FUSION DE COLLECTIONS ET DE DONNÉES EN RECHERCHE D'INFORMATIONS: MODÈLES BASÉS SUR LA RÉGRESSION LOGISTIQUE

Thèse

présentée à la Faculté des sciences de Neuchâtel

spécialité informatique

pour l'obtention du grade de Docteur ès sciences

par

ANNE LE CALVÉ

Directeur de thèse :

Prof. Jacques SAVOY, Université de Neuchâtel, Suisse

Co-directeur de thèse :

Prof. Pierre-Jean ERARD, Université de Neuchâtel, Suisse

Experts:

Prof. Yves CHIARAMELLA, Université Joseph Fourier, Grenoble, France

Prof. Jianyun NIE, Université de Montréal, Canada

© Anne LE CALVÉ 1998

IMPRIMATUR POUR LA THÈSE

Fusion de collections et de données en recherche d'informations: modèles basés sur la régression logistique

de Mme Anne Le Calvé

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel sur le rapport des membres du jury,

MM. P.-J. Erard (co-directeur de thèse), J. Savoy (co-directeur de thèse), Y. Chiaramella (Grenoble) et J. Nie (Montréal)

autorise l'impression de la présente thèse.

Neuchâtel, le 23 octobre 1998

Le doyen:

F. Stoeckli

F. Stoeckli

Je tiens à remercier,

Le professeur Jacques SAVOY de l'Université de Neuchâtel, qui m'a guidée dans l'élaboration de cette thèse. Sa disponibilité, ses remarques constructives et la qualité de son encadrement m'ont été très précieuses.

Les professeurs Yves CHIARAMELLA, Pierre-Jean ERARD et Jianyun NIE pour le temps consacré à l'expertise de cette thèse.

C. Buckley de l'université de Cornell et de "SABIR Research", pour l'utilisation du système SMART.

Mes collègues et amis Justin PICARD, Miriam SCAGLIONE et Dana VRAJITORU, pour leurs conseils et leur aide.

Silvia JACQUET et Vanessa GALESNE pour leur soutien et leurs encouragements.

Cette recherche a été soutenue en partie par le Fonds national suisse de la recherche scientifique avec les subventions 20-43'217.95 et 20-50'578.97.

A mes parents.

TABLE DES MATIERES

INTRODUCTION

I. LA RECHERCHE D'INFORMATIONS 9

I.1 Définitions et principes 9
I.2 Les principaux systèmes de recherche 14
I.3 L'évaluation 16

II. PROBLÉMATIQUE GÉNÉRALE ET PLAN 17

II.1 Introduction au problème de fusion 17
II.2 Objectifs 19
II.3 Plan de la thèse 20

PARTIE 1 : LE PROBLEME DE FUSION

I. INTRODUCTION 22

II. LA FUSION DE COLLECTIONS 22

II.1 Présentation 22
II.2 Les stratégies existantes 24

III. LA FUSION DE DONNÉES 26

III.1 Présentation 26
III.2 Les stratégies existantes 28

PARTIE 2 : LA REGRESSION LOGISTIQUE APPLIQUEE A LA FUSION DE COLLECTIONS

I. PRÉSENTATION DES DONNÉES 31

1.1 Les collections-tests 31
I.2 Les variables utilisées 32

II. LE MODÈLE MATHÉMATIQUE PROPOSÉ 38

II.1 La régression logistique 38
II.2 Développement des modèles 43
II.3 Interprétation des modèles 47

III. ÉVALUATION ET COMPARAISON 49
III.1 Évaluation de plusieurs systèmes de dépistage de l'information 49
III.2 Fusion de collections avec systèmes de recherche similaires 50
III.3 Fusion de collections avec des systèmes de recherche différents 52
IV. CONCLUSION 55

PARTIE 3 : LA REGRESSION LOGISTIQUE APPLIQUEE A LA FUSION DE DONNÉES

I. PRÉSENTATION DES DONNÉES 58
II. LE MODÈLE MATHÉMATIQUE PROPOSÉ 60
II.1 La régression logistique 60
II.2 Développement des modèles 66
III. ÉVALUATION ET COMPARAISON 68
IV. QUELQUES RÉFLEXIONS SUR LA FUSION DE DONNÉES 69
V. CONCLUSION 72

CONCLUSION 74

ANNEXES 78

BIBLIOGRAPHIE 90

TABLE DES TABLEAUX

Tableau 1 Classification et exemples de stratégies de fusion de collections 26

Tableau 2 Statistiques sur les collections utilisées 32

Tableau 3 Coefficients de corrélation de Pearson (ρ) entre les variables Sim, SimDécomp, et ln(rang) 35

Tableau 4 Statistiques descriptives des variables 36

Tableau 5 Statistiques sur les documents extraits (WSJ90-OKAPI, 186 requêtes) 37

Tableau 6 Valeurs des coefficients de la régression logistique pour le Modèle 1 43

Tableau 7 Exemple de fusion basée sur le Modèle 1 44

Tableau 8 Valeurs des coefficients de la régression logistique pour le Modèle 2 45

Tableau 9 Valeurs des coefficients de la régression logistique pour le Modèle 3 46

Tableau 10 Ratio $r(x_i)$ obtenu par le Modèle 1 (WSJ91-LNU, 171 requêtes) 47

Tableau 11 Rangs des documents correspondant à des valeurs particulières du ratio $r(x_i)$ dans le Modèle 1 48

Tableau 12 Évaluation de plusieurs systèmes de recherche 50

Tableau 13 Évaluation de plusieurs stratégies de fusion de collections utilisant le même système de dépistage 51

Tableau 14 Évaluation de plusieurs stratégies de fusion de collections utilisant des systèmes de recherche différents 52

Tableau 15 Résumé des résultats des modèles logistiques 55

Tableau 16 Statistiques descriptives des variables(WSJ, 186 requêtes) 59

Tableau 17 Coefficients de corrélation de Pearson (ρ) entre les variables et g(x) (OKAPI, WSJ, 186 requêtes) 61

Tableau 18 Statistiques descriptives des variables dans le modèle combiné (WSJ, 186 requêtes) 64

Tableau 19 Valeurs des coefficients de la régression logistique pour le Modèle 1 66

Tableau 20 Valeurs des coefficients de la régression logistique pour le Modèle 2 67

Tableau 21 Valeurs des coefficients de la régression logistique pour le Modèle 3 68

Tableau 22 Évaluation de plusieurs stratégies de fusion de données 69

Tableau 23 Évaluation de fusion de données par somme sur plusieurs combinaisons de systèmes 70

Tableau 24 Exemple de combinaisons de trois systèmes 71

TABLE DES FIGURES

Figure 1 Principales composantes des systèmes de recherche d'informations 10

Figure 2 Représentation graphique du modèle vectoriel 15

Figure 3 Principe de la fusion de collections 23

Figure 4 Principe de la fusion de données 27

Figure 5 Relation empirique entre le degré de similarité moyen et le rang (WSJ90-OKAPI, 170 requêtes) 33

Figure 6 Relation empirique entre le degré de similarité moyen et ln(rang) (WSJ90-OKAPI, 170 requêtes) 34

Figure 7 Nuage de points et estimation linéaire et logistique d'une variable binaire Y en fonction d'une variable explicative X 39

Figure 8 Relation empirique entre la variable ln(rang) et la probabilité de pertinence (WSJ90-OKAPI, 170 requêtes) 41

Figure 9 Intervalle de confiance à 95% pour les dix estimateurs $\hat{\beta}$ 54

Figure 10 Intervalle de confiance à 95% pour les dix estimateurs $\hat{\alpha}$ 54

Figure 11 Relation empirique entre la variable ln(rang) et la probabilité de pertinence (OKAPI, WSJ, 186 requêtes) 60

Figure 12 Modèle séparé de fusion de données 63

Figure 13 Modèle combiné de fusion de données 65

Figure 14 Évolution de la précision moyenne de fusion par somme pour plusieurs combinaisons de systèmes (WSJ, 186 requêtes) 71

TABLE DES ANNEXES

ANNEXE 1. 30 mots les plus fréquents en français 79
ANNEXE 2. Pondérations 80
ANNEXE 3. Exemples de requête 81
ANNEXE 4. Exemple de document 82
ANNEXE 5. Quelques statistiques sur les collections utilisées 84
ANNEXE 6. Coefficient de corrélation linéaire de Pearson 85
ANNEXE 7. Intervalle de confiance 86
ANNEXE 8. Histogrammes des variables pour wsj90-okapi 87

INTRODUCTION

I. LA RECHERCHE D'INFORMATIONS 9
I.1 Définitions et principes 9
I.2 Les principaux systèmes de recherche 14
I.3 L'évaluation 16
II. PROBLÉMATIQUE GÉNÉRALE ET PLAN 17
II.1 Introduction au problème de fusion 17
II.2 Objectifs 19
II.3 Plan de la thèse 20

I. LA RECHERCHE D'INFORMATIONS

I.1 Définitions et principes

De plus en plus d'informations sont mémorisées sur support électronique et sont rendues accessibles par des réseaux locaux (INTRANET) ou encore par des réseaux métropolitains ou à grande distance (INTERNET). Ces informations peuvent être représentées aussi bien sous forme de textes, de tables, d'images, de graphiques, de séquences vidéo que de sons. Dans le cadre de cette recherche, nous nous limiterons à l'information textuelle comme, par exemple, des articles de journaux, des articles scientifiques, des rapports d'entreprises, ou encore une norme technique. Devant l'ampleur de ce volume d'information, il devient de plus en plus difficile d'y retrouver le(s) fragment(s) désiré(s).

Le but de **la recherche d'informations** (RI) est alors le suivant :

Étant donné :

- une collection d'objets (documents dans notre cas), et
- un utilisateur pouvant exprimer son besoin d'information sous forme d'une requête,

$\Rightarrow$ comment retrouver rapidement tous les documents et seulement les documents qui répondent à la demande de l'usager exprimée de manière imprécise par la requête?

Sur INTERNET, une technique couramment utilisée pour répondre à cette question, était de naviguer à travers les sites, en suivant les liens qui mènent d'une page à l'autre. Cette méthode hasardeuse et fastidieuse, critiquée déjà en 1988 par Halasz ([88]), a laissé la place à des techniques plus spécialisées en recherche d'informations. Cependant, cette démarche s'avère utile lorsque l'utilisateur ne sait pas exprimer son besoin d'information, ou qu'il souhaite se faire une idée du contenu d'une collection (ou d'un site) donnée. Sur le Web, une structure arborescente permet, en partant d'un sujet très général, de resserrer sa recherche en sélectionnant de plus en plus précisément le sujet de l'information désirée. L'utilisateur peut également transposer son besoin d'information en mots, et soumettre une **requête** à des **systèmes automatiques de recherche d'informations** (SRI). Après une étape intermédiaire d'indexation, ces systèmes, dont la structure est schématisée dans la Figure 1, interrogent une collection de documents et retournent les documents considérés pertinents par la machine.

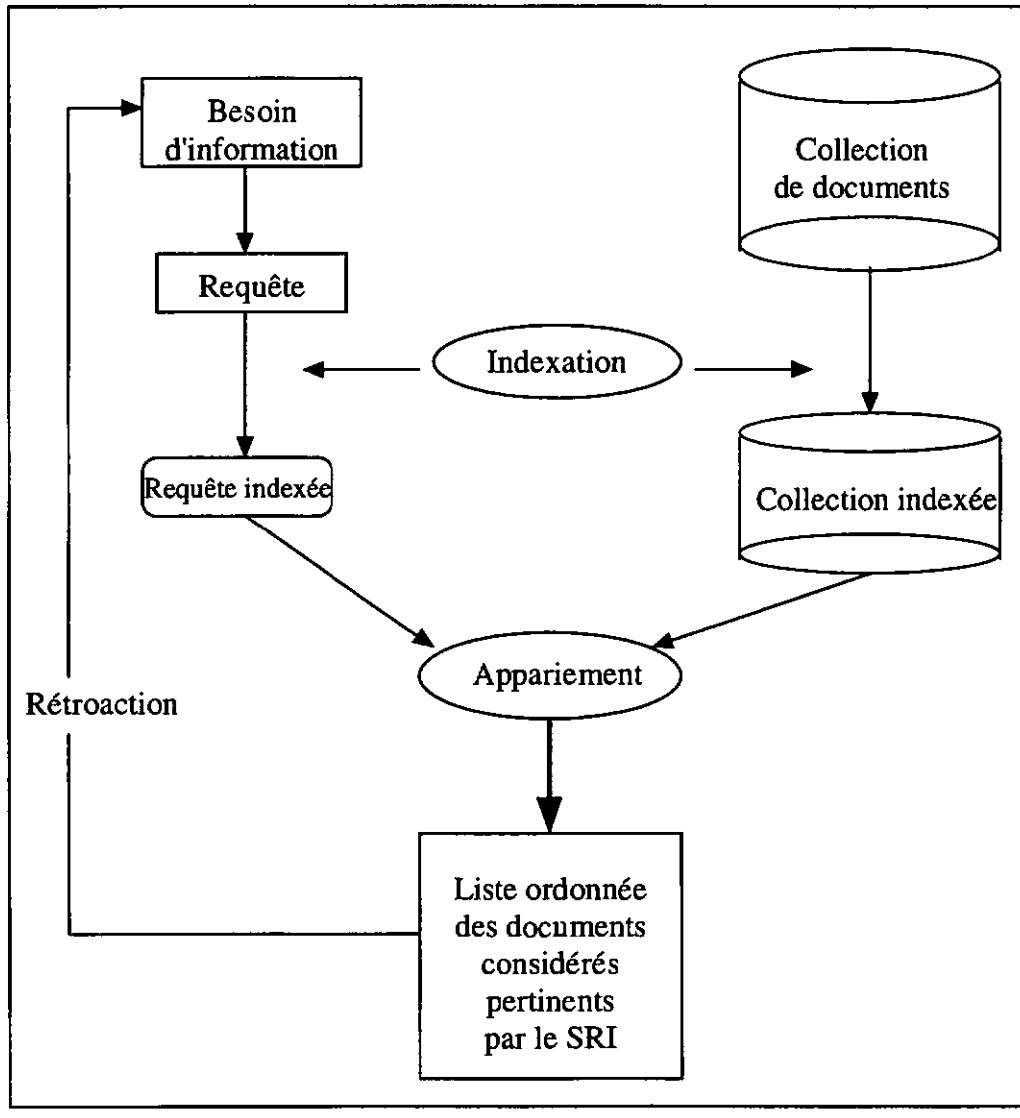

Figure 1: Principales composantes des systèmes de recherche d'informations

Plus précisément, le fonctionnement des systèmes de recherche d'informations peut se découper en trois étapes principales: l'indexation, l'appariement, la sélection et présentation des résultats. Les SRI intègrent parfois une quatrième étape : la rétroaction (ou "feedback").

1. L'indexation

Les systèmes de recherche d'informations ne travaillent habituellement pas directement sur les documents, mais sur une représentation de ces objets. De chaque document (qui peut être un article complet, un chapitre, une section, un paragraphe, ...) est extraite une liste de termes. Nous avons opté ici pour une représentation à l'aide d'une série de mots-clés pondérés, mais d'autres formes de représentations sont possibles. Le travail d'indexation peut être manuel ou automatique, cependant il a été démontré ([Salton 69], [Salton 72]) que l'indexation automatique est, en moyenne, tout aussi performante que l'indexation manuelle.

Dans le cadre des systèmes étudiés dans cette thèse, l'indexation automatique s'effectue selon l'algorithme général suivant:

pour chaque document D_i, i=1...n :

1. segmentation du texte en mots (suite de lettres et/ou de chiffres);
2. élimination des mots peu ou pas porteurs d'information (souvent les mots fréquents comme les articles, les pronoms,...);

   (exemple: liste de trente mots les plus fréquents en français qui recouvrent 38% des formes présentes dans les documents, annexe ANNEXE 1)
3. suppression des suffixes, ou plus précisément de certaines séquences terminales, dans l'espoir de ramener toutes les variations d'un mot à une forme commune;
4. détermination d'un poids w_{ij} qui indique l'importance relative de chaque terme T_j (mot, syntagme nominal) dans la description du document D_i.

Cette dernière pondération peut se calculer selon un large éventail de formules. La plus simple d'entre elle, nommée indexation binaire, consiste à accorder un poids unitaire si le terme est présent dans le document, et un poids nul dans le cas contraire. Mais souvent la pondération tient également compte du nombre d'occurrences (noté tf_{ij}), d'un terme T_j dans le document D_i, du nombre de documents dans lequel le terme T_j apparaît (df_j), ou encore de la longueur du document mesurée par le nombre de termes d'indexation associés au document D_i (nt_i). Ainsi, si un terme est très fréquent dans la collection, il faut alors lui accorder un poids plus faible car il ne permettra guère de distinguer un document d'un autre.

A la suite du travail d'indexation, les traits sémantiques importants des documents sont mémorisés sous forme de fichiers inversés. Dans ces fichiers, à chaque terme de l'indexation est associée la liste des documents dans laquelle il apparaît. Les SRI travaillent ensuite sur ses fichiers inversés, et non plus sur la collection de départ.

Les requêtes peuvent être considérées soit comme des conditions que doit satisfaire un document pour être retourné (modèles booléens), soit comme de petits documents (modèles vectoriels et probabilistes). Les requêtes passent alors également par une étape d'indexation, comme le montre le petit exemple suivant:

requête de départ: "les stratégies en fusion de collections"

Étapes de l'indexation:

1. les stratégies en fusion de collections
2. ~~les~~ stratégies ~~en~~ fusion ~~de~~ collections
3. stratégie~~s~~ fusion collection~~s~~
4. stratégie (1), fusion (1), collection (1)

2. L'appariement

Le système de recherche d'informations compare ensuite les termes de la requête avec ceux des représentants des documents, puis extrait les documents répondant aux critères de la requête (système booléen) ou évalue la similitude entre ces représentants et la requête. La forme et la mesure de ce degré de similitude varient d'un système à l'autre (différentes approches ont été proposées dans le cadre des modèles vectoriels et probabilistes).

3. Sélection et présentation

Le système présente enfin à l'utilisateur la liste des documents jugés répondre le mieux à la requête. Ces documents sont sélectionnés et ordonnés habituellement en fonction de la mesure d'appariement effectuée à l'étape précédente (sauf pour le système booléen qui fournit une réponse binaire, et non un degré de pertinence permettant un classement). De nombreux efforts sont faits en vue d'améliorer cette présentation des résultats. Certains systèmes mettent en surbrillance les termes responsables de la sélection des documents, ou mettent en évidence le passage contenant le plus de termes communs avec la requête. D'autres présentent une ligne de couleur dont la nuance reflète le degré de pertinence du document par rapport à la requête. Ces efforts permettent à l'utilisateur de retrouver plus facilement l'information désirée parmi les résultats. En effet, la recherche d'informations étant loin d'être parfaite, les premiers documents présentés ne sont pas toujours ceux qui correspondent réellement au besoin d'information de l'utilisateur.

4. Rétroaction

Comme la première recherche n'aboutit pas toujours aux résultats escomptés, une dernière étape possible consiste à rediriger la recherche à partir des jugements de pertinence que fournit l'usager sur les documents proposés. Ainsi, la requête peut être modifiée puis soumise une nouvelle fois avec un effort minimal de l'utilisateur:

- soit en ajustant le poids des termes de la requête en fonction du poids des termes des documents jugés pertinents;
- soit en lui ajoutant de nouveaux attributs;
- soit en choisissant comme nouvelle requête les termes d'un document considéré très pertinent;
- ou encore en combinant ces trois possibilités.

Ce principe d'apprentissage permet de mieux cibler le besoin d'information de l'utilisateur et permet très souvent d'améliorer la performance du système.

Les étapes du fonctionnement d'un système de recherche, décrites ci-dessus, ne sont que des étapes très générales, et les SRI peuvent être très différents. Ils se distinguent, par exemple, selon :

- la façon de considérer la requête (vue comme de petits documents ou comme une condition à satisfaire);
- la manière de représenter les documents (liste de termes, formules logiques, ...);
- le choix de l'indexation (manuelle ou automatique);
- la pondération des termes des documents et/ou de la requête;
- la mesure d'appariement entre les documents et la requête (mesure binaire ou degré de similitude);
- la sélection et la présentation générale des documents (ordonnancement ou non);
- ...

Les principaux systèmes de recherche et leurs caractéristiques sont présentés ci-dessous.

I.2 Les principaux systèmes de recherche

Différents modèles de systèmes de recherche d'informations existent, se distinguant chacun selon la manière de représenter la requête, les documents et la fonction d'appariement choisie. Trois grands systèmes classiques se dégagent : les modèles booléens, vectoriels et probabilistes. D'autres modèles, non présentés ici, ont été développés tels que les modèles logiques ou les modèles linguistiques.

- **Le modèle booléen classique**

Le modèle booléen classique utilise une indexation binaire pour représenter les documents. Dans ce système, un document est représenté comme un ensemble de termes indépendants et une requête booléenne comme une expression logique de termes reliés entre eux par les opérateurs ET, OU, ET NON. La requête exprime les conditions que doivent satisfaire les documents afin d'être retournés à l'usager. Une telle formulation de la requête peut paraître difficile pour des utilisateurs peu entraînés.

L'appariement entre la requête et les documents est très simple: le document remplit ou ne remplit pas les conditions logiques de la requête. Dans ce modèle booléen classique, les documents retrouvés ne sont donc pas ordonnés selon un degré de similitude mais selon des attributs externes tels que la date de publication du document, le titre ou l'auteur du document.

Plusieurs variations du modèle booléen classique existent, introduisant par exemple un degré de pertinence, une pondération non binaire des termes, ou encore s'appuyant sur la logique floue.

- **Le modèle vectoriel**

Dans le modèle vectoriel ([Salton 71]), d'inspiration essentiellement géométrique, les documents comme les requêtes, sont représentés sous forme de vecteurs dont chaque composante correspond au poids du terme dans le document ou la requête (voir Figure 2). Plus l'angle entre le vecteur document et le vecteur requête est petit, et plus le document répond à la requête.

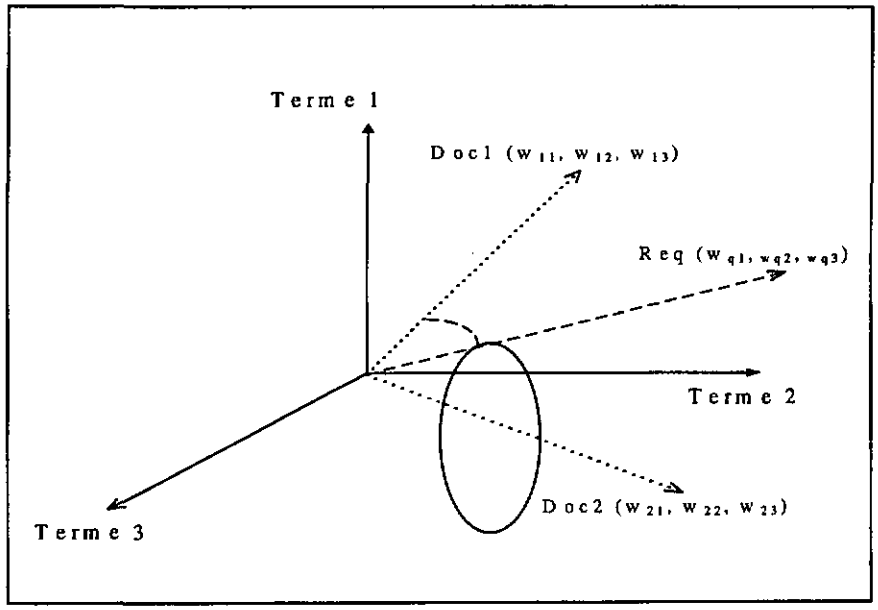

Figure 2: Représentation graphique du modèle vectoriel

Contrairement au modèle booléen, une requête s'écrit en langue naturelle (phrase ou suite de mots). La pondération des termes peut se faire selon diverses formules (cf. annexe ANNEXE 2). Le système calcule ensuite le degré de **similarité**, c'est-à-dire la ressemblance entre les vecteurs des documents et le vecteur de la requête, et sélectionne les documents qui possèdent le plus haut degré de similarité. De même que pour la pondération des termes, la similarité peut se calculer de différentes façons. Une formule largement répandue, car son interprétation géométrique est très aisée, est celle du cosinus. Son évaluation est la suivante:

$$\mathrm{Sim}(Q,D_i) = \frac{\sum_{k=1}^{n} W_{ik} \cdot W_{qk}}{\sqrt{\sum_{k=1}^{n} W_{ik}^2 \cdot \sum_{k=1}^{n} W_{qk}^2}}$$

où w_{ik} est le poids du terme T_k dans le document D_i et w_{qk} le poids du terme T_k dans la requête Q.

