

Assessing Effectiveness in Video Retrieval^{*}

Alexander Hauptmann and Wei-Hao Lin

Language Technologies Institute,
Carnegie Mellon University,
5000 Forbes Avenue,
Pittsburgh, PA 15213, USA
{alex, whlin}@cs.cmu.edu

Abstract. This paper examines results from the last two years of the TRECVID video retrieval evaluations. While there is encouraging evidence about progress in video retrieval, there are several major disappointments confirming that the field of video retrieval is still in its infancy. Many publications blithely attribute improvements in retrieval tasks to the different techniques without paying much attention to the statistical reliability of the comparisons. We conduct an analysis of the official TRECVID evaluation results, using both retrieval experiment error rates and ANOVA measures, and demonstrate that the difference between many systems is not statistically significant. We conclude the paper with the lessons learned from both results with and without statistically significant difference.

1 Introduction

More people use video and television as information sources, yet video retrieval still lags far behind text retrieval in terms of effectiveness. In recent years, video retrieval research has been a very active field, and many different approaches to video retrieval have been proposed both in the non-interactive setting [1, 2, 3] and the interactive setting [3, 4, 5].

What makes a difference in video retrieval? This paper examines the question by looking at the TRECVID 2003 and TRECVID 2004 official results, which evaluated video retrieval systems in a variety of conditions and on a variety of topics. Our initial goal was to examine the rankings of the evaluated systems, analyze the descriptions and establish which approaches were more effective than others. However, as soon as we took a closer look at the differences in evaluation scores between the systems, it quickly became clear that many evaluation score differences are very small, hinting that any conclusions drawn from the score difference alone may be inconclusive. We began to investigate which differences were statistically significant when comparing results in the TRECVID evaluation. Much to our surprise, the significant differences were relatively sparse.

^{*} This work was supported in part by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037.

Since it also turned out to be very difficult to determine from the descriptions which approaches were part of the system versions that made a real difference, we reduced our analysis to comparing versions of our own submitted systems, when they produced statistically significant results.

The paper is organized as follows. We first describe the NIST TRECVID evaluations in 2003 and 2004 in Section 2, followed by an analysis based on retrieval experiment error rate (REER) to determine whether a system is reliably better than the other in Section 3. An alternative analysis is provided in Section 4 through analysis of variance (ANOVA) and the Newman-Keuls pairwise significance tests, where we also introduce the idea of pseudo-grouping to summarize statistically significant comparisons succinctly. We present the empirical results of our analysis on various TRECVID submission conditions and describe, based on our own participation, which approaches seemed to be effective in Section 5. Finally, Section 6 provides a discussion and summary of the findings.

2 NIST TREC Video Retrieval Evaluations in 2003 and 2004

TREC Video Retrieval Evaluations (TRECVID) is an independent evaluation forum devoted to research in content-based retrieval of digital video [6]. Its goal is to encourage research in information retrieval from large amounts of videos by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. TRECVID focuses on the single shot as the unit of information retrieval rather than the scene or story/segment/movie. The TRECVID test corpora for 2003 and 2004 consisted of broadcast news from ABC, CNN, and C-SPAN (for 2003), with 32,318 reference shots in the test video corpus for 2003 and 33,367 reference shots in 2004.

The nontrivial size of the corpus, its definitions of sets of information needs (topics), and human-determined truth for the topics provide a starting point for scientifically valid evaluations and comparisons. Taking advantage of this framework, a total of 135 runs were submitted for search results in 2003, and participation grew to 219 runs in 2004. For the search tasks, there were 61 interactive runs in 2004 on 23 topics (37 on 25 topics in 2003), 52 “manual” search runs in 2004 (38 in 2003), where a manual run gives the researcher 15 minutes per topic to “translate” the information need into a form suitable for the system. Finally, there were 23 fully automatic runs in 2004 (this condition was not not evaluated in 2003). More detailed information can be found at the NIST TREC Video Track web site, where interested readers are referred to the complete descriptions on the TRECVID guidelines[7].

