Reusing Annotation Labor for Concept Selection

Robin Aly
University Twente
P.O. Box 217,
7500AE Enschede, The Netherlands
r.aly@ewi.utwente.nl

Djoerd Hiemstra
University Twente
P.O. Box 217,
7500AE Enschede, The Netherlands
hiemstra@ewi.utwente.nl

Arjen de Vries
CWI, Amsterdam
TUD, Delft
The Netherlands
arjen@acm.org

ABSTRACT

Describing shots through the occurrence of semantic concepts is the first step towards modeling the content of a video semantically. An important challenge is to automatically select the right concepts for a given information need. For example, systems should be able to decide whether the concept "Outdoor" should be included into a search for "Street Basketball". In this paper we provide an innovative method to automatically select concepts for an information need. To achieve this, we provide an estimation for the occurrence probability of a concept in relevant shots, which helps us to quantify the helpfulness of a concept. Our method reuses existing training data which is annotated with concept occurrences to build a text collection. Searching in this collection with a text retrieval system and knowing about the concept occurrences allows us to come up with a good estimate for this probability. We evaluate our method against a concept selection benchmark and search runs on both the TRECVID 2005 and 2007 collections. These experiments show that the estimation consistently improves retrieval effectiveness.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Performance

Keywords

Video Information Retrieval, Semantic Concept Selection

1. INTRODUCTION

Concept based video retrieval systems have to select useful semantic concepts for an information need from their lexicon before using them for searching. Example semantic concepts are *Outdoor* or *Basketball* and they are sometimes also referred to visual concepts or high level features. In this paper we limit us to the textual representation of the query. Even then, it is hard to establish a link between the query keywords and helpful concepts since they are often not mentioned in the query. The Mutual Information between the concept and relevance has been proposed to quantify the helpfulness of a concept [6]. An important figure for calculating the Mutual Information is the occurrence probability of the concept in relevant shots. The contribution of this paper is to provide an innovative method to estimate this probability without requiring prior relevance information or user interaction. The result of this estimation can then be used to quantify the helpfulness of a concept.

Currently, the Video IR research community focuses on improving the detectors for semantic concepts. However, regardless of the detection technique every concept based retrieval system needs to answer the following questions: (1) What are good concepts for an information need? (2) Can they be effectively detected? And (3) how to employ them in search? We present an estimation method for the occurrence probability of a concept in relevant shots. With this probability we can approximate the Mutual Information of a concept and relevance. Therefore, the estimation helps answering two of the three questions: (1) Concepts which have a high Mutual Information are useful for search and (3) by providing weights which can be used for retrieval models that either depend on binary concept classification or probabilistic concept occurrence. However, this paper does not address question (2), whether a concept can be detected.

The problem of selecting concepts for an information need poses the following new challenges: First, a mapping between the query terms and the available concepts has to be found since many concepts will not directly be named in the query. Second, a measure for the helpfulness of each concept has to be estimated. Finally, the search system has to choose which concepts should be used. For example, in the query "Street Basketball" the unmentioned concept *Outdoor* will virtually always occur in relevant shots (nearly all relevant shots for "Street Basketball" are outdoor) while it will occur less in the rest of the collection. As a result, the Mutual Information between this concept and relevance will be high. Additionally, detectors for the concept *Outdoor* are often reliable due to the vast amount of training examples. Therefore, searching for shots which are *Outdoor* will help answering the query.

To train concept detectors it is a common practice to create a development collection in which humans annotate each

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIVR '09, July 8-10, 2009 Santorini, GR.
Copyright 2009 ACM 978-1-60558-480-5/09/07 ...$5.00.

shot with the occurrence or absence of all concepts from a lexicon. In order to estimate the occurrence probability of a concept in relevant shots we derive an artificial text collection from this development collection. For each shot we create one document which textually describes the content of the shot. Afterwards, we index the text collection using a standard Text IR system. At query time, the search process is divided into following steps:

1. Execution of the original text query on the artificial text collection,
2. Using the returned scores together with the human annotated concept occurrences to estimate the occurrence probability of a concept in relevant shots,
3. Selecting the concepts to use in the search,
4. Run a concept based search on the actual search collection using the estimated weights and the detector output.

Our method is similar to pseudo relevance feedback methods. However, for the initial search it uses a different search technique and a collection about which we have more accurate knowledge (the concept occurrences).

This paper is structured as follows. In Sec. 2 we present related work to this approach. Section 3 describes our estimation method and in Sec. 4 we describe the results of our experiments. The Sec. 5 presents our conclusions and proposes future work.

