EXPLOITING IMPLICIT USER FEEDBACK IN INTERACTIVE VIDEO RETRIEVAL

Stefanos Vrochidis[1,2], Ioannis Kompatsiaris[2] and Ioannis Patras[1]

[1]Queen Mary University of London, London, UK
[2]Informatics and Telematics Institute, Thessaloniki, Greece

ABSTRACT

This paper describes a video retrieval search engine that exploits both video analysis and user implicit feedback. Video analysis (i.e. automatic speech recognition, shot segmentation and keyframe processing) is performed by employing state of the art techniques, while for implicit feedback analysis we propose a novel methodology, which takes into account the patterns of user-interaction with the search engine. In order to do so, we introduce new video implicit interest indicators and we define search subsessions based on query categorization. The main idea is to employ implicit user feedback in terms of user navigation patterns in order to construct a weighted graph that expresses the semantic similarity between the video shots that are associated with the graph nodes. This graph is subsequently used to generate recommendations. The system and the approach are evaluated with real user experiments and significant improvements in terms of precision and recall are reported after the exploitation of implicit user feedback.

1. INTRODUCTION

The availability of large amount of audiovisual content places the demand for advanced multimedia search engines. However, video retrieval still remains one of the most challenging research topics. One of the proposed ways to improve performance of video search engines is to take advantage of the implicit and explicit feedback provided by users of video retrieval systems. Explicit user feedback is typically requested in Relevance Feedback (RF) approaches, but the main drawback is that users are usually reluctant to provide such feedback. For that reason, it is important to take into account the implicit user feedback. As such is considered any action or behavior of the user during a retrieval task including patterns of user-computer interaction (e.g. mouse clicks, etc), as well as user physiological and neurological reactions (e.g. eye movements, heart rate, etc) to the presentation of multimedia material. These could be used in order to reason about the levels of interest, emotional state, attitude or deduce the relevance of the presented material to a query.

Implicit feedback approaches based on user-computer interaction have been proposed in the context of textual retrieval. In [1] the definition of "Implicit Interest Indicators" was introduced by proposing specific user actions that can be considered as meaningful implicit feedback. In [2], the authors performed a comparison between an explicit and an implicit feedback system concluding that substituting the former with the latter could be feasible. A particularly interesting approach to exploit user feedback during video retrieval interactive sessions was to extend the idea of "query chains" [3] by constructing a graph that describes the user search and navigation actions and convert it to a weighted graph, in which video shots are interlinked with weights that express the semantic similarity of the corresponding nodes. More specifically, in [4] a video retrieval system enhanced by a recommendation generator based on such a graph structure is presented, while in [5] the authors evaluate 4 different recommendation algorithms for a similar system. These approaches consider only textual queries, while basic video retrieval options as visual and temporal based search are ignored. In [6], a video retrieval framework is presented, which employs RF and multimodal fusion of different sources (textual, visual and mouse click data) to generate recommendations. However, the implicit information is not sufficiently exploited as no sequence of query actions is taken into account, failing in that way to semantically connect subsequent queries and shots.

In this work we focus on exploiting past user-computer interaction by introducing new implicit interest indicators for video search and constructing a semantic affinity graph that expresses the semantic similarity between video shots. This graph is utilized to generate recommendations and is constructed in two steps. First an action graph that describes the user navigation pattern is generated by employing a novel methodology that defines search subsessions (i.e. parts of sessions in which the user searches a specific topic) based on query categorization. Then a set of action graphs is converted to a single weighted graph by aggregating the action graphs and assigning weights to the user actions based on the definition of the implicit interest indicators. We evaluate the approach by conducting real user experiments with two different video search engines: a baseline version that supports only video analysis retrieval options and the enhanced version that exploits also user implicit feedback. The contributions of this work are summarized in the introduction of new implicit interest indicators for video search, as well as in the proposed methodology of graph analysis based on query categorization and the definition of subsessions.

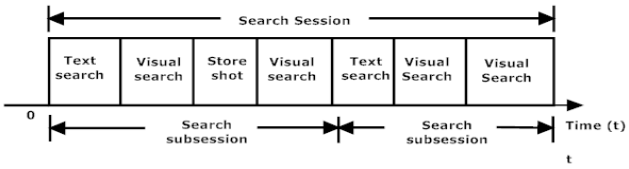

Fig. 1. Search session divided into search subsessions.

This paper is structured as follows: section 2 describes the implicit feedback analysis, while section 3 presents the implemented search engine, experiments and results. Finally, section 4 concludes the paper.

