Search Trails using User Feedback to Improve Video Search

***Frank Hopfgartner** ***†David Vallet** ***Martin Halvey** ***Joemon Jose**

*Department of Computing Science,
University of Glasgow,
Glasgow, United Kingdom.

†Universidad Autónoma de Madrid,
Escuela Politécnica Superior Ciudad Universitaria de
Cantoblanco, 28049 Madrid, Spain.

{hopfgarf, halvey, jj} @ dcs.gla.ac.uk, david.vallet@uam.es

ABSTRACT

In this paper we present an innovative approach for aiding users in the difficult task of video search. We use community based feedback mined from the interactions of previous users of our video search system to aid users in their search tasks. This feedback is the basis for providing recommendations to users of our video retrieval system. The ultimate goal of this system is to improve the quality of the results that users find, and in doing so, help users to explore a large and difficult information space and help them consider search options that they may not have considered otherwise. In particular we wish to make the difficult task of search for video much easier for users. The results of a user 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.

Categories and Subject Descriptors

H.5.1 Multimedia Information Systems, H.5.3 Group and Organization Interfaces

General Terms

Experimentation, Human Factors.

Keywords

Video, search, collaborative, community, feedback, recommender, user studies.

1. INTRODUCTION

With the improving capabilities and the decreasing prices of current hardware systems, there are ever growing possibilities to store and manipulate videos in a digital format. In addition to this, with ever increasing broadband capabilities it is now feasible to view video online at home as easily as text-based pages were viewed when the Web first appeared. People now build 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 personal collection. However the systems that currently exist to organise and retrieve theses videos are not sufficient to deal with such large and rapidly growing volumes of video. In particular there is an ever increasing need to develop tools and techniques to assist users in the complex task of searching for video clips. Current state of the art systems rely on using annotations provided by users, methods that use the low level features available in the videos or on an existing representation of concepts associated with the retrieval tasks. None of these methods are sufficient enough to overcome the problems associated with video search (see Section 2.1 for a full discussion).

In order to alleviate some these problems associated with video search we have developed a video retrieval system that uses the actions involved in previous user searches to help and inform future users of the system, through video recommendation. Our system does not require users to alter their normal searching behaviour, provide annotations or provide any other supplementary feedback. We achieve this outcome by utilising the available information about user interactions. In addition to this, our system does not require a representation of the concepts in the video that the user wishes to retrieve, while still offering a work around for the problems associated with the semantic gap [11]. We believe that the use of this system can result in a number of desirable outcomes for users. In particular, improved user performance in terms of task completion, it can aid user exploration of the collection and can also increase user satisfaction with their search and their search results. An evaluative study was conducted, in order to examine and validate these assumptions. A baseline system that provides no recommendations was compared with our system that provides recommendations. The two systems and their respective performances were evaluated both qualitatively and quantitatively.

The remainder of this paper is organised as follows: We will provide a rationale for our work, and describe the state of the art in a video search. Subsequently, in Section 3 we will describe our approach for using implicit feedback to provide recommendations. Section 4 will describe two systems which were used in our study. In Section 5 we will then describe our experimental methodology, which is followed by the results of our experiments. Finally we will provide a discussion of our work and some conclusions.

2. BACKGROUND AND MOTIVATION

2.1 Interactive Video Retrieval

Interactive video retrieval refers to the process of users formulating and carrying out video searches, and subsequently

MM'08, October 26-31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-303-7/08/10…$5.00.

reformulating queries and results based on the previously retrieved results. As video is extremely rich content there are a number of different ways that users can query video retrieval systems. The use of the low-level features that are available in images and videos, such as colour, texture and shape to retrieve results, is one common approach. This approach is often used for query by example, where users provide sample images or video clips as examples to retrieve similar images or video clips. While this approach seems reasonable it also presents a number of problems. It requires a representation and extraction of all of the features required from all of the videos presenting issues of efficiency. Also the difference between the low-level data representation of videos and the higher level concepts users associate with video, commonly known as the semantic gap [11], provide difficulties. Bridging the semantic gap is one of the most challenging research issues in multimedia information retrieval today. In an attempt to bridge this semantic gap, a great deal of interest in the multimedia search community has been invested in search by concept. The idea is that semantic concepts such as "vehicle" or "person" can be used to aid retrieval; an example of this is the Large Scale Ontology for Multimedia (LSCOM) [14]. However query by concept also has a number of issues that hinder its use, it requires a large number of concepts to be represented and to date has not been deployed on a large scale for general usage.