2. RELATED WORK

The estimation of the occurrence probability of a term w in relevant documents $P(w|R)$ has been investigated for decades in Probabilistic Text IR research; see Fuhr et al. [4] or Sparck-Jones et al. [17] for reference. The occurrence probability of a concept C in relevant shots, $P(C|R)$, which is estimated in this paper, is similar to this setting. Rocchio's Relevance Feedback Mechanism in [14] could be used in case we would know about concept occurrences and a user would give us relevance feedback on some documents. Croft and Harper investigate in [2] the case of an initial query where no relevance information is available. The method treats the top n documents as relevant and learns the search parameters from these documents. As our problem is similar to the above, this suggests that we could use the same techniques. However, the difference in our setting is that it is not obvious how to perform an initial search since we would need a set of concepts. Furthermore, if there was a good initial search method the occurrences of concepts in the ranking would not be observable which would make the estimation of the probability difficult.

There is little research on the evaluation of concept selection methods. Hauff et al. [5] were the first to use a human benchmark to evaluate concept selection. Here, humans had to select concepts they expected to be useful for the search for a topic. The evaluation was done by assuming these concepts were "relevant" concepts and a ranked list of concepts from a concept selection algorithm was evaluated by the Mean Average Precision (MAP) against these relevance judgments. A more extensive user study is provided by Huurnink et al. [8]. The additionally proposed collection benchmark establishes an order of the concepts by the descending Mutual Information to an information need. The proposed set agreement and rank correlation measures try to answer of following questions: (1) "Is this concept also selected by the benchmark" (set agreement) and (2) "How correlated is the order of concepts compared to the benchmark" (rank correlation). Until now there is no measure which quantitively evaluates the concept weights, such as the occurrence probability of a concept in relevant shots, which are employed for ranking. However, this would be worthwhile since the value of the weight plays a big role for the retrieval process.

Several works address the concept selection problem for a query. Natsev et al. present a comprehensive overview of the current methods in [11]. There, query expansion is equivalent to our concept selection. We recapitulate here the most related approaches. The WordNet thesaurus [3] is often used to find useful concepts, see for example [5] among others. Here, humans insert concepts into the graph of the thesaurus. The terms of a query are then used to find the semantic distance of the query to each concept. However, this distance reflects the semantic meaning and not the relationship of the concept occurrence in relevant shots. Therefore it will be difficult to transform it into the probability $P(C|R)$.

Hauff et al. use in [5] Text IR in a collection of textual concept descriptions to select good concepts. Wikipedia articles and concatenated WordNet Glosses are investigated as description sources. The score of a Text IR system is then used to measure the "closeness" of the concept to the relevant shots. Apart from the question of completeness of the concept descriptions, another unsolved problem of this method is that the score provided by the Text IR system – similar to the semantic distance – has to be transformed to a probability, should it be used in probabilistic retrieval methods. The proposed method in this paper extend the work from Hauff et al. by using artificial descriptions of whole video shots rather than descriptions of single concepts.

Furthermore, Natsev et al. [11] also propose a query expansion method (called "statistical corpus analysis") which learns the useful concepts from the term concept cooccurrences in the search collection. For this, the detector outputs are transformed to binary classifications and the ASR output is taken as the actual uttered words. Then, a likelihood ratio test is used to calculate a correlation coefficient of the cooccurrences. In the following, concepts which are strongly correlated to the query terms are selected for the search and a normalized version of the correlation coefficient is used as a concept weight. There are following differences to our method: First, the forced binary classification strongly depends on the detector performance. Second, in [11] the query terms are considered independently from each other. And finally, many concept will not occur together with the query terms since many queries ask for visual events, which will have no manifestation in the ASR.

We use the following two concept combination methods to examine the influence of the concept selection method: Zheng et al. propose in [19] a concept based retrieval method which uses the probability $P(C|R)$. They estimate the probability that a concept occurs in relevant shots through the probability of concepts occurring in the positive examples which are provided with the query. However, this approach neglects that a concept, which occurs in the example shot, might occur only at random or concepts which do not occur in the example might be useful. Since the authors did

not name the model we refer to this model as the Entropy method.

Recently, Aly et al. presented in [1] PRFUBE, a probabilistic retrieval method for the uncertain occurrence of concepts in shots, which estimates the probability of relevance, given a shot description of concept occurrences. In contrast to the Entropy method it considers both, the case in which the concept occurs and in which it does not.

3. ANNOTATION DRIVEN CONCEPT SELECTION

In this section, we present our method to estimate the occurrence probability of a concept given relevance for arbitrary concepts and textual queries which express the information need of a user. Let C be the random variable "concept occurrs" and R the "relevance to the information need" of a shot. As both are uncertain in the search collection we resort for the estimation of $P(C|R)$ to the development collection which has been completely annotated with the occurrences of all available concepts. This collection had to be created for the development of concept detectors. We assume that the concept occurrence in the development collection is similarly distributed as in the search collection which is a reasonable assumption for video data of the same domain. Clearly, in the case that the search collection is from another domain this assumption is strong. Nevertheless, we believe it is the best we can do in the situation.

