, \bar{d}_j\}$;

θ est l'ensemble des configurations des parents de la requête;

θ^l une configuration possible de θ .

D'une manière générale, le processus d'évaluation des documents étant donnée une requête est déclenché pour tous les documents de la collection. L'instanciation « positive » du document D_2 , $D_2 = d_2$ entraîne le développement suivant :

$$\begin{aligned} \Pi(Q \mid T_2 T_3 T_6) &= \max(\\ &\quad \Pi(Q \mid t_2 t_3 t_6) \Pi(t_2 \mid d_2) \Pi(t_3 \mid d_2) \Pi(t_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid t_2 t_3 \bar{t}_6) \Pi(t_2 \mid d_2) \Pi(t_3 \mid d_2) \Pi(\bar{t}_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid t_2 \bar{t}_3 t_6) \Pi(t_2 \mid d_2) \Pi(\bar{t}_3 \mid d_2) \Pi(t_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid t_2 \bar{t}_3 \bar{t}_6) \Pi(t_2 \mid d_2) \Pi(\bar{t}_3 \mid d_2) \Pi(\bar{t}_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid \bar{t}_2 t_3 t_6) \Pi(\bar{t}_2 \mid d_2) \Pi(t_3 \mid d_2) \Pi(t_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid \bar{t}_2 t_3 \bar{t}_6) \Pi(\bar{t}_2 \mid d_2) \Pi(t_3 \mid d_2) \Pi(\bar{t}_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid \bar{t}_2 \bar{t}_3 t_6) \Pi(\bar{t}_2 \mid d_2) \Pi(\bar{t}_3 \mid d_2) \Pi(t_6 \mid d_2) \times \Pi(d_2), \\ &\quad \Pi(Q \mid \bar{t}_2 \bar{t}_3 \bar{t}_6) \Pi(\bar{t}_2 \mid d_2) \Pi(\bar{t}_3 \mid d_2) \Pi(\bar{t}_6 \mid d_2) \times \Pi(d_2) \\ &\quad) \end{aligned} \quad (4.37)$$

et le même développement est produit lorsque $D_2 = \overline{d_2}$ et dans ce cas, nous remplaçons d_2 par $\overline{d_2}$ dans 4.37.

La requête est composée de 3 termes de domaines binaires chacun ce qui conduit à 2^3 configurations possibles des termes de la requête (lorsque la requête est instanciée positivement). Les possibilités conditionnelles des parents des termes de la requête, $\Pi(T_i | D_j)$, sont données dans les tableaux 4.12, 4.13, 4.14. Ces valeurs sont les possibilités conditionnelles des termes étant donnée l'instanciation d'un parent document donné telles que définies dans le tableau 4.3. Nous ne donnons pas d'exemple de la pondération utilisant la théorie de

TAB. 4.12 – Possibilités conditionnelles $\Pi(T_2 | D_j)$

	d_2	$\overline{d_2}$	d_3	$\overline{d_3}$	d_4	$\overline{d_4}$	d_5	$\overline{d_5}$
t_2	1	0.861	1	0.861	0.067	0.991	1	0.861

Dempster-Shafer. Nous explicitons et discutons celle-ci dans le chapitre des expérimentations (Chapitre 8).

Le terme T_6 n'indexe que le document D_2 . La possibilité conditionnelle de ce

TAB. 4.13 – Possibilités conditionnelles $\Pi(T_3 | D_j)$

	d_2	$\overline{d_2}$	d_3	$\overline{d_3}$	d_4	$\overline{d_4}$	d_5	$\overline{d_5}$
t_3	0.5	0.931	1	0.861	1	0.861	1	0.861

terme est alors donnée dans le contexte de son unique noeud parent (tableau 4.14).

Nous considérons que l'utilisateur est plus intéressé par les documents longs

TAB. 4.14 – Possibilités conditionnelles $\Pi(T_6 | D_2)$

	d_2	$\overline{d_2}$
t_6	0.25	0.75

que par les documents courts. Nous définissons la possibilité *a priori* des documents telle que définie par 4.29. Le tableau 4.15 donne la possibilité marginale des noeuds documents qui sont en fait des noeuds racines.

La longueur du document est égale à la somme de ses fréquences. La longueur est normalisée par rapport à la longueur maximale parmi les documents de la collection. Le document D_2 est de longueur maximale, ce qui implique que sa

longueur normalisée vaut 1. Dans le tableau 4.16, sont données les valeurs des

TAB. 4.15 – Possibilités *a priori* des documents $\Pi(D_j)$

	d_1	d_2	d_3	d_4	d_5
$\Pi(d_j)$	0.178	1	0.053	0.304	0.804

possibilités marginales des noeuds termes racines. Ces valeurs sont obtenues par le calcul de l'entropie. Le facteur ndf_{3_i} est le pouvoir d'un terme t_i à discriminer entre les documents de la collection. Plus la valeur de ce facteur pour un terme t_i est basse, plus son absence de la représentation d'un document donné diminue la pertinence du document en question. Le terme t_6 lorsqu'il

TAB. 4.16 – Possibilités marginales des termes $\Pi(T_k)$

	t_2	t_3	t_6
ndf_{3_i}	0.812	1	0.105

est présent dans la requête pointerait vers le sous ensemble de documents de la collection le contenant. Il est très spécifique dans cette collection.

Agrégation booléenne Le tableau 4.17 donne les valeurs de la possibilité conditionnelle de la requête Q dans le contexte de ses parents. Les valeurs sont proposées pour une agrégation booléenne de type conjonctive, *ET*, et disjonctive, *OU* pour chaque configuration possible des parents. Lorsque la

TAB. 4.17 – Possibilités conditionnelles des parents de Q

$T_2T_3T_6$	<i>ET</i>	<i>OU</i>
$t_2t_3t_6$	1	1
$t_2t_3\bar{t}_6$	0	1
$t_2\bar{t}_3t_6$	0	1
$t_2\bar{t}_3\bar{t}_6$	0	1
$\bar{t}_2t_3t_6$	0	1
$\bar{t}_2t_3\bar{t}_6$	0	1
$\bar{t}_2\bar{t}_3t_6$	0	1
$\bar{t}_2\bar{t}_3\bar{t}_6$	0	0

requête est une conjonction de termes, il n'existe qu'une seule configuration possible qui la satisfait, à savoir t_2, t_3, t_6 . Dans l'exemple que nous présentons, le seul document de la collection contenant les trois termes à la fois est le document D_2 . Dans ce cas, le maximum sur les configurations des parents de Q est obtenu pour la configuration t_2, t_3, t_6 . Ainsi :

$$\Pi(Q \wedge d_2) = \Pi(Q | t_2 t_3 t_6) \times \Pi(t_2 | d_2) \times \Pi(t_3 | d_2) \times \Pi(t_6 | d_2) \times \Pi(d_2), \text{ et} \quad (4.38)$$

$$\Pi(Q \wedge \overline{d_2}) = \Pi(Q | t_2 t_3 t_6) \times \Pi(t_2 | \overline{d_2}) \times \Pi(t_3 | \overline{d_2}) \times \Pi(t_6 | \overline{d_2}) \times \Pi(\overline{d_2})$$

Ainsi la possibilité de pertinence du document d_2 étant donnée la requête Q est égale à 0.208 et sa nécessité de pertinence vaut 0. La possibilité de pertinence n'est pas maximale parce que les termes de la requête ne sont pas tous fortement présents dans ce document, ce qui explique aussi la nécessité nulle. La liste des documents restitués est composée d'un seul document D_2 en réponse à la requête conjonctive Q . Lorsque la requête est une disjonction de termes, s'il existe au moins un terme instancié tel que dans la requête, alors la requête est considérée comme satisfaite. Soit le calcul de la pertinence du document D_3 . Nous avons :

$$\begin{aligned} \Pi(Q \wedge d_3) = \max(& \\ & \Pi(Q | t_2 t_3 t_6) \Pi(t_2 | d_3) \Pi(t_3 | d_3) \Pi(t_6) \times \Pi(d_3), \\ & \Pi(Q | t_2 t_3 \overline{t_6}) \Pi(t_2 | d_3) \Pi(t_3 | d_3) \Pi(\overline{t_6}) \times \Pi(d_3), \\ & \Pi(Q | t_2 \overline{t_3} t_6) \Pi(t_2 | d_3) \Pi(\overline{t_3} | d_3) \Pi(t_6) \times \Pi(d_3), \\ & \Pi(Q | t_2 \overline{t_3} \overline{t_6}) \Pi(t_2 | d_3) \Pi(\overline{t_3} | d_3) \Pi(\overline{t_6}) \times \Pi(d_3), \\ & \Pi(Q | \overline{t_2} t_3 t_6) \Pi(\overline{t_2} | d_3) \Pi(t_3 | d_3) \Pi(t_6) \times \Pi(d_3), \\ & \Pi(Q | \overline{t_2} t_3 \overline{t_6}) \Pi(\overline{t_2} | d_3) \Pi(t_3 | d_3) \Pi(\overline{t_6}) \times \Pi(d_3), \\ & \Pi(Q | \overline{t_2} \overline{t_3} t_6) \Pi(\overline{t_2} | d_3) \Pi(\overline{t_3} | d_3) \Pi(t_6) \times \Pi(d_3), \\ & \Pi(Q | \overline{t_2} \overline{t_3} \overline{t_6}) \Pi(\overline{t_2} | d_3) \Pi(\overline{t_3} | d_3) \Pi(\overline{t_6}) \times \Pi(d_3) \\ &) \end{aligned} \quad (4.39)$$

Le même calcul est opéré pour l'instanciation $D_3 = \overline{d_3}$. Dans ce cas, nous remplaçons d_3 par $\overline{d_3}$ dans 4.39. La configuration $\overline{t_2 t_3 t_6}$ peut être supprimée des calculs parce qu'elle ne répond pas à la disjonction demandée par la requête. $\Pi(Q \wedge d_3) = 1$ et $\Pi(Q \wedge \overline{d_3}) = 1$, donc une possibilité de pertinence maximale avec une nécessité nulle. Nous obtenons pour les documents d_2, d_4, d_5 les possibilités données dans le tableau 4.18. Aucun document n'est nécessairement pertinent en réponse à la requête Q . Les documents sont restitués dans ce cas

TAB. 4.18 – Possibilité de pertinence des documents

	$\Pi(D_j Q)$	$N(D_j Q)$
D_1	0	0
D_2	1	0
D_3	0.06	0
D_4	0.54	0
D_5	0.9	0

par ordre décroissant de leur possibilité de pertinence, D_2 , D_5 , D_4 , D_3 . Nous remarquons deux points intéressants :

1. La nécessité de pertinence est fonction des termes utilisés dans la requête. Plus ces termes ont un pouvoir à discriminer entre les documents, et plus ils sont fortement présents dans un document donné et plus ce document est pertinent ;
2. Ce modèle est flexible dans la mesure où lorsque nous ne sommes pas certains de la décision de la pertinence d'un document, nous savons qu'au moins, ou plausiblement, il y aurait des chances de satisfaire la requête par ceux qui ont un degré de plausibilité de pertinence supérieure à zéro. De plus, cette flexibilité permet d'éviter, dans la mesure du possible, de restituer une liste vide de documents.

Agrégation par Noisy Or Pour ce type d'agrégation, nous considérons les termes de la requête comme ayant des importances différentes. D'une manière générale, l'utilisateur peut en formulant son besoin donner différentes importances (ou pondérations) aux termes qu'il utilise. Dans notre approche, il est aussi possible d'utiliser la spécificité du terme. Différentes mesures de spécificité du terme peuvent être utilisées, telles que idf , $nidf$, df_i . Le tableau 4.19 donne les possibilités conditionnelles des configurations possibles des noeuds parents de Q . Ces valeurs sont obtenues pour une agrégation de type *NOISY OR* comme définie par la formule 4.16.