Query by text is the most popular method of searching for video. It is used in many large scale video retrieval systems such as YouTube and GoogleVideo, and is also the most popular query method at TRECVID [2]. Query by text is simple and users are familiar with this paradigm from text based searches, in addition to this, query by text does not require a representation of concepts or features associated with a video. Query by text does however rely on the availability of sufficient textual descriptions of the video and its content. Textual descriptions in some cases may be extracted from closed captions or through automatic speech recognition; however a study of a number of state of the art video retrieval systems [10] concludes that the availability of these additional resources varies for different systems. Where these resources are available, they may not always be reliable, due to limitations in automatic speech recognition or language differences for example. More recent state of the art online systems, such as YouTube and Google Video, rely on using annotations provided by users to provide descriptions of videos. However, quite often users can have very different perceptions about the same video and annotate 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 an abundance of annotations unless there is some benefit to the user [7].

While each of these methods outlined above have problems, they have been used in conjunction with each other in a number of systems, including MediaMill [20] and Informedia [2]. These systems have been amongst the most successful systems at recent TRECVID interactive search evaluations. However, these top results are for "expert" users, who establish the idealistic performance of users [3], also a combination of these approaches requires a vast amount of metadata to be extracted and stored for each individual video clip.

As we have seen there are a number of different ways in which a user can query a video retrieval system; as has been shown these include query by text, query by example and query by concept. Each of these approaches have had limited success, and to date none of these approaches has provided an adequate solution to providing the tools to facilitate video search [2]. With this in mind we are proposing an approach that utilises the actions involved in previous users' searches to provide feedback to help future searches. This collaborative feedback based approach does not require any additional representation of video clips, unlike query by example, or any additional metadata, unlike query by concept or query by text, but instead uses the actions that users would carry out naturally while searching for video, to improve their search results.

2.2 Collaborative Information Access

Many of the earliest collaborative techniques emerged online in the 1990's [6], [15], [17]. Since those early days collaborative or community based methods have evolved and been used to aid browsing [22], e-learning [5] and in collaborative search engines [19]. These techniques rely on user feedback. 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 [25]. However, we believe that such techniques are insufficient where there is a lack of associated information [7] and will also suffer from problems associated with the semantic gap [11]. 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.

Traditionally explicit relevance feedback has been used to provide feedback for these methods; 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 [21] or when they are unfamiliar with the document collection [16]. Also previous evaluations have found that users of explicit feedback systems often do not provide sufficient levels of feedback for adaptive retrieval algorithms to work [8]. With this in mind in our system we concentrate on implicit relevance feedback.

Implicit feedback has been shown to be a good indicator of interest in a number of areas in IR [12]. Hopfgartner et al. [9] have suggested that implicit relevance feedback can aid users searching in digital video library systems. White et al. [23] use the concept of "search trails", meaning the search queries and document interactions sequences performed by the users during a search session, to enhance web search. Craswell and Szummer [4] apply a random walk on a graph of user click data, to help retrieve relevant documents for user searches. Liu et al. [13] used a graph representation based on the textual features associated with a video to improve result list ranking. Yang et al. [25] provide a multi-modal content-based video recommender system that's uses a combination of textual similarity over ASR and OCR data, visual similarity and aural similarity, in conjunction with relevance feedback. The relevance feedback is applied when fusing the multimodal rankings, to give more to a particular feature depending on negative and positive relevance example. For example, if a user search for "Mercedes" and clicks on recommended videos which share visual similarities, the recommendation system will give more weight on the visual feature for the recommendations of the current search session. This recommendation approach is content-based, whereas the approach that we use in this paper is based on click through data.

