# A Probabilistic Ranking Framework using Unobservable Binary Events for Video Search

Robin Aly[1], Djoerd Hiemstra[1], Arjen de Vries[2] and Franciska de Jong[1]
[1] CS Department, University Twente, Enschede, the Netherlands
[2] CWI, Amsterdam, the Netherlands
[1] {r.aly,hiemstra,f.m.g.dejong}@ewi.utwente.nl, [2] arjen@acm.org

## ABSTRACT

Recent content-based video retrieval systems combine output of concept detectors (also known as high-level features) with text obtained through automatic speech recognition. This paper concerns the problem of search using the noisy concept detector output only. Unlike term occurrence in text documents, the event of the occurrence of an audiovisual concept is only indirectly observable. We develop a probabilistic ranking framework for unobservable binary events to search in videos, called *PR-FUBE*. The framework explicitly models the probability of relevance of a video shot through the presence *and* absence of concepts. From our framework, we derive a ranking formula and show its relationship to previously proposed formulas. We evaluate our framework against two other retrieval approaches using the TRECVID 2005 and 2007 datasets. Especially using large numbers of concepts in retrieval results in good performance. We attribute the observed robustness against the noise introduced by less related concepts to the effective combination of concept presence *and* absence in our method. The experiments show that an accurate estimate for the probability of occurrence of a particular concept in relevant shots is crucial to obtain effective retrieval results.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Retrieval models

## General Terms

Algorithms, Experimentation, Performance, Theory

## Keywords

MultiMedia Information Retrieval, Probabilistic Information Retrieval, Concept Based Search

## 1. INTRODUCTION

With the ever growing amounts of multimedia information becoming available in digital format, the Information Retrieval (IR) Community is increasingly challenged to offer solutions that support the search of multimedia content sets. In this paper we address the information needs that focus on the visual characteristics of the material only. In such cases, results obtained through automatic speech recognition (ASR) are unlikely to be useful. For instance, a request like "Find shots of a person talking on a telephone" (TRECVID topic 202) is unlikely to be handled adequately by relying on the spoken content of a video. For such cases, search methods have to exploit visual features. In this paper, we propose a new framework for probabilistic, purely concept based search of video data.

Concepts denote categories of clearly defined real world objects or non-physical semantic primitives (for instance "outdoor") which can be instantiated visually in a video shot. Using concepts for video indexing creates three main problems. Firstly, image concept detectors can be faulty, secondly, the concepts detected might not be fully applicable to a specific query, and finally the use of queries involving multiple concepts is difficult. Our approach, called PR-FUBE, tackles these problems by: 1) taking into account all possible combinations of occurrence and absences of concepts; 2) using the probability that a concept occurs given relevance, i.e., concepts might contribute only partly to relevance, even if we know for certain that they occur (which we normally don't); and 3) combining the detector output for multiple concepts in a coherent way. The main contribution of this paper is a probabilistic framework to handle queries using the occurrence of concepts. We also attempt to explain why the assumptions made in other models have caused problems. Furthermore, we apply and extend a previously developed method to select applicable concepts and verify our ideas in a prototype system.

We limit the scope of this paper to search on the output of concept detectors, but the approach would be applicable to the directly unobservable occurrence of terms as well. Instead of relying on the results of automatic speech recognition, as if it was certain text, by taking the most probable output, called the 1-best [8], some approaches try to take other, also probable output into account, for example by using the full speech lattice [23]. We expect that our framework can be beneficial in this domain as well.

A commonly adopted procedure in concept based search involves three steps: 1) detection of the occurrence of (a set) of concepts in the video (typically done offline), 2) selection of query-related concepts and 3) search result scoring using the combination of these concepts. Between step 2) and 3) we propose to make an estimate of the "impact" for each selected concept for a specific query and use the outcome for the score calculation. This step is often not modeled separately.

To clarify matters, let us consider Topic 149 from the TRECVID 2005 topic set, "Find shots of Condoleezza Rice". We assume that

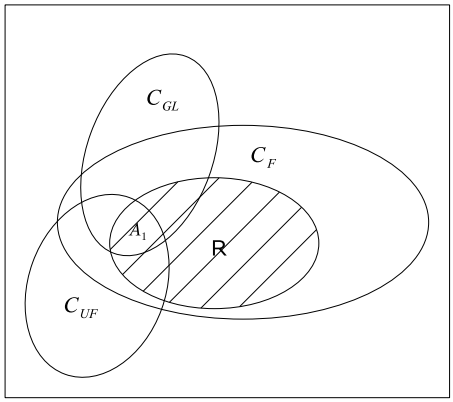



**Figure 1: Topic 149 TRECVID 2005: Venn diagram of Condoleezza Rice $R$) and Three Related Concepts $C_F$, $C_{GL}$and $C_{UF}$**

*Condoleezza Rice* is not present as a concept in the set of concepts detected during indexing time. We consider the set of relevant shots $R$ to be an instantiation of the information need *Condoleezza Rice*. Furthermore, suppose an algorithm for the previously mentioned step 2) selects three concepts as related to the topic: *Female* ($C_F$), *Government Leader* ($C_{GL}$) and *U.S. Flag* ($C_{UF}$). The bracketed symbols denote the set of all occurrences of this concept. In Figure 1 a fictive distribution of occurrences of these concepts and the event of relevance, over all shots of a dataset, are shown as a Venn diagram. While being a fictive distribution, the picture should make clear, that trying to determine whether $C_F$, $C_{GL}$and $C_{UF}$occur together and then estimate the probability that they co-occur with the information is likely to not return good results. The reason is that in the example above $75\%$ of *Condoleezza Rice* appear without a $C_{GL}$or $C_{UF}$.

This paper proposes a framework which exploit all evidence provided by the probabilistic occurrences of a certain set of concepts. In other words, we consider the probability of relevance of shots under presence and absence of related concepts. We therefore follow a quote from Allan Smeaton: "In video information retrieval, we need all evidence we have", made during the Opening Talk of the Search Task of TRECVID 2007.

This paper is organized as follows: In Section 2 we introduce notation which we will use throughout the paper. Section 3 relates previous work to our proposed framework is summarized. Next, in Section 4 we develop our framework for combining concepts and a concrete scoring method. Section 5 describes experiments which evaluate the implementation to demonstrate the effectiveness of the framework. Section 6 concludes the paper and sketches directions of possible future work.

## 2. NOTATION

Throughout the paper we will use the following notation.