Les documents qui possèdent un degré de similarité élevé sont sélectionnés et présentés à l'usager par ordre décroissant de cette similarité.

- **Le modèle probabiliste**

Dans les modèles probabilistes, possédant un fondement théorique plus avancé ([Robertson 77]), les documents retrouvés sont classés selon leur probabilité de pertinence par rapport à la requête. La probabilité est considérée sous sa forme conditionnelle et est estimée, à partir du théorème de Bayes, selon la formule générale suivante.

$$P(\text{Pert}|D,Q) = \frac{P(D,Q|\text{Pert}) \cdot P(\text{Pert})}{P(D,Q|\text{Pert}) \cdot P(\text{Pert}) + P(D,Q|\text{NonPert}) \cdot P(\text{NonPert})}$$

où :

- P(Pert) et P(NonPert) sont les probabilités a priori d'obtenir un document pertinent, respectivement non pertinent;
- P(D,Q | Pert) et P(D,Q | NonPert) sont les probabilités d'obtenir le document D, et une requête Q, sachant que le document est pertinent, respectivement non pertinent

Plusieurs estimations de ces probabilités sont possibles et conduisent à des modèles différents. Très souvent, par souci de simplification, les termes sont considérés indépendants.

Comme de nombreux modèles existent, il s'avère intéressant de connaître leur performance respective. Pour cela, plusieurs méthodes d'évaluation ont été développées, et l'une des plus courantes est introduite ci-dessous.

I.3 L'évaluation

Plusieurs éléments peuvent entrer en compte pour juger de l'efficacité d'un système de recherche. L'évaluation peut tenir compte par exemple du temps de réponse du système, de sa facilité d'utilisation ou encore du degré de couverture de la collection. Deux mesures sont couramment utilisées en recherche d'informations pour juger de la qualité de la réponse d'un système: la précision et le rappel.

La **précision** correspond au rapport entre le nombre de documents pertinents retrouvés sur le nombre de documents retrouvés.

$$P = \frac{|\text{Pertinents} \cap \text{Retoumés}|}{|\text{Retoumés}|} \qquad \text{avec } |\text{Retournés}| \neq 0$$

Le **rappel** indique la proportion de documents pertinents retrouvés sur l'ensemble des documents pertinents.

$$R = \frac{|\text{Pertinents} \cap \text{Retournés}|}{|\text{Pertinents}|} \qquad \text{avec } |\text{Pertinents}| \neq 0$$

Ces deux notions apportent des informations complémentaires. Cependant, lorsqu'il s'agit de comparer plusieurs systèmes de recherche, il n'est guère aisé de devoir manipuler deux indices. Une troisième mesure, issue de la combinaison de la précision et du rappel, est donc utilisée pour déterminer la performance d'un système: **la précision moyenne à onze points fixes de rappel**, et que nous appellerons par la suite "précision moyenne" ([van Rijsbergen 79]). Pour calculer cette mesure d'évaluation, reprise dans les expériences de la conférence TREC, il faut fixer le rappel à onze valeurs (0, 0.1, ..., 0.9, 1), et prendre la moyenne des précisions obtenues pour ces valeurs de rappel. Un système A est jugé significativement meilleur qu'un système B si la précision moyenne de A est supérieure d'au moins 5% à celle de B. Une différence de 10% est considérée comme très significative ([Sparck Jones 77]). Cette règle ad hoc est généralement admise mais de récents travaux tentent de proposer une méthode plus rigoureuse ([Savoy 97]).

D'un côté pratique, le temps de réponse ou la manière optimale de stocker l'information dans les fichiers inversés sont également des mesures intéressantes. Cependant, si ces mesures sont dignes d'intérêt, elles ne correspondent pas à notre but premier et ne seront pas considérées ultérieurement dans ce travail de recherche.

Pour obtenir la précision et le rappel, il est nécessaire de savoir quels sont les documents pertinents pour la requête en cours. Bien entendu cette information n'est pas connue en situation réelle, mais dans le cadre d'un travail de recherche, nous disposons de collections-tests qui possèdent, pour chaque requête, ces jugements de pertinence. De telles collections sont très coûteuses à mettre sur pied et, si leur taille est parfois jugée faible par rapport aux informations disponibles sur le Web, elle est très raisonnable pour bien des applications (par exemple 200'000 documents pour la collection Wall Street Journal, une des collections de TREC).

II. PROBLÉMATIQUE GÉNÉRALE ET PLAN

II.1 Introduction au problème de fusion

Devant le volume sans cesse croissant d'information disponible sur les réseaux locaux (INTRANET) ou sur INTERNET, il devient indispensable de posséder des systèmes automatiques performants de recherche, capables d'extraire les documents pertinents à une requête, et de rejeter les autres. Dans notre recherche, nous avons choisi d'utiliser des modèles vectoriels et probabilistes, laissant de côté les modèles booléens possédant une performance moyenne nettement inférieure à celle des modèles retenus. Cependant, si les modèles retenus offrent en moyenne une performance similaire sur l'ensemble des requêtes, la qualité de la réponse pour une requête donnée est en fait assez variable d'un modèle à l'autre. Il est alors très difficile de déterminer a priori quelle stratégie de recherche répondra au mieux à la requête en cours ([Savoy 96]).

Plutôt que de se limiter à une performance moyenne "raisonnable" obtenue par une stratégie donnée, le domaine de la **fusion de données** suggère de combiner les résultats de plusieurs modèles de recherche interrogeant la même collection de documents. Le but consiste à fusionner adéquatement les différentes listes de documents proposées par chaque modèle pour créer une nouvelle et unique liste finale, en moyenne plus performante que les premières.

Ce premier problème de fusion n'est pas le seul, et le monde INTERNET nous pose celui de la **fusion de collections**. Ainsi, après avoir sélectionné les serveurs les plus prometteurs et leur avoir soumis une requête, le système de recherche doit être en mesure de combiner, fusionner les listes de résultats provenant de ces différents sites pour n'en proposer qu'une seule à l'usager. Dans ce cadre, les listes à fusionner sont obtenues par des systèmes de dépistage, souvent différents, qui interrogent des collections différentes.

Jusqu'à présent la fusion de collections et la fusion de données ont été analysées comme deux problèmes distincts, et des solutions différentes ont été proposées. Si ces deux problèmes possèdent leurs propres caractéristiques et difficultés sous-jacentes, ils n'en soulèvent pas moins les mêmes questions:

- combien de documents faut-il extraire de chacune des listes?
- quels documents doivent être sélectionnés?
- quels critères (degré de similarité et/ou rang) sont utiles et efficaces pour aider le système à sélectionner les documents?
- sur la base de ces critères, quelle stratégie doit être utilisée pour combiner les différentes listes de documents?

Dès lors, notre optique diffère des travaux déjà réalisés dans ces deux domaines car nous considérons la fusion de collections et la fusion de données comme une seule et même question: comment combiner de manière optimale des listes de documents? Nous voulons donc aborder les deux problèmes sous le même angle, et rechercher une stratégie globale de combinaison d'évidences multiples.

De plus, les stratégies proposées jusqu'ici, tant en fusion de collections qu'en fusion de données, reposent très souvent sur des heuristiques. Il nous semble important d'adopter une approche plus théorique qui puisse ainsi se justifier et s'employer quelles que soient les modalités pratiques de la fusion.

Ces réflexions nous ont donc conduits à la volonté de déterminer, d'une part, un modèle théorique robuste commun aux deux problèmes de fusion, et d'autre part, à évaluer nos propositions et celles suggérées par d'autres auteurs.

II.2 Objectifs

Cette thèse développe une nouvelle stratégie de fusion basée sur la régression logistique afin de répondre conjointement au problème de la fusion de données et à celui de la fusion de collections. Pour cela, nous avons développé plusieurs modèles qui répondent chacun à différentes situations réelles. Leur évaluation montre la performance et la robustesse d'une telle stratégie par rapport aux solutions déjà existantes. Les objectifs initiaux qui ont permis de développer cette stratégie sont les suivants:

1. Développer un nouveau modèle théorique justifiant clairement la manière de combiner plusieurs évidences:
   - **1.1** trouver un modèle général pouvant s'appliquer aisément quelles que soient les informations disponibles pour effectuer la fusion;
   - **1.2** obtenir des paramètres stables et fiables pour notre modèle;
   - **1.3** ordonner clairement les documents retrouvés afin d'améliorer leur sélection lors de la fusion;
   - **1.4** en fusion de collections, savoir combien de documents il faut sélectionner dans chaque liste.
2. Évaluer empiriquement notre modèle et obtenir une performance significativement supérieure à celle des autres stratégies.

Pour atteindre ces objectifs, nous avons également tenu compte de plusieurs réflexions:

1. Les problèmes de fusion de données et de collections doivent être traités selon la même méthodologie;
2. Le rang des documents peut être un critère suffisant à lui seul pour effectuer la fusion (cette information pouvant être la seule disponible).

Nous ne désirons pas nous limiter au rang et nous avons également traité des situations dans lesquelles d'autres d'informations sont disponibles, comme le degré de similarité entre le document et la requête. Dans de tels cas, nous désirons vérifier si:

3. De nouvelles variables calculées à partir de la similarité et du rang peuvent apporter de meilleurs résultats.

II.3 Plan de la thèse

Après cette introduction, une première partie présente, avec plus de détails, la fusion de collections et la fusion de données, ainsi que les travaux d'autres auteurs s'y rapportant. Une deuxième partie est consacrée au développement de notre modèle théorique, adapté au problème de la fusion de collections. Dans ce but, cette partie présente tout d'abord les données sur lesquelles notre recherche s'est fondée, puis introduit et justifie le modèle statistique, et enfin évalue la solution proposée en la comparant avec les résultats d'autres stratégies. Le même plan est suivi dans la troisième partie consacrée à notre stratégie de fusion, adaptée cette fois à la fusion de données.

PARTIE 1

LE PROBLÈME DE FUSION

I. INTRODUCTION 22

II. LA FUSION DE COLLECTIONS 22

II.1 Présentation 22

II.2 Les stratégies existantes 24

III. LA FUSION DE DONNÉES 26

III.1 Présentation 26

III.2 Les stratégies existantes 28

I. INTRODUCTION

Le problème de fusion consiste à combiner différentes listes de documents en une liste unique et optimale à retoumer à l'usager. Ces listes peuvent provenir des résultats de l'interrogation de divers serveurs d'information ou de diverses sous-collections (**fusion de collections**) d'une part, et d'autre part, d'une même collection mais par des systèmes de recherche différents (**fusion de données**). Pour chacun de ces deux types de fusion, nous verrons tout d'abord une présentation du problème, puis les principaux travaux qui ont déjà été menés dans ces domaines.

II. LA FUSION DE COLLECTIONS

II.1 Présentation

Afin d'exploiter au mieux les divers sites d'INTERNET, l'utilisateur peut naviguer lui-même à travers les pages proposées ou utiliser des outils de recherche d'informations qui interrogent eux-mêmes ces sources. Mais le problème est très similaire sur les réseaux locaux (INTRANET) qui permettent d'accéder à une collection très volumineuse distribuée sur un thème donné, ou sur différentes collections abordant des sujets très différents et réparties sur plusieurs machines. Ces outils de recherche d'informations adaptés à cet environnement distribué, comportent les étapes principales décrites ci-dessous et schématisées dans la Figure 3.

1. **la sélection de collections**: choisir les sites (les collections) à interroger. Pratiquement, ce problème est souvent ignoré et toutes les sources disponibles sont interrogées.
2. **l'interrogation multiple**: soumettre la requête, convertie au préalable dans un format adéquat, aux systèmes de recherche associés aux différents serveurs;
3. **la fusion**: trier et sélectionner les documents les plus pertinents obtenus des différents serveurs et présenter une liste finale unique.

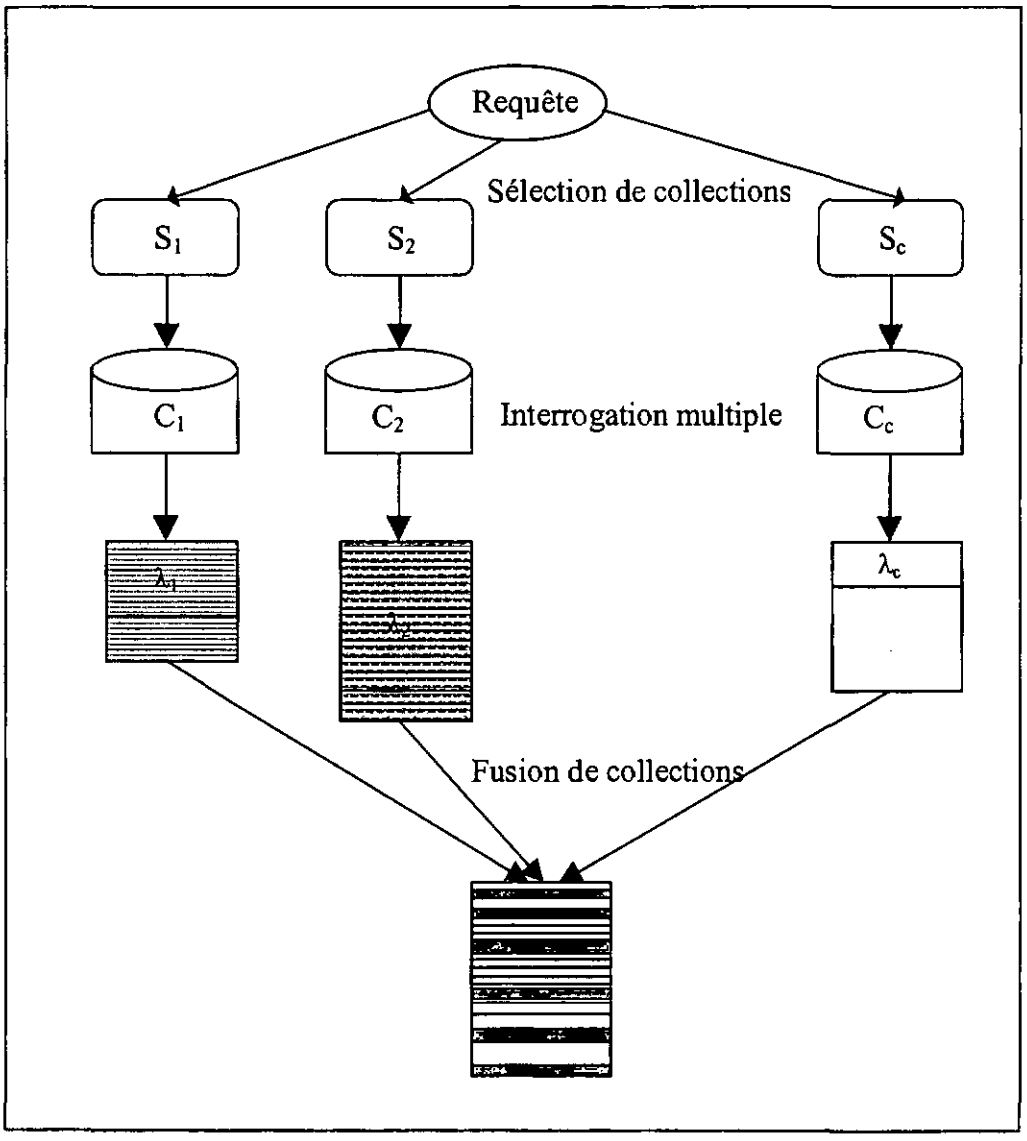

Figure 3 : Principe de la fusion de collections

Plus formellement, la **fusion de collections** peut s'énoncer de la façon suivante:

Étant donné:

- une requête Q,
- c collections de documents (ou serveurs ou sous-collections),

$\Rightarrow$ extraire respectivement λ_1, λ_2, ... λ_c documents de chaque collection, tel que le nombre total de documents pertinents retrouvés soit maximal, et que l'ordre de présentation soit optimal.

Les problèmes principaux posés par la fusion de collections sont:

- Les diverses collections n'étant pas toujours interrogées par le même système de recherche, quelles informations peuvent être utiles pour effectuer la fusion?
- Combien de documents faut-il extraire de chaque liste?
- Ce nombre est-il le même pour chaque serveur? pour chaque requête?
- Selon quel(s) critère(s) les documents sélectionnés dans la liste finale doivent-ils être ordonnés?
- Détecter et supprimer les duplicata et les quasi-duplicata. Ces duplicata peuvent être soit une même publication apparaissant dans plusieurs sites, soit différentes versions d'un même article (rapport technique, version d'une conférence, version apparue dans un journal).

II.2 Les stratégies existantes

De récents travaux ont suggéré différentes stratégies pour répondre à ces problèmes. Dans la littérature du domaine, nous rencontrons les cinq propositions suivantes, qui correspondent chacune à des situations différentes.

• **Chacun son tour** (ou "round-robin"). Cette première approche considère que chaque serveur renvoie un nombre presque identique de documents pertinents, et que leur classement dans les listes produites par les divers serveurs s'avère très similaire. Si ces deux hypothèses se vérifient, la solution consiste à prendre les documents l'un après l'autre dans chacune des listes et à les interclasser pour former la liste finale.

• **Chacun son tour pondéré.** Voorhees et al. ([95a], [95b], [96]) ont montré qu'il est possible d'améliorer cette première solution car les deux hypothèses sous-jacentes ne se vérifient pas en pratique. Ainsi, ils suggèrent de pondérer la sélection des documents en fonction du nombre de documents pertinents retrouvés par le serveur considéré lors des requêtes passées et de la similarité de la requête courante avec les requêtes passées (apprentissage automatique).

Ces deux premières approches ne tiennent compte que du rang des documents alors que bien souvent les systèmes de recherche renvoient, en plus du titre, une indication supplémentaire qui est le degré de similarité entre le document et la requête. Les trois stratégies suivantes utilisent ce degré de similarité (ou score).

• **Fusion par le score** (ou "raw-score merging"). Une troisième approche part de l'hypothèse que tous les serveurs d'informations appliquent la même stratégie de recherche (ou des stratégies très similaires), et qu'ainsi les valeurs de similarité qu'ils renvoient sont comparables ([Kwok 95], [Moffat 95]). Les documents sont alors triés et sélectionnés selon ces valeurs de similarité.

Cependant, Dumais ([94]) montre que les statistiques des documents varient fortement entre les diverses collections et qu'ainsi l'hypothèse précédente est invalide. De plus, lorsque les systèmes de recherche sont différents, les similarités ne peuvent plus être comparées entre elles.

• **Fusion par score normalisé.** Pour pallier à ce problème, une solution consiste à normaliser les degrés de similarité en les divisant par la similarité maximale obtenue par chaque système. Ces nouvelles valeurs sont désormais comparables et la fusion peut alors s'effectuer selon la stratégie "fusion par le score".

• **Fusion par score pondéré.** Pour améliorer cette stratégie, Callan et al. ([95]) proposent de pondérer les valeurs de similarité obtenues dans chaque collection en fonction de la qualité de la réponse de la sous-collection. Cette approche résout ainsi non seulement le problème de fusion, mais aussi celui de la sélection de collections.

Nous pourrions également imaginer une stratégie de fusion avec échange d'informations permettant aux serveurs de communiquer entre eux différentes statistiques sur les termes de la requête et simulant ainsi l'interrogation d'une collection unique. Une telle stratégie très coûteuse en communication peut s'évaluer en considérant les sous-collections comme une seule et grande collection.

La plupart de ces stratégies de fusion de collections sont des stratégies ad hoc et correspondent à des situations différentes (Tableau 1). Les approches peuvent se différencier selon le degré de coopération entre les systèmes de recherche. Ainsi, nous nommerons **stratégie de fusion isolée** si seul le rang est fourni, et **stratégie intégrée de fusion** si le degré de similarité est disponible. Lorsque les systèmes de recherche sont différents et que le rang des documents est la seule information disponible, aucune stratégie n'est suggérée. Or, cette situation particulièrement difficile se rencontre très couramment dans la pratique (par exemple sur les serveurs Web tels que AltaVista, Yahoo, InfoSeek, HotBot). Nous nous proposons dans cette thèse de déterminer un modèle de fusion de collections pouvant s'adapter à tous ces cas, et possédant un solide fondement théorique et une meilleure performance que les approches proposées jusqu'à présent.

	systèmes de recherche très similaires	systèmes de recherche différents
rang seul	Chacun son tour Chacun son tour pondéré	
rang, similarité	Fusion par score	Fusion par score normalisé Fusion par le score pondéré

Tableau 1 : Classification et exemples de stratégies de fusion de collections

III. LA FUSION DE DONNÉES

III.1 Présentation

La fusion de données et la fusion de collections sont des domaines relativement proches. La différence essentielle repose sur le fait qu'en fusion de données, les systèmes de recherche interrogent la même collection. En effet, si certains systèmes de recherche sont connus pour fournir, *en moyenne*, de meilleurs résultats pour un ensemble de requêtes, ils sont cependant moins performants pour certaines requêtes. Comme il est très difficile de prédire en fonction de la requête quel est le système le plus approprié ([Savoy 96]), il paraît alors judicieux de combiner plusieurs systèmes de recherche afin d'augmenter la qualité de la réponse finale du système.

Le problème de **fusion de données**, schématisé dans la Figure 4, s'énonce comme suit:

Étant donné

- une requête Q,
- c systèmes de recherche différents,

$\Rightarrow$ extraire de toutes les réponses proposées (listes de documents), les documents qui seront jugés les plus pertinents pour la requête, et tels que la performance finale soit significativement supérieure aux performances individuelles.

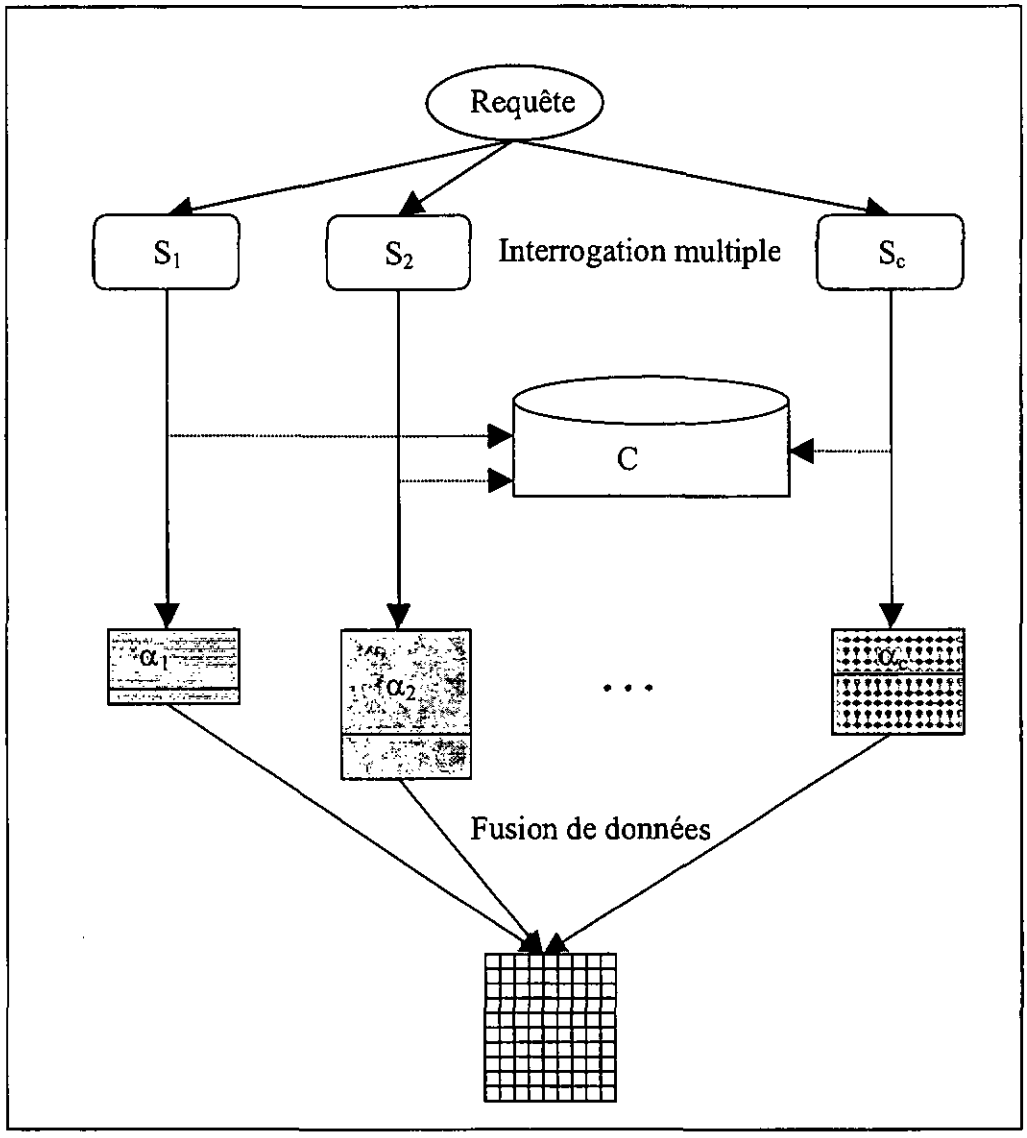

Figure 4: Principe de la fusion de données

La fusion de données se trouve confrontée à la contrainte que tous les systèmes travaillent sur la même collection, et qu'ils retrouvent souvent les mêmes documents. Il faut donc faire attention à ne pas comptabiliser deux fois le même document. De plus, Saracevic et Kantor ([Saracevic 88]) ont démontré qu'un document retrouvé par plusieurs systèmes de recherche possède une probabilité plus grande d'être pertinent qu'un document dépisté par un seul système. Nous retrouvons donc ici le problème d'une combinaison d'évidences entre différents experts (en voyant chaque système de repérage comme un expert). Ainsi, les problèmes particuliers à la fusion de données sont les suivants:

- comment combiner et exploiter au mieux les informations fournies par les systèmes sur le même document?
- quelle décision prendre par rapport à un document retrouvé par un seul système de recherche mais classé parmi les premiers documents?
- que faire si un document est retrouvé par plusieurs systèmes mais pas par tous?

