Indexation multimédia

Caractérisation du déséquilibre entre les modalités texte et parole

Mémoire - DEA Informatique

Benoit Favre
Sous la direction de Jean-François Bonastre et Patrice Bellot
Laboratoire d'Informatique d'Avignon - Université d'Avignon

20 juin 2003

Résumé

La recherche d'information multimédia est un thème émergeant notamment grâce à des technologies de caractérisation du contenu (reconnaissance de la parole, d'image, ...). Nous abordons cette problématique à travers la recherche documentaire appliquée à un corpus mélangeant des documents textuels et des transcriptions automatiques de parole.

En étudiant le comportement des méthodes classiques pour la recherche documentaire (modèles vectoriel et probabiliste), nous faisons apparaître une évolution du besoin en information multimédia : la couverture en modalités (ici texte et parole) est à considérer conjointement à la précision du résultat d'une recherche.

Nous caractérisons le déséquilibre quantitatif et qualitatif des données impliqué par ce besoin en étudiant le pouvoir discriminant des mots sur les documents défini par le concept d'*inverse document frequency* (*idf*). Différentes perspectives issues de ces résultats sont présentées. Nous décrivons plus précisément une nouvelle voie d'investigation pour l'équilibrage des modalités au cours de l'expansion de requête.

1 Introduction

La quantité d'information rendue disponible par les réseaux croit fortement chaque jour. Cette information représente une grande richesse dès lors qu'elle est structurée et accessible. L'indexation et la recherche d'information sont devenues des tâches primordiales pour réaliser ces objectifs. De fortes avancées en recherche d'information textuelle ont été observées dans les dernières années. L'apparition de nombreux documents multimédia, l'augmentation des capacités, des débits et de la puissance de calcul vont de pair avec l'émergence d'un besoin de recherche documentaire multimédia apportant de nouvelles problématiques.

L'information multimédia est formée d'images, de vidéos et de bandes audio en plus du texte. Alors que l'extraction d'informations de haut niveau à partir des images fixes et animées n'en est qu'à ses débuts, l'utilisation de moteurs de reconnaissance de la parole pour transcrire et indexer les bandes sonores est suffisamment développée pour obtenir de bonnes performances en recherche documentaire audio. L'indexation de la parole a été très étudiée, notamment lors des campagnes d'évaluation NIST (*National Institute for Science and Technology*) *Spoken Document Retrieval* pour lesquelles elle a été un succès [Garofolo 2000].

Nous proposons d'étudier le comportement des méthodes classiques de recherche documentaire sur des corpus multimédia : mélange de transcriptions de *parole* et documents *textuels*. Plus précisément, nous montrons qu'il existe un déséquilibre sur la quantité de données et les sujets abordés dans ces deux modalités et que le besoin d'information, exprimé par l'utilisateur, évolue pour faire apparaître l'importance de la couverture en modalités des résultats d'une recherche. Un compromis entre couverture en modalités et précision des résultats devient nécessaire pour satisfaire cette évolution du besoin. Nous proposons pour répondre à cette problématique différentes pistes visant à l'équilibrage en modalités, notamment dans l'expansion de requête en aveugle.

Dans la première partie de ce mémoire, nous revenons sur les principaux concepts de la recherche documentaire classique, qui dirigent notre analyse de l'information multimédia. Nous mettons l'accent sur le modèle vectoriel et le modèle probabiliste, servant de base à nos travaux. Dans une seconde partie, nous caractérisons les différences fondamentales entre l'information textuelle et les informations multimédia parole et texte. A travers cette caractérisation, un besoin spécifique émerge, correspondant à l'équilibrage en modalités des résultats de la recherche documentaire. En application à cette étude, nous envisageons la perspective d'une méthode d'équilibrage en modalités dans la troisième partie de notre travail. Fondée sur le modèle probabiliste, elle s'insère dans l'expansion de requête en aveugle. La méthode proposée permet d'utiliser autant de documents pseudo-pertinents dans chaque modalité et de fixer *a priori* la probabilité pour que les documents provenant d'une modalité soient pertinents. Enfin, le document est suivi d'une conclusion synthétisant l'ensemble des résultats auxquels nous sommes parvenus.

2 La recherche documentaire

Dans cette partie, nous présentons l'ensemble des outils qui nous sont nécessaires pour étudier la recherche documentaire sur des données mélangées issues du texte et de la parole.

2.1 Notion d'information

La recherche documentaire relève d'une problématique mal définie et difficile : répondre au besoin en information d'un utilisateur. Ce besoin est exprimé à travers une requête intrinsèquement incomplète et approximative. L'information est contenue dans des documents qui répondent plus ou moins à la requête de l'utilisateur. Un document est "pertinent" à une requête donnée s'il a été *aimé*[1] par l'utilisateur, s'il répond totalement ou en partie à son besoin en information. Il est dénoté "non pertinent" dans le cas contraire.

Posons les différents éléments sur lesquels s'appuie la recherche documentaire. Un *document* est défini comme un regroupement d'*informations* à propos d'un ou plusieurs *sujet*. Il est bon de noter qu'il est difficile de définir où s'arrête un *sujet*, c'est pourquoi un *document* peut parler de plusieurs *sujets*. Sont extraits des documents des *attributs* qui permettent de déterminer leurs *sujets* et les *informations* qu'ils contiennent. Pour la recherche documentaire classique (textuelle) les *attributs* sont appelés *termes* ou *concepts*, extraits du langage et le plus souvent représentés par les mots eux-mêmes.

En général, les systèmes de recherche documentaire répondent à une requête par une liste de documents classés, par pertinence décroissante. L'utilisateur a une plus forte probabilité de trouver une réponse à son besoin en information dans les premiers éléments de cette liste que dans la suite.

2.2 Evaluation

La meilleure méthode pour évaluer la qualité des réponses au besoin de l'utilisateur est de demander s'il a *aimé* ou non les documents trouvés. Nous voyons dans cette section l'évaluation de la qualité de la réponse au besoin en information d'un utilisateur.

2.2.1 La mesure précision/rappel

L'idéal pour pouvoir comparer plusieurs systèmes de recherche documentaire est de connaître, sur une collection de documents donnée, quels sont les documents pertinents à chaque requête. Connaître cet ensemble est un problème récurrent car la collection de documents est en général grande et il n'est pas possible de juger manuellement tous les documents[2]. A partir de

[1] Le verbe *aimer* est choisi pour exprimer la satisfaction de l'utilisateur car il définit en anglais l'événement L pour *Liked* utilisé dans la formulation des modèles en recherche documentaire.

[2] Swanson explique dans [Swanson 1997] que les premiers jugements de pertinence faits dans les années cinquante avaient abouti à 70% d'accord entre les juges et que ce problème restera caractéristique de la recherche documentaire quelles qu'en soient les évolutions :

cet ensemble de documents, deux mesures de la qualité d'une recherche sont définies :

- La *précision* est le nombre de documents pertinents retrouvés par rapport au nombre de documents retrouvés ($Precision = \frac{|P \cap R|}{|R|}$, où P est l'ensemble des documents pertinents, R l'ensemble des documents retrouvés, $\cap$ l'intersection ensembliste, $|X|$ le cardinal de l'ensemble X) ;
- Le *rappel* est le nombre de documents pertinents retrouvés par rapport au nombre de documents pertinents dans la collection ($Rappel = \frac{|P \cap R|}{|P|}$).

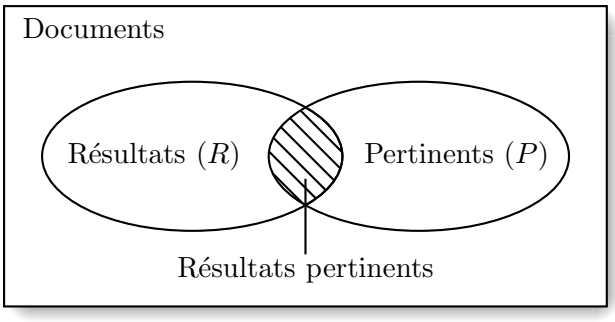


FIG. 2.1 – Répartition des documents dans la réponse à une requête.

Une courbe de précision/rappel, issue de l'analyse de la liste de résultats, permet d'évaluer la qualité de la réponse à une requête donnée. Cette courbe est majoritairement utilisée mais elle est aussi très contestée [Raghavan 1989] car l'utilisateur n'est intéressé que par les premiers documents du classement (les plus pertinents) et une mesure de précision à x documents peut être plus représentative. Une mesure générale est nécessaire pour comparer les systèmes sur des ensembles de requêtes et obtenir une valeur unique. La précision moyenne extrapolée sur 11 points de rappel (0% de rappel, 10%, ..., 100%), moyennée sur l'ensemble des requêtes est couramment utilisée.

Ces outils sont utilisés pour déterminer si une méthode de recherche documentaire améliore de façon significative une autre méthode sur un corpus de documents et des requêtes données. L'accent est mis dans [Hull 1993] sur les tests statistiques qui permettent de caractériser une amélioration statistiquement significative. Ces tests ne sont pas très répandus dans la littérature et les auteurs s'accordent en général, comme dans [Bellot 2003], sur le fait qu'une amélioration de cinq points est significative sans préciser sur quelles données ni quelles requêtes. Hull note aussi que les résultats requête par

chaque utilisateur est satisfait de manière différente.

requête peuvent énormément varier selon la méthode de recherche documentaire utilisée. Il est important d'étudier soigneusement les résultats d'une évaluation pour comprendre pourquoi certaines requêtes sont dégradées et d'autres améliorées alors que la précision moyenne semble augmenter.

2.2.2 La piste *Adhoc* des campagnes d'évaluation TREC

La plus grande et la plus connue des campagnes d'évaluation en recherche d'information est la campagne TREC organisée par le NIST et DARPA (*Defense Advanced Research Project Agency*). Les pistes évaluées par TREC sont nombreuses. Celles qui se rapprochent le plus de notre étude sont la piste *Adhoc* et la piste *Spoken Document Retrieval* (*SDR*). En effet, elles sont similaires dans le sens où les requêtes sont dans les deux cas du type *topic matching*, à savoir trouver les documents qui se rapportent au même *sujet* que celui de la requête[3].

