Information retrieval on mixed written and spoken documents

Benoit Favre, Patrice Bellot, Jean-François Bonastre
{benoit.favre,patrice.bellot,jean-francois.bonastre}@lia.univ-avignon.fr
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 09
Fax : +33 (0) 4 90 84 35 01

January 5, 2004

Abstract

While advances have been made in structuring, indexing and retrieval of multimedia documents, we propose to study the unexplored problematics 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 propose a method to balance modalities within the query expansion process of the probabilistic model. As there has never been experiments in this domain, we suggest that building evaluation data for the addressed medias (text and speech) as well as other medias (image...) is important for the multimedia information retrieval community.

1 Introduction

The amount of information available over networks grows every days. This information worths beeing accessed and structured. Indexation and information retrieval are essential tasks to realize these objectives.

Major advances in textual information retrieval where observed within the last years. A new need for multimedia information retrieval bringing new problematics is implied by the rising production of multimedia documents, the growth of capacities, rates and computation power.

Multimedia information is composed of images, videos and audio streams in addition to textual documents. Whereas high level information extraction from still and animated images are outbreaking, automatic speech recognition is used to transcribe and index the spoken content of multimedia documents and performs well enough to achieve Spoken Document Retrieval (SDR). Speech indexing has been much studied during the SDR NIST (National Institute of Science and Technology) evaluation campains [GAV00].

We propose to study the behaviour of classical information retrieval methods on multimedia document collections composed of textual documents and speech transcripts. This multimedia information retrieval task deals with heterogeneous document sets and, as far as we know, has never been explored before. We notice that the user's information need implies the consideration of the modality coverage in the search results. This leads to a balance problem between modalities that can be resolved using an ad-hoc query expansion method that we provide in section 5.2.1. We conclude on the need for evaluation data for this new kind of information retrieval.

2 Motivations

During the last years, information retrieval has been studied on separated media, but as multimedia documents surround us, the medias are now studied together. Links are made in the information space between media. For instance, image retrieval using low level features like color histograms and texture gives low retrieval efficiency [SC01]. Therefore, text is captured around an image (caption...) to get high level concepts related to the image. This kind of information retrieval binds multiple medias in single documents and, to get back to our previous example, an image won't be retrieved without its caption.

We study in this article an approach of multimedia information retrieval where the document collection is made of documents of multiple medias. The study is reduced to textual documents and speech transcripts because these modalities have been well studied during the past years and may be retrieved using similar methods. Textual documents and speech transcripts share the same representation of the content (words, sentences...) but the later enables the retrival of audio documents. We are particularly interested in knowing how the user would deal with a multimedia database containing textual and audio information. The user may be looking for textual only or audio only documents and on the contrary he may want to compare textual to audio content. Hence an information retrieval system working on this kind of database will have to deal with two ways of formulating the user information need.

Information retrieval on mixed media corpus is an important step toward mulitmedia information retrieval and does not seem (as far as we know) to have been studied before.

3 Related work

3.1 Textual information retrieval

There are several approaches to information retrieval given the studied media. Text retrieval has been the first approch to information retrieval and emerged from a difficult problematic : find the documents meeting the user's information need. This need is not well defined and biasedly expressed via a textual query. A document will be *relevant* (or *not relevant*) to the user's query when he *liked* (*disliked*) it. This way of partitionning search results helps in evaluating an information retrieval system using precision and recall metrics. The *precision* is the number of relevant documents retrieved compared to the number of retrieved documents and the *recall* is the number of relevant documents retrieved compared to the number of relevant documents in the target collection.

The Text REetrieval Conference (TREC) is an evaluation campagn organised by NIST and DARPA to provide test data and follow the progress of information retrieval systems. Its *Adhoc* track purposes to evaluate the following task : given a document set and a query set, the system must retrieve the more relevant documents and the less irrelevant documents as possible. Taken for more than ten years, this track has been a success and the very good results lead on other tracks like the *Question Answering* track (more focused on the piece of information relevant to a question), multilingual and translingual tracks.

