Community Based Feedback Techniques to Improve Video Search

David Vallet[2], Frank Hopfgartner[1], Martin Halvey[1] and Joemon Jose[1]

[1] Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, United Kingdom.
[2]Universidad Autónoma de Madrid, Escuela Politécnica Superior Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain.
david.vallet@uam.es, {hopfgarf, halvey, jj} @ dcs.gla.ac.uk

Abstract. In this paper we present a novel approach to aid users in the difficult task of video search. We use a graph based model based on implicit feedback mined from the interactions of previous users of our video search system to provide recommendations to aid users in their search tasks. This approach means that users are not burdened with providing explicit feedback, while still getting the benefits of recommendations. The goal of this approach is to improve the quality of the results that users find, and in doing so also help users to explore a large and difficult information space. In particular we wish to make the challenging task of video search much easier for users. The results of our evaluation indicate that we achieved our goals, the performance of the users in retrieving relevant videos improved, and users were able to explore the collection to a greater extent.

Keywords: Video, search, collaborative, implicit, feedback, user study

1 Introduction

With the growing capabilities and the falling prices of current hardware systems, there are ever increasing possibilities to store and manipulate videos in a digital format. Also with ever increasing broadband capabilities it is now possible to view video online as easily as text-based pages were viewed when the Web first appeared. People are now creating their own digital libraries from materials created through digital cameras and camcorders, and use a number of systems to place this material on the Web, as well as store them as their own individual collections [20]. However, the systems that currently are used to manage and retrieve these videos are insufficient for dealing with such large and swiftly increasing volumes of video. Current state of the art systems rely on either using annotations provided by users or on methods that use the visual low level features available in the videos. As experimental results in state of the art video retrieval show, neither of these approaches is sufficient to overcome the difficulties associated with video search (see Section 2 for more details).
We believe that many of these problems associated with browsing and searching large collections of video can be alleviated through the use of recommendation techniques.

Recommendation techniques based on implicit actions do not require users to alter their normal behaviour, while all of the actions that users carry out can be used to improve their retrieval results. To test these assertions, we have developed a graph based model that utilises the implicit actions involved in previous user searches. This model can provide recommendations to support users in completing their search tasks. This approach is flexible as it allows us to use a combination of explicit and implicit feedback, or implicit feedback alone, to help users. We believe that this approach can result in a number of desirable outcomes. We achieved an improved user performance in terms of task completion, proving that we can aid user exploration of the collection and can also increase the user satisfaction with their search and their search results. An evaluative study was performed, in order to examine and validate these assumptions. Two systems were compared. The first system is a baseline system that provides no recommendations. The second system is a system that provides recommendations based on our model of implicit user actions. The two systems and their respective performances were evaluated both qualitatively and quantitatively. It was found that our approach increases the accuracy of videos retrieved by users, allows users to navigate a video collection to a greater extent and that users were more at ease using our recommendation system in comparison with a baseline. The remainder of this paper is organised as follows: In Section 2, we will provide a rationale for our work, and describe the state of the art in a number of areas that have inspired this work. Subsequently, in Section 3 we will describe our approach for using implicit feedback to provide recommendations. Section 4 will describe the two systems used in our study. In Section 5 we will describe our experimental methodology, which is followed by the results of our experiments in Section 6. Finally, in Section 7, we provide a discussion of our work and some conclusions.

2 Background and Motivation

The work that we have undertaken is multi-faceted and draws inspiration from various research areas. We will discuss a number of these research areas and how they are related to our work. We begin by discussing video retrieval, and in particular interactive video retrieval.

2.1 Interactive Video Retrieval

Interactive video retrieval refers to the process of users formulating and carrying out video searches, and subsequently reformulating queries based on the previously retrieved results to retrieve more results. Most state of the art video retrieval engines, interactive or otherwise can be divided into two major components [32], an indexing engine and a retrieval engine, which accesses the index through queries. The indexing engine is responsible for providing a quickly accessible index of information that is available in a corpus of video. The indexing begins with shot segmentation, in this step a larger video is segmented into a sequence of shorter video shots. A shot is a sequence of the video which is visually related, the length of a shot can vary depending on the video content. Boundaries between shots are typically marked by a cut or fade in the video sequence. For most video retrieval systems, a shot is the element of retrieval. Each shot is indexed separately by the video retrieval system. The shots may have text associated with them and can be represented by individual frames. In addition to this, for each shot, a number of example frames which provide the best representation of the shot are calculated, these are called keyframes. When a search is carried out, the results of the search are normally presented as a list of shots, which are represented by keyframes.

In addition to this basic representation of videos, there are additional pieces of metadata or additional information that can be extracted from video to aid search and be added to the index, e.g. speech, closed captions, etc. A study of a number of state of the art video retrieval systems [15] concludes that the availability of these additional resources varies for different systems. For example Heesch et al. [13] include Closed Caption transcripts (CC), Automatic Speech Recognition (ASR) and Object Character Recognition (OCR) output in their index, whereas Foley et al. [7] include index speech recognition output only. As these additional resources are not always available, they cannot always be relied upon to aid user searches. Despite this uncertainty and lack of resources there has been a great deal of research carried out into their use. For example, Huang [16] has argued that speech contains a great deal of semantic information that can be used to help search. Further research from Chang et al. [3] found that text extracted from speech data can be an important feature for extracting additional information from a video; it can be used for named identity extraction, annotating concepts in video and topic change extraction in a video. The indexing procedure, together with the retrieval engine, constitutes the “backend” of a video retrieval system.

The main component that allows the interactivity in interactive video retrieval is the “front end”; this is the interface between the computer and the user. The interface allows a user to compose queries and interact with the results. The users can then re-formulate queries if they wish and get a new set of results. There are a number of different ways in which a user can query a video retrieval system; these include query by text, query by example and query by concept. None of these approaches have as of yet provided an adequate solution to providing the tools to facilitate video search. Query by text is one of the most popular methods of searching for video. It is simple and users are familiar with this paradigm from text based searches. However, query by text relies on the availability of sufficient textual descriptions. Textual descriptions may be extracted from closed captions or through automatic speech recognition; however, the availability of these additional resources varies for different systems [15]. More recent online state of the art systems rely on using annotations from users to provide textual descriptions for videos, as well as other media, e.g. images, Web pages, podcasts etc. However, annotations also provide a number of problems. Users can have different perceptions about the same video and tag that video differently. This can result in synonyms, polysemy and homonymy, which makes it difficult for other users to retrieve the same video. It has also been found that users are reluctant to provide a large number of tags [11]. Query by example allows the users to provide sample images or video clips as examples to retrieve more results. This approach uses the low-level features that are available in images and videos, such as colour, texture and shape to retrieve results. The approach is also inadequate for video retrieval because of the difference between low-level data representation of videos and the higher level concepts users associate with video, commonly known as the semantic gap [17]. Bridging the semantic gap is one of the most challenging research issues in multimedia information retrieval today. In an attempt to bridge this gap, a great deal of interest in the multimedia search community has been invested in search by concept. Semantic concepts such as "vehicle" or "person" can be used to aid retrieval; an example of this semantic application has been explored by the Large Scale Ontology for Multimedia (LSCOM) initiative [23]. Search by concept requires a large number of concepts to be represented and a number of training examples to represent those concepts. While each of these methods alone is inadequate, they have been used in conjunction with each other in a number of systems, including MediaMill [31] and Informedia [4]. MediaMill and Informedia have been amongst the most successful systems at recent TRECVID interactive search evaluations [24]. However, these results are for "expert" users, who are supposed an idealistic performance compared to the common user [5].

2.2 Personal Multimedia Search and Sharing

In recent years there has been a rapid increase in the number of photographs and videos that individuals have stored in personal collections and shared online. This has led to the emergence of some interesting research that has investigated user interaction with multimedia. Kirk et al. present their research on "photowork" [19] and "videowork" [20]; photowork and videowork are the activities that users engage in with their digital photos or videos prior to sharing. These user activities form some of the context for the browsing and searching that users carry out. Kirk et al. find that search as we know it may have much less relevance than new paradigms and ways of browsing, for the design of new digital photo tools. Ethnographic-style studies [2] have also found that there are huge similarities between the ways in which participants used personally captured photos and commercially purchased music. Halvey and Keane [11] provide analyses of people's linking and search behaviour using YouTube. Initial results show that page views in the video context deviate from the typical power-law relationships seen on the Web. There are also clear indications that tagging and textual descriptions play a key role in making some video-pages more popular than others. However, a number of studies have shown that users cannot be relied upon to provide large amounts of textual annotations [10, 11], thus we need another form of information to bridge the semantic gap. We believe that using collaborative or community based methods to aid users' video search is one of the best solutions to the problems associated with video search.