Piste originelle de TREC, la piste *Adhoc* évalue des systèmes de recherche documentaire sur leur capacité à retrouver les documents textuels parlant du même sujet que la requête. Les requêtes sont subdivisées en plusieurs parties : un *titre*, une *description* et une *narration*. Nous ne nous intéressons qu'à la partie *description* des requêtes.

2.2.3 La piste *SDR* de TREC

Il est intéressant de suivre le tour d'horizon de la piste *SDR* proposé dans [Garofolo 2000] pour bien situer le départ de nos travaux.

La recherche documentaire audio ou recherche documentaire sur de la parole (*Spoken Document Retrieval*), a été introduite dans les évaluations TREC sous forme d'une recherche documentaire sur des informations corrompues (contenant des erreurs dues à la transcription automatique).

En effet, des techniques similaires à celles mises en œuvre pour la tâche *Adhoc* sont appliquées pour faire de la recherche documentaire audio, mais appliquées aux transcriptions générées par les moteurs de reconnaissance de la parole. Ces transcriptions contiennent des erreurs provoquées par la suppression, le remplacement ou l'ajout de mots. La qualité des transcriptions est mesurée par le taux d'erreurs sur les mots, *Word Error Rate* (WER).

Les différentes évaluations TREC ont montré que le taux d'erreur dans les transcriptions n'affectait pas trop les performances en recherche documentaire (une chute de 10% de la précision pour un WER de 25% et une chute de 15% pour un WER de 50% [Allan 2002], sachant que les performances aux récentes évaluations *Rich Transcription 2003* (RT'03) sont aux

[3]Se reporter à [Voorhees 1999] pour une présentation de la campagne TREC.

alentours d'un WER de 10,5%). En revenche, il existe certains cas où ces résultats ne sont plus aussi optimistes, notamment lors de l'utilisation de requêtes courtes. De plus, les requêtes utilisées dans cette piste sont des requêtes textuelles qui ne sont pas soumises à la reconnaissance de la parole pour limiter la dégradation de l'expression du besoin d'information. Cette piste introduit un certain nombre de différences peu exploitées par rapport à la piste *Adhoc*.

Caractéristiques spécifiques de la piste *SDR* : La reconnaissance de la parole apporte des informations en supplément de la transcription textuelle qui pourraient améliorer la recherche d'information comme pour [Allan 2002] :

- les scores des mots (aux niveaux acoustique et linguistique) ;
- l'utilisation d'hypothèses multiples ;
- la prosodie ;
- l'identification du locuteur ;
- l'harmonisation des modèles de langages utilisés dans la reconnaissance de la parole et dans la recherche documentaire.

Nous n'étudions pas ces caractéristiques dans le cadre de ce mémoire mais il est important de noter qu'elles ont une incidence sur la recherche documentaire audio.

Problématiques ouvertes de la piste *SDR* : La recherche d'information audio a été estimée "résolue" après la campagne d'évaluation TREC-9 par [Garofolo 2000] estimant qu'il serait judicieux de passer plus de temps sur des problèmes moins évidents. Mais [Allan 2002] liste les possibilités qu'il faudrait explorer :

- adapter l'indexation à des requêtes courtes ;
- créer des interfaces utilisateur encourageant à des requêtes plus longues ;
- aider l'utilisateur à reconnaître les documents pertinents à sa requête ;
- résumer la parole pour discriminer les réponses d'un moteur de recherche comme c'est le cas pour le texte ;
- adapter les modèles de langage utilisés dans la reconnaissance de la parole au corpus indexé ;
- faire des recherches sur des documents textuels et audio mélangés ;
- indexer des documents audio de type dialogue, réunion ou cours ;
- adapter l'indexation à des environnements à fort taux d'erreur (50%) ;
- créer des interfaces textuelles à des systèmes audio (*voicemail*).

L'énonciation de ces pistes exploratoires nous permet de situer notre travail et d'en limiter le cadre à celui défini dans [Sanderson 2002] : **faire des recherches sur des documents textuels et audio mélangés**.

2.3 Modèles

Les modèles que nous utilisons lors de notre étude de l'information multimédia sont présentés dans cette section.

2.3.1 Le modèle vectoriel

Les *attributs* utilisés le plus fréquemment en indexation sont les *termes* définis par les mots des documents après filtrage des mots à valeur sémantique faible (*Stop word stripping*) et suppression des suffixes (*Stemming*) pour relier mots et concepts sous-jacents. Le modèle vectoriel est l'un des modèles les plus utilisés car il est simple et efficace. Dans ce modèle, les documents sont représentés dans un espace $\mathcal{D}$ dont les dimensions sont les *attributs* $a_i \in \mathcal{A}$ qui les composent. L'espace $\mathcal{D}$ a autant de dimensions qu'il existe d'*attributs* différents.

$$\vec{d_j} \in \mathcal{D}, \qquad \vec{d_j} = (w_{1,j}, ..., w_{n,j}) \qquad \text{où} \qquad n = card(\mathcal{A}) \tag{1}$$

Les $w_{i,j}$ sont les poids associés à chacuns des attributs a_i pour le document représenté par le vecteur $\vec{d_j}$. En général, un document n'est lié qu'à un petit nombre d'attributs, donc la plupart des composantes du vecteur $\vec{d_j}$ sont nulles. Les requêtes sont représentées dans ce même espace selon les attributs qui permettent de les qualifier.

$$\vec{q} \in \mathcal{D}, \qquad \vec{q} = (w_{1,q}, ..., w_{n,q}) \tag{2}$$

Il a été prouvé qu'il existe une distorsion entre les espaces [Cui 2002] de représentation des documents et des requêtes. Nous faisons l'hypothèse implicite dans la littérature que cette distortion est négligeable pour pouvoir travailler dans un espace commun. Les documents sont classés selon leur similarité à une requête. La similarité *cosine* est utilisée fréquemment, définie par le cosinus de l'angle entre le vecteur document $\vec{d_j}$ et celui de la requête $\vec{q}$.

$$s(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}||\vec{q}|} \tag{3}$$

où $|\vec{x}|$ représente la norme de $\vec{x}$ et $\cdot$ est le produit scalaire. On en déduit :

$$s(\vec{d_j}, \vec{q}) = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \tag{4}$$

Il existe differentes façons de pondérer les *attributs* dans les documents[4].

[4]Voir [Bellot 2003] pour un tour d'horizon des différentes pondérations dans le modèle vectoriel.

Si ces *attributs* sont des *termes*, la pondération $tf \times idf$ (*term frequency* $\times$ *inverse document frequency*) est utilisée. Elle peut prendre la forme suivante :

$$w_{i,j} = tf_{i,j} idf_i = \log(tf_{i,j} + 1) \log \frac{N}{n_i} \tag{5}$$

où $tf_{i,j}$ est le nombre d'occurrences de l'*attribut* a_i dans le document $\vec{d_j}$, N est le nombre de documents de la collection et n_i est le nombre de documents dans lequel l'*attribut* a_i apparaît. Cette pondération, appelée LTC[5], est facilement interprétable. En effet, le tf représente l'importance d'un *attribut* dans un document alors que l'*idf* représente son pouvoir discriminant par rapport aux autres termes.

Ce modèle est très simple, rapide et offre des capacités au moins aussi bonnes que les autres modèles dans la plupart des circonstances.

2.3.2 Le modèle probabiliste[6]

Le modèle probabiliste a été introduit par Robertson et Spärck Jones. Ses développements sont énoncés dans [Spärck Jones 1998]. La présentation se trouvant dans cette section se veut rapide, une présentation complète est donnée en annexe. Ce modèle est fondé sur le besoin de l'utilisateur qui souhaite retrouver tous les documents pertinents pour sa requête. Le modèle fait donc l'hypothèse que l'ensemble des documents est partitionnable en deux sous ensembles : les documents pertinents et les documents non pertinents. Le modèle cherche à déterminer si un document (associé à l'événement D_j) est pertinent (événement noté L, comme *Liked*) dans le cadre d'une requête. La règle de décision appliquée peut être vue comme une fonction de classement :

$$score(D_j) = \frac{P(L|D_j)}{P(\overline{L}|D_j)} \tag{6}$$

où $P(L|D_j)$ est la probabilité que l'utilisateur *aime* le document D_j et $P(\overline{L}|D_j)$ est la probabilité qu'il ne l'*aime* pas. Le théorème de Bayes, nous permet de réécrire les probabilités conditionnelles :

$$score(D_j) = \frac{P(D_j|L)P(L)}{P(D_j|\overline{L})P(\overline{L})} \tag{7}$$

D_j peut être représenté par les attributs a_i qui le composent. Pour simplifier les calculs, on suppose que les attributs sont indépendants. Cette

[5] La première lettre représente la transformation appliquée au *tf*, $\log(x+1)$, la seconde le facteur multiplicatif appliqué, $tf \cdot idf$ et la troisième la normalisation, *cosine*.