**Relevance** $R$ Binary random variable expressing whether a shot is relevant ($R = 1$) or not ($R = 0$) with the same semantics as in [16].

**Concept** $C_i$ Binary random variable expressing the occurrence of a concept $C_i$.

**Detector Output** $D_i$ Real valued random variable expressing the strength of belief of a detector that $C_i = 1$ holds.

**Realization** $\vec{C}$ A vector of occurrences and non-occurrences $(c_1, ..., c_n)$ with $c_i \in \{0, 1\}$. for the random variables $C_1, ..., C_n$ of a particular shot.

**All instances** $C$ Set off all possible realizations: $\{\vec{C} = (c_1, ..., c_n) | c_i \in \{0, 1\}, \forall 1 \leq i \leq n\}$

In unambiguous cases, for a binary random variable $A$ we write $A$ instead of $A = 1$ and $\bar{A}$ instead of $A = 0$. We assume $P(C_i|D_i = d_i) = d_i$ and $P(\bar{C}_i|D_i = d_i) = (1 - d_i)$. That means that the output of the detector reflects the probability of occurrence. The probability that it does not occur is then calculated as usual.

## 3. RELATED WORK

The first step, in order to combine concepts, is to select the ones related to the query. A common approach to use is a large term ontology such as Wordnet [5] to find the "semantic relatedness" [4] between the concept and query terms. This distance is then used to determine the relatedness of the concept to the query. Other methods, for example by Hauff et al. [9], use textual descriptions of concept and perform standard text retrieval using the original query text on these descriptions. The output is a ranked list of document identifiers, which point to the concept to use. Both types of methods have the disadvantage that the impact of a concept on a query cannot be directly deduced from the score produced by the method.

Starting from the publication of Probabilistic Ranking Principle [16], by Robertson in 1977, there has been a lot of research in probabilistic Text Information Retrieval. A good overview is provided by Sparck-Jones et al. [22]. Text retrieval systems base their score calculations on features, which they extracted from the document or from other sources. Sparck-Jones et al. [22] refer to these as attributes. Common features are the frequency of a term in a document, the number of documents that contain a term (document frequency), or the frequency of the term occurring in the whole collection. Other features, which consider the context of a document, are for example the successful PageRank [15] algorithm. In video retrieval the kinds of features are even more numerous. Early video retrieval systems employed mainly low-level features like color vectors. In recent years, mainly text from ASR and detection output from concept detectors came to use.

A lot of work has been done on the most effective way to combine several sources of information. Fox and Shaw show in [6] several methods of how to combine different text search scores. In [3] among others the two principle methods *Add* and *Mult* were investigated. In *Add* the retrieval status value (RSV) is calculated as: ($RSV = \sum_i P(C_i)$). The combination models a OR operation among the concepts. Figure 1 will for example return a score above 1 if a $C_F$is present with a $C_{GL}$. *Mult* ($RSV = \prod_i P(C_i)$) captures the co-occurrences of concepts. As shown in the motivating example from Figure 1, this method can easily return results, which are not about *Condoleezza Rice*. Moreover, a lot of relevant shots won't be ranked correctly, as substantial amounts of *Condoleezza Rice* might not have, for example, a *U.S. Flag* on them. Furthermore, for both methods there is no mechanism to prioritize more influential concepts.

Recently presented work at TRECVID 2007 [20] include interesting attempts to include the specificity of concepts, calculated as $1 - P(C_i)$, into the scoring formula. The measure has a similar effect as the famous inverse document frequency measure in text retrieval - penalizing concepts which occur very often. This approach is based on re-ranking of an existing, trusted ranking of the search result. The ranking from a standard text based retrieval is used as the trusted ranking.

Yan [24] bases his work on the availability of general real-valued features $f_i$ ($1 \leq i \leq n$). His basic probabilistic model requires coefficients $\lambda_i$, which are learned from a training set with relevance judgments for queries. The features are selected based on the $\chi^2$ test with relevance judgments. He then defines his RSV as $P(y_+|D,Q) = [1 + exp(-\sum_i (\lambda_i f_i))]^{-1}$, where $y_+$ is the event of relevance. The difference to our model is three-fold: (1) Our framework follows the methodology of probabilistic Text IR closer. (2) We model the event of relevance under the absence of a related concept explicitly and (3) Our method does not require training.

Zheng et al. [25] take a similar approach to ours. Starting from information theory they derive a formula which is supposed to rank shots in the same way as the probability of relevance. The event of relevance under absence of a concept is not modeled. The ranking formula is built like this:

$$P(Y(d) = y_1) = \alpha \sum_i \log\Big(\frac{P(x_{i1}|y_1)}{P(x_{i1})}\Big) P(X_i(d) = x_i 1) \quad (1)$$

Here, $Y$ is the binary random variable of relevance and $y_1$ is the value for relevance of this shot (here $d$). $x_{i_1}$ stands for the occurrence of a concept $i$. $X_i(d)$ is the random variable that the concept $i$ occurs. As we are using a different notation we allow us to reformulate the formula into our notation:

$$P(R|S) = \alpha \sum_i \log\Big(\frac{P(C_i|R)}{P(C_i)}\Big) P(C_i|D_i) \quad (2)$$

For concept based search they rely on example images: They run concept detectors on them and see what concepts are detected. Afterwards, they use the score of the detector as $P(C_i|R)$. We regard this step as problematic: the fact that a detector detects a concept - to a certain percentage - and a concept occurs to a certain probability given relevance are two different things. For example, the detector might have estimated $0.9$ for the occurrence of a house, but that does not mean that $90\%$ of the relevant shots contain a house.

# 4. PR-FUBE

Following the Probability Ranking Principle [16], we want to order shots of a video collection by their probability of relevance to a query of a user. This way we can base our work on years of successful probabilistic text information retrieval (Text IR) research, and work on a solid statistical foundation.

First, we adapt the problem formulation from Sparck-Jones et al. [22] to the video search setting. Therefore, our aim is to calculate the probability of relevance given a vector of known attributes of a shot. Here, we use only the detector output for concepts as our only attributes. Therefore, let $\vec{D} = (d_1, ..., d_n)$ be the vector of the output of the detectors for a certain shot $S$, where each component $D_i$ corresponds to the detection of single concept $C_i$. Now, we formulate according to [22] the probability of relevance of this shot as follows:

$$P(R|S) = P(R|\vec{D}) \quad (3)$$

In the following Section 4.1 we first show why traditional Text IR methods can not directly be applied here. Section 4.2 presents the main contribution of this paper: a new framework to formulate the probability of relevance given unobservable occurrences of concepts for video search. Section 4.3 presents an implementable ranking criterion which is following this framework. The Section 4.4 investigates key steps in concept selection and its problems. It also describes our basic attempts to this part of the search procedure. Section 4.5 is concerned with the in Text IR widely used relevance feedback method, to improve parameter estimation.