2.3 Collaborative Information Access

Many of the earliest collaborative techniques emerged online in the 1990's [9, 25, 28] and focused on the notion of collaborative filtering. Collaborative filtering was first developed in the Tapestry system to recommend e-mails to users of online newsgroups [9]. Collaborative filtering aims to group users with similar interests, with a view to treating them similarly in the future. So, if two users have consistently liked or disliked the same resources, then chances are that they will like or dislike future resources of that type. Since those early days collaborative or community based methods have evolved and been used to aid browsing [36], e-learning [8] and in collaborative search engines [30]. More recently there has also been some recent initial research into carrying out collaborative video search [1]. This work, however, concentrated on two users carrying out a search simultaneously rather than using the implicit interactions from previous searches to improve future searches, which is not the focus of our research.

Traditionally, explicit relevance feedback has been used for a number of collaborative systems; however, there are a number of problems with this approach. Providing explicit feedback can be a cognitively taxing process. Users are forced to update their need constantly and this can be a difficult process when their information need is vague [34] or when they are unfamiliar with the document collection [27]. Also, previous evaluations have found that users of explicit feedback systems often do not provide sufficient levels of feedback in order for adaptive retrieval algorithms to work [12]. With this in mind we are using the implicit action of users as the basis for our system. Implicit feedback has been shown to be a good indicator of interest in a number of areas [18]. Hopfgartner et al. [14] have suggested that implicit relevance feedback can aid users searching in digital video library systems. Using click through data as implicit feedback, White et al. [37] use the concept of "search trails", meaning the search queries and document interaction sequences performed by the users during a search session, to enhance Web search. Craswell and Szummer [6] apply a random walk on a graph of user click through data, to help retrieve relevant documents for user searches. Specifically for multimedia search, relevance feedback based on the content of video has also been used in conjunction with related information, e.g. tags, to provide video search recommendations to users [21, 38]. However, we believe that such techniques are insufficient where there is a lack of associated information and will also suffer from problems associated with the semantic gap [17]. In response to a number of the problems that have been outlined in the previous sections, we have developed our own graph based model of implicit actions, which we use to provide recommendations. This approach uses the previous work of White et al. [37] and Craswell and Szummer [6] as a guide. This model and the recommendation techniques that we use are described in the following section.

3 Implicit Feedback: A Graph Based Approach

For our recommendation model based on user actions, there are two main desired properties of the model for action information storage. The first property is the representation of all of the user interactions with the system, including the search trails for each interaction. This allows us to fully exploit all of the interactions to provide richer recommendations. The second property is the aggregation of implicit information from multiple sessions and users into a single representation, thus facilitating the analysis and exploitation of past implicit information. To achieve these properties we opt for a graph based representation of the users' implicit information. We take the concept of trails from White et al. [37]; however unlike White et al. we do not limit the possible recommended documents to those documents that are at the end of the search trail. We believe that during an interactive search the documents that most of the users with similar interaction sequences interacted with, are the documents that could be most relevant for recommendation, not just the final document in the search trail. Similar to Craswell and Szummer [6], our approach represents queries and documents in the same graph, however we represent the whole interaction sequence, unlike their approach where the clicked documents are linked directly to the query node. Once again we want to recommend potentially important documents that are part of the interaction sequence and not just the final document of this interaction. Another difference between our approach and previous work is that we take into consideration other types of implicit feedback actions, related to multimedia search, e.g. length of play time, browsing keyframes etc., as well as click through data. This additional data allows us to provide a richer representation of user actions and potentially better recommendations. Overall our representation exploits a greater range of user interactions in comparison with other approaches [6, 21, 37]. This results in a more complete representation of a wide range of user actions that may facilitate better recommendations. These properties and this approach result in two graph-based representations of user actions. The first representation utilises a Labelled Directed Multigraph (LDM) for the detailed and full representation of implicit information. The second graph is a Weighted Directed Graph (WDG), which interprets the information in the LDM and represents it in such a way that is exploitable for a recommendation algorithm. The recommendations that are provided are based on three different techniques based on the WDG. The two graph representation techniques and the recommendation techniques are described in detail in the following sections.

3.1 Labelled Directed Multigraph

A user session s can be represented as a set of queries Q_s, which were input by the user u, and a set of multimedia documents D_s the users interacted with during the search session. Queries and documents are represented as nodes $N_s = \{Q_s \cup D_s\}$ of our graph representation, $G_s = (N_s, A_s)$. The interactions of the user during the search session are represented as a set of action arcs $A_s(G) = \{n_i, n_j, a, u, t\}$. Each action arc indicates that, at a time t, the user u performed an action of type a that lead the user from the query or document node n_i to node n_j, $n_i, n_j \in N_s$. Note that n_j is the object of the action and that actions can be reflexive. For instance, when a user clicked to view a video and then navigate through it. Action types depend on the kind of actions recorded by the implicit feedback system. In our system we recorded playing a video, navigating through a video, highlighting a video to get additional metadata and selecting a video. Links can contain extra associated metadata, as type specific attributes, e.g. length of play in a play type action. The graph is multilinked, as different actions can have the same source and destination nodes. The session graph $G_s = (N_s, A_s)$ will then be constructed by all the accessed nodes and linking actions, and will represent the whole interaction process for the user's session s. Finally, all session-based graphs can be aggregated into a single graph $G = G(N, A)$, $N = \bigcup_s N_s$, $A = \bigcup_s A_s$ which represents the overall pool of implicit information.

Quite simply, all of the nodes from the individual graphs are mapped to one large graph, and then all of the action edges are mapped onto the same graph. This graph may not be fully connected, as it is possible, for instance, that users selected different paths through the data or entered a query and took no further actions etc. While the LDM gives a detailed representation of user interaction with the collection, it is extremely difficult to use to provide recommendations. The multiple links make the graph extremely complex. In addition to this all of the actions are weighted equally. This is not always a true representation; some actions may be more important than others and should be weighted differently.

3.2 Weighted Directed Graph

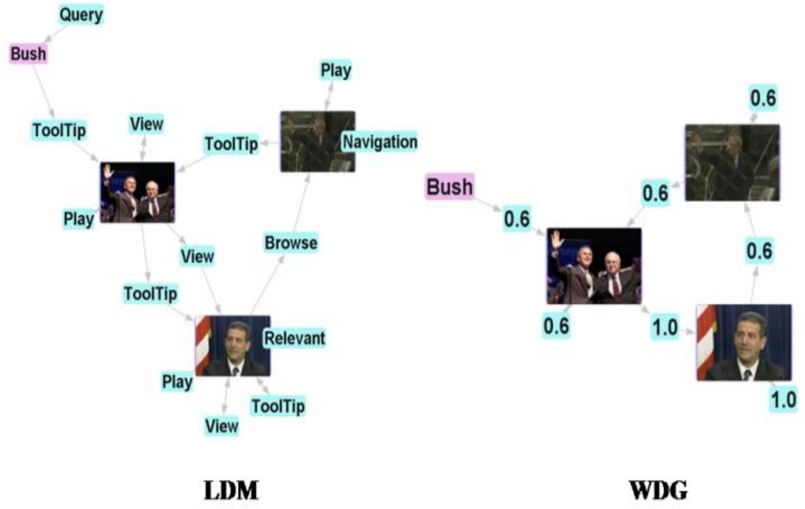
In order to allow our recommendation algorithm to exploit the LDM representation of user actions, we convert the LDM to a WDG by collapsing all links interconnecting two nodes into one single weighted edge. This process is carried out as follows. Given the detailed LDM graph of a session s, $G_s = (N_s, A_s)$, we compute its correspondent weighted graph $G_s = (N_s, W_s)$. Links $W_s = \{n_i, n_{j,} \mathrm{w_s}\}$ indicate that at least one action lead the user from the query or document node n_i to n_j. The weight value $\mathrm{w_s}$ represents the probability that node $n_{j,}$ was relevant to the user for the given session, this value is either given explicitly by the user, or calculated by means of the implicit evidence obtained from the interactions of the user with that node:

$$w_s(n_i, n_j) = \begin{cases} 1, & \text{iff explicit relevance for } n_j \\ -1, & \text{iff explicit irrelevance for } n_j \\ lr(n_j) \in [0,1], & \text{otherwise (i. e. implicit relevance)} \end{cases}$$

In the case that there is only implicit evidence for a node n, the probability value is given by the *local relevance* $lr(n)$. $lr(n)$ returns a value between 0 and 1 that approximates a probability that node n was relevant to the user given the different interactions that the user had with the node. For instance if the user opened a video and played it for the whole of its duration, this can be enough evidence that the video has a high chance of being relevant to the user. Following this idea, and based on previous work on the impact of implicit feedback importance weights [14], the local relevance function is defined as $lr(n) = 1 - \frac{1}{x(n)}$, where $x(n)$ is the total of added weights associated to each type of action in which node n is an object of. This subset of actions is defined as $A_s(G_s, n) = \{n_i, n_j, a, u, t | n_j = n\}, n \in N_s$. These weights are natural positive values returned by a function $f(a): A \to \mathbb{N}$, which maps each type of action to a number. These weights are higher for an action that is understood to give more evidence of relevance to the user. In this way, $lr(n)$ is closer to 1 as more actions are observed that involve n and the higher the associated weight given to each action type. In our weighting model some of the implicit actions are weighted nearly as highly as explicit feedback. The accumulation of implicit relevance weights can thus be calculated as $x(n) = \sum_{a \in A_s(G_s,n)} f(a)$. Table 1 shows an example of function f, used during our evaluation process; all of these actions are part of the system described in Section 4. This system considers the following actions: 1) playing a video during a given interval of time (Play); 2) clicking a search result in order to view its contents (View); 3) navigating through the contents of a video (Navigate) ; 4) browsing to the next or previous video keyframe (Browse R/L) and 5) tooltiping a search result by leaving the mouse pointer over the search result. The weights assigned to these actions are based on previous work by Hopfgartner et al. [14], where they carried out a study which simulated users carrying out interactive video searches. As part of this work a number of different weighting schemes were evaluated in order to determine which weighting scheme was the most effective for implicit feedback for interactive video search. The weights that we use are based on the results of this previous work. Figure 1 shows an example of LDM and its correspondent WDG for a given session.