We proceed as follows: In Sec. 3.1 we describe the construction of an artificial text collection from concept descriptions and the annotated development collection. Afterwards, in Sec. 3.2, we present the actual estimation method of the occurrence probability of a concept given relevance based on the scores of the query against the previously described text collection. Finally, in Sec. 3.4 we give an example of an estimation process.

3.1 Text Collection from Annotations

We create for each shot s in the development collection a textual description D_s. The descriptions in the text collection should be made in a way that a Text IR system can return a good ranking of the documents with regards to the relevance of the underlying video shot. Ideally, a text description meets two criteria: (1) it is precise (unambiguous) and (2) exhaustive, so that all words that a user could use to express his/her information need will be properly represented. Unfortunately, (1) and (2) contradict each other since a longer text inevitably introduces more ambiguity.

We now describe the components for the shot description D_s which are used in this paper. Clearly, words being said during the shot have a descriptive nature for its content so we include the output of an Automatic Speech Recognition System (ASR) in the description. However, many information needs will be concerned with the visual content of a shot so that the spoken words will be only of limited help. Therefore, we also add a textual description DC_i of each concept C_i which occurs in the shot – which we know from the human annotations. In this paper we investigate following concept descriptions: (1) The name and definition of the concept and (2) the content of a hand selected Wikipedia article about this concept. For the selection we used the article with the same name as the concept and if this resulted in an disambiguation page or a nonexistent page we manually picked an appropriate article. Therefore, for a given shot s in which the concepts $C_1, ..., C_n$ occur and with ASR output A the shot description is the concatenation of these components $D_s = (DC_1, ..., DC_n, A)$. Note that this is only one possible way of creating shot descriptions, we expect that it can be improved in future work.

After the creation of the text collection it is indexed by a standard Text IR system. This index is used during query time to efficiently find shots through their description, this fulfills the first step of the search procedure which was described in the introduction. The next section elaborates on how our method estimates the occurrence probability of a concept given relevance based on this ranking, which is the second step of the execution scheme in the introduction.

3.2 Probability of a Concept Given Relevance

For a particular query, we now have for each shot in the development collection its rank, the returned score from the Text IR system and the concepts occurring in it. In the following let C also be the set of all shots in the collection which contain the concept C. Furthermore, let $s_1, ..., s_m$ be a ranking of all m shots and $score(s_i)$ the score of shot s_i. In a perfect ranking with r relevant shots, the best estimate we can produce is defined as:

$$P(C|R) = \frac{\sum_{i=1, s_i \in \mathrm{C}}^{r} 1}{\sum_{i=1}^{r} 1} \tag{1}$$

That is, we divide the number of relevant shots containing the concept by the total number of relevant shots r, which is the definition of $P(C|R)$ for the development collection. However, in reality we will have, at best, a ranking where relevant shots are more often at higher ranks than irrelevant shots and r is unknown. Nevertheless, similar to Croft et al. in [2], we fix r to an empirical value. We refer to the estimation which assumes the first r shots are relevant as "certainty based". However, a high r will also take more irrelevant shots into account and probably misguides the following search in the actual search collection. Therefore we take here the score of the shots into consideration. This is done under the assumption that shots with a higher score are more likely to be relevant:

$$P(C|R) = \frac{\sum_{i=1, s_i \in C}^{r} score(s_i)}{\sum_{i=1}^{r} score(s_i)} \tag{2}$$

In (2) the sum of all scores of the first r shots in which the concept occurs is divided by the sum of the scores of the first r shots. The estimate is always normalized. This has the effect that shots which occur later in the ranking are considered but have less influence on the estimation. We name our estimation method "score-based". Now, a set of concepts can be selected and a Concept Based IR method, such as PRFUBE [1] or Entropy [19], can use these values to execute the third step of the retrieval procedure – to search on the actual search collection.

Clearly, the choice of the text retrieval model used in the initial query will play a substantial role in the performance of the concept selection. In the experiment section we evaluate two different retrieval models.

3.3 Order of Concepts

It is possible to calculate the Mutual Information of R and the C with the three probabilities $P(C|R)$, $P(C)$ and $P(R)$ using the law of total probability. The estimation of the first probability is subject of this paper. The second can be accurately estimated from training data. Therefore if we fix the prior probability of relevance to a small value, we arrive at an estimate for the Mutual Information. The calculation follows directly from its definition in [8]. We use this estimate to sort the concepts for a query and fixed number of the top-ranked concepts for search.