La dernière ligne du tableau est ignorée puisque le maximum nous intéresse et qu'aucun terme de la requête n'est satisfait.

Les *meilleures* configurations des parents de la requête sont celles qui contiennent le terme T_6 instancié positivement, $T_6 = t_6$. Ce résultat n'est pas surprenant dans la mesure où ce terme est plus spécifique que les termes T_2 et T_3 . Ce-

TAB. 4.19 – Possibilités conditionnelles des parents de Q , $\Pi(Q | \theta)$

$T_2T_3T_6$	<i>NOISYOR</i>
$t_2t_3t_6$	1
$t_2t_3\bar{t}_6$	0.222
$t_2\bar{t}_3t_6$	0.905
$t_2\bar{t}_3\bar{t}_6$	0.112
$\bar{t}_2t_3t_6$	0.905
$\bar{t}_2t_3\bar{t}_6$	0.112
$\bar{t}_2\bar{t}_3t_6$	0.8089
$\bar{t}_2\bar{t}_3\bar{t}_6$	0

pendant, pour ce modèle il est plus intéressant d'avoir le terme T_6 et un autre terme de la requête que le terme T_6 seul.

Le classement des documents dans le tableau 4.20 est identique à celui obtenu par le classement des modèles vectoriel et probabiliste. Comme présenté dans le tableau 4.18 les nécessités de pertinence des documents valent 0. Les documents sont alors restitués par ordre décroissant de leur possibilité de pertinence. Nous montrerons dans le chapitre des expérimentations (*Chapitre 6*) des réponses constituées des pertinences nécessaires et plausibles et nous discuterons leurs effets sur les performances du SRI.

Nous ne pouvons pas encore affirmer à ce stade que nous obtenons toujours

TAB. 4.20 – Classement des documents

D_2
D_5
D_4
D_3

le même classement que ces modèles, puisqu'il ne s'agit ici que d'une base de test artificielle. Cet exemple a été déroulé sur tous les modèles de RI présentés principalement pour mieux expliciter la définition et le calcul de la pertinence d'un document en réponse à une requête.

4.6 Conclusion

Nous avons décrit dans ce chapitre un nouveau modèle pour la Recherche d'Information. Ce modèle traite l'incertitude d'une manière originale basée sur la théorie des possibilités et particulièrement les Réseaux possibilistes. Les noeuds dans ce réseau représentent les documents, les termes d'indexation ainsi que le besoin utilisateur. Les arcs reliant chaque couple de noeuds décrivent une relation de dépendance et sont quantifiés par deux mesures : la nécessité et la possibilité. Quel que soit le type de la relation décrite par un arc entre deux noeuds, sa quantification est opérée par deux mesures. Alors que la première est utile pour écarter certaines informations, la seconde mesure renforce les informations restantes. Nos contributions peuvent être orientées essentiellement en trois directions :

1. La modélisation de la pertinence ;
2. La pondération des termes d'indexation ;
3. La prise en compte des termes de la requête absents des documents lors du calcul des scores de pertinence.

Il est indéniable que les points cités ci-dessus sont étroitement liés. La pertinence est une notion vague et il est difficile de la définir d'une manière précise. La définition de « plusieurs pertinences » permet de gérer des *points particuliers et précis* de la pertinence. Nous ne prétendons pas maîtriser toute la sémantique liée à la pertinence dans nos travaux. Nous avons uniquement essayé de montrer que la prise en compte de différents types d'information peut apporter de la précision à la pertinence. Ainsi, la prise de décision quant à la restitution des documents pertinents en réponse à la requête est facilitée.

La pondération dépend généralement des poids des termes de la requête et des documents. Nous avons défini sous deux angles différents la représentativité d'un terme dans un document donné. Cette représentativité a été appréhendée selon deux approches. La première stipule qu'un terme est représentatif d'un document au moins au degré de sa présence dans ce document. Le degré de présence est quantifié par la fréquence normalisée. Nous avons utilisé la théorie de l'évidence pour cette méthode. La seconde manière de faire est de considérer que les termes absents des documents permettent d'écarter les documents des réponses possibles à la requête. Cette méthode a été quantifiée par la théorie des possibilités.

Finalement, nous avons tenté de proposer des poids aux termes dans le but de

calculer le degré de spécificité dans une collection donnée. Ces poids ont été utilisés dans notre approche pour mesurer l'absence des termes de la requête des documents lors du calcul des scores de pertinence. Ces poids peuvent aussi être utilisés pour mesurer l'importance des termes présents. Cette utilisation serait identique à celle du facteur de discrimination probablement le plus connu *idf*.

D'autre part, nous avons considéré que la restitution d'un document en réponse à une requête utilisateur peut être considérée dans un cadre d'inférence. En effet, la restitution d'un document est « causée » par la soumission d'une requête au système. Les données sur lesquelles se basent les modèles de la littérature pour restituer une liste de documents en réponse à un besoin utilisateur sont pauvres, incertaines et imprécises. La logique possibiliste se prête naturellement à ce genre d'application. Nous avons pu déterminer deux types de pertinence : la nécessaire et la plausible. La première permet de renforcer « nos croyances » vis à vis des résultats de la recherche et la seconde permet d'éviter de restituer une liste de documents vides à une requête utilisateur et d'en écarter ceux qui ne sont pas intéressants. La combinaison de la représentation par réseaux et de l'utilisation de la théorie des possibilités nous ont permis de répondre à un tel type de pertinence. La requête introduit de l'information qui change nos croyances sur les noeuds termes d'indexation ainsi que leurs noeuds parents. La liste de documents restitués contient les documents nécessairement pertinents en haut de la liste, puis les documents plausiblement pertinents.

Chapitre 5

Expérimentations

Introduction

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

L'objectif de ces expérimentations est de mesurer les performances et la viabilité de notre approche. Nous avons conçu un système de recherche d'information basé sur le modèle possibiliste que nous avons proposé.

Les expérimentations peuvent être subdivisées en deux classes. La première classe évalue l'impact des différents paramètres manipulés par le système sur les performances du système. La seconde classe s'articule autour de la comparaison de notre système avec le système actuel le plus performant en RI à savoir le système *OKAPI*. Nous évaluons ainsi l'apport de la double mesure de pertinence du cadre possibiliste par rapport à la pertinence probable du système *OKAPI*.

Dans la première section de ce chapitre nous décrivons brièvement les collections de tests utilisées pour évaluer notre système. Nous rappelons le protocole d'évaluation que nous avons suivi dans la seconde section. Nous fixons et présentons dans la troisième section les valeurs prises par les paramètres du modèle de base. Ce modèle est dit « de base » dans la mesure où il nous a permis d'obtenir les meilleures performances. De plus, ses performances (en termes de précision et de rappel) sont la base de comparaison pour le reste des expérimentations. L'impact des paramètres du système est évalué et présenté

dans la quatrième section. Dans la dernière section nous présentons une comparaison en terme de précisions et de rappel de notre système avec le système *OKAPI*.

5.1 Collection de tests

Nous avons utilisé dans nos expérimentations une collection de tests standard issue du programme *CLEF*. Cette collection fournit des protocoles unifiés qui permettent l'évaluation des SRI. Elle comporte :

- un ensemble de documents et un ensemble de requêtes ;
- une liste de documents pertinents pour chaque requête ;

La collection utilisée dans ces expérimentations, en l'occurrence *LeMonde 1994* est une sous collection de *CLEF*. Elle comporte des articles du journal français *Le Monde*. Cette collection est composée de 44013 documents et de 40 requêtes, le tout formant 154 MB de données. Parmi ces requêtes, 6 d'entre elles contiennent des termes qui ne figurent dans aucun document de la collection. Ces requêtes ne sont pas évaluées par notre système.

5.2 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 de ce manuscrit (*Chapitre 1*).

Le but de ces expérimentations est d'évaluer nos différentes contributions. Elles s'articulent essentiellement autour de 3 points :

1. les nouveaux facteurs de discrimination et l'impact de leurs normalisations ;
2. les méthodes utilisées pour la pondération des termes ;
3. la modélisation de la pertinence et l'apport d'une double mesure de pertinence.

Les paramètres dans notre système représentent les informations considérées pour calculer la nécessité et la possibilité de pertinence d'un document étant donnée une requête.

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 :

$$\%C = \frac{B - A}{A} \times 100 \quad (5.1)$$

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

Nous rappelons dans ce qui suit les paramètres manipulés par le système et comment nous proposons de les évaluer :

Facteurs de discrimination df_i Pour pouvoir mesurer le premier point il aurait été plus judicieux de faire une étude statistique plus approfondie de la distribution des termes dans la collection. Nous avons uniquement mesuré la viabilité de ces facteurs en termes de rappel et précision. Ces facteurs peuvent dans notre cas intervenir à 2 niveaux de notre modèle. Au niveau de la pondération des termes racines ou bien au niveau de la pondération des termes d'indexation présents dans le document. Plus précisément :

- dans le cas des termes racines, ils servent à diminuer le score de pertinence des documents ne contenant pas les termes de la requête. Plus le terme absent (du document) est doté d'un pouvoir discriminant (fort) plus la pertinence du document est pénalisée. Dans le modèle proposé ceci consiste à quantifier les termes racines du réseau ($\Pi(t_k) = df_i$; t_k appartient à la requête mais n'appartient pas au document). Nous l'appellerons dans la suite discrimination « négative » ;

- dans le cas des termes présents dans le document, ils servent à augmenter le score de pertinence des documents contenant de tels termes de la requête. Il s’agit ici de mesurer pour un terme t_i qui apparaît dans un document d_j , $\Pi(t_i | \overline{d_j}) = 1 - \phi_{ij}$. Nous désignons dans la suite de ce manuscrit cette utilisation des facteurs de discrimination par *discrimination positive*. Dans notre approche générale, $\phi_{ij} = ntf_{ij} \times nidf_i$, permet de mesurer cette discrimination positive. Nous la signalons dans nos expérimentations uniquement lorsque le facteur df_i remplace le facteur idf .

Pondération des termes d’indexation ($\Pi(T_i | D_j)$) Nous avons proposé deux approches pour pondérer les termes d’indexation des documents :

- la pondération « positive » qui affecte le maximum de croyance à un terme lorsqu’il est présent dans les documents ($\Pi(t_i | d_j) = 1$). Nous avons défini cette approche initialement dans le cadre de la théorie de Dempster-Shafer ;
- la pondération « négative » qui mesure la représentativité d’un terme dans un document d’une manière proportionnelle à son nombre d’apparitions dans ce document ($\Pi(t_i | d_j) = ntf_{ij}$). Cette approche a été définie initialement dans le cadre de la théorie des possibilités.

Longueur des documents ($\Pi(D_j)$) Il a été montré pour la majorité des modèles existant expérimentés sur les collections de tests que nous utilisons que les systèmes obtiennent de meilleures performances lorsque la longueur des documents est prise en compte dans le calcul des scores de pertinence [108]. Nous montrons dans notre système l’impact de la prise en compte de l’information apportée par la longueur affectée aux possibilités *a priori* des documents.

Agrégation des termes de la requête ($\Pi(Q | \theta^l)$) Les termes de la requête dans les collections de tests que nous avons utilisés ne sont pas agrégés explicitement par l’utilisateur. Cependant, les résultats obtenus par des agrégations

booléennes et quantifiée n'ont pas été concluants. Ainsi, pour toutes les variations des paramètres utilisés par notre approche, nous agrégeons les termes de la requête par l'opérateur *Noisy Or*. Cet opérateur a permis d'obtenir les meilleures performances.