## 4.1 Text IR Procedure

We show that continuing Formula 3 with the common steps in Text IR, for reference refer to [7, 22], is problematic. The steps are as follows: Firstly, we have to apply Bayes' Theorem on the right term of the equation. Afterwards, the odds of relevance are calculated from this result. In Text IR, this has the merit that the probability of $P(\vec{D})$ is canceled out and the general probability of relevance $P(R)$ can be ignored because it is not affecting the ranking. Furthermore, conditional independence of the output of the detectors given relevance ($P(D_i = d_i|R)$) is assumed. Therefore we would have:

$$P(R|\vec{D}) \underbrace{\simeq}_{rank} \frac{P(\vec{D}|R)}{P(\vec{D}|\bar{R})} \underbrace{=}_{cond.indep} \prod_i \frac{P(D_i = d_i|R)}{P(D_i = d_i|\bar{R})} \quad (4)$$

The next step in Probabilistic Information Retrieval is to determine the probabilities $P(D_i = d_i|R)$. We remind the reader that $d_i$ is real-valued number, so we need to answer the question "What is the probability that a particular detector $D_i$ outputs $d_i$, given the underlying shot is relevant". This estimate will depend heavily on the detection procedure and video footage, and has to be re-estimated for every change in the detector.

However, concepts were introduced to bridge, with their defined semantic, the often cited semantic gap [19]. Of course, the reliable detection of the presence of these concepts in a video shot is a very hard problem. We assume that concept detectors provide a probability of the concept being present in a shot, or, $P(C_i|D_i = d_i) = d_i$ and $P(\bar{C_i}|D_i = d_i) = 1 - d_i$. Like in text retrieval, we make the (incorrect) assumption that concepts occur independently from each other, such that $P(\vec{C}|\vec{D}) = \prod_i P(C_i|D_i)$.

For now, let us assume we knew $P(C_i|R), P(C_i|\bar{R})$ and we had perfect classifiers for the concepts $C_i$ with output $\{0, 1\}$. In this case, we could continue using most of the techniques in Text IR, by ranking with:

$$\log O(R) \simeq \sum_i^n \log \frac{p_i(1 - \bar{p_i})}{\bar{p_i}(1 - p_i)} \quad (5)$$

where $p_i = P(Ci|R)$ and $\bar{p_i} = P(C_i|\bar{R})$, using the notation from Sparck-Jones et al. [22]. However, previous research has shown that the classifiers used for concept detection are not very accurate, and shots wrongly classified with $C_i = 0$, would be strongly penalized. This penalization would be worse than in traditional Text IR as the vocabulary of concepts will be much smaller as then number of words. Also, extensions of the basic Text IR model that model term eliteness from their within-document frequency cannot be applied on the binary shot-based detector outputs. Therefore, the remainder of this paper modifies the basic probabilistic text retrieval model for the specific case of noisy binary concept detectors, which is the main contribution of the paper.

## 4.2 The Framework

Let us consider again our start Formula 3. As we can not estimate $P(\vec{D}|R)$, we calculate the probability of relevance given a vector $\vec{C}$ of certain occurrences and absences of $n$ concepts. Afterwards, we multiply this probability of relevance with the probability that $\vec{C}$ was present. We assume now that the probability of relevance only depends on this realization of events (which is of course only reasonable for larger values of $n$, the number of different concepts

used in for scoring). However, one particular $\vec{C}$ is only one out of $2^n$ possible combination of occurrence and absences of $n$ concepts. To calculate the original probability of relevance $P(R|\vec{D})$ from Formula 3 we have to sum over all possible realizations, resulting in:[1]

$$P(R|\vec{D}) = \sum_{\vec{C} \in C} P(R|\vec{C}) P(\vec{C}|\vec{D}) \quad (6)$$

Reasoning about a distribution for $P(R|\vec{C})$ is difficult, so, as common in the related Text IR models, we apply Bayes' Theorem:

$$P(R|\vec{D}) = \sum_{\vec{C} \in C} \frac{P(\vec{C}|R)P(R)}{P(\vec{C})} P(\vec{C}|\vec{D}) \quad (7)$$

Unfortunately, taking the odds of relevance given $\vec{C}$ would not help simplify the term, because of the weighted sum over the now multiple possible representations of a shot. This is why the separate odds from each $P(R|\vec{C})$ term do not preserve the ranking of the original probability of relevance $P(R|\vec{D}$. We can only bring the concept-independent factor $P(R)$ outside the sum to obtain:

$$P(R|\vec{D}) = P(R) \sum_{\vec{C} \in C} \frac{P(\vec{C}|R)}{P(\vec{C})} P(\vec{C}|\vec{D}) \quad (8)$$

Formula 8 represents our framework for unobservable binary events by which we rank their probabilistic occurrence, consistent with the Probability Ranking Principle. However, the formula, like it is presented, has a complexity of $O(2^n)$, and is therefore only usable for small $n$.

## 4.3 Ranking Function

Like in many cases in Text IR, this basic framework cannot be implemented directly into a retrieval system. Estimating $P(\vec{C}|R)$ should be easier than estimating the continuous variant $P(\vec{D}|R)$, but the sheer number of possible realizations of $\vec{C}$ makes reliable estimation of this distribution infeasible in practice. We therefore adopt the first order approximation, common in Text IR, that assumes conditional independence of all $C_i$ given $R$. We actually need the stronger independence assumption among all $C_i$ occurrences, to be able to compute the normalizing constant $P(\vec{C})$. Finally, we assume also the independence of $P(C_i|D_i = d_i)$ for all $i$ in order to calculate $P(\vec{C}|\vec{D})$. Let $(c_1, ..., c_n)$ be the values of the components in $\vec{C}$. Then we have the resulting Formula 9:

$$P(R|\vec{D}) = P(R) \sum_{\vec{C} \in C} \prod_i \frac{P(C_i = c_i|R)}{P(C_i = c_i)} P(C_i = c_i|D_i) \quad (9)$$

Because all $C_i$ are binary, we can apply the generalized distributive law given by Aji [2]:

$$P(R|\vec{D}) =$$

$$P(R) \prod_{i=1}^{n} \Big[ \underbrace{\frac{P(C_i|R)}{P(C_i)} P(C_i|D_i)}_{C_i \text{ occurs}} + \underbrace{\frac{P(\bar{C}_i|R)}{P(\bar{C}_i)} P(\bar{C}_i|D_i)}_{C_i \text{ is absent}} \Big] \quad (10)$$

[1]This formulation bears similarity to Hofmann's Probabilistic Latent Semantic Indexing (PLSI) of documents by latent (term) classes [11], where the concepts in our work play a similar role as the latent classes in PLSI.