Another track of TREC is closely related to our research : the Spoken Document Retrieval track. It is basically the same as the Adhoc track but using speech transcripts to retrieve audio documents containg spoken content. Speech transcripts are generated using Automatic Speech Recognition (ASR) which contain transcription errors that can penalize the retrieval process introducing false informations. [JJSW00, WR99, SCH$^+$99] have proved that a word error rate up to 20% could be compensated using query expansion technics. This track was estimated to be really successful [GAV00] after TREC-9 but [All02] lists the following open problematics that should be explored :

- the query length problem (adapt indexing to short queries and encourage the user to formulate longer queries)
- relevant information localization (help the user to recognize relevant documents, summarizing audio

content for instance [HF00])

- adapt indexing to various environments (dialogues, meetings, spoken queries, and environments with a high error rate)
- improve indexing using speech recognition (using the same language models [FII01], word confidence scores, speaker identity, prosody, N-best hypothesis...)
- search mixed text and audio documents.

Simple and efficient models for information retrieval are *bag-of-words* models which see words unconnected (the vector model, the probabilistic model...). But more and more models involve Natural Language Processing (NLP) technics that parse the documents to extract *understood* content [Voo99]. These models are suitable for written documents but the errorfull transcripts of spoken documents will confuse the parsers.

3.2 Multimedia information retrieval

Information retrieval research is extending to address the multimedia content of the growing quantity of documents accessible through networks. The next generation information retrieval systems will have to deal with structured and unstructured media from miscellaneous sources. The information will be represented by different means (text, speech, video...) from which extracting low and high level features really differs. Therefore, each channel of the multimedia data has to be corelated to extract retrievable information. Structuring, browsing and searching the data is the key of multimedia information retrieval. The TREC video track is an example of how every subdomains of multimedia information retrieval are gathered in an evaluation campaign [BCG+03].

Full multimedia information retrieval involving every information that can be extracted from multimedia document is only outbreaking, but domains can be limited to retrieve multiple medias. A media can be used to retrieve other medias, for instance, speech transcripts are used to retrieve video content. Then, the different medias used in a same document (representing the same information) can be used together to achive better performence (typically image retrieval using low level features and textual captions). Finally, information retrieval using heterogeneous document sets containting documents of different medias can be retrieved using specific models.

The last point brings new problematics that may be explored. As information is distributed in heterogeneous document sets, the coverage of the results has to be considered as much as their precision to satisfy the user. In fact, the user may want information from every modality whereas this is not possible using classic information retrieval models. This problematic will be explored in the next section.

4 An analysis of mixed modality information retreival

We study mixed modality information retrieval in this section. The impact of using the vector model with heterogeneous document sets containing written and spoken documents is analysed.

4.1 The vector model

The vector model is the most widely used model for information retreival because it is simple and efficient. Documents are represented in this model in a space $\mathcal{D}$ whose dimensions are the attributes $a_i \in \mathcal{A}$ composing the documents. Hence, the $\mathcal{D}$ space has the same number of dimensions than the number of different attributes. In general, attributes, called index *terms* in [Hul96], are words extracted from the documents after stripping stop words and stemming them.

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

The $w_{i,j}$ weigth of every attribute a_i of the document D_j represented by the vector $\vec{d_j}$. Queries are represented in the same space than documents, as if they where real documents.

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

Documents are ranked according to their similarity to a query. The *cosine* similarity is widely used and quite efficient. It is defined as the cosine of the angle between the document vector and the query vector.

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

where $|\vec{x}|$ is the length of $\vec{x}$ and $\cdot$ is the dot product. Hence,

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

There are many ways to weight attributes in documents [SB88], and the $tf \times idf$ (*term frequency* $\times$ *inverse document frequency*) is used in most situations because it balances corpus and document statistics to get a good estimation of the information carried by a *term* (when stopwords are stripped). It can be computed using the formula below :

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

where $tf_{i,j}$ is the number of occurencies of the attribute a_i in the document $\vec{d_j}$, N is the number of documents in the collection and n_i is the number of documents where a_i appears. In a way, tf represents the relevance of a *term* for a document whereas *idf* represents its discriminant power against other *terms*.

4.2 Data sets

There are no mixed media document sets available to the information retrieval community as far as we know. Therefore, we choosed to gather documents from the TREC evaluation campains since queries and references are provided for their respective modality.

The *Adhoc* TREC track evaluates the retrieval of textual documents using textual queries, whereas the *SDR* track evaluates the retrieval of spoken documents using textual queries. The two tracks provide journalistic data and the same kind of queries[1]

The TREC-8 *Adhoc* and *SDR* track document and query sets are used for the above experiments. Since there are no references for *Adhoc* queries against spoken documents and for *SDR* queries against written documents, a way of analysing how the user information need could be met has to be explored.