III.2 Les stratégies existantes

Les travaux effectués en fusion de données se distinguent tout d'abord selon deux points de vue: celui de la requête et celui des systèmes.

Une première approche part de l'observation que des utilisateurs distincts, ayant cependant le même besoin d'information, ne formulent pas leur requête de façon similaire et obtiennent par là même des documents différents.

• **Formulation multiple de la requête.** Ainsi, afin d'augmenter la performance finale d'un système, certains auteurs suggèrent de soumettre plusieurs formulations d'un même besoin d'information c'est-à-dire d'encourager l'usager à écrire plusieurs requêtes, chacune évaluée par un système de recherche. Les résultats obtenus doivent ensuite être combinés. Certains auteurs ont évalué la combinaison de requêtes booléennes / modèle P-NORM ([Belkin 95]), ou booléennes / vectorielles ou probabilistes à l'intérieur d'un système de recherche basé sur les réseaux bayesiens ([Turtle 91]).

• **Classification des systèmes.** Une deuxième approche est de considérer que les propriétés intrinsèques des modèles de pondération permettent de retrouver des types de documents différents ([Lee 95], [Lee 97]), et qu'ainsi la fusion des résultats issus de divers systèmes de recherche est plus performante que les systèmes pris de manière individuelle. Fox et Shaw ([Shaw 95]) proposent quelques stratégies ad hoc pour combiner les résultats, alors que Lee [95] suggère une combinaison selon une classification effectuée à partir des caractéristiques propres d'indexation.

Dans la majorité des cas (les travaux de Turtle faisant exception), il faut ensuite choisir une méthode pour combiner les listes de documents retrouvés par les systèmes de recherche. La plupart des travaux suggèrent de se limiter exclusivement à la similarité comme critère pour la fusion. Des études semblent démontrer que le rang des documents n'apporte pas de résultats probants ([Shaw 95], [Belkin 95]). Une méthode largement répandue pour fusionner les c listes de documents est de définir, pour chaque document D_i, une nouvelle valeur de similarité selon la formule suivante:

$$\mathrm{Sim}(D_i) = \bigoplus_{j=1}^{c} \alpha_j \cdot \mathrm{Sim}_j(D_i)$$

où $\mathrm{Sim}_j(D_i)$ représente le degré de similarité (ou score) fourni par le système j pour le document D_i, α_j un facteur de pondération attribué à chaque système, et $\oplus$ l'opérateur de fusion.

Parmi les différents opérateurs $\oplus$ qui ont été proposés (le maximum, le minimum, la médiane, la moyenne, la somme), la somme est celui qui semble offrir les meilleurs résultats empiriques ([Shaw 95]).

Mais, si les systèmes de recherches possèdent des caractéristiques assez divergentes, les degrés de similarité n'ont alors pas de valeurs comparables. Dans ce cas, il est suggéré de procéder à une normalisation des scores.

Comme pour l'opération de fusion de collections, les approches proposées sont ad hoc, se justifiant uniquement par de bons résultats empiriques. Une telle démarche n'est guère satisfaisante à nos yeux, et nous voulons développer un modèle basé sur une méthodologie théorique plus robuste et rigoureuse.

PARTIE 2

LA REGRESSION LOGISTIQUE APPLIQUEE

A LA FUSION DE COLLECTIONS

I. PRÉSENTATION DES DONNÉES 31
I.1 Les collections-tests 31
I.2 Les variables utilisées 32
II. LE MODÈLE MATHÉMATIQUE PROPOSÉ 38
II.1 La régression logistique 38
II.2 Développement des modèles 43
II.3 Interprétation des modèles 47
III. ÉVALUATION ET COMPARAISON 49
III.1 Évaluation de plusieurs systèmes de dépistage de l'information 49
III.2 Fusion de collections avec systèmes de recherche similaires 50
III.3 Fusion de collections avec des systèmes de recherche différents 52
IV. CONCLUSION 55

I. PRÉSENTATION DES DONNÉES

I.1 Les collections-tests

Dans le cadre de ce travail de recherche, nous utilisons la collection-test "Wall Street Journal" (WSJ) issue du corpus de la conférence TREC. Cette collection de documents est composée de 74'520 articles (rédigés en anglais) durant les années 1990 à 1992, de 186 requêtes et des jugements de pertinence correspondants (voir en annexes ANNEXE 3 et ANNEXE 4 des exemples de requêtes et de documents).

Pour simuler l'existence de plusieurs collections de documents, nous avons découpé ce corpus en trois sous-collections, nommées WSJ90, WSJ91 et WSJ92, selon l'année de publication des articles (respectivement 1990, 1991 et 1992). Ces sous-collections ainsi obtenues n'ont aucun document en commun. Comme le montre le Tableau 2, elles sont de tailles très différentes, et ne possèdent pas le même nombre de documents pertinents par requête. D'autre part, certaines requêtes ne possèdent pas de jugements de pertinence pour l'une ou l'autre des années. Un tel phénomène est assez naturel car, avant le début de 1998, qui s'intéressait à Nagano, et en 1990, qui accordait autant d'importance à un gouverneur nommé Bill Clinton? Le numéro de ces requêtes ainsi que certaines statistiques liées aux différentes années sont présentées en annexe ANNEXE 5.

Collection	WSJ90	WSJ91	WSJ92	WSJ
Taille	73 Mb	146 Mb	35 Mb	254 Mb
Nb de documents	21 705	42 652	10 163	74 520
Nb de doc. pertinents	2 074	3 563	831	6 468
Nb de requêtes	170	171	140	186

Tableau 2: Statistiques sur les collections utilisées

Pour simuler une situation réelle, nous avons appliqué des systèmes de recherche différents sur chacune des trois collections. Dans la suite de nos travaux, nous avons opté pour les choix suivants:

1. le modèle probabiliste OKAPI ([Robertson 95]) sur WSJ90;
2. le modèle vectoriel LNU-LTC ([Buckley 96]) sur WSJ91;
3. le modèle vectoriel LNC-LTC sur WSJ92.

Ces couples collection/système sont respectivement référés par la suite "WSJ90-OKAPI", "WSJ91-LNU" et "WSJ92-LNC". Lors des évaluations, chacun des trois systèmes renvoie, pour chaque requête, une liste de 1000 documents triés par ordre décroissant de leur degré de similarité (les précisions moyennes à onze points fixes de rappel sont calculées par le logiciel TREC_EVAL). Nous pouvons donc nous appuyer sur le rang et/ou le degré de similarité des documents par rapport à la requête pour définir les variables à utiliser lors de la fusion de collections.

I.2 Les variables utilisées

Après l'interrogation des divers serveurs, les seules informations dont nous disposons pour effectuer la fusion de collections sont le degré de similarité (**Sim.**) et le **rang** des documents retrouvés. Or, comme les systèmes de recherche utilisés dans notre étude sont différents, les valeurs de similarité ne sont pas comparables entre elles. La fusion ne peut donc se faire directement à partir de ces valeurs. D'autre part, une de nos réflexions (2) consiste à fusionner les résultats en choisissant comme unique critère le rang des documents. Cependant, le rang, n'est peut-être pas une indication assez fiable pour être une bonne clé de fusion. Ainsi, et pour suivre une autre de nos réflexions (3) nous définissons les nouvelles variables suivantes à partir de la similarité et du rang.

• **SimMax.** Certains auteurs préconisent de recourir au degré de similarité (ou score), en le divisant par le score maximum (c'est-à-dire par le score du document placé au premier rang). Une nouvelle variable ordinale, appelée ici **SimMax**, est calculée pour chaque document D_i selon l'équation suivante:

$$\mathrm{SimMax}(D_i) = \frac{\mathrm{Sim}(D_i)}{\mathrm{Sim}(D_1)}$$

Les valeurs de SimMax, comprises entre 0 et 1, sont alors comparables entre les différents systèmes de dépistage. Cette variable, en plus de conserver l'ordre des documents retrouvés, respecte les variations du degré de similarité lorsque l'on considère des documents toujours plus loin dans la liste des documents extraits.

• **ln(rang).** Lorsque le système de dépistage ne fournit pas le degré de similarité mais uniquement le rang, une indication, à priori précieuse, est perdue: la différence des scores entre les documents est inconnue. En effet, la variable "rang" est codée 1, 2, 3, ... ,999, 1000, et accorde donc toujours un saut d'une unité entre chaque document retrouvé, que les documents considérés soient placés en tête de liste ou pas. Or, une analyse de l'évolution du score selon le rang des documents, (par exemple le modèle WSJ90-OKAPI en Figure 5), montre bien que la similarité décroît plus vite pour les premiers documents que pour les suivants. Cette baisse est encore plus lente pour les derniers documents retrouvés. Afin de tenir compte de cette évolution, nous suggérons de transformer la variable rang par son logarithme népérien (**ln(rang)**), les propriétés du logarithme conservant une proportionnalité entre la baisse de la similarité et la hausse de ln(rang) (cf. Figure 6).

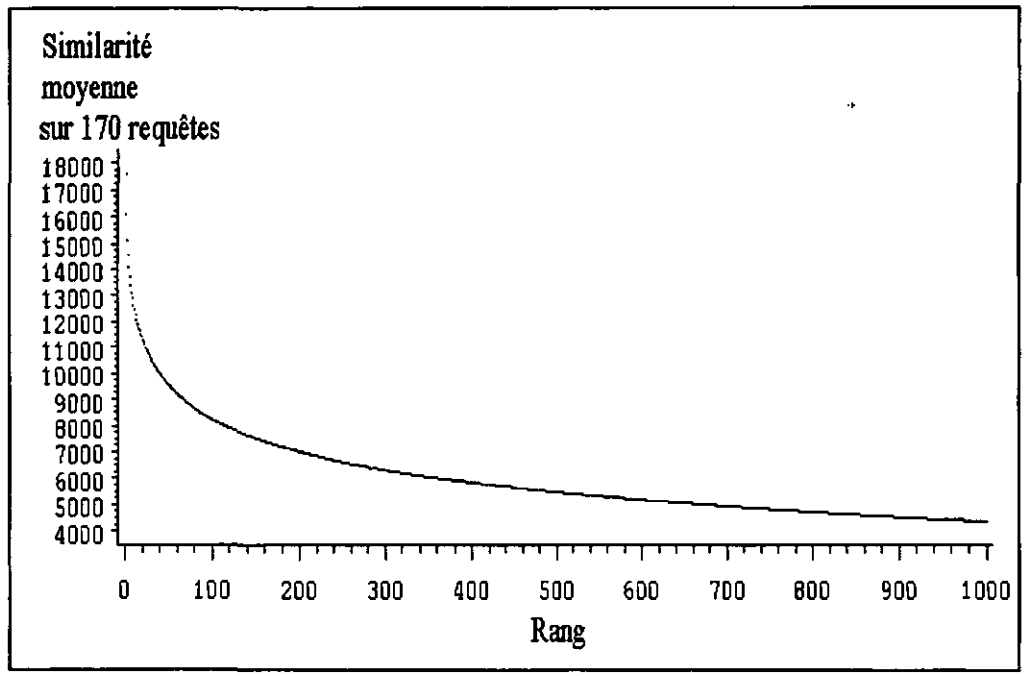

Figure 5: Relation empirique entre le degré de similarité moyen et le rang (WSJ90-OKAPI, 170 requêtes)

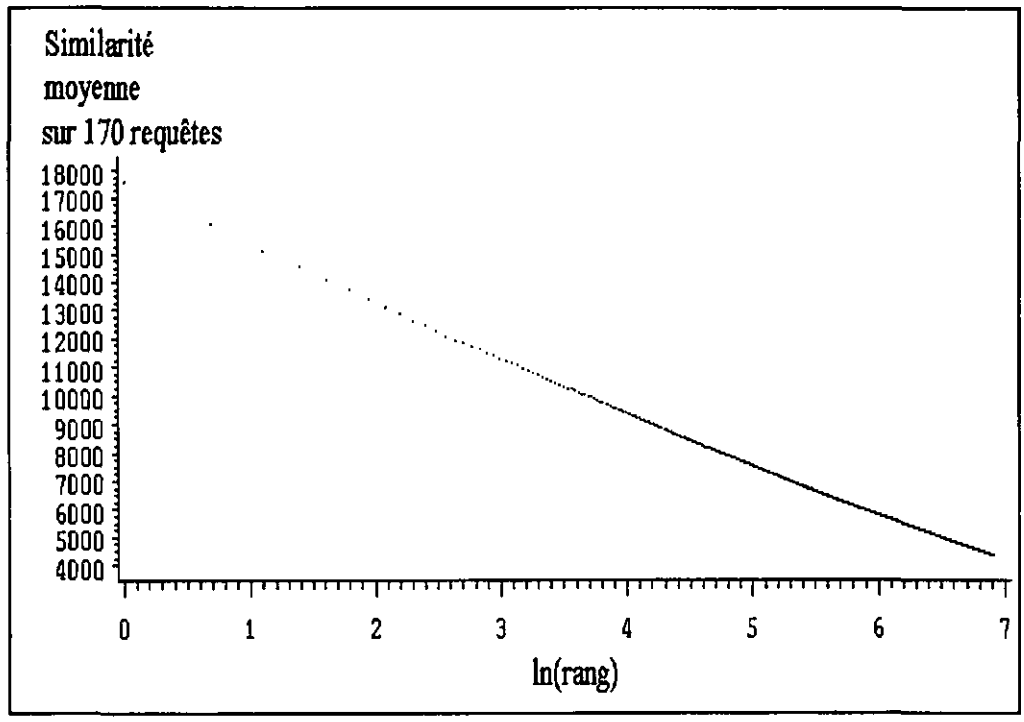

Figure 6: Relation empirique entre le degré de similarité moyen et ln(rang) (WSJ90-OKAPI, 170 requêtes)

• **SimDécomp.** Les degrés de similarité des documents retrouvés contiennent deux informations complémentaires :

– la qualité de réponse d'un document par rapport à une requête;

– l'évolution de la dégradation de cette qualité entre les documents successifs.

Si le degré de similarité est utilisé comme critère dans la fusion de collections, l'une de ces deux informations intrinsèques risque de voiler l'autre. Nous proposons donc de dissocier ces deux informations et de les utiliser comme deux critères distincts lors de la fusion de collections. La première partie de l'information modélisant la baisse du score d'un document à l'autre, est reprise par la variable ln(rang). La seconde partie de l'information, représentant la qualité de réponse du document, est extraite de la similarité par le biais d'une nouvelle variable. Cette nouvelle variable, nommée **SimDécomp**, indique alors si le degré de confiance que le système attribue au document retrouvé est, en moyenne, supérieur aux autres documents retrouvés au même rang dans d'autres requêtes. Dans ce but, le degré de similarité est centré et réduit selon l'équation ci-dessous :

$$\text{SimDécomp}_{r,q}(D_i) = \frac{\text{Sim}_{r,q}(D_i) - \hat{\mu}(\text{Sim}_r)}{\hat{\sigma}(\text{Sim}_r)}$$

$\text{Sim}_{r,q}(D_i)$ représente le degré de similarité du document D_i retourné au rang r pour la requête q.

$\hat{\mu}(\mathrm{Sim}_r)$ et $\hat{\sigma}(\mathrm{Sim}_r)$ sont les estimations de la moyenne et de l'écart type du degré de similarité des documents du même rang *r*, sur l'ensemble des requêtes passées pour un même système de recherche.

Alors que la similarité et le rang des documents sont des variables fortement corrélées, ln(rang) et SimDécomp sont désormais deux variables totalement distinctes, et leur coefficient de corrélation (cf. annexe ANNEXE 6) est considéré comme significativement égal à zéro (cf. Tableau 3).

Coefficient de corrélation linéaire de Pearson ρ			
		ln(rang)	
		ρ	Prob(ρ=0)
WSJ90-OKAPI	Sim	-0.63	0.0001
	SimDécomp	0	1
WSJ91-LNU	Sim	-0.60	0.0001
	SimDécomp	0	1
WSJ92-LNC	Sim	-0.71	0.0001
	SimDécomp	0	1

Tableau 3: Coefficients de corrélation de Pearson (ρ) entre les variables Sim, SimDécomp, et ln(rang)

Cette première analyse des données nous a permis de définir, à partir du rang et du degré de similarité, trois nouvelles variables (SimMax, ln(rang) et SimDécomp) pouvant servir de critères lors de la fusion de collections. Le Tableau 4 présente quelques statistiques descriptives de ces cinq variables obtenues à partir des résultats des trois systèmes de dépistage OKAPI, LNU et LNC appliqués respectivement sur les collections WSJ90, WSJ91 et WSJ92. Pour exemple, les histogrammes de la distribution de ces cinq variables pour WSJ90-OKAPI sont présentés en annexe ANNEXE 8. Ces statistiques démontrent bien que l'ordre de grandeur de la variable "Sim" est très différent d'une collection à l'autre, tandis que sa transformation par la variable "SimMax" permet d'obtenir des valeurs comparables entre les différents systèmes.

Le modèle WSJ90-OKAPI retrouve 85% des documents pertinents, WSJ91-LNU en retourne 79%, et le modèle WSJ92-LNU près de 90%. Mais si les systèmes retrouvent la plupart des documents pertinents, ces documents sont très souvent noyés dans la masse de tous les documents retournés. En effet, ces pourcentages ne tiennent pas compte du nombre de documents extraits (1000 par requêtes), ni de la position des documents pertinents (premier ou millième rang). Ainsi, pour WSJ90-OKAPI, seulement 1% des documents retoumés est pertinent (1765/170000), 1.6% pour WSJ91-LNU et 0.5% pour le modèle WSJ92-LNU.

Collection		WSJ90 170 requêtes	WSJ91 171 requêtes	WSJ92 140 requêtes
Modèle		OKAPI	LNU - LTC	LNC - LTC
# de documents pertinents		2074	3563	831
# de documents pertinents retrouvés		1765 (85%)	2811 (79%)	745 (90%)
Précision moyenne		24.13	22.88	22.52
Rang	Minimum	1		
	Maximum	1000		
	Moyenne	500.5		
	Écart-type	228.67		
Ln(Rang)	Minimum	0		
	Maximum	6.91		
	Moyenne	5.91		
	Écart-type	0.98		
Sim	Minimum	1.152	0.002	0.011
	Maximum	47.176	0.0065	0.055
	Moyenne	6.013	0.027	0.435
	Écart-type	2.741	0.0026	0.027
SimMax	Minimum	0.09	0.09	0.06
	Maximum	1	1	1
	Moyenne	0.36	0.40	0.31
	Écart-type	0.14	0.14	0.13
SimDécomp	Minimum	-2.27	-1.78	-2.01
	Maximum	5.59	5.10	4.11
	Moyenne	0	0	0
	Écart-type	1.00	1.00	1.00

Tableau 4: Statistiques descriptives des variables

Une analyse des distributions des documents pertinents et de ceux sans intérêt (Tableau 5 pour le système WSJ90-OKAPI), démontre qu'ils forment deux distributions distinctes avec des comportements différents. Ainsi, les documents pertinents se situent principalement parmi les premiers rangs. En effet, l'intervalle de confiance à 95% (cf. annexe ANNEXE 7) de la moyenne des rangs est [138.59;158.66] pour les documents pertinents contre [502.82;505.56] pour les autres. Le système de dépistage répond somme toute à nos attentes et place en tête de liste plutôt des documents pertinents. Les autres variables présentent également une différence de comportement entre les documents pertinents et non pertinents, les intervalles de confiance des deux ensembles étant toujours nettement éloignés.

WSJ90- OKAPI 170 requêtes		Documents pertinents	Documents sans intérêt
Rang	Minimum	1	
	Maximum	1000	
	Moyenne	148.62	504.19
	Écart-type	215.02	287.07
	IC 95%	[138.59 ; 158.66]	[502.82 ; 505.56]
ln(Rang)	Minimum	0	
	Maximum	6.91	
	Moyenne	3.84	5.93
	Écart-type	1.74	0.95
	IC 95%	[3.76 ; 3.92]	[5.93; 5.94]
Sim	Minimum	2.1	1.15
	Maximum	47.18	44.51
	Moyenne	11.37	5.96
	Écart-type	6.14	2.62
	IC 95%	[11.08 ; 11.66]	[5.94 ; 5.99]
SimMax	Minimum	0.15	0.09
	Maximum	1	1
	Moyenne	0.60	0.36
	Écart-type	0.19	0.13
	IC 95%	[0.59 ; 0.60]	[0.35 ; 0.36]
SimDécomp	Minimum	-2.03	-2.27
	Maximum	5.59	5.52
	Moyenne	0.40	-0.004
	Écart-type	1.12	0.99
	IC 95%	[0.35 ; 0.45]	[-0.01 ; 0.0006]

Tableau 5: Statistiques sur les documents extraits (WSJ90-OKAPI, 186 requêtes)

Les trois variables SimMax, SimDécomp, et ln(rang), ainsi définies, comportent donc des propriétés plus intéressantes que les variables isolées du rang et du degré de similarité:

- valeurs comparables entre les systèmes pour la variable SimMax;
- évolution non linéaire de la variable ln(rang), reflétant mieux les variations du degré de similarité;

- absence de corrélation entre SimDécomp et ln(rang), permettant de dégager la qualité de réponse d'un document par rapport à la requête.

Ces variables vont servir de critères pour la fusion dans notre modèle basé sur la régression logistique.

II. LE MODÈLE MATHÉMATIQUE PROPOSÉ

II.1 La régression logistique

Le problème de la fusion de collection peut se résumer ainsi:

- décider, en fonction de certains critères tels que le rang et/ou le degré de similarité, si un document d'une liste est pertinent ou non;
- ordonner les articles retenus du plus pertinent au moins pertinent.

Pour résoudre ce problème, nous suggérons d'utiliser la **régression logistique** ([Cox 1989], [Hosmer 1989]). Si cet outil statistique a déjà été utilisé en recherche d'information ([Bookstein 92] [Cooper 92], [Gey 93], [Gey 94], [Fuhr 94]), il ne l'a encore jamais été en fusion de collections ou de données. Voyons dans un premier temps les aspects théoriques essentiels de la régression logistique, puis dans un deuxième temps la justification d'un tel modèle par rapport à nos données.

- **Théorie de la régression logistique**

La régression logistique est une méthodologie statistique employée habituellement pour prédire la réalisation d'une variable binaire en fonction d'une ou plusieurs variables explicatives. Pour plus de clarté, l'explication suivante ne porte que sur le cas univarié, la méthode pouvant se généraliser au cas multivarié.

Comme le présente la Figure 7, il n'est guère approprié d'utiliser une régression linéaire classique pour estimer une variable binaire. En effet, le système devrait répondre "1" ou "0" selon la valeur calculée. De plus, un des désavantages du modèle linéaire est que l'estimation par une droite conduit dans certains cas à des valeurs estimées supérieures à 1, ou inférieures à 0, ce qui est clairement non réaliste et logiquement impossible dans le cas d'une variable binaire.

Une des solutions est de déformer le modèle linéaire en incurvant la droite lorsqu'elle se rapproche des valeurs 0 et 1 afin qu'elle converge désormais vers ces valeurs sans jamais les atteindre (Figure 7). Cette nouvelle courbe est dite "courbe en S". La régression logistique va alors prédire non pas directement la variable binaire mais sa probabilité de réalisation.