3.4 Example

In the following we give an example to illustrate the two steps described in Sec. 3.1 and Sec. 3.2: First, for the creation of the text collection let us assume we have three shots in the development collection with the following occurrence annotations (the output of the ASR system is ignored here):

	Basketball	Indoor	Outdoor	Sport
shot_1	1	0	1	1
shot_2	0	1	0	1
shot_3	1	1	0	1

Therefore, in Shot 1 the concepts *Basketball*, *Outdoor* and *Sports* occur. Shot 2 contains *Indoor* and *Sport* but not *Basketball*. Shot 3 shows *Basketball*, *Indoors* and *Sport*. The concept descriptions DC of the concepts could be as follows:

Basketball Basketball is a team sport [...]

Indoor Inside a building [...]

Outdoor Outside people walk on the street [...]

Sport Sport is an activity [...]

Now, we create for each shot a textual description by concatenating the descriptions of the occurring concepts. The resulting text documents look as follows:

Shot	Description
shot_1	Basketball is a team sport [...] Outside people walk on the street [...] Sport is an activity [...]
shot_2	Inside a building [...] Sport is an activity [...]
shot_3	Basketball is a team sport [...] Inside a building[...] Sport is an activity [...]

Afterwards, the estimation procedure for a concrete query proceeds as follows: Suppose the user entered the query "Street Basketball" and only shot_1 is relevant to the user (from hindsight). The Text IR system calculates the retrieval scores for the three shots $(0.9, 0.2, 0.4)$. Therefore the denominator of Eq. (1) is 3 and for Eq. (2) is 1.5. The estimates of the probability of the concepts occurring with in relevant shots are estimated as follows:

Probability	"Certainty-Based"	"Score-Based"	From Judgments
$P(Basketball\|R)$	0.66	0.93	1.00
$P(Indoor\|R)$	0.66	0.40	0.00
$P(Outdoor\|R)$	0.33	0.60	1.00
$P(Sport\|R)$	1.00	1.00	1.00

Table 1: Data Set Statistics

	TRECVID 2005	TRECVID 2007
Number of Shots	45765	18142
Domain	News Broadcast	General Dutch Television
Concept Vocabulary	MediaMill [16]	LSCOM [10]
Detector Output	MediaMill [16]	Vireo [9]
Number of Detectors	101	374
Number of Queries	24	24

We see that $P(Basketball|R)$ is set too low due to the incorrect assumption that the last two shots are also relevant. However, the "score-based" estimation is closer to the correct value than the "certainty-based" estimation. The concepts *Indoor* and *Outdoor* are both not set precisely, but the "score-based" method is more precise. The probability for the concept *Sport* is estimated correctly.

4. EXPERIMENTS

In this section we describe the experiments we performed to assess the effectiveness of our concept selection method. First, we describe in Sec. 4.1 the set up of the experiments. Second, Sec. 4.2 presents the evaluation of our concept selection method independently from the actual search, Sec. 4.3 evaluates the estimates on two TRECVID collections using real detector outputs. Sec. 4.4 investigates the effect of different shot descriptions. We end the experiments with a discussion of the results in Sec. 4.5.

4.1 Experiment Setup

Our experiments are based on the concept annotations from LSCOM [10] and MediaMill [16] on the development collection of TRECVID 2005 containing 43907 shots, see Over et al. [12]. Please refer to Tab. 1 for further statistics on the experiment data. We will use the development collection to create the text collection for the parameter estimation. Unfortunately we had to use two different detector sets (from MediaMill and Vireo), because for none were predictions for both search collections available. This makes it more difficult to interpret the results since the concept lexicon and the quality of the concept detectors differs. The search on the TRECVID 2007 collection demonstrates the performance of using a development collection from another domain for the estimation of the performance.

We evaluate our concept selection method through the average precision of the ranking according to [5] and the benchmark from [8]. We assume that all concepts with a positive Mutual Information for a query are "relevant". We do not report the set agreement measure because it does not contain any ranking information. Furthermore, we evaluate the results of the estimation through the overall search performance on two official TRECVID search collections from 2005 and 2007 and finally we investigate the influence of the chosen shot description on the search performance.