2. IMPLICIT FEEDBACK ANALYSIS

2.1. Implicit Interest Indicators for Video Search

In this section, we aim at defining implicit interest indicators [1] that measure aspects of the user-computer interaction, in order to exploit the information content that the latter carries about the user's perception of the presented multimedia material. Based on advanced retrieval functionalities in video search, which extend beyond the classical text-based queries already included in existing systems [7], we define the following minimum set of user actions that can be considered as the main implicit interest indicators:

1. Text-based query (TQ): the user inserts a keyword and submits the query. We assume that when a user submits a keyword as a search term, this keyword satisfies the query (or at least part of it) with a very high probability.
2. Visual query (VQ): the user selects a shot from previous results and submits a visual query by example. We assume that when a user selects a keyframe and searches for visually similar images, then there is also interest in the used example with a high probability.
3. Side-shot query (SQ): the user selects a shot in order to view the temporally adjacent shots and the associated textual description. This action can be interpreted as a declaration of interest in the selected shot.
4. Video-shot query (VSQ): the user selects a shot and retrieves all the shots of the same video. In this case we consider that the user is interested in the initial shot to a certain extend.
5. Submit a shot (SS): the user marks a shot as relevant. In this case we assume that the user is very interested in this shot.

2.2. Action Graphs

We exploit the implicit feedback information inspired by the graph construction methodology proposed in [4]. However, while [4] considers only text-based queries, we deal with a more complex situation, where visual-based and temporal-based queries are also included. First, we define as "search session" the time period that a certain user is active in using the search engine. Then, we propose to classify the query actions involved in a search session into two main categories: a) the autonomous queries, which comprise any query action not depending on previous results and b) the dependent queries, which take as input results from previous search actions. To construct an action graph, we propose a novel methodology, where we exploit the properties of autonomous and dependent queries to divide each search session in subsessions generating in that way several subgraphs. During a search session, the user may search for a specific topic, however it is possible to perform search having a very broad or complex topic in mind or decide to change the search topic during the session. For this reason, we propose that such sessions should not be analyzed as whole, but should be first decomposed into smaller subsessions. Assuming that every autonomous query could initiate a different topic search, we divide the search sessions into "search subsessions" using as break points the autonomous queries.

Taking into account the corresponding functionalities of the introduced implicit interest indicators, only the text-based search can be denoted as autonomous query, while the other queries are considered as dependent. In such a case the text-based query is utilized as a break point between the subsessions as illustrated in the example of Fig. 1. In the general case, a search subsession S consists of a set of actions A_s that includes one autonomous and a number of dependent query actions. The proposed subgraph G_s is comprised by a set of nodes (i.e. shots and keywords that represent inputs and outputs of a set of actions A_s) and links that represent the corresponding actions $a_i \epsilon A_s$, where $i\epsilon\{1,..,N_S\}$ and N_S is the cardinality of the elements of A_s.

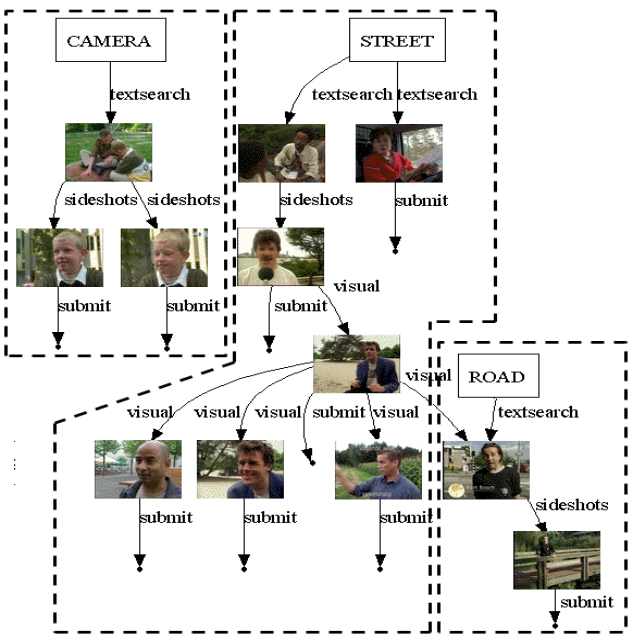

Fig. 2. Construction of action graph utilizing the main implicit interest indicators for video search. The different search subsessions are denoted by the dashed rectangles.

Fig. 3. Interactive video retrieval engine interface.