Double mesure de pertinence ($N(T_i | D_j)$ et $\Pi(T_i | D_j)$) La distinction des documents nécessairement pertinents de ceux possiblement pertinents : nous montrons l'importance de la séparation de ces notions pour mieux quantifier la pertinence. Nous évaluons à quel point et dans quelles mesures la prise en compte de deux types de pertinence constitue une approche intéressante pour la RI ;

Nous comparons nos résultats de recherche à un des systèmes les plus performants actuellement à savoir le système *OKAPI* (présenté dans le *Chapitre 1* de ce manuscrit).

Les résultats ou performances du système sont évaluées principalement au moyen des précisions. Nous ajoutons les résultats aux points de rappel pour renforcer le comportement d'un paramètre donné.

5.3 Le modèle de base

Nous décrivons dans cette section les instanciations prises par les paramètres du modèle de base qui est optimal. Les paramètres ont été fixés pour ce modèle tels que décrits dans le tableau 5.1.

Dans le tableau 5.1 la « longueur des documents » et le « terme racine » sont

TAB. 5.1 – Possibilités conditionnelles et marginales

$\Pi(T_i D_j)$	d_j	\bar{d}_j	<i>Noisy Or</i>	t_i	\bar{t}_i
t_i	$nf_{t_{ij}}$	$1 - \phi_{ij}$	<i>Q</i>	$1 - q_i$	1
\bar{t}_i	1	1			
$\Pi(T_i)$	<i>Terme racine</i>	$\Pi(D_j)$	<i>Longueur des documents</i>		
t_i	ndf_{3_i}	d_j	nl_{d_j}		
\bar{t}_i	1	\bar{d}_j	1		

les possibilités marginales définies pour les documents ($\Pi(D_j)$) et les termes racines ($\Pi(T_k)$) respectivement. La représentativité d'un terme d'un document est mesurée par la possibilité conditionnelle ($\Pi(T_i | D_j)$). La pondération $\Pi(t_i | \overline{d_j})$ est telle que fixée dans le chapitre des contributions (mesure 4.25 dans le *Chapitre 4*). Nous le signalons dans ce qui suit uniquement lorsque nous modifions ce paramètre. Les termes d'agrégation de la requête sont agrégés par la possibilité conditionnelle ($\Pi(Q | T_i)$).

Les valeurs prises par les paramètres dans le modèle de base et présentées dans le tableau 5.1 peuvent être synthétisées comme suit :

1. utilisation du facteur ndf_3 pour une discrimination négative des termes de la requête absents des documents ;
2. prise en compte de la longueur des documents ;
3. la pondération négative des termes d'indexation : le poids affecté aux termes est tel que défini dans le cadre de la théorie des possibilités dans le chapitre des contributions (*Chapitre 4*).

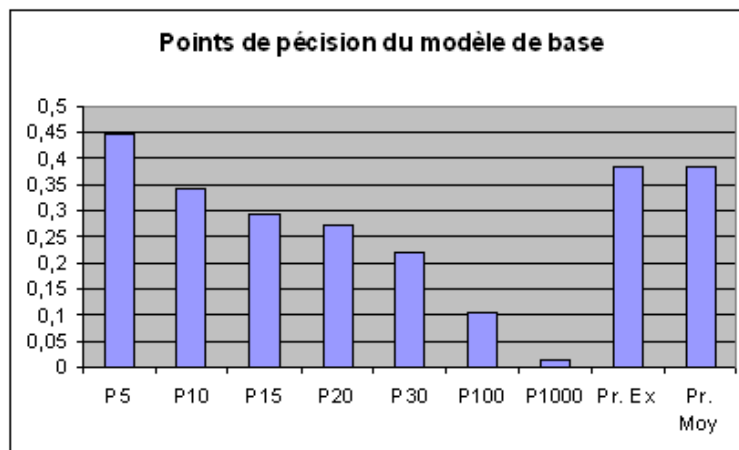


FIG. 5.1 – Points de précision

La figure 5.1 présente les valeurs des points de précision obtenues pour les 34 requêtes évaluées. La précision exacte et moyenne de ce modèle de base sont de 0.3661 et 0.3821 respectivement. Nous remarquons dans la figure 5.1 que l'écart entre les points de précision P_5 et P_{10} est assez élevé comparé aux écarts entre les autres points de précision pris deux à deux. Une explication possible est que notre approche, grâce à cette notion de nécessité de pertinence, permet de restituer les meilleurs documents en début de liste. Cette approche permet

de faire de la « haute précision ».

Nous illustrons dans ce qui suit les résultats obtenus lorsque la discrimination négative des termes est effectuée. Nous avons choisi de commencer par ces facteurs dans le but d'évaluer l'impact des autres paramètres par rapport à ces facteurs.

5.4 Expérimentations et résultats

Nous avons effectué plusieurs expérimentations afin d'évaluer les différents paramètres considérés dans notre modèle. Nous présentons dans ce qui suit les résultats les plus significatifs.

5.4.1 Impact des facteurs de discrimination « df »

Un des apports de notre travail concerne la proposition de trois nouveaux facteurs de discrimination df_i , $i = 1, \dots, 3$. Comme ces facteurs peuvent être normalisés de deux manières différentes : une normalisation orientée documents, notée DF_i , et une normalisation orientée collection, notée NDF_i , nous évaluons leurs impacts dans ce contexte. Nous évaluons ici, l'apport de ces facteurs en considérant les paramètres suivants :

1. DF et NDF pour une discrimination négative des termes racines $\Pi(T_k)$;
2. pondération négative pour $\Pi(T_i | D_j)$;
3. la longueur des documents $\Pi(D_j)$.

La figure 5.2 présente les résultats de ces expérimentations. Les meilleurs résultats (en termes de précision) ont été obtenus par les facteurs (NDF_i). En effet, ces facteurs sont normalisés par rapport à tous les termes de la collection. Ces facteurs s'intéressent au pouvoir d'un terme à discriminer entre les documents de la collection. Il n'est pas surprenant que le pouvoir d'un terme étant donné tous les termes de la collection soit plus significatif (par rapport à sa spécificité dans la collection) que celui d'un terme comparé uniquement à sa distribution dans les documents qu'il indexe. La normalisation orientée collection du facteur

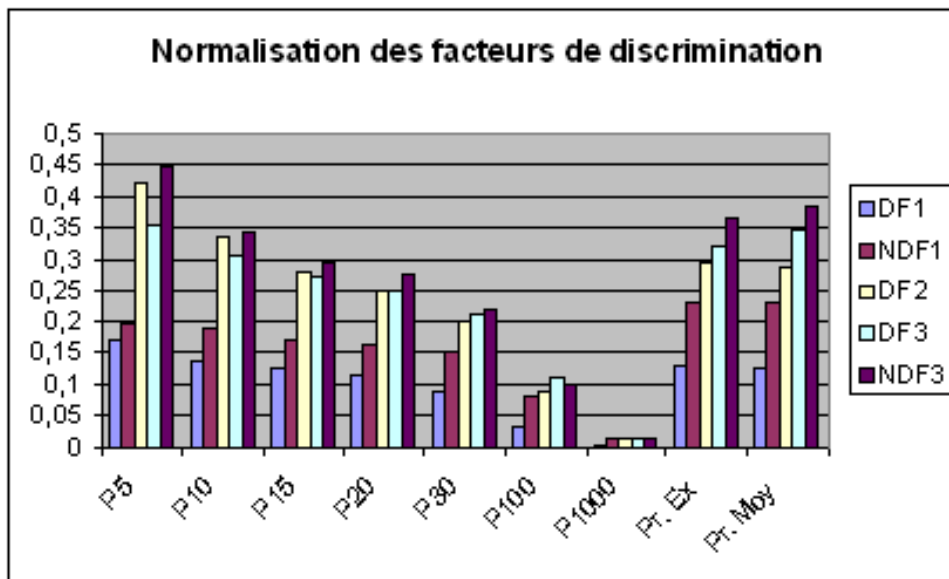


FIG. 5.2 – Normalisation des facteurs de discrimination

DF_2 n'a pas été intégrée dans la figure 5.2 du fait que sa normalisation n'a pas d'impact sur les précisions.

Le meilleur facteur de discrimination semblerait être le facteur NDF_3 . Les autres facteurs qu'ils soient normalisés par DF_i ou NDF_i n'atteignent pas les performances produites par ce facteur. La précision moyenne obtenue par le facteur NDF_3 est supérieure de 40% et 25% aux précisions moyennes obtenues par les facteurs NDF_1 et DF_2 respectivement.

Les précisions de chaque requête prise indépendamment apportent des informations supplémentaires. Ainsi, le facteur NDF_3 améliore les précisions de 19 requêtes parmi 34 évaluées comparés aux facteurs NDF_1 et DF_2 .

Les facteurs NDF_3 et NDF_1 trouvent des valeurs de précision identiques pour la précision à 5 de 14 requêtes. Ce résultat s'explique par le fait que ces requêtes contiennent des termes *relativement peu spécifiques* de la collection. Lorsque les requêtes contiennent des termes spécifiques les précisions du facteur NDF_3 sont meilleures.

Le facteur DF_2 améliore légèrement les précisions comparativement au facteur NDF_1 .

L'écart d'un point de précision à l'autre pour les facteurs DF_2 et DF_3 (normalisé collection ou documents) semble être assez élevé. Cependant, pour le

facteur DF_1 l'écart est moins important. Lorsque les facteurs DF_2 , DF_3 et NDF_3 sont utilisés le système semble restituer, en haut de liste des documents restitués, les documents les plus pertinents.

5.4.2 Impact des techniques de pondération $\Pi(T_i | D_j)$

Comme nous l'avons souligné précédemment, le paramètre $\Pi(T_i | D_j)$ peut être mesuré selon deux approches : l'approche *négative* et l'approche *positive*. Nous avons retenu pour ces expérimentations les paramètres suivants :

1. facteurs de discrimination : nous avons gardé les 3 facteurs NDF_i pour la discrimination négative afin de mieux évaluer cette pondération dans ces trois contextes ;
2. pondération positive vs négative des termes ;
3. la longueur des documents.

Les figures 5.3, 5.4 et 5.5 illustrent les précisions obtenues lorsque les pondérations « positive » et « négative » respectivement sont opérées.