[6] appelé *Binaray Independence Retrival* (BIR) puis *linked dependency assumption* dans [Cooper 1995] après quelques corrections.

hypothèse n'est pas forcément justifiée mais elle permet de réduire la complexité du modèle. [Van Rijsbergen 1979] propose des moyens pour contourner cette hypothèse. Soit A_i, l'événement associé à un *attribut* a_i :

$$score(D_j) = \frac{\prod_i P(A_i|L)}{\prod_i P(A_i|\overline{L})} \frac{P(L)}{P(\overline{L})} \tag{8}$$

Cette fonction de classement est réécrite dans le domaine logarithmique en supprimant les constantes[7] invariantes dans le contexte d'une requête donnée :

$$score_{log}(D_j) = \sum_{a_i \in D_j} weight_{a_i} \qquad \text{avec} \qquad weight_{a_i} = \log \frac{P_i(1-\overline{P_i})}{\overline{P_i}(1-P_i)} \tag{9}$$

où P_i est la probabilité que l'attribut a_i soit présent dans le document lorsque ce dernier est *aimé* par l'utilisateur, et $\overline{P_i}$ est la probabilité que l'attribut soit présent lorsque le document n'est pas *aimé*. Le problème de cette formulation est que le poids des *attributs* dans un document n'est pas pris en compte. Par exemple, quand les *attributs* sont les *termes* des documents, plus un *terme* apparaît souvent dans un document, plus ce dernier a des chances de parler du *sujet* relié à ce terme. Cette notion introduite par [Robertson 1997] est appellée "élitisme" d'un *attribut*. Bien que compliquée dans la théorie, elle peut être utilisée dans la formulation simplifiée suivante :

$$weight\_elite_{a_i} = \frac{freq_{a_i,D_j}(k_1+1)}{K + freq_{a_i,D_j}} weight_{a_i} \tag{10}$$

$$K = k_1((1-b) + b\frac{length(D_j)}{\frac{1}{N}\sum_k length(D_k)}) \tag{11}$$

$freq_{a_i}$ correspond à l'importance attribuée à A_i (soit le tf du modèle précédent lorsque les attributs sont des *termes*). k_1 et b sont des constantes ajustables pour lesquelles le modèle ne préconise pas de valeur. $weight_{a_i}$ peut être estimé selon un processus itératif :

$$weight_{a_i} = idf_i = \log \frac{N}{n_i} \quad \text{à la première itération ;} \tag{12}$$

$$weight_{a_i} = \log \frac{(r_i+0.5)(N-R-n_i+r_i+0.5)}{(n_i-r_i+0.5)(R-r_i+0.5)} \quad \text{avec un } a \textit{ priori} \text{ sur la pertinence des documents ;} \tag{13}$$

où N, R, n_i, r_i sont respectivement le nombre de documents de la collection, le nombre de documents connus comme étant pertinents, le nombre de documents contenant l'*attribut* a_i, et le nombre de documents pertinents contenant l'*attribut* a_i.

[7] $\frac{P(L)}{P(\overline{L})}$ est identique pour tous les documents.

L'ensemble des documents pertinents est constitué de façon itérative. C'est l'utilisateur, ou un processus en aveugle, qui déterminent la partie de l'ensemble des documents pertinents servant à l'itération suivante de la recherche. [Van Rijsbergen 1979] propose même un seuil pour délimiter cet ensemble. Ce seuil étant rarement utilisé, nous ne le verrons pas dans ce mémoire.

Finalement, le modèle probabiliste et le modèle vectoriel sont très proches dans leur formulation mais très éloignés dans la théorie. Malgré de bonnes performances, l'hypothèse de l'indépendance des *attributs* est difficilement justifiable. Ce modèle sert de base à nos travaux car il est relativement facile d'introduire de nouveaux éléments, liés aux modalités, dans la formulation probabiliste. En effet, nous développons dans ce mémoire les perspectives d'une méthode d'équilibrage en modalités durant le processus d'expansion de requête du modèle probabiliste.

3 Étude de l'information multimédia : parole et texte

Nous présentons dans cette partie l'étude de l'indexation conjointe de textes et de transcriptions de parole, en utilisant les méthodes classiques en recherche d'information. Cette étude montre que le besoin en information évolue lorsque les données proviennent de plusieurs modalités et qu'il en résulte un déséquilibre entre modalités dans les résultats d'une requête.

3.1 Les modalités parole et texte

Cette section présente succinctement les spécifités des données multimédia *texte* et *parole*.

3.1.1 Un corpus multimodal

Nous avons vu dans la section 2.1 que l'*information* était répartie dans des *documents* caractérisés par des *attributs*. Pour la recherche documentaire multimédia, nous ajoutons le concept de *modalité*. Une *modalité* est un moyen d'exprimer l'information comme le texte, l'image [Smeaton 2002] et la parole [Garofolo 2000]. Les *documents* peuvent réunir plusieurs modalités sur un ou plusieurs sujets. Un document est toujours représenté par des *attributs* mais ces derniers appartiennent à l'une où l'autre des modalités.

A partir de ces éléments, les différents types de recherche multimédia sont déterminés selon la quantité de modalités adressées et les relations qui les lient.

La recherche unimodale n'adresse qu'une modalité. Historiquement, le texte est la modalité la plus étudiée. Mais, depuis quelques années, il est possible de faire des recherches sur la parole, l'image et la musique. Il se peut que la modalité dans laquelle est formulée la requête ne soit pas la même que celle des documents. Dans ce cas, il existe un espace commun de représentation des attributs qui permet les recherches.

Le deuxième type de recherche est la recherche en documents multimodaux. Par exemple un ensemble d'images annotées par du texte [Srihari 2000b, Srihari 2000a] constitue un corpus où chaque document représente la même information selon plusieurs modalités.

La recherche se fait sur un corpus multimodal lorsque les documents peuvent être de modalités différentes mais qu'ils ne contiennent pas plusieurs modalités. Nous prendrons comme exemple un corpus contenant du texte et des transcriptions de discours audio. Il faut noter que la problématique est différente selon que les attributs disposent ou non d'un espace de représentation commun (les mots dans notre exemple) pour les modalités.

Le dernier type de recherche est la recherche multimédia qui combine toutes les caractéristiques des autres types : des documents multimodaux dans un corpus multimodal et ne disposant pas forcément d'espace de représentation commun des attributs[8].

Bien qu'ayant mené la réflexion sur la recherche d'information de façon générique, nous nous concentrons dans ce travail sur la recherche dans un corpus multimodal adressant les modalités *parole* et *texte*, dans un espace de représentation commun des attributs. Les requêtes étudiées sont uniquement dans la modalité *texte*.

3.1.2 Présentation des données

Pour étudier la recherche documentaire sur corpus multimodal, nous avons réuni une source de données dans chacune des modalités. La langue étudiée est l'anglais et les données proviennent des évaluations TREC.

Pour le texte, TREC propose 6,4 Go d'articles journalistiques. Nous avons utilisé le sous-ensemble de données et les requêtes correspondant à TREC-8, dernière évaluation de la piste *AdHoc*. Pour l'audio, TREC ne propose que 500 heures de transcriptions de journaux radiophoniques, manuelles et automatiques selon plusieurs taux d'erreur. Un récapitulatif de la composition des collections de documents est présenté dans l'annexe B.

Ces collections présentent de nombreux inconvénients parmi lesquels :
– les données ne sont pas du tout équilibrées en terme de quantité, notamment entre l'audio et le texte ; c'est une caractéristique classique

[8] Voir pour une indexation multimédia de vidéos [Smoliar 1997, Smeaton 2002].

des données multimédia hétérogènes car il est beaucoup plus coûteux en temps de transcrire des documents audio que de réunir des documents textuels ;
- les données ne datent pas de la même époque ; cela a néanmoins l'intérêt de faire ressortir l'évolution des sujets abordés d'une époque à l'autre, ceci ayant des répercussions lors de l'évaluation des requêtes.

Pour ce qui est des requêtes, nous avons employé celles qui correspondent à la huitième édition de TREC en *AdHoc* et *SDR*. Les requêtes TREC disposent de plusieurs champs et il est reconnu que l'utilisation de ces différents champs a des répercussions sur les résultats de la recherche. Comme seul le champ *DESC*[9] est fourni pour l'audio, nous n'avons utilisé que ce champ pour les requêtes provenant du texte.

Le nombre de requêtes utilisées n'est pas très élevé, mais elles ne sont pas biaisées car elles ont été choisies sans avoir connaissance du contenu des collections de données [Voorhees 1999]. Nous notons qu'il y a deux sous-ensembles dans ces requêtes :
- les requêtes ayant des réponses dans les deux modalités car le sujet est abordé dans les deux cas ;
- les requêtes n'ayant pas de réponse dans l'une ou l'autre des modalités car les différences d'époques font que le sujet n'est pas abordé dans l'une des modalités.

Ces deux sous-ensembles ne sont pas connus *a priori* ; ils n'ont pas fait l'objet de traitements spécifiques.

3.1.3 Evolution du besoin

Depuis les quelques années qu'existent les moteurs de recherche, il faut se demander si le besoin de l'utilisateur n'a pas été influencé par leur implémentation. En effet, les modèles actuels utilisés en recherche documentaire proposent une liste classée de documents, les premiers ayant une plus grande probabilité de répondre au besoin énnoncé par l'utilisateur. Cela nous éloigne de la problématique ensembliste qui ne fait que deux classes : les documents pertinents et les documents non pertinents.

En regardant cet aspect au niveau de la recherche multimodale, l'utilisateur peut vouloir une liste classée par modalité ou une liste classée toutes modalités confondues, lui permettant de comparer les modalités entre elles. En d'autres termes, l'utilisateur s'intéresse peut-être à la diversité des docu-

[9] Exemple de requête TREC *Adhoc* complète :
title : Ireland, peace talks
desc : How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence ?
narr : Any interruptions to the peace process not directly attributable to acts of violence are not relevant.

ments retournés, à savoir le rappel en modalités et en sous-sujets. Une autre possibilité est qu'il souhaite pouvoir associer à chaque document d'une modalité ses plus proches voisins dans les autres modalités.

Pour résumer, nous classons ce besoin en trois catégories :
- la recherche en *modalités séparées* pour laquelle la requête est soumise à chacune des modalités ;
- la recherche en *modalités croisées*, les réponses à une modalité sont soumises aux autres modalités ;
- la recherche en *modalités mélangées*, l'espace des résultats est commun.

Le besoin que nous adressons dans ce mémoire est le besoin en modalités mélangées dans lequel les documents sont retournés au sein d'une même liste de résultats. Face à ce besoin, la problématique est bien d'équilibrer les résultats de la recherche. Nous avons vu dans la section précédente que des requêtes pouvaient avoir des réponses dans une seule des modalités. Pour mettre en œuvre un équilibrage efficace, il faudra que la couverture en modalités ait un poids moins important lors de requêtes de ce type.

3.2 Caractérisation du déséquilibre entre les modalités

Le déséquilibre en modalités s'exprime en termes qualitatifs mais également quantitatifs.