4.3 Inverse Document Frequency graphs

Inverse document frequency (*idf*) is used in many information retrieval models (the vector model in our experiments) as an estimation of discriminant power of words for documents. The less a concept is tackled in the document set the more it will be important in a document. This mesure is really helpfull because it can balance words significance in a query.

Graph 4.1 compares *idf* of every *term* in each modality. This figure does not really show the density of *terms* because there are a lot of word whose frequency in the collection lead to equal *idf*. Density follows 3 tendencies : it rises up with *idf*, there are many significant *terms* in both modalities (d); it is relativelly low in the central region (c); it is high for the unimodal *terms* on the $x = 0$ and $y = 0$ axes.

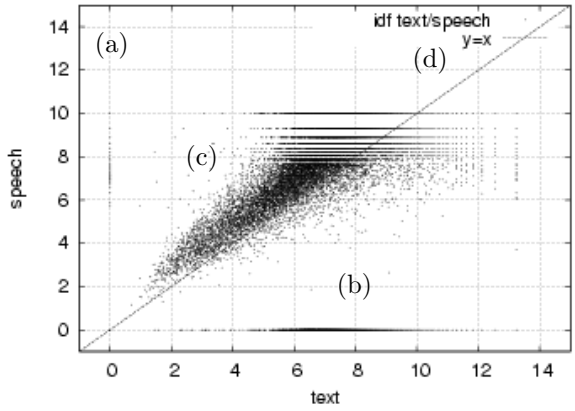
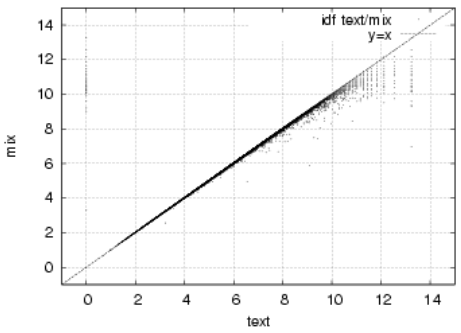
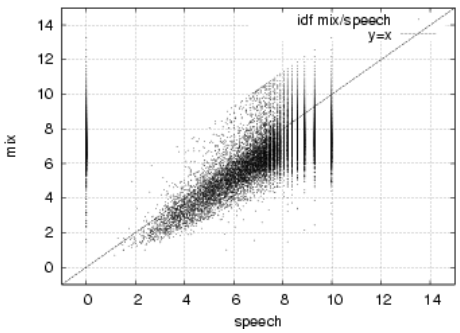


Figure 4.1: *idf* graph between spoken and written words; the central cloud (c,d) represents common *terms* of the modalities, the further a dot is from the $y = x$ axis, the less modalities are balanced; dots of high *idf* (d) seem to be distributed among levels, this phenomemon is due to the inverse of integer part of the *idf* and represents the most uncommon and discriminant *terms*; *terms* that only appear in one modality lay on the $x = 0$ (a) and $y = 0$ (b) axes.



text compared to the speech and text mix



speech compared to the speech and text mix

Figure 4.2: *idf* graphs comparing the modalities; written text *idf* are closer of the speech and text mix than speech *idf*; this phenomenon is implied by quantity balancing problems.

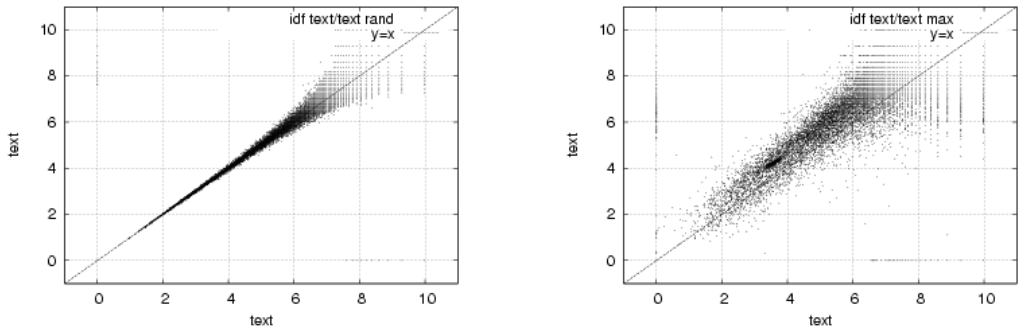
This plot takes all its sense when it is compared to the (4.2) and (4.3) figures realized on subcollections extracted from the original data.

idf are still not balanced if the collections are of same sizes, this means that the balance problem is implied by quality as well as quantity issues. For instance, *idf* of the mixed modalities are far closer from the textual modality than from the audio modality.