3 Retrieval Experiment Error Rate

As a first approach, we examined the TRECVID evaluation results with the retrieval experiment error rate, as suggested by Voorhees and Buckley [8]. Re-

trieval experiment error rate (REER) is motivated to evaluate the reliability of evaluation results in a retrieval experiment like TRECVID. REER is defined as the probability of making opposite effectiveness judgments about two systems over two sets of topics based on a common evaluation metric, like Mean Average Precision (MAP). If we make an effectiveness statement about two retrieval systems (or two submission runs in the TRECVID setting) based on that evidence that one system has higher MAP over a set of topics, REER is the likelihood that the effectiveness judgment will be reversed, i.e. an experiment error, if we compare the two systems on another set of topics. Intuitively, if two video retrieval systems are equally effective, we would expect to observe that one system performs better than the other only half of them time, i.e. REER is 0.5, when two systems are repeatedly over different sets of topics. Therefore, only when REER is much lower than 0.5 can we have much confidence that one retrieval system is significantly more effective than the other. By calculating the REER of retrieval experiments, we can obtain better insights into the reliability of the score difference between systems instead of assuming that System X is better than System Y merely because the MAP of System X is minimally larger. When it may not be appropriate to make normality assumptions¹, REER provides an alternative tool to objectively evaluate if the score different is meaningful. Note that REER does not make assumptions on the score distributions and is not designed to be a statistical test.

REER can be estimated in the frequentist manner [8] by counting how often retrieval experiment errors occur, but this method requires large amounts of retrieval experiment results, and an extra curve fitting step is needed in order to extrapolate a prediction of the REER for larger topic set sizes. Instead, we estimate REER by directly following the theoretical analysis in [9] and estimate REER in the following equation,

$$\text{REER} = 2\Phi\left(\frac{-(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{|\mathcal{T}|}}}\right) \left(1 - \Phi\left(\frac{-(\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{|\mathcal{T}|}}}\right)\right) \quad (1)$$

where μ_X, μ_Y and σ_X^2, σ_Y^2 are the means and variances of MAP probability distribution of the two systems X and Y, respectively, \mathcal{T} is the set of topics in the retrieval experiment, Φ is the standard normal distribution function.

3.1 REER of TRECVID 2003 and 2004

We estimate REER of the TRECVID 2003 and 2004 evaluation results based on Equation 1, and the REER curves of various MAP differences are plotted

¹ We perform Anderson-Darling test for normality on TRECVID 2003 and 2004 retrieval submissions. After controlling the False Discovery Rate at the level of 0.05 using Benjamini-Hochberg procedure, 92.82% of the 209 runs are rejected to be normally distributed.

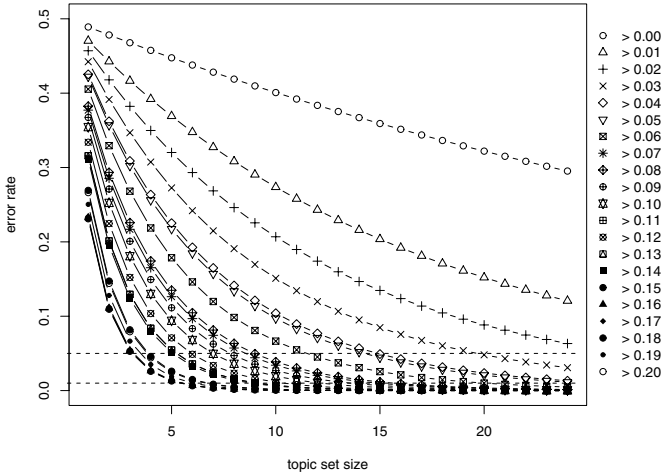


Fig. 1. The Retrieval Experiment Error Rate (REER) curves are estimated from all search submission runs in TRECVID 2003 and 2004. Each REER curve represents different MAP difference as a function of the topic set size. The topmost curve stands for the MAP difference greater than 0 but less than 0.01, and the second curve stands for the difference greater than 0.01 but less than 0.02, and so on. Two horizontal dashed lines are drawn at the REER of 0.01 and 0.05, respectively

in Figure 1. We consider only submission runs that answer all the topics. In order to independently sample two sets of topics of equal size and calculate the MAP difference between two systems, a set of 12 topics for each experiment is the maximum number we can draw from typical 25 search topics in one TRECVID year. The sample means and variances of two systems in (1) is hence estimated at the topic set size of 12, and REER is extrapolated to topic set size of 25².

If we follow the dashed line of REER 0.05 and look the MAP difference at the topic set size of 25, which is typical in TRECVID, we can see that the MAP difference must be greater than 0.02. Many MAP difference between submission runs in TRECVID, especially for manual runs (see Figures 4,5, and 7), are less than 0.02, and REER suggests that conclusions drawn from MAP differences less than 0.02 are unlikely to hold in other retrieval experiments at the error rate of 0.05. If we want to be really confident and make comparisons at the stringent error rate of 0.01, the MAP difference between two systems must be greater than 0.05, which renders most video retrieval systems indistinguishable in terms of effectiveness.