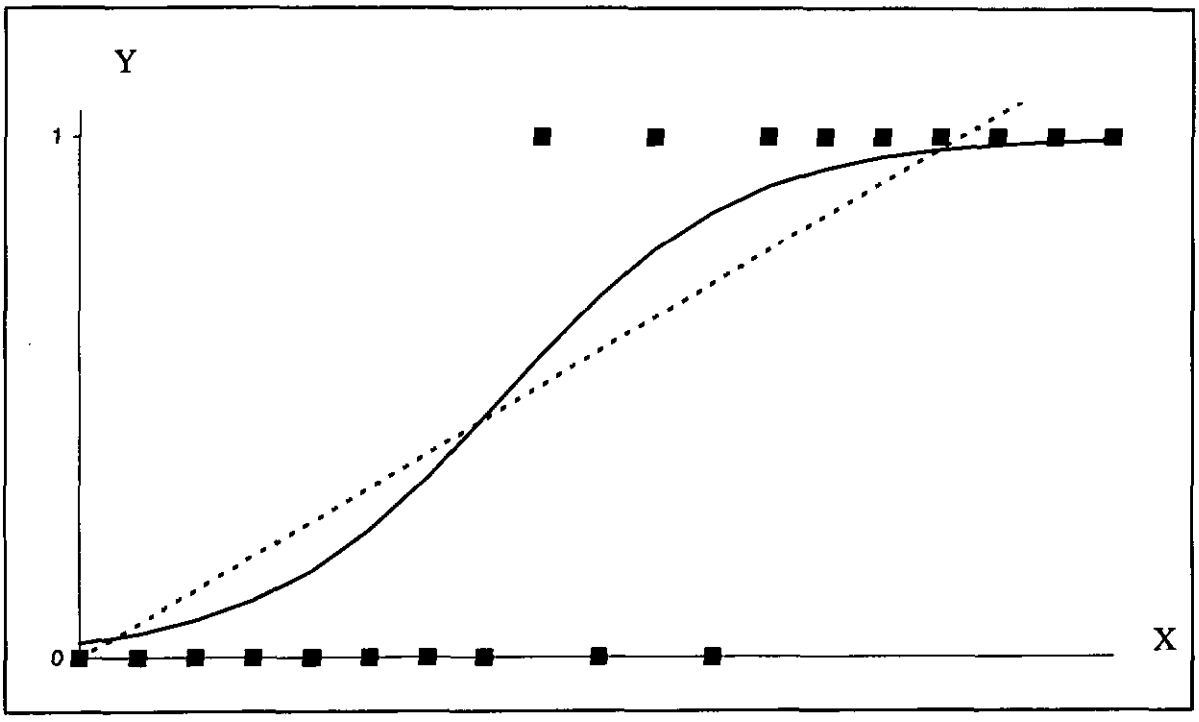

Figure 7 : Nuage de points et estimation linéaire et logistique d'une variable binaire Y en fonction d'une variable explicative X

Si plusieurs fonctions ont été proposées pour ajuster cette courbe en S, une solution flexible et dont les propriétés mathématiques sont intéressantes, est de choisir une fonction de distribution logistique. Ainsi, la probabilité de réalisation de la variable Y en fonction de la valeur de la variable explicative X s'exprime à partir de l'équation suivante:

$$\text{Prob}[y = 1|x] = \pi(x_i) = \frac{e^{\alpha+\beta \cdot x_i}}{1+e^{\alpha+\beta \cdot x_i}}$$

Une des propriétés centrale de l'utilisation de la fonction logistique est que le logarithme du rapport entre $\pi(x_i)$ et $1-\pi(x_i)$, noté $g(x_i)$, est linéaire en X :

$$g(x_i) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \alpha + \beta \cdot x_i$$

Ainsi, à partir de cette relation, les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ des coefficients α et β sont estimés par la méthode du maximum de vraisemblance. Cette méthode consiste à maximiser le logarithme de la vraisemblance, c'est-à-dire à maximiser la probabilité de réalisation de l'échantillon.

La vraisemblance est définie comme suit :

$$L = \prod_{i=1}^{n} \left[\pi(x_i)^{y_i} \cdot (1 - \pi(x_i))^{y_i} \right]$$

Comme il est mathématiquement plus facile de travailler sur le logarithme de la vraisemblance, la méthode du maximum de vraisemblance cherche alors à maximiser:

$$\ln(L) = \sum_{i=1}^{n} \left[yi \cdot \ln(\pi(xi)) + (1 - yi) \cdot \ln(\pi(xi)) \right]$$

en résolvant les équations :

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n} (y_i - \pi(x_i)) = 0$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} x_i \cdot (y_i - \pi(x_i)) = 0$$

Ces deux équations n'étant pas linéaires en α et β, leur résolution fait appel à des méthodes numériques itératives. Le logiciel SAS, utilisé dans nos expériences, se sert de l'algorithme particulier IRLS (Iteratively Reweighted Least Squares) (cf. [SAS v6]).

Afin de vérifier la validité et la pertinence du modèle, deux tests sont effectués, l'un sur le modèle, l'autre sur les coefficients estimés,.

Le premier de ces tests permet de juger si le modèle est significatif, et compare pour cela le modèle estimé à un modèle sans variable explicative. Un test du χ^2 permet d'évaluer si ces deux modèles sont significativement différents. Les hypothèses du test sont donc équivalentes à:

H_0: tous les coefficients du modèle logistique sont significativement égaux à zéro

H_1: le modèle logistique possède au moins un des coefficients significativement différent de zéro (peut-être plusieurs, voire tous).

Pour s'assurer que chacune des valeurs estimées des coefficients du modèle est significativement différente de zéro, un second test basé sur la statistique de Wald est effectué sur chacun des paramètres. Sous l'hypothèse nulle H_o (la valeur estimée du paramètre est égale à zéro), cette statistique, calculée comme le paramètre estimé divisé par son écart-type estimé, suit une loi du χ^2 à un degré de liberté. L'hypothèse nulle est alors rejetée ou acceptée selon les résultats du test du χ^2.

- **Utilisation de la régression logistique**

La régression logistique semble très bien adaptée au cas de la fusion de collections car il s'agit de prédire la probabilité de réalisation d'un événement binaire, soit la pertinence, en fonction d'une ou plusieurs variables explicatives. De plus, comme la régression prédit la probabilité de pertinence des documents, cette valeur estimée s'utilisera comme clé de tri pour ordonner les documents retenus.

Cependant, avant de conclure trop hâtivement à l'adéquation de la régression logistique au problème de fusion de collections, il est tout d'abord important de regarder si les données ne sont pas en contradiction avec l'utilisation d'un modèle logistique.

La régression logistique estime la probabilité d'une variable binaire selon une courbe logistique en S. En prenant l'exemple du modèle WSJ90-OKAPI (résultats similaires pour WSJ91-LNU et WSJ92-LTN), la Figure 8 présente la probabilité de pertinence des documents retrouvés en fonction de la variable ln(rang). Cette probabilité empirique est calculée pour chaque valeur de ln(rang), comme le nombre de documents pertinents retrouvés pour cette valeur, divisé par le nombre de documents retrouvés pour cette valeur. Au vue de ce nuage de points, l'estimation de la probabilité de pertinence par une fonction logistique ne paraît pas dénuée de sens. En effet, la baisse de la distribution est tout d'abord lente puis s'accélère en son milieu pour finalement tendre vers la valeur nulle. Cette première approche de nos données nous permet de ne pas rejeter l'utilisation de la régression logistique comme modèle de fusion.

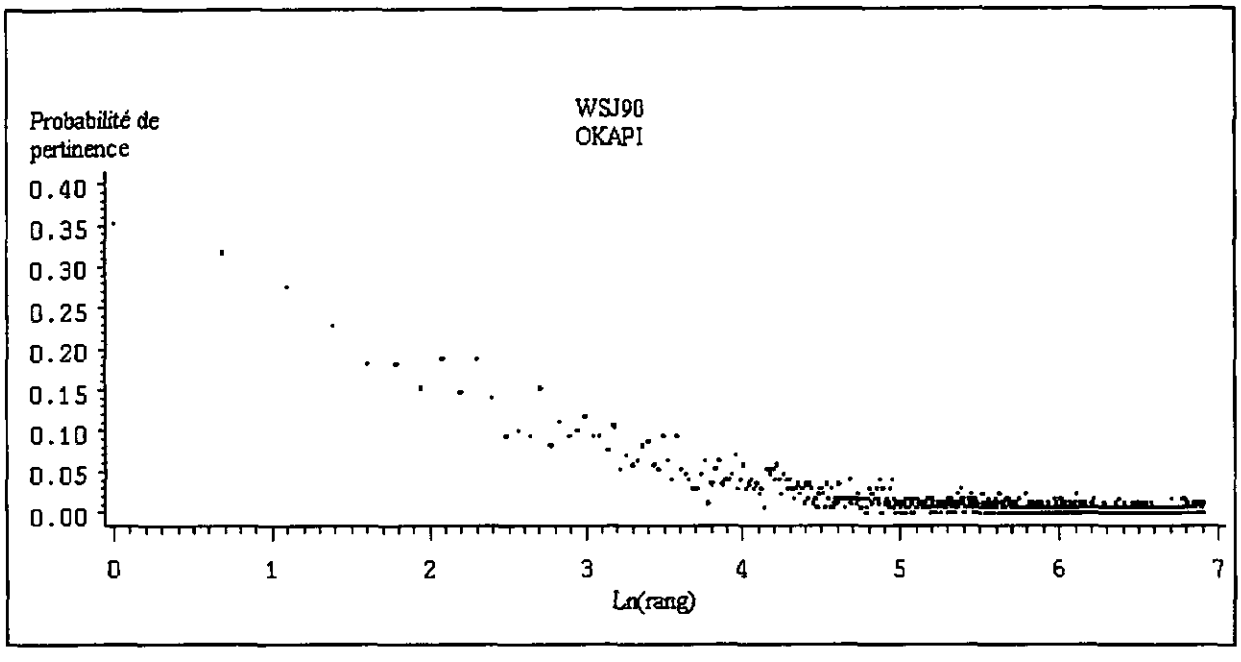

Figure 8: Relation empirique entre la variable ln(rang) et la probabilité de pertinence (WSJ90-OKAPI, 170 requêtes)

Une seconde approche des données renforce l'intuition d'utiliser un modèle logistique. Reprenons pour cela l'une des propriétés de la régression logistique: le modèle admet que la fonction g(x), c'est-à-dire le logarithme du rapport de la probabilité de pertinence sur la probabilité de non pertinence, est une fonction linéaire. Or, cette linéarité se retrouve sur nos données. En effet, en gardant le système WSJ90-OKAPI, la variable ln(rang), possède une forte corrélation linéaire avec g(x) (dans ce cas, le coefficient de corrélation de Pearson est égal à -0.86, et est considéré comme significatif au seuil de 5%).

Enfin, le recours à la régression logistique pour résoudre le problème de fusion de collections rejoint plusieurs de nos objectifs:

1. *Développer un nouveau modèle théorique justifiant clairement la manière de combiner plusieurs évidences.*
   Notre modèle est issu d'une méthodologie statistique reconnue;

   1.3 *Ordonner clairement les documents retrouvés.*

   L'ordonnancement se fait désormais selon les probabilités de pertinence qui sont des valeurs comparables entre les différents systèmes de dépistage de l'information;

   1.5 *En fusion de collections, savoir combien de documents il faut sélectionner dans chaque liste.*

   Le nombre de documents n'est plus un problème car les documents sont sélectionnés selon leur probabilité, et donc indépendamment de leur collection d'origine.

Tant pour des raisons théoriques que pratiques, la méthode de la régression logistique est tout à fait appropriée pour la fusion de collections. Notre stratégie de fusion correspond donc aux étapes suivantes:

- déterminer les variables explicatives du modèle;
- appliquer une régression logistique pour chacune des différentes listes fournies par les systèmes;
- ordonner tous les documents selon leur probabilité de pertinence prédite par la régression;
- sélectionner les documents qui possèdent les probabilités les plus fortes.

En suivant cette stratégie, nous avons conçu trois modèles logistiques de fusion de collections qui diffèrent par le choix des variables explicatives retenues pour effectuer les régressions.

II.2 Développement des modèles

L'un des objectifs de cette recherche (**1.1**) est de proposer une méthode de fusion de collections qui s'adapte aux diverses situations de la réalité. A partir de la stratégie de fusion définie précédemment, nous développons plusieurs modèles basés sur la régression logistique. Dans un premier temps, nous nous plaçons dans le cadre d'une stratégie de fusion isolée (Modèle 1) puis, dans une stratégie intégrée de fusion (Modèle 2 et 3).

- **Modèle 1**

Si le rang des documents retrouvés par chaque système s'avère être la seule information disponible, nous suggérons d'utiliser non pas directement la valeur du rang, mais son logarithme népérien. Ainsi, le premier modèle, nommé Modèle 1, est défini comme suit:

$$\text{Prob}\left[\text{Pert}(D_i)\middle|\ln(\text{rang}(D_i))\right] = \frac{e^{\alpha+\beta\cdot\ln(\text{rang}(D_i))}}{1+e^{\alpha+\beta\cdot\ln(\text{rang}(D_i))}}$$

Les estimateurs des coefficients α et β ainsi que les statistiques de chaque régression sont calculés par le logiciel SAS et sont présentés dans le Tableau 6.

		valeur estimée, IC à 95%	écart type estimé	test de Wald sur chaque paramètre	test χ^2 sur le modèle
WSJ90-OKAPI	$\hat{\alpha}$	0.322 [0.199 ; 0.445]	0.063	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.949 [-0.976 ; -0.922]	0.014	p=0.0001	
WSJ91-LNU	$\hat{\alpha}$	0.634 [0.526 ; 0.742]	0.055	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.902 [-0.926 ; -0.878]	0.012	p=0.0001	
WSJ92-LNC	$\hat{\alpha}$	-0.310 [-0.475 ; -0.145]	0.084	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.976 [-1.015 ; -0.937]	0.020	p=0.0001	

Tableau 6: Valeurs des coefficients de la régression logistique pour le Modèle 1

Afin de vérifier la validité des résultats de la régression, le logiciel SAS applique le test du χ^2 sur l'ensemble du modèle et le test basé sur la statistique Wald pour chaque valeur estimée des paramètres ([SAS v6]). Le Tableau 6 présente dans les deux dernières colonnes les probabilités p d'acceptation de l'hypothèse nulle de ces tests. Dans tous les cas, cette hypothèse nulle a une infime chance d'être vraie (p=0.0001), si faible que nous devons la rejeter et donc accepter les hypothèses alternatives impliquant que tous les modèles sont significatifs et que la valeur de chaque coefficient est significativement différente de zéro.

Plus précisément, pour WSJ91-LNU, la valeur estimée des coefficients est $\hat{\alpha}$=0.634 et $\hat{\beta}$=-0.902. Les valeurs de ces deux paramètres sont jugées toutes deux significativement différentes de zéro (colonne "test de Wald"). L'intervalle de confiance à 95% pour la valeur de $\hat{\alpha}$ est égal à $0.634 \pm 1.96 \cdot 0.055$ ([0.526 ; 0.742]), et celui pour la valeur $\hat{\beta}$ est $-0.902 \pm 1.96 \cdot 0.012$, c'est-à-dire [-0.926 ; -0.878].

WSJ90-OKAPI		WSJ91-LNU		WSJ92-LNC	
rang	$\hat{\pi}(x_i)$	rang	$\hat{\pi}(x_i)$	rang	$\hat{\pi}(x_i)$
1	0.580	1	0.653	1	0.423
2	0.417	2	0.502	2	0.272
3	0.327	3	0.412	3	0.201
4	0.270	4	0.351	4	0.159
5	0.230	5	0.306	5	0.132
6	0.201	6	0.273	6	0.113
...		...		...	

Fusion			
rang final	$\hat{\pi}(x_i)$	Issus de	rang initial
1	0.653	WSJ91-LNU	1
2	0.580	WSJ90-OKAPI	1
3	0.502	WSJ91-LNU	2
4	0.423	WSJ92-LNC	1
5	0.417	WSJ90-OKAPI	2
6	0.412	WSJ91-LNU	3
7	0.351	WSJ91-LNU	4
8	0.327	WSJ90-OKAPI	3
9	0.306	WSJ91-LNU	5
10	0.273	WSJ91-LNU	6
...	...		

Tableau 7: Exemple de fusion basée sur le Modèle 1

Les valeurs estimées des coefficients permettent de calculer les probabilités de pertinence de tous les documents retrouvés dans chaque collection. La fusion s'opère alors en choisissant les documents ayant la plus forte probabilité (voir un exemple dans le Tableau 7). Parmi les dix premiers documents retenus lors de la fusion, six proviennent de la collection WSJ91 avec le système de recherche LNU, 3 de la collection WSJ90 (OKAPI) et un seul est issu de WSJ92 (LNC). Ces chiffres reflètent assez bien la proportion de documents pertinents que comporte chaque collection. En effet, en reprenant les statistiques des collections, WSJ91 contient 55% de l'ensemble des documents pertinents, la collection WSJ90 en comprend 32%, et WSJ92 seulement 12%.

- **Modèle 2**

Si le degré de similarité des documents extraits est également disponible, il s'avère possible de recourir à la variable ordinale SimMax comme unique variable explicative de notre modèle de régression logistique. Les paramètres de ce Modèle 2 sont présentés dans le Tableau 8. Ce modèle est, dans son ensemble, significatif (colonne "test χ^2") et les valeurs des coefficients sont toutes significativement différentes de zéro.

		valeur estimée, IC à 95%	écart type estimé	test de Wald sur chaque paramètre	test χ^2 sur le modèle
WSJ90-OKAPI	$\hat{\alpha}$	-8.231 [-8.380 ; -8.082]	0.076	p=0.0001	p=0.0001
	SimMax $\hat{\beta}$	8.024 [7.777 ; 8.271]	0.126	p=0.0001	
WSJ91-LNU	$\hat{\alpha}$	-7.813 [-7.944 ; -7.682]	0.067	p=0.0001	p=0.0001
	SimMax $\hat{\beta}$	7.675 [7.455 ; 7.895]	0.112	p=0.0001	
WSJ92-LNC	$\hat{\alpha}$	-8.273 [-8.465 ; -8.081]	0.098	p=0.0001	p=0.0001
	SimMax $\hat{\beta}$	7.406 [7.083 ; 7.729]	0.165	p=0.0001	

Tableau 8: Valeurs des coefficients de la régression logistique pour le Modèle 2

- **Modèle 3**

Dans le but d'exploiter au mieux toute l'information contenue dans le degré de similarité des documents extraits en utilisant des variables indépendantes, notre dernière stratégie repose sur les deux variables explicatives ln(rang) et SimDécomp. Le Modèle 3 est construit à partir l'équation suivante:

$$\text{Prob}\left[\text{Pert(Di)} \middle| \ln(\text{rang(Di)}), \text{SimDécomp(Di)}\right] = \frac{e^{\alpha+\beta_1 \cdot \ln(\text{rang(Di)})+\beta_2 \cdot \text{SimDécomp(Di)}}}{1+e^{\alpha+\beta_1 \cdot \ln(\text{rang(Di)})+\beta_2 \cdot \text{SimDécomp(Di)}}}$$

De même que pour les deux premiers modèles, la statistique de Wald indique que la valeur des coefficients peut être considérée comme significativement différente de zéro et que le modèle est, dans son ensemble, significatif.

		valeur estimée, IC à 95%	écart type estimé	Wald test sur chaque paramètre	test χ^2 sur le modèle
WSJ90-OKAPI	$\hat{\alpha}$	0.297 [0.174 ; 0.420]	0.063	p=0.0001	p=0.0001
	Ln(rang) $\hat{\beta}_1$	-0.959 [-0.986 ; -0.932]	0.014	p=0.0001	
	SimDécomp $\hat{\beta}_2$	0.380 [0.337 ; 0.423]	0.022	p=0.0001	
WSJ91-LNU	$\hat{\alpha}$	0.614 [0.504 ; 0.724]	0.056	p=0.0001	p=0.0001
	Ln(rang) $\hat{\beta}_1$	-0.916 [-0.940 ; -0.892]	0.012	p=0.0001	
	SimDécomp $\hat{\beta}_2$	0.428 [0.393 ; 0.463]	0.018	p=0.0001	
WSJ92-LNC	$\hat{\alpha}$	-0.376 [-0.545 ; -0.207]	0.086	p=0.0001	p=0.0001
	Ln(rang) $\hat{\beta}_1$	-0.990 [-1.029 ; -0.951]	0.020	p=0.0001	
	SimDécomp $\hat{\beta}_2$	0.507 [0.440 ; 0.574]	0.034	p=0.0001	

Tableau 9: Valeurs des coefficients de la régression logistique pour le Modèle 3

Nos trois modèles de fusion, estimés par le logiciel SAS, sont significatifs et leur coefficients sont tous significativement différents de zéro. A ce stade de la recherche, il est déjà possible de tirer quelques interprétations intéressantes des résultats de nos différents modèles de régression logistique.

II.3 Interprétation des modèles

Une première interprétation des modèles est d'analyser le *ratio* de la probabilité de pertinence d'un document sur la probabilité de non pertinence. Dans le cas univarié, ce ratio noté $r(x_i)$ est calculé comme suit:

$$r(x_i) = \frac{\pi(x_i)}{1-\pi(x_i)} = e^{\alpha+\beta \cdot x_i}$$

Si $r(x_i) = 1$, alors le document a autant de chance d'être pertinent que sans intérêt pour l'usager (du moins dans le jugement posé par la machine). Lorsque $r(x_i) = 0.5$, le document a deux fois plus de chance d'être sans utilité que pertinent.

Rang i	$\hat{\pi}(x_i)$	$\hat{r}(x_i)$
1	0,653	1.885
2	0,502	1.009
3	0,412	0.700
4	0,351	0.540
5	0,306	0.442
6	0,273	0.375
7	0,246	0.326
8	0,224	0.289
9	0,206	0.260
10	0,191	0.236
...		
100	0,029	0.030
...		
200	0,016	0.016

Tableau 10: Ratio $r(x_i)$ obtenu par le Modèle I (WSJ91-LNU, 171 requêtes)

Dans le cadre du Modèle 1, le plus simple à interpréter car nous pouvons nous ramener directement au rang des documents, le Tableau 10 reprend quelques valeurs pour WSJ91-LNU. Dès le troisième rang, la probabilité que le document soit pertinent devient plus petite que celle d'être non pertinent ($r(x_i)<1$), et dès le quatrième rang, le document possède deux fois moins de chance d'être pertinent que non pertinent ($r(x_i) \approx 0.5$). Ainsi, dans une recherche où la précision s'avère très importante, le système ne devrait retourner, dans ce cas précis, que les deux premiers documents. Cette limite varie bien entendue selon la collection et le système de recherche utilisé. De plus, entre le deuxième et le troisième rang, la baisse du ratio s'élève à 60%, et une baisse similaire se retrouve entre les rangs 100 et 200. Ce mouvement correspond à l'effet de la transformation de la variable rang par son logarithme népérien et du choix d'un modèle logistique (courbe en S).

Certaines valeurs particulières du ratio $r(x_i)$ sont plus intéressantes à analyser, par exemple 1.5, 1 et 0.5. Le Tableau 11 nous montre ainsi que pour les deux autres collections, la chute de la probabilité de pertinence est encore plus rapide puisque pour WSJ90-OKAPI, le deuxième rang a déjà plus de chance d'être sans intérêt que pertinent, et pour WSJ92-LNC, ce phénomène apparaît aux alentours du rang 1. Ces résultats peuvent sembler surprenants, mais ils sont dus au faible nombre de documents pertinents parmi l'ensemble des documents retrouvés par les systèmes. N'ayant comparativement que peu de données sur les documents pertinents, les régressions ne peuvent accorder un poids plus grand aux probabilités de pertinence au delà de ces premiers rangs.

		WSJ90 OKAPI		WSJ91 LNU		WSJ92 LNC	
$r(x_i)$	$p(x_i)$	Ln(rang)	rang	Ln(rang)	rang	Ln(rang)	rang
1.5	0.6	-0,088	0,92	0.253	1.29	-0,733	0,48
1	0.5	0,339	1,40	0.703	2.02	-0,318	0,73
0.5	0.33	1,069	2,91	1.472	4.36	0,393	1,48

Tableau 11 : Rangs des documents correspondant à des valeurs particulières du ratio $r(x_i)$ dans le Modèle 1

Une seconde interprétation intéressante du modèle logistique se base sur la valeur du coefficient β. Dans une régression linéaire classique, le coefficient de la pente indique la variation de la variable expliquée lorsque la variable explicative augmente d'une unité. En régression logistique, le coefficient β est égal à:

$$\beta = g(x+1) - g(x) = \ln\left(\psi(1)\right)$$

où, pour rappel :

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta \cdot x$$

et la fonction $\psi(c)$ représente ainsi l'effet d'un changement de c unités de la variable explicative sur la variable expliquée, selon la formule suivante:

$$\psi(c) = e^{g(x+c)} - e^{g(x)} = \frac{\frac{\pi(x+c)}{1-\pi(x+c)}}{\frac{\pi(x)}{1-\pi(x)}} = e^{c\beta}$$

Ainsi, pour le Modèle 1 (collection WSJ90 et système OKAPI), entre les rangs 1 et 10, la probabilité d'être pertinent varie de $\psi(\ln(10) - \ln(1)) = 0.112$, ou, en d'autres termes, le dixième document a 1/0.112=8.9 fois moins de chance d'être pertinent que le premier. Pour WSJ91-LNU, cette baisse est de 7.97 et pour WSJ92-LNC de 9.46. De part la transformation logarithmique de la variable rang, cette baisse n'est donc pas linéaire, et une même augmentation de 9 rangs mais cette fois entre les rangs 101 et 110 entraîne pour WSJ90-OKAPI, que le $110^{ème}$ document a 1.08 fois moins de chance d'être pertinent que le $101^{ème}$ (1.08 pour WSJ91-LNU, et 1.09 pour WSJ92-LNC).