Using some of this previous work that uses click through data [23], [4] as a basis, we have developed our own graph based model of implicit actions and recommendation strategy, which we use to provide recommendations. This model is described in detail in a following section, but first we provide a description of our video retrieval system.

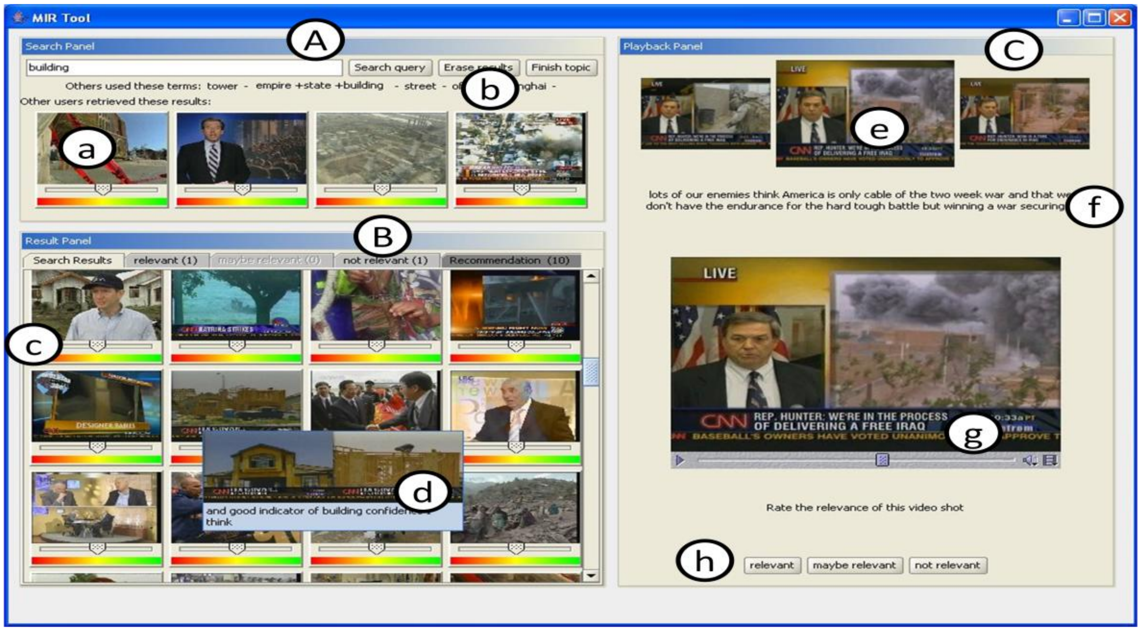
3. SYSTEM DESCRIPTION

Our collaborative 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 system consists of four main components, a search interface, a keyframe index, a retrieval engine and our recommendation model. The keyframes in our keyframe index were indexed based on automatic speech recognition transcript and machine translation output. The retrieval engine is based on the Okapi BM25 retrieval model, which was used to rank retrieval results that were returned to the user by text searches. 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 4 for details on the recommendation graph).

The interface for this system is shown in Figure 1 and can be divided into three main panels, search panel (A), result panel (B) and playback panel (C). The search panel (A) is where users formulate and carry out their searches. Users 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 or irrelevant by using a sliding bar (c). In the result panel additional information about each video shot can be retrieved. Hovering the mouse tip over a video keyframe, will result in that keyframe being 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 the 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 [18]; 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 recommendation system is the provision of keyframe recommendations (a).



Others used these terms: tower - empire +state +building - street - of nghai -

Other users retrieved these results:

and good indicator of building confiden 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 1: Interface of the video retrieval system

4. FEEDBACK BASED RECOMMENDATION: A GRAPH BASED REPRESENTATION

For the implementation of 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. [23]; however unlike White et al. [23] we do not limit the possible recommended documents to those documents that are at the end of the search trail. The reason for this is that 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. Thus the main difference between our search trail and that of White et al. [23] is that ours is a more complex representation. Similar to Craswell and Szummer [4], 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. The approach from Crasswell and Szummer [4] does not represent search trails, their approach is based on finding correlations query-clicked document. We use search trails because once again we want to recommend potentially important documents that are part of the interaction sequence. 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 [4], [23], [25], this results in a more full representation of a wide range of user actions that may facilitate better recommendations. In addition while these other approaches [4], [23] have been successful in other domains they have not been applied to video search. These properties and this approach results in two graph-based representations of user actions. The first uses 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 given by the LDM and represents it in such a way that is more easily exploitable for a recommendation algorithm. In our system the recommendations are based on three different analyses techniques based on the WDG. The two graph representation techniques and the recommendation techniques are described in detail in the following sections.

