Supporting Relevance Feedback in Video Search

Cathal Gurrin[1,2], Dag Johansen[1], Alan F. Smeaton[2]

[1] Dept. of Comp. Sci, Universitetet i Tromsø, 9037 Tromsø, Norway
dag@cs.uit.no

[2] Centre for Digital Video Processing, Dublin City University
Glasnevin, Dublin 9, Ireland
{cgurrin, alan.smeaton}@computing.dcu.ie

Abstract. WWW Video Search Engines have become increasingly commonplace within the last few years and at the same time video retrieval research has been receiving more attention with the annual TRECVid series of workshops. In this paper we evaluate methods of relevance feedback for video search engines operating over TV news data. We show for both video shots and TV news stories, that an optimal number of terms can be identified to compose a new query for feedback and that in most cases; the number of documents employed for feedback does not have a great effect on these optimal numbers of terms.

1 Introduction

Within the last few years we have seen the major search engines provide video searching and we are now able to search through large collections of video as if searching for web pages. At the same time, video retrieval research has continued apace, fostered to a great extent, by the TRECVid series of workshops. One aspect of interactive video retrieval systems, both research systems and WWW video search engines, has been the facility for a user to engage in relevance feedback. The research we present in this paper evaluates the effect of query size on relevance feedback performance for video archives of TV news shots and TV news stories where, like WWW video search engines, retrieval uses text surrogates of the video data. We examine scenarios where a user may choose to feedback one video document, or more than one video document in a relevance feedback process. Typically single document feedback is employed in WWW video search, whereas multi-document feedback has been employed primarily in research systems. For recommendation of video content based on user histories, the ability to automatically generate meaningful (and optimal) queries based on multiple user history documents is an important consideration and motivates this research.

Major search engines (such as Google and Yahoo!) have recently begun to provide video retrieval services. In addition there are a number of dedicated video search engines such as Truveo.com and Blinkx.com. WWW video search engines normally operate using a text surrogate of a video and process textual user queries. One possible option for generating text surrogates is the ASR (Automatic Speech Recognition)

text from the audio track of the video; however the most widely used technique uses the surrounding text from a web page, in a similar manner to the WWW image search engines. Search and relevance feedback is then supported using these text surrogates.

Research into video retrieval has been ongoing since the early 90s and two of the best known projects are Informedia [1] from CMU the Físchlár [2] Digital Video suite from DCU. Since 2001, the annual TRECVid Workshop [3] has fostered and encouraged such research by providing video test collections and a comparison and evaluation framework for participants. Many research video retrieval systems (e.g. [2]) support single video document feedback, but also 'more like these' relevance feedback where more than one document can be selected for feedback. Conventional relevance feedback techniques, when presented with a video document (a text surrogate), or many video documents, can then select terms to append to a query or compose a new query. We are interested in identifying the optimal number of terms (for both video shots and video news stories) used to compose a new query from text surrogates of video documents and how the number of feedback video documents influences this.

2 Relevance Feedback Experiment from Digital Video Libraries

The data used for this experiment was the TRECVid 2004 test collection, (33,367 video shots from TV news video and 24 topics). We represented each video shot by a document (textual surrogate) generated from the ASR transcript. In addition, we constructed a similar test collection of 1,757 news stories (with relevance judgments) from the TRECVid 2004 video shots, using predefined manually generated story boundaries (which excluded story transition shots). The basic text retrieval engine employed for this work implemented BM25 [4] with parameters trained on the TRECVid 2003 collection, which was similar in nature and size to TRECVid 2004. A custom stopword list was employed, based on the SMART list, but employing thirteen additional terms and the Porter stemmer was applied.

To evaluate relevance feedback, we automatically modeled a user selecting from one to nine video documents for feedback and examined system performance when between one and thirty terms were selected from these surrogates (270 evaluations). This feedback process generated a new query, not an expanded version of the original query. We assumed that the user would only feedback relevant video documents and that the user's information need, as expressed in the TRECVid topic did not change during the feedback process. Therefore, we only selected relevant (judged) video documents for feedback from the top ranked videos returned by the BM25 retrieval engine for each of the 24 topics and evaluated performance using the relevance judgments (which excluded the video documents already chosen for feedback). For feedback of 1 to 3 video documents we evaluated 5 different random combinations of documents from the top 5 relevant documents and averaged the results[1]. Feedback of

[1] E.g. two document feedback: 5 random pairs of unique documents from the top 5 relevant documents were chosen and evaluated for all 1-30 terms with the results averaged across the five pairs of documents producing 30 results (1-30 terms) for two document feedback.

4 to 9 video documents was performed similarly, though we evaluated 10 different random combinations from the top 10 relevant documents.