² Extrapolation may be avoided by combining two years' results together, resulting in total 50 topics. However, most retrieval systems changes across years, and we have no way to tell which run in 2003 is the same run in 2004 based on descriptions in the workshop papers only.

4 ANOVA and Pair-Wise Significance Tests

As an alternative method, we apply an Analysis of Variance (ANOVA) approach to determine how well TRECVID evaluation results can be explained by topics and systems. Instead of applying multiple t -test and suffering from the multiple testing problem, where random differences appear significant if enough experiments are preformed, the Newman-Keuls test is used to estimate is the pairwise MAP difference between two systems is statistically significant. ANOVA has been shown to be very robust to violations of the assumption that errors are normally distributed, which is why it is so heavily used in psychology³.

4.1 The ANOVA Model

We use a standard Analysis of Variance (ANOVA) repeated measurements design [10] to analyze the data for statistical significant differences. ANOVA models the average precision scores $Y_{i,j}$ of System j for Topic i as a combination of effects in the following formula,

$$Y_{i,j} = M + t_i + r_j + e_{i,j} \quad (2)$$

where M is the global mean for all topics and systems, t_i is Topic i mean for all systems, r_j is System j mean average precision for all topics, and $e_{i,j}$ is the error term, which “explains” the rest of the $Y_{i,j}$ score as due to random measurement noise.

ANOVA allows us to compute the probability that this model can explain the data. The resulting confidence probability p asserts the rejection of the null hypothesis, i.e. that all data comes from the same distribution according to the model.

4.2 The Newman-Keuls Test of Pairwise Significance

For TRECVID data, we generally find that there is a significant effect due to topic and system differences overall, but we also want to find out which pairs of differences are significant, and which are not. The method we used for this is the Neumann-Keuls post-hoc test of pairwise significance. While neither the most conservative or generous test, Newman-Keuls has the advantage that it takes the number of pairwise comparisons into account when computing the significance and adjusts the significance criterion. The reasoning is that if you make many pairwise comparisons on randomly selected data, some will seem to be significant, and the Neumann-Keuls test raises the bar for each additional comparison. This avoids a situation where several hundred t -test are performed at the $p < 0.05$ significance level, and some appear significant due to random sampling effects. Alternative (and in many ways comparable) tests would be Tukey’s test or Scheffe’s test.

³ In our case, the actual estimation is complicated by the fact that we only have one value in each cell.

The Neumann-Keuls test first arranges all means in descending order. According to the statistic, different cells now have different “critical differences”, depending on the mean square error, the degrees of freedom and a so-called r value. The r value is obtained from the difference in the number of comparison between compared cells.

4.3 Pseudo-grouping

As a practical matter, we find that many pairwise differences are significant [11], as shown in Figure 2 which shows just the top 30 submissions for 2004 interactive search, and many others are not, with no easy way to spot “groups” of equally effective systems.

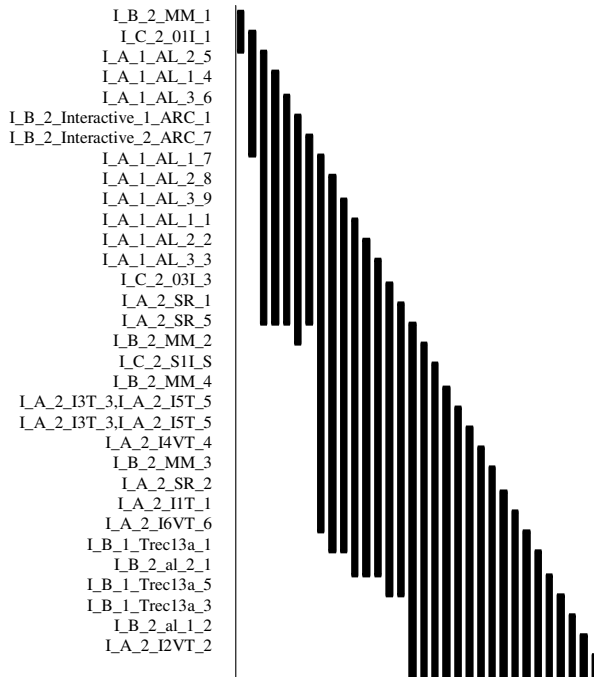


Fig. 2. The top interactive video retrieval systems in TRECVID 2004 ranked in descending order by MAP. Systems covered by the same vertical bar have no significant differences