The initial Text IR runs were performed with the general purpose retrieval system PF/Tijah [7]. We used the NLLR [15] and BM25 [13] retrieval models to rank the artificial text collection. We found that the text search in the

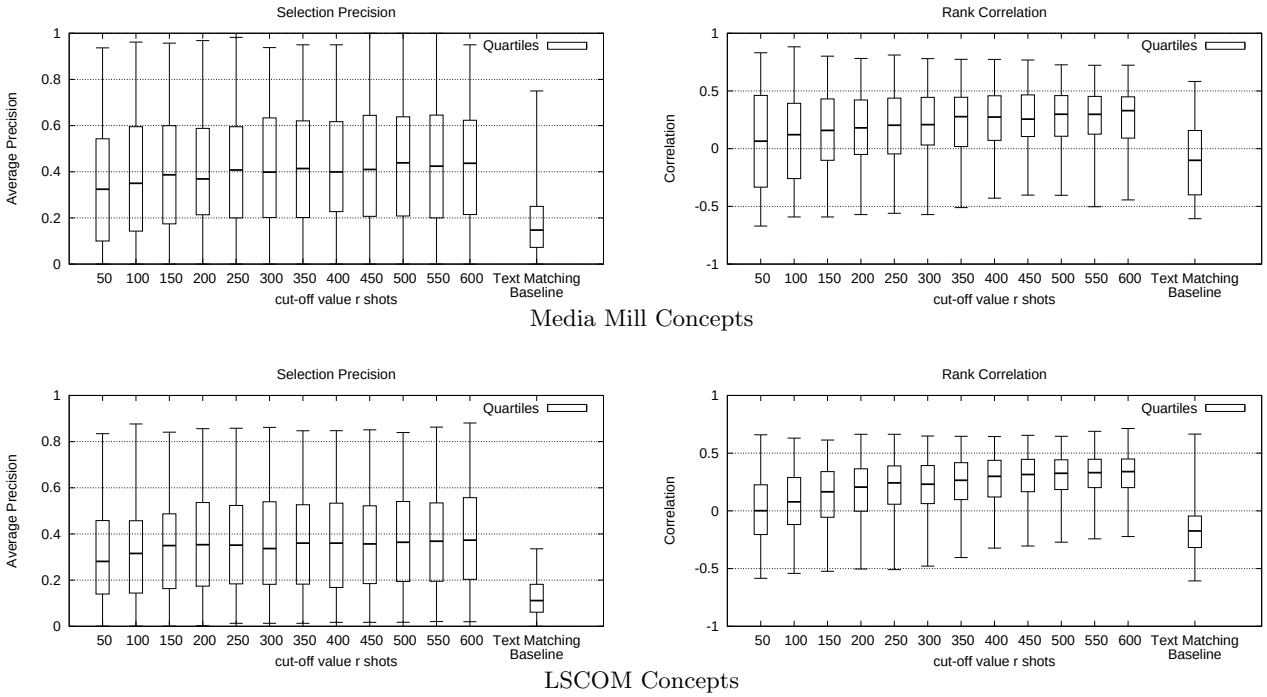

Figure 1: Concept Selection Evaluation

shot descriptions from Vireo concepts (Step 1 in the search procedure) resulted in poor performance. Therefore we use only the shot descriptions created from the MediaMill concepts in the following. Note that we can use the shot descriptions from MediaMill, resulting in a text collection, but perform our estimations using another concept lexicon as we are only using the text ranking from the shot descriptions of the first concept lexicon.

For the Concept Based runs we use the recently proposed retrieval models PRFUBE [1] and Entropy [19]. For a baseline we choose the method from Hauff et al. [5] where the Text IR scores for concept descriptions of single concepts are linearly scaled to form the probability $P(C|R)$. The parameters for the scaling are set according to [1] where the best setting was 0.05 for the lowest possible probability and a range of 0.6 for the TRECVID 2005 collection. For the TRECVID 2007 collection we used a range of 0.20. As a second baseline we use probabilities resulting from a survey of 21 users which were asked to select useful concepts was useful for each query [5]. Here, the probability was calculated as follows:

$$P(C|R) = \frac{numUser(C)}{numUser()}$$

where $numUser(C)$ is the number of users who selected the concept C and $numUser()$ is the total number of users.

Our approach depends on a good precision of the initial Text IR run. As a preliminary indicator we evaluated the MAP of the 24 TRECVID 2005 queries on the development collection [1]. This resulted in 0.26 MAP. We consider this measure sufficient but the overall search would probably benefit from better results.

[1]Relevance judgments were kindly provided by Rong Yan formally at Carnegie Mellon University [18]

4.2 Concept Selection

Figure 1 shows a comparison of our "score-based" estimation method and the "Text Matching" baseline method from [8] using the collection benchmark. We choose the collection benchmark since our method aims to find concepts for a particular collection and not concepts which a users finds useful. We report the performance measures separately for the MediaMill and the LSCOM concepts.