Nous constatons dans ces figures, que la pondération « positive » ($\Pi(t_i | d_j) =$

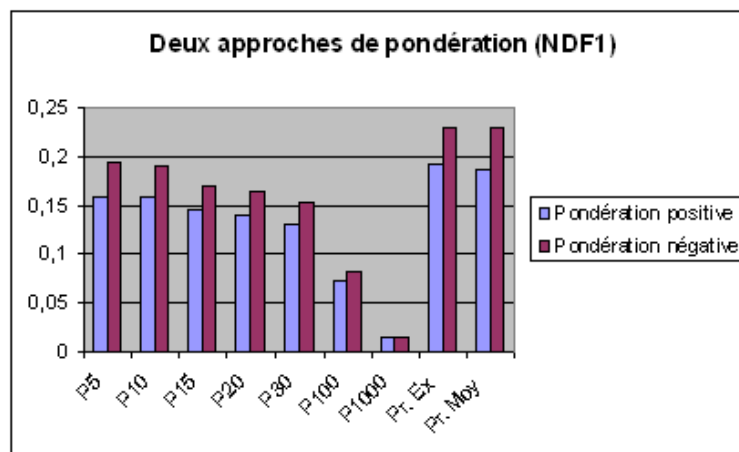


FIG. 5.3 – Pondération positive vs négative

1) décroît les performances du système. Cette constatation est vraie quel que soit le facteur de discrimination utilisée (5.3, 5.4 et 5.5). Ce résultat n'est pas surprenant dans la mesure où la pondération positive est trop large. Un document qui contient les termes de la requête avec un faible nombre d'apparitions

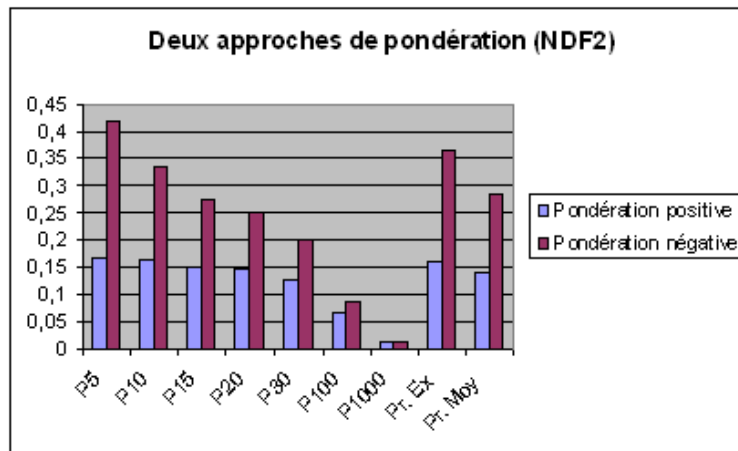


FIG. 5.4 – Pondération positive vs négative

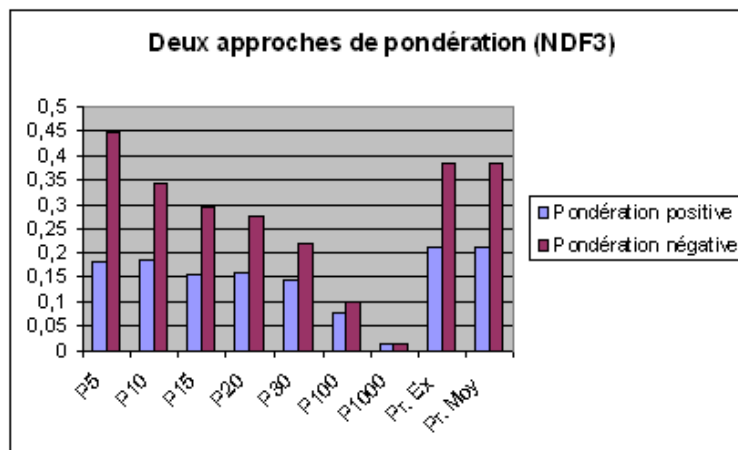


FIG. 5.5 – Pondération positive vs négative

peut se retrouver en haut de liste des documents restitués, puisque cette information (nombre d'apparitions) n'est pas considérée. La pondération positive se base uniquement sur la présence absence des termes. Ce document peut se retrouver au même rang dans la liste des documents restitués qu'un document qui contient « fortement » les termes de la requête.

Le tableau 5.2 mesure le pourcentage d'écart sur les précisions entre la pondération se référant à l'information négative et celle se référant à l'information positive¹. Nous affectons aux variables définies par la mesure 5.1 : $B = NDF_i^{neg}$, $A = NDF_i^{pos}$ et $\%C = \%NDF_i$ Avec :

NDF_i^{neg} et NDF_i^{pos} désignent respectivement le point de précision au point

¹Nous précisons au-delà de 2 chiffres après la virgule pour éviter les confusions

TAB. 5.2 – Pourcentage de perte liée à l'utilisation de la pondération « positive »

	P_5	P_{10}	P_{15}	P_{20}	P_{30}	P_{100}	P_{1000}	$Pr.Ex$	$Pr.Moy$
NDF_1^{neg}	0,19	0,18	0,16	0,16	0,15	0,08	0,0146	0,22	0,22
NDF_1^{pos}	0,15	0,15	0,14	0,13	0,13	0,07	0,0142	0,19	0,18
$\%NDF_1$	18,18	15,62	13,93	14,39	14,21	12,22	2,73	16,31	19,31
NDF_2^{neg}	0,42	0,33	0,27	0,24	0,20	0,08	0,01	0,36	0,28
NDF_2^{pos}	0,16	0,16	0,14	0,14	0,12	0,06	0,01	0,15	0,14
$\%NDF_2$	60,30	51,00	45,97	40,91	36,99	27,75	0	56,40	50,21
NDF_3^{neg}	0,44	0,34	0,29	0,27	0,22	0,10	0,0149	0,38	0,38
NDF_3^{pos}	0,18	0,18	0,15	0,16	0,14	0,07	0,0146	0,212	0,21
$\%NDF_3$	59,20	45,69	45,65	41,08	34,22	24,33	2,01	44,46	45,04

P_i obtenu par la pondération négative et positive. Ce pourcentage est calculé pour chaque facteur NDF .

Nous constatons que la précision moyenne diminue d'au moins 20% lorsque les poids sont obtenus par l'approche positive (tableau 5.2). Les facteurs les plus touchés par cette perte de performance sont essentiellement les facteurs NDF_2 et NDF_3 . Les variations sur les précisions du facteur NDF_1 (figure 5.3) sont faibles. L'unique information utilisée pour calculer ce facteur est sa densité normalisée dans la collection par rapport au maximum des densités des termes de la collection. Il semblerait que dans ce cas que l'impact de la fréquence normalisée devient minime.

On constate par ailleurs que la différence de performances a tendance à s'estomper pour les points de précision élevés. Nous pouvons penser que l'information portant sur le nombre d'apparitions des termes pour les documents classés en bas de liste n'a pas d'importance pour ces documents. En effet, a priori les termes de la requête ont des chances d'avoir de faibles nombres d'apparitions dans ces documents. L'information de présence-absence d'un terme donné de la requête pour ces documents suffirait peut être à caractériser la pertinence de ces documents.

L'écart entre les précisions à 5 et à 10 est peu élevé lorsque la pondération positive est adoptée (le nombre de documents pertinents restitués augmente d'une manière proportionnelle au nombre de documents restitués). Au vu des

valeurs des points de précision, il semblerait que cet écart augmente lorsque le nombre de documents restitués augmente (pour les précisions $(P_{15}, \dots, P_{1000})$).

5.4.3 Impact de la longueur des documents

Les modèles actuels de la RI semblent privilégier les documents longs aux courts. Il a déjà été montré, que les évaluations dans les collections de tests favorisent les documents longs de la collection [108]. Nous montrons ici, l'impact de la prise en compte de la longueur des documents sur les performances du système.

Nous retenons pour ces expérimentations les paramètres suivants :

- les facteurs NDF_i pour la discrimination négative ;
- la pondération négative des termes d'indexation ;

Nous présentons dans la figure 5.6 les précisions obtenues lorsque la longueur est supprimée du calcul des scores de pertinence. Nous constatons que les meilleures performances du système pour les précisions P_5, \dots, P_{20} ne sont pas atteintes lorsque le facteur NDF_3 est utilisé (ce qui était le cas jusqu'ici) mais plutôt par le facteur NDF_2 . Cependant, la précision moyenne du facteur NDF_3 est la plus élevée comparée à celle obtenue par les facteurs NDF_1 et NDF_2 . Dans

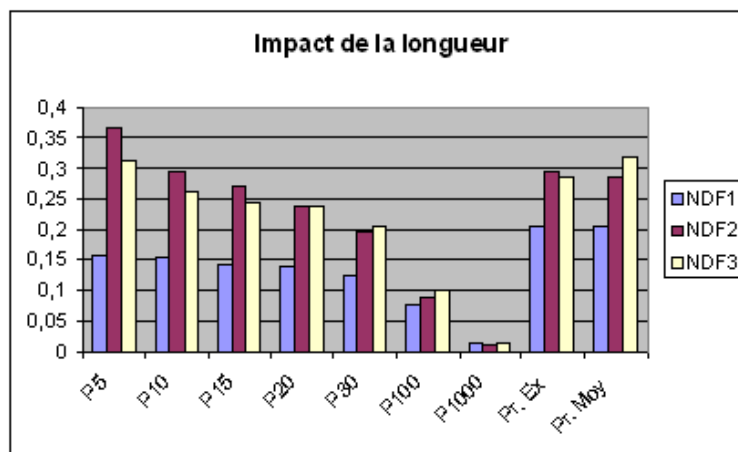


FIG. 5.6 – Elimination de la longueur du calcul des scores de pertinence

le tableau 5.3 $SL NDF_i$ désigne le point de précision lorsque la longueur n'est pas intégrée dans le calcul des scores de pertinence, et, $AL NDF_i$ désigne le point de précision lorsque la longueur est considérée.

L'élimination de la longueur des documents du calcul des scores de pertinence a un impact négatif sur les performances du système. En effet, dans le tableau 5.3 nous pouvons constater que les valeurs des précisions obtenues lorsque la longueur est éliminée sont plus faibles que lorsque la longueur est considérée. L'impact le plus important de la longueur se produit lorsque les facteurs de discrimination utilisés sont NDF_1 et NDF_3 respectivement. Le calcul du facteur NDF_2 tient compte essentiellement de la longueur. Pour illustrer de manière claire les performances selon les NDF_i , nous présentons dans la figure 5.7 les résultats obtenus par chacun d'eux sous forme de courbes rappel-précision.

Nous gardons dans la figure 5.7 la même légende que celle utilisée dans le

TAB. 5.3 – Impact de la longueur sur les points de précisions

	P_5	P_{10}	P_{15}	P_{20}	P_{30}	P_{100}	P_{1000}	$Pr.Ex$	$Pr.Moy$
$AL\ NDF_1$	0,19	0,18	0,16	0,16	0,15	0,08	0,01	0,22	0,22
$SL\ NDF_1$	0,15	0,15	0,14	0,13	0,12	0,07	0,01	0,20	0,20
$AL\ NDF_2$	0,42	0,33	0,27	0,24	0,20	0,08	0,01	0,36	0,28
$SL\ NDF_2$	0,36	0,29	0,26	0,24	0,19	0,08	0,01	0,29	0,28
$AL\ NDF_3$	0,44	0,34	0,29	0,27	0,22	0,10	0,01	0,38	0,38
$SL\ NDF_3$	0,31	0,26	0,24	0,23	0,20	0,09	0,014	0,28	0,31

tableau 5.3.

Les meilleurs résultats en termes de rappel sont atteints par le système lorsque la longueur est utilisée dans le calcul des scores de pertinence en combinaison avec le facteur NDF_3 . La courbe qui correspond à cette combinaison, $AL\ NDF_3$, est supérieure à tous les points de rappel des autres courbes.

Pour certains points de rappel, l'utilisation de la combinaison $AL\ NDF_2$ et $SL\ NDF_3$ est presque identique.