Thus we introduce serial pseudo-grouping of systems, where each (pseudo) group, going in order from highest MAP to lowest, has no significant differences **and** there cannot be any overlap between groups. For example, if there is no difference between System 1 and System 3 but there is difference between System 1 and System 4, then System 1 through 3 are in one group, and the next group starts might at System 4. The complete data of interactive runs from Figure 2

is plotted this way in Figure 3. In this format all differences, significant or not, of systems in the middle of a group with others in the middle of another group, get ignored. This provides a fairly concise summary of the data, emphasizing the distinctions among the top systems, which is usually what researchers care about. However, the pseudo-groups might mislead readers to think all systems in one group are equivalent **and** better than all systems in the next group, when instead, the interpretation should be that there are no significant differences within the group, and at there is least one significant difference between the best member of the group and the best member of the next group.

5 Analysis of Results

We perform ANOVA and the Newman-Keuls tests on the TRECVID 2003 and 2004 search evaluation results, and summarize the pairwise significance results in the pseudo groups. The analysis of variance finds strong significant effects for topics and systems in all 2003 and 2004 tests at $p < 0.001$. The red bars indicate the runs that are CMU submissions, for which we can distinguish what aspects of retrieval made a difference. Unfortunately, we can only speculate what happened in other systems' submissions, but we can describe with certainty what among our own approaches made a significant difference.

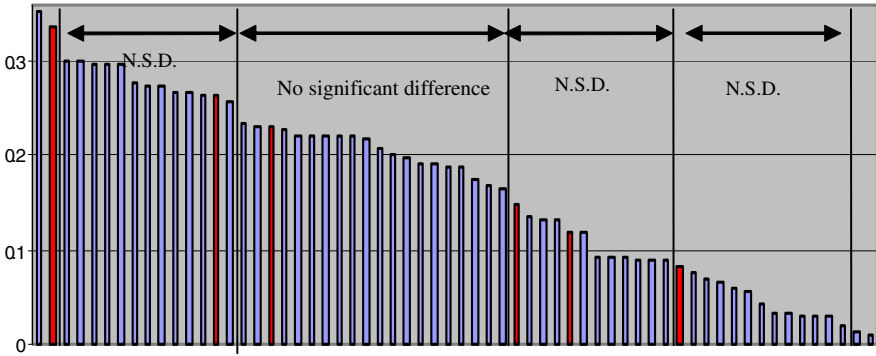


Fig. 3. TRECVID 2004 Interactive Search Results. Systems are ordered by descending MAP. Arrows show pseudo-groups without significant differences

TRECVID 2004 Interactive Search (see Figures 2, 3): Looking at the pseudo-groupings of pairwise differences, for the 2004 systems, we find that the top 2 interactive systems are not significantly different, followed by the next group of runs that are not significantly different ranging down to rank 15. Notice that the highest MAP difference between adjacent pseudo groups is around 0.5, which stands for REER 0.01. Therefore, the statistically significant MAP difference between the two runs using the Newman-Keuls test is consistent with REER.

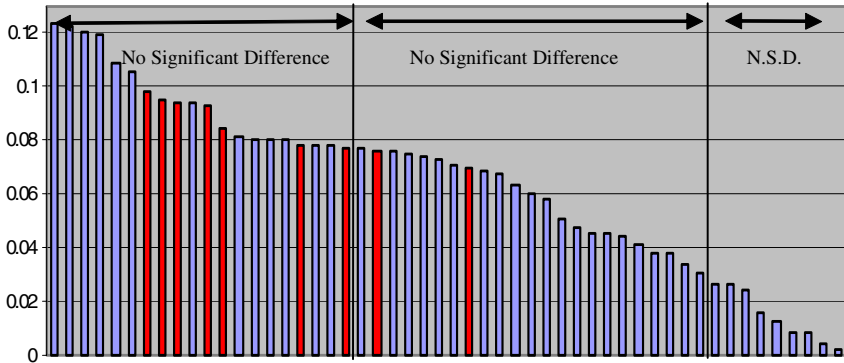


Fig. 4. TRECVID 2004 Manual Retrieval Results

In the case of the interactive runs, we find that our top ranked system was by an expert using the full features of the Informedia system[?]. The second red bar, significantly worse than the expert, is the submission by a novice user with the full system. The last two red bars, again significantly worse than the other CMU submissions, are both expert and novice users, but using a system that did not exploit any textual information such as speech recognition transcripts.