The action graph of a search session is composed of several subgraphs, which reflect the respective subsessions and have as parent nodes the autonomous queries. The above are illustrated in the example of Fig. 2, where an action graph for a search session, which includes the query actions defined as implicit interest indicators, is presented. Here, the user is searching for shots, in which people talking to the camera in an outdoor scene are depicted. We observe that the three keywords that were used to start the search (i.e. camera, street and road) are considered as the parents for new subgraphs, which correspond to different subsessions. In this way, concepts with different semantic meaning are not interconnected (e.g. 'camera' with 'street'), while keywords with similar semantic meaning (i.e. 'street' and 'road') are eventually interconnected due to the visual similarity between two shots in different subgraphs. Then we construct a single action graph aggregating the action graphs from the different user sessions.

2.3. Weighted Graphs

Once the single action graph is formed, we construct the weighted graph by a) linking the relevant results to the parent query, b) collapsing the multiple links between the same nodes into one and c) translating actions into weights. As suggested in [4], the final weight w for a link n between two nodes k and m is given by the formula:

$$w(n) = 1 - \frac{1}{x(n)} \qquad (1)$$

where $x(n)$ is the sum of the weights for each action a_i that is connecting nodes k and m. That is:

$$x(n) = \sum_{a_i \epsilon U_{km}} f(a_i) \qquad (2)$$

where f is the function that maps each action $a_i \epsilon U_{km}$ to an implicit weight, U_{km} comprises the set of actions between the nodes k, m, $i \epsilon \{1, .., M_{km}\}$ and M_{km} is the cardinality of the elements of U_{km}. Following the analysis in section 2.1, we assign indicative values (between 0 and 10) that quantify the levels of interest of the user to the multimedia material (shot/keyword) by associating a weight to the introduced implicit interest indicators (Table 1).

$Actions(a_i)$	$f(a_i)$	$Actions(a_i)$	$f(a_i)$
Text-based query (TQ)	8	Visual query (VQ)	8
Side-shot query (SQ)	7	Submit a shot (SS)	9
Video-shot query (VSQ)	6		

Table 1. Assigned weights for each action.

In [5] several recommendation algorithms based on such a weighted graph were proposed; however, in most of the cases, the best performing algorithm was depending on the search topics. Here, we employ a straightforward algorithm that initiates recommendations based on the distances on the weighted graph. The latter are calculated as the shortest path between two nodes.

3. EXPERIMENTS AND RESULTS

In order to evaluate the approach, we implemented a video search engine[1] (Fig 3.), which supports basic video retrieval options such as text, visual and temporal queries based on the system of [7]. The data used, is the test video set of TRECVID[2] 2008, which includes about 100 hours of video (news, documentaries, etc) segmented into about 30.000 shots. Then, we utilized the search engine to conduct an evaluation experiment, which was divided into 3 phases. In the first phase, 16 search sessions, each lasting 15 minutes, took place, in which 16 users searched for 4 different topics (i.e. 4 users for each topic) and their actions were recorded. Then, we constructed the weighted graph based on the proposed methodology. In the second phase, we recruited 4 different users, who searched for another 4 topics: 2 relevant (but not identical) to the ones of the first part and 2 irrelevant. For example, two relevant topics were: "Find people sitting at a table" and "Find shots of food" (i.e. topic 4 in Fig. 6,7), while two irrelevant were: "Find water scenes" and "Find people with horses" (i.e. topic 1 in Fig. 6,7). In this case, each user searched for all the 4 topics.

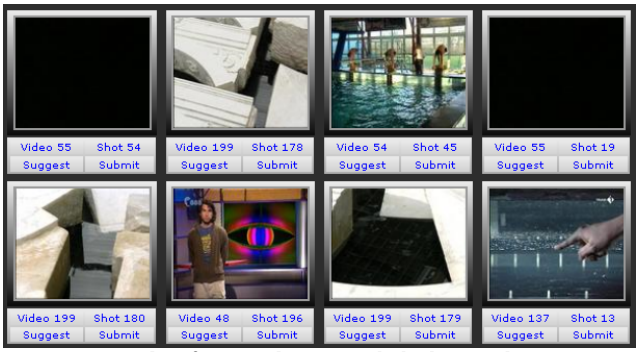

Fig. 4. Results of a textual query with the keyword "water".

[1] Demo available at: http://mklab-services.iti.gr/lelantus
[2] TRECVID: http://www-nlpir.nist.gov/projects/trecvid/

Fig. 5. Results of the recommendation module (keyword "water").

During these search sessions, the users were not allowed to use the recommendation module based on implicit feedback. Finally, another 4 users performed a search for the topics of the previous phase. These users were able to use not only the basic retrieval options of the system, but also the recommendation functionality. The duration for each session was 10 minutes for the last two experimental phases.