4.1 Labelled Directed Multigraph (LDM)

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 actions 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. Actions 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 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 the 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 fully connected, as it is possible 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.

4.2 Weighted Directed Graph (WDG)

In order to exploit the previous representation by our recommendation algorithm, 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 the interpreted 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 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:

$$\mathrm{w}_s(n_j) = \begin{cases} -1, & iff\ explicit\ irrelevance\ for\ n_j \\ (0,1) = lr(n_j), & implicit\ relevance\ for\ n_j \\ 1, & iff\ explicit\ relevance\ for\ n_j \end{cases}$$

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 [9], 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 $f(a): A\to\mathbb{N}$ mapping 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 were described in the system description (see Section 3). As was stated earlier these weights are based on previous work on implicit feedback for video search [9]. Figure 2 shows an example of LDM and its correspondent WDG for a given session.

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

Table 1: Values for f function for each action type used in the system.

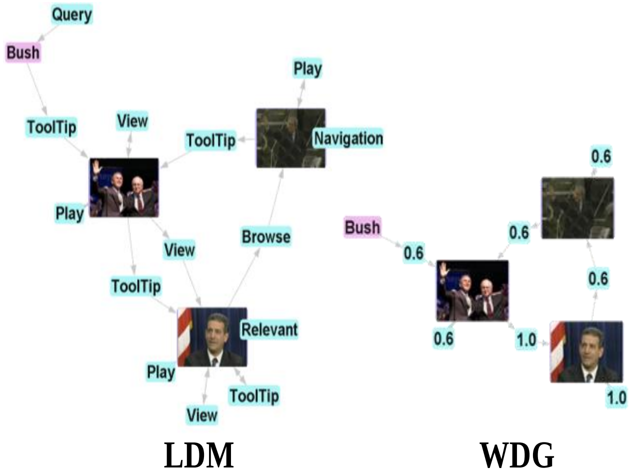


Figure 2: Node based graph representation vs. weight based representation for a search for "Bush"

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 the probabilities of all 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, which actions lead from node n_i to n_j, gave to node n_j. Figure 3 shows an example of implicit relevance pool.

4.3 Relevance Pool Based Recommendation

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 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:

$$NH(n)=\{n_1,\ldots,n_M|distance(n,n_m)<D_{MAX},n_m\in N\}$$

which are the nodes that are within a distance D_{MAX} of n, not taking link directionality into account. 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 buy 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 current session of the user and the implicit relevance pool we can then define the node recommendation value as:

$$nr(n,N_{s'})=\sum_{n_i\in N_{s'}} lr'(n_i)\cdot or(n)|n\in NH(n_i)$$

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 recommends using the documents instead. The last recommendation component is based on the users' 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:

$$ir(n,N_{s'})=\sum_{n_i\in N_{s'}}\left((lr'(n_i)\cdot\xi^{l-1}\cdot w)\middle|\begin{matrix}\exists\, p=n_i\rightsquigarrow n_j\to n\\ w\in\{n_j,n,w\}\\ l=length(p)\\ l<L_{MAX}\end{matrix}\right)$$