La courbe obtenue par le facteur NDF_2 est discutable selon deux points de vue :

1. la discrimination NDF_2 privilégie les documents longs et l'ajout de l'information de la longueur permet de « booster » davantage les performances du système. Ceci expliquerait le peu d'écart entre les courbes

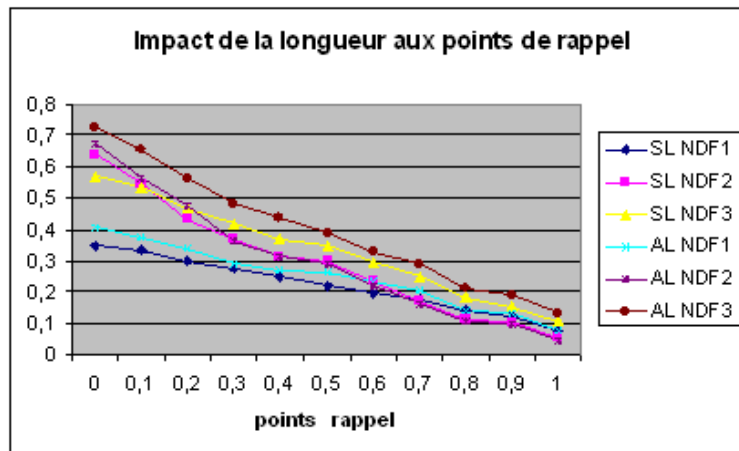


FIG. 5.7 – Courbes de rappel des NDF avec et sans longueur

$AL\ NDF_2$ et $SL\ NDF_3$;

- le facteur NDF_2 privilégie les documents courts et l'ajout de l'information de la longueur permet d'équilibrer les performances du système. Ceci expliquerait les résultats quasi équivalents obtenus par les courbes $AL\ NDF_2$ et $SL\ NDF_2$.

5.4.4 Comparaison idf et NDF

L'objectif de cette expérimentation est de comparer les facteurs de discrimination proposés avec le facteur communément utilisé en RI, à savoir le facteur idf (cf *Chapitre 1*). Comme nous l'avons souligné dans le protocole d'évaluation (section 5.2), ces facteurs peuvent être utilisés pour discriminer positivement et/ou négativement. Les paramètres que nous utilisons pour ces expérimentations sont :

- facteurs de discrimination positive et négative, quantifiés avec NDF et/ou idf ;
- pondération négative ;
- longueur des documents.

Nous souhaitons tout d'abord évaluer l'impact des NDF_i utilisés pour discriminer positivement et négativement. Nous illustrons dans la figure 5.8 l'impact des facteurs que nous proposons sur les précisions.

Lorsque les NDF_i sont utilisés aussi bien pour discriminer positivement que

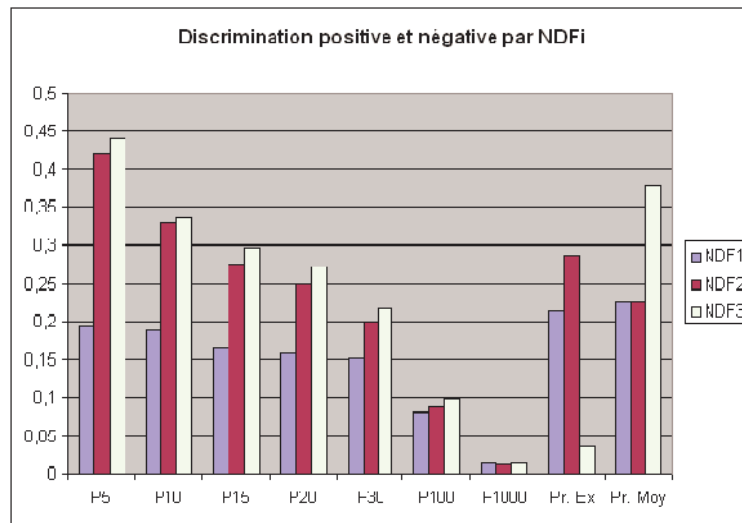


FIG. 5.8 – Impact des facteurs de discrimination sur la présence et l’absence des termes

négativement, le facteur NDF_3 reste le meilleur facteur. Le facteur NDF_1 est encore une fois le facteur le moins performant.

Cependant, lorsque les facteurs NDF_i sont utilisés pour discriminer négativement et $nidf$ pour « discriminer positivement », les performances de notre système sont plus intéressantes. Le tableau 5.4 montre que les précisions moyennes et exactes obtenues dans les deux cas :

- $NDF_i - nidf$ le facteur NDF est utilisé pour discriminer négativement et la facteur $nidf$ pour discriminer positivement et ;
- $NDF_i - NDF_i$ lorsque le facteur NDF_i est utilisé pour discriminer à la fois positivement et négativement.

TAB. 5.4 – Discrimination positive vs discrimination négative

	<i>Pr.Ex</i>	<i>Pr.Moy</i>
$NDF_1 - nidf$	0,2298	0,2299
$NDF_1 - NDF_1$	0,2151	0,2259
$NDF_2 - nidf$	0,3661	0,2852
$NDF_2 - NDF_2$	0,2872	0,28
$NDF_3 - nidf$	0,3661	0,3821
$NDF_3 - NDF_3$	0,3590	0,3772

Nous focalisons la comparaison uniquement sur les précisions moyennes et exactes parce qu'elles sont les plus significatives.

Les valeurs de précision décroissent lorsque les facteurs NDF_i sont utilisés pour mesurer la présence des termes. Nous pourrions penser que ces facteurs sont moins performants que le facteur $nidf$. Cependant, les facteurs que nous proposons sont plus « fins » dans la discrimination. Deux termes qui apparaissent dans le même nombre de documents peuvent avoir la même valeur de discrimination par le facteur $nidf$ mais des valeurs différentes par les facteurs NDF_i . Ainsi, une fine discrimination est plus adéquate pour mesurer l'information *négative* qui traduit dans ce contexte l'absence d'un terme de la requête d'un document donné.

5.4.5 Elimination des facteurs NDF_i du modèle de base

La non prise en compte des facteurs de discrimination dans notre modèle décroît considérablement les performances du système. Nous retenons pour ces expérimentations les paramètres suivants :

- la pondération négative des termes d'indexation ;
- prise en compte de la longueur des documents.

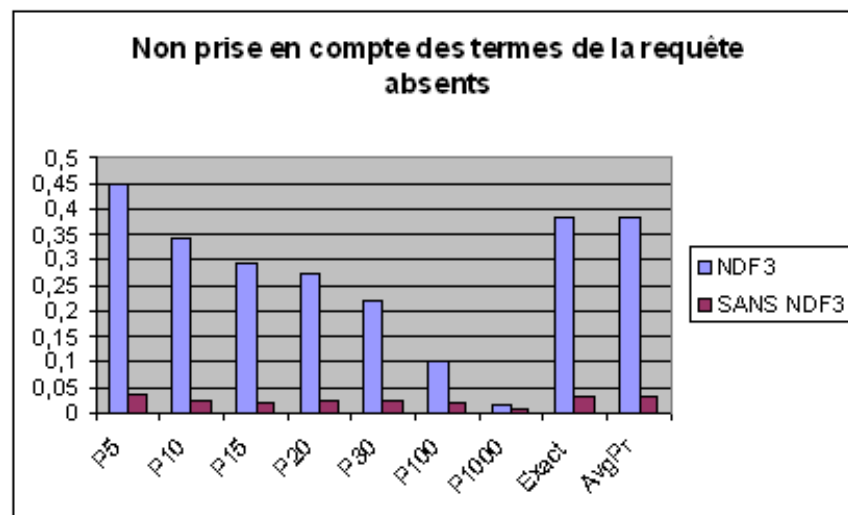


FIG. 5.9 – Non prise en compte des termes de la requête absents

Dans la figure 5.9 nous comparons le modèle initial désigné par NDF_3 aux résultats obtenus par le non prise en compte de la discrimination négative. Nous ne jugeons pas utile de représenter les résultats obtenus par les facteurs NDF_2 et NDF_1 . Nous cherchons à montrer ici, que notre modèle devient totalement non intéressant lorsque la discrimination négative n'est pas prise en compte.

5.5 Comparaison avec *OKAPI*

Un des apports de notre approche consiste à modéliser d'une nouvelle manière la pertinence. Nous avons défini la pertinence possible d'un document vis à vis d'une requête et sa pertinence nécessaire. La pertinence possible vise à éliminer les documents non pertinents, la pertinence nécessaire à renforcer la pertinence des documents non éliminés par la possibilité.

Cette double mesure de pertinence est censée aider le système dans sa décision concernant les documents à restituer ainsi que de leur ordre de restitution. Pour ce faire, nous comparons les performances de notre système à un des systèmes les plus performants actuellement à savoir le système *OKAPI*. Ce système repose sur un modèle probabiliste et est présenté dans le *Chapitre 1* de ce manuscrit. La pondération des termes utilisés est donc celle définie dans le modèle de Poisson.

Une première constatation au vu des points de précision est que notre système obtient de meilleures performances. Nous présentons un comparatif des points de précision dans la figure 5.10. Nous remarquons une nette amélioration des performances par rapport aux documents restitués en haut de liste. En effet, au vu de ces résultats, il est clair que les valeurs des points de précisions P_5, \dots, P_{20} obtenues par notre système sont plus élevées. Nous obtenons une amélioration de plus de 14% pour la précision à 5 (P_5). D'une manière générale, comme présenté dans le tableau 5.5, les précisions P_5, \dots, P_{20} obtenus par l'utilisation de notre approche sont supérieurs de plus de 5% au modèle *OKAPI*. $P_i^{Possibiliste}$ et $P_i^{Probabiliste}$ désignent la précision au point P_i obtenue respectivement par notre approche et celle d'*OKAPI*.

La précision moyenne obtenue par notre système est supérieure de plus de 8% à celle obtenue par *OKAPI*. Nous remarquons aussi que l'augmentation des nombres de documents restitués décroît les précisions de l'approche possibiliste. La figure 5.11 compare les performances en terme de rappel des deux

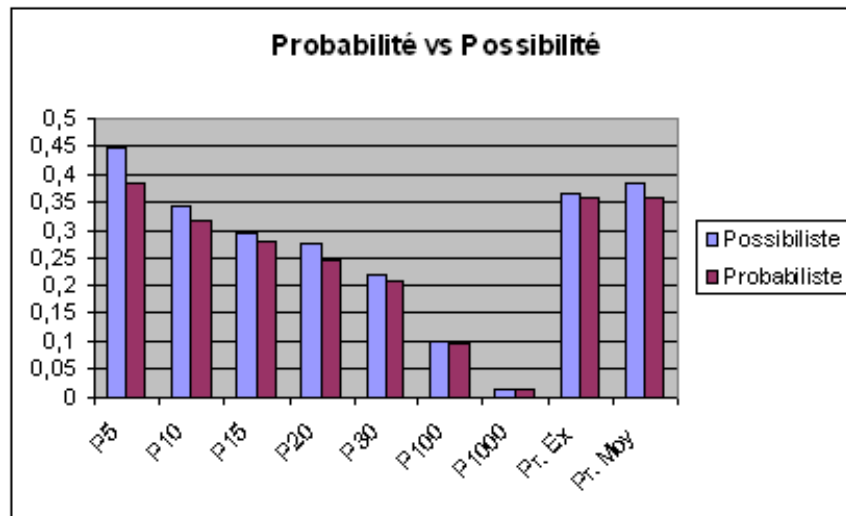


FIG. 5.10 – Comparatif des deux systèmes : Possibiliste et OKAPI

TAB. 5.5 – Pourcentage d'amélioration de notre approche comparée à l'approche probabiliste

	P_5	P_{10}	P_{15}	P_{20}	P_{30}	P_{100}	P_{1000}	$Pr.Moy$
$P_i^{Probabiliste}$	0,38	0,31	0,27	0,24	0,20	0,09	0,01	0,35
$P_i^{Possibiliste}$	0,44	0,34	0,29	0,27	0,22	0,10	0,01	0,38
$\%Am$	16,91	7,43	4,95	11,47	6,15	4,53	2,05	8,02

systèmes. Nous remarquons également que d'une manière générale la courbe de notre système est souvent au-dessus de celle d'OKAPI avec un net écart pour les premières valeurs de rappels qui correspondent aux premiers documents sélectionnés.