The left plot shows for each concept lexicon the distribution of the average precisions of the selected concepts at the cut-off of r shots. We see that for both concept lexicon the median of the "Text Matching" baseline lies beneath the first quartile of every cut-off value. In the MediaMill lexicon there is always at least one query where the selected set of concepts does not contain any concepts of the benchmark. With the LSCOM concept lexicon all queries have at least a small average precision with the selected set by the benchmark. Furthermore, the maximum precision is almost constant over different cut-off values for both concept vocabularies.

The right plot shows for each concept lexicon the rank correlation measure quartiles per cut-off value. We can see that from a cut-off value of 400 for the MediaMill concepts and 150 for the LSCOM concepts 75% of the selected concepts through the "score-based" estimation have a higher rank correlation with the benchmark than the "Text Matching" baseline. The maximum rank correlation stays almost constant for all cut-off values. The same holds for the minimum rank correlation with the MediaMill concept set. However, for the LSCOM concept set the minimum rank correlation increases monotonically from a cut-off value of 300.

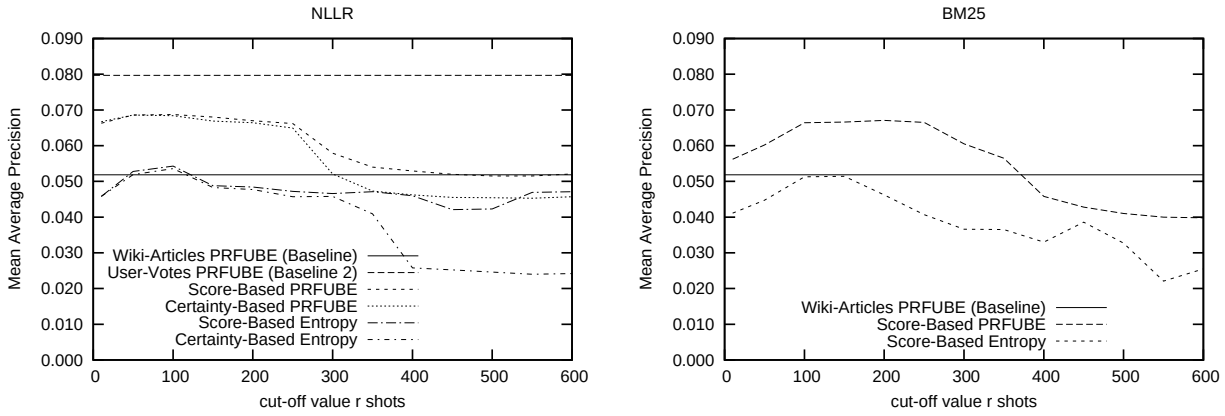

Figure 2: TRECVID 2005 / Media Mill Detectors

4.3 Official TRECVID Runs

The left part of Fig. 2 shows results of the two PRFUBE and two Entropy runs for the TRECVID 2005 search task using the NLLR text retrieval model. The chosen number of concepts used here was experimentally set to produce the optimum performance which was 10 concepts for the PRFUBE model and 7 for the Entropy method. On the X axis we see the cut-off value r which is the number of shots we consider at the top of the Text IR ranking. The Y axis depicts the mean average precision (MAP). Here, PRFUBE performs best with both estimation alternatives and achieves a MAP of 0.070 at a cut-off value of $r = 150$ shots. Afterwards, around $r = 250$ both alternatives decline. The "score-based" alternative stabilizes around 0.055 MAP while the "certainty-based" alternative drops further and reaches a MAP of 0.046 at a cut-off value of $r = 600$. The "score-based" estimations help the Entropy method to achieve a MAP of 0.054 at cut-off $r = 150$. Afterwards, it stabilizes around 0.046. The "certainty-based" method causes the same maximum but drops to 0.020 afterwards. The baseline from Hauff et al. [5] achieves a 0.051 MAP. The second baseline, where the parameters are estimated through user votes, achieved 0.079 MAP, is with 0.079 the best achieved result. Both baselines do not depend on the cut-off value, therefore they are drawn as constant lines in the graph. The improvement of the estimations for PRFUBE at a cut-off value of $r = 150$ against the baseline from Hauff et al. [5] was successfully tested for significance using a Wilcoxon signed-rank test with a significance level of 0.05.

In the right part of Fig. 2 we repeat the experiment using BM25 text retrieval model. For readability we only show the "score-based" alternatives. For cut-off values of $r < 200$ the graphs look similar. However, with increasing r the MAP of both ranking models, PRFUBE and Entropy, decreases more than with the NLLR model.

In left part of Fig. 3 we show the results of the two PRFUBE and two Entropy runs for TRECVID 2007. The number of concepts was chosen in the same ways as for the TRECVID 2005 dataset. Here, the optimum was 45 concepts for the PRFUBE model and 15 for the Entropy model. The axes are also the same as in Fig. 2. We can see that PRFUBE with the "score-based" estimation method benefits from a cut-off value up to $r = 2500$ with 0.036 MAP.

