Recherche d'information dans un mélange de documents écrits et parlés

Benoit Favre, Jean-François Bonastre, Patrice Bellot

Laboratoire d'Informatique d'Avignon - Université d'Avignon
339, chemin des Meinajaries - Agroparc BP 1228 84911 AVIGNON Cedex 9, France
Tél. : +33 (0)4 90 84 35 77 - Fax. : +33 (0)4 90 84 35 01
Courriel : {benoit.favre,jean-francois.bonastre,patrice.bellot}@lia.univ-avignon.fr

ABSTRACT

While advances have been made in structuring, indexing and retrieval of multimedia documents, we propose to study the less explored problematic of information retrieval on heterogeneous media sets composed of written and spoken documents. The coverage of modalities in retrieved results seems to be an important part of the user's information need. We show that this problematic is not satisfied by the usual *bag-of-words* models and we propose a method to balance modalities within the query expansion process of the probabilistic model. Few experiments have been carried out in this domain and we suggest that building evaluation data for the addressed media (text and speech) as well as other media (image...) is worthy to the multimedia information retrieval community.

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.

Avec l'apparition de nombreux documents multimédias, l'augmentation des capacités, des débits et de la puissance de calcul, un besoin de recherche documentaire multimédia émerge, apportant de nouvelles problématiques.

La recherche documentaire sur des documents parlés a été rendue possible en utilisant la reconnaissance automatique de la parole pour indexer les transcriptions grâce à des méthodes textuelles. Nous étudions la recherche documentaire sur un corpus hétérogène, mélange de documents écrits et parlés. Cette dernière particularité implique de prendre en compte conjointement la couverture et la précision des résultats d'une recherche, afin de satisfaire l'utilisateur.

Une rapide introduction revient sur les concepts de la recherche d'information, puis un déséquilibre entre les modalités texte et parole est mis en évidence lorsqu'on utilise les modèles du type "sac de mots". Enfin, une méthode d'équilibrage dans l'expansion de requête est présentée, avant de conclure sur le besoin de données d'évaluation pour ce nouveau domaine de la recherche d'information.

2. Recherche d'information

La recherche d'information explore une problématique simple mais finalement mal définie : "répondre au besoin en information d'un utilisateur". Dans le large champ d'application de la recherche d'information, la recherche documentaire textuelle a été le domaine le plus étudié. Elle consiste à retrouver les documents remplissant le besoin en information d'un utilisateur. Celui-ci exprime le plus souvent ce besoin à l'aide d'une requête écrite (thème, expression, question...)

Des ensembles de données d'évaluation (documents, requêtes et référentiels) sont mis à disposition lors de campagnes visant à mesurer les performances des modèles et systèmes de recherche d'information. Les campagnes les plus connues sont les TExt Retrieval Conferences (TREC[1]) organisées par le National Institute of Science and Technology (NIST), USA [11].

Les référentiels d'évaluation sont des listes de documents, constituées manuellement, séparant les documents répondant à une requête (appelés pertinents) de ceux qui n'y répondent pas. Le plus souvent, les systèmes de recherche documentaire renvoient un classement des documents, les premiers étant les plus susceptibles de répondre à la requête concernée. Les deux mesures des performances les plus répandues en recherche documentaire sont la précision (pourcentage de documents pertinents pour un nombre de documents retrouvés) et le rappel (pourcentage de documents pertinents retrouvés par rapport au nombre total de documents pertinents).

2.1. Recherche documentaire audio

L'objectif de la recherche documentaire audio est de retrouver les documents parlés satisfaisant un utilisateur. Évaluée lors de la piste TREC *Spoken Document Retrieval* (*SDR*), elle propose d'indexer les transcriptions de documents contenant de la parole journalistique. Les transcriptions sont réalisées par des systèmes de reconnaissance automatique de la parole et contiennent une part d'erreurs qui fait diminuer les performances de la recherche documentaire textuelle classique.

Des méthodes d'enrichissement de requête [8] ont permis de rehausser les résultats sur des transcriptions erronées au niveau de ceux des transcriptions manuelles lors de l'utilisation de modèles du type "sac de mots". Un certain nombre de problématiques liées à la nature intrinsèque des documents parlés n'ont pas été suffisamment explorées lors de ces évaluations [2] :

- la longueur des requêtes (il faut pousser l'utilisateur à formuler des requêtes plus longues) ;
- la localisation de l'information pertinente (navigation et résumé de parole) ;
- l'indexation dans des environnements variés (conversations, requêtes parlées, fort taux d'erreur...) ;
- l'utilisation des spécificités de la reconnaissance de la parole (modèles de langage, identité du locuteur, scores de confiance, prosodie...) ;
- la recherche documentaire sur des contenus hétérogènes et multimédias.

[1] http ://trec.nist.gov

2.2. Recherche multimédia

La recherche d'information s'oriente vers le traitement des documents multimédias. L'information est alors contenue dans le texte, l'audio et les images (fixes ou animées) qu'il faut analyser, structurer et indexer afin de pouvoir les exploiter. L'extraction d'informations de bas et haut niveau diffère beaucoup selon le média traité. Il faut alors corréler les médias pour pouvoir en retirer de l'information. La piste vidéo [6] de TREC est un bon exemple de campagne d'évaluation explorant certains domaines de la recherche d'information multimédia.

Divers types de recherche multimédia peuvent être envisagés. Il est possible d'utiliser un média pour en retrouver un autre : ceci a permis d'améliorer significativement les performances de la recherche d'information sur les images, en associant aux caractéristiques de bas niveau, tels que les couleurs ou la texture, les concepts extraits du texte entourant les images [9].

Une recherche documentaire sur des corpus hétérogènes regroupant des documents de médias différents doit aussi être envisagée. Ce dernier type de recherche d'information apporte de nouvelles problématiques et pose notamment la question du taux de couverture des différents médias dans les résultats. En effet, en prenant l'exemple du mélange de documents textuels et de transcriptions de parole, le besoin en information de l'utilisateur demande la considération conjointe des notions de précision et de couverture en médias des résultats. Afin que l'utilisateur soit correctement informé, il faudra lui présenter à la fois les articles de presse et les interviews radiodiffusées correspondant à sa requête.

Il ne semble pas exister de données d'évaluation pour la recherche documentaire sur des corpus hétérogènes comme ceux que nous étudions. Nous avons donc choisi de rassembler les documents, requêtes et référentiels fournis pour les pistes *Adhoc* et *SDR* de TREC-8 [2] car ils sont de nature similaire. Les référentiels d'une modalité sur l'autre n'étant pas fournis, un moyen de déterminer si la couverture en modalités est respectée doit être trouvé.

3. Déséquilibre entre texte et parole

Les modèles du type "sac de mots" tel le modèle vectoriel sont les plus utilisés pour leur simplicité et leur robustesse. Leur faiblesse provient du fait qu'ils n'utilisent que les statistiques d'apparition des mots dans les documents sans vraiment prendre en compte la sémantique créée par leur enchaînement. Ces modèles permettent d'indexer conjointement des documents textuels et des documents parlés.

3.1. Analyse de l'information

Nous analysons, à travers le pouvoir discriminant des mots, le déséquilibre provoqué par les nouvelles problématiques de couverture induites par le besoin en information de l'utilisateur.

Le modèle vectoriel [1] : Les mots composant les documents sont utilisés comme entrées de l'index. Ils sont appelés *attributs* ou *termes* d'indexation. Pour améliorer la recherche documentaire, les mots à faible valeur sémantique hors contexte (*stop words*) sont supprimés alors que les autres sont réduits à leur racine (*stemming*) [3]. Dans ce modèle, les documents sont représentés dans un espace $\mathcal{D}$ dont les dimensions sont les *attributs* qui les composent.

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

Les $w_{i,j}$ sont les poids associés à chacun des attributs $a_i \in \mathcal{A}$ (ensemble des attributs) pour le document représenté par le vecteur $\vec{d_j}$. Les requêtes sont représentées dans ce même espace selon les attributs qui permettent de les qualifier.

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

Les documents sont classés selon leur similarité à une requête. La similarité *cosine* est fréquemment utilisée, définie par le cosinus de l'angle entre le vecteur document $\vec{d_j}$ et celui de la requête $\vec{q}$.

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

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

Il existe différentes façons de pondérer les *attributs* dans les documents [5]. 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 :

$$w_{i,j} = tf_{i,j} idf_i = \log(tf_{i,j} + 1) \log \frac{N}{n_i} \quad (4)$$

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. tf représente l'importance d'un *attribut* dans un document alors qu'*idf* représente son pouvoir discriminant dans la collection.

Les graphes d'*idf* : Les *idf* ($idf_i = \log \frac{N}{n_i}$) permettent d'équilibrer le poids des *termes* d'indexation dans un corpus en définissant leur pouvoir discriminant pour les documents qui les contiennent. Cette mesure est utilisée dans de nombreux domaines de la recherche d'information pour saisir des propriétés globales des corpus. Nous comparons les modalités parole et texte à travers les graphes d'*idf* présentés dans les figures 1, 2 et 3.

Dans la figure 1, le nuage central (c,d) représente les *termes* communs aux deux modalités, plus un point est éloigné de l'axe $y = x$, plus le déséquilibre est grand ; 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).

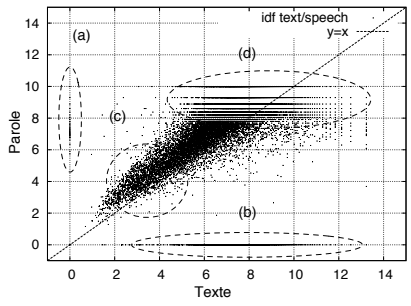
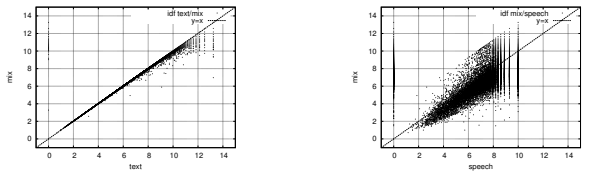


Fig. 1: graphe d'*idf* entre les mots issus de l'audio et ceux du texte

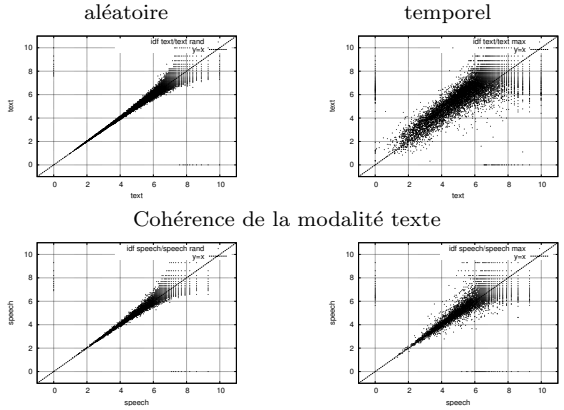
Dans la figure 2, les *idf* du texte sont beaucoup plus proches de ceux du mélange que les *idf* de la parole ; ce phénomène est dû en partie à la différence de quantité entre les modalités.



texte comparé au mélange parole comparée au mélange

Fig. 2: Graphes d'*idf* pour comparer les modalités

Dans la figure 3, les sous-collections choisies aléatoirement sont très représentatives de l'homogénéité des modalités (figures de gauche) ; les sous-collections de périodes temporelles différentes (figures de droite) montrent que le déséquilibre d'*idf* est dû aux sujets abordés dans les sous-collections.



Cohérence de la modalité texte

Cohérence de la modalité parole

Fig. 3: Cohérence des modalités

Pour compléter l'étude, une évaluation manuelle des requêtes sur les modalités croisées est opérée. Les résultats sur 20% des requêtes proposées dans TREC-8 *SDR* et *Adhoc* sont présentés dans la table 1. L'évaluation est réalisée en utilisant le moteur de recherche SMART[2] (disponible pour la recherche et évalué lors de nombreuses campagnes [4]).

Tab. 1: Précision à 30 documents après évaluation de 20% des requêtes : la modalité parole est très peu retrouvée.

	Adhoc	*SDR*
documents écrits	0,41	0,28
documents parlés	0,09	0,31
documents mélangés	0,41	0,39

Les résultats de l'analyse des modalités texte et parole montrent un déséquilibre dont la source n'est pas aisément appréciable. Nous suggérons que ce déséquilibre est dû à l'inadéquation des données au niveau de leur répartition temporelle (les sujets abordés dans l'actualité évoluent dans le temps) et leur déséquilibre quantitatif. Ces conclusions soulignent la nécessité de constituer des données d'évaluation dédiées à ce domaine particulier.

3.2. Rééquilibrage dans l'expansion de requête

Nous proposons une méthode d'équilibrage entre les modalités texte et parole à travers l'expansion de requête. Bien que le processus d'expansion de requête soit possible pour le modèle vectoriel, il est plus facile d'introduire les différences entre les modalités dans le modèle probabiliste (un autre modèle du type "sac de mots" présentant des performances similaires mais un cadre théorique plus développé).

Le modèle probabiliste [7] : L'ensemble des documents est partitionné en deux sous-ensembles : les documents pertinents et les documents non pertinents. On cherche à déterminer si un document (choisi lors de 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 :

$$score(D_j) = \frac{P(L|D_j)}{P(\overline{L}|D_j)} \tag{5}$$

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 permet de récrire les probabilités conditionnelles :

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

D_j peut être représenté par les attributs a_i qui le composent. Pour simplifier les calculs, on suppose que les attributs sont indépendants [10]. Cette hypothèse n'est pas forcément justifiée mais elle permet de réduire la complexité du modèle. Soit A_i, l'événement associé à un *attribut* a_i :

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

L'ensemble des documents pertinents est constitué de façon itérative. C'est l'utilisateur, ou un processus en aveugle, qui détermine la partie de l'ensemble des documents pertinents servant à l'itération suivante de la recherche.

Expansion de requête [1] : Le comportement itératif du modèle probabiliste mène directement à l'expansion de requête. Ce processus a été mis au point en remarquant que l'utilisateur passe beaucoup de temps à reformuler ses requêtes. Il s'agit d'ajuster automatiquement le poids des *termes* de la requête et de l'étendre à d'autres *termes* reliés. Lorsque l'expansion se fait en aveugle, les interactions avec l'utilisateur sont réduites, mais la qualité de recherche est tributaire de la première itération.

Une méthode de rééquilibrage : La fonction de pondération proposée par Robertson pour l'expansion de requête dans le modèle probabiliste est reformulée de façon à prendre en compte les modalités et leurs différences.

La pondération d'un attribut est définie par :

$$weight_i = \log \frac{P_i(1-\overline{P_i})}{\overline{P_i}(1-P_i)} \tag{8}$$

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

$$P_i = P(t_i|L) \quad \text{estimée par} \quad p_i = \frac{r_i}{R} \tag{9}$$

$$\overline{P_i} = P(t_i|\overline{L}) \quad \text{estimée par} \quad \overline{p_i} = \frac{n_i - r_i}{N - R} \tag{10}$$

où L est l'événement *Liked*, $\overline{L}$ 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. Soit $\mathcal{M}$ l'ensemble des modalités. P_i est exprimée en fonction des modalités

[2] ftp ://ftp.cs.cornell.edu/pub/smart

$M \in \mathcal{M}$:

$$P_i = \frac{\sum_M P(t_i \wedge M \wedge L)}{P(L)} = \sum_M P(t_i|M \wedge L)P(M|L) \quad (11)$$

$P(M|L)$ 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.

$$\forall M \in \mathcal{M}, \qquad P(M|L) = P(M|\overline{L}) = \frac{1}{|\mathcal{M}|} \quad (12)$$

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 :

$$p(t_i|M \wedge L) = \frac{r_{i,M}}{R_M} \quad (13)$$

$$p(t_i|M \wedge \overline{L}) = \frac{n_{i,M} - r_{i,M}}{N_M - R_M} \quad (14)$$

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 $\overline{P_i}$:

$$p_i = \frac{1}{|\mathcal{M}|} \sum_M \frac{r_M}{R_M} \quad (15)$$

$$\overline{p_i} = \frac{1}{|\mathcal{M}|} \sum_M \frac{n_M - r_M}{N_M - R_M} \quad (16)$$

L'expansion se fait en aveugle, les R premiers résultats étant considérés comme pertinents pour construire l'ensemble $\mathcal{R}$, ensemble des documents pertinents. 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 $R_M = R$, $\forall M \in \mathcal{M}$.

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 :

$$offer\_weight_i = (p_i - \overline{p_i})w_i \quad (17)$$

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 $weight\_mod_i$ (obtenu en remplaçant p_i et $\overline{p_i}$ dans $weight_i$). Si p_i est pris dans (15), $\frac{1}{|\mathcal{M}|R_M}$ est le même pour tous les termes, d'où :

$$offer\_weight\_mod_i = \sum_M r_M weight\_mod_i \quad (18)$$

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. L'évaluation de cette méthode nécessite d'élaborer des données de test. Nous sommes actuellement en phase de recherche de partenaires pour réunir des données d'évaluation pour la recherche documentaire sur des corpus hétérogènes tels que ceux utilisés dans nos travaux.

4. Conclusions et Perspectives

Nous avons abordé dans cet article la recherche d'information multimédia sur des corpus hétérogènes contenant des documents écrits et des transcriptions de documents parlés. Ce domaine, peu exploré, de la recherche documentaire apporte de nouvelles problématiques comme celle de la couverture en modalités des résultats d'une recherche. En analysant le pouvoir discriminant des mots, nous avons mis en lumière le déséquilibre entre les modalités et proposé une méthode d'équilibrage au travers de l'expansion de requête. Regrettant l'absence de données d'évaluation dans ce domaine, nous proposons d'engager la voie d'une campagne d'évaluation dédiée à cette problématique.

Nous avons étudié les modèles de type "sac de mots" dans cet article, pour leur robustesse et leur simplicité. Les méthodes utilisées en Traitement Automatique de la Langue (TAL) peuvent donner de bien meilleurs résultats sur les documents écrits lorsque l'information recherchée est plus ciblée, mais demandent à être adaptées à l'oral. La parole admet de nombreuses spécificités, qu'il serait bon d'étudier pour la recherche documentaire, notamment le suivi de locuteur dans une conversation (structure de l'argumentation, éléments contextuels), les phénomènes du type hésitations, reprises, coupures, bégaiements ou la prosodie. Développer des méthodes prenant en compte ces spécificités permettra d'aborder les problématiques de questions/réponses et de résumé automatique de parole tout en approfondissant celles de la recherche documentaire.

Références

[1] Ricardo Baeza–Yates and Berthier Ribiero–Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[2] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The trec spoken document retrieval track : A success story. In *The Eighth Text REtrieval Conference*, 2000.

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

[4] G. Salton. The smart retrieval system – experiments in automatic document processing, 1971.

[5] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 1988.

[6] Alan F. Smeaton, Paul Over, and R. Taban. The TREC-2002 video track report. In *The Eleventh Text REtrieval Conference*, 2002.

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

[8] Karen Spärck Jones, P. Jourlin, S. E. Johnson, and P. C. Woodland. The Cambridge Multimedia Document Retrieval Project : summary of experiments. Technical report, University of Cambridge, Computer Laboratory, 2001.

[9] Rohini K. Srihari, Aibing Rao, Benjamin Han, Srikanth Munirathnam, and Xiaoyun Wu. A model for multimodal information retrieval. In *IEEE International Conference on Multimedia and Expo (II)*, pages 701–704, 2000.

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

[11] Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (trec-8). In *The Eighth Text REtrieval Conference*, 1999.