3.2.1 Quantitatif

Nous avons commencé par étudier le déséquilibre quantitatif au moment de l'indexation. Nous rappelons que les *termes* sont les mots conservés pour l'indexation après la suppression des *mots outils* et suppression des suffixes[10]. Ces *termes* sont ceux utilisés habituellement en recherche documentaire [Buckley 1995, Walker 1999].

Les statistiques présentées dans l'annexe B montrent :
- la quantité de documents dans les modalités ;
- le nombre de termes indexés dans les modalités ;
- la différence de taille entre les vocabulaires des deux modalités ;
- la longueur moyenne des documents.

Ces statistiques montrent qu'il existe un grand déséquilibre en faveur du *texte* pour le nombre de documents (25 fois plus) et la taille du vocabulaire (18 fois plus grand) alors que ce dernier est moins marqué pour la taille moyenne des documents (2 fois plus grande). Dans le modèle vectoriel, ces différences ont essentiellement un impact sur l'*idf* que nous étudions dans la section suivante ; par contre l'impact est difficile à caractériser dans la normalisation *cosine*. Dans le modèle probabiliste, l'*idf* est aussi utilisé, mais

[10] La supression des suffixes a été faite grâce l'algorithme de Porter [Hull 1996].

la normalisation sur la longueur des documents est connue pour favoriser des documents courts, comme ceux de la modalité parole.

3.2.2 Qualitatif

Nous avons voulu examiner le déséquilibre de façon plus *qualitative*. En plus des erreurs provoquées par la reconnaissance de la parole, les structures grammaticales et sémantiques de la parole spontanée peuvent différer de celles du texte.

Nous pouvons nous demander si ces caractéristiques ont des répercussions dans les mesures classiques utilisées en recherche documentaire et si elles s'appliquent à la parole journalistique que nous étudions. La fréquence des termes dans la collection est fortement utilisée à travers l'*idf* (*inverse document frequency*). Nous rappelons que ce dernier permet de déterminer le pouvoir discriminant des termes les uns par rapport aux autres : $idf(terme) = \log \frac{\text{Nombre de documents}}{\text{Nombre de documents contenant le terme}}$. En comparant les facteurs d'*idf* des termes dans chacune des modalités, nous pouvons mettre en valeur leur différence de pondération lors du calcul du score.

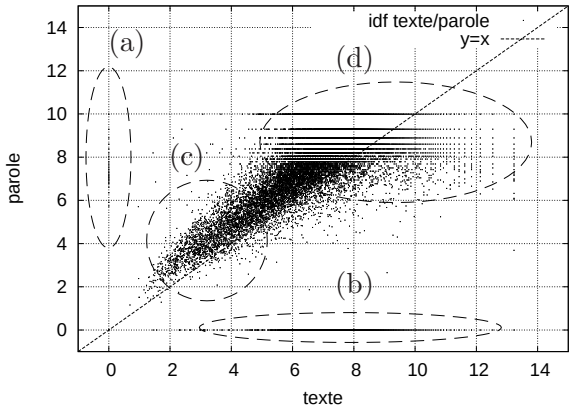


FIG. 3.2 – graphe d'*idf* entre les mots issus de l'audio et ceux du texte, les régions (a) à (d) sont décrite dans le corps du texte.

La figure 3.2 compare l'*idf* de chaque *terme* entre les modalités *texte* et *parole*. Avant de l'interpréter, nous devons en décrire chacune des particularités. Tout d'abord, trois types de points sont observables :

– le nuage central représente les *termes* communs aux deux modalités, plus un point est éloigné de l'axe $y = x$, plus il y a une différence entre l'*idf* dans chaque modalité ;

– les points d'*idf* élevé (d) semblent être répartis selon des paliers précis, ce phénomène est dû à l'inverse d'un nombre entier dans l'*idf* et représente les *termes* les plus rares donc les plus discriminants ;
– les *termes* dont l'un des *idf* est nul n'apparaissent pas dans une des deux modalités et sont situés sur les axes $x = 0$ (a) et $y = 0$ (b).

Cette figure ne montre pas la densité des points dans l'espace, mais il ne faut pas oublier que de nombreux termes ont le même *idf*, par exemple tous les termes qui n'apparaissent qu'une fois dans les deux modalités sont représentés par le point d'abscisse et d'ordonnée maximales. La densité est répartie selon trois tendances :

– elle croit vers les termes d'*idf* fort ; il y a donc beaucoup de termes rares et importants dont le pouvoir discriminant est souvent identique (d) ;
– elle est très forte pour les termes unimodaux, notamment pour le texte (*idf* haut et proche de la droite $y = 0$ (b)) ;
– elle est relativement faible dans le nuage central (c).

Il y a une différence d'échelle entre les deux modalités compte tenu de la différence en quantité de documents (ainsi, l'*idf* maximal est plus grand dans le texte que dans l'audio). Le nuage central n'est pas centré sur l'axe $y = x$ mais décalé vers la modalité audio, les termes d'*idf* moyen sont moins fréquents dans l'audio que dans le texte.

Observer cette figure prend toute sa signification lorsque nous la comparons aux figures présentées en annexe, réalisées en utilisant des sous collections extraites des données étudiées :

– graphe d'*idf* entre texte et parole sur des sous collections de même taille (même nombre de documents, figure C.4) ;
– graphe d'*idf* entre texte et modalités mélangées (figure C.5) ;
– graphe d'*idf* entre parole et modalités mélangées (figure C.5) ;
– graphe d'*idf* entre des sous collections de la modalité texte (figure C.6) ;
– graphe d'*idf* entre des sous collections de la modalité parole (figure C.7).

Si les collections sont de même taille (le déséquilibre quantitatif est réduit), il existe toujours une grande disparité des *idf* entre les modalités. Lorsque chaque modalité est comparée à une collection contenant le mélange des données, nous apercevons que les *idf* du mélange sont beaucoup plus proches de ceux du texte : les spécificités de la parole sont noyées par le déséquilibre quantitatif.

Nous avons aussi testé la cohérence des *idf* à l'intérieur des modalités. Il y a différentes façons de choisir les sous collections intra-modales. Les documents peuvent être pris de façon à avoir des classes proches et similaires

car très représentatives de la collection. Pour cela, les documents sont pris au hasard, ou en suivant l'axe temporel, un document sur deux appartenant à la même sous collection. Les documents peuvent aussi être sélectionnés de façon à maximiser la différence entre les classes. Le plus évident est d'utiliser une frontière temporelle pour que tous les documents antérieurs à une période soient dans la même classe. Lorsque les collections sont prises au hasard, les *idf* sont très proches dans chaque sous collection. Lorsqu'elles sont temporellement distantes, les *idf* sont beaucoup plus dispersés. Nous en déduisons que les disparités d'*idf* sont avant tout dues aux changements de *sujet* à travers le temps. Les données que nous étudions sont journalistiques et il est clair que l'actualité se focalise sur certains thèmes selon des durées limitées. Nous suggérons que les différences intrinsèques aux deux modalités sont en partie masquées par ces évolutions de *sujet*.

3.3 Comportement des méthodes classiques

3.3.1 Modalités séparées

Nous étudions dans cette section les résultats du moteur SMART[11] disponible pour le modèle vectoriel et dans lequel nous avons implémenté le modèle probabiliste[12]. Ces résultats sont évalués selon les modalités de la campagne TREC-8, la piste Adhoc pour le texte et la piste SDR pour la parole.

TREC *Adhoc* : Le système SMART a participé à toutes les campagnes TREC [Buckley 2000, Buckley 1995, Buckley 1996] sur la piste Adhoc [Spärck Jones 1999], mais la version disponible, sur laquelle nous avons travaillé, date des premières éditions de ces évaluations. Nous n'avons pas implémenté les améliorations qui ont été apportées au cours des années et avons utilisé le modèle vectoriel en place avec la pondération *ltc* classique. Nous prenons comme référence [Walker 1999] où Buckley présente un récapitulatif de l'ensemble des méthodes qu'il a proposées depuis le début de TREC. Nous avons pu reproduire des résultats similaires aux siens pour TREC 6, 7 et 8.

Pour le modèle probabiliste, nous avons implémenté dans SMART les pondérations proposées par Robertson dans [Spärck Jones 1998] et obtenons des résultats comparables à ceux publiés dans [Walker 1999]. Les paramètres

[11] disponible sur `ftp://ftp.cs.cornell.edu/pub/smart`.

[12] Le modèle a été implémenté sur la base de la structure de fichier inversé de SMART et de la formulation proposée dans [Spärck Jones 1998]. La librairie ajoutant le support du modèle probabiliste dans SMART est disponible sur `http://www.lia.univ-avignon.fr/downloads/favre`.

k_1 et b (formules 10 et 11) que nous avons employés ne sont pas optimaux : nous les avons fixés empiriquement à respectivement 1.0 et 0.5.

La figure D.9 présente les courbes de précision-rappel que nous comparons à celles produites lors de recherches multimodales. Rappelons que ces courbes correspondent à l'évaluation des requêtes *texte* sur les documents de la modalité *texte* que nous avons présentée dans la section 3.1.2.

TREC SDR : Nous avons utilisé SMART pour faire l'évaluation TREC-8 SDR avec les mêmes méthodes que pour le texte. Dans la piste SDR sont fournies les transcriptions manuelles et automatiques des données audio. Lors des campagnes TREC-8 et TREC-9, [Singhal 1999b] et [Spärck Jones 2001] ont montré que la chute de précision provoquée par les erreurs des transcriptions automatiques était compensable par des méthodes classiques d'expansion de requêtes et de documents dont nous parlerons dans la section 4.2. Ceci justifie que nous n'utilisions que les transcriptions manuelles pour nos expériences préliminaires.

Les résultats présentés dans la figure D.10 sont nettement meilleurs que ceux observés pour le texte car les requêtes associées à l'audio sont plus faciles, elles couvrent des sujets plus larges dans une collection moins grande. Ces résultats ont une grande variance car il existe des requêtes pour lesquelles le système ne retrouve aucun document comme il en existe pour lesquelles il retrouve tous les documents.