In order to show the improvement of the performance, when implicit feedback is taken into account, we present visual examples from interaction modes, as well as evaluation of the results utilizing precision and recall metrics. First, we present a usage scenario, in which the user is looking for scenes that a water body is visible by typing the keyword "water" (Fig. 4). As text retrieval is performed on the noisy information provided by Automatic Speech Recognition (ASR), only some of the results depict water scenes. Conducting the same query utilizing the graph with the past interaction data, we get a clearly better set of results (Fig. 5). Performance in terms of precision and recall for the 2nd and 3rd phases of the experiment is illustrated in Fig. 6 and Fig. 7, where these metrics are calculated against annotated results for each topic. The average improvement in recall for the first two topics 1 and 2 (i.e. the irrelevant to the initial ones) is about 5%, while precision seems to slightly drop by an average of 2%. As expected, the major improvement is reported in the topics 3 and 4 (i.e. the relevant to the initial queries), in which recall and precision are increased by an average of 72% and 9,8% respectively. The low absolute recall values are due to the fact that the many shots that were relevant for each query-topic, could not possibly be retrieved in the requested time duration of the experimental search sessions.

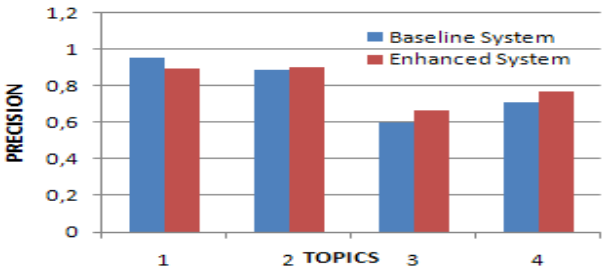

Fig. 6. Precision for the results of the last 2 experiments.

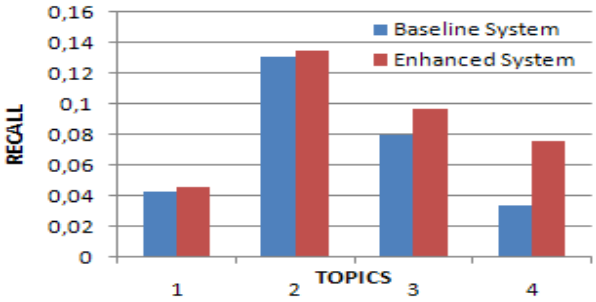

Fig. 7. Recall for the results of the last 2 experimental phases.

4. CONCLUSIONS

In this paper we have introduced new implicit interest indicators for video search and proposed a novel methodology to construct a content similarity graph based on the implicit indicators of patters of interaction of a user with a search engine. As it is shown by the results, the past user data can be of added value in modern video retrieval engines as rich user implicit feedback can become available.

5. ACKNOWLEDGMENTS

This work was supported by the projects CHORUS (FP6-045480) and PetaMedia (FP7-216444).

6. REFERENCES

[1] M. Claypool, P. Le, M. Waseda and D. Brown, "Implicit Interest Indicators" in *Proc. of ACM Intelligent User Interfaces Conference*. 2001, pp. 14–17, New Mexico, USA.

[2] MR. White, I. Ruthven, J. M. Jose, "The Use of Implicit Evidence for Relevance Feedback in Web Retrieval" in *Proc. of 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*. 2002, pp. 93–109, Glasgow, UK.

[3] F. Radlinski, T. Joachims,, "Query chains: learning to rank from implicit feedback" in *Proc. of the eleventh ACM SIGKDD*. 2005, pp. 239–248, Chicago, USA.

[4] F. Hopfgartner, D. Vallet, M. Halvey, J. M. Jose, "Search trails using user feedback to improve video search" *ACM Multimedia 2008*, pp. 339–348.

[5] D. Vallet, F. Hopfgartner, J. M. Jose, "Use of Implicit Graph for Recommending Relevant Videos: A Simulated Evaluation" in *Proc. of ECIR*. 2008, pp. 199–210, Chicago, USA.

[6] B. Yang, T. Mei, X-S. Hua, L. Yang, S-Q. Yang, M. Li, "Online video recommendation based on multimodal fusion and relevance feedback" in *Proc. of CIVR*. 2007, pp. 73–80, Amsterdam, The Netherlands.

[7] S. Vrochidis, P. King, L. Makris, A. Moumtzidou, V. Mezaris, I. Kompatsiaris, "MKLab Interactive Video Retrieval System" in *Proc. of CIVR*. 2008, pp, 563–563, Niagara Falls, Canada.