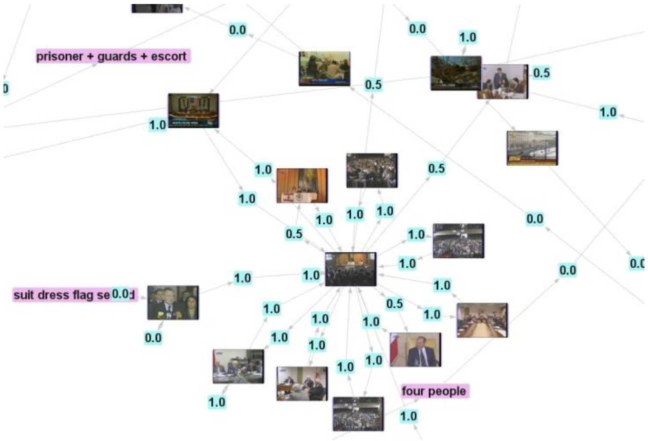


Figure 3: Graph illustrating implicit relevance pool

where p is the path between any node n_i and node n, taking into consideration the link directionality. l is the length of the path (counted as the number of links) and the distance is lower than a maximum length L_{MAX}. Finally, ξ is a length reduction factor, this was set to 0.8 in our system for all of our evaluations. This length reduction factor allows us to give more importance to those documents that directly follow the interaction sequence, however if a document with high levels of interaction occurs two or three steps away it will be recommended as well.

In a final step, we obtain the three recommendation lists from each recommendation component and merge them into a single final recommendation list. For this we use a rank-based aggregation approach, the scores of the final recommendations are the sum of the rank-based normalised score 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.

5. EXPERIMENTAL METHODOLOGY

5.1 Hypothesis

In order to measure the effectiveness of our proposed approach we conducted a user-centred evaluation. The goal of our evaluation was to investigate the effect of using community based implicit feedback to aid search in a video search paradigm. There are a number of potential benefits of our approach, which we would like to test:

- The performance of the system, in terms of precision of retrieved videos, will improve with the use of recommendations based on implicit 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, serendipitously.
- 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

With the purpose of determining the effects of implicit feedback users were required to carry out a number of video search tasks based on the TRECVID 2006 evaluations [18]. For our evaluation we focused on search tasks from the interactive search track. In 2006 the TRECVID collection contained 79,848 shots from English, Arabic, and Chinese news video. This data collection is noisy and hence the state of the art retrieval systems do not do achieve the same P/R as for text. In the TRECVID 2006 interactive search evaluations there were a total of 24 tasks. For our evaluation we are limiting the number of tasks that the users carry out to 4. Limiting the number of tasks allowed us to carry out more evaluations, as 24 individual search topics did not have to be carried out for each participant. For this evaluation we chose the four tasks for which the median precision in the 2006 TRECVID workshop was the worst. In essence these are the most difficult tasks. The four tasks were:

- Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible (Task 1)
- Find shots with one or more soldiers, police, or guards escorting a prisoner (Task 2)
- Find shots of a group including at least four people dressed in suits, seated, and with at least one flag (Task 3)
- Find shots of a greeting by at least one kiss on the cheek (Task 4)

The users were given the topic and a maximum of fifteen minutes to find shots relevant to the topic. The users could carry out text based queries. 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. In order 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, 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 each participant was allowed to carry out training tasks. 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. The purpose of using this experimental methodology was to validate our three hypotheses.

6. RESULTS

24 participants took part in our evaluation. The participants were mostly postgraduate students and researchers at a university. 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. The participants indicated that they regularly interacted with and searched for multimedia. They were paid a sum of £10 for their participation in the experiment, which took

approximately 2 hours. The results of the user trials were analysed with respect to our hypotheses that were given in the previous section. The evidence for and against each of these benefits is laid out in the following sections.

6.1 Task Performance

Since we were using the TRECVID collection and tasks, we were able to calculate precision and recall values for all of the tasks. Figure 4 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.

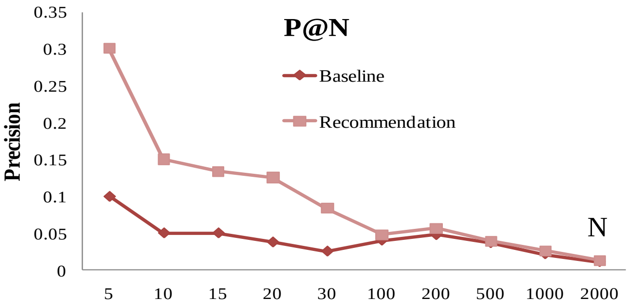


Figure 4: P@N for the Baseline and Recommendation Systems

The results show that the system that uses recommendations outperforms the baseline system in terms of precision. It can be seen quite clearly from Figure 4 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). It can also be seen over the next 100-2000 shots that 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, 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 to users.