3.3.2 Modalités mélangées

Les deux modalités sont facilement mélangeables dans SMART, dans le but d'observer le comportement du système. Par contre, il n'existe pas de référenciel pour les documents d'une modalité soumis aux requêtes de l'autre modalité. Dans un premier temps, nous avons considéré ces documents comme non pertinents lors du calcul de la courbe de précision-rappel. Cela fait apparaître clairement le déséquilibre entre les modalités. En effet, en soumettant les requêtes SDR et en considérant les documents *textuels* comme non pertinents, la courbe de précision-rappel chute, signifiant que les documents audio sont beaucoup moins ramenés que les documents texte.

Il faut tout de même remarquer que, si les documents qui ne sont pas dans la modalité des requêtes ne sont plus considérés comme non pertinents, mais simplement ignorés dans le calcul des courbes de précision-rappel, les courbes sont approximativement les mêmes que lorsque la modalité est seule. Il faut retenir que le classement intra-modal est relativement bien conservé lors d'une recherche multimodale.

Il ne nous a pas été possible de juger l'intégralité des documents dans

les deux modalités pour créer des références complètes. Par contre, nous avons utilisé une mesure plus rapide à évaluer que la précision-rappel : la précision à x documents. Nous avons développé une interface d'évaluation afin de juger la pertinence des 30 premiers documents retournés dans chaque modalité pour les requêtes dont nous n'avions pas les jugements. Cette interface[13] utilise les résultats du moteur SMART pour présenter aux utilisateurs les premiers résultats d'un série de requêtes. Après lecture de chaque document, l'utilisateur doit juger s'il est pertinent ou non. La phase de jugement peut se faire en plusieurs sessions car le temps de lecture des document est relativement long.

La première observation à faire est que parmi les 30 premiers résultats des requêtes *Adhoc*, il y a en moyenne 2% issus de transcriptions de parole, alors que sur les 30 premiers résultats des requêtes *SDR*, il y en a 17%. Ceci s'explique en partie par le fait que les requêtes *Adhoc* sont globalement plus difficiles et correspondent rarement à des sujets abordés dans la modalité *parole*. Par contre, le taux de documents audio sur les requêtes *SDR* est faible et beaucoup de documents hors sujet sont trouvés dans la modalité *texte*.

Le cadre d'évaluation mis en place permettra d'obtenir des résultats de référence pour construire des méthodes d'équilibrage. Au vu du temps nécessaire pour juger les documents et mener à bien l'évaluation des méthodes classiques sur les modalités mélangées, nous n'avons pas pu finaliser ces tests mais les résultats préliminaires, sur 20% des requêtes, confortent notre vision du déséquilibre.

	requêtes *Adhoc*	requêtes *SDR*
documents texte	0.41	0.28
documents audio	0.09	0.31
documents mélangés	0.41	0.39

TAB. 1 – Précision à 30 documents après évaluation de 20% des requêtes.

Pour conclure cette approche de l'information multimédia, nous retiendrons qu'il est facile de faire apparaître le déséquilibre en modalités mais qu'il n'est pas du tout évident d'en trouver la source sur les données que nous étudions. Nous proposons en section 4 quelques solutions pour équilibrer les résultats de recherches documentaires dans un corpus multimodal.

[13]L'interface d'évaluation est disponible sur `http://www.lia.univ-avignon.fr/downloads/favre`.

4 Perspectives pour l'équilibrage en modalités

Dans cette partie, nous explorons différentes méthodes d'équilibrage. Ce sont des perspectives car elles n'ont pas été validées par une évaluation complète sur les données étudiées.

La première idée pour l'équilibrage est de considérer chacune des modalités de façon indépendante et d'utiliser les paramètres locaux dans les pondérations. Par exemple, utiliser l'*idf* de la modalité ou les facteurs de normalisation de la modalité. Utiliser ces facteurs locaux revient à faire des recherches sur les modalités séparément et à fusionner les listes de résultats. Une normalisation sur les résultats est utilisée habituellement pour effectuer la fusion, mais elle n'est pas toujours efficace [Lee 1997].

4.1 Équilibrage quantitatif sur les *idf*

Le graphe d'*idf* vu lors de notre analyse de l'information multimédia montrait le déséquilibre provoqué par les *idf*. En effet, la valeur maximale de l'*idf* est $\log(N_M)$, où N_M est le nombre de documents de la modalité M, lorsqu'un terme n'apparaît que dans un document. Une première méthode d'équilibrage est d'ajuster les *idf* pour normaliser cette valeur maximale. La formule utilisée est :

$$idf_M(t) = \frac{\log N_{\overline{M}}}{\log N} \log \frac{N_M}{n_M} \tag{14}$$

où $N_{\overline{M}}$ est le nombre de documents qui ne sont pas dans la modalité M, N le nombre de documents total, N_M le nombre de documents dans la modalité M et n_M le nombre de documents dans lesquels apparaît le terme t.

Cet équilibrage n'est pas suffisant car le vocabulaire partagé par les deux modalités est relativement faible et les termes unimodaux sont des termes peu fréquents, dont l'*idf* est élevé. Nous allons maintenant traiter ces termes grâce à l'expansion de requête.

4.2 L'expansion de requête

L'expansion de requête est un concept qui fut introduit en observant que l'utilisateur d'un système de recherche documentaire passait beaucoup de temps à reformuler sa requête pour améliorer les résultats de ses recherches. Il est observé dans [Cui 2002, Mitra 1998] qu'il existe une distorsion entre l'espace de représentation des requêtes et celui des documents. C'est ce qui explique la difficulté d'exprimer un besoin d'information à travers une requête. Cette distorsion est poussée à l'extrême par les moteurs de recherche *web* dans lesquels les utilisateurs ne définissent une requête en moyenne que par deux ou trois mots.

Le principe de l'expansion de requête est de repondérer les termes de cette requête et d'en ajouter selon deux méthodes : l'analyse locale des résultats d'une première recherche et l'analyse globale de la collection de documents. Différents types d'analyses fondées sur le partitionnement thématique des documents sont présentés dans [Baeza–Yates 1999].

Des méthodes d'expansion de requêtes et de documents, pour l'indexation audio, sont présentées dans [Johnson 2000, Walker 1999, Singhal 1999a]. [Johnson 2000] a montré qu'il n'y avait pas d'amélioration significative entre l'utilisation combinée de ces techniques et l'utilisation d'une seule d'entre elles.

Le travail présenté dans [Johnson 2000] se rapproche en partie du notre car il effectue son expansion de requête en utilisant un corpus parallèle constitué de documents textuels journalistiques. Dans son cas, les requêtes sont étendues sur la collection, la collection parallèle et les collections mélangées. Les meilleurs résultats sont obtenus en mélangeant les collections. Ces résultats tendent aussi à montrer que la taille de la collection parallèle importe plus que l'homogénéité des sujets abordés dans les deux collections. Les désavantages inhérents aux données que nous utilisons s'en voient légèrement diminués.

4.2.1 Équilibrage dans l'expansion de requête en aveugle

En nous inspirant d'une idée introduite dans [Baumgarten 2000] pour la recherche d'information distribuée [Abbaci 2002], nous allons reformuler la fonction de pondération proposée par Robertson pour l'expansion de requête dans le modèle probabiliste de façon à prendre en compte les modalités et leurs différences.

La fonction de pondération d'un attribut est définie par :

$$weight_i = \log \frac{P_i(1 - \overline{P_i})}{\overline{P_i}(1 - P_i)} \tag{15}$$

avec, lorsque les documents sont tous dans la même modalité :

$$P_i = P(t_i|L) \quad \text{estimée par} \quad p_i = \frac{r_i}{R} \tag{16}$$

$$\overline{P_i} = P(t_i|\overline{L}) \quad \text{estimée par} \quad \overline{p_i} = \frac{n_i - r_i}{N - R} \tag{17}$$

où L est l'événement *Liked*[14] , $\overline{L}$ l'événement *not Liked*, r_i est le nombre de documents pertinents où apparaît le terme t_i, n_i le nombre de documents où apparaît t_i, R le nombre de documents pertinents et N le nombre de documents de la collection.

[14]L'événement *Liked* (*not Liked*) apparaït lorsque l'utilisateur a *aimé* (n'a pas *aimé*) les documents retournés par le système à une requête donnée.

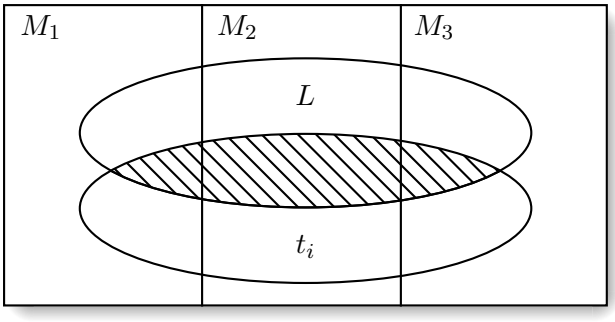


FIG. 4.3 – Les événements t_i et L sur une partition en 3 modalités

Soit $\mathcal{M}$ l'ensemble des modalités. P_i est exprimée en fonction des modalités $M \in \mathcal{M}$, d'après la figure 4.3 :

$$\begin{aligned} P_i = \frac{\sum_M P(t_i \wedge M \wedge L)}{P(L)} &= \sum_M \frac{P(t_i \wedge M \wedge L)}{P(L)} \frac{P(M \wedge L)}{P(M \wedge L)} \\ &= \sum_M P(t_i|M \wedge L) P(M|L) \end{aligned} \tag{18}$$

$P(M|L)$ est ainsi isolée et correspond à la probabilité qu'un document soit d'une modalité donnée quand il est pertinent. Nous pouvons fixer cette probabilité en prenant pour hypothèse qu'elle est la même pour toutes les modalités.