Action a	f(a)	Action a	f(a)
Play	3	Navigate Browse R/L	2
View	10	Tooltip	1

Table 1: Values for function f() used during the experiment.



LDM

WDG

Figure 1: Correspondence between the LDM (left) and WDG (right) models

Similarly to the detailed LDM graph, the session-based WDGs can be aggregated into a single overall graph $G = (N, W)$, which will be called the implicit relevance pool, as it collects all the implicit relevance evidence of all users across all sessions. The nodes of the implicit pool are all the nodes involved in any past interaction $N = \bigcup_s N_s$, whereas the weighted links combine all of the session-based values. In our approach we opted for a simple aggregation of these probabilities, $W = \{n_i, n_j, w\}$, $w = \sum_s w_s$. Each link represents the overall implicit (or explicit, if available) relevance that all users, whose actions lead from node n_i to n_j, gave to node n_j. Figure 2 shows an example of the implicit relevance pool.

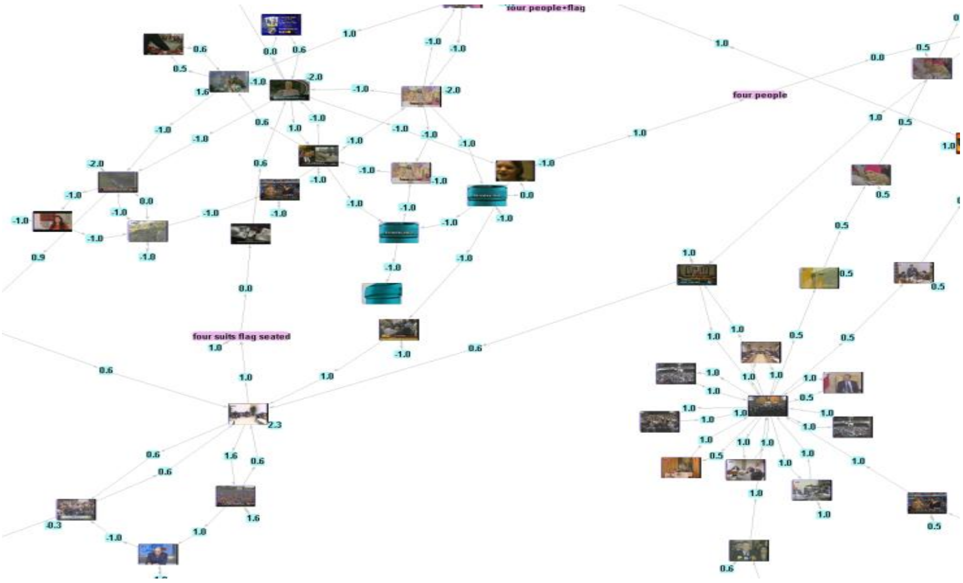


Figure 2: Graph illustrating implicit relevance pool

3.3 Implicit Relevance Pool Recommendation Techniques

In our system we recommend both queries and documents to the users. These recommendations are based on the status of the current user session. As the user interacts with the system, a session-based WDG is constructed. The current user's session is thus represented by $G_s{}' = (N_{s'}, W_{s'})$. This graph is the basis of the recommendation algorithm which has three components; each component uses the implicit relevance pool in order to retrieve similar nodes that were somehow relevant to other users. The first two components are neighbourhood based. A neighbourhood approach is a way of obtaining related nodes; quite simply we define the node neighbourhood of a given node n, as the nodes that are within a distance $D_{MAX}\, d$ of n, without taking the link directionality into consideration. These nodes are somehow related to n by the actions of the users, either because the users interacted with n after interacting with the neighbour nodes, or because they are the nodes the user interacted with after interacting with n. More formally as a way of obtaining related nodes, we define the node neighbourhood of a given node n as:

$$NH(n) = \{m \in N | \delta(n, m) < D_{MAX}\}$$

where $\delta(n, m)$ is the shortest path distance between nodes n and m, and D_{MAX} is the maximum distance in order to take into consideration a node as a neighbour. The best performing setting for this value, in our experiments, was $D_{MAX} = 3$.

Using the properties derived from the implicit relevance pool, we can calculate the overall relevance value for a given node. This value indicates the aggregation of implicit relevance that users gave historically to n, when n was involved with the users' interactions. Given all the incident weighted links of n, defined by the subset $W_s(G_s, n) = \{n_i, n_j, w | n_j = n\}, n \in N_s$, the overall relevance value for n is calculated as follows:

$$or(n) = \sum_{w \in W_s(G_s, n)} w$$

Given the ongoing user session s, and the implicit relevance pool we can then define the node recommendation value as:

$$nh(n, N_s) = \sum_{n_i \in N_s, n \in NH(n_i)} lr'(n_i) \cdot or(n)$$

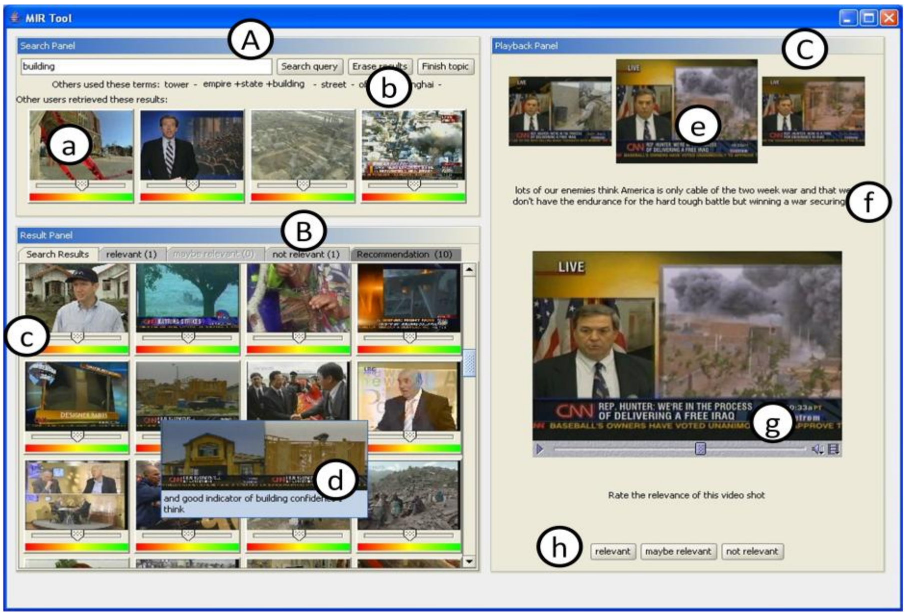
where $lr'(n_i)$ is the local relevance computed for the current session of the user $G_{s'}$, so that the relevance of the node to the current session is taken into consideration. We can then define the first recommendation value $r_1(n, N_{s'}) = nr(n, Q_{s'}) | Q_{s'} \in N_{s'}$, i.e. the node recommendation value for the queries related to the current session. Similarly, we can define the second recommendation value $r_2(n, N_{s'}) = nr(n, D_{s'}) | D_{s'} \in N_{s'}$, which exploits the session-related documents instead. The last recommendation component is based on the user's interaction sequence. The interaction sequence recommendation approach tries to take into consideration the interaction process of the user, with the scope of recommending those nodes that are following this sequence of interactions. For instance, if a user has opened a video of news highlights, the recommendation could contain the more in-depth stories that previous users found interesting to view next. The recommendation value $r_3(n, N_{s'})$, called interactive recommendation, can thus be defined as follows:

$$r_3(n, N_s) = \sum_{\substack{n_i \in N_s \\ p = n_i \rightsquigarrow n_j \to n \\ lengt\,h(p) < L_{MAX}}} lr'(n_i) \cdot \xi^{lengt\,h(p)-1} \cdot w\left(n_j, n\right)$$

where $n_i \rightsquigarrow n_j$ denotes the existence of a path from n_i to n_j in the graph, $n_j \to n$ means that n is adjacent to n_j, and the same notation is used as a shorthand to define p as any path between n_i and n, taking into consideration the link directionality. $length(p)$ is counted as the number of links in path p, which must be less than a maximum length L_{MAX}. Finally, ξ is a length reduction factor, set to 0.8 in our experiments, value which allowed weight to be propagated significantly up to six degrees of separation. This length reduction factor allows giving more importance to those documents that directly follow the interaction sequence, though if a document with high levels of interaction occurs two or three steps away it may still be recommended. In a final step, we obtain the three recommendation lists from each recommendation component and merge them into a single final recommendation lists. For this we use a rank-based aggregation approach, the scores of the final recommendations are the sum of the rank-based normalised scores of each of the recommendation list, i.e. using a score $\frac{1}{r(n)}$ where $r(n)$ is the position of n in the recommended list. The final list is then split into recommended queries and recommended documents; these are then presented to the user.