Figure 5 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.

It can be seen quite clearly that the MAP of the shots that the participants selected using the recommendation system is higher than the MAP of the shots that the participants selected using the baseline system. We verified that the difference between the two sets of results were statistically significant using a pair wise t-test (p = 0.0028, t = 6.5623). The general trend is that the MAP 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. These results show that, as well as participants finding more related shots in the data set, that they are finding new and diverse relevant shots in the data set.

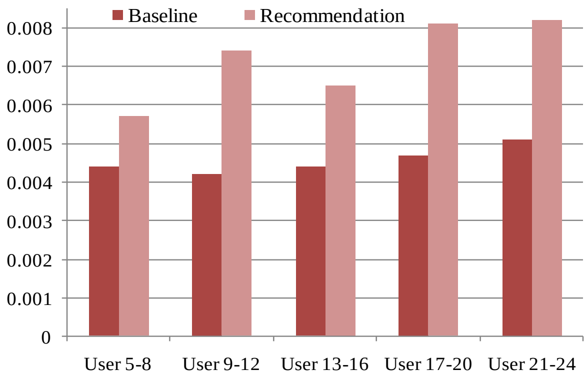


Figure 5: Mean Average Precision for Baseline and Recommendation Systems for Different Groups of Users

However, these findings are not quite borne out by the recall values for the tasks. Despite having higher precision values for the recommender system in comparison with the baseline, the recall for the tasks is still quite low. 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. For the 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. This demonstrates the validity of our first hypothesis. In the following section we will discuss user exploration of the collection in more detail.

6.2 User Exploration

6.2.1 Analysis of Interaction Graph

To begin our investigation of user exploration we analysed the graph of interactions. The number of nodes, the number of unique queries and the number of links that were present in the graph, at each stage where the graph had additional information for the previous four users added, were analysed. Table 2 shows the results of that analysis; it can be seen that the number of new interactions with the collection increases with the number of participants.

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

Table 2: Number of graph elements in graph after each group of four users.

The majority of nodes in our graph are video shots, 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 users 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 user 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 (see Section 6.2.2) with the additional users. These results give an indication that further participants are not just using the recommendations to mark relevant videos, but also interacting with further shots.

6.2.2 Text Queries

In both the baseline and the recommender based systems the participants were also 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 users expected them to be. Whereas the queries recommended by the recommender system were queries that previous users had used. One participant stated that "The query expansion terms didn't have any meaning." Another participants 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 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. In fact in the top ten queries for each task query expansions only occur twice, both for the same topic. 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. In fact only 4 queries occur 10 times or more, and they were all for the same task.

The results in this section indicate that the users explore the collection to greater extent using the recommendations. Later users 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 through out the evaluation (see table 2). Also there was very little query repetition, and newer users used new and diverse query terms. These results 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 analysed the user perceptions of the results and systems, this analysis is presented in the following section.

6.3 User Perceptions

In order to provide further validation for our second hypothesis and to validate our third hypothesis, we analysed the post task and post experiment questionnaires that our participants filled out.

6.3.1 Retrieved Videos

In post search task questionnaires we 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. Some of these are contradictions and some of the scales were inverted to reduce bias. The scales and differentials were: "I had an idea of which kind of videos were relevant for the topic before starting the search" (Initial Idea), "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) and the following Semantic differentials : The videos I have received through the searches were: "relevant" / "irrelevant", "appropriate" / "inappropriate", "complete" / "incomplete", "surprising" / "expected". Table 3 presents the average responses for each of these scales and differentials, using the labels after each of the Likert scales in the bulleted list above. The values for the four semantic differentials are included at the bottom of the table. The most positive response across for each system is shown in bold.

Differential	Baseline	Recommendation
Initial Idea	3.625	**4.175**
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 3: Perceptions of System (Higher = Better)

From the results in Table 3 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 Section 6.2 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 shot 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 recommendations are providing benefits in terms of exploration and user perception.