$$\forall M \in \mathcal{M}, \qquad P(M|L) = P(M|\overline{L}) = \frac{1}{|\mathcal{M}|} \tag{19}$$

où $|\mathcal{M}|$ est le nombre de modalités. $P(t_i|M \wedge L)$ et $P(t_i|M \wedge \overline{L})$ peuvent être estimées par :

$$p(t_i|M \wedge L) = \frac{r_{i,M}}{R_M} \tag{20}$$

$$p(t_i|M \wedge \overline{L}) = \frac{n_{i,M} - r_{i,M}}{N_M - R_M} \tag{21}$$

où $r_{i,M}$, $n_{i,M}$, R_M et N_M sont définis comme précédemment mais dans la modalité M. Ceci permet d'obtenir les estimations de P_i et $\overline{P_i}$:

$$p_i = \frac{1}{|\mathcal{M}|} \sum_M \frac{r_M}{R_M} \tag{22}$$

$$\overline{p_i} = \frac{1}{|\mathcal{M}|} \sum_M \frac{n_M - r_M}{N_M - R_M} \tag{23}$$

La pondération est retrouvée en remplaçant P_i et $\overline{P_i}$ dans (15) :

$$weight\_mod_i = \log \frac{\frac{1}{|\mathcal{M}|} \sum_M \frac{r_{i,M}}{R_M} (1 - \frac{1}{|\mathcal{M}|} \sum_M \frac{n_{i,M} - r_{i,M}}{N_M - R_M}}{\frac{1}{|\mathcal{M}|} \sum_M \frac{n_{i,M} - r_{i,M}}{N_M - R_M} (1 - \frac{1}{|\mathcal{M}|} \sum_M \frac{r_{i,M}}{R_M})} \tag{24}$$

Lors d'une expansion de requête, il faut déterminer l'ensemble $\mathcal{R}$, ensemble des documents pertinents. Usuellement ceci est fait lors d'une interaction avec l'utilisateur, ce dernier jugeant les premiers documents du classement. Dans la pratique, l'expansion se fait en aveugle : les R premiers résultats sont considérés comme pertinents pour construire cet ensemble. La qualité de l'expansion de requête est alors relativement dépendante de la qualité de la première itération de la recherche [Cui 2002].

Il faut équilibrer les résultats en prenant l'hypothèse que l'utilisateur a *aimé* autant de documents de chaque modalité lors de l'itération précédente d'une stratégie de recherche. Ceci peut se traduire par $R_M = R$, $\forall M \in \mathcal{M}$, ce qui signifie prendre autant de documents de chaque modalité lors de l'expansion en aveugle. Cette hypothèse est fondée car nous avons vu précédemment que le classement des documents d'une modalité n'est pas détérioré lors d'une recherche multimodale.

Le second aspect de l'expansion de requête est l'ajout de termes à la requête d'origine. Une valeur de sélection permet de décider quels termes ajouter à la requête et quels poids leur donner. Cette valeur est définie par Robertson comme étant :

$$offer\_weight_i = (p_i - \overline{p_i})w_i \tag{25}$$

Il précise que $\overline{p_i}$ peut être ignoré car il est très petit devant p_i, ainsi que w_i est une pondération du terme interprétable par $weight_m od_i$. Si p_i est pris dans (22), $\frac{1}{|\mathcal{M}|R_M}$ est le même pour tous les termes, d'où :

$$offer\_weight\_mod_i = \sum_M r_M weight\_mod_i \tag{26}$$

Grâce à ces formulations, il est possible de construire une requête pour la nouvelle itération d'expansion en aveugle, capable d'équilibrer les résultats en modalités. Le nombre de termes à ajouter est une inconnue théorique, mais il semble intéressant de le relier à la taille du vocabulaire de chaque modalité. Le processus d'expansion de requête que nous proposons peut très bien être utilisé comme expansion de document, en employant chaque document comme requête et en lui appliquant les repondérations et ajouts de termes.

Nous n'avons pas encore évalué l'équilibrage en modalités dans l'expansion de requête car cette évaluation est très coûteuse en temps et en ressources humaines. La prise en compte des modalités dans l'expansion de requête n'en est pas moins prometteuse pour l'équilibrage en modalités.

5 Conclusions

La recherche d'information multimédia est un domaine émergeant dans lequel il reste beaucoup de pistes ouvertes. Dans ce mémoire, nous nous sommes attachés à l'étude des caractéristiques de l'information multimédia, en nous concentrant sur le problème de la recherche documentaire appliquée à des corpus multimodaux.

Lorsque les modèles classiques de recherche documentaire sont appliqués à un mélange de documents textuels et de transcriptions automatiques de parole, un déséquilibre dû notamment à une évolution dans le besoin en information de l'utilisateur est observé. Ce besoin demande un compromis entre précision et couverture en modalités. La couverture est plus importante pour les requêtes ayant des réponses dans le texte et les transcriptions de parole alors que la précision est plus importante pour les requêtes n'ayant pas de réponse dans l'une ou l'autre des modalités.

Compte tenu du déséquilibre quantitatif des données utilisées, nous avons étudié leur déséquilibre qualitatif grâce aux graphes d'*idf* et localisé la source du déséquilibre dans les *sujets* abordés dans chaque modalité. Il apparaît que les différences de localisation temporelle des données provoquent une grande partie du déséquilibre.

L'étude du comportement des méthodes classiques lors d'une recherche sur ces données a néanmoins montré que les documents audio sont très peu retrouvés parmi la masse de documents textuels. Nous avons noté qu'ils conservent malgré tout un bon classement relatif au sein des résultats, information qu'il faudrait pouvoir mieux exploiter.

Par conséquent, nous avons étudié les perspectives d'un équilibrage en modalités à travers l'expansion de requête du modèle probabiliste, capable d'influencer la couverture en modalités et d'utiliser autant de documents pour étendre la requête dans chaque modalité. Afin de valider l'équilibrage en modalités, nous avons mis en place les protocoles et outils permettant de réaliser les évaluations qui détermineront la qualité de la réponse au besoin.

Pour obtenir un compromis précision/couverture capable de satisfaire l'utilisateur, il faut prendre en compte les requêtes dont le *sujet* n'apparaît que dans une seule des modalités. Les documents retournés dans l'autre modalité traitent d'autres *sujets* sans rapport avec la requête. Nous avons envisagé de déterminer à quel point une requête a des réponses dans les deux modalités en comparant les résultats en modalités séparées. Si les documents rapportés sont proches dans les deux modalités, alors la requête a des réponses dans les deux modalités. Sinon, les résultats d'une des modalités ne correspondent pas au besoin de l'utilisateur. Une des perspectives de ce travail est de déterminer la méthode permettant le meilleur compromis précision/couverture et donc un bon équilibrage.

En ajoutant à cet équilibrage l'introduction de facteurs non lexicaux comme peuvent en fournir les moteurs de reconnaissance de la parole, la différence entre les modalités texte et parole orientera notre travail vers une recherche réellement multimédia.

Remerciements

Ce stage de DEA a été réalisé grâce à la collaboration des équipes Parole et Traitement Automatique du Langage Naturel du Laboratoire d'Informatique d'Avignon. Je tiens particulièrement à remercier Pierre Jourlin et Corinne Fredouille, ainsi que tous les membres du LIA, dont l'aide précieuse m'a beaucoup apporté.

A Formulation complète du modèle probabiliste

La littérature n'est pas toujours très claire dans la formulation du modèle probabiliste. Nous avons jugé intéressant de le développer de la façon la plus complète possible. Vous pourrez vous référer à [Spärck Jones 1998, Van Rijsbergen 1979, Robertson 1997] pour des points particuliers.

$Ev(X)$ représente l'occurrence d'un événement X. $P(Ev(X))$ est la probabilité d'occurrence de cet événement. Dans ce cadre, par commodité, l'occurrence de X est notée X, à savoir $P(X) := P(Ev(X))$.

Le système cherche à déterminer si un document D est pertinent (événement noté L) dans le cadre d'une requête. La régle de décision appliquée est :

$$\left[P(L|D) > P(\overline{L}|D)\right] \tag{27}$$

Le critère de décision est transformable en fonction de classement :

$$rank(D) = \frac{P(L|D)}{P(\overline{L}|D)} \tag{28}$$

Le théorème de Bayes nous permet de basculer les probabilités conditionnelles :

$$rank(D) = \frac{P(D|L)P(L)}{P(D|\overline{L})P(\overline{L})} \tag{29}$$

Un document est composé d'attributs A_i, chacun pouvant être présent ($a_i = 1$) ou absent ($a_i = 0$).

$$D = \bigcap_i A_i \qquad \text{avec} \qquad A_i := Ev(a_i = 1) \cup Ev(a_i = 0) \tag{30}$$

Une hypothèse majeure et pas toujours justifiée est faite : les A_i sont indépendants. D'où,

$$rank(D) = \frac{\prod_i P(A_i|L)}{\prod_i P(A_i|\overline{L})} \frac{P(L)}{P(\overline{L})} \tag{31}$$

Ce qui peut être traduit en utilisant les événements de présence/absence a_i en :

$$rank(D) = \frac{\prod_{a_i=1} P(a_i = 1|L) \prod_{a_i=0} P(a_i = 0|L)}{\prod_{a_i=1} P(a_i = 1|\overline{L}) \prod_{a_i=0} P(a_i = 0|\overline{L})} \frac{P(L)}{P(\overline{L})} \tag{32}$$

Les indices sont factorisés :

$$rank(D) = \frac{\prod_i P(a_i = 1|L)^{a_i} P(a_i = 0|L)^{1-a_i}}{\prod_i P(a_i = 1|\overline{L})^{a_i} P(a_i = 0|\overline{L})^{1-a_i}} \frac{P(L)}{P(\overline{L})} \tag{33}$$

Qu'il faut voir sous la forme :

$$rank(D) = \prod_i \left(\frac{P(a_i = 1|L)P(a_i = 0|\overline{L})}{P(a_i = 1|\overline{L})P(a_i = 0|L)} \right)^{a_i} \frac{P(a_i = 0|L)}{P(a_i = 0|\overline{L})} \frac{P(L)}{P(\overline{L})} \tag{34}$$

Après passage au log, le classement est en fonction de a_i :

$$rank_{log}(D) = \sum_i a_i \log \left(\frac{P(a_i = 1|L)P(a_i = 0|\overline{L})}{P(a_i = 1|\overline{L})P(a_i = 0|L)} \right) + \underbrace{\sum_i \log \frac{P(a_i = 0|L)}{P(a_i = 0|\overline{L})} + \log \frac{P(L)}{P(\overline{L})}}_{\text{constant pour une requête donnée}} \tag{35}$$

Notons pour plus de commodité :

$$P_i = P(a_i = 1|L) \tag{36}$$

$$\overline{P_i} = P(a_i = 1|\overline{L}) \tag{37}$$

En ignorant les termes constants quel que soit le document pour une requête donnée, sachant que $P(a_i = 0|L) + P(a_i = 1|L) = 1$ et $P(a_i = 0|\overline{L}) + P(a_i = 1|\overline{L}) = 1$, nous obtenons une pondération :

$$rank_{log}(D) = \sum_i weight_i \qquad \text{avec} \qquad weight_i = \log \frac{P_i(1 - \overline{P_i})}{\overline{P_i}(1 - P_i)} \tag{38}$$

Un terme présent lorsque le document est pertinent ou absent lorsque le document est non pertinent contribuera positivement au score, alors que s'il est présent en cas de non pertinence ou absent en cas de pertinence, il aura une contribution négative.