Two feedback techniques were examined for this study, TF-IDF and a variation on Robertson's Relevance Weight formula [4]. These algorithms are used to select the *N* most useful terms from the feedback documents to compose a new query. The TF-IDF algorithm employed *log* normalised TFs. The second algorithm incorporates a *log* nomalised TF weight into Robertson's RW formula and will be called TFRW. Having (*r=R=0*) where *N* is the size of the collection and *n* is the number of segments that term *i* occurs in, the formula is:

$$TFRW_i = \log(TF_i) \times \log((0.5/(N-n+0.5))/((n+0.5)\times 0.5). \quad (1)$$

Our findings suggest that there is no significant difference between the performances of these techniques in the experiments we present, and therefore we focus on TFRW in our results, which performed marginally better than TFIDF.

2.1 Shot-level and Story-level Feedback

The shot-level search engine achieved a MAP of 0.0465 over the 24 TRECVid 2004 topics (optimal parameter MAP is 0.0511). While low in absolute terms this is comparable to the expected performance of an automatic system on the TRECVid 2004 data.

Examining the results in detail (Fig 1), it is clear that performance increases significantly for relevance feedback of shots as terms are added up to a maximum of 7-8 terms, after which performance decreases or remains relatively static, regardless of the number of video documents chosen for feedback. The addition of any additional terms will not only affect query response time but also effectiveness.

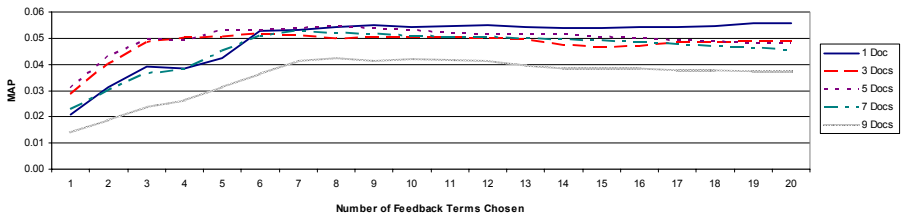

Fig. 1. Plot of 1,3,5,7 & 9 video shots chosen for feedback. The average MAP is shown as the number of feedback terms increased up to twenty (2,4,6 &8 removed for clarity)

The 2004 story-level search engine achieved a MAP of 0.2310 when using the 2003 parameters (optimal parameter MAP is 0.2318). With TV news story video, the optimal performance occurs between 10 and 13 terms in the feedback query (see Fig 2). The notable exception is when a single video document is chosen for feedback, when the optimal number of terms was found to be 30, though only a minor improvement in performance (6%) was noted over queries comprised of the top 13 terms. Adding additional terms above the top 30 has a negative effect on MAP. This

Adding additional terms above the top 30 has a negative effect on MAP. This negative effect increases with the number of video documents chosen for feedback.

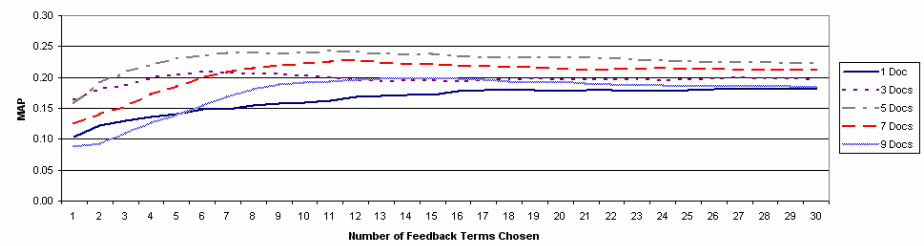

Fig. 2. Plot of 1,3,5,7 & 9 TV news story videos chosen for feedback. The average MAP is shown as the number of feedback terms increased up to twenty (2,4,6 & 8 removed for clarity)

3 Conclusions and Future Work

The purpose of this experiment was to evaluate the influence of query size on relevance feedback for video retrieval systems that index video shots or news stories. We have shown that for shots a system will perform at or near its peak when 7-8 terms are used to generate a new feedback query and for TV news stories that the peak can be found in most cases when 10-13 terms comprise the query. The number of video documents chosen for feedback does not affect these optimal numbers of terms noticeably (except for a single news story). The addition of more terms (beyond the optimal) from feedback video documents will be expected to hamper performance, while also having a negative effect on processing time. This is an important consideration for commercial WWW video search engines for whom processing time for each query is an important consideration. Future work planned includes optimising the feedback algorithms for general video data and we also plan to evaluate video search and relevance feedback on real-world WWW video content, with real users.

References

1. Hauptmann, A., Thornton, S., Houghton, R., Qi, Y., Ng, T.D., Papernick, N., Jin, R. Video Retrieval with the Informedia Digital Video Library System. In: Proceedings of the Tenth Text Retrieval Conference (TREC'01), Gaithersburg, Maryland, November 13-16, (2001).
2. Gurrin, C., Lee, H., Smeaton, A.F. Físchlár @ TRECVID2003: System Description. In 12th ACM International Conference on Multimedia 2004, New York, NY, 15-16 October (2004) 938-939.
3. TRECVid Workshop. WebLink: http://www-nlpir.nist.gov/projects/trecvid/. Last Visited 21st Nov. 2005.
4. Robertson, S. E., Sparck Jones, K. Simple, proven approaches to text retrieval. Tech. Rep. TR246, University of Cambridge, (1997).