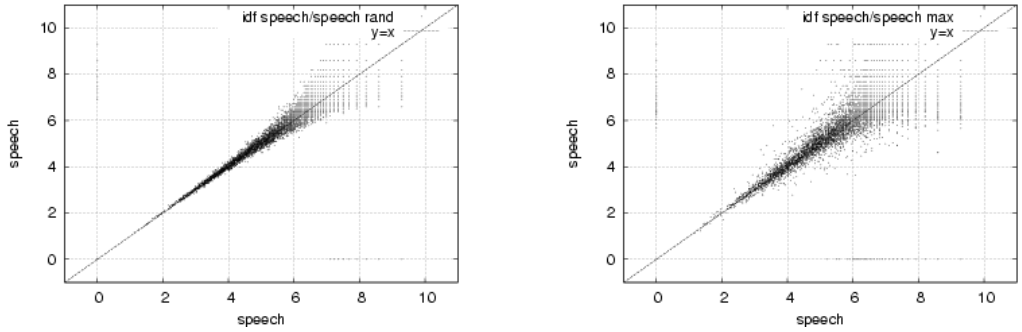
We also tested inner modality coherence. If documents are randomly extracted to build sub collections, the statistical coherency of modalities is shown. On the contrary, if similar documents are taken for each sub collection, *idf* are much more scattered. This suggest that modality differencies are caused by the topic differencies of the data sets. Hence, data dealing with similar topics are needed to achieve this analysis.

TREC queries were submitted to each modality and to the modality mix using the SMART information retrieval system [BMAC00]. We manually evaluated a random set of 20% of the TREC queries for the

[1]Only the *DESC* field of the query will be used for the *Adhoc* track.



Written text inner consistancy



Spoken text inner consistancy

Figure 4.3: Modality consistancy; randomly chosen subcollections represent the homogeneity of modalities (on the left), whereas temporally different collections (on the right) show the subject wide *idf* balancing problems.

Adhoc and *SDR* tracks. Results are given in the below table and show that the spoken documents are not easily retrieved.

	Adhoc queries	*SDR* queries
written documents	0.41	0.28
spoken documents	0.09	0.31
mixed documents	0.41	0.39

Table 1: 30-document precision after 20% manual query evaluation on cross modality experiments

5 Modality balancing

We have seen that the modality coverage in search results is not satisfied using conventional models. An overview of the probabilistic model is given in this section to introduce a balancing method in the query expansion process.

5.1 The probabilisic model

The probabilistic model was introduced by Roberstson and Spärck Jones. It is developped in [SWR98]. This model is built upon the user's information need. Hence, the assumption that the document set is devided in two parts : given a query, a document is relevant or not. Robertson says a document is

relevant when the user *liked* it (the L event) and not relevant in case of dislikeness (the $\overline{L}$ event). A decision rule is derived that can be seen as a ranking function :

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

Where $P(L|D_j)$ is the probability that the user liked the document D_j and $P(\overline{L}|D_j)$ is the probability that he did not like it. Then, the Bayes Theorem is applied to rewrite the conditionnal probabilites :

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

D_j can be represented by its attributes a_i (most of the time the words it contains). We consider the attributes to be independent events to simplify computations althought it is not the case in reality[2]. Let A_i be the event bound to an *attribute* a_i :

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

This ranking function is transposed in the logarithmic domain removing constants[3] for a given query :

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

where P_i is the probability that the attribute a_i is present in the document when it is *liked* by the user, and $\overline{P_i}$ is the probability for the attribute to be present when the document was not liked. The drawback of this formulation is that the score of a document does not involves *attribute* weights. For instance, word (as attribute) weight can be computed using its occurencies. This notion was introduced in [RW97] and is called *attribute eliteness*. It is estimated using *Poisson* distributions and leads to formulations that can be simplified as :

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

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

$freq_{a_i}$ is the relevance of A_i. k_1 and b are tuning constants as theory cannot give them a value.

$$weight_{a_i} = idf_i = \log \frac{N}{n_i} \quad \text{for the first iteration;} \tag{12}$$

$$weight_{a_i} = \log \frac{(r_i+0.5)(N-R-n_i+r_i+0.5)}{(n_i-r_i+0.5)(R-r_i+0.5)} \quad \text{with an } a \text{ } priori \text{ on document relevence;} \tag{13}$$

where N, R, n_i, r_i are respectively : the number of documents in the collection, the number of relevant documents, the number of documents containing a_i, and the number of relevant documents containing a_i. The relevant document set is built iterratively. The user or a blind process gather a part of the relevant document set that drives the next iteration of the search process ([Van79] provides a threshold value to determine the amount of documents to use).