Il s'agit maintenant d'estimer ces probabilités pour calculer le rang des documents. Voyons comment sont répartis les documents pour un attribut donné :

	Pertinent	Non pertinent	
attribut présent	r_i	$n_i - r_i$	n_i
attribut absent	$R - r_i$	$N - R - (n_i - r_i)$	$N - n_i$
	R	$N - R$	N

où N, R, n_i, r_i sont respectivement le nombre de documents de la collection, le nombre de documents connus comme pertinents, le nombre de documents contenant l'*attribut* a_i, et le nombre de documents pertinents contenant l'*attribut* a_i. $weight(i)$, le poids apporté par un attribut aux documents qui le contiennent, est estimé :

$$weight_i = \log \frac{r_i(N - R - n_i + r_i)}{(n_i - r_i)(R - r_i)} \tag{39}$$

Cette estimation n'est bonne qu'en cas de connaissance *a priori* les documents pertinents et non pertinents. Le processus dans lequel elle est utilisée est itératif et Robertson propose d'ajouter une incertitude au poids de l'attribut avec cette formulation :

$$weight_i = \log \frac{(r_i + 0.5)(N - R - n_i + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)} \tag{40}$$

C'est l'utilisateur ou un processus en aveugle qui s'occupe de déterminer l'ensemble de documents pertinents qui sera utilisé pour pondérer les attributs de l'itération suivante. Lors de la première itération, aucun *a priori* sur cet ensemble n'est connu, l'*idf* classique est utilisé pour estimer ce poids.

$$idf_i = \log \frac{N}{n_i} \tag{41}$$

Le problème de cette formulation est la binarité de la contribution des attributs dans les documents. Pour cela, le concept d'élitisme d'un attribut est introduit [Robertson 1997]. L'événement $Ev(E)$ signifie que l'attribut est une *élite* pour le document, c'est à dire qu'il est *impliqué* dans le *sujet* de ce document.

$$P(A_i = x|L) = P(A_i = x|E)P(E|L) + P(A_i = x|\overline{E})P(\overline{E}|L) \tag{42}$$
$$P(A_i = x|\overline{L}) = P(A_i = x|E)P(E|\overline{L}) + P(A_i = x|\overline{E})P(\overline{E}|\overline{L}) \tag{43}$$

L'hypothèse est faite que la fréquence d'apparition d'un attribut lorsqu'il est (respectivement n'est pas) une *élite* pour un document suit une loi de poisson de moyenne λ (respectivement μ). Cette hypothèse n'est valable que si la longueur des documents est la même quel que soit le document.

Nous verrons par la suite comment introduire des documents de longueur différente.

$$p(A_i = x|E) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{44}$$

$$p(A_i = x|\overline{E}) = \frac{\mu^x e^{-\mu}}{x!} \tag{45}$$

Ces estimations sont introduites en utilisant (42) dans (38) et en posant $\chi = p(E|L)$ et $\overline{\chi} = p(\overline{E}|L)$:

$$\begin{aligned} weight\_elite_i &= \log \frac{P(A_i = x|L)P(A_i = 0|\overline{L})}{P(A_i = x|\overline{L})P(A_i = 0|L)} \\ &= \log \frac{(\chi\lambda^x e^{-\lambda} + (1-\chi)\mu^x e^{-\mu})(\overline{\chi}e^{-\lambda} + (1-\overline{\chi})e^{-\mu})}{(\overline{\chi}\lambda^x e^{-\lambda} + (1-\overline{\chi})\mu^x e^{-\mu})(\chi e^{-\lambda} + (1-\chi)e^{-\mu})} \end{aligned} \tag{46}$$

λ, μ, χ et $\overline{\chi}$ sont les paramètres de cette distribution dont nous n'avons aucune estimation apparente. Robertson et Walker en ont étudié les propriétés pour en déduire une estimation bien plus simple prenant en compte la longueur du document.

$$weight\_elite_i = \frac{freq_{a_i}(k_1 + 1)}{K + freq_{a_i}} weight_i \tag{47}$$

$$K = k_1((1-b) + b\frac{length(D)}{\frac{1}{N}\sum_k length(D_k)}) \tag{48}$$

$length(D)$ est la longueur du document, $freq_A$ est le nombre d'apparitions de l'événement A dans le document ; k_1 et b sont des constantes inconnues qu'il faudra fixer pour chaque collection.

C'est ainsi que sont développées les pondérations utilisées dans le modèle probabiliste dont il faut rappeler qu'il est proche du besoin de l'utilisateur malgré une hypothèse dure : l'indépendance des attributs.

B Statistiques sur le corpus multimodal

Répartition des données dans les modalités :

sources	années	taille	documents	nmm[15]
Trec Disk 4				
Financial Times	1991-1994	564 Mo	210158	412,7
FR94	1994	395 Mo	55630	644,7
CR	1993	235 Mo	27922	1373,5
Trec Disk 5				
FBIS	1994	470 Mo	130471	543,6
LA	1989-1990	475 Mo	131896	526,5
Total sur le texte	1989-1994	2,1 Go	556077	541.8
Trec SDR99				
ABC	1998	2,8 Mo	1827	223
CNN	1998	12 Mo	13528	119,5
Public Radio International	1998	5,1 Mo	2407	326
Voice of America	1998	6 Mo	3992	214
Total sur l'audio	1998	26 Mo	21754	168.5

Les requêtes des deux modalités :

type	identificateur TREC	longueur min (en mots)	longueur max (en mots)	longueur moyenne (en mots)
texte (trec-8)	401-450	5	32	13,76
audio (sdr99)	74-123	8	28	13,74

Nombres de documents :

documents au total	577831
documents textuels	556077
documents audio	21754
ratio	1/25

Nombres de termes (entrées de l'index au sens de Robertson) indexés :

termes au total (en millions)	148,5
termes dans le texte (en millions)	146,7
termes dans l'audio (en millions)	1,7
ratio	1,18%
termes dans texte mais pas audio (en millions)	14,9
termes dans audio mais pas dans texte (en millions)	0,027

Taille du vocabulaire :

[15] nombre moyen de mots par document.

vocabulaire total	525850
vocabulaire texte	522928
vocabulaire audio	29284
ratio	1/18
vocabulaire dans texte mais pas audio	496566
vocabulaire dans audio mais pas texte	2922

Taille des documents (en termes) :

taille moyenne des documents	144.48
taille moyenne des documents texte	149.66
taille moyenne des documents audio	81.76

C Les graphes d'idf

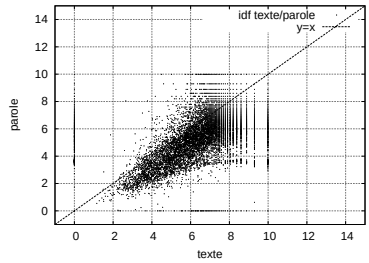
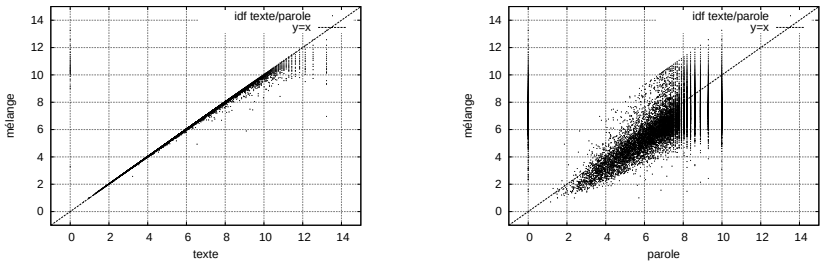


FIG. C.4 – graphe d'*idf* entre texte et parole sur des sous collections de même taille.



texte comparé au mélange audio comparé au mélange

FIG. C.5 – Graphes d'*idf* pour comparer les modalités.

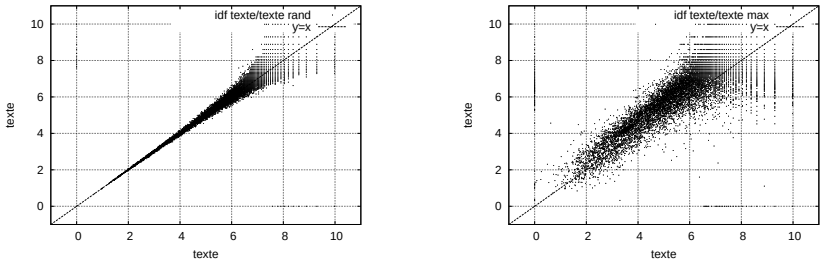


FIG. C.6 – Cohérence de la modalité *texte* selon que les sous collections sont choisies aléatoirement ou de façon à maximiser la différence.