Ces premières interprétations du Modèle 1 montrent que la probabilité de pertinence décroît très rapidement pour les premiers documents retrouvés, puis se stabilise pour les derniers documents dépistés. Ce phénomène s'explique par la répartition des documents pertinents dans les listes extraites par les systèmes de dépistage de l'information.

III. ÉVALUATION ET COMPARAISON

Dans ce dernier chapitre, nous présentons tout d'abord des évaluations de différents systèmes de recherche appliqués sur les trois sous-collections retenues. Nous évaluons ensuite différentes stratégies de fusion se basant sur des systèmes de recherche identiques. Enfin, nous évaluons et comparons nos modèles logistiques avec d'autres stratégies de fusion de collections dans le cas où des systèmes de recherche différents sont utilisés.

III.1 Évaluation de plusieurs systèmes de dépistage de l'information

Le Tableau 12 présente l'évaluation du modèle probabiliste OKAPI, ainsi que différentes variantes du modèle vectoriel avec les sous-collections WSJ90, WSJ91, WSJ92 et sur la collection entière du Wall Street Journal (WSJ). Ces variantes se différencient selon la pondération utilisée dans l'indexation des documents et des requêtes. Comme l'indexation peut être différente pour les articles et les requêtes, le système SMART propose un premier triplet de lettres représentant la formulation d'indexation appliquée aux documents et un second triplet, pour la pondération de la requête. Les formules exactes de ces indexations se trouvent en annexe ANNEXE 2.

	Précision moyenne (% variation)			
collection modèle	WSJ90 170 requêtes 2'074 doc. pert.	WSJ91 171 requêtes 3'563 doc. pert.	WSJ92 140 requêtes 831 doc. pert.	WSJ 186 requêtes 6'468 doc. pert.
OKAPI - NPN	24.13	22.74	28.16	20.32
LNU - LTC	24.31 (+0.7 %)	22.88 (+0.6 %)	27.95 (-0.7 %)	20.27 (-0.2 %)
LTN - NTC	23.89 (-1.0 %)	22.49 (-1.1 %)	24.13 (-14.3 %)	19.16 (-5.7 %)
HTN - BNN	23.21 (-3.8 %)	21.46 (-5.6 %)	26.00 (-7.7 %)	19.09 (-6.1 %)
ATN - NTC	22.27 (-7.7 %)	20.50 (-9.9 %)	24.11 (-14.4 %)	18.85 (-7.2 %)
ANN - NTC	21.03 (-12.8 %)	18.81 (-17.3 %)	22.20 (-21.2 %)	16.78 (-17.4 %)
LNC - LTC	18.39 (-23.8 %)	17.72 (-22.1 %)	22.52 (-20.0 %)	15.76 (-22.4 %)
LTC - LTC	17.55 (-27.3 %)	17.98 (-20.9 %)	22.94 (-18.5 %)	15.06 (-25.9 %)
ANC - LTC	14.12 (-41.5 %)	14.02 (-38.3 %)	19.06 (-32.3 %)	11.94 (-41.2 %)
ANN - ANN	14.70 (-39.1 %)	11.43 (-49.7 %)	15.53 (-44.9 %)	10.76 (-47.0 %)
LNC - LNC	11.89 (-50.7 %)	10.69 (-53.0 %)	14.23 (-49.5 %)	9.69 (-52.3 %)
BNN - BNN	9.37 (-61.2 %)	5.84 (-74.3 %)	7.82 (-72.2 %)	4.85 (-76.1 %)
NNN - NNN	6.36 (-73.6 %)	5.80 (-74.5 %)	7.94 (-71.8 %)	4.53 (-77.7 %)

Tableau 12: Évaluation de plusieurs systèmes de recherche

Les résultats de cette première évaluation montrent que le modèle OKAPI et la variante LNU-LTC du modèle vectoriel possèdent une performance significativement meilleure que les autres systèmes, tant sur les sous-collections que sur la collection entière. Ces deux modèles ne sont cependant pas significativement différents l'un de l'autre (les variations de performance moyenne passent de +0.7% à -0.2%). Le système NNN-NNN, recourant uniquement à la fréquence d'apparition des termes, et le modèle d'indexation binaire BNN-BNN, possèdent une précision moyenne très mauvaise. Il est également intéressant de remarquer que les performances absolues des systèmes varient légèrement d'une collection à l'autre.

III.2 Fusion de collections avec systèmes de recherche similaires

Une des hypothèses émises par certains auteurs afin de résoudre le problème de fusion de collections est d'admettre que les systèmes appliqués sur les différentes collections sont rigoureusement identiques. C'est le cas par exemple lors de la distribution d'une collection importante sur différents serveurs reliés par un réseau local. Le Tableau 13 présente des stratégies de fusion où les mêmes systèmes ont interrogé les trois collections WSJ90, WSJ91 et WSJ92.

collection modèle	Précision moyenne (% variation)					
	WSJ entière base	séparation en WSJ90- WSJ91- WSJ92				
		chacun son tour		fusion par le score		fusion par le score & sélection optimale	
	186 requêtes	186 requêtes		186 requêtes		186 requêtes	
OKAPI - NPN	20.32	17.34	(-14.7 %)	20.37	(+0.2 %)	21.93	(+7.9 %)
LNU - LTC	20.27	17.28	(-14.8 %)	20.09	(-0.9 %)	22.10	(+9.0 %)
LTN - NTC	19.16	17.46	(-8.9 %)	18.76	(-2.1 %)	21.53	(+12.4 %)
HTN - BNN	19.09	16.55	(-13.3 %)	18.79	(-1.6 %)	20.76	(+8.7 %)
ATN - NTC	18.85	16.24	(-13.8 %)	18.13	(-3.8 %)	20.31	(+7.7 %)
ANN - NTC	16.78	14.80	(-11.8 %)	16.62	(-1.0 %)	18.74	(+11.7 %)
LNC - LTC	15.76	12.90	(-18.1 %)	15.60	(-1.0 %)	16.82	(+6.7 %)
LTC - LTC	15.06	12.87	(-14.5 %)	14.91	(-1.0 %)	16.04	(+6.5 %)
ANC - LTC	11.94	9.92	(-16.9 %)	11.77	(-1.4 %)	12.95	(+8.5 %)
ANN - ANN	10.76	9.17	(-14.8 %)	10.81	(+0.5 %)	11.51	(+7.0 %)
LNC - LNC	9.69	8.24	(-15.0 %)	9.76	(+0.7 %)	10.46	(+7.9 %)
BNN - BNN	4.85	5.58	(+15.1 %)	4.87	(+0.4 %)	6.18	(+27.4 %)
NNN - NNN	4.53	3.97	(-12.4 %)	4.53	(0.0 %)	4.92	(+8.6 %)

Tableau 13: Évaluation de plusieurs stratégies de fusion de collections utilisant le même système de dépistage

La stratégie très simple de fusion de collections "chacun son tour" montre une dégradation de l'ordre de 14% par rapport à une interrogation sur la collection entière (comme par exemple dans le cas d'une fusion avec échange d'informations entre les sous-collections). Ces résultats confirment des études précédentes ([Callan 95]), qui ont trouvé une dégradation allant jusqu'à 50%. Par contre, la stratégie de fusion par le score offre une performance légèrement supérieure, mais non significativement, à celle obtenue sur la collection entière WSJ. Cette stratégie de fusion est nettement plus performante que la technique "chacun son tour" (+17.5% pour le système OKAPI). Dans ces deux premières stratégies de fusion de collections, les 186 requêtes ont toutes été testées, alors que certaines d'entre elles n'ont pas de document pertinent pour l'une ou l'autre des sous-collections (cf. annexe ANNEXE 5). La troisième stratégie évaluée ici ne retient que les listes extraites des sous-collections sous la condition que cette sous-collection possède au moins un document pertinent. En conservant la méthode de fusion par le score, cette sélection optimale offre des performances significativement meilleures que les méthodes "chacun son tour" et fusion par le score, de même qu'une performance supérieure à celle qui traite la collection comme une seule unité.

Cependant, de notre point de vue, il est rare que dans la réalité (par exemple sur INTERNET) les mêmes systèmes de recherche soient appliqués sur les collections, et les résultats des stratégies de fusion présentées ci-dessus ne sont peut-être pas très réalistes.

III.3 Fusion de collections avec des systèmes de recherche différents

a) Évaluation et comparaison de notre stratégie

Nous avons développé nos modèles de fusion de collections en interrogeant trois collections avec des systèmes de recherche différents, situation plus réaliste que la précédente. Afin de juger la performance de nos modèles, nous avons également évalué les stratégies "chacun son tour", fusion par le score, et par le score normalisé (Tableau 14).

	Précision moyenne (% variation)		
Collection modèle	WSJ90 OKAPI 170 requêtes	WSJ91 LNU - LTC 171 requêtes	WSJ92 LNC - LTC 140 requêtes
précision moyenne	24.13	22.88	22.52
nombre de documents pertinents	2074	3563	831
nombre de doc. pertinents retrouvés	1765	2811	745
	WSJ 186 requêtes		
rang seul			
Chacun son tour	16.96		
Modèle 1 (ln(rang))	18.40 (+8.49 %)		
rang et/ou similarité			
Fusion par le score	8.75 (-48.41 %)		
Fusion par le score normalisé	15.42 (-9.08 %)		
Modèle 2 (SimMax)	17.34 (+2.24 %)		
Modèle 3 (ln(rang, SimDécomp)	19.01 (+12.09 %)		

Tableau 14: Évaluation de plusieurs stratégies de fusion de collections utilisant des systèmes de recherche différents

Dans ces circonstances, l'approche simple "chacun son tour" est significativement meilleure que les stratégies de fusion par le score. Ce résultat s'explique aisément pour la fusion par le score car les valeurs de similarités n'ont pas le même ordre de grandeur et sont difficilement comparables. La fusion par le score normalisé est ainsi plus performante que la fusion par le score simple mais reste cependant significativement moins bonne que la stratégie "chacun son tour", ce qui s'explique aisément car les valeurs des degrés de similarité des trois systèmes retenus n'ont pas le même ordre de grandeur. Le Modèle 2 qui utilise la similarité normalisée est légèrement plus performant que la stratégie "chacun son tour" mais pas de manière significative. Notre Modèle 1 de stratégie de fusion isolée et notre Modèle 3 de stratégie intégrée de fusion présentent des performances significativement meilleures que la stratégie de base "chacun son tour".

Cependant, cette évaluation est peut-être biaisée car nous avons utilisé le même jeu de requêtes pour estimer la valeur des coefficients des modèles et pour évaluer leur performance. Pour vérifier que ce biais n'a pas d'incidence marquée sur nos résultats, nous évaluons, dans la section suivante, la performance moyenne à l'aide de la méthode de la validation croisée ([Stone 74]).

b) Validation croisée

La validation croisée est une méthodologie d'évaluation dont le but est de séparer clairement les observations utilisées pour l'évaluation de celles utilisées pour l'estimation des paramètres. La démarche générale de la validation croisée est la suivante:

1. diviser les données de départ en n sous-ensembles;
2. Pour i=1 jusqu'à n, répéter:
   2.1 enlever le $i^{ème}$ sous-ensemble d'observations et seulement celui-ci;
   2.2 estimer le modèle à partir des n-1 sous-ensembles d'observations restants (c'est-à-dire tous sauf le $i^{ème}$);
   2.3 évaluer la performance P_i du modèle avec le $i^{ème}$ sous-ensemble qui avait été retiré;
3. calculer la moyenne arithmétique des n mesures de performance P_i, i=1...n.

Pour tester la validité de la valeur estimée des coefficients d'une part, et d'autre part, de son influence sur la performance de nos modèles, nous avons utilisé une validation croisée en subdivisant en dix l'ensemble des requêtes. La Figure 9 et la Figure 10 présentent les intervalles de confiance à 95% de chacune des estimations des coefficients (lignes verticales) du Modèle 1 pour WSJ91-LNU. Les valeurs des estimateurs $\hat{\alpha}$ et $\hat{\beta}$, calculées sur l'ensemble de la collection WSJ91, sont symbolisées par une ligne horizontale qui passe très près du centre de chaque intervalle de confiance. Des conclusions similaires peuvent être établies pour les deux autres sous-collections. Ainsi, l'objectif de stabilité et de fiabilité des coefficients du modèle est atteint (**1.2**).

L'évaluation du Modèle 1 par la méthode de la validation croisée retourne une précision moyenne à onze points fixes de rappel de 18.40. Ce chiffre, aux arrondis près, est le même que celui trouvé par l'évaluation sur l'ensemble des données (évaluation obtenue en utilisant les mêmes requêtes pour l'estimation et l'évaluation). Ce phénomène confirme ainsi que la performance obtenue par notre première évaluation ne possède pas de biais marqué.

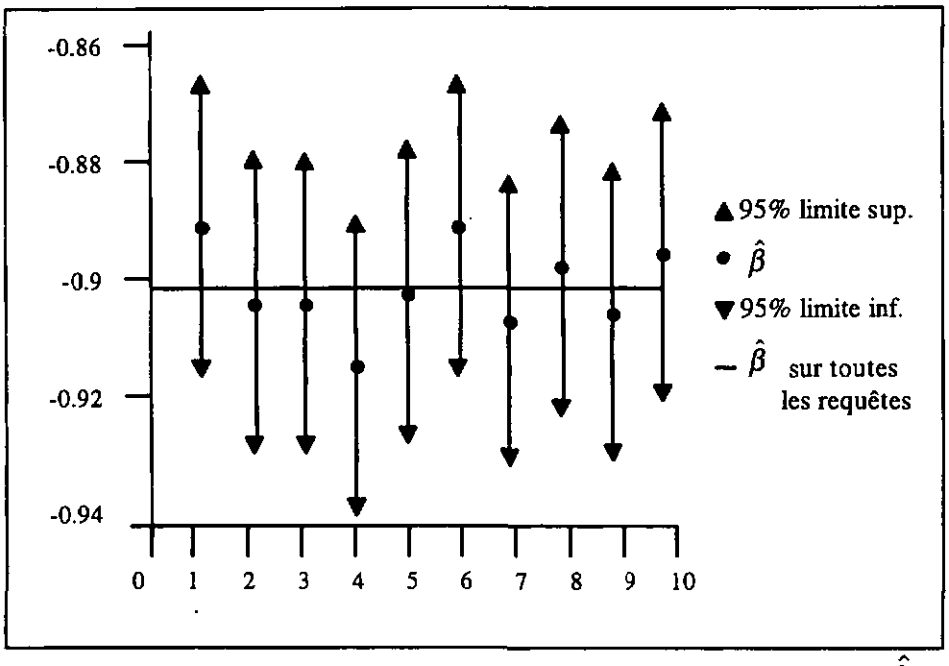

Figure 9: Intervalle de confiance à 95% pour les dix estimateurs $\hat{\beta}$ obtenus par validation croisée (WSJ91-LNU, 171 requêtes)

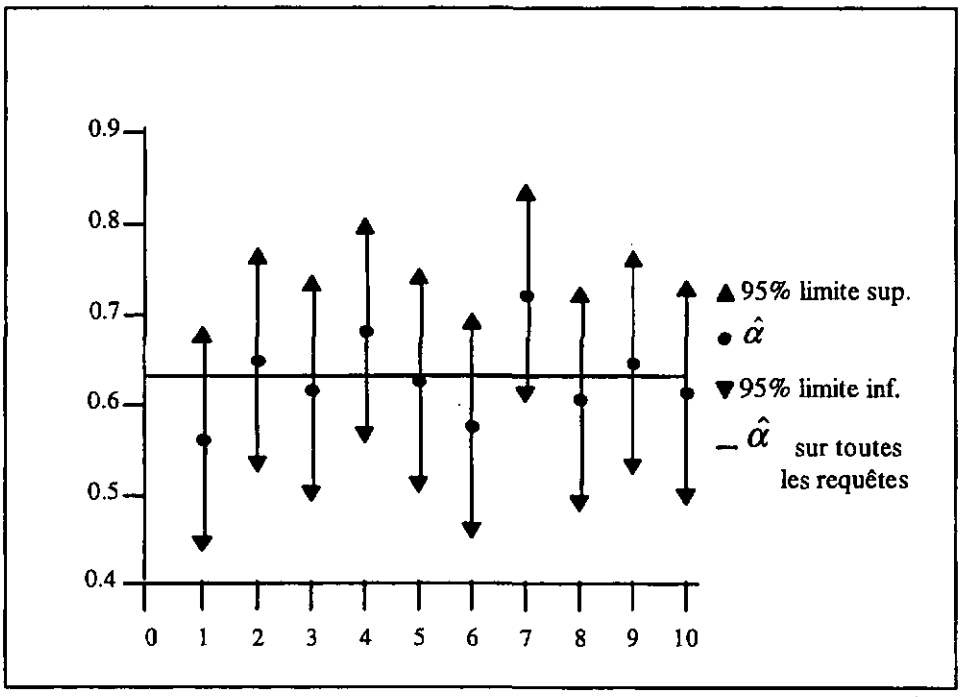

Figure 10: Intervalle de confiance à 95% pour les dix estimateurs $\hat{\alpha}$ obtenus par validation croisée (WSJ91-LNU, 171 requêtes)

IV. CONCLUSION

Nous avons proposé une stratégie de fusion de collections basée sur la régression logistique. Trois modèles ont été développés, correspondant chacun à des situations différentes, comme le résume le Tableau 15.

	Information disponible	Variables utilisées	% d'amélioration par rapport à la meilleure stratégie: "chacun son tour"
Modèle 1	rang	ln(rang)	+ 8.49%
Modèle 2	degré de similarité	SimMax	+2.24%
Modèle 3	rang et degré de similarité	ln(rang) et SimDécomp	+12.09%

Tableau 15: Résumé des résultats des modèles logistiques

Nous avons atteint nos objectifs fixés au départ:

1 *Développer un nouveau modèle théorique justifiant clairement la manière de combiner plusieurs évidences;*

Notre stratégie de fusion de collections est issue directement de la méthodologie de la régression logistique.

1.1 *Trouver un modèle général pouvant s'adapter aisément à des situations différentes, selon les informations disponibles pour effectuer la fusion;*

Notre stratégie présente ici deux modèles qui correspondent à des cas réels différents, celui de la stratégie de fusion isolée et celui de la stratégie intégrée de fusion. Pour répondre à d'autres cas de figures, il suffit de choisir des variables explicatives appropriées, le modèle de fusion restant le même.

1.2 *Obtenir des paramètres stables et fiables pour notre modèle;*

La validation croisée prouve que les paramètres sont stables et que l'évaluation recourant aux mêmes requêtes pour l'estimation et le test demeure dans notre cas une approche sans biais important.

1.3 *Ordonner clairement les documents retrouvés afin d'améliorer leur sélection lors de la fusion;*

Les documents, grâce à la régression logistique, sont aisément ordonnés selon leur probabilité de pertinence.

1.5 *En fusion de collections, savoir combien de documents il faut sélectionner dans chaque liste;*

Notre stratégie de fusion élimine ce problème de nombre de documents car les documents peuvent, par l'intermédiaire des probabilités, être directement comparés puis classés.

2 *Évaluer empiriquement notre modèle et obtenir une performance significativement supérieure à celle des autres stratégies;*

Les modèles 1 (ln(rang)) et 3 (ln(rang), SimDécomp) sont significativement plus performants que les méthodes proposées jusqu'à présent. Le Modèle 2 (SimMax) peut-être mis de coté car ses résultats n'apportent pas d'amélioration suffisante.

Plusieurs de nos réflexions de départ se sont également avérées justes:

2 *Le rang des documents peut être un critère suffisant à lui seul pour effectuer la fusion;*

Le Modèle 1, se basant uniquement sur le rang des documents retrouvés, offre de bons résultats.

3 *De nouvelles variables calculées à partir de la similarité et du rang pourraient apporter de meilleurs résultats;*

Dans les deux modèles retenus, l'introduction de deux nouvelles variables apporte de très bons résultats, à savoir le logarithme népérien du rang (ln(rang)), et la variable SimDécomp calculée selon la formule suivante:

$$\text{SimDécomp}_{rq} = \frac{\text{Sim}_{rq} - \mu_{(\text{Sim}_r)_q}}{\sigma_{(\text{Sim}_r)_q}}$$

où μ et σ, sont respectivement la moyenne et l'écart type, sur l'ensemble des requêtes q, de la similarité des documents au même rang r.

Nous avons donc proposé un modèle général de fusion de collections basé sur la régression logistique. Si ce modèle apporte de très bons résultats, son implantation nécessite cependant de connaître a priori des jugements de pertinence sur quelques requêtes, afin de pouvoir calculer les coefficients de la régression. Dans cette recherche, nous avons estimé les coefficients des régressions à l'aide des 1000 premiers documents retrouvés par requête. Cependant, il est rare dans la réalité de pouvoir disposer de jugements de pertinence sur autant de documents. Nous avons donc estimé les coefficients du Modèle 1 sur les 30 premiers documents retrouvés, puis avons appliqué ces valeurs pour la fusion en demandant 1000 documents par requête. La précision moyenne de cette fusion est de 18.14, soit très faiblement inférieure à celle obtenue par le Modèle 1 sur les 1000 documents (18.40). Ainsi, l'implantation d'un modèle logistique en fusion de collections ne nécessite pas de connaître un très grand nombre de jugements de pertinence.

PARTIE 3

LA REGRESSION LOGISTIQUE APPLIQUEE

A LA FUSION DE DONNÉES

I. PRÉSENTATION DES DONNÉES 58

II. LE MODÈLE MATHÉMATIQUE PROPOSÉ 60

II.1 La régression logistique 60

II.2 Développement des modèles 66

III. ÉVALUATION ET COMPARAISON 68

IV. QUELQUES RÉFLEXIONS SUR LA FUSION DE DONNÉES 69

V. CONCLUSION 72

I. PRÉSENTATION DES DONNÉES

Pour présenter notre modèle de fusion de données, nous avons choisi de fusionner les réponses des trois systèmes de dépistage les plus performants parmi ceux évalués sur la collection WSJ dans le Tableau 12, à savoir:

1. le modèle probabiliste OKAPI
2. le modèle vectoriel LNU-LTC
3. le modèle vectoriel LTN-LTC

Chacun de ces trois systèmes interroge le corpus WSJ (74'520 documents, 186 requêtes) et retourne, pour chaque requête, une liste de 1000 documents ordonnés selon leur score (principes adoptés par la conférence TREC).

Chacun des systèmes de dépistage retenus nous fournit deux variables pouvant servir de critères lors de la fusion de données: le rang et le degré de similarité. Dans notre modèle, nous utilisons trois variables, déjà introduites dans la partie précédente et issues de ces deux informations:

- **Sim** qui est le degré de similarité d'un document retourné par rapport à la requête;

- **SimMax** qui est le degré de similarité d'un document divisé par le degré de similarité maximal de tous les documents de la requête. Cette variable normée représente la baisse du degré de similarité d'un document à l'autre;

• **ln(rang)** qui est le logarithme népérien du rang de chaque document retrouvé par une méthode (l'utilisation de cette variable s'est montrée très attractive pour la fusion de collections).

Le Tableau 16 présente les principales statistiques descriptives de ces variables, pour chacun des systèmes de recherche interrogeant la collection WSJ. Les trois systèmes retrouvent des proportions de documents pertinents très similaires. Les valeurs des variables Sim sont très différentes d'un système à l'autre. De plus, pour les systèmes OKAPI et LNU-LTC, ces valeurs sont en moyenne légèrement supérieures à celles retrouvées lorsqu'ils interrogent les sous-collections WSJ90 et WSJ92 (voir pour comparaison le Tableau 4). Enfin, la transformation du score par la variable SimMax permet d'obtenir des valeurs comparables.