[2] this independency assumption is replaced by the linked dependecy assumption by [Coo95]

[3] $\frac{P(L)}{P(\overline{L})}$ is the same for every documents.

5.2 Query expansion

The concept of query expansion was introduced observing the user search strategies. Obviously, the user spends a long time reformulating his query to meet his information need. [CWNM02, MSB98] observed a distortion between query and document representation spaces that explain the difficulty for the user to express his information need. This distortion leads to an average 3 word query length occulting contextual information.

The purpose of query expansion is to automatically reweight *terms* and add to the query related words from the document set. Local analysis methods only use the first few retrieved documents but stick to the user's information need, and global analysis methods retrieve related words from the whole document set and from external semantic networks [BB99].

Query expansion was applied to spoken document retrieval to cancel the effects of recognition errors [JJSW00, WR99, SCH+99]. [JJSW00] Showed that there where no significative performance gain using all techniques together instead of only one. Indeed, he applied query expansion to a parallel corpus containing textual data and spoken transcripts to improve the performance of spoken document retrieval. The results showed that the size of the parallel collection matters more than the homogeneousity of the documents topics.

5.2.1 Balancing modalites in query expansion

The weighting function proposed by Robertson for query expansion in the probabilistic model is reformulated in order to introduce modality specificities.

This function is defined for an attribute as :

$$weight_i = \log \frac{P_i(1-\overline{P_i})}{\overline{P_i}(1-P_i)} \tag{14}$$

with, when all documents are in the same modality :

$$P_i = P(t_i|L) \quad \text{estimated by} \quad p_i = \frac{r_i}{R} \tag{15}$$

$$\overline{P_i} = P(t_i|\overline{L}) \quad \text{estimated by} \quad \overline{p_i} = \frac{n_i - r_i}{N - R} \tag{16}$$

where each symbol refers to the model previously presented.

Let $\mathcal{M}$ be the modality set. P_i is expressed across all modalities $M \in \mathcal{M}$:

$$\begin{aligned} P_i = \frac{\sum_M P(t_i \wedge M \wedge L)}{P(L)} &= \sum_M \frac{P(t_i \wedge M \wedge L)}{P(L)} \frac{P(M \wedge L)}{P(M \wedge L)} \\ &= \sum_M P(t_i|M \wedge L) P(M|L) \end{aligned} \tag{17}$$

$P(M|L)$ is isolated and corresponds to the probability for a document to be in a modality when it is relevant. We can set this probability to be equal for every modality :

$$\forall M \in \mathcal{M}, \qquad P(M|L) = P(M|\overline{L}) = \frac{1}{|\mathcal{M}|} \tag{18}$$

where $|\mathcal{M}|$ is the number of modalities. $P(t_i|M \wedge L)$ and $P(t_i|M \wedge \overline{L})$ can be estimated as :

$$p(t_i|M \wedge L) = \frac{r_{i,M}}{R_M} \tag{19}$$

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

where $r_{i,M}$, $n_{i,M}$, R_M and N_M are defined as in the above model but for a given modality. This leads to estimations of P_i and $\overline{P_i}$:

$$p_i = \frac{1}{|\mathcal{M}|} \sum_M \frac{r_M}{R_M} \tag{21}$$

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

If P_i and $\overline{P_i}$ are replaced in (14), the weighting function becomes :

$$weight\_mod_i = \log \frac{\frac{1}{|\mathcal{M}|} \sum_M \frac{r_{i,M}}{R_M} (1 - \frac{1}{|\mathcal{M}|} \sum_M \frac{n_{i,M} - r_{i,M}}{N_M - R_M}}{\frac{1}{|\mathcal{M}|} \sum_M \frac{n_{i,M} - r_{i,M}}{N_M - R_M} (1 - \frac{1}{|\mathcal{M}|} \sum_M \frac{r_{i,M}}{R_M})} \tag{23}$$

$\mathcal{R}$, the set of relevant documents of the current query expansion iteration, is determined blindly using the R first results of the preceding search.