Equation 10 is our final ranking formula. Its complexity is $O(n)$ with low constants (6 multiplications and 1 addition). Therefore, it is usable for all realistic sizes of concept sets.

One benefit of our framework is, that it allows to explicitly model "discouraging" concepts. In the example from Figure 1, in a perfect system $P(C_F|R)$ equals 1. However, the user might be presented with a lot of irrelevant shots containing another female person *Ms Bhutto* $C_B$, with Detector $D_B$. Therefore, here the system - or the user - can specify $P(\bar{C_B}|R) = 1$ to filter out these shots. Of course this attempt would also not display *Ms. Rice* together with *Ms Bhutto*. To our knowledge, there has been no work on the identification of "discouraging" concepts. This paper focuses on the framework itself; therefore we leave this subject to future work.

Note, if we leave out the term marked with $C_i$ is absent in Formula 10 and set $P(R) = \alpha$, we have exactly the Formula 2 from Zheng et al. [25], because the sum of logarithms is equal to the product. This suggests that Zheng's formula only considers the case $P(\vec{C}|R)$ with $\vec{C} = (1, .., 1)$, in other words, the calculated quantity is the probability of relevance given the occurrence of all $n$ concepts times the probability that all these $n$ concepts occur.

## 4.4 Concept Selection

In Sparck-Jones' terminology, the selection of the attributes $A_i$ for shots in video search is harder than for documents in Text IR. In the latter attributes are commonly query terms. However, in video search, the query is a sequence of terms from the users' large vocabulary. This sequence has to be mapped to a set of applicable concepts. Certainly, a one to one mapping from one term to a concept with the same name will not be sufficient, as many concepts will not be available.

In particular, we identify three challenges in the selection of concepts for a query $q$:

1. Identification of a set of concepts which occurrences differentiate between relevant and non relevant shots,
2. Estimation of probability of occurrence given relevance $P(C_i|R)$,
3. Extracting general properties of concepts and detectors like the probability of occurrence $P(C_i)$ or the reliability of the detector output.

As this paper is focused on the development of the general framework for $n$ concepts, we only propose a preliminary method to the mentioned challenges and leave the improvement of these methods to future work. In the following, we give examples for the challenges and explain how this was done in our framework.

To give an example for the challenges of the identification step (1), suppose it yielded for the query *Condoleezza Rice*, only the concept $C_F$. Given perfect detection, we would find all relevant shots randomly distributed over the returned shots which contain $C_F$. This is clearly not a good solution. On the other hand, taking as many concepts as possible might not be advisable either, as mistakes in the following steps might lead to noise in the ranking. Therefore, the selection method has to find a tradeoff between these two extremes.

Our concept identification step is done according to [9]. For each concept we gather descriptive text and put it into a document with the document identifier equal to the concept name. Afterwards, a standard Text IR system indexes the documents. The user's full-text query is executed in the Text IR system. The result, a ranked list of concept identifiers, is taken as a list of related concepts. For now, we do not have a well motivated, formal method on how to identify

$$\frac{P(C_x|R)}{P(C_x)} = \frac{0.3}{0.2} = 1.5 \quad \frac{P(\bar{C_x}|R)}{P(\bar{C_x})} = \frac{0.7}{0.8} = 0.88$$

$$\frac{EP}{P(C_x)} = \frac{0.4}{0.2} = 2 \quad \frac{1-EP}{P(\bar{C_x})} = \frac{0.6}{0.8} = 0.75$$

**Table 1: Effects of wrong Estimation of $P(C_i|R)$**

a good number of concepts to select from this list. Therefore, the experiment section investigates the performance in dependency of the number of used concepts.

The second challenge is the estimation of $P(C_i|R)$. In Text IR this subject has been researched for decades. Here, especially the conditional independence of concepts, given relevance, makes a precise estimation crucial. For example let us consider a concept $C_x$ and assume $P(C_x|R) = 0.3$ and $P(C_x) = 0.2$. And suppose the estimate $EP$ of the probability of $P(C_x|R)$ is 0.1 too high. Then, the differences of the two weights in Formula 10 are shown in Table 1. One can see that the probability of occurrence of $C_x$ is incorrectly emphasized by 25%. In the example of $C_{UF}$this would mean that a lot of shots with *U.S. Flags* are displayed, which do not contain Condoleezza Rice.

Here, we take a straightforward approach. We transform the scores, from the Text IR system, linearly into an interval and take these numbers to be the probability $P(C_i|R)$.

$$P(C_i|R) = \frac{score - min}{max - min} range + lowest \tag{11}$$

Here, $score$ is the score from Text IR, $min$ and $max$ are the minimum and maximum score in the returned ranking for the query. Furthermore, $range$ is the interval in which the found scores should be situated and $lowest$ is the begin of the interval. We experimented with multiple different settings for $range$ and $lowest$.

Possible properties of concepts and their detectors are the estimated frequency of their occurrences and the quality of their detectors. A much related concept might be excluded from the set if it is known that its detector faulty. At the moment, we only take the frequency of the concept occurrence into account. However, we plan to include measures the performance of detectors in the future.

## 4.5 User Study and Relevance Feedback

We first tested the performance of human concept selection, through a user study. The users were asked to select all concepts, which they thought are related to a query. After the study, we estimated the probability that a concept occurred with relevance, $P(C_i|R)$, by the number of users which selected the concept $C_i$ ($numUser(C_i)$) over the total number of users having participated in the study ($numUser$):

$$P(C_i|R) = \frac{numUser(C_i)}{numUser} \tag{12}$$

We assume here that more users will select the concept if it occurs more often in relevant shots. Due to constraints in resources and the big number of concepts in the Vireo concept set, we limited the study to the TRECVID 2005 topics. However, the application of such a method is only feasible, for example, in collaborative search scenarios as demonstrated by Adcock et al. in [1] at TRECVID 2007.