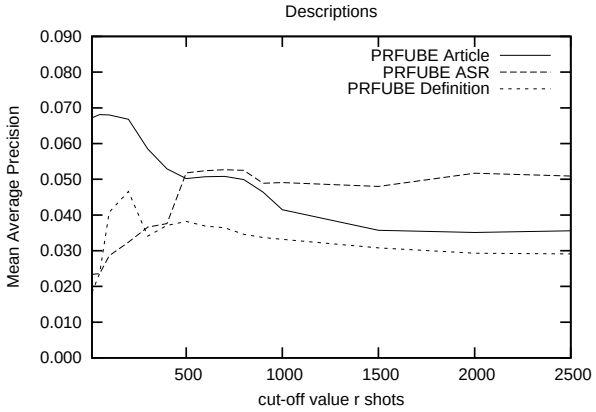

Figure 4: Comparison of Description Sources

Afterwards the performance drops until 0.030 MAP and stabilizes there. The "certainty-based" method helps PRFUBE to achieve the maximum of 0.034 MAP at $r = 2500$ but degrades afterwards to around 0.025. Both estimation methods cause the Entropy model to degrade quickly from a MAP of 0.023 to virtually zero. The baseline only achieves 0.013. The improvement of PRFUBE through the estimations compared to the baseline was significant according to a Wilcoxon signed-rank test with significance level 0.05.

The right part of Fig. 3 shows the performance of the PRFUBE and Entropy ranking model using the "score-based" estimation alternative using the BM25 ranking model to score the text collection. The performance of the PRFUBE retrieval model slowly decreases from a MAP of 0.025 to a MAP of 0.020. The Entropy method has a performance peak at a cut-off value of $r = 2500$. Afterwards the MAP decreases to 0.010.

4.4 Shot Description Sources

In this Section we investigate the impact of the used shot description on the search performance. Figure 4 shows search runs using the NLLR text retrieval model and the PRFUBE combination method using following description types: wikipedia articles and ASR and definitions (label: Article), only ASR

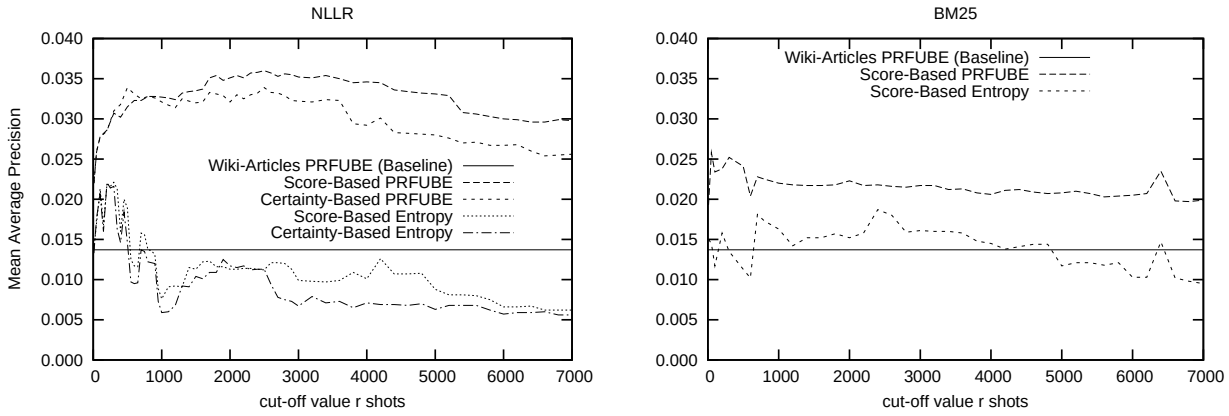

Figure 3: TRECVID 2007 / Vireo Detectors

output and only the concept definitions. The axis are the same as in the previous experiments. The combination of all descriptions shows the best result. The ASR description show worse performance but stabilize on a higher level with high cut-off values. The definitions show always the worst performance.

4.5 Discussion

The experiments carried out here show the success of the proposed methods to select useful concepts. The selection method is evaluated by using a concept selection benchmark and through the search performance of two retrieval systems against the isolated Wikipedia article approach. The evaluation of the Text IR search on the development collection showed reasonable results which is a requirement for the retrieval from the search collection to be successful.

The evaluation of our concept selection strategy with the benchmark from [8] shows that with a growing cut-off value, the distribution of average precisions stays stable and the rank correlation improves. However, since both measures only consider the ranking of the concepts and not a numeric weight, here the occurrence probability of a concept in relevant shots, these results can only be used as preliminary indications of the quality of the concept selection.