Collection WSJ 186 requêtes 6'468 documents pertinents		OKAPI	LNU - LTC	LTN - LTC
nombre de doc. pertinents retrouvés		4606 (71.21%)	4672 (72.23%)	4554 (70.41%)
Précision moyenne		20.32	20.27	19.16
Sim	Minimum	2.56	0.002	1.22
	Maximum	52.19	0.03	36.00
	Moyenne	8.13	0.007	5.27
	Écart-type	3.41	0.003	2.08
SimMax	Minimum	0.11	0.12	0.07
	Maximum	1	1	1
	Moyenne	0.43	0.43	0.33
	Écart-type	0.14	0.14	0.14
ln(Rang)	Minimum	0		
	Maximum	6.91		
	Moyenne	5.91		
	Écart-type	0.98		

Tableau 16: Statistiques descriptives des variables(WSJ, 186 requêtes)

Ces trois variables ainsi définies servent de critères de fusion dans notre stratégie de fusion de données. Le modèle mathématique de cette stratégie repose sur l'utilisation de la régression logistique pour prédire les probabilités de pertinence des documents retrouvés par les différents systèmes de dépistage de l'information.

II. LE MODÈLE MATHÉMATIQUE PROPOSÉ

II.1 La régression logistique

Si la fusion de collections et la fusion de données ont toujours été traitées séparément, notre point de vue est qu'il s'agit dans les deux cas d'un problème de combinaison: comment combiner de façon optimale diverses listes de documents? Notre optique est alors que la même méthodologie de fusion peut s'appliquer à ces deux problèmes de fusion (réflexion **1**). Nous développons donc, comme pour la fusion de collections, un modèle de fusion de données basé sur la régression logistique afin de combiner et de sélectionner les documents retrouvés par différents systèmes de dépistage de l'information.

L'utilisation de cette méthodologie statistique pour le problème de fusion de données se justifie par des raisons similaires à celles avancées pour la fusion de collections:

- la régression logistique est particulièrement adaptée à la prédiction d'une variable binaire, c'est-à-dire la pertinence (cf. Figure 7);
- les données ne démentent pas l'ajustement par une courbe en S de la distribution de la probabilité de pertinence des documents retrouvés (cf. exemple avec ln(rang) en Figure 11) .

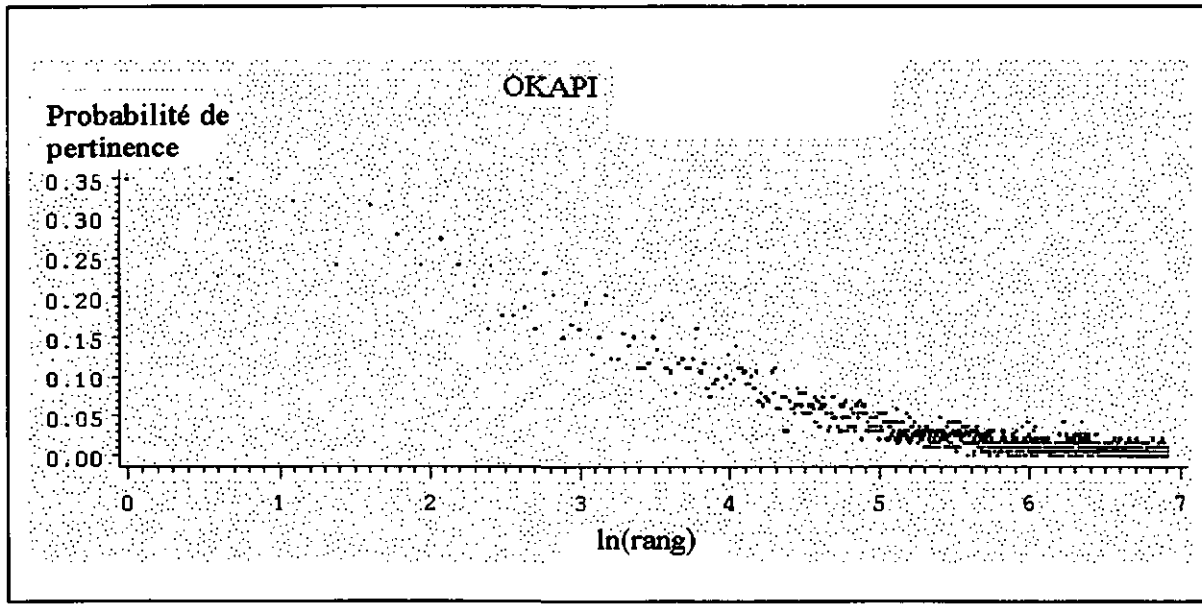

Figure 11: Relation empirique entre la variable ln(rang) et la probabilité de pertinence (OKAPI, WSJ, 186 requêtes)

- la fonction g(x), qui représente le rapport de la probabilité de pertinence divisée par la probabilité de non pertinence, est fortement corrélée linéairement avec les variables, comme le montre le Tableau 17 pour le système OKAPI.

Coefficient de corrélation de Pearson ρ	g(x)	
	ρ	Prob(ρ=0)
Sim	0.47	0.0001
SimMax	0.58	0.0001
ln(rang)	-0.88	0.0001

Tableau 17: Coefficients de corrélation de Pearson (ρ) entre les variables et g(x) (OKAPI, WSJ, 186 requêtes)

D'autre part, l'emploi de cette méthode nous permet d'atteindre plusieurs de nos objectifs :

1 *Développer un nouveau modèle théorique justifiant clairement la manière de combiner plusieurs évidences*
La régression logistique est une méthodologie statistique reconnue et apporte un fondement théorique au modèle de fusion de données.

1.1 *Trouver un modèle général pouvant s'appliquer aisément quelles que soient les informations disponibles pour effectuer la fusion*
Le modèle logistique est un modèle général dont la forme particulière peut varier selon le choix des variables explicatives. Ainsi, si le rang est la seule information fournie par les systèmes de dépistage, le modèle de fusion ne comportera qu'une seule variable explicative basée sur ce rang. Si d'autres informations sont également retournées (comme le degré de similarité), il est alors possible de les intégrer comme variables explicatives supplémentaires.

1.3 *Ordonner clairement les documents retrouvés afin d'améliorer leur sélection lors de la fusion*
La régression logistique permet de calculer les probabilités de pertinence des différents documents. Ces probabilités sont des valeurs comparables entre elles et permettent donc un classement clair des documents.

Les deux problèmes de fusion de données et de collections ne sont pas identiques, et la fusion de données comporte une forte contrainte sur les duplicata. En effet, comme les systèmes de dépistage interrogent strictement la même collection, les résultats de leur recherche contiennent des documents communs.

Sur l'ensemble des 186'000 documents retrouvés pour toutes les requêtes, le système OKAPI et le système LNU-LTC possèdent 34'373 documents communs, soit 15.6% de l'ensemble des documents retrouvés par les deux systèmes. Les systèmes OKAPI et LTN-LTC ont 28.56% de documents en commun, et les systèmes LNU-LTC et LTN-LTC 25.27%. Cette contrainte est à intégrer dans notre stratégie de fusion, car le traitement de chaque liste ne peut plus s'effectuer de manière totalement indépendante.

La régression logistique nous permet d'aborder la fusion de données et sa contrainte de duplicata sous deux formes différentes. Nous développons ainsi deux modèles principaux: le **modèle séparé** et le **modèle combiné**. L'adjectif "séparé" indique qu'une régression logistique est appliquée séparément sur chaque liste de document. Dans le modèle "combiné", une seule régression est effectuée sur l'ensemble des divers résultats des systèmes de repérage préalablement regroupés (combinés).

Le modèle séparé:

Le modèle séparé, schématisé dans la Figure 12, est très proche de notre modèle de fusion de collections. Cette stratégie de fusion suit les principales étapes suivantes:

1. pour chaque liste de documents retournés par les systèmes de repérage, estimer la probabilité de pertinence des documents par une régression logistique;
2. pour l'ensemble des documents retrouvés par tous les systèmes, déterminer une note finale, calculée à partir des différentes probabilités de pertinence estimées à l'étape précédente;
3. ordonner tous les documents en fonction de leur note finale;
4. sélectionner les documents possédant les notes les plus fortes.

Pour déterminer la note finale à attribuer à chaque document, nous avons choisi de calculer la moyenne arithmétique des probabilités fournies par les diverses régressions. Ainsi:

$$\text{note}(D_i) = \frac{1}{c}\sum_{j=1}^{c} P_j(D_i)$$

où $P_j(D_i)$ est la probabilité de pertinence du document D_i estimée à l'aide de la régression logistique pour le système *j* (avec *c* systèmes).

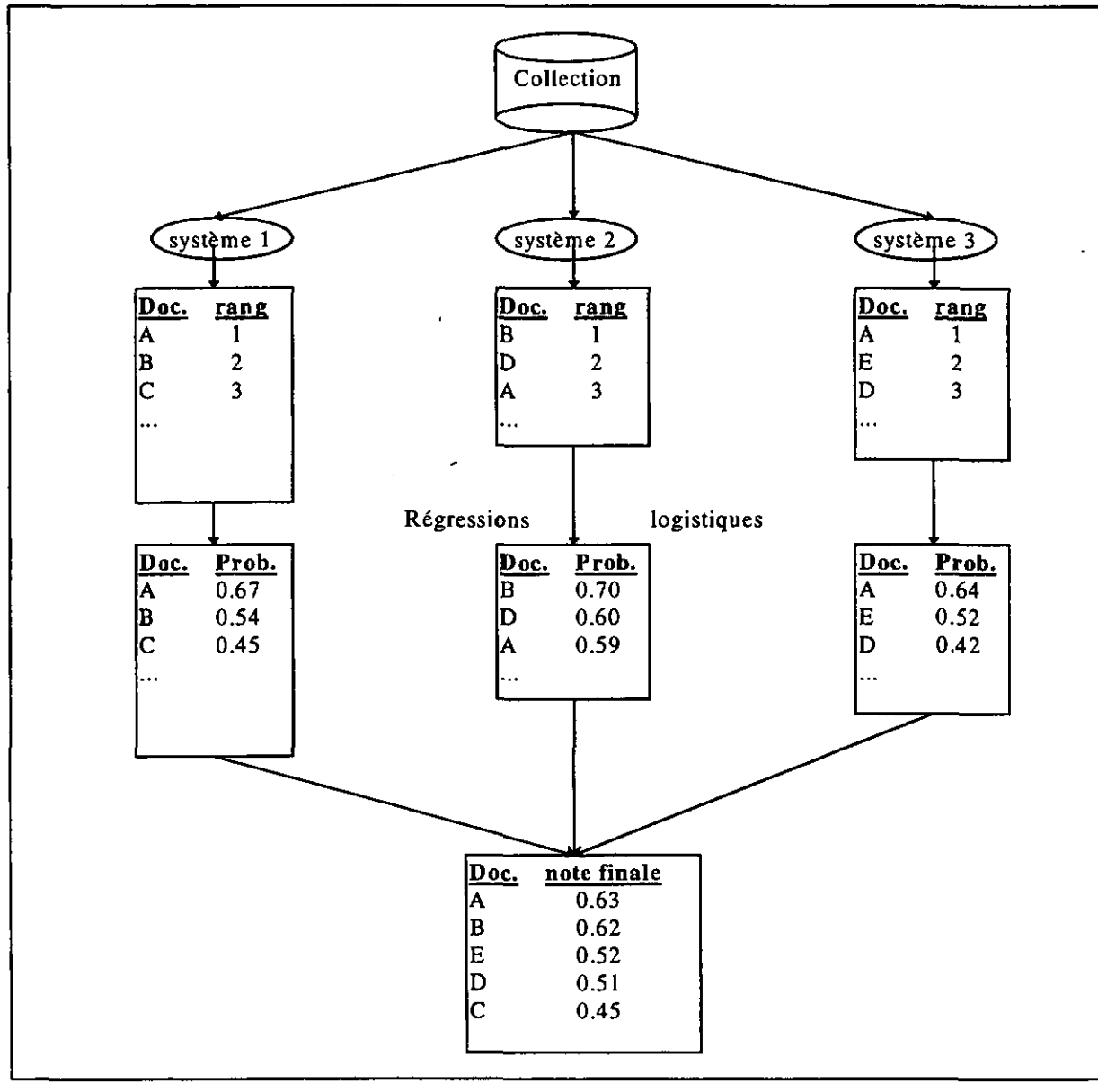

Figure 12 : Modèle séparé de fusion de données

Le Modèle combiné

Le modèle combiné (schématisé dans la Figure 13) repose sur l'idée que les scores et/ou rangs fournis par les différents systèmes sont des caractéristiques importantes de chaque document retrouvé. Ces variables deviennent alors, pour chaque document distinct, une des variables explicatives de la régression logistique. La fusion s'effectue donc comme suit:

1. combiner les listes de documents en un seul jeu de données:
   - pour chaque document, attribuer une variable pour chaque information fournie par les systèmes de repérage (le nombre de variables est donc égal au nombre de systèmes utilisés multiplié par le nombre de variables retenues par système);
   - si un document n'est pas retrouvé par un système, attribuer une valeur nulle à sa variable pour appuyer le fait qu'il n'a pas été retrouvé par le système;
2. estimer la probabilité de pertinence de tous les documents en appliquant une régression logistique sur cet ensemble de données;
3. ordonner les documents selon leur probabilité de pertinence;
4. sélectionner les documents ayant les plus fortes probabilités.

Dans le cas d'un modèle combiné, le Tableau 18 présente les statistiques descriptives des variables SimMax et Sim pour les trois systèmes OKAPI, LNU-LTC, et LTN-LTC appliqués à la collection WSJ. Nous pouvons constater une baisse de la moyenne et une hausse de l'écart-type des variables par rapport aux statistiques d'un modèle séparé (Tableau 16). Ces différences s'expliquent par l'attribution de la valeur nulle aux variables des systèmes qui ne retrouvent pas un document dépisté par d'autres systèmes.

279'460 doc. différents	Minimum	Maximum	Moyenne	Écart-type
Sim OKAPI	0	52.19	5.41	4.74
Sim LNU	0	0.03	0.005	0.004
Sim LTN	0	36.00	3.51	3.01
SimMax OKAPI	0	1	0.29	0.23
SimMax LNU	0	1	0.28	0.23
SimMax LTN	0	1	0.22	0.20

Tableau 18: Statistiques descriptives des variables dans le modèle combiné (WSJ, 186 requêtes)

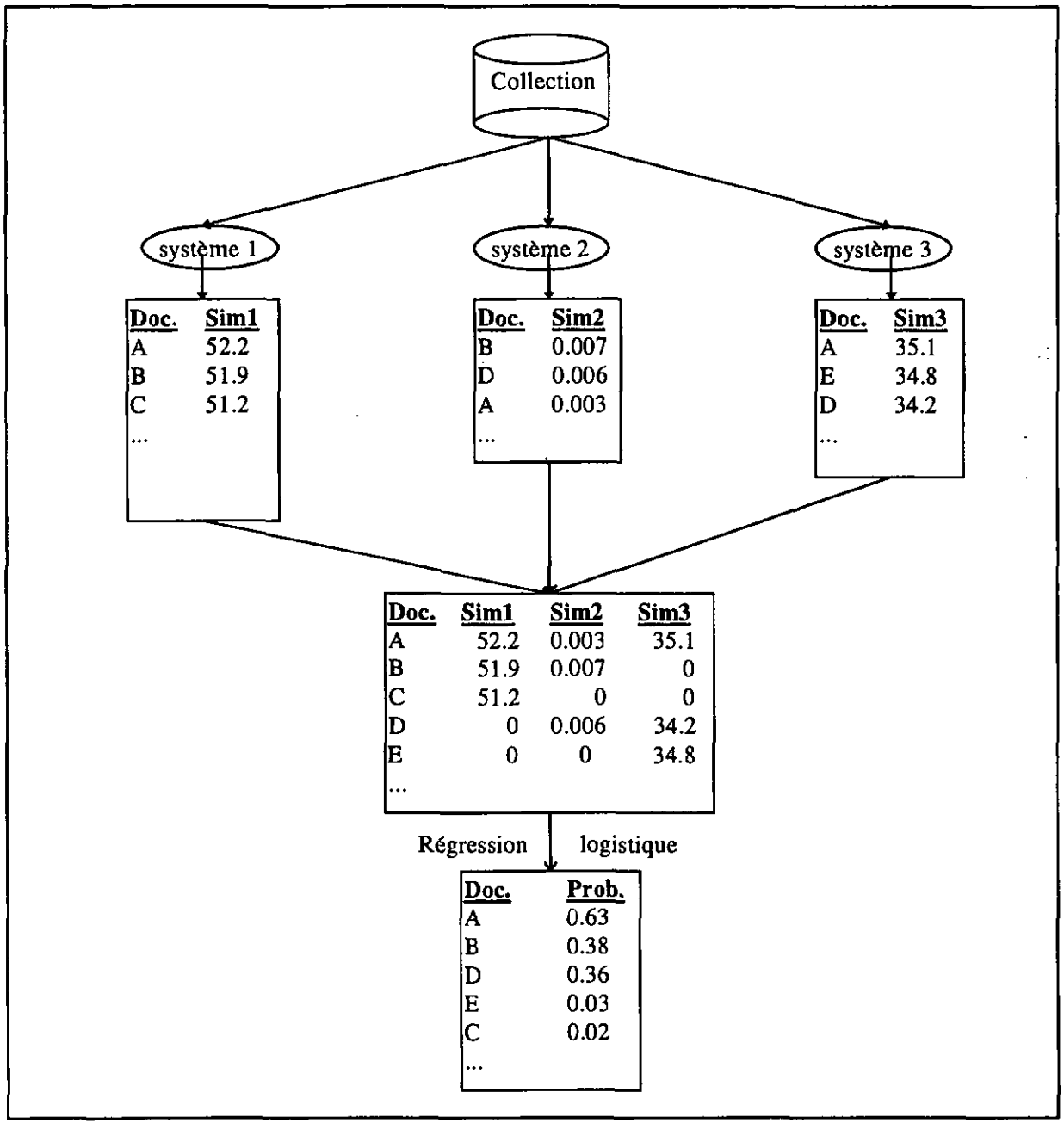

Figure 13: Modèle combiné de fusion de données

Le modèle séparé et le modèle combiné utilisent tous deux la régression logistique pour traiter le problème de fusion de données mais selon une approche quelque peu différente. De même que pour la fusion de données, nous avons voulu étudier les situations correspondant au cas où seul le rang des documents est disponible, et à celui où le degré de similarité l'est aussi.

II.2 Développement des modèles

Nous avons développé trois modèles de fusion de données. Le premier modèle proposé est un modèle séparé, tandis que les deux autres (Modèle 2 et Modèle 3) sont des modèles combinés.

- **Modèle 1**

Si le rang des documents retrouvés est la seule information fournie par les systèmes de repérage, nous suggérons de recourir à un modèle séparé basé sur le logarithme népérien des rangs. La variable ln(rang), dans un tel type de modèle, a en effet fourni de très bons résultats en fusion de collections.

Le Tableau 19 présente les valeurs des coefficients calculés par les trois régressions logistiques, ainsi que les résultats du test de Wald sur les paramètres et du test du χ^2 sur les modèles. Toutes les valeurs estimées des coefficients sont considérées significativement différentes de zéro et les trois modèles sont jugés significatifs.

		valeur estimée, IC à 95%	écart type estimé	test de Wald sur chaque paramètre	χ^2 sur le modèle
OKAPI	$\hat{\alpha}$	0.697 [0.603 ; 0.791]	0.048	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.815 [-0.835 ; -0.795]	0.010	p=0.0001	
LNU-LTC	$\hat{\alpha}$	0.708 [0.614 ; 0.802]	0.048	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.814 [-0.834 ; -0.794]	0.010	p=0.0001	
LTN-LTC	$\hat{\alpha}$	0.536 [0.440 ; 0.632]	0.049	p=0.0001	p=0.0001
	ln(rang) $\hat{\beta}$	-0.784 [-0.804 ; -0.764]	0.010	p=0.0001	

Tableau 19: Valeurs des coefficients de la régression logistique pour le Modèle 1

- **Modèle 2**

Le Modèle 2 est un modèle combiné qui utilise les degrés de similarité des documents fournis par les divers systèmes de repérage. Nous avons choisi, pour ce modèle, de n'utiliser que la variable SimMax, c'est-à-dire le degré de similarité divisé par le degré de similarité maximal. Les probabilités de pertinence des documents sont alors calculées selon l'équation suivante:

$$\text{Prob}[\text{Pert}(Di)] = \frac{e^{\alpha+\beta_1 \cdot \text{SimMax}_{\text{OKAPI}}(Di)+\beta_2 \cdot \text{SimMax}_{\text{LNU}}(Di)+\beta_3 \cdot \text{SimMax}_{\text{LTN}}(Di)}}{1+e^{\alpha+\beta_1 \cdot \text{SimMax}_{\text{OKAPI}}(Di)+\beta_2 \cdot \text{SimMax}_{\text{LNU}}(Di)+\beta_3 \cdot \text{SimMax}_{\text{LTN}}(Di)}}$$

Les statistiques de Wald sur chaque paramètre estimé par la régression (Tableau 20) montrent que les valeurs estimées des coefficients sont significativement différentes de zéro (p=0.00001). Le modèle dans son ensemble est également jugé significatif.

	valeur estimée, IC à 95%	écart type estimé	test de Wald sur chaque paramètre	χ^2 sur le modèle
$\hat{\alpha}$	-6.482 [-6.558 ; -6.406]	0.039	p=0.0001	p=0.0001
SimMax OKAPI $\hat{\beta}_1$	2.683 [2.426 ; 2.940]	0.131	p=0.0001	
SimMax LNU $\hat{\beta}_2$	1.121 [0.852 ; 1.390]	0.137	p=0.0001	
SimMax LNC $\hat{\beta}_3$	3.012 [2.847 ; 3.177]	0.084	p=0.0001	

Tableau 20: Valeurs des coefficients de la régression logistique pour le Modèle 2

- **Modèle 3**

Le dernier modèle (combiné) repose sur l'utilisation des variables SimMax et des degrés de similarité (Sim) des documents retrouvés. Les valeurs estimées des coefficients sont toutes significativement différentes de zéro, et le test du χ^2 sur le modèle montre que le modèle est significatif (Tableau 21).

	valeur estimée, IC à 95%	écart type estimé	test de Wald sur chaque paramètre	χ^2 sur le modèle
$\hat{\alpha}$	-6.481 [-6.559 ; -6.403]	0.040	p=0.0001	
SimMax OKAPI $\hat{\beta}_1$	1.962 [1.643 ; 2.281]	0.163	p=0.0001	
SimMax LNU $\hat{\beta}_2$	-0.110 [-0.518 ; 0.298]	0.208	p=0.0001	
SimMax LNC $\hat{\beta}_3$	-0.400 [-0.755 ; -0.045]	0.181	p=0.0001	p=0.0001
Sim OKAPI $\hat{\beta}_4$	0.00002 [0.00001 ; 0.00003]	0.000005	p=0.0001	
Sim LNU $\hat{\beta}_5$	0.091 [0.073 ; 0.109]	0.009	p=0.0001	
Sim LNC $\hat{\beta}_6$	0.0002 [0.00018 ; 0.00022]	0.00001	p=0.0001	

Tableau 21 : Valeurs des coefficients de la régression logistique pour le Modèle 3

III. ÉVALUATION ET COMPARAISON

Afin de juger la performance de nos trois modèles de fusion de données, nous avons évalué, sur la même collection-test et avec les mêmes systèmes de repérage, la stratégie classique de fusion basée sur la somme.

Ces évaluations, présentées dans le Tableau 22, montrent tout d'abord l'efficacité d'une fusion de données. En effet, les quatre stratégies de fusion de données sont toutes significativement plus performantes (de 5% à 8%) que le meilleur système de recherche (OKAPI). En regardant de plus près les performances de chaque stratégie de fusion, aucune différence significative en terme de précision moyenne n'existe entre la fusion par somme et nos modèles logistiques (-2.5% pour le Modèle 1, -0.5% pour le Modèle 2, et +0.05% pour le Modèle 3). Parmi nos trois modèles, le modèle séparé (Modèle 1) présente de moins bons résultats que les modèles combinés. Les différences de stratégie entre combiné et séparé font que dans le modèle séparé chaque système est optimisé séparément avant la fusion, tandis que dans le modèle combiné les résultats sont fusionnés selon les caractéristiques des différents systèmes. Cette optimisation plus globale du modèle combiné prend en compte la différence de qualité de réponse entre les systèmes, ce que ne fait pas le modèle séparé.