6.3.2 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. The participants 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) and "find more effective for the tasks you performed" (Effective). The users were also given some space to provide any feedback that they felt may be useful.

Differential	Recommendation	Baseline	Same
Best	**16**	2	1
Learn	7	2	**11**
Easier	5	2	**13**
Prefer	**17**	1	2
Perception	**11**	3	6
Effective	**14**	3	3

Table 4: User preferences for each system.

It can be seen clearly 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 further 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. In this section it has been seen that the users have a definite preference for the recommendation 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. Thus validating part three of our hypothesis, and also helping validate part two of our hypothesis. The results of our analysis have addressed all of the points of our hypotheses and have demonstrated that we have achieved our goals.

6.4 Follow Up Evaluation

In order to expand on some of our results 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 (see Section 3), the same dataset (see Section 5.2) and the same experimental methodology (see Section 5.3); however we use four different tasks. Two of these tasks were related to tasks that had been carried out in the first evaluation, and two further tasks that were not related. Some of these tasks were not from TRECVID 2006 so we cannot perform all of the same evaluations that we have 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 human judges judged the shots that were marked as relevant, so that we could perform some analysis. 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 difference 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. 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.

In another body of related work we simulated over 7200 search sessions using our graph based approach for recommendations and compared the results with the search trails approach from White et al. [23] and the random walk from Craswell and Szummer [4]. The full details of the experimental setup are available in Vallet et al. [24]. Figure 6 shows the MAP for each approach with respect to the number of elements that have been added to the graph.

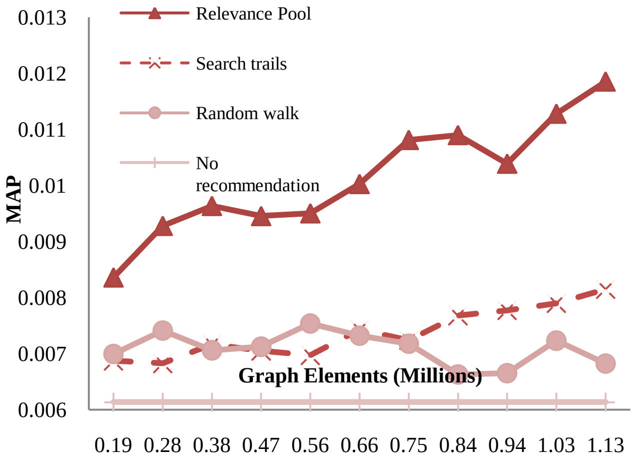


Figure 6: MAP for different recommendation approaches with respect to graph size

It can be seen quite clearly from Figure 6 that our approach (represented by relevance pool) outperforms the other approaches in terms of MAP. Figure 6 also illustrates the scalability of our approach; the relevance pool consistently gained performance as more users were added to the implicit graph (up to a test corpus of 1.25M total graph elements). For this evaluation we did not provide a direct comparison with other graph based work from Yang et al. [25] that was cited earlier in this paper. As was pointed out earlier (see Section 2.2) this recommendation approach is content-based, whereas ours is based on sole click through data much like the work of White et al. [23] and Crasswell and Szummer [4], so a direct comparison would not be appropriate. The following section will provide some final conclusions and a discussion of our findings.

7. DISCUSSION AND CONCLUSIONS

We have presented a novel video retrieval system, which uses feedback from previous users to inform and aid users of a video search system. The recommendations provided are based on user actions and on the previous interaction pool. 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 (see Section 6.1), 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. This demonstrates that the performance of users of the recommendation system will improve with the use of recommendations based on implicit feedback. The statistics presented in Section 6.2, 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. This indicates 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. This second hypothesis is further validated in Section 6.3 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. These results successfully demonstrate the potential of using implicit feedback to aid multimedia search, and that this area deserves further investigation to be fully developed. To this end we carried out some brief follow up experiments to investigate some of our findings further. It was shown that the recommendations are useful for related tasks, while not hindering unrelated tasks. These follow up evaluations demonstrated further uses of our approach, however there is future work that can be carried out. In particular, these techniques could be extended with other types of querying, e.g. query by example, to provide even more improved query results for users. In conclusion, the results of the evaluation, for our system that uses a collection of user actions has highlighted the promise of this approach to alleviate the major problems that users have while searching for multimedia, thus presenting a potential work around to the semantic gap [11] and other problems associated with video search.