4 System Description

Our implicit feedback approach has been implemented in an interactive video retrieval system. This allows us to have actual end users test our system and approach. The keyframes in our index were indexed based on automatic speech recognition transcript and machine translation output. The Okapi BM25 retrieval model [26] was used to rank retrieval results. In addition to the ranked list of search results, the system provides users with additional recommendations of video shots that might match their search criteria based on our recommendation graph (see Section 3 for details on the recommendation graph).



Search field: building

Others used these terms: tower - empire +state +building - street - o[b]nghai -

Other users retrieved these results:

and good indicator of building confide[d]
think

lots of our enemies think America is only cable of the two week war and that we
don't have the endurance for the hard tough battle but winning a war securing

Rate the relevance of this video shot

Figure 3: Interface of the video retrieval system.

Figure 3 shows a screen shot of the recommendation system. The interface can be divided into three main panels: the search panel (A), the result panel (B) and the playback panel (C). The search panel (A) is where users formulate and carry out their searches. Users can enter a text based query in the search panel (A) to begin their search. The users are presented with text based recommendations for search queries that they can use to enhance their search (b). The users are also presented with recommendations of video shots that might match their search criteria (a). Each recommendation is only presented once, but may be retrieved by the user at a later stage if they wish to do so. The result panel is where users can view the search results (B). This panel is divided into five tabs, the results for the current search, a list of results that the user has marked as relevant, a list of results that the user has marked as maybe being relevant, a list of results that the user has marked as irrelevant and a list of recommendations that the user has been presented with previously. Users can mark results in these tabs as being relevant by using a sliding bar (c). Additional information about each video shot can be retrieved by hovering the mouse tip over a video keyframe, that keyframe will be highlighted, along with neighbouring keyframes and any text associated with the highlighted keyframe (d). The playback panel (C) is for viewing video shots (g). As a video is playing it is possible to view the current keyframe for that shot (e), any text associated with that keyframe (f) and the neighbouring keyframes. Users can play, pause, stop and can navigate through the video as they can on a normal media player, and also make relevance judgements about the keyframe (h). Some of these tools in the interface allow users of the system to provide explicit and implicit feedback, which is then used to provide recommendations to future users. Explicit feedback is given by users by marking video shots as being either relevant or irrelevant (c, h). Implicit feedback is given by users playing a video (g), highlighting a video keyframe (d), navigating through video keyframes (e) and selecting a video keyframe (e).

In order to provide a comparison to our recommendation system, we also implemented a baseline system that provides no recommendations to users. The baseline system has previously been used for the interactive search task track at TRECVID 2006 [35], the performance of this system was average when compared with other systems at TRECVID that year. A tooltip feature which shows neighbouring keyframes and the transcript of a shot was added to this system to improve its performance. Overall the only difference between the baseline and the recommendation system is the provision of keyframe recommendations (a).

5 Experimental Methodology

In order to determine the usefulness of our approach we carried out a user-centred evaluation of our system and approach. In this section we will give details on our hypothesis, experimental methodology, video collections and tasks that were used.

5.1 Hypothesis

The goal of our evaluation was to investigate the effect of using community based feedback to aid search in a video search system. There are a number of potential benefits of our approach:

- The performance of the users of the system, in terms of precision of retrieved videos, will improve with the use of recommendations based on feedback.
- The users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered.

- The users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search.

5.2 Collection and Tasks

In order to determine the effects of implicit feedback, users were required to carry out a number of video search tasks based on the TRECVID 2006 collection and tasks [29]. The TRECVID evaluation meetings are an on-going series of workshops focusing on a list of different information retrieval research areas in content based retrieval of video. For our evaluation we focus on the interactive search tasks. Interactive search tasks involve the use of low level content based search techniques and feedback from users of the video search system.

Topic Number 0179

Find shots of Saddam Hussein with at least one other person's face at least partially visible

Image examples:

Figure 4: TRECVID 2006 task example

For the interactive search task users are given a specific query and a maximum of fifteen minutes to find shots relevant to that query. Figure 4 shows an example query for TRECVID 2006. In interactive tasks the users can use a combination of text and shots to form an initial search for relevant videos to the query. Users can interact with the systems and examine the results, and can also re-formulate their queries and retrieve new results to continue their search. Users then mark as many shots as possible as being relevant to the topic. The shots that the participants or system marked as relevant are then compared with a set of relevant shots which has been created based on pooling [33] to determine the accuracy and recall of the results. In 2006 there were 79,848 shots in the TRECVID test collection, and a total of 24 tasks. For our evaluation we are limiting the number of tasks that the users carry out to 4. Although for the TRECVID evaluations 24 tasks are carried out, it was not practical for our evaluation to do this. The user evaluations are very expensive in terms of time and costs. In our case, each user required 15 minutes per topic plus another 5-10 minutes to complete questionnaires per user, thus it would not have been possible to carry out a systematic study with a large number of topics. In addition, the goals of this evaluation were not the same as TRECVID. We felt it was not necessary to use all of the 24 tasks. For this evaluation we chose the four tasks for which the average precision in the 2006 TRECVID workshop was the worst. In essence these are the most difficult tasks. The 4 tasks chosen were chosen as in general these are tasks for which there are less relevant documents. Indeed the mean average precision (MAP) values show that it is extremely difficult to find these documents. We feel that any improvement which may be gained on these difficult tasks with few documents will be reflected on less difficult tasks with larger numbers of relevant documents. The same cannot be said about the gains made for easier tasks being borne out in more difficult tasks. Moreover, due to the difficult nature of these topics, different users had to use a different search query, which ensures that users do not just follow other users' search trails (this is shown in subsequent sections). As can be seen below there were very few relevant shots in the collection for these task, 98 shots out of 79,848 shots for one of the tasks. In addition to this, not all of the relevant shots have text associated with them. As the most popular form of search is search by textual query [5], finding these shots becomes even more difficult. The four tasks that were used for this evaluation were:

1. Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible (142 relevant shots, 53 with associated text)

2. Find shots with one or more soldiers, police, or guards escorting a prisoner (204 relevant shots, 106 with associated text)

3. Find shots of a group including at least four people dressed in suits, seated, and with at least one flag (446 relevant shots, 287 with associated text)

4. Find shots of a greeting by at least one kiss on the cheek (98 relevant shots, 74 with associated text)

The users were given the topic and a maximum of fifteen minutes to find shots relevant to the topic. The users could only carry out text based queries, as this is the normal method of search in most online and desktop video retrieval systems and also the most popular search method at TRECVID [5]. The shots that were marked as relevant were then compared with the ground truth in the TRECVID collection.

5.3 Experimental Design

For our evaluation we adopted 2-searcher-by-2-topic Latin Square designs. Each participant carried out two tasks using the baseline system, and two tasks using the recommendation system. The order of system usage was varied as was the order of the tasks; this was to avoid any order effect associated with the tasks or with the systems. To determine the effect of adding more implicit actions to the implicit pool, participants in the experiment were placed in groups of four. For each group, the recommendation system used the implicit feedback from all of the previous users. At the beginning of the evaluation there was no pool of implicit actions, therefore the first group of four users received no recommendations; their interactions formed the training set for the initial evaluations. Using this experimental model we can evaluate the effect of the implicit feedback within a group of participants, and also the effect of additional implicit feedback across the entire group of participants. In addition to this, the ground truth provided in the TRECVID 2006 collection allowed us to carry out analyses that we may not have been able to do with other collections. Each participant was given five minutes training on each system and carried out a training task with each system. These training tasks were the tasks for which participants had performed the best at TRECVID 2006. For each participant their interaction with the system was logged, the videos they marked as relevant were stored and they also filled out a number of questionnaires at different stages of the experiment.

5.4 Participants

24 participants took part in our evaluation and interacted with our two systems. The participants were mostly postgraduate students and research assistants. The participants consisted of 18 males and 6 females with an average age of 25.2 years (median: 24.5) and an advanced proficiency with English. Students were paid a sum of £10 for their participation in the experiment. Prior to the experiment the participants were asked to fill out a questionnaire so that we could ascertain their proficiency with and experience of dealing with multimedia. We also asked participants about their knowledge of news stories, as the video collection which the participants would be dealing with consists of mainly news videos. It transpired that the participants follow news stories/events once or twice a week and also watch news stories online. The majority of participants deal with multimedia regularly (once or twice a day) and are quite familiar with creating multimedia data (images, videos). The participants also had a great deal of experience of searching for various types of multimedia. These activities were mainly carried out online, with Flickr, Google or YouTube being cited as the most commonly used online services. The most common search strategy that users mentioned was searching for data by using initial keywords and then adapting the query terms to narrow down the search results based on the initial results they get. Using the recommendations provided by some of these services was also mentioned by a number of users. Although the participants often searched for multimedia data, they stated that they rarely use multimedia management tools to organise their personal multimedia collection. The most common practice amongst the participants is creating directories and files on their own personal computer. Categorising videos and images according to the content and time when this data was produced, is the most popular method of managing media. However, when asked how a system could support the own search strategy, many participants mentioned that it would be helpful to sort or retrieve multimedia based on their semantic content. The following section outlines the results of our evaluation.