Nous nous intéressons dans ce qui suit aux pertinences des documents restitués par le système que nous proposons. Nous tentons de trouver ainsi, une corrélation entre la nouvelle approche de modélisation de la pertinence avec l'amélioration des performances pour les documents restitués en haut de la liste. Pour ce faire, nous commençons par étudier les requêtes qui améliorent ces résultats.

Parmi les 34 requêtes évaluées pour les 2 systèmes, le système possibiliste améliore les précisions à 5 (P_5) de 14 d'entre elles, et obtient les mêmes valeurs pour 13 d'entre elles. Le système OKAPI obtient de meilleures valeurs P_5 pour 7 d'entre elles.

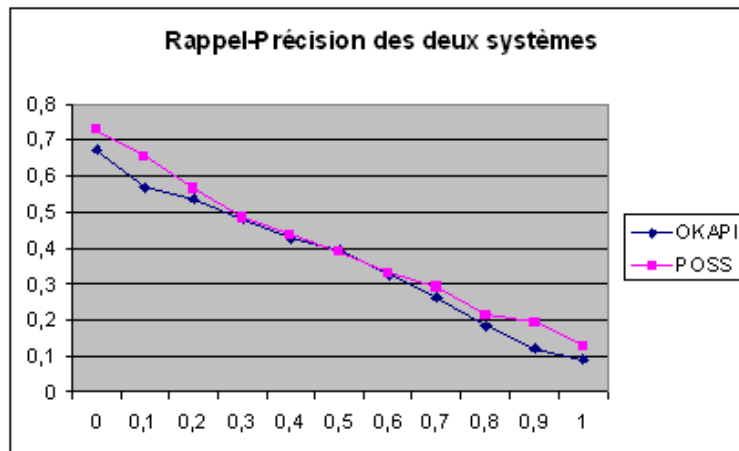


FIG. 5.11 – Comparatif des deux systèmes en terme de rappel précision : Possibiliste et OKAPI

Intuitivement, notre approche de classement des documents restitués en réponse à une requête utilisateur semble au vu de ces résultats intéressante. Le « découpage » entre les documents certainement (ou nécessairement) pertinents et possiblement pertinents permet de classer les meilleurs documents en haut de la liste. Cependant, les valeurs de précision plus élevées ne sont pas uniquement le résultat de la pertinence nécessaire. A contrario, le système a restitué des documents nécessairement pertinents pour certaines requêtes sans que leurs valeurs de précision ne soit supérieures à celles d'*OKAPI*.

5.6 Conclusion

Nous avons décrit dans ce chapitre les expérimentations effectuées pour mesurer la viabilité de notre système. D'une manière générale ces expérimentations peuvent être répertoriées en deux classes :

1. la première classe s'articule autour de l'impact des paramètres sur les performances du système. Nous évaluons notamment les performances lorsque les facteurs de discrimination sont utilisés pour pénaliser les documents ne les contenant pas. Nous mesurons aussi l'apport des ces facteurs lorsqu'ils ont la même utilisation que le facteur *idf*. Ces facteurs contribuent à de meilleures performances de notre système lorsqu'ils sont

utilisés pour pénaliser les documents ne contenant pas des termes discriminants. Cependant, ces facteurs ne sont pas propres au modèle proposé. Ils peuvent à notre sens être adapté dans les autres modèles de RI, comme une alternative à *idf*. De plus nous avons adapté l'agrégation des termes de la requête par l'opérateur du *Noisy Or* qui a aussi permis d'obtenir de meilleurs résultats que des agrégations booléennes simples de ces termes. Les résultats de recherche en utilisant les autres opérateurs se sont avérés non intéressants et nous ne les avons pas présentés dans ce manuscrit. De plus, nous avons évalué l'impact de deux approches de pondération. L'approche positive considère un terme (totalement) possiblement représentatif d'un document lorsque le terme apparaît dans ce document. Cette approche est large et ne peut pas discriminer entre deux termes présents avec des nombres d'apparitions différents dans un même document. Pour l'approche négative un terme représente un document d'une manière proportionnelle à son nombre. Finalement, nous renforçons l'idée de l'importance de la longueur des documents dans ce type de collections de tests ;

2. la seconde classe concerne la double mesure de pertinence. Nous comparons dans ce type d'expérimentations les performances de notre système à celles du système *OKAPI*. La précision moyenne que nous obtenons est supérieure à celle obtenue par *OKAPI* de plus de 8%.

Quatrième partie

Conclusion générale

Conclusion Générale

Synthèse

Les travaux présentés dans ce manuscrit s'inscrivent dans le cadre de la Recherche d'Information (RI). Nous nous sommes principalement intéressés à la définition d'un modèle de RI permettant une « meilleure » modélisation de la notion de pertinence, élément fondamental en RI. Notre modèle trouve ses fondements théoriques dans les réseaux possibilistes. Plus précisément, le modèle que nous proposons est basé sur un réseau pour lequel les noeuds représentent les documents, leurs termes d'indexation et la requête. La topologie du réseau permet de prendre en compte naturellement les relations de dépendance entre ces noeuds.

L'utilisation des réseaux, qu'ils soient probabilistes ou neuronaux en RI s'est avérée intéressante grâce notamment à leur puissance pour inférer la pertinence des documents vis à vis d'une requête ainsi qu'à leur capacité de représenter de manière naturelle les différents liens existants entre les objets manipulés en RI, à savoir les termes, les documents et la requêtes. Cependant, le cadre probabiliste dans lequel ces réseaux ont été définis traduit mal les nuances imprécises sous jacentes aussi bien à la notion de pertinence qu'à la représentativité des termes dans les documents. En effet, cette théorie permet uniquement de mesurer la certitude d'un événement et de son contraire. Dans ces modèles la pertinence et la représentativité d'un terme dans un document sont binaires. Un document donné est pertinent ou non vis à vis d'une requête à un certain degré. Un terme est représentatif d'un document ou non à un certain degré. Nous avons proposé pour notre part un modèle de RI basé sur les réseaux possibilistes. L'utilisation du cadre possibiliste permet à notre sens de mieux traduire les différentes nuances liées à la pertinence ainsi qu'à la représentativité

d'un terme dans un document. En effet, la notion de pertinence d'un document étant donnée une requête est modélisée par une double mesure. La **pertinence nécessaire** permet de focaliser sur les documents à restituer ainsi que de renforcer la nécessité de les faire figurer parmi les premiers de la liste des résultats restitués en réponse à une requête. La **pertinence possible** permet de rejeter les documents non pertinents à une requête donnée. Les arcs reliant des paires de noeuds et sont quantifiées par des degrés de possibilité et de nécessité. Ces degrés mesurent d'une manière générale le degré de possibilité et de nécessité de l'information véhiculée par les arcs. Cette information concerne la représentativité d'un terme dans un document et permet de quantifier la pertinence d'un document étant donnée une requête. Il s'agit à notre connaissance de la première application de ces réseaux dans le domaine de la RI.

L'évaluation de la pertinence d'un document vis à vis d'une requête est effectuée par un processus de propagation à travers les noeuds termes reliés à cette requête. Les termes de la requête absents dans les représentations des documents sont donc naturellement et explicitement considérés dans le calcul des scores de pertinence contrairement aux systèmes actuels de RI.

Compte tenu de l'intérêt que nous avons accordé à cette notion de représentativité d'un terme dans un document, nous avons proposé trois nouveaux facteurs permettant de mieux quantifier l'importance d'un terme selon sa distribution dans la collection. Ces facteurs affectent un pouvoir discriminant aux termes leurs permettant de pointer vers un sous ensemble particulier de documents. De tels termes dotés d'un pouvoir discriminant lorsqu'ils figurent dans la requête contribuent à la restitution des documents pertinents. Nous avons évalué ces facteurs en adaptant leurs utilisations à notre modèle. Cependant ces facteurs peuvent être utilisés par tous les types de modèles.

Nous avons conçu un système basé sur ce modèle. Afin d'évaluer la viabilité de notre approche, nous avons expérimenté ce système sur une des collections standards de RI, à savoir *CLEF*. Pour ce faire, nous avons opéré des variations sur les principaux paramètres intervenant lors du calcul des scores de pertinence.

Nous avons proposé deux approches pour la pondération des termes d'indexation. L'approche dite négative quantifie la représentativité d'un terme dans un document donné en fonction de son nombre d'apparitions dans ce document. L'approche positive est plus large que l'approche négative. Le poids d'un

terme quantifiant la représentativité d'un terme présent dans le document veut 1. L'évaluation de ces deux approches de pondération a montré l'efficacité de la pondération négative. En effet, cette approche (négative) quantifie d'une manière « nuancée » la représentativité d'un terme donné. Il est ainsi plus aisé de différencier la représentativité de deux termes dans un même document. La seconde approche est large puisqu'elle repose sur une vision booléenne pour « quantifier » la représentativité d'un terme (présence absence du terme dans le document).

De plus, nous avons évalué l'impact des facteurs de discrimination que nous avons proposés. Ces facteurs peuvent être utilisés selon deux approches de discrimination. L'affectation de ces facteurs aux termes de la requête, dans l'approche positive, consiste à augmenter les scores de pertinence des documents contenant ces termes (utilisation identique au facteur inverse d'un terme, *idf*). Dans la discrimination négative, ces poids sont affectés aux termes de la requête dans le but de pénaliser les scores de pertinence des documents ne les contenant pas. La pénalisation et l'augmentation des scores sont proportionnelles au pouvoir des termes à discriminer entre les documents de la collection. Les facteurs que nous avons proposés sont plus « fins » que le facteur *idf*, puisque la distribution des termes dans la collection de documents ne dépend pas seulement de la présence absence des termes dans les documents de la collection (comme *idf*) mais de la distribution de leur densité dans les documents de la collection. Ces mesures se sont avérées efficaces pour la discrimination négative, comparé notamment à *idf*.

Enfin, nous avons comparé les résultats obtenus par notre système avec ceux obtenus par le système *OKAPI*. *OKAPI* est un des modèles, voire le modèle de référence en RI. Les résultats de cette comparaison sont très encourageants. En effet, nous avons obtenu avec notre système une précision moyenne supérieure de plus de 8% par rapport à celle obtenue par *OKAPI*. Les travaux sur la viabilité d'un modèle de RI, s'accordent à affirmer qu'à partir d'une amélioration de 5% des précisions, les résultats peuvent être considérés intéressants.

Perspectives

De nombreuses perspectives découlent de nos travaux.
A court terme nous prévoyons :

- d’intégrer un processus itératif à la recherche pour la reformulation de requêtes. Pour ce faire, deux techniques existant dans les modèles basés sur les réseaux Bayésiens probabilistes pourraient être adaptées à notre approche. La première préconise l’ajout des noeuds ou d’arcs dans le réseau pour recalculer les distributions de possibilité. Cette technique permet ainsi d’ajouter des relations de dépendance entre des termes et la requête. Ces termes peuvent être issus des documents jugés par l’utilisateur ou les termes des n premiers documents restitués initialement par le système. La seconde technique considère la requête reformulée comme une nouvelle information à introduire dans le système ;
- de définir les relations de dépendance dans un cadre qualitatif. Les valeurs affectées à ces relations traduiraient des ordres partiels de préférence. La théorie des possibilités offre deux cadres de travail. Le cadre qualitatif ou ordinal et le cadre numérique. Nous avons proposé notre modèle dans un cadre numérique. Nous proposons ici de traduire ce modèle dans un cadre ordinal. Ainsi, des préférences pourraient être définies entre les termes d’indexation pour représenter les documents et/ou la requête. Ces préférences peuvent être données par des experts, ou par des études statistiques sur le texte, etc. Ces préférences permettraient par la suite, de restituer des documents classés par préférence de pertinence. Il serait possible dans un tel cadre de mesurer le point auquel un document d_1 est préféré au document d_2 ou de mesurer la préférence du document d_1 par rapport à un ensemble de documents $\{d_3, d_4\}$.