Collection WSJ 186 requêtes 6'468 documents pertinents	Précision moyenne (% variation)		
	OKAPI	LNU-LTC	LTN-LTC
	20.32 (base)	20.27 (-0.25 %)	19.16 (-5.71 %)
Fusion par somme	22.03 (+8.42 %)		
Modèle séparé Modèle 1 (ln(rang))	21.47 (+5.66 %)		
Modèle combiné Modèle 2 (SimMax) Modèle 3 (SimMax, Sim)	21.92 (+7.87 %) 22.04 (+8.46 %)		

Tableau 22: Évaluation de plusieurs stratégies de fusion de données

Pour l'évaluation de ces stratégies de fusion de données, nous avons choisi d'utiliser les résultats des trois meilleurs systèmes de repérage appliqués à la collection-test WSJ. La fusion de données de cette combinaison particulière de systèmes permet d'obtenir une augmentation significative de la précision moyenne par rapport à l'emploi du meilleur système. Cependant, faut-il en conclure aussitôt que toute combinaison de système fournit automatiquement de meilleurs résultats? En effet, l'efficacité de la fusion de données ne dépendrait-elle pas également du choix des systèmes à combiner, et du nombre de systèmes à combiner? La partie suivante donne quelques réflexions sur ce sujet et ouvre la discussion.

IV. QUELQUES RÉFLEXIONS SUR LA FUSION DE DONNÉES

La fusion de données part de l'idée que si un même document est retrouvé par plusieurs systèmes de recherche, alors il possède plus de chance d'être pertinent qu'un document dépisté par un seul système ([Saracevic 88]). L'évaluation précédente de plusieurs modèles de fusion de données semble démontrer l'efficacité d'une telle combinaison. Cependant, ce principe ne peut s'étendre indéfiniment et si la fusion de données combine un trop grand nombre de systèmes, la performance peut en être affaiblie (les documents pertinents se noyant alors dans la masse de documents retrouvés). Deux questions se posent alors:

- Existe-t-il un nombre optimal de systèmes à combiner, au delà duquel la fusion de données n'est plus performante?
- Tous les systèmes sont-ils intéressants à combiner?

Pour obtenir quelques éléments de réponse à cette première question, nous avons fusionné, selon le modèle simple de fusion par somme, les résultats de plusieurs systèmes de repérage en combinant un système de plus à chaque fois (les systèmes ont été ajoutés selon l'ordre décroissant de leur performance évaluée dans le Tableau 12). Le Tableau 23 présente ainsi en première colonne le nombre et le nom des systèmes combinés (le système LNC-LTC est référencé par LNC2, et LNC-LNC par LNC). La troisième colonne indique le pourcentage de variation de la performance de la fusion par rapport à la fusion précédente, et en dernière colonne la différence en pourcentage entre la précision de la fusion effectuée et celle du meilleur système (OKAPI).

	Collection WSJ 186 requêtes	précision moyenne	% variation	% OKAPI
1	OKA	20,32		
2	OKA-LNU	20,32	0,00	0,00
3	OKA-LNU-LTN	22,03	8,42	8,42
4	OKA-LNU-LTN-HTN	21,85	-0,82	7,53
5	OKA-LNU-LTN-HTN-ATN	22,24	1,78	9,45
6	OKA-LNU-LTN-HTN-ATN-ANN	22,2	-0,18	9,25
7	OKA-LNU-LTN-HTN-ATN-ANN-LNC2	22,2	0,00	9,25
8	OKA-LNU-LTN-HTN-ATN-ANN-LNC2-LTC	22,23	0,14	9,40
9	OKA-LNU-LTN-HTN-ATN-ANN-LNC2-LTC-BNN	20,36	-8,41	0,20
10	OKA-LNU-LTN-HTN-ATN-ANN-LNC2-LTC-BNN-LNC	20,41	0,25	0,44
11	OKA-LNU-LTN-HTN-ATN-ANN-LNC2-LTC-BNN-LNC-NNN	12,7	-37,78	-37,50

Tableau 23 : Évaluation de fusion de données par somme sur plusieurs combinaisons de systèmes

Nous constatons tout d'abord que la fusion de deux systèmes (OKAPI et LNU) n'apporte aucune amélioration par rapport à l'interrogation de la collection par le système OKAPI seul. Par contre, une différence significative apparaît dès qu'un troisième système (LTN) est intégré (+8.42%). Si les combinaisons de 5, 6, 7, et 8 systèmes offrent de très bonnes performances, elles n'apportent cependant pas d'augmentation significative de la précision par rapport à la combinaison des trois premiers systèmes. Enfin, combiner plus de 8 systèmes entraîne, dans notre exemple, une diminution considérable de la performance moyenne. L'évolution de la valeur de la précision moyenne de ces combinaisons est reprise dans la Figure 14, où les traits épais indiquent une variation significative.

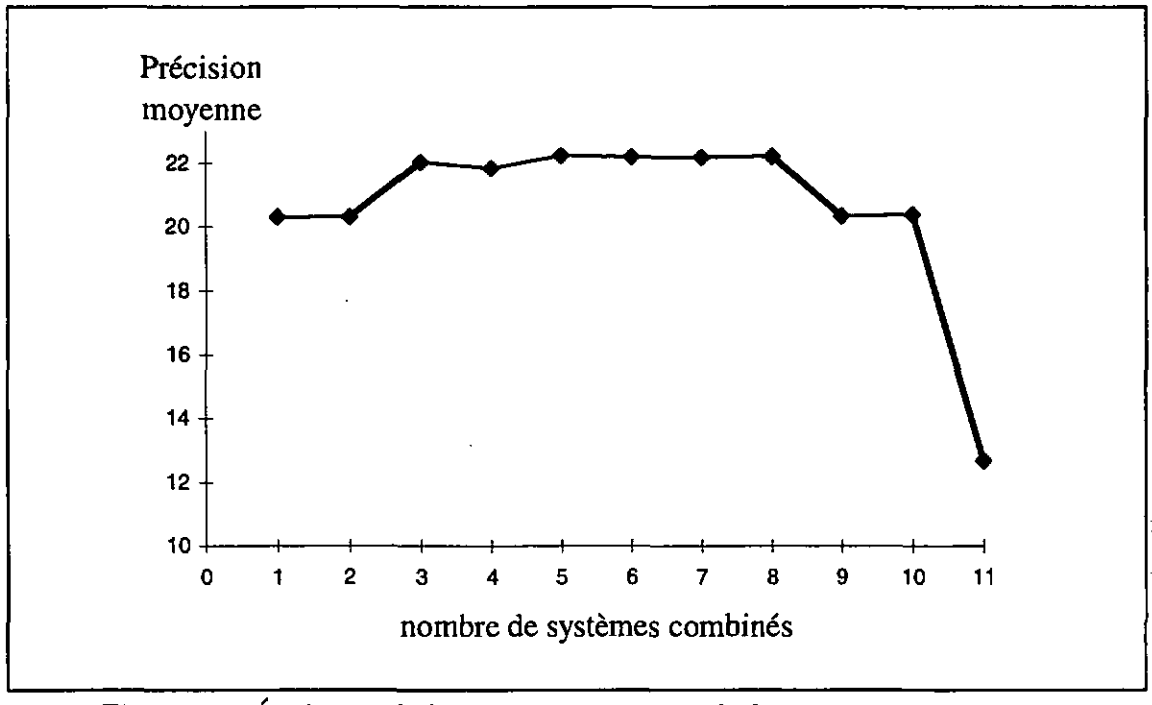

Figure 14: Évolution de la précision moyenne de fusion par somme pour plusieurs combinaisons de systèmes (WSJ, 186 requêtes)

Après cette première réflexion, nous pouvons nous demander si les différences de performance entre les diverses combinaisons proviennent uniquement du nombre de systèmes utilisés, ou bien également du choix des systèmes à combiner. Un petit exemple de trois autres combinaisons de systèmes (Tableau 24) présente trois situations différentes. La première fusion de données effectuée avec les systèmes LNU, LTN, et HTN est moins performante (de manière non significative) qu'une simple interrogation par le meilleur système (LNU). La deuxième fusion (OKA-LNU-HTN) présente une précision moyenne supérieure à celle du meilleur des trois systèmes combiné (OKAPI). Enfin, la dernière combinaison produit des résultats équivalents à une interrogation simple par OKAPI. Ce petit exemple montre bien que le choix des systèmes lors d'une fusion de données a une influence sur la performance finale.

Collection WSJ, 186 requêtes					
Système individuel	OKA	LNU	LTN	HTN	LNC2
Précision	20,32	20,27	19,16	19,09	18.39
Fusion par somme	Précision	% variation (par rapport au meilleur système combiné)			
LNU-LTN-HTN	19,59	-3,35%	(LNU)		
OKA-LNU-HTN	21,15	+4,08%	(OKAPI)		
OKA-LNU-LNC2	20,32	0	(OKAPI)		

Tableau 24: Exemple de combinaisons de trois systèmes

Pour choisir les systèmes, Lee ([95]) a proposé de les combiner selon leurs caractéristiques de pondération, et a défini dans ce but trois classes de systèmes de dépistage :

- Classe C : normalisation par le cosinus;
- Classe M : normalisation par le maximum;
- Classe N : autres.

Il suggère ensuite de combiner des systèmes de la classe C, avec des systèmes des autres classes.

Cependant, cette approche ne se vérifie pas dans notre exemple puisque la demière combinaison qui utilise un système issu de la classe C (LNC2), et deux systèmes de la classe N (OKAPI et LNU), fournit des résultats équivalents à ceux d'une interrogation simple par OKAPI. Ceci peut s'expliquer car la classification de Lee ne tient pas compte des propriétés des requêtes ni de la performance des systèmes.

Ces quelques réflexions montrent que la fusion de données ne permet pas toujours d'améliorer la performance d'une recherche d'informations. Son efficacité peut dépendre en effet du nombre et du choix des systèmes à combiner.

V. CONCLUSION

Pour aborder le problème de la fusion de données, nous avons développé deux stratégies de modèles basées sur la régression logistique :

- **un modèle séparé** où les probabilités de pertinence des documents sont tout d'abord estimées par une régression logistique sur la base des résultats de chaque système, puis combinées.

- **un modèle combiné** où les résultats des divers systèmes de dépistage sont préalablement regroupés en un seul jeu de données. Une seule régression logistique est ensuite appliquée sur l'ensemble de ces données pour fournir les probabilités de pertinence des différents documents.

Nous avons testé trois modèles (un séparé et deux combinés) qui reflètent des situations différentes :

- **Modèle 1 (séparé) :** lorsque le rang est la seule information retournée par les systèmes, la variable explicative de chaque régression est le logarithme népérien des rangs (ln(rang)).

• **Modèle 2 (combiné) :** lorsque le degré de similarité est disponible, les variables explicatives du modèle sont les variables SimMax (similarité du document divisée par similarité maximale) de chaque système.

• **Modèle 3 (combiné) :** les variables explicatives du modèle sont le logarithme népérien des rangs et les degrés de similarité fournis par les systèmes.

Ces trois modèles ont été évalués puis comparés à la stratégie de fusion par somme. Aucun de nos modèles logistiques ne s'est montré significativement plus performant que la fusion par somme.

Cependant, toutes ces stratégies de fusion permettent une amélioration significative de la précision moyenne par rapport à une interrogation de la collection à l'aide d'un seul système de dépistage. Ainsi, si nos modèles de fusion de données ne sont pas significativement supérieurs à des stratégies existantes, les performances fournies n'en sont pas moins encourageantes. L'idée d'une stratégie de fusion de données basée sur la régression logistique n'est pas à rejeter, sans doute faut-il encore affiner les modèles.

Nos réflexions sur la fusion de données montrent que la performance d'une fusion de données dépend du nombre et du choix des systèmes à combiner. De plus, les divers systèmes de repérage utilisés ici pour évaluer et tester nos modèles de fusion de données sont des modèles relativement similaires. En effet, les différences essentielles portent sur le choix de la formule d'indexation des documents et de la requête. L'utilisation de modèles radicalement différents apporterait peut-être des résultats plus encourageants. Nous laissons cette question ouverte pour des travaux ultérieurs.

CONCLUSION

Dans le cadre de cette thèse, nous avons recherché une méthodologie et une modélisation communes aux problèmes de fusion de collections et de données. Nous avons voulu que les modèles de recherche obtenus reposent sur un fondement théorique solide, et qu'ils apportent de meilleurs résultats que les stratégies existantes. Pour cela, nous avons développé une approche basée sur la régression logistique qui utilise des critères tels le rang et/ou le degré de similarité comme variables explicatives pour déterminer les probabilités de pertinence des documents retournés. La fusion, tant de collections que de données, se fait ensuite en fonction de ces probabilités estimées.

- **Fusion de collection**

A l'issue de cette recherche, nous proposons en fusion de collections deux modèles dont les performances sont attractives et significativement supérieures aux stratégies existantes. Ces deux modèles suivent la même stratégie de fusion:

- déterminer les variables explicatives du modèle;
- appliquer une régression logistique pour chacune des différentes listes de documents retournées par les systèmes;
- ordonner tous les documents selon leur probabilité de pertinence prédite par la régression;
- sélectionner les documents qui possèdent les probabilités les plus fortes.

Nos deux modèles se distinguent par le choix de la ou des variable(s) explicative(s) utilisée(s).

Le premier de ces modèles (Modèle 1) n'utilise que le rang des documents retrouvés par les divers systèmes. Plus exactement, la régression logistique prend le logarithme népérien du rang (ln(rang)) comme variable explicative du modèle pour déterminer les probabilités de pertinence des documents. Ce modèle permet une augmentation de 8.49 % de la précision moyenne par rapport à la meilleure des autres stratégies évaluées (stratégie "chacun son tour").

Le deuxième modèle (Modèle 3) utilise conjointement les informations contenues dans le rang des documents et leur degré de similarité avec la requête. Pour cela, nous définissons une nouvelle variable, nommée SimDécomp qui reflète la qualité de réponse du document. Les variables explicatives du modèle logistique sont alors ln(rang) et SimDécomp. Ce modèle permet une augmentation de 12.09 % de la précision moyenne par rapport à la stratégie "chacun son tour".

Le développement de ces deux modèles répond à tous les objectifs que nous nous étions fixés:

1. Développer un nouveau modèle théorique justifiant clairement la manière de combiner plusieurs évidences:
   - **1.1** trouver un modèle général pouvant s'appliquer aisément quelles que soient les informations disponibles pour effectuer la fusion;
   - **1.2** obtenir des paramètres stables et fiables pour notre modèle;
   - **1.3** ordonner clairement les documents retrouvés afin d'améliorer leur sélection lors de la fusion;
   - **1.4** en fusion de collections, savoir combien de documents il faut sélectionner dans chaque liste.
2. Évaluer empiriquement notre modèle et obtenir une performance significativement supérieure à celle des autres stratégies.

De plus, deux des réflexions initiales se sont avérées exactes:

2. Le rang des documents peut être un critère suffisant à lui seul pour effectuer la fusion. Cette information est souvent la seule disponible, le titre du document étant souvent la seconde information connue des systèmes;
3. De nouvelles variables calculées à partir de la similarité et du rang peuvent apporter de meilleurs résultats.

- **Fusion de données**

Nous avons voulu analyser la fusion de données sous le même angle que la fusion de collections. Pour cela, nous avons développé une stratégie basée sur la régression logistique. Afin de résoudre le problème des duplicata dans les différentes listes de documents retrouvés par les systèmes, deux modèles ont été proposés : un modèle séparé (qui estime, par la régression logistique, les probabilités de pertinence des documents pour chaque système, puis les combine pour ordonner les documents), et un modèle combiné (qui regroupe tout d'abord les diverses listes de documents, puis effectue une régression logistique sur cet ensemble).

Ces deux approches ont été testées à travers trois modèles logistiques. Un premier modèle (Modèle 1, séparé) utilise le logarithme népérien des rangs comme seule variable explicative des régressions. Les deux derniers modèles sont des modèles combinés dont l'un (Modèle 2) se sert des variables SimMax de chacun des systèmes comme variables explicatives, et l'autre (Modèle 3) utilise les variables SimMax et degré de similarité.

Ces modèles de fusion sont tous significativement meilleurs qu'une interrogation de la collection à l'aide d'un seul système. Cependant, aucune différence significative n'est apparue entre nos modèles et la stratégie classique de fusion par somme. Nos modèles sont soit équivalents (Modèle 3), soit moins bons (Modèle 1 et 2) que cette stratégie. Ces résultats décevants, par rapport à ceux de la fusion de collections, ne doivent cependant pas écarter l'approche de la régression logistique pour la fusion de données. En effet, les modèles doivent sans doute être affinés.

Nous laissons les questions ouvertes suivantes pour des travaux ultérieurs:

- existe-t-il un nombre optimal de systèmes à combiner en fusion de données;
- quels sont les systèmes les plus intéressants à combiner en fusion de données (tenir compte pour cela des propriétés des requêtes et des performances des systèmes);
- comment affiner notre modèle logistique de fusion de données;
- les coefficients de nos modèles logistiques de fusion de collections, obtenus sur la collection Wall Street Journal, sont-ils valides sur d'autres corpus?
- comment intégrer l'apprentissage fourni par des systèmes avec rétroaction dans notre modèle de fusion de collections?

ANNEXES

ANNEXE 1. 30 MOTS LES PLUS FRÉQUENTS EN FRANÇAIS

3'200'000 mots, 94'132 mots différents, Corpus OFIL et INIST				
rang	mot	effectif	fréquence	fréq. cumulée
1	de	184'249	0,0576	0,0576
2	la	100'431	0,0314	0,0890
3	la	75'103	0,0235	0,1124
4	le	70'751	0,0221	0,1345
5	à	63'572	0,0199	0,1544
6	et	62'916	0,0197	0,1741
7	les	62'517	0,0195	0,1936
8	des	59'899	0,0187	0,2123
9	d	55'257	0,0173	0,2296
10	en	45'602	0,0143	0,2438
11	du	41'688	0,0130	0,2569
12	un	37'295	0,0117	0,2685
13	une	33'301	0,0104	0,2789
14	a	28'319	0,0088	0,2878
15	est	27'954	0,0087	0,2965
16	dans	26'207	0,0082	0,3047
17	que	24'658	0,0077	0,3124
18	il	24'158	0,0075	0,3200
19	qui	23'227	0,0073	0,3272
20	pour	23'075	0,0072	0,3344
21	par	21'714	0,0068	0,3412
22	au	20'879	0,0065	0,3477
23	sur	16'598	0,0052	0,3529
24	pas	14'920	0,0047	0,3576
25	plus	13'761	0,0043	0,3619
26	qu	12'786	0,0040	0,3659
27	sur	12'745	0,0040	0,3699
28	ne	12'039	0,0038	0,3736
29	ce	12'038	0,0038	0,3774
30	se	11'973	0,0037	0,3811

tableau A.1 : 30 mots les plus fréquents extraits des corpus français OFIL et INIST

ANNEXE 2. PONDÉRATIONS

Afin de pondérer l'importance attribuée à un terme T_j dans un document D_i (ou une requête), différentes formulations sont possibles comme le montre le tableau A.2.

BNN	$w_{ij} = 1$	NNN	$w_{ij} = tf_{ij}$
ANN	$w_{ij} = 0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}$	ATN	$w_{ij} = \left[0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}\right] \cdot idf_j$
NPN	$w_{ij} = \left[0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}\right] \cdot idf_j$	LTN	$w_{ij} = \left[\ln(tf_{ij}) + 1\right] \cdot idf_j$
HTN	$w_{ij} = \frac{\ln(tf_{ij} + 1) \cdot idf_j}{\ln(nt_i)}$	OKAPI	$w_{ij} = \sum_{k=1}^{t} \frac{2 \cdot tf_{ik}}{C + tf_{ik}}$
LNC	$w_{ij} = \frac{\ln(tf_{ij}) + 1 \cdot}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	NTC	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$
ANC	$w_{ij} = \frac{0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}}{\sqrt{\sum_{k=1}^{t} \left[0.5 + 0.5 \cdot \frac{tf_{ik}}{\max tf_{i.}}\right]^2}}$	LTC	$w_{ij} = \frac{\left[\ln(tf_{ij}) + 1\right] \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left[\ln(tf_{ik}) + 1\right] \cdot idf_k\right)^2}}$
LNU	$w_{ij} = \frac{\frac{1 + \ln(tf_{ij})}{1 + \ln(mean(tf_{i.}))}}{(1-c) \cdot mean(nt_.) + c \cdot nt_i}$		

tableau A.2: Formulations de pondération

Dans cette table, la valeur tf_{ij} indique le nombre d'occurrences du terme T_j dans le document D_i, n représente le nombre de documents D_i dans la collection et df_j le nombre de documents dans lesquels le terme T_j apparaît. La mesure idf_j, inversement proportionnelle à la fréquence du terme dans la collection, se calcule comme $\ln[n/df_j]$. La longueur d'un document peut aussi jouer un rôle et se mesure par nt_i correspondant au nombre de termes d'indexation associés au document D_i. Dans nos expériences, la constante c a été fixée à 0.2 et C a été calculée comme $0.5 + 1.5 \cdot [nt_i / mean(nt_.)]$, soit le rapport entre la longueur du document D_i et la longueur moyenne de tous les documents de la collection, moyenne notée $mean(nt_.)$.

ANNEXE 3. EXEMPLES DE REQUÊTE

<top>

<num> 003

<dom> International Economics

<title> Joint Ventures

<desc> Document will announce a new joint venture involving a Japanese company.

<narr> A relevant document will announce a new joint venture and will identify the partners (one of which must be Japanese) by name, as well as the name and activity of the new company.

<con>

1. joint venture, tie up
2. partner, cooperation, joint management, agreement
3. cooperate, work together, jointly manage, jointly own, jointly produce

<fac>

<nat> Japan

<time> Current

</fac>

<def> Joint Venture - An international business undertaking defined as the commitment, for more than a very short duration, of funds, facilities, and services by two or more legally separate interests to an enterprise for their mutual benefit.

</top>

<top>

<num> 156

<title> Efforts to enact Gun Control Legislation

<desc> Documents will relate to efforts to enact Gun Control Legislation. They may also discuss implication upon and interpretation of the Second Amendment to the U.S. Constitution as it relates to these efforts.

<narr> Any discussion of Gun Control -- Pro or Con -- in editorials, letters to editor, news items, etc. This will include efforts at gun control at all political levels -- city, county, state, and federal. Of particular interest are discussions which relate these efforts to the broader issue of just what relevance does the Second Amendment have to the question. This search will be limited to U.S. efforts.

</top>

ANNEXE 4. EXEMPLE DE DOCUMENT

<DOCNO> WSJ920103-0136 </DOCNO>
<DOCID> 920103-0136. </DOCID>
<HL>
Dividend Increases
Rose in December
From Year Earlier

By Neal Lipschutz
Special to The Wall Street Journal
</HL>
<DATE> 01/03/92 </DATE>
<SO> WALL STREET JOURNAL (J), PAGE C8 </SO>
<IN> DIVIDENDS (DIV) </IN>
<NS> DIVIDENDS (DIV) </NS>
<RE>
NORTH AMERICA (NME)
UNITED STATES (US)
</RE>
<LP>
NEW YORK -- More U.S. companies raised dividends in December than in the year-earlier month. It was the second consecutive month of such positive dividend news and the third time in the final four months of 1991 that there were more dividend increases than in the year-earlier month, according to data from Standard & Poor's Corp. Until September, there had been 22 consecutive months of negative year-to-year dividend comparisons.
</LP>
<TEXT>
"December was perhaps a little better than we thought, but the comparison was pretty easy" because in December 1990 the economy was in recession and the U.S. was headed toward war in the Persian Gulf, said Arnold Kaufman, editor of S&P's Outlook newsletter. S&P said 102 companies raised their dividends in December, compared with 88 in the year-earlier month and 137 in December 1989.

Despite the better trend near the end of the year, 1991 was a dismal one for dividends. S&P said 1,086 companies raised their dividends in 1991, down from 1,263 in 1990 and 1,656 in 1989.

The 1991 dividend increase total was the lowest since 1971, when 885 companies raised dividends. But because dividends were subject to federal wage and price controls in 1970 and 1971, Mr. Kaufman said, a more valid comparison would go all the way back to 1962, which also saw 1,086 companies raising dividends.

Sixteen companies cut their dividends last month, compared with 15 a year earlier. For all of 1991, 187 companies cut dividends, compared with 143 in 1990. Mr. Kaufman said the 1991 total was the highest since 258 dividends were cut in 1982, which also was a recession year.

Eighteen companies omitted dividends in December, compared with 32 a year earlier. In all of 1991, there were 250 omissions, compared with 266 in 1990. Mr. Kaufman is fairly upbeat about the dividend outlook for 1992, though he points out that year-to-year comparisons will be made with a weak 1991. He forecasts a 3% to 5% increase in dividends of companies in the S&P 500 index and says it is likely that the 1990 total of companies raising dividends, 1,263, will be matched in 1992.

The dividend-increase forecast is linked to expectations that corporate earnings will rebound from last year's depressed levels as the benefits of lower interest rates and cost-cutting take hold, he said.