6 Results

6.1 Task Performance

As we were using the TRECVID collection and tasks, we were able to calculate precision and recall values for all of the tasks. Figure 5 shows the P@N for the baseline and recommendation systems for varying values of N. P@N is the ratio between the number of relevant documents in the first N retrieved documents and N.

The P@N value focuses on the quality of the top results, with a lower consideration on the quality of the recall of the system.

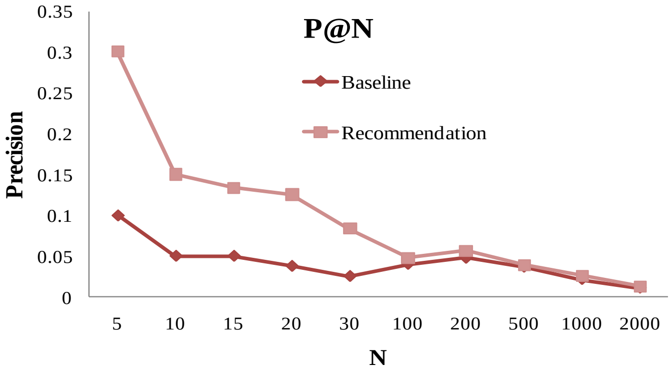


Figure 5: P@N for the baseline and recommendation systems for varying values of N

Figure 6 shows the mean average precision (MAP) for baseline and recommendation systems for different groups of users. Each group of four users also had additional feedback from previous participants, which the previous group of four users did not have. MAP is the average for the 11 fixed precision values of the PR (Precision and Recall) metric, and is normally used for a simple and convenient system's performance comparison. The values in Figure 5 and Figure 6 were calculated using the evaluation tools provided by the organisers of TRECVID.

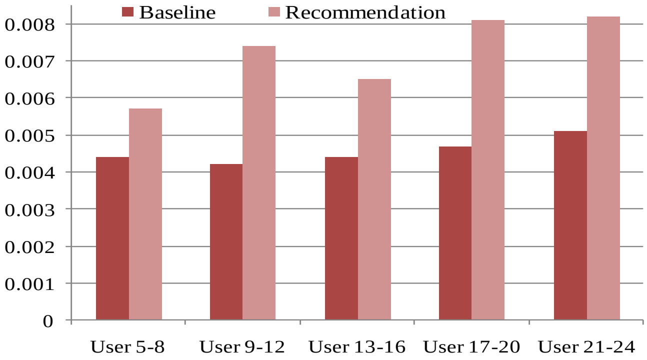


Figure 6: Mean Average Precision (MAP) for baseline and recommendation systems for different groups of users

Figure 7 shows the average time in seconds that it takes a user to find the first relevant shot for both the baseline and the recommender systems.

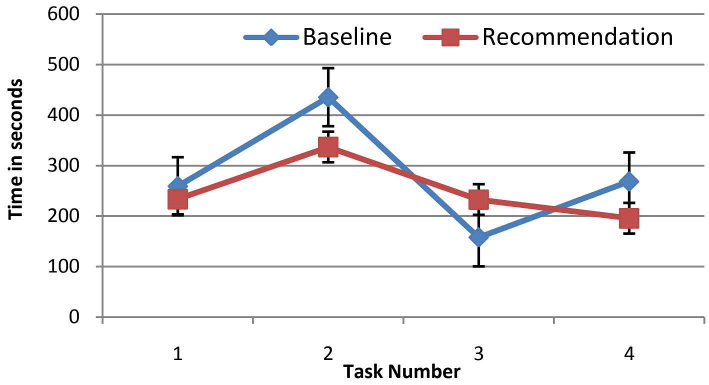


Figure 7: Average time in seconds to find first relevant shot for baseline and recommendation systems.

The results indicate that the system that uses recommendations outperforms the baseline system in terms of precision. It can be seen quite clearly from Figure 5 that the shots returned by the recommendation system have a much higher precision over the first 5-30 shots than the baseline system. We verified that the difference between the two P@N values for values of *N* between 5 and 100 was statistically significant using a pair wise t-test (p = 0.0214, t = 3.3045). Over the next 100-2000 shots the difference is negligible. However, it is unlikely that a user would view that number of shots; given that in total our 24 participants viewed 3034 shots (see Table 2), in the entire trial, 24 hours of video viewing. This demonstrates that the use of the implicit feedback can improve the retrieval results of the system, and thus be of greater assistance for the users.

Figure 6 shows that the MAP values of the shots the participants selected using the recommendation system are higher than the MAP values of the shots that the participants selected using the baseline system. We verified that the difference between the two sets of results was statistically significant using a pair wise t-test (p = 0.0028, t = 6.5623). The general trend is that the MAP values of the shots found using the recommendation system is increasing with the amount of training data that is used to propagate the graph based model. There is a slight dip in one group; however, this may be due to the small sample groups that we are using. The MAP values over all sessions for the baseline system was 0.0057, whereas for the recommender system it was 0.0082. These results show that participants are at the same time finding related, new and diverse relevant shots in the data set. However, these findings are not quite borne out by the recall values for the tasks. The recall for the tasks is quite low. In general the recall is low for all of the systems for all of the tasks at TRECVID 2006; the main focus is on the precision values. While recall is an important aspect we feel that it is more important that the users found accurate results and that they perceived that they had explored the collection, as they had found a heterogeneous set of results. While the results in Figure 5 and Figure 6 show that the users are seeing more accurate results and finding more accurate results, this is not telling the full story. In a number of scenarios users will just want to find just one result to satisfy their information need. Figure 7 shows that for three of the four tasks the users using the recommendation system find their first relevant result more quickly than the users using the baseline system. The one task for which the baseline system outperforms the recommendation system is due to the actions of two users who did not use the recommendations. We do not know why these two users did not use the recommendations, as they did utilise the recommendations for the other task which they carried out using the recommendation system. A closer examination of the users who did use the recommendations found that three users found relevant shots in less than one minute, none of the users using the baseline system managed to find relevant shots in less than a minute. Overall the difference in values is not statistically significant, but a definite trend can be seen.

The results presented so far have shown that users do achieve more accurate results using the system that provides recommendations. We measured P@N and MAP values; it has been shown that the recommendation system outperforms the baseline system, and that this difference is statistically significant. It can be seen that overall the system that is providing recommendations is returning more accurate results to the user. As a result of this, the users are interacting with more relevant videos and find more accurate results. In addition to this, users are finding relevant videos more quickly using the recommendation system (see Figure 7). This demonstrates the validity of our first assumption, that the performance of the users of the system, in terms of precision of retrieved videos, has improved with the use of recommendations based on implicit feedback. In the following sub-section we will discuss user exploration of the collection in more detail.

6.2 User Exploration

6.2.1 User interactions

As was outlined in our system description (Section 4) there are a number of ways that the participants could interact with our system. Once a participant enters an initial query in our system, all of the available tools may be used to browse or search the video collections. We begin our investigation of user exploration by briefly analysing these interactions. Table 2 outlines how many times each action available was used across the entire experimental group. During the experiments, the participants entered 1083 queries; many of these queries were unique. This indicates that the participants took a number of different approaches to the tasks, indicating that their actions were not determined by carrying out the same tasks. The figures in Table 2 also show that participants play shots quite often. However, if a video shot is selected then it plays automatically in our system. This makes it more difficult to determine whether participants are playing the videos for additional information or if the system is doing so automatically. To compensate for this we only count a play action if a video plays for more than 3 seconds. Another feature that was widely used in our system was the tooltip feature. The tooltip highlighting functionality allowed the users to view neighbouring keyframes and associated text when moving the mouse over one keyframe. This meant that the participants could get context and a feel for the shot without actually having to play that shot. This feature was used on average 42.3 (with a median of 38) times per participant per task when viewing a static shot.