As the estimation of $P(C_i|R)$ is difficult, we investigate the help of user feedback [17] to improve the estimation. This common paradigm is often used to re-estimate the parameters of IR systems. We proceeded as follows: Given the best ranking without feedback

| | TRECVID 2005 | TRECVID 2007 |
|---|---|---|
| Number of Shots | 45765 | 18142 |
| Domain | News Broadcast | General Television |
| Concept Set | MediaMill [21] | LSCOM [13] |
| Detector Output | MediaMill [21] | Vireo [12] |
| Number of Detectors | 101 | 374 |
| Number of Queries | 24 | 24 |

**Table 2: Data Set Statistics**

for TRECVID 2005 and 2007 we consider the first $i$ entries and extract the relevant shots, using the relevance assessments. This selection of relevant shots would have been performed by the user in reality. As we did not have the ground truth of the occurrence of concepts for the search set, we had to assume that the detectors, given relevance, performed on well. We then recalculate the parameters $P(C_i|R)$ through the probability of occurrence of $C_i$ in the selected, relevant shots $SR$, through the following formula:

$$P(C_i|R) = \frac{\sum_{r \in SR} P(C_i|D = d_{ir})}{|R|} \tag{13}$$

Here, $d_{ir}$ is the estimation value for concept $i$ of shot $r$. Afterwards, we perform a reordering of the concepts in the original Text IR ranking in decreasing order of $P(C_i|R)$. Therefore, the concepts with highest expectation to occurrence in a relevant shot are used first. Then, the scoring was performed on all other remaining shots.

# 5. EXPERIMENTS

In this section we present the experiments we performed to assess the performance of our framework. As the focus of this paper was on the development of the general framework, some employed methods can be improved.

## 5.1 Experimental Setup

We evaluated our framework on the datasets of TRECVID 2005 and 2007. Statistics over the collection, detector sets and queries are gathered in Table 2. A preliminary assessment of the accuracy of the output of MediaMill's detectors, on the provided test set, showed that the $d_i$ values reflected the probability of occurrence well. Unfortunately, as there was not test set for the Vireo Detectors available, such an evaluation was not possible. The reader is reminded that our method does not only depend on a good ranking of occurring and non-occurring shots (a relative ordering), but on the absolute estimated probability of occurrence, which we assume to be $d_i$ here. We used the standard TRECVID topics (24 for each dataset) which were provided with relevance judgments for both datasets.

We compared our method *PR-FUBE* against the methods *Mult* and *Entropy*, as described in the section of related work 3. We take *Mult* as the baseline. The results of the *Entropy* method, which can be seen as a simplification of our method, should demonstrate the improvement of taking also the possibility into account that a shot could be relevant with the absence of a concept, which is one of our main contributions in this paper.

## 5.2 Experiments Execution

This Section explains the results of our experiments. In all graphs, the X axis depicts the number of concepts which were used in the combination function, and the Y axis shows the achieved mean average precision (MAP) of the particular method. However, for the

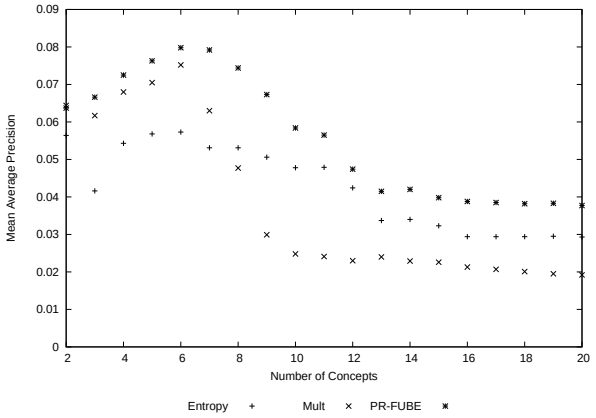



**Figure 2: Run Concept Selection and Parameter Estimation From User-Study**

| Concept | Text IR Score | $P(C_i\|R)$ estimate | Oracle $P(C_i\|R)$ |
|---|---|---|---|
| Tennis | $1.2527E^{-15}$ | 0.2500 | 0.3470 |
| Court | $4.662E^{-18}$ | 0.0507 | 0.0016 |
| Baseball | $4.1023E^{-18}$ | 0.0507 | 0.0001 |
| Basketball | $4.1023E^{-18}$ | 0.0507 | 0.0224 |

**Table 3: Top-4 Concepts for TRECVID 2005 Topic *0156 Tennis players on the court***

Relevance Feedback experiment we used on the X axis the amount of shots virtually considered by the user.

### 5.2.1 User-Study

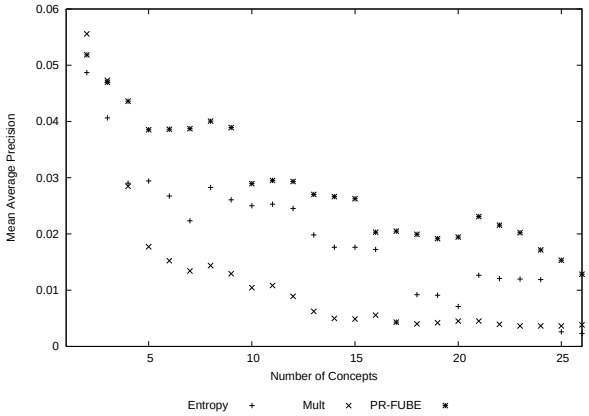
Figure 2 shows the results of the combination of the concepts selected through the users in our study. Note, that we plotted the graph up to 20 concepts, which was the maximum number of selected concepts per single query. Until concept 7 all methods increase their performance monotonically. However, from the eighth concept onwards the performance decreases for all methods. Investigations revealed that the eighth concept, *Crowd*, for the most influential topic $0156$ did appear less often in relevant than in none relevant shots. Nevertheless, *PR-FUBE* performs in all cases better than the other methods.