TRECVID 2004 Manual Retrieval (see Figure 4): In the manual retrieval condition, we find that the top 11 runs are not significantly different. Among the red CMU submissions, we find that none of the runs are significantly different from each other, including the baseline of retrieval based only on the transcript text from speech recognition. This holds for all our submissions, as confirmed by a full pairwise analysis between our run, despite the fact that they appear in different pseudo-groups. None of the more sophisticated video retrieval techniques provide a significant boost over text baseline.

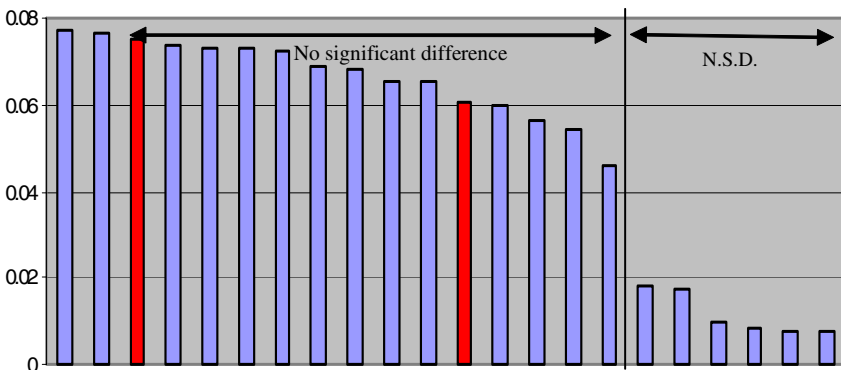


Fig. 5. TRECVID 2004 Automatic Retrieval Results

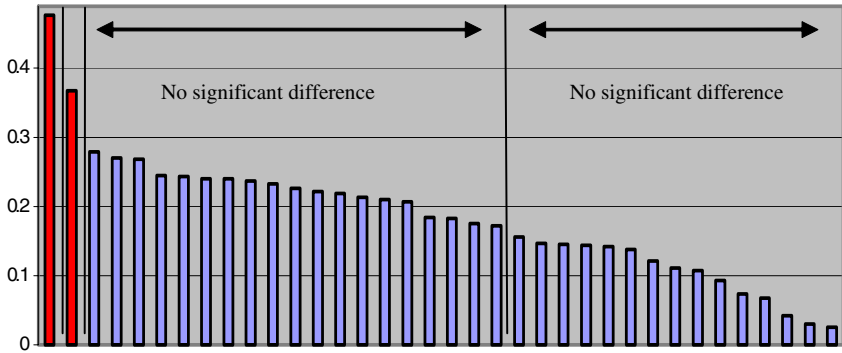


Fig. 6. TRECVID 2003 Interactive Search Results

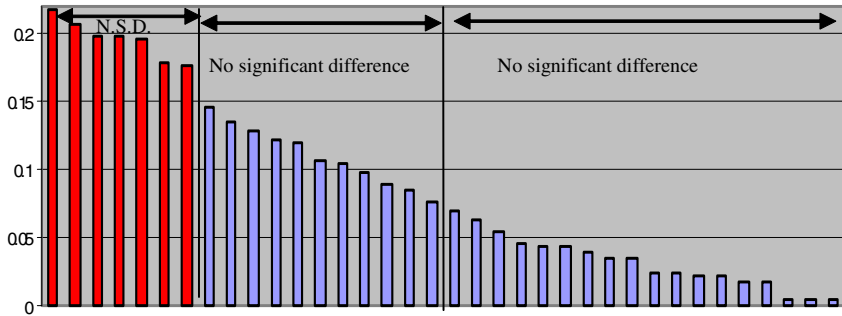


Fig. 7. TRECVID 2003 Manual Search Results

TRECVID 2004 Automatic (see Figure 5): The top 16 automatic runs in 2004 were also not significantly different. This includes the two CMU submissions.

TRECVID 2003 Interactive Search (see Figure 6): For 2003, our expert interactive system at rank 1 is significantly different from our non-expert system at rank 2. Both of these are significantly different from a number of other submissions, whereas the next 19 systems are statistically indistinguishable.