Action Type	Occ.	Action Type	Occ.
Query	1083	Play (For more than 3 sec)	7598
Mark Relevant	1343	Browse keyframes	814
Mark Maybe Relevant	176	Navigate within a video	3794
Mark Not Relevant	922	Tooltip	4795
View	3034	Total Actions	23559

Table 2: Action type and the number of occurrences during the experiment

6.2.2 Analysis of Interaction Graph

In order to gain further insight into the user interactions a number of different aspects of the interaction graph were analysed. In particular we were interested in investigating changes in the graph structure as additional users used our system. These aspects include the number of nodes, the number of unique queries and the number of links that were present in the graph. Table 3 shows the results of this analysis. It can quite clearly be seen in Table 3 that the number of new interactions increases as the number of participants also increases. The majority of nodes in our graph are video shots (apart from query nodes), as the number of participants increases so does the number of unique shots that have been viewed. On further investigation of the graph and logs it was found that, overall, 49% of documents selected by users 1-12 were selected at least by one user in 13-24. Users 1-12 clicked 1050 unique documents, whereas users 13-24 clicked 596 unique documents. Also, users 1-12 produced 1737 clicks, whereas users 13-24 produced 1024. This can be interpreted as users 13-24 were satisfied more quickly than users 1-12. It was also found that the number of unique queries also increases with the additional users. These results give an indication that later participants are not just using the recommendations to mark relevant videos, but also interacting with further new and unique shots.

Users	Number of Nodes	Number of Queries	Number of Edges	Total Graph Elements
1-4	1001 (28.31%)	115 (18.51%)	2505 (23.09%)	3621 (24.13%)
1-8	1752 (49.56%)	258 (41.54%)	4645 (42.81%)	6655 (44.35%)
1-12	2488 (70.38%)	388 (62.48%)	7013 (64.63%)	9989 (66.57%)
1-16	3009 (85.12%)	452 (72.79%)	8463 (78%)	11924 (79.46%)
1-20	3313 (93.72%)	550 (88.57%)	9868 (90.95%)	13731 (91.5%)
1-24	3535 (100%)	621 (100%)	10850 (100%)	15006 (100%)

Table 3: Number of graph elements in graph after each group of four users.

6.2.3 Top Retrieved Videos

Figure 8 shows the probability that a particular shot is relevant plotted against a relevance value that is assigned to that document from our graph representation. The relevance value on the x-axis thus represents the sum of the weights of all of the edges leading to a particular node. The average interaction value was just 1.23, with irrelevant documents having an average value of 1.13 and relevant documents having an average of 2.94. This result is encouraging as it shows that relevant documents do receive more interaction from the users of the system. It can be seen that up until a certain point as the interactions from previous users increase so does the probability of the document being relevant. It was also found that for some of the documents with higher relevance values the probability tails off slightly. Further investigation found that there were two main reasons that a number of irrelevant documents had high relevance values. Firstly, there were shots that seemed relevant at first glance but upon further investigation were not relevant; however, for participants to investigate this required some interaction with the shot thus giving it a high interaction value. Secondly, there were a number of shots that appeared in the top of the most common queries before any recommendations were given, thus increasing the chances of participants interacting with those videos. It should also be noted that on average only 5.49% of nodes in the graph relate to relevant shots. This indicates that users are exploring and interacting with large portions of the collection that are not relevant, to help them find relevant shots. However, even with this kind of sparse and difficult data the performance of the users is improved with the recommendations presented to the users. It was found that as the amount of information in the graph increased so did the proportion of recommendations selected by users; users 5-8 selected 9.77% of the recommendations, whereas users 21-24 selected 18.67% of the recommendations.

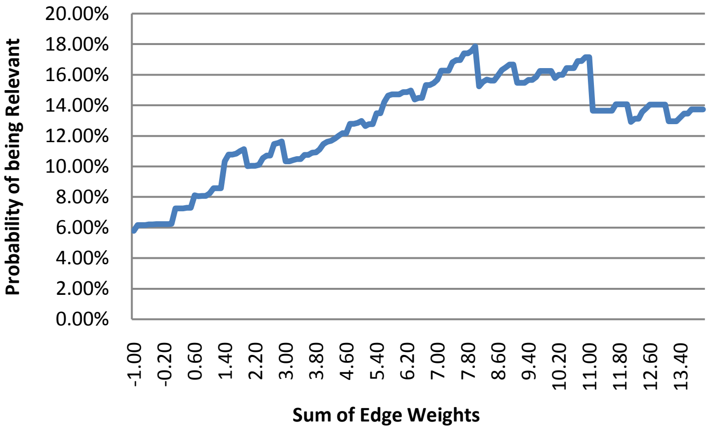


Figure 8: Probability of a document being relevant given a certain level of interaction. The y-axis represents the probability that the video is relevant and the x-axis represents the assigned interaction value in our graph.

6.2.4 Text Queries

In both the baseline and recommendation systems the participants were presented with query expansion terms that they could use to enhance their queries. We found however, that the majority of participants chose not to use the query expansion terms provided by the baseline system as they found them confusing. The query terms returned by the baseline system were stemmed and normalised and hence were not in the written form as users expected them to be, where as the queries recommended by the recommendation system were queries that previous users had used. One participant stated that "The query expansion terms didn't have any meaning." Another participant said that the "query expansion did not focus on real search task". This can be explained in part by specificities of some of the chosen topics, for example, in Task 1, when a user enters the name of a city ("New York") to get a shot of the city's sky line, the query expansion terms did not help to specify the search query. The top 5 queries for each topic are presented in Table 4. The top 15 unique terms across all four tasks are shown in Table 5. Across all 24 users a number of terms were repeated extensively. There were 130 unique terms and combinations of these were used to create a number of unique queries, on average the participants used 2.21 terms per query. However, it can be seen that across the 24 users and 4 topics there is relatively little repetition of the exact same queries, there were 621 unique queries out of 1083 total queries (57%). In fact only 4 queries occur 10 times or more, and they were all for the same task. This task had fewer facets to it than the others, and thus there was less scope for the users to use different search terms. This shows that despite the fact that users are carrying out the same task they are searching in differing ways, as the search tasks are multi-faceted and the participants are providing their own context. The results in this section indicate that the users explore the collection to a greater extent using the recommendations. Users of the system did not merely interact with videos that the previous users had interacted with, but instead could see what previous users had done and explore new video shots. Nodes were added to the graph of implicit actions throughout the evaluation (see Table 3). Also there was very little query repetition, and newer users used new and diverse query terms. These results give an indication that further participants are not just using the recommendations to mark relevant videos, but also interacting with further shots. These results also give an indication that we are achieving the second benefit of our approach; that users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered. However, this finding has not been fully validated. In order to do this we must analyse the users' use perceptions of the tasks, this analysis is presented in the following section.

Task 1		Task 2	
City	9	Jail	5
Building	8	prisoner guards	4
new York	8	Prisoner	4
tall buildings	8	Police	4
Tower	7	prisoner escorted	3
Task 3		Task 4	
Flag	8	Kiss	22
meeting flag	7	greeting kiss	20
Conference	5	Greeting	10
meeting	5	Kiss cheek	10
group flag	5	Cheek	6

Table 4: 5 most popular queries for each topic

Kiss	175	Police	60	City	37
Flag	101	People	50	Soldier	36
Prisoner	100	Tall	48	Greet	34
Greeting	72	Cheek	45	Four	31
Building	62	Meeting	39	Buildings	31

Table 5: 15 most commonly used keywords across all four tasks

6.3 User Perceptions

In order to provide further validation for our second hypothesis that “the users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered”, and to validate our third hypothesis, that “the users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search ”, we analysed the post task and post experiment questionnaires that our participants filled out.

6.3.1 Search Tasks

To begin with, we wished to gain insight into the participants’ perceptions of the two systems and also of the tasks that they had carried out. In the post-search questionnaires, we asked subjects to complete four 5-point semantic differentials indicating their responses to the attitude statement: “The search we asked you to perform was”. The paired stimuli offered as responses were: “relaxing”/“stressful”, “interesting”/“boring”, “restful”/“tiring” and “easy”/“difficult”. The average obtained differential values are shown in Table 6 for each system Each cell in Table 6 represents the responses for 20 participants (the four participants in the initial training set were not included as they did not use the recommendation system). The most positive response across all system and task combinations for each differential pair is shown in bold.

Differential	Baseline	Recommendation
Easy	1.9	**2.65**
Restful	**2.7**	2.575
Relaxing	2.725	**3.175**
Interesting	2.325	**2.75**

Table 6: Perceptions of search process for each system (Higher = Better)

The trends in Table 6 indicate that the users gave more positive responses for the recommendation system. It was found that the participants perceived some tasks as more easy, relaxing, restful and interesting than others. It can also be seen in Table 6 that there is a slight preference towards the system that provides recommendations amongst the participants. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks. We found that how easy and relaxing the participants found the tasks was system dependent (p< 0.14 and p<0.134 respectively for the significance of the system), whereas the user interest in the task was more dependent on the task that they were carrying out (p<0.194 for the significance of the system).

6.3.2 Retrieved Videos

In post search task questionnaires we also solicited subjects' opinions on the videos that were returned by the system. We wanted to discover if participants explored the video collection more based on the recommendations or if it in fact narrowed the focus in achievement of their tasks. The following Likert 5-point scales and semantic differentials were used 21:

- "During the search I have discovered more aspects of the topic than initially anticipated" (Change 1)
- "The video(s) I chose in the end match what I had in mind before starting the search" (Change 2)
- "My idea of what videos and terms were relevant changed throughout the task" (Change 3)
- "I believe I have seen all possible videos that satisfy my requirement" (Breadth)
- "I am satisfied with my search results" (Satisfaction)
- Semantic differentials : The videos I have received through the searches were: "relevant" / "irrelevant", "appropriate" / "inappropriate", "complete" / "incomplete", "surprising" / "expected".