A long terme nous prévoyons :

- d’intégrer des relations de dépendance entre des paires de termes d’indexation ou des paires de documents. Cette perspective peut être en relation avec la perspective précédente. Dans ce contexte, les arcs sont mesurés par des valeurs numériques traduisant des quantités et non pas des ordres partiels. Afin de quantifier ces relations, nous pourrions nous baser sur la connaissance représentée dans une ontologie. Une ontologie permet de formaliser des liens sémantiques entre des concepts unités de sens. Définie dans un cadre possibiliste, elle pourrait ajouter de l’information pertinente à considérer lors du processus de propagation déclenchée par la requête. Le réseau serait composé d’un sous réseau documents et d’un sous réseau requête. Ces sous réseaux pourraient être reliées à travers une ontologie.

-
- d'intégrer des relations entre paire de documents dans un cadre numérique ou ordinal. Les relations de dépendance entre paires de documents pourraient traduire des liens sémantiques ou statistiques évaluant les distributions des termes communs à des paires ou ensembles de documents. Les termes ou les documents peuvent ainsi être regroupés dans des classes communes ;
 - d'étendre notre modèle pour représenter des types de documents particuliers à savoir les documents *XML*. L'architecture du réseau se prête naturellement à ce type de représentation. En effet, la structure du document pourrait être traduite par la topologie du réseau, les noeuds intermédiaires correspondant aux balises du document et les noeuds feuilles aux termes des granules. L'application d'une telle approche à notre système permettrait d'évaluer le niveau de granularité (partie de documents, documents ou ensemble de documents) nécessairement et possiblement pertinents étant donnée une requête.

Bibliographie

- [1] ACID, S., DE CAMPOS, L., FERNANDEZ, J., AND HUETE, J. An information retrieval model based on simple bayesian networks. *International Journal of Intelligent Systems* 18, 2 (2003), 251–265.
- [2] BAEZA-YATES, R., AND RIBEIRO-NETO, B. *Modern information retrieval*. ACM Press / Addison-Wesley, 1999.
- [3] BALPE, J., LELU, A., AND SALEH, I. *Hypertextes et hypermédias : réalisations, outils et méthodes*. Paris : Hermès, 1995.
- [4] BENFERHAT, S., DUBOIS, D., GARCIA, L., AND PRADE, H. Possibilistic logic bases and possibilistic graphs. In *Proc. of the Conference on Uncertainty in Artificial Intelligence* (1999), pp. 57–64.
- [5] BERGER, A., AND LAFFERTY, J. Information retrieval as statistical translation. Research and development in information retrieval. In *Proc. of the International ACM-SIGIR Conference* (1999), pp. 222–229.
- [6] BOOKSTEIN, A., AND SWANSON, D. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science (JASIS)* 25 (1974), 312–318.
- [7] BOOKSTEIN, A., AND SWANSON, D. A decision theoretic foundation for indexing. *Journal of the American Society for Information Science (JASIS)* 26 (1975), 45–50.
- [8] BORGELT, C., GEBHARDT, J., AND KRUSE, R. Possibilistic graphical models. *Computational Intelligence in Data Mining, Courses and Lectures 408*, Springer, Wien 26 (2000), 51–68.
- [9] BORLUND, P., AND INGWERSEN, P. Measures of relative relevance and ranked half-life : performance indicators for interactive ir. In *Proc. of the International ACM-SIGIR conference* (1998), pp. 24–28.

-
- [10] BOUGHANEM, M. Systèmes de recherche d'informations : d'un modèle classique à un modèle connexionniste, 1992. Thèse de doctorat, Univ. Paul Sabatier, Toulouse III.
- [11] BOUGHANEM, M. Formalisation et spécification des systèmes de recherche et de filtrage d'information, 2000. Hdr, Univ. Paul Sabatier, Toulouse III.
- [12] BOUGHANEM, M., AND BRINI, A. Introduction de la gradualité dans le jugement utilisateur. In *Actes de la conférence Extraction et Gestion des Connaissances (EGC)* (2003), vol. 17, pp. 343–348.
- [13] BRINI, A., AND BOUGHANEM, M. Relevance feedback : introduction of partial assessments for query expansion. In *Proc. of the Conference of the European Society for Fuzzy Logic and Technology, (EUSFLAT)* (2003), pp. 67–72.
- [14] BRINI, A., BOUGHANEM, M., AND DUBOIS, D. Towards a possibilistic approach for information retrieval. In *Proc. of the conference EURO-FUSE, Data and Knowledge Engineering* (2004), pp. 92–102.
- [15] BRINI, A., BOUGHANEM, M., AND DUBOIS, D. Une approche possibiliste pour la recherche d'information. In *Logique Floue et ses Applications, (LFA 2004)* (2004), pp. 51–58.
- [16] BRINI, A., BOUGHANEM, M., AND DUBOIS, D. Vers une approche possibiliste pour la recherche d'information. In *Veille Stratégique Scientifique et Technologique, (VSST 2004)* (2004), pp. 55–65.
- [17] BRINI, A., BOUGHANEM, M., AND DUBOIS, D. A model for information retrieval based on possibilistic networks. In *Proc. of the symposium on String Processing and Information REtrieval (SPIRE 2005), LNCS, Springer* (2005), pp. 271–282.
- [18] BRINI, A., CAMPOS, L., DUBOIS, D., AND BOUGHANEM, M. Query propagation in possibilistic information retrieval networks. In *Proc. of the Conference of the European Society for Fuzzy Logic and Technology, (EUSFLAT 2005)* (2005).
- [19] BRUZA, P., AND VAN DER GAAG, L. Index expression belief networks for information disclosure. *International Journal of Expert Systems* 7, 2 (1994), 107–138.
- [20] CALADO, P., CRISTO, M., DE MOURA, E., ZIVIANI, N., RIBEIRO-NETO, B., AND GONÇALVES, M. A. Combining link-based and content-based methods for web document classification. In *Proc. of ACM*

- Conference on Information and Knowledge Management (CIKM)* (2003), pp. 394–401.
- [21] CALLAN, J., CROFT, W., AND HARDING, S. The INQUERY retrieval system. In *Proc. of International Conference on Database and Expert Systems Applications (DEXA)* (1992), pp. 78–83.
- [22] CLEVERDON, C. Progress in documentation. evaluation of information retrieval systems. *Journal of Documentation* 26 (1970), 55–67.
- [23] CLEVERDON, C. Comparative evaluation of searching by controlled and natural language in a nasa database, 1977. European Space Agency Report. 1/432.
- [24] CLEVERDON, C. W. <http://www.gnu.org/>, 1960. slib Cranfield research project : Report on the first stage of an investigation into the comparative efficiency of indexing systems. Cranfield : The College of Aeronautics.
- [25] COOPER, G. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence* 42, 2-3 (1990), 393–405.
- [26] COUSINS, S., CHEN, W., AND FRISSE, M. Creating a query network for a new medical domain. In *Proc. of the Symposium on Computer Applications in Medical Care (UW-SIG)* (1990), IEEE Computer Society, pp. 800–804.
- [27] COUSINS, S., FRISSE, M., CHEN, W., AND HASSAN, S. Approaches to information filtering in medicine, 1991. Technical Report. Medical Informatics Laboratory, Washington University.
- [28] COUSINS, S., SILVERSTEIN, J., AND FRISSE, M. Query networks for medical information retrieval—assigning probabilistic relationships. In *Proc. of the Symposium on Computer Applications in Medical Care (UW-SIG)* (1991), IEEE Computer Society Press, pp. 803–807.
- [29] CRESTANI, F. Comparing neural and probabilistic relevance feedback in an interactive information retrieval system. In *Proc. of the IEEE International Conference on Neural Networks (ICNN)* (1994), pp. 3426–3430.
- [30] CRESTANI, F., DE CAMPOS, L. M., FERNÁNDEZ-LUNA, J. M., AND HUETE, J. F. Ranking structured documents using utility theory in the bayesian network retrieval model. In *Proc. of the symposium on String Processing and Information REtrieval (SPIRE)* (2003), pp. 168–182.

- [31] CROFT, W., AND BRUCE, W. Approaches to intelligent information retrieval. *Information Processing and Management : an International Journal* 23, 4 (1987), 249–254.
- [32] CROFT, W., AND TURTLE, H. R. Text retrieval and inference. *Text-Based Intelligent Systems. Current Research and Practice in Information Extraction and Retrieval* (1992), 127–155.
- [33] DAS-NEVES, F. A tri-valued belief network model for information retrieval, 2001. Technical Report TR-01-25. Computer Science Department. Virginia Polytechnic Institute and State University.
- [34] DE CAMPOS, L., FERNÁNDEZ-LUNA, J., AND HUETE, J. Two term-layers : an alternative topology for representing term relationships in the bayesian network retrieval model. *Advances in soft computing-engineering, design and manufacturing*, J. Benitez, O. Cordon, F. Hoffmann, et R. Roy (Eds.), Springer-Verlag (1988), 213–224.
- [35] DE CAMPOS, L., FERNÁNDEZ-LUNA, J., AND HUETE, J. A layered bayesian network model for document retrieval. In *Proc. of the 24th BCS-IRSG European Colloquium on IR Research : Advances in Information Retrieval* (2002), pp. 169–182.
- [36] DE CAMPOS, L., FERNÁNDEZ-LUNA, J., AND HUETE, J. Improving the efficiency of the bayesian network retrieval model by reducing relationships between terms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11, Suplément (2003), 101–116.
- [37] DENOYER, L., WISNIEWSKI, G., AND GALLINARI, P. Document structure matching for heterogeneous corpora, 2004. Proc. of the International ACM-SIGIR Conference : Workshop on XML and Information Retrieval.
- [38] DOMINICH, S. *Mathematical Foundations of Information Retrieval*. Kluwer Academic Publishers, Dordrecht, Boston, London, 2001.
- [39] DUBOIS, D., AND PRADE, H. *Possibility Theory*. Plenum, 1988.
- [40] FARRADANE, J. Relational indexing part I. *Information Science* 1, 5 (1980), 267–276.
- [41] FOX, C. Lexical analysis and stoplists. *Information Retrieval : Data Structures and Algorithms* (1992), 102–130.
- [42] FRAKES, W. Stemming algorithms. *Information Retrieval : Data Structures and Algorithms* (1992), 131–160.
- [43] FRISSE, M. Searching for information in a hypertext medical handbook. *Communications of the ACM (CACM)* 31, 7 (1988), 880–886.