### 5.2.2 Wiki Concept Selection

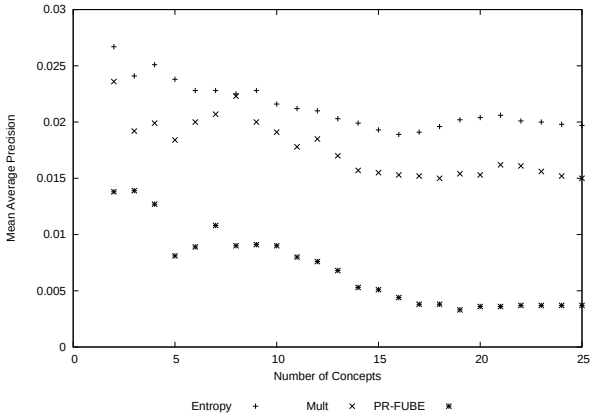
Furthermore, we also investigated an automatic concept selection technique from Wikipedia articles as descriptive text, as described in Section 4.4. We downloaded from Wikipedia[2] for every concept the corresponding article with the same name. Afterwards, we used the PF/Tijah [10] Text IR system to perform the queries on the set of documents. The result was for every query a ranked list of concept names with scores. Clearly, the scores are dependent on the ranking function of the Text IR system. To attain $P(C_i|R)$ we performed a transformation of the Text IR score according Formula 11 with $lowest = 0.05$ and $range = 0.20$.

Figure 3 (a) and (b) show the results using the ranking of the concepts by the results of the Text IR results with linear interpolation to estimate the probability of occurrence of a concept given relevance ($P(C_i|R)$). We plot up to a maximum of 25 concepts

[2]http://en.wikipedia.org




(a) TRECVID 2005

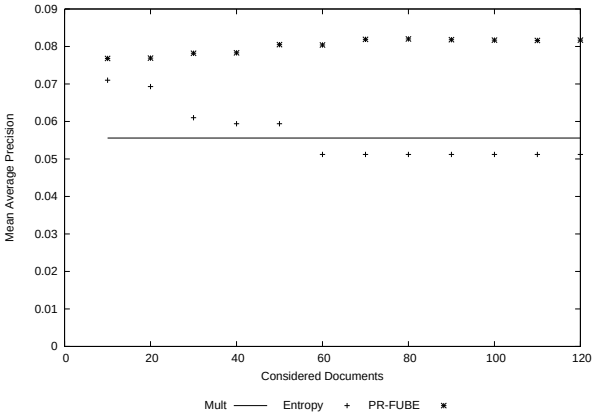



(b) TRECVID 2007

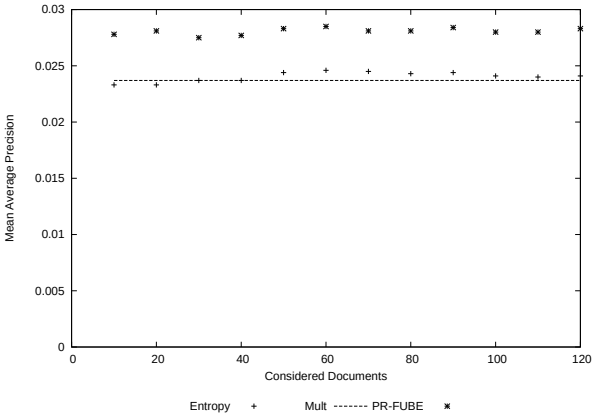
**Figure 3: Wiki Concept Ranking with Linear Transformation of Text IR Score to $[0.05...0.25]$ for Parameter $P(C_i|R)$**

as displaying a higher number of concepts did not add new information and worsened the overview. Unfortunately, the resulting picture is different from the user study. Sub-figure 3 (a) shows the results for the MediaMill detector set on TRECVID 2005 data. Here, the best result was achieved by method *Mult* with the combination of only two concepts. From there the results constantly degrade. However, our *PR-FUBE* method performs better than the other methods from the third concept onwards. An investigation of the top-5 ranked concepts revealed, that for the by far best performing topic *0156 Find me shots of tennis players* the concepts shown in Table 3 were selected. The second concept *Court* was selected as the download from Wikipedia resulted in a description of a tennis court. However, the detector from MediaMill was designed to detect *legal courts*, where trials are hold. The method *Mult* did not get as much affected by this confusion as it was treating both concepts independently. The first concept *Tennis* was often sufficient to answer the query while the concept *Court* appeared only seldom. Furthermore, the next two concepts, *Basketball* and *Baseball*, got also a comparatively high value for $P(C_i|R)$ assigned, where it is very unlikely that shots with tennis players also contain basketball or baseball. In fact, they can be seen as discouraging concepts.

Sub-figure 3 (b) shows the results of the same experiment with Vireo Detectors on the TRECVID 2007 dataset. Unfortunately the *PR-FUBE* method performs always the worst and our comparison method *Entropy* is performing the best for all number of concepts.

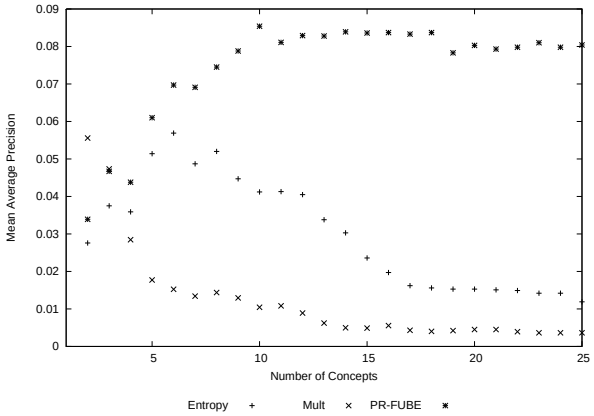


(a) TRECVID 2005

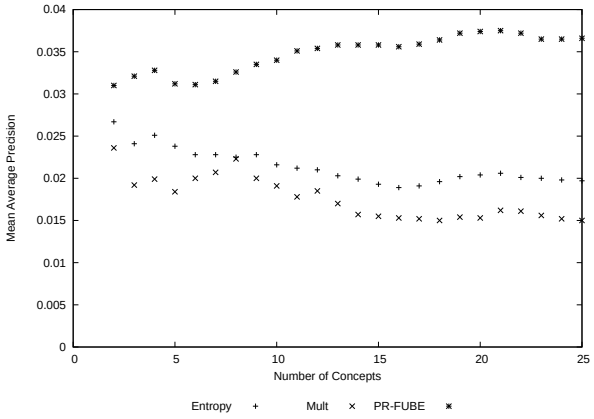


(a) TRECVID 2007

**Figure 4: Relevance Feedback: first $n$ Concepts with Parameter $P(C_i|R)$ Estimation from $i$ Shots. For *PR-FUBE* $n = 12$ and for *Entropy* $n = 2$**



(a) TRECVID 2005



(b) TRECVID 2007

**Figure 5: First $n$ Concepts through Text IR Scores and Retrospective Parameter Estimation for $P(C_i|R)$**

Furthermore, it shows the best performance with only two concepts, which suggests that, the combination of more than two concepts is not useful.