Table 7 shows the average responses for each of these scales and differentials, using the labels after each of the Likert scales in the bulleted list above, for each system. The values for the four semantic differentials are included at the bottom of the table. The most positive response across all system and task combinations is shown in bold.

Differential	Baseline	Recommendation
Change 1	3.1	**3.5**
Change 2	3.475	**3.725**
Change 3	2.725	**3.05**
Breadth	2.625	**3.075**
Satisfaction	2.95	**3.4**
Relevant	1.925	**2.55**
Appropriate	3.125	**3.775**
Complete	2.225	**2.5**
Surprising	1.55	**1.725**

Table 7: Perceptions of the retrieval tasks for each system (Higher = Better)

The general trends that can be seen in Table 7 show that the users gave more positive responses for the recommendation system. It appears that participants have a better perception of the video shots that they found during their tasks using the recommendation system. It also appears that the participants believe more strongly that this system changed their perception of the task and presented them with more options. This would back up the findings in the previous section that the participants explored the collection to a greater extent when presented with the recommendations. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks to test these assertions. The initial ideas that the participants had about relevant shots were dependent on the task ($p < 0.019$ for significance of task). The changes in their perceptions were more dependent on the system that they used rather than the task, as was the participants belief that they had found relevant shots through the searches ($p < 0.217$ for significance of system). This demonstrates that the recommendation system helped the users to explore the collection to a greater extent, and also indicates that the users have a preference for the recommendation system. This finding strengthens the argument that our recommendation is providing benefits in terms of exploration and user perception.

6.3.3 System Support

We also wanted to determine the participants' opinion about the systems' support of their retrieval tasks. Firstly we asked them if the system had helped them to complete their task (satisfied). Participants were then asked to complete a further five 5-point Likert scales indicating their responses to the following statement: "The system helped me to complete my task because…". The criteria of the responses were:

- "explore the collection" (explore)
- "find relevant videos" (relevant)
- "detect and express different aspects of the task" (different)
- "focus my search" (focus)
- "find videos that I would not have otherwise considered" (consider)

Table 8 presents the average responses for each of these scales, using the labels after each of the Likert scales above for each system. The most positive response is shown in bold. Some of the scales were inverted to reduce bias in the questionnaires.

Differential	Baseline	Recommendation
Satisfied	3.3	**3.6**
Explore	3.775	**3.9**
Relevant	3	**3.4**
Different	2.925	**3.275**
Focus	2.625	**3.25**
Consider	3.075	**3.375**

Table 8: Perceptions of system support for each system (Higher = Better)

Once again it appears that participants have a better perception of the video shots that they found during their tasks using the system with recommendations, and that they believe the system helped them to explore the collection of shots more thoroughly using this system. We applied two-way analysis of variance (ANOVA) to each differential across both systems and the four tasks to test our hypotheses; none of the dependencies were significant. From our analysis of the results, however, there is a trend that the focus of the search, the ability to express different aspects of the task and the change in videos considered is more dependent on the task, rather than the system.

6.3.4 Ranking of Systems

After completing all of the tasks and having used both systems we attempted to discover whether the participants preferred the system that provided recommendations or the system that did not. The participants were asked to complete an exit questionnaire where they were asked which system they preferred for particular aspects of the task, they could also indicate if they found no difference between the systems. The participants were asked, “Which of the systems did you…”:

- “find best overall” (Best)
- “find easier to learn to use” (Learn)
- “find easier to use” (Easier)
- “prefer” (Prefer)
- “find changed your perception of the task” (Perception)
- “find more effective for the tasks you performed” (Effective)

The users were also given some space where they could provide any feedback on the system that they felt may be useful.

Differential	Baseline	Recommendation	No Difference
Best	2	**16**	1
Learn	2	7	**11**
Easier	2	5	**13**
Prefer	1	**17**	2
Perception	3	**11**	6
Effective	3	**14**	3

Table 9: Users preferences for the two different systems.

It can be seen clearly in Table 9 that the participants had a preference for the system that provided the recommendations. It is also encouraging that the participants found there to be no major difference in the effort and time required to learn how to use the recommendations that are provided by the system with recommendations. This indicates that users were more satisfied with the system that provides recommendation, thus realising the third goal of our system will be more satisfied with the system that provides feedback. Users have a definite preference for the recommendation system. 17 out of 20 users preferred the recommendation system, with one user preferring the baseline system. The participants also indicated in their post task questionnaires that the system that provided recommendations helped them to explore the task and find aspects of the task that they otherwise would not have considered, in comparison with the baseline system.

The results of our analysis have addressed all of the points of our hypotheses and have demonstrated that we have achieved our goals. The following section will provide a discussion of some follow up evaluations that were carried out to help validate and expand our results and findings.

6.4 Follow Up Evaluation

In order to expand on some of our results and to alleviate any of doubts surrounding the validity of our approach we performed a follow up evaluation. The goal of this evaluation was to validate our approach using related but not identical tasks. For this evaluation we used the same two systems that have been described earlier in this paper, the same dataset and the same experimental methodology; however we use four different tasks. Two of these tasks were related to tasks that had been carried out in the first evaluation, whereas two further tasks were not. The four tasks were:

- Find shots of one or more people entering or leaving a vehicle (Task 5, not related)
- Find shots of a natural scene – with, for example, fields, trees, sky, lake, mountain, rocks, rovers, beach, ocean, grass, sunset, waterfall, animals or people; but no buildings, no roads, no vehicles (Task 6, not related)
- Find shots of a skyline with tall buildings visible (Task 7, related)
- Find shots of a greeting between two or more people (Task 8, related)

Some of these tasks were from TRECVID 2006 and some were not, so we cannot perform all of the same evaluations that we have carried out for the main user-centred evaluation presented in this paper, as we do not have the ground truth data that we had available for the initial evaluation. However, we can get an indication of user task performance and perceptions, when users are not repeating the same tasks. The pool of implicit actions from the previous experiment was used to provide recommendations for this evaluation. Three independent persons judged the shots that were marked as relevant, so that we could perform some analysis. Even though we have a set of relevant shots for the tasks that were from TRECVID, the assessments of shots were performed on all four tasks, in order to maintain consistency. Four users carried out the new evaluation, as this evaluation was to validate findings to date and not to re-test the hypotheses.

After the experiment was completed it was found that for the two related tasks the users retrieved more video shots using the recommendation system in comparison with the baseline system. For the unrelated task the participants retrieved slightly less videos with the recommendation system, however the difference was not significant. In terms of precision, for one of the related tasks the precision of the results is increased three fold using the recommendation system, for the second related task the precision is slightly lower, in this case the difference was not significant. In terms of the unrelated tasks the precision was greater for one of the tasks with the recommendation system, and lower for the other, again this was not significant. The participants indicated in their post task questionnaires that the system that provided recommendations helped them to explore the task and find aspects of the task that they otherwise would not have considered. All of the participants had a preference for the recommendation system.

Some of the variations in these results may be due to using such a small sample of users, but overall the trends support the conclusions found in the first evaluation, without repeating the same tasks. It appears that overall the use of recommendations does not hinder performance on unrelated tasks, while still helping users with related but not identical tasks. These findings once again support the hypotheses that were made at the beginning of this paper.

8 Conclusions

In this paper we have presented a novel approach for combining implicit and explicit feedback from previous users to inform and help users of a video search system. The recommendations provided are based on user actions and on the previous interaction pool. This approach has been realised in a video retrieval system, and this system was tested by a pool of users on complex and difficult search tasks. There are a number of conclusions that can be made about using community based implicit feedback to provide recommendations. For the results of task performance, whether users retrieve more videos that match their search task, we measured P@N and MAP values. It was shown that the recommendation system outperforms the baseline system, in that the users of the recommendation system retrieve more accurate results overall and that this difference is statistically significant. It was also seen that users are finding relevant results more quickly using the recommendation system. These results validate our first hypothesis, that the performance of users of the recommendation system will improve with the use of recommendations based on implicit feedback. The statistics presented on user exploration, show that the users are pursuing the tasks sufficiently differently. They were able to explore the collection to a greater extent and find more relevant videos. Nodes were added to the graph of implicit actions throughout the evaluation, indicating that users are not just using the same queries and marking the same results, but they are exploring new parts of the collection (see Table 2). These results give an indication that further participants are not just using the recommendations to mark relevant videos, but also interacting with further shots. This also indicates that we have achieved our second goal; that users will be able to explore the collection to a greater extent, and also discover aspects of the topic that they may not have considered. In addition to demonstrating the validity of our second hypothesis, these findings also illustrate the validity of our approach and experimental methodology. The tasks that were chosen for the experiment were multi-faceted and ambiguous. As the tasks are multi-faceted we believed that participants would carry out their searches in differing ways and use numbers of different query terms and methodologies, thus providing their own context. This belief has been demonstrated by these findings. This second hypothesis was validated by our analysis of user perceptions of the system where the users gave an indication that the recommendation system helped them to explore the collection. The participants indicated in their post task questionnaires that the system that provided recommendations helped them to explore the task and find aspects of the task that they otherwise would not have considered, in comparison with the baseline system. It is also shown that the users have a definite preference for the recommendation system. 17 out of 20 users preferred the recommendation system, while one user preferred the baseline system. These findings indicate that we have achieved goal three of our hypothesis; that users will be more satisfied with the system that provides feedback, and also be more satisfied with the results of their search. These results successfully demonstrate the potential of using this implicit feedback to aid multimedia search, and that this area deserves further investigation to be fully developed. In addition to our main evaluation, a follow up evaluation was conducted. In this evaluation we asked users to once again use our baseline and recommendation systems, using the same graph of implicit actions, but performing two tasks related to the tasks from the original evaluation and two unrelated tasks. It was found that even for the unrelated tasks users still preferred the recommendation system, and that the recommendations could still potentially aid users in their tasks.