8. ACKNOWLEDGEMENTS

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

9. REFERENCES

[1] Adcock, J., Pickens, J., Cooper, M., Anthony, L., Chen, F. and Qvarfordt, P. FXPAL Interactive Search Experiments for TRECVID 2007. In Proc. TRECVID 2007.

[2] Christel, M.G, and Conescu, R.M. Mining Novice User Activity in TRECVID Interactive Retrieval Tasks. In Proc CIVR 2006, 21-30.

[3] Christel, M.G. Establishing the Utility of Non-Text Search for News Video Retrieval with Real World Users. In Proc ACM MM 2007, 707-716.

[4] Craswell, N. and Szummer, M., Random walks on the click graph. In Proc. SIGIR 2007, ACM Press (2007), 239-246.

[5] Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B. and Coyle, M. Collecting Community Wisdom: Integrating Social Search and Social Browsing. In Proc. IUI 2007, ACM Press (2007), 52-61.

[6] Goldberg, D., Nichols, D., Oki, B.M., and Douglas, T. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM 35, 12 (1992), 61-70.

[7] Halvey, M. and Keane, M.T. Analysis of Online Video Search and Sharing. In Proc. ACM HT 2007, ACM Press (2007), 217-226.

[8] Hancock-Beaulieu, M. and Walker, S. An evaluation of automatic query expansion in an online library catalogue. Journal of Documentation 48, 4 (1992), 406–421.

[9] Hopfgartner, F., Urban, J., Villa, R. and Jose, J. Simulated Testing of an Adaptive Multimedia Information Retrieval System. In Proc. CBMI 2007, IEEE (2007), 328-335.

[10] Hopfgartner, F. Understanding Video Retrieval. VDM Verlag (2007)

[11] 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 MIR, ACM Press (2005), 3–8.

[12] Kelly, D., and Teevan, J. Implicit feedback for inferring user preference: A bibliography. SIGIR Forum 32, 2 (2003), 18-28.

[13] Liu, J., Lai, W., Hua, X-S., Huang, Y. and Li, S. Video Search Re-Ranking via Multi-Graph Propagation, In Proc ACM MM, 208-217.

[14] Naphade, M., Smith, J.R., Tesic, J., Chang, J-S., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J. Large-Scale Ontology for Multimedia. In IEEE MultiMedia 13(3), 2006, 86-91.

[15] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (1994).GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proc. CCSCW 1994, 165-173.

[16] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. Readings in information retrieval (1997), 355–364.

[17] Shardanand, U. and Maes, P. Social Information Filtering: Algorithms for Automating "Word of Mouth". In Proc. CHI 1995, ACM Press (1995), 210-217.

[18] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proc. MIR 2006, ACM Press (2006), 321-330.

[19] 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. UMUAI 14, 5 (2004). 383-423.

[20] Snoek, C., Worring, M., Koelma, D., and Smeulders, A. Learned Lexicon-Driven Interactive Video Retrieval. In Proc CIVR 2006, 11-20.

[21] Spink, A, Greisdorf, H., and Bateman, J. From highly relevant to not relevant: examining different regions of relevance. Inf. Process. Management 34, 5 (1998), 599–621.

[22] Wexelblat, A. and Maes, P. Footprints: History rich tools for information foraging. In Proc. CHI 1999, ACM Press (1999), 270-277.

[23] White, R., Bilenko, M. and Cucerzan, S., Studying the use of popular destinations to enhance web search interaction. In Proc. SIGIR 2007, ACM Press (2007), 159-166.

[24] Vallet, D., Hopfgartner, F., and Jose, J. Use of Implicit Graph for Recommending Relevant Videos: A Simulated Evaluation. In Proc ECIR 2008.

[25] 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. SIGIR 2007, ACM Press (2007), 73-80.