The "score-based" estimation method significantly improved the results of the PRFUBE retrieval model compared to the baseline. For the TRECVID 2005 search collection, which is similar to the development collection, cut-off values of 150 show the best improvements. The Entropy ranking model does not benefit from this estimation method as much.

Considering TRECVID 2007, it is interesting that the performance in collections from another domain than the development collection increases until a high cut-off value of 2500 shots. However, the performance is much lower than for the TRECVID 2005 task. The reason is the domain change from broad cast news to general television, which causes a performance drop of the detectors. This is an acknowledged fact in the TRECVID community. Furthermore, as the detectors from MediaMill and Vireo are built using different techniques the general performance might be different. The reason for the high cut-off values might be that there are few really relevant shots for the TRECVID 2007 topics in the TRECVID 2005 development collection and the estimation method only finds related concepts, which might occur more scattered over the ranking.

The Entropy model was not improved as much by neither estimation method. The reason most likely is that the method ignores the possibility that a shot might be relevant in the absence of a concept. This effect gets stronger with a higher number of concepts. The presented estimation alternatives are stable against the use of different retrieval models for the text search. Furthermore, the evaluation of the textual description types revealed that the combination of Wikipedia articles, ASR and concept descriptions.

The second baseline, created by asking users, beats our estimation method in the search task of TRECVID 2005. At this point, we can only propose that the concepts selected by users are better transferable to the search collection. However, due to the labor costs of such a concept selection method this method is hardly usable in practice.

5. CONCLUSION AND FUTURE WORK

This paper introduced a novel method to select useful concepts for an information need. It does so by estimating the occurrence probability of a concept in relevant shots. It is based on the construction of an artificial text collection derived from the development collection which is used to train concept detectors. First, the textual query is evaluated on this collection to estimate for each concept the occurrence probability of this concept given relevance. The "certainty-based" alternative assumes a fixed number of r shots to be relevant. The "score-based" estimation method uses the Text IR score as a confidence weight of the shot being relevant. According to the concept selection benchmark from [8] the "score-based" alternative outperforms the "Text Matching" baseline provided by the benchmark creators. In both assessed search collections, TRECVID 2005 and 2007, the selection enable the recently proposed ranking models PRFUBE to perform stable and increase its MAP significantly at low values of r. The consideration of the retrieval score in the "score-based" estimation consistently gave better results.

We have limited ourselves to using Wikipedia resources, which are freely and easily available. However, we expect other – especially more domain related – text sources to show better results.

Finally, in future work, we will look into the possibility of

estimating the probability of combinations of multiple concepts given relevance by extending our estimation method. This could prevent unwanted side effects through incorrectly assumed conditionally independence assumptions.

6. REFERENCES

[1] Robin Aly, Djoerd Hiemstra, Arjen P. de Vries, and Franciska de Jong. A probabilistic ranking framework using unobservable binary events for video search. In *CIVR '08: Proceedings of the International Conference on Content-Based Image and Video Retrieval 2008*, pages 349–358, New York, NY, USA, 2008. ACM.

[2] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.

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

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

[5] Claudia Hauff, Robin Aly, and Djoerd 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.

[6] A. Hauptmann, Rong Yan, Wei-Hao Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. In *IEEE Transactions on Multimedia*, volume 9-5, pages 958–966, August 2007.

[7] Djoerd Hiemstra, Henning Rode, Thomas van Os, Roel, and Jan 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.

[8] Bouke Huurnink and Maarten de Rijke Katja Hofmann. Assessing concept selection for video retrieval. In *Proceedings of the first MIR Conference'08*, Oktober 2008.

[9] Yu-Gang Jiang, Chong-Wah Ngo, and Jun 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.

[10] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[11] Apostol (Paul) Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 991–1000, New York, NY, USA, 2007. ACM.

[12] W. Kraaij A. F. Smeaton P. Over, T. Ianeva. Trecvid 2005 an overview, 2005.

[13] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[14] J. J. Rocchio. *The SMART Retrieval System - Experiments in Automatic Document Processing.*, chapter Relevance feedback in information retrieval., pages 313–323. Prentice Hall., 1971.

[15] Henning Rode and Djoerd Hiemstra. Using query profiles for clarification. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *ECIR*, volume 3936 of *Lecture Notes in Computer Science*, pages 205–216. Springer, 2006.

[16] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold 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.

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

[18] Rong Yan and Alexander G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484, October 2007.

[19] Wujie Zheng, Jianmin Li, Zhangzhang Si, Fuzong Lin, and Bo 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.