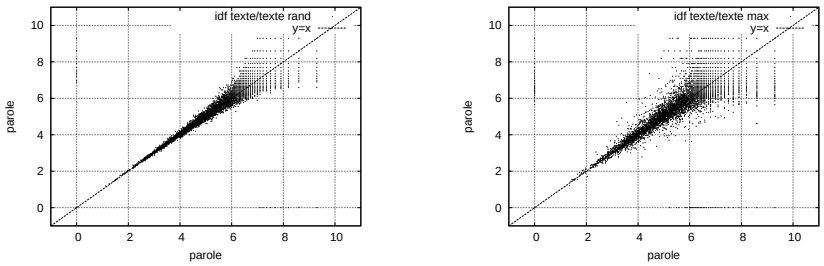


FIG. C.7 – Cohérence de la modalité *audio* selon que les sous collections sont choisies aléatoirement ou de façon à maximiser la différence.

D Courbes d'évaluation

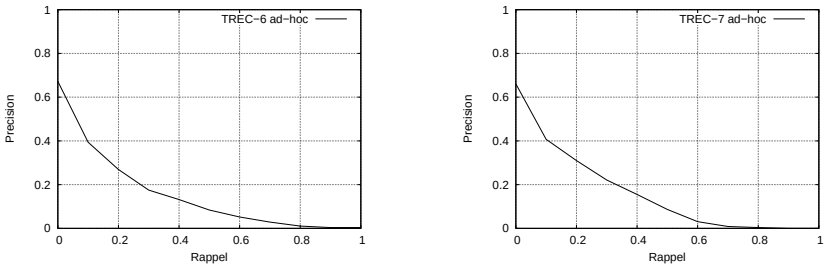
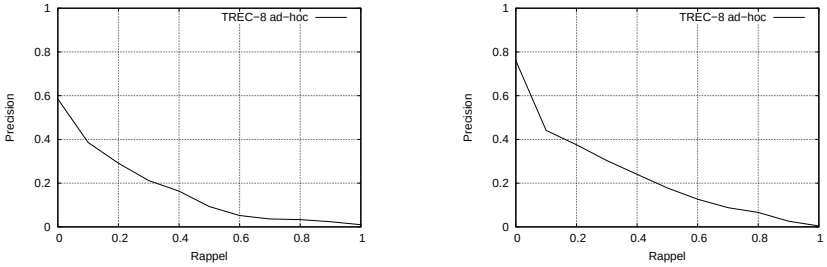
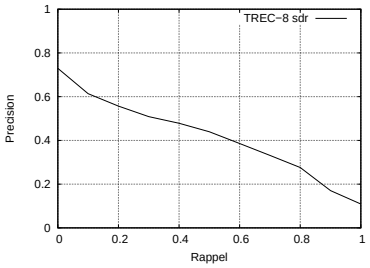


FIG. D.8 – Courbes de précision-rappel sur TREC-6 et TREC-7 *Adhoc* (modèle vectoriel).

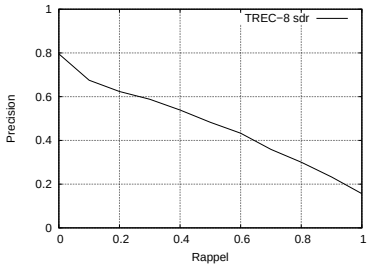


modèle vectoriel modèle probabiliste

FIG. D.9 – Courbes de précision-rappel sur les requêtes de TREC-8 *Adhoc*.

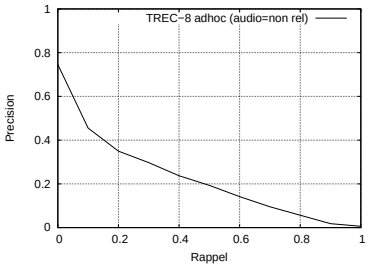


modèle vectoriel

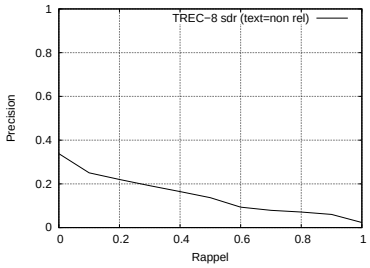


modèle probabiliste

FIG. D.10 – Courbes de précision-rappel sur les requêtes de TREC-8 *SDR*.



requêtes *Adhoc*



requêtes *SDR*

FIG. D.11 – Courbes de précision-rappel lorsque seuls les documents de la modalité de la requête peuvent être pertinents.

Références

[Abbaci 2002] Abbaci F., Savoy J. et Beigbeder M., A methodology for collection selection in heterogeneous contexts, dans *proceedings of ITCC2002, International Conference on Information Technology : Coding and Computing*, pages 529–535, 2002.

[Allan 2002] Allan J., *Information Retrieval Techniques for Speech Applications*, chapitre Perspectives on Information Retrieval and Speech, Anni R. Coden and Eric W. Brown and Savitha Srivinvasen, 2002.

[Baeza–Yates 1999] Baeza–Yates R. et Berthier Ribiero–Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[Baumgarten 2000] Baumgarten C., Retrieving information from a distributed heterogeneous document collection, *Information Retrieval*, 3(3) :253–271, 2000.

[Bellot 2003] Bellot P., Pondérations dans le système vectoriel pour la recherche documentaire, Rapport technique, Laboratoire Informatique d'Avignon, 2003.

[Buckley 2000] Buckley C., Mitra M., A. Walz J. et Cardie C., Using clustering and superconcepts within SMART : TREC 6, *Information Processing and Management*, 36(1) :109–131, 2000.

[Buckley 1995] Buckley C., Salton G., Allan J. et Singhal A., Automatic query expansion using SMART : TREC 3, dans *The Third Text REtrieval Conference*, 1995.

[Buckley 1996] Buckley C., Singhal A., Mitra M. et Salton G., New retrieval approaches using SMART : TREC 4, dans *The Fourth Text REtrieval Converence*, pages 25–48, 1996.

[Cooper 1995] Cooper W. S., Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval, dans *ACM Transactions on Information Systems*, 1995.

[Cui 2002] Cui H., Wen J.-R., Nie J.-Y. et Ma W.-Y., Probabilistic query expansion using query logs, dans *Proceedings of the eleventh international conference on World Wide Web*, pages 325–332, 2002.

[Garofolo 2000] Garofolo J. S., Auzanne C. G. P. et Voorhees E. M., The trec spoken document retrieval track : A success story, dans *The Eighth Text REtrieval Conference*, 2000.

[Hull 1993] Hull D. A., Using statistical testing in the evaluation of retrieval experiments, dans *Research and Development in Information Retrieval*, pages 329–338, 1993.

[Hull 1996] Hull D. A., Stemming algorithms : A case study for detailed evaluation, *Journal of the American Society of Information Science*, 47(1) :70–84, 1996.

[Johnson 2000] Johnson S. E., Jourlin P., Spärck Jones K. et Woodland P. C., Spoken document retrieval for TREC-8 at cambridge university, dans *The Eighth Text REtrieval Conference*, pages 197–206, 2000.

[Lee 1997] Lee J. H., Combining multiple evidence from different relevant feedback networks, dans *Database Systems for Advanced Applications'97, Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA)*, tome 6, pages 421–430, 1997.

[Mitra 1998] Mitra M., Singhal A. et Buckley C., Improving automatic query expansion, dans *Research and Development in Information Retrieval*, pages 206–214, 1998.

[Raghavan 1989] Raghavan V. V., Jung G. S. et Bollmann P., A critical investigation of recall and precision as measures of retrieval system performance, dans *ACM Transactions on Office and Information Systems*, pages 205–229, 1989.

[Robertson 1997] Robertson S. E. et Walker S., Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, dans *Readings in Information Retrieval*, 1997.

[Sanderson 2002] Sanderson M. et Shou X. M., *Information Retrieval Techniques for Speech Applications*, chapitre Speech and Hand Transcribed Retrieval, Anni R. Coden Eric W. Brown Savitha Srinivasan, 2002.

[Singhal 1999a] Singhal A., Choi J., Hindle D., Lewis D. D. et Pereira F. C. N., ATT at TREC-8, dans *Text REtrieval Conference*, 1999.

[Singhal 1999b] Singhal A. et Pereira F., Document expansion for speech retrieval, dans *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, 1999.

[Smeaton 2002] Smeaton A. F., Over P. et Taban R., The TREC-2002 video track report, dans *The Eleventh Text REtrieval Conference*, 2002.

[Smoliar 1997] Smoliar S. W. et Wilcox L. D., Indexing the content of multimedia documents, dans *Second International Conference on Visual Information Systems*, pages 53–60, 1997.

[Spärck Jones 1999] Spärck Jones K., Summary performance comparisons trec-2 through TREC-8, dans *The Eighth Text REtrieval Converence*, 1999.

[Spärck Jones 2001] Spärck Jones K., Jourlin P., Johnson S. E. et Woodland P. C., The Cambridge Multimedia Document Retrieval Project : summary of experiments, Rapport technique, University of Cambridge, Computer Laboratory, 2001.

[Spärck Jones 1998] Spärck Jones K., Walker S. et Robertson S. E., A probabilistic model of information retrieval : development and status, Rapport technique, Computer Laboratory, University of Cambridge, 1998.

[Srihari 2000a] Srihari R. K., Rao A., Han B., Munirathnam S. et Wu X., A model for multimodal information retrieval, dans *IEEE International Conference on Multimedia and Expo (II)*, pages 701–704, 2000.

[Srihari 2000b] Srihari R. K., Zhang Z. et Rao A., Intelligent indexing and semantic retrieval of multimodal documents, *Information Retrieval*, 2(2/3) :245–275, 2000.

[Swanson 1997] Swanson D. R., Historical note : Information retrieval and the future of an illusion, dans *Readings in Information Retrieval*, 1997.

[Van Rijsbergen 1979] Van Rijsbergen C. J., *Information Retrieval.*, Butterworths, 1979.

[Voorhees 1999] Voorhees E. M. et Harman D., Overview of the eighth text retrieval conference (trec-8), dans *The Eighth Text REtrieval Conference*, 1999.

[Walker 1999] Walker S. et Robertson S. E., Okapi/Keenbow at TREC-8, dans *NIST Special Publication 500-246 : The Eighth Text REtrieval Conference (TREC-8)*, 1999.