In conclusion, our results have demonstrated the huge potential of using a collection of implicit actions from a community to help relieve some the major problems that ordinary users have while searching for multimedia, thus presenting a potentially significant stride towards bridging the semantic gap [17].

Acknowledgements

This research work was partially supported by the EC under contracts: K-Space (FP6-027026) and SEMEDIA (FP6-045032).

References

1. Adcock, J., Pickens, J., Cooper, M., Anthony, L., Chen, F. and Qvarfordt, P.: FXPAL Interactive Search Experiments for TRECVID 2007. Paper presented at the 5th TREC Video Retrieval Evaluation (TRECVID) workshop, Gaithersburg, Maryland, 5-6 November 2007
2. Bentley, F., Metcalf, C. and Harboe, G.: Personal vs. commercial content: the similarities between consumer use of photos and music. In: Proc. SIGCHI conference on Human Factors in computing systems, Montréal, Québec, Canada, pp. 667-676, 2006
3. Chang, S.-F., Manmatha, R., and Chua, T.-S.: Combining Text and Audio-Visual Features in Video Indexing. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA, pp. 1005-1008, 2005

4. Christel, M.G, and Conescu, R.M.: Mining Novice User Activity in TRECVID Interactive Retrieval Tasks. In: Proc. International Conference on Image and Video Retrieval, Tempe, Arizona, USA, pp. 21-30, 2006
5. Christel, M.G.: Establishing the Utility of Non-Text Search for News Video Retrieval with Real World Users. In: Proc. ACM Multimedia, University of Augsburg, Germany, pp. 707-716, 2007.
6. Craswell, N. and Szummer, M.: Random walks on the click graph. In: Proc. Annual International ACM SIGIR Conference, Amsterdam, pp. 239-246, 2007
7. Foley, E., Gurrin, C., Jones, G., Lee, H., McGivney, S., O'Connor, N.E., Sav, S., Smeaton A.F. and Wilkins, P.: TRECVid 2005 Experiments at Dublin City University. Paper presented at the 3th TREC Video Retrieval Evaluation (TRECVID) workshop, Gaithersburg, Maryland, 14-15 November 2007
8. Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B. and Coyle, M.: Collecting Community Wisdom: Integrating Social Search and Social Browsing. In: Proc. International Conference on Intelligent User Interfaces, Honolulu, Hawaii, pp. 52-61, 2007
9. Goldberg, D., Nichols, D., Oki, B.M., and Douglas, T.: Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM. **35**(12), 61-70 (1992)
10. Golder, S.A. and Huberman, B.A.: Usage Patterns in Collaborative Tagging Systems. Journal of Information Science. **32**(2), 198-208 (2006)
11. Halvey, M. and Keane, M.T.: Analysis of Online Video Search and Sharing. In: Proc. ACM Conference on Hypertext and Hypermedia, Manchester, UK, pp. 217-226, 2007
12. Hancock-Beaulieu, M. and Walker, S. An evaluation of automatic query expansion in an online library catalogue. Journal of Documentation. **48**(4) 406–421 (1992)
13. Heesch, D., Howarth, P., Magalhaes, J., May, A., Pickering, M., Yavlinski, A., and Rueger, S. Video Retrieval Using Search and Browsing. Paper presented at the 13th Text REtrieval Conference, Gaithersburg, Maryland, 16-19 November 2007
14. Hopfgartner, F., Urban, J., Villa, R. and Jose, J. Simulated Testing of an Adaptive Multimedia Information Retrieval System. In: Proc. International Workshop on Content-Based Multimedia Indexing, Bordeaux, France, pp. 328-335, 2007
15. Hopfgartner, F. Understanding Video Retrieval. VDM Verlag (2007)
16. Huang, C.-W. Automatic Closed Caption Alignment Based on Speech Recognition Transcripts. Technical Report, University of Columbia, 2003
17. Jaimes, A., Christel, M., Gilles, S., Ramesh, S., and Ma, W-Y. Multimedia Information Retrieval: What is it, and why isn't anyone using it? In: Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval, Singapore, pp.3–8, 2005.
18. Kelly, D., and Teevan, J.: Implicit feedback for inferring user preference: A bibliography. SIGIR Forum. **32**(2), 18-28 (2003)
19. Kirk, D., Sellen, A. and Rother, C. and Wood, K.: Understanding photowork. In: Proc. SIGCHI conference on Human Factors in computing systems, Montréal, Québec, Canada, pp. 761-770, 2006
20. Kirk, D., Sellen, A., Harper, R., and Wood, K.: Understanding videowork. In: In: Proc. SIGCHI conference on Human Factors in computing systems, San Jose, California, USA, pp. 61-70, 2007
21. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology. **140,** 1-55 (1932)
22. Mei, T., Hua, X-S, Yang, L., Yang, S-Q and Li, S.: VideoReach: An Online Video Recommendation System. In: Proc. Annual International ACM SIGIR Conference, Seattle, WA, USA, pp.767-768, 2006.

23. Naphade, M., Smith, J.R., Tesic, J., Chang, J-S., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J.: Large-Scale Ontology for Multimedia. IEEE MultiMedia. **13**(3), 86-91 (2006)
24. National Institute of Standards Technology. NIST TREC Video retrieval Evaluation Online Proceedings. http://www-nlpir.nist.gov/projects/tvpubs/tv/pubs.org.html
25. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In: Proc. National CSCW Workshop pp. 165-173, 1994
26. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M.: Okapi at TREC-3. Paper presented at the 4th Text REtrieval Conference, Gaithersburg, Maryland, 1994
27. Salton, G. and Buckley, C.: Improving retrieval performance by relevance feedback. In: Readings in information retrieval, pp. 355–364, Morgan Kauffman, San Francisco, (1997)
28. Shardanand, U. and Maes, P.: Social Information Filtering: Algorithms for Automating “Word of Mouth”. In: Proc. SIGCHI conference on Human Factors in computing systems, Denver, Colorado, USA, pp. 210-217, 1995
29. Smeaton, A. F., Over, P., and Kraaij, W. 2006.: Evaluation campaigns and TRECVid. In: Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval, Santa Barbara, California, USA, pp. 321-330, 2006
30. Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., and Boydell, O.: Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine. User Modeling and User-Adaptated Interaction. **14**(5), 383-423 (2004)
31. Snoek, C., Worring, M., Koelma, D., and Smeulders, A.: Learned Lexicon-Driven Interactive Video Retrieval. In: Proc. International Conference on Image and Video Retrieval, Tempe, Arizona, USA, pp. 11-20, 2006
32. Snoek, C.G., Worring, M., Koelma, D.C.. and Smeulders, A.W.M.: A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. Transcations on Multimedia. **9**(2), 280-292 (2007)
33. Sparck Jones K., and Van Rijsbergen C.: Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. In: British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge (1975).
34. Spink, A, Greisdorf, H., and Bateman, J.: From highly relevant to not relevant: examining different regions of relevance. Inf. Process. Management **34**(5), 599–621 (1998)
35. Urban, J., Hilaire, X., Hopfgartner, F., Villa, R., Jose, J., Chantamunee, S., and Gotoh,Y.: Glasgow University at TRECVid 2006. Paper presented at the 4th TREC Video Retrieval Evaluation (TRECVID) workshop, Gaithersburg, Maryland, 13-14 November 2006
36. Wexelblat, A. and Maes, P.: Footprints: History rich tools for information foraging. ”. In: Proc. SIGCHI conference on Human Factors in computing systems, Pittsburgh, PA, USA, pp. 270-277, 1999
37. White, R., Bilenko, M. and Cucerzan, S.: Studying the use of popular destinations to enhance Web search interaction. In: Proc. Annual International ACM SIGIR Conference, Amsterdam, pp.159-166, 2007
38. Yang, B., Mei, T., Hua, X-S, Yang, L., Yang, S-Q and Li, M.: Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In: Proc. Annual International ACM SIGIR Conference, Amsterdam, pp.73-80, 2007