### 5.2.3 Relevance Feedback

We evaluated the relevance feedback experiment with the score freezing method following Salton [18]. Therefore, we created a base run using our *PR-FUBE* method with two concepts. Afterwards, we then fixed the first $i$ entries in the ranking, and took the relevance information from provided relevance judgments. In other words we act like a user that sees the top $i$ shots of a ranking and selects the relevant ones, consistent with the judgments. For the set of selected relevant judgments $SR$ in the first $i$ shots, we re-estimated the parameter $P(C_i|R)$ by Formula 13. We then reorder the concepts by decreasing $P(C_i|R)$. This prioritizes concepts which occur most often in relevant shots in the selection. We then performed the search again with the new parameters, ignore all already considered $i$ shots and appended the result to the previous ranking.

The results of the relevance feedback experiments are shown in Figure 4. Both sub-figures show on the X axis the number of considered shots $i$ in the base ranking and the MAP that the method achieved with using this information. Note, as the *Mult* method does not depend on the parameter $P(C_i|R)$. Therefore, we only plot here the best results from the original run as a reference, see Figure 3. Figure 4 (a) shows the performance of relevance feedback for TRECVID 2005. From the changed ranking we use 12 concepts for *PR-FUBE* and 2 *Entropy*. We do this because the performance of *Entropy* deteriorates quickly with more than two concepts. The methods *PR-FUBE* performs much better than without the feedback. Already with ten considered shots, the performance is better than all other methods measured for TRECVID 2005.

In Figure 4 (b) the results of user feedback for TRECVID 2007 are shown. Here, the results are also very different from using no feedback. Our method *PR-FUBE* is even always better than the other two methods, also compared to the previous runs with pure parameter estimates from Text IR scores.

### 5.2.4 Retrospective Experiments

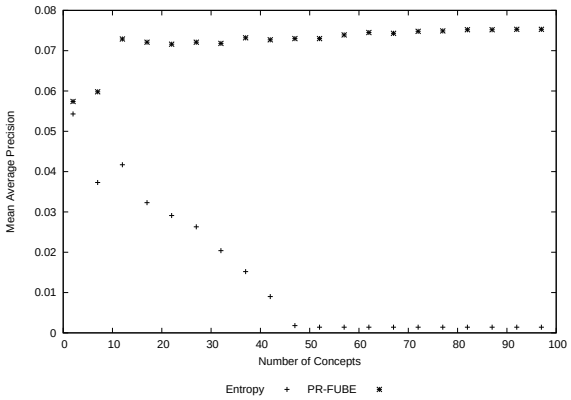
We conducted following retrospective experiments to reveal what would have been achievable, given a good estimation for the probability that a concept occurs given relevance $P(C_i|R)$ from the start (without relevance feedback). Note that the some detectors showed biases. For example, for query *0150 Find shots of Iyad Allawi* from the TRECVID 2005 topic set the average of $d_i$ for the concept *Allawi* was only $0.007$ , whereas it should have been, by definition 1.

To be able to compare the results of the retrospective experiments, we first used the ranking from Text IR and estimated $P(C_i|R)$

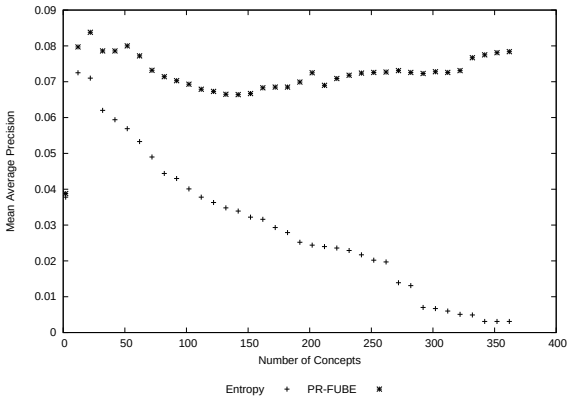
through Formula 13, where we use the whole set of relevant shots $R$ as $SR$. This simulates the possibility that we selected the order of the concepts through Text IR and an "Oracle", which has access to the relevance judgments, tells us $P(C_i|R)$.

Figure 5 (a) shows the performance of the MediaMill detector set using the Text IR ranking from the Wikipedia articles. *Entropy* and *PR-FUBE* use Parameter Estimates from the Oracle, while *Mult* is same from the first experiment with Text IR ranking. Initially both re-estimated methods *PR-FUBE* and *Entropy* are worse than the *Mult* method. However, they both improve until six concepts. Afterwards, the *Entropy* method deteriorates nearly monotonically. In contrast to that the *PR-FUBE* increases further. A maximum of $8.5\%$ is reached at ten concepts. Then, with more concepts, the MAP of *PR-FUBE* stabilizes around $8\%$, which is a absolute improvement of $2.5\%$ against the best result from the original method. In Figure 5 (b) the results on the TRECVID 2007 dataset with the Vireo detectors are shown. The performance of the *Entropy* method decreases steadily. On the other hand, our *PR-FUBE* method is in all measurements better than both other methods and reaches a maximum of $3.75\%$ at 21 concepts.

The second retrospective experiment used the result from the "Oracle" and ordered the concepts in decreasing order of $\frac{P(C_i|R)}{P(C_i)}$. Therefore, the concepts, which occurred more often in relevant shots than in the rest of the collection, were selected first. Figure 6 show the results of this experiment. To increase the clarity we only plotted for TRECVID 2005 in an interval of five concepts because the plot of using every single increase of $n$ did not reveal any information. For TRECVID 2007 we used ten concepts as the interval. For the MediaMill dataset in sub-figure (a) one can see that our method *PR-FUBE* is always better than the *Entropy* method. Around ten concepts there is big performance gain. Afterwards MAP only increases very slightly. The *Entropy* method quickly falls stabilizes close to $0\%$ MAP. This is probably the case because of absences of concepts, contributing to relevance. For the performance of the Vireo detector set, see Sub-figure 6 (b). Here the combination of only two concepts show already a huge improvement for the two method *Entropy* and *PR-FUBE*. Our *PR-FUBE* reaches a maximum MAP at 22 concepts of $8.3\%$. From $50$ until around $150$ concepts the performance decreases again. However, from $150$ concepts onwards mounts again and stabilizes at $8\%$ The *Entropy* method drops nearly monotonically towards $0\%$ MAP.