- [44] FRISSE, M., AND COUSINS, S. Information retrieval from hypertext : Update on the dynamic medical handbook. In *In Proc. of ACM Hypertext Conference (1989)*, pp. 199–211.
- [45] FROEHLICH, T., AND EISENBERG, M. Special topic issue on relevance research. *Journal of the American Society for Information Science (JASIS)* 45, 3 (1994), 124–134.
- [46] FUHR, N. Probabilistic models in information retrieval. *The Computer Journal* 35, 3 (1992), 243–255.
- [47] FUHR, N. Language models and uncertain inference in information retrieval, 2001. In *Proc. of the Language Modeling and IR workshop*.
- [48] FUNG, R. M., AND FAVERO, B. D. Applying bayesian networks to information retrieval. *Communications of the ACM (CACM)* 38, 3 (1995), 42–48.
- [49] GHAZFAN, D., INDRAWAN, M., AND SRINIVASAN, B. Toward meaningful bayesian networks for information retrieval systems. In *Proc. of the International Conference in Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)* (1996), pp. 841–846.
- [50] HARMAN, D. Relevance feedback and other query modification techniques. In *Information Retrieval : Data Structures and Algorithms* (1992), William B. Frakes and Ricardo Baeza-Yates, editors, Prentice Hall, Englewood, Cliffs, NJ, pp. 241–263.
- [51] HARTER, S. A probabilistic approach to automatic keyword indexing. part ii. an algorithm for probabilistic indexing. *Journal of the American Society for Information Science (JASIS)* 35, 3 (1975), 280–289.
- [52] HARTER, S. Psychological relevance and information science. *Journal of the American Society for Information Science (JASIS)* 43, 9 (1992), 602–615.
- [53] HIEMSTRA, D., AND KRAAJ, W. Twenty-one at trec-7 : Ad hoc and cross language track. In *Proc. of the Text REtrieval Conference (TREC-7)* (1998), pp. 227–238.
- [54] IDE, E. New experiments in relevance feedback, 1971. In Salton G., editor, *The Smart System - Experiments in Automatic Document Processing*, Englewood Cliffs, NJ, Prentice-Hall Inc.
- [55] IJDENS, J., BRUZA, P., AND HARPER, D. Probabilistic inference experiments using the ECLAIR system. Tech. rep., 1995.

- [56] INDRAWAN, M., GHAZÍON, D., AND SRINIVASAN, B. Using bayesian networks as retrieval engines. In *Proc. of the Text REtrieval Conference (TREC-6)* (1996), pp. 437–444.
- [57] INDRAWAN, M., SRINIVASAN, B., WILSON, C., AND REDPATH, R. Optimising bayesian belief networks : the case study of information retrieval systems. In *Proc. of the IEEE Conference on System, Man and Cybernetics* (1998), pp. 2273–2278.
- [58] INDRAWAN, M., WILSON, C., AND SRINIVASAN, B. Relevance feedback in an information retrieval system based on bayesian networks. In *Proc. of the International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA)* (1999), pp. 266–271.
- [59] JENSEN, F. V. *Bayesian networks and decision graphs*. Springer Verlag, 2000.
- [60] JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1972), 111–121.
- [61] JONES, K. S. Does indexing exhaustivity matter? *Journal of the American Society for Information Science (JASIS)* 24 (1973), 313–316.
- [62] JONES, K. S. Exhaustivity and specificity. In *Journal of Documentation* 60, 5 (2004), 493–502.
- [63] JONES, K. S., WALKER, S., AND ROBERTSON, S. A probabilistic model of information retrieval : development and comparative experiments, parts 1 & 2. *Information Processing & Management (IPM)* 36, 6 (2000), 779–808,809–840.
- [64] KEEN, E., AND DIGGER, J. *Report of an Information Science Index Languages Test*. Aberystwyth College of Librarianship, Wales, 1972.
- [65] KEKÄLÄINEN, J., AND JÄRVELIN, K. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In Bruce, H., Fidel, R., P. Ingwersen, P. Vakkari, eds. *Emerging Frameworks and Methods, Seattle, Colerado : Libraries Unlimited* (2002), pp. 253–270.
- [66] KWOK, K. A neural network for probabilistic information retrieval. In *Proc. of the International ACM-SIGIR Conference* (1989), pp. 21–30.
- [67] LAFFERTY, J., AND ZHAI, C. *Probabilistic relevance models based on document and query generation.*, vol. 13. Kluwer Academic, 2003.
- [68] LANCASTER, F., AND WARNER, A. *Information retrieval today*. Arlington : Information Resources Press, 1993.

- [69] LELU, A., AND FRANÇOIS, C. Information retrieval based on neural unsupervised extraction of thematic fuzzy clusters. In *Proc. of the International Conference, Neural Networks and their Applications* (1992), pp. 93–104.
- [70] LUHN, H. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 4, 1 (1957), 309–317.
- [71] LUHN, H. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 24, 2 (1958), 159–165.
- [72] MANDALA, R., TOKUNAGA, T., AND TANAKA, H. Combining multiple evidence from different types of thesaurus for query expansion. *Proc. of the International ACM-SIGIR Conference*, 1 (1972), 191–197.
- [73] MANIEZ, J., AND DE GROLIER, E. A decade of research in classification. *International Classification* 18, 2 (1991), 73–77.
- [74] MARON, M. Automatic indexing : an experimental enquiry. *Journal of the ACM* 24, 8 (1961), 404–417.
- [75] MARON, M., AND KUHN, J. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery* 7 (1960), pages 216–244.
- [76] MIZZARO, S. Relevance : the whole history. *Journal of the American Society for Information Science (JASIS)* 48, 9 (1997), 810–832.
- [77] PEARL, J. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann San Mateo, Ca, 1988.
- [78] PONTE, J. M., AND CROFT, W. B. A language modeling approach to information retrieval. research and development in information retrieval. In *Proc. of the International ACM-SIGIR Conference* (1998), Proc. of the International ACM-SIGIR Conference, pp. 275–281.
- [79] PORTER, M. F. An algorithm for suffix stripping. Program 14, 1980.
- [80] QIU, Y., AND FREI, H. Concept based query expansion. In *Proc. of the International ACM-SIGIR Conference* (1993), pp. 160–169.
- [81] RIBEIRO-NETO, B., AND MUNTZ, R. R. A belief network model for ir. In *Proc. of the International ACM-SIGIR Conference* (1996), pp. 253–260.
- [82] RIJSBERGEN, C. V. A theoretical basis for the use of co-occurrence data in information retrieval. In *Journal of Documentation*, 33 (1977), 106–119.

-
- [83] RIJSBERGEN, C. V. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, 1979.
- [84] ROBERTSON, S., AND JONES, K. S. Relevance weighting of search terms. *Journal of the American Society for Information Science (JASIS)* 27, 3 (1976), 129–146.
- [85] ROBERTSON, S., MARON, M. E., AND COOPER, W. S. Probability of relevance : a unification of two competing models for information retrieval. *Information Technology - Research and Development* 1 (1982), 1–21.
- [86] ROBERTSON, S., VAN RIJSBERGEN, C., AND PORTER, M. Probabilistic models of indexing and searching. *Information retrieval research, (Ed. W.R. Oddy et al), London :Butteworths* (1981), 36–65.
- [87] ROBERTSON, S., AND WALKER, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of the International ACM-SIGIR Conference* (1994), pp. 232–241.
- [88] ROBERTSON, S., WALKER, S., AND BEAULIEU, M. Experimentation as a way of life : Okapi at trec. In *Information Processing and Management*, 36 (2000), 95–108.
- [89] ROBERTSON, S., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M., AND GATFORD, M. Okapi at trec-3. In *Proc. of the Text REtrieval Conference (TREC-3)* (1995), NIST Special Publication, pp. 109–126.
- [90] ROCCHIO, J. J. Relevance feedback in information retrieval. In *In G. Salton, editor, The SMART Retrieval System : Experiments in Automatic Document Processing* (1971), Prentice-Hall, Englewood Cliffs, NJ, pp. 313–323.
- [91] SALTON, G. *Automatic Information Organization and Retrieval*. New York : McGraw.Hill Book Company, 1968.
- [92] SALTON, G. *The Smart retrieval system-experiments*. Automatic Document Processing, Prentice Hall Inc, 1971.
- [93] SALTON, G. A theory of indexing. In *Technical report No. TR74-203* (1974), Department of Computer Science, Cornell University, Ithaca, New York, pp. 109–126.
- [94] SALTON, G. Syntactic approaches to automatic book indexing. In *Proc. of the annual meeting on Association for Computational Linguistics (ACL)* (1988), Department of Computer Science, Cornell University, Ithaca, New York, pp. 204–210.

-
- [95] SALTON, G., AND BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management (IPM)* 24, 5 (1988), 513–523.
- [96] SALTON, G., AND BUCKLEY, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science (JASIS)* 44, 4 (1990), 288–297.
- [97] SALTON, G., AND MCGILL, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [98] SALTON, G., WONG, A., AND YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM (CACM)* 18, 11 (1975), 613–620.
- [99] SALTON, G., AND YANG, C. On the specification of term values in automatic indexing. *In Journal of Documentation*, 29 (1973), 351–372.
- [100] SALTON, G., YANG, C., AND YU, C. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science (JASIS)*, 26 (1975), 33–44.
- [101] SARACEVIC, T. Relevance : A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science (JASIS)* 32, 2 (1975), 321–343.
- [102] SARACEVIC, T. Relevance reconsidered. In *Information science : Integration in perspectives* (1996), Proc. of the Conference on Conceptions of Library and Information Science, pp. 201–218.
- [103] SAVOY, J. A learning scheme for information retrieval in hypertext. *Information Processing and Management : an International Journal* 30, 4 (1994), 515–533.
- [104] SCHAMBER, L., EISENBERG, M., AND NILAN, M. A re-examination of relevance : Toward a dynamic, situational definition. *Information Processing and Management : an International Journal* 26, 6 (1990), 755–776.
- [105] SHAFER, G. *A mathematical theory of evidence*. Princeton Univ. Press, 1976.
- [106] SHANNON, C. E. The mathematical theory of communication. *Bell System Technical Journal* 27, 2 (1948), 379–423; 623–656.
- [107] SILVA, I., RIBEIRO-NETO, B. A., CALADO, P., DE MOURA, E., AND ZIVIANI, N. Link-based and content-based evidential information in a belief network model. In *Proc. of the International ACM-SIGIR Conference* (2000), pp. 96–103.

-
- [108] SINGHAL, A., BUCKLEY, C., AND MITRA, M. Pivoted document length normalization. *Proc. of the International ACM-SIGIR Conference 32*, 2 (1996), 21–29.
- [109] SINGHAL, A., MITRA, M., AND BUCKLEY, C. Learning routing queries in a query zone. In *In Proc. of the International ACM-SIGIR Conference* (1997), pp. 25–32.
- [110] SINGHAL, A., SALTON, G., MITRA, M., AND BUCKLEY, C. Document length normalization. *Information Processing & Management (IPM)* 32, 5 (1996), 619–633.
- [111] TURTLE, H. Inference networks for document retrieval, 1991. Ph.D. thesis, University of Massachusetts.
- [112] TURTLE, H., AND CROFT, W. Inference networks for document retrieval. In *Proc. of the International ACM-SIGIR Conference* (1990), pp. 1–24.
- [113] TURTLE, H. R., AND CROFT, W. B. Evaluation of an inference network-based retrieval model. *ACM Transaction on Information Systems* 9, 3 (1991), 7187–222.
- [114] VAN RIJSBERGEN, C. J. A non-classical logic for information retrieval. *In Computer Journal* 29, 6 (1986), 481–485.
- [115] VAN RIJSBERGEN, C. J. Towards an information logic. In *In Proc. of the International ACM-SIGIR Conference* (1989), pp. 77–86.
- [116] YAGER, R. R., AND LARSEN, H. L. Retrieving information by fuzzification of queries. *Journal of Intelligent Information Systems* 2, 4 (1993), 106–119.
- [117] YANG, J. J., AND KORFHAGE, R. Query optimization in information retrieval using genetic algorithms. In *Proc. of the International Conference on Genetic Algorithms* (1993), pp. 603–611.
- [118] ZADEH, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1 (1978), 3–28.
- [119] ZIPF, H. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge, Massachusetts, 1949.