TRECVID 2003 Manual Retrieval (see Figure 7): In the 2003 manual submission runs, we find that the top seven runs are not significantly different from each other, even though much time has been spent interpreting the results of different multimedia combination schemes [4]. The text baseline used for these runs distinguishes it from the following pseudo-group, but none of the additional multimedia analysis techniques result in a significant improvement. The next (pseudo-)group of ten systems is also not significantly different from each other.

6 Discussion and Conclusions

There are several lessons to be learned from this analysis of the data. The first one is, of course, that one should not believe all the hype surrounding effective techniques in video retrieval. Too often small differences are interpreted as substantial, even though they may just reflect uncertainty in measurement. Both the retrieval experiment error rate and ANOVA analysis give a strongly consistent interpretation of the results, and MAP difference of 0.05 between two retrievals is the minimal value to have a meaningful difference. Our data provides consistent evidence, across two years, that there are no clearly distinguished effective techniques for either manual or automatic video retrieval. Perhaps the relatively small number of topics is to blame; compared to the standard text retrieval evaluations, 25 and 23 search topics per year makes it very difficult to ascertain significant differences. If we take the risk of over-generalizing results in Figure 1 and continue the REER curves, we could justify 0.02 MAP difference at the error rate level of 0.01 if we conduct retrieval experiments with 50 topics, but this will pose a significant burden on the TRECVID organizers.

What is disappointing about our analysis is that we repeatedly find that none of the multimedia analysis and retrieval techniques provide a significant benefit over retrieval using only textual information such as ASR transcripts or closed captions. This is actually consistent with findings in the earlier TRECVID evaluations in 2001 and 2002, where the best systems were based exclusively on retrieval using automatic speech recognition. However, we should also point out that it is not the case that “nothing works” here. In interactive systems, we do find significant differences among the top systems, indicating that interfaces can make a huge difference for effective video search. Not surprisingly, from comparisons of our own data, we find that expert users significantly outperform novice users, and visual only systems that do not exploit broadcast news speech transcripts are significantly inferior to systems that exploit all available knowledge. While in 2003, there were big, significant gaps between the top systems, that difference shrunk in the 2004 TRECVID interactive submissions, indicating that the knowledge about effective interactive search systems is more broadly disseminated.

References

1. Ianeva, T., Boldareva, L., Westerveld, T., Cornacchia, R., Hiemstras, D., de Vries, A.P.: Probabilistic approaches to video retrieval. [6]
2. Chua, T.S., Neo, S.Y., Li, K.Y., Wang, G., Shi, R., Zhao, M., Xu, H.: TRECVID 2004 search and feature extraction task by NUS PRIS. [6]
3. Amir, A., Argillander, J.O., Berg, M., Chang, S.F., Hsu, W., Iyengar, G., Kender, J.R., Lin, C.Y., Naphade, M., Natsev, A.P., Smith, J.R., Tesic, J., Wu, G., Yan, R., Zhang, D.: IBM research TRECVID-2004 video retrieval system. [6]
4. Yan, R., Yang, J., Hauptmann, A.G.: Learning query-class dependent weights in automatic video retrieval. In: Proceedings of the Twelfth ACM International Conference on Multimedia. (2004) 548–555

5. Hauptmann, A., Chen, M.Y., Christel, M., Huang, C., Lin, W.H., Ng, T., Papernick, N., Velivelli, A., Yang, J., Yan, R., Yang, H., Wactlar, H.D.: Confounded expectations: Informedia at TRECVID 2004. [6]
6. Proceedings of the TREC Video Retrieval Evaluation 2004. In: Proceedings of the TREC Video Retrieval Evaluation 2004. (2004)
7. NIST: Guidelines for the TRECVID 2004 evaluation. Webpage (2004) <http://www-nlpir.nist.gov/projects/tv2004/tv2004.html>.
8. Voorhees, E.M., Buckley, C.: The effect of topic set size on retrieval experiment error. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press (2002) 316–323
9. Lin, W.H., Hauptmann, A.: Revisiting the effect of topic set size on retrieval error. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2005)
10. Myers, J.L.: Fundamentals of Experimental Design. Allyn and Bacon, Boston, MA (1972)
11. Braschler, M.: CLEF 2001 - Overview of Results. In: Evaluation of Cross-Language Information Retrieval Systems : Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers. Volume 2406 of Lecture Notes in Computer Science. Springer-Verlag GmbH (2002) 9–26