(a) TRECVID 2005



(b) TRECVID 2007

**Figure 6: First $n$ Concepts through Retrospective Ordering of concepts by descending $\frac{P(C_i|R)}{P(C_i)}$ and Retrospective Parameter Estimation for $P(C_i|R)$**

## 5.3 Discussion

Our method performs well with user selected concepts (based on majority of users). However, this is only applicable in bigger collaborative search environments. The more realistic selection technique through Text IR still shows unsatisfying results. The reasons is - as other experiments show - an inadequate estimation of the parameter $P(C_i|R)$. However, if we simulate user feedback from the relevance judgments, *PR-FUBE* performs best among all three methods. In several retrospective experiments we show that an improvement of the $P(C_i|R)$ estimation can yield big enhancements in MAP. Taking these results our method performs for the TRECVID 2007 dataset only $0.5$ worse than the best full automatic search method from TRECVID 2007 by Tao et al. [14]. However, we expect that we can further improve our method by including other sources like the output of ASR, which we do not yet exploit. A positive property of our method is also its relative stability, especially in higher number concepts. This makes it robust against the choice of suboptimal concepts.

Unfortunately, the evaluation methodology in the TRECVID workshop does not allow a separate assessment of the different steps in the search system. So far, only the concept detection process is separately analyzed. The following steps are concept selection, impact or $P(C_i|R)$ estimation and combination. They are solely executed on the uncertain detector output. Therefore, the sources of bad performance are hard to trace.

We propose that the evaluation of the other steps of video search will help. In fact, big data sets with judged occurrence of concepts exist, for the evaluation of the detectors. Therefore, only relevance judgments for real queries on this data set are missing. The relevance of shots and the occurrences of a concept $C_i$ allow the calculation of $P(C_i|R)$. Measures like the mean square error could be used as evaluation criteria. The question, whether an adequate set of concepts was selected, could be assessed by the divergence of $P(R|\vec{C})$ and $P(R|S)$, which can be taken from the relevance judgments. Furthermore, the combination method can be benchmarked on the judged occurrences of the concepts and the ranking of $P(R|\vec{C})$.

# 6. CONCLUSION

We presented in this paper a probabilistic framework for unobservable binary events, which is directly derived from the Probabilistic Ranking Principle of Robertson [16]. We limited the scope to the occurrence of concepts and left the incorporation of ASR data for future work. A main contribution of the framework is the explicit model of relevance during the absence of concepts. We

motivated the need for this by an example of a TRECVID about Condoleezza Rice, where the shots with the simultaneous presence of three concepts were unlikely to give a satisfactory answer.

A crucial aspect of this framework is - besides the quality of the detector output - the estimation of the probability of occurrence of a concept given relevance. We estimated this probability either through a user selection, from Text IR scores of the concept description or through user feedback.

The experiments showed that a collaborative selection of concepts by multiple users achieved the best results. However, when using pure text IR scores our method performed suboptimal. We showed that this can be greatly improved through relevance feedback. In two retrospective experiments we demonstrated the performance of our model given better parameter estimations. With good estimations the performance proofs to be stable over the usage of many concepts.

# 7. REFERENCES


[1] J. Adcock. Fxpal interacive search experiments for trecvid 2007. In *Proceedings of the 7th TRECVID Workshop*, Gaithersburg, USA, October 2007.

[2] S. M. Aji and R. J. McEliece. The generalized distributive law. *Information Theory, IEEE Transactions on*, 46(2):325–343, 2000.

[3] R. B. N. Aly, D. Hiemstra, and R. J. F. Ordelman. Building detectors to support searches on combined semantic concepts. In *Proceedings of the Multimedia Information Retrieval Workshop, Amsterdam, The Netherlands*, pages 40–45, Amsterdam, August 2007. Yahoo! Research.

[4] A. Budanitsky and G. Hirst. Semantic distance in wordnet: an experimental, application-oriented evaluation of five measures. In *In Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, PA*, June. 2001.

[5] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. The MIT Press, 1998.

[6] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC*, pages 243–252, 1993.

[7] N. Fuhr. Probabilistic models in information retrieval. *Comput. J.*, 35(3):243–255, 1992.

[8] J. S. Garofolo, C. Auzanne, and E. M. Voorhees. The trec spoken document retrieval track: A success story. In *TREC*, 1999.

[9] C. Hauff, R. B. N. Aly, and D. Hiemstra. The effectiveness of concept based search for video retrieval. In *Workshop Information Retrieval (FGIR 2007), Halle, Germany*, volume 2007 of *LWA 2007 Lernen - Wissen Adaption*, pages 205–212, Halle-Wittenberg, 2007. Gesellschaft fuer Informatik.

[10] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. Pftijah: text search in an xml database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR), Seattle, WA, USA*, pages 12–17. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2006.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[12] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501, New York, NY, USA, 2007. ACM.

[13] L. Kennedy and A. Hauptmann. Lscom lexicon definitions and annotations (version 1.0). Technical report, Columbia University, March 2006.

[14] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, Y. Liu, Z. Gu, G.-J. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu. Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search. In *Proceedings of the 7th TRECVID Workshop*, Gaithersburg, USA, October 2007. To be published.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] S. Robertson. The probability ranking principle in ir. *J. Documentation*, 33:294–304, 1977.

[17] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.

[18] G. Salton. *The SMART Retrieval System; Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[19] N. Sebe. The state of the art in image and video retrieval. In *Image and Video Retrieval*, volume Volume 2728/2003, pages 1–8. Springer Berlin / Heidelberg, 2003.

[20] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. van Liempt, O. de Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. C. Thean, C. J. Veenman, and M. Worring. The mediamill trecvid 2007 semantic video search engine. In *Proceedings of the 7th TRECVID Workshop*, Gaithersburg, USA, October 2007. To be published.

[21] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.

[22] K. Sparck-Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments - part 2. *Information Processing and Management*, 36(6):809–840, 2000.

[23] L. van der Werff and W. Heeren. Evaluating asr output for information retrieval. In F. de Jong, D. Oard, R. Ordelman, and S. Raaijmakers, editors, *Proceedings of the ACM SIGIR Workshop âĂŹSearching Spontaneous Conversational Speech*, Enschede, 2007. CTIT.

[24] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Canegie Mellon University, 2006.

[25] W. Zheng, J. Li, Z. Si, F. Lin, and B. Zhang. Using high-level semantic features in video retrieval. In *Image and Video Retrieval*, volume Volume 4071/2006, pages 370–379. Springer Berlin / Heidelberg, 2006.