</TEXT>

</DOC>

ANNEXE 5. QUELQUES STATISTIQUES SUR LES COLLECTIONS UTILISÉES

Collection	WSJ90	WSJ91	WSJ92	WSJ
Taille	73 Mb	146 Mb	35 Mb	254 Mb
# de documents	21,705	42,652	10,163	74,520
# de topics	170	171	140	186
# doc. pertinents	2,074	3,563	831	6,468
Basé sur les termes indexés				
moyenne	142.8	138.8	134.2	139.37
écart-type	123.0	119.4	117.0	120.86
médiane	90	88	83	88
maximum	1532	1738	1425	1738
minimum	3	5	5	3
Basé sur tf_{ij}				
moyenne	246.6	239.6	232.2	240.64
écart-type	236.5	230.4	228.0	231.90
médiane	145	141	134	141
maximum	4924	5791	4343	5791
minimum	3	5	5	3

tableau A.3: Statistiques associées aux collections

Dans la collection WSJ90, les requêtes {65, 80, 81, 101, 102, 104, 139, 140, 146, 201, 213, 214, 221, 225, 227, 232, 234, 236, 252, 260, 263, 272, 277, 278, 279, 280, 281, 292, 295, 296} n'ont aucun documents pertinents, dans la collection WSJ91, les requêtes {66, 69, 80, 81, 103, 104, 105, 121, 141, 144, 146, 201, 210, 213, 214, 220, 231, 232, 236, 253, 260, 262, 263, 267, 276, 278, 279, 281, 296} ont été retirées de l'évaluation. Dans le corpus WSJ92, il s'agit des requêtes {54, 58, 63, 64, 69, 70, 77, 80, 81, 91, 93, 99, 101, 103, 104, 105, 121, 126, 127, 129, 130, 131, 133, 139, 140, 144, 146, 201, 210, 213, 214, 217, 220, 229, 232, 236, 238, 239, 252, 253, 256, 258, 262, 263, 265, 266, 267, 268, 271, 275, 276, 278, 279, 280, 281, 288, 293, 295, 296, 300}. Enfin, sur la collection entière WSJ, les requêtes {80, 81, 104, 146, 201, 213, 214, 232, 236, 263, 278, 279, 281, 296} peuvent être ignorées.

ANNEXE 6. COEFFICIENT DE CORRÉLATION LINÉAIRE DE PEARSON

Soit X et Y deux variables aléatoires, et $x=(x_1, ...,x_n)$ et $y=(y_1,...,y_n)$ deux réalisations de ces variables.

Le coefficient de corrélation linéaire de Pearson mesure la liaison linéaire entre deux variables X et Y, et se calcule comme suit:

$$\boxed{\rho = \frac{Cov_{xy}}{S_x S_y}}$$

où $Cov_{x,y}$ est la covariance observée :

$$Cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})$$

et S_x et S_y sont respectivement les écarts-types de x et y :

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})} \qquad S_y = \sqrt{\frac{1}{n}\sum_{i=t}^{n}(y_i - \overline{y})}$$

Le coefficient de corrélation linéaire est égal à ± 1 s'il y une relation linéaire entre les variables x et y. La nullité du coefficient exclut la relation linéaire.

ANNEXE 7. INTERVALLE DE CONFIANCE

•INTERVALLE DE CONFIANCE

Soit α e [0,1], on appelle intervalle de confiance d'un paramètre θ, de niveau de confiance $1-\alpha$, l'intervalle de valeur [a,b] tel que :

$$\text{Prob}(a \leq \theta \leq b) = 1 - \alpha$$

•INTERVALLE DE CONFIANCE DE LA MOYENNE :

•σ connu :

$\overline{X}$ est le meilleur estimateur de μ et $\overline{X}$ suit une loi normale $N(\mu, \sigma)$.

L'intervalle de confiance symétrique au seuil α est :

$$\overline{x} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$u_{\alpha/2}$ se lit dans la table de la loi normale. Si a=5%, alors $u_{\alpha/2}$=1.96.

•σ inconnu :

Le rapport $\frac{\sqrt{n}(\overline{X} - \mu)}{S_x}$ suit une loi de Student à (n-1) degrés de liberté, et l'intervalle de confiance symétrique au seuil α est alors :

$$\overline{x} - t_{\alpha/2}\frac{S_x}{\sqrt{n-1}} \leq \mu \leq \overline{x} + t_{\alpha/2}\frac{S_x}{\sqrt{n-1}}$$

$t_{\alpha/2}$ se lit dans la table de la loi de Student.

R :Q. : Le théorème central-limite a pour conséquence que les intervalles précédents sont valables pour estimer l'espérance d'une loi quelconque lorsque *n* est grand.

ANNEXE 8. HISTOGRAMMES DES VARIABLES POUR WSJ90-OKAPI

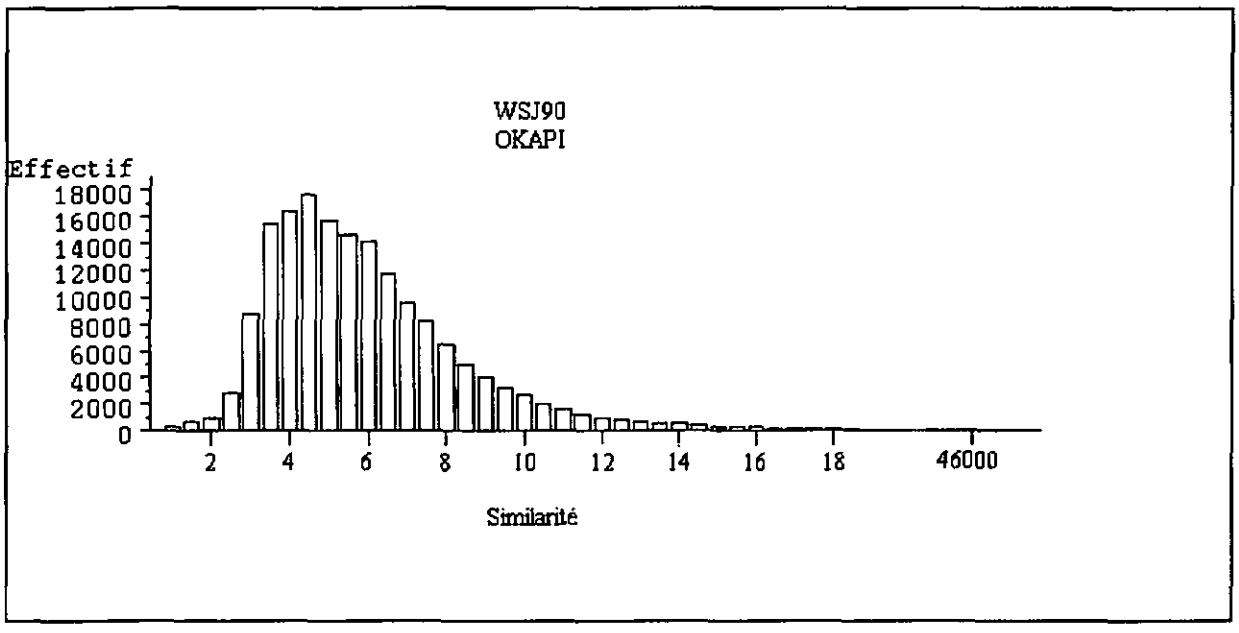

figure A.1 : Histogramme de la variable Sim (degré de similarité) (WSJ90-OKAPI, 170 requêtes)

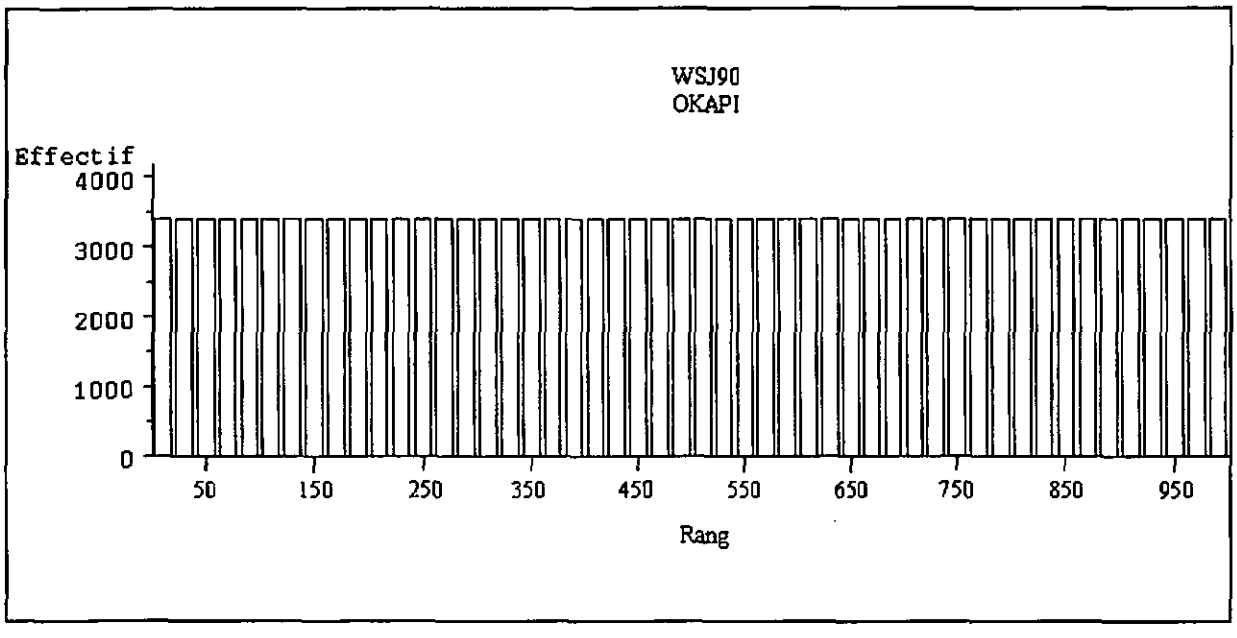

figure A.2 : Histogramme de la variable rang (WSJ90-OKAPI, 170 requêtes)

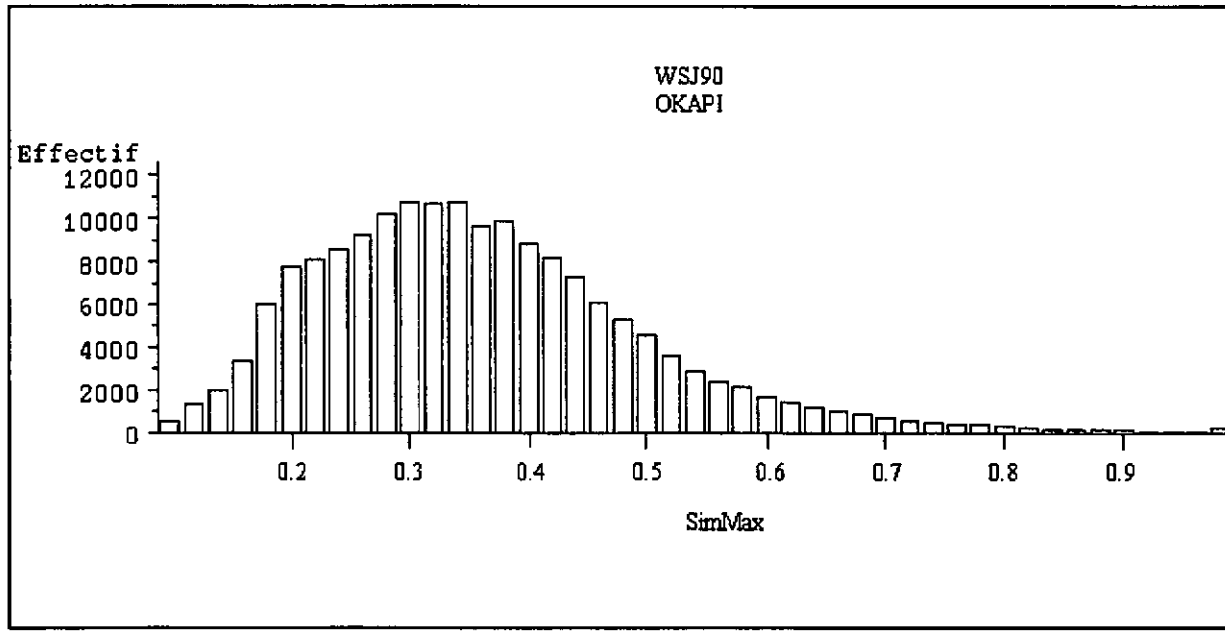

figure A.3: Histogramme de la variable SimMax (WSJ90-OKAPI, 170 requêtes)

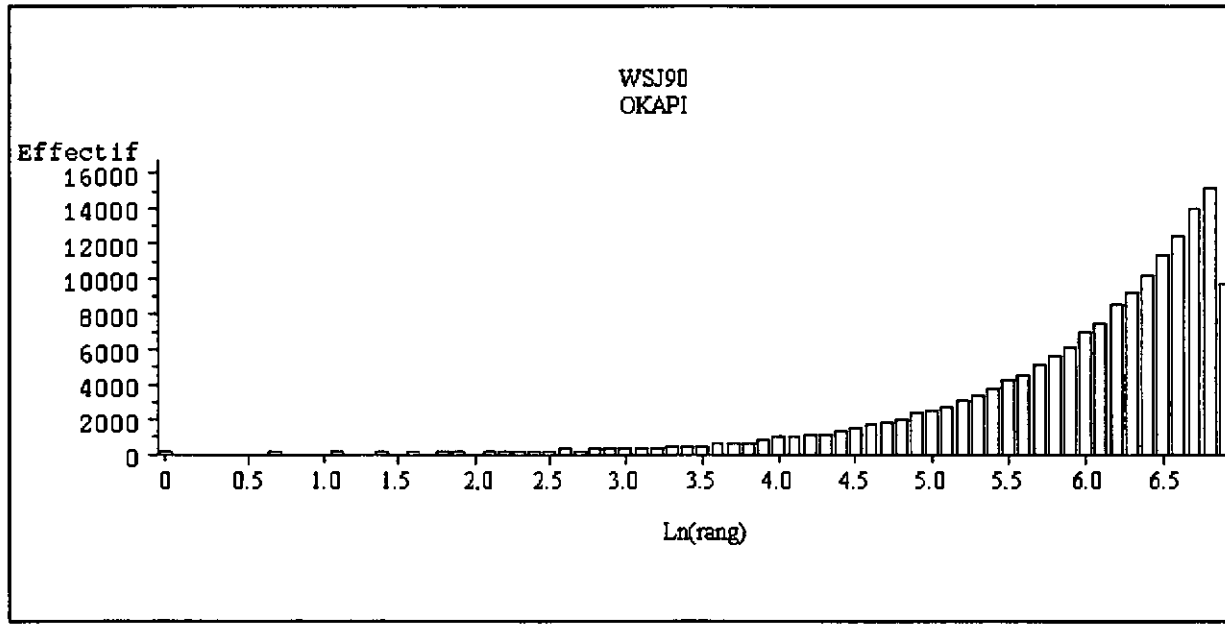

figure A.4: Histogramme de la variable ln(rang) (WSJ90-OKAPI, 170 requêtes)

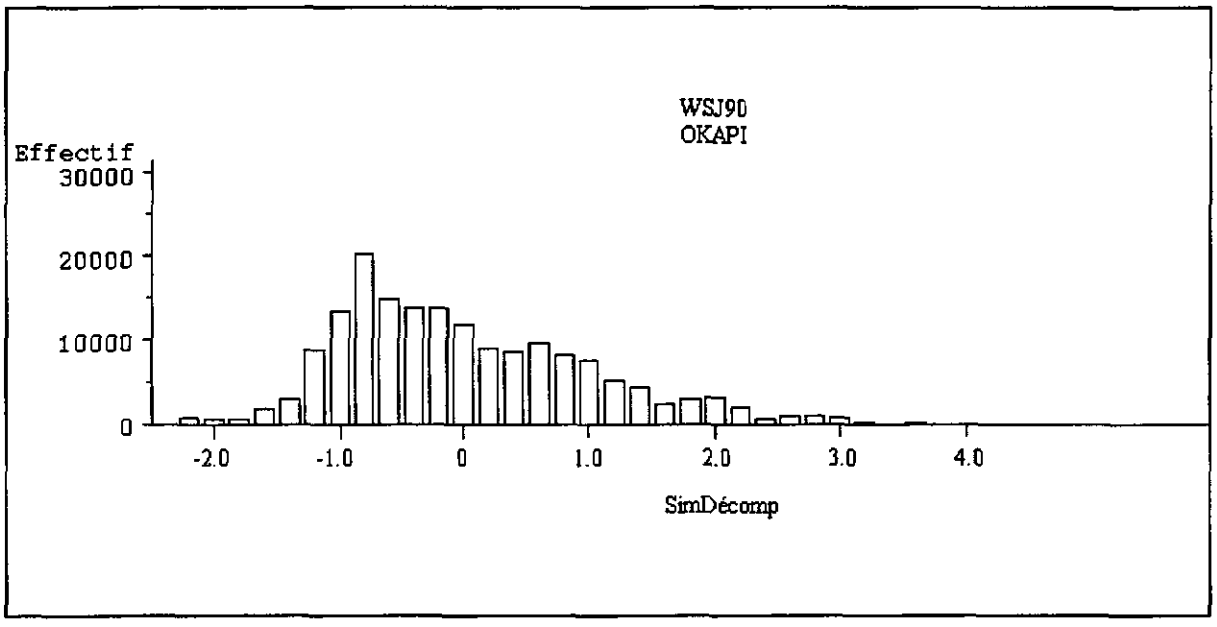

figure A.5 : Histogramme de la variable SimDécomp (WSJ90-OKAPI, 170 requêtes)

BIBLIOGRAPHIE

[Belkin 95] Belkin N. J., Kantor P., Fox E. A., Shaw J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3), 431-448.

[Bookstein 92] Bookstein A., O'Neil E., Dillon M. & Stephens D. (1992). Applications of loglinear models for informetric phenomena. *Information Processing & Management*, 28(1), 75-88.

[Buckley 96] Buckley C., Singhal A., Mitra M. & Salton G. (1996, October). *New retrieval approaches using SMART.* Proceedings of the TREC'4, Gaithersburg, MD, NIST publication 500-236, 25-48.

[Callan 95] Callan J. P., Lu Z. & Croft W. B. (1995, June). *Searching distributed collections with inference networks.* Proceedings of the 18th International Conference of the ACM-SIGIR'95, Seattle, WA, 21-28.

[Cooper 92] Cooper W., Gey F. C., Gabney D. P. (1992, June). *Probabilistic retrieval based on staged logistic regression.* Proceedings ACM-SIGIR'92 ACM-SIGIR'92, Copenhagen, 198-210.

[Cox 89] Cox D. R. & Snell E. J. (1989). *Analysis of binary data.* 2nd ed., London, UK: Chapman and Hall.

[Dumais 94] Dumais S. T. (1994, March). *Latent semantic indexing (LSI) and TREC-2.* Proceedings of TREC'2, Gaithersburg, MD, NIST Publication #500-215, 105-115.

[Fuhr 94] Fuhr N. & Pfeifer U. (1994). *Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions.* ACM Transactions on Information Systems, 12(1), 92-115.

[Gey 93] Gey F. C. (1993). *Probabilistic Dependence and Logistic Inference in Information Retrieval.* Ph. D. Thesis, University of California at Berkeley.

[Gey 94] Gey F. C. (1994, July). *Inferring probability of relevance using the method of logistic regression.* Proceedings of the 17th International Conference of the ACM-SIGIR'94, Dublin, Ireland, 222-231.

[Halasz 88] Halasz F. G. (1998) : Reflections on NoteCards : seven issues for the next generation of hypermedia systems. *Communications of the ACM*, 31(7), 836-852.

[Hosmer 89] Hosmer D. & Lemeshow S. (1989). *Applied logistic regression.* New-York, NY: John Wiley & Sons.

[Kwok 95] Kwok K. L., Grunfeld L. & Lewis D. D. (1995, April). *TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS.* Proceedings of TREC'3, Gaithersburg, MD, NIST Publication #500-225, 247-255.

[Lee 95] Lee J. H. (1995). *Combining multiple evidence from different properties of weighting schemes.* Proceedings of the 18th International Conference of the ACM-SIGIR'95, Seattle, WA, , 180-188.

[Moffat 95] Moffat A. & Zobel J. (1995, April). *Information retrieval systems for large document collections.* Proceedings of TREC'3, Gaithersburg, MD, NIST Publication #500-225, 85-93.

[van Rijsbergen 79] van Rijsbergen C. J. (1979). *Information Retrieval.* Butterworths, 2nd ed., London.

[Robertson 77] Robertson S. E. (1997). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294-304.

[Robertson 95] Robertson S. E., Walker S. & Hancock-Beaulieu M. M. (1995). Large test collection experiments on an operational, interactive system: OKAPI at TREC. *Information Processing & Management*, 31(3), 345-360.

[Salton 69] Salton G. (1969). A comparison between manual and automatic indexing methods. *American Documentation.* 20(1), 61-71.

[Salton 71] G. Salton (1971). *The SMART retrieval system - experiments in automatic document processing.* Prentice-Hall Inc., Englewood Cliffs (NJ).

[Salton 72] Salton G. (1972). A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*. 23(2), 74-84.

[Sparck Jones 77] Sparck Jones K. & Bates R. G. (1977). *Research on automatic indexing 1974-1976.* Technical Report, Computer Laboratory, University of Cambridge (UK).

[Saracevic 88] Saracevic T., Kantor P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. *Journal of the American Society for Information Science*, 33(4), 495-512.

[Savoy 96] Savoy J., Ndarugendamwo M., Vrajitoru D. (1996, October). *Report on the TREC-4 Experiment: combining probabilistic and vector-space schemes.* Proceedings of TREC'4, Gaithersburg, MD, NIST Publication 500-236, 537-547.

[Savoy 97] Savoy J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 31(3), 495-512.

[SAS v6] SAS Institute . *SAS/STAT User's Guide, Volume 2, GLM-VARCOMP.* SAS Institute Inc. , Version 6, Fourth edition, 1071-1126.

[Shaw 95] Shaw J. A., Fox E. A. (1995). *Combination of multiple searches.* Proceedings of TREC'3, Gaithersburg, MD, NIST Publication #500-225, 105-108.

[Stone 74] Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B,* 36(2),111-147.

[Turtle 91] Turtle H., Croft W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), 187-222.

[Voorhees 95a] Voorhees E. M., Gupta N. K. & Johnson-Laird B. (1995, April). *The collection fusion problem.* Proceedings of TREC'3, Gaithersburg, MD, NIST Publication #500-225, 95-104.

[Voorhees 95b] Voorhees E. M., Gupta N. K. & Johnson-Laird B. (1995, July). *Learning collection fusion strategies.* Proceedings of the 18th International Conference of the ACM-SIGIR'95, Seattle, WA, 172-179.

[Voorhees 96] Voorhees E. M. (1996, October). *Siemens TREC-4 report: Further experiments with database merging.* Proceedings of TREC'4, Gaithersburg, MD, NIST Publication 500-236, 121-130.

Anne LE CALVE
Institut interfacultaire d'informatique
Pierre à Mazel 7
2000 NEUCHÂTEL
SUISSE
Tél. (+41 32) 718 13 73
e-mail : Anne.Lecalve@seco.unine.ch

Née le 07-10-1970
Nationalité : Française

Diplômes :

1995 - 1994 : Maîtrise M.A.S.S. (Mathématiques appliquées et Sciences Sociales) à l'Université de Haute-Bretagne, Rennes, France.

1994 - 1993 : Licence M.A.S.S. à l'Université de Haute-Bretagne, Rennes, France.

1991 - 1990 : Diplôme Universitaire de Technologie en Statistique et Traitement Informatique des Données, à l'I.U.T. de Vannes, France.

Publications :

Le Calvé A., Savoy J. (1997, Avril). *Un nouveau modèle de recherche d'informations pour INTERNET.* Actes JST97 FRANCIL 1997, Avignon, France, 99-103.

Savoy J., Le Calvé A., Vrajitoru D. (1997, November) *Report on the TREC-5 Experiment: Data Fusion and Collection Fusion.* Proceedings TREC'5, Gaithersburg (MD), NIST Publication 500-238, 489-502.

Publications internes :

Le Calvé A., Savoy J. (1998, Juin). *Database merging strategy based on logistic regression.* Université de Neuchâtel, (à paraître).

Savoy J., Le Calvé A. (1997, Avril). *Rapport d'expérience AMARYLLIS 1997.* Université de Neuchâtel, CR-I-97-03.

Savoy J., Le Calvé A. (1996, Juin). *Recherche d'informations dans un environnement distribué.* Université de Neuchâtel, CR-I-96-01.

Le Calvé A., Savoy J. (1996, Octobre). *Data Fusion and Collection Fusion.* Université de Neuchâtel, CR-I-96-03.