The intramodality ranking of the first results is assumed to be of a quality good enough to use as many blindly relevant documents from each modality in the expansion process. This can be seen as $R_M = R$, $\forall M \in \mathcal{M}$.

The second aspect of query expansion is the choice of *terms* to add to the user query. These *terms* are extracted according to their *offer weigth*, defined by Robertson :

$$offer\_weight_i = (p_i - \overline{p_i}) w_i \tag{24}$$

$\overline{p_i}$ can be ignored because it is very small considering p_i. w_i is a *term* weighting function that can be interpertted as $weight\_mod_i$.

$$offer\_weight\_mod_i = \sum_M r_M weight\_mod_i \tag{25}$$

These formulations make it possible to expand the query across modalities in a balanced way. The number of terms to add to the query seems to be theorically unknown, but can be guessed using experimentation. The process of the balanced query expansion is applied to documents themself in a process known as document expansion.

There is no better experimental way to validate the proposed method than evaluating the results using query and reference sets like in the usual evaluation campaings. As far as we know, there is no evaluation data for mixed modality information retrieval available. Queries and references could be built on our own, but it is an expensive effort that we cannot assume. We are currently looking for partners to build up a full evaluation campaign on mixed modality information retrieval.

6 Conclusions

We have presented in this article a sub domain of multimedia information retrieval which involves heterogeneous document collections : information retrieval on mixed written and spoken documents sets. This subdomain has outbroken an evolution of the user's information need, the coverage in modality of the retrieved documents has an important role in the task of satisfying the user.

The *bag-of-words* models that give the best results on Spoken Document Retrieval are shown to be inefficient on mixed modality collections since coverage of the results is not respected. Specially, the *idf* graphs comparing the discriminant power of words across modalities show the disparities between written and spoken document retrieval. Consequently, we propose a balancing method that takes place in the blind query expansion process of the probabilistic model. The probability for words to be added to the query can be the same for every modality and relevant documents must not appear in the first results.

As there has never been experiments in this domain, providing evaluation data for the adressed medias (and other media like images...) appears to be the next step of this work.

References

[All02] James Allan. *Information Retrieval Techniques for Speech Applications*, chapter Perspectives on Information Retrieval and Speech. Anni R. Coden and Eric W. Brown and Savitha Srivinvasen, 2002.

[BB99] Ricardo Baeza–Yates and Berthier Ribiero–Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[BCG+03] P. Browne, C. Czirjek, G. Gaughan, C. Gurrin, G. Jones, S. Lee H. Marlow, K. McDonald, N. Murphy, N. O'Connor, N. O'Hare, A. Smeaton, and J. Ye. Dublin city university video track experiments for TREC 2003, 2003.

[BMAC00] Chris Buckley, Mandar Mitra, Janet A. Walz, and Claire Cardie. Using clustering and superconcepts within SMART: TREC 6. *Information Processing and Management*, 36(1):109–131, 2000.

[Coo95] William S. Cooper. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. In *ACM Transactions on Information Systems*, 1995.

[CWNM02] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Probabilistic query expansion using query logs. In *Proceedings of the eleventh international conference on World Wide Web*, pages 325–332, 2002.

[FII01] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. *Lecture Notes in Computer Science*, 2273, 2001.

[GAV00] 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.

[HF00] Chiori Hori and Sadaoki Furui. Automatic speech summarization based on word significance and linguistic likelihood, 2000.

[Hul96] David A. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science*, 47(1):70–84, 1996.

[JJSW00] S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland. Spoken document retrieval for TREC-8 at cambridge university. In *The Eighth Text REtrieval Conference*, pages 197–206, 2000.

[MSB98] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Research and Development in Information Retrieval*, pages 206–214, 1998.

[RW97] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Readings in Information Retrieval*, 1997.

[SB88] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 1988.

[SC01] C. Sable and K. W. Church. Using bins to empirically estimate term weights for text categorization. In *Proceedings of the 2001 Conference on Empirical Methods in Natuarl Language Processing (EMNLP-01)*, 2001.

[SCH+99] Amit Singhal, John Choi, Donald Hindle, David D. Lewis, and Fernando C. N. Pereira. ATT at TREC-8. In *Text REtrieval Conference*, 1999.

[SWR98] 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.

[Van79] C. J. Van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[Voo99] Ellen M. Voorhees. Natural language processing and information retrieval, 1999.

[WR99] S. Walker and S. E. Robertson. Okapi/Keenbow at TREC-